Horticulture Research

*Horticulture Research*, 2023, **10**: uhad058
https://doi.org/10.1093/hr/uhad058

Article

# Haplotype resolved chromosome level genome assembly of *Citrus australis* reveals disease resistance and other citrus specific genes

Upuli Nakandala[1,2], Ardashir Kharabian Masouleh[1,2], Malcolm W. Smith[3], Agnelo Furtado[1,2], Patrick Mason[1,2], Lena Constantin[1,2] and Robert J. Henry[1,2,*]

[1]Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane 4072, Australia
[2]ARC Centre of Excellence for Plant Success in Nature and Agriculture, University of Queensland, Brisbane 4072, Australia
[3]Department of Agriculture and Fisheries, Bundaberg Research Station, Bundaberg, Queensland 4670, Australia
*Corresponding author. E-mail: robert.henry@uq.edu.au

## Abstract

Recent advances in genome sequencing and assembly techniques have made it possible to achieve chromosome level reference genomes for citrus. Relatively few genomes have been anchored at the chromosome level and/or are haplotype phased, with the available genomes of varying accuracy and completeness. We now report a phased high-quality chromosome level genome assembly for an Australian native citrus species; *Citrus australis* (round lime) using highly accurate PacBio HiFi long reads, complemented with Hi-C scaffolding. Hifiasm with Hi-C integrated assembly resulted in a 331 Mb genome of *C. australis* with two haplotypes of nine pseudochromosomes with an N50 of 36.3 Mb and 98.8% genome assembly completeness (BUSCO). Repeat analysis showed that more than 50% of the genome contained interspersed repeats. Among them, LTR elements were the predominant type (21.0%), of which LTR Gypsy (9.8%) and LTR copia (7.7%) elements were the most abundant repeats. A total of 29 464 genes and 32 009 transcripts were identified in the genome. Of these, 28 222 CDS (25 753 genes) had BLAST hits and 21 401 CDS (75.8%) were annotated with at least one GO term. Citrus specific genes for antimicrobial peptides, defense, volatile compounds and acidity regulation were identified. The synteny analysis showed conserved regions between the two haplotypes with some structural variations in Chromosomes 2, 4, 7 and 8. This chromosome scale, and haplotype resolved *C. australis* genome will facilitate the study of important genes for citrus breeding and will also allow the enhanced definition of the evolutionary relationships between wild and domesticated citrus species.

## Introduction

Citrus is one of the most valuable fruit crops in the world and is widely grown in more than 100 countries under tropical, subtropical, and Mediterranean climatic conditions [1]. There are six citrus species, all of which are limes, that are native to Australia. One species is endemic to the Northern Territory whilst the other five species are primarily found in Queensland. Other populations of these species are found in New South Wales and South Australia [2]. *Citrus australis* is a slow growing, hardy plant which is naturally found in southeast Queensland [3]. *C. australis* is commonly known as Australian round lime, Dooja, Gympie Lime or Native lime. Characteristically, the trees are moderately frost tolerant, the fruits are globose or subglobose with pulp vesicles bearing large masses of oil and the seeds are monoembryonic. Although the raw fruits can be eaten, the fruits are more suitable to be used for the preparation of sauces, jams, cordials and as a flavouring agent [3].

Huanlongbing (HLB) or greening disease is caused by a vector-transmitted bacteria (*Candidatus Liberibacter*) leading to severe economic losses to the citrus industry around the world [4]. Most commercial citrus are known to be susceptible to HLB [5]. However, *C. australis* and other Australian native lime species including *Citrus australasica*, *Citrus glauca* and *Citrus inodora* and their derived hybrids have shown different degrees of resistance to HLB, providing highly valuable genetic resources in breeding HLB resistant cultivars, and for use as rootstocks or interstocks [6]. Recently, a novel class of small antimicrobial peptides (SAMPs) were isolated from *C. australasica* and other close relatives which can suppress the growth of HLB causing bacteria and promote host immunity in citrus [7]. However, SAMPs have not yet been identified in *C. australis* or other Australian wild limes which are resistant to HLB. The identification and characterization of genetic loci encoding the peptides and those conferring resistance against HLB is very important for the breeding of resistant citrus. Complete, high-quality genomes of resistant species will, therefore, provide enormous benefits in developing resistant cultivars.

High-quality reference genomes are a key resource for plant breeding, providing highly accurate prediction of genes in plants and supporting gene discovery [8]. PacBio circular consensus sequencing which generates Pacbio HiFi long reads (≥ 15 kb) with high base level accuracy (99.9%) outperforms the short reads and earlier long read technologies [Pacbio and Oxford Nanopore technology (ONT)] in assembling more contiguous genomes [9–11]. Hi-C based genomic scaffolding technologies have now enabled the

**Received:** 20 December 2022; **Accepted:** 27 March 2023; **Published:** 3 April 2023 **Corrected and Typeset:** 1 May 2023

generation of haplotype resolved chromosome level genomes in combination with Pacbio Hifi data [12]. Haplotype phasing which is still in its infancy provides unprecedented genomic resources to capture the structural variations of individual haplotypes and reveal haplotype specific variations regulating important traits. The loss of haplotype specific information in consensus genome assemblies limits their utility of informing breeding operations in highly heterozygous species [13].

Several citrus species have been sequenced and assembled over the past few years including two species (*Citrus limon* and *Citrus sinensis*) that have been anchored at the chromosome level and are haplotype resolved [14, 15]. However, reference genomes of Australian limes have not yet been reported. Here we present the *de novo* chromosome-scale haplotype resolved genome assembly and annotation of genes of *C. australis* which will be a valuable resource for citrus improvement through genomic-assisted breeding approaches.

# Results

## Comparisons of assembly size with k-mer approaches and flow cytometry estimates

Genome estimates varied depending on the k-mer value employed in the analysis. K-mer 21 was used for genomescope in one approach as it was the most widely executed k-value in other research [14, 16] and the recommended length by Genemoscope based on the computational accuracy and speed [17]. The k-mer depth distribution histogram was a bimodal profile which is the typical nature of heterozygous genomes with a short peak around 75X coverage and high peak around 150X coverage (Supplementary Fig. S1a). The estimated genome size with k-mer 21 was 297 Mbp with 0.503% heterozygosity. The kmergenie approach generated abundance histograms for different values of k ranging from k = 17–121 and predicts the best k value as 97. The predicted best k = 97 was then used for genome size estimation using genomescope which dramatically increased the genome size to 318 Mbp (Supplementary Fig. S1b). The results indicated that higher values of k resulted in higher genome sizes in genomescope. The genome size was estimated to be 340.5 Mb ± 0.5416% CV using flow cytometry (Supplementary Fig. S2). The estimated genome assembly size of the nuclear DNA content of the collapsed genome was 331.1 Mb (9 pseudochromosomes and unplaced scaffolds).

## Hifiasm assembly

The two PacBio SMRT cells yielded 30.6 Gb (90X) and 27.8 Gb (81X) of Hifi reads with Q32 median read quality (Supplementary Table S1). The genome assembly was performed using Hifiasm in 3 modes as mentioned in the methodology. The contig assembly generated from HiFi reads with default parameters (default option) produced a phased assembly with two phased haplotypes. BUSCO analysis revealed that the collapsed assembly covered 98.8% universal single copy genes with an N50 of 29.5 Mb. The assembly contained 4678 contigs with a total length of 485 Mb. The two phased haplotypes; hap1 and hap2 contain a total of 4410 and 1401 contigs respectively. The total lengths of the two phased haplotypes were 470 Mb and 380 Mb. Hap1 covered 95.1% of the single copy orthologs with an N50 of 29.4 Mb, whilst hap2 covered 96.7% single copy orthologs with an N50 of 27.1 Mb (Supplementary Table S2).

The primary/alternate mode generated a primary and an alternate assembly with HiFi reads. The primary assembly is comprised of 4637 contigs with 487 Mb assembly size. The BUSCO revealed 98.8% single copy orthologs for the primary assembly and the N50 was 29.7 Mb which was slightly higher than the contig assembly generated with default parameters. The alternate assembly was composed of alternate contigs that were discarded in the primary assembly. The alternate assembly was highly fragmented (N50 is 0.95 Mb) and did not cover most of the universal single copy genes (complete BUSCOs = 54.5%, missing BUSCOs = 42.9%) making it is less useful for further analysis (Supplementary Table S2).

Hifiasm generated a Hi-C integrated assembly, comprising of a collapsed and a pair of phased assemblies with paired end Hi-C reads in Hi-C mode. The collapsed/consensus assembly was made up of 4639 contigs with a total length of 486 Mb. It covered 98.8% complete BUSCOs with an N50 of 29.7 Mb. The assembly contiguity and completeness in the Hi-C mode were similar to the primary assembly and the contiguity was higher than that of the collapsed assembly generated from Hifiasm default option. The two phased haplotype assemblies generated from this option; hap1 and hap2 have 4410 and 1499 contigs respectively. The two haplotype assemblies covered 98.8% and 97.4% complete BUSCOs and have 29.4 Mb and 28.3 Mb N50s respectively. Assembly statistics of two individual haplotype assemblies revealed that the contiguity and completeness of Hi-C integrated assembly are greater in comparison to the Hifi reads only assembly (Supplementary Table S2).

The collapsed assembly from Hi-C mode can be characterized in four groups. There are 10 contigs greater than 13 Mb in length and 9 contigs greater than 1 Mb in length. The assembly had 36 contigs greater than 0.1 Mb and 4584 contigs less than 0.1 Mb. The dotplot analysis showed a lower sequence similarity between the phased haplotypes generated from Hifi reads only assembly (Fig. 1a), however a higher similarity between the phased haplotypes produced from the Hi-C partition options (Fig. 1b). Based on the assembly statistics, it is clear that the Hifiasm outputs the best contig assembly with the Hi-C partition options for phased haplotypes.

## Chromosomal scale pseudochromosome generation using hi-C data

The three contig assemblies was subjected to scaffolding to further understand assembly contiguity and completeness at the scaffold level. Hi-C proximity ligation libraries produced in two lanes generated a total of 656 M paired-end reads. Hi-C scaffolding of the first option (Hifi reads in default) generated 4663 scaffolds with 29.7 Mb N50 and 98.8% of complete BUSCOs. In the second option, the primary assembly generated 4618 scaffolds with 31.3 Mb N50 and 98.8% complete BUSCOs. The third option (Hi-C mode) generated 4642 total number of scaffolds with 31.3 Mb and 98.8% complete BUSCOs (Supplementary Table S3). Hi-C scaffolding revealed that the second and third options generated assemblies with similar contiguity which outperformed the first option where only Hifi reads were used in default.

Based on the assembly statistics and the ability to generate phased assemblies, we selected the Hi-C integrated Hifiasm assembly as the best assembly to be used for subsequent downstream analysis. The dotplot between Hi-C scaffolds against Hi-C integrated HiFiasm contigs assembly is shown in Fig. 2a. The 4642 scaffolds could be characterized into three subgroups including large scaffolds (from 24 Mb - 48 Mb), medium sized scaffolds (from 1.5 Mb - 4.5 Mb) and small sized scaffolds (less than 1.5 Mb). The scaffolding generated nine main pseudomolecules corresponding to the nine chromosomes with a total length of 311.5 Mb, an N50 of 36.3 Mb and 98.8% completeness of conserved single copy

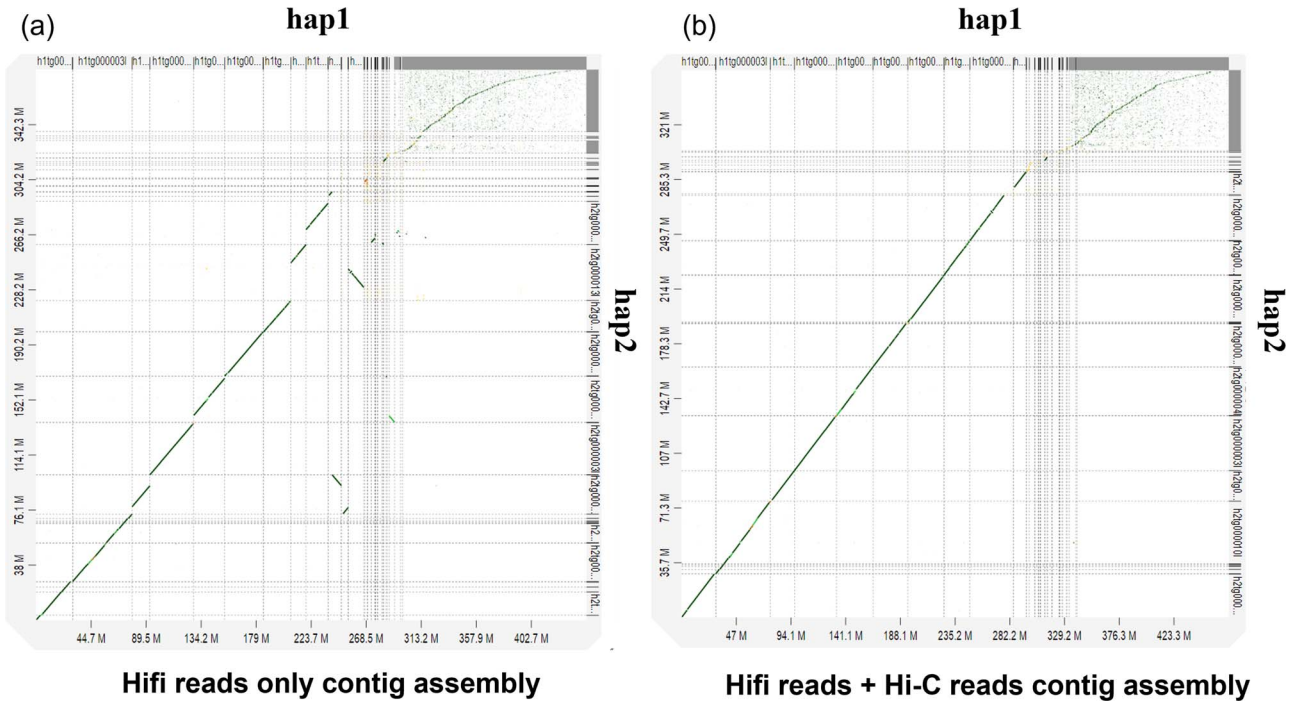


**Figure 1.** Phased haplotypes generated from two different options in Hifiasm. (a) hap1 vs hap2 generated only with Hifi reads in default. The sequence similarity was less between the two phased haplotypes. (b) hap1 vs hap2 generated with Hifi reads and Hi-C reads. The two phased haplotypes had high sequence similarities

orthologs. The pseudochromosome numbers were assigned to the corresponding scaffolds based on synteny between the present assembly of *C. australis* and other three genomes: *C. sinensis* (sweet orange) (Fig. 3a), *Citrus maxima* (pummelo) (Fig. 3b) and *C. limon* (lemon) (Fig. 3c).

The assembly containing the nine pseudo chromosomes was considered as the final version of the assembly (Fig. 4). Out of the nine pseudochromosomes, seven (Chr1, Chr2, Chr3, Chr4, Chr 5, Chr7, Chr9) were covered by one single HiFi contig and two (Chr6 and Chr8) were represented by two Hifi contigs (Fig. 2b, Supplementary Table S4). The nine pseudochromosomes were in the range between 24.8 Mb – 48.1 Mb, where chromosome five was the longest (48.1 Mb) and chromosome six was the shortest (24.8 Mb). Among the nine pseudomolecules, pseudomolecules one, three, five and seven had telomere repeats at both ends representing full chromosomes while the other pseudomolecules had telomere repeats only at one end. Pseudomolecules six and eight were spanned by two contigs and only one peripheral chromosomal region of the two pseudomolecules had telomeres (Supplementary Table S4). There were also five medium sized scaffolds which did not belong to the nine chromosomes ranging from 1.5 Mb – 4.5 Mb (Total size = 19.6 Mb) (Fig. 2a).

The scaffolds of individual haplotypes were assigned with chromosome numbers corresponding with the nine pseudo chromosomes of *C. australis* collapsed assembly based on dotplots (Fig. 5). Pseudo chromosome five was the largest of the two haplotypes (hap1–46.9 Mb, hap2–47.4 Mb) and the pseudochromosome 6 was the smallest (hap1–24.4 Mb, 24.6 Mb).

## Organelle genome analysis

Alignments of the complete *C. australis* chloroplast genome (Supplementary Fig. S3) with the small sized scaffolds showed that large parts of the scaffolds 27–4642 showed high similarities with the chloroplast genome (Supplementary Fig. S4c). Among the medium sized scaffolds, only scaffolds 12 and 13 contain small fragments of the chloroplast genome (Supplementary Fig. S4b). Sequence similarities with some parts of the top 9 scaffolds with the chloroplast genome indicate the insertion of chloroplast sequences within the nuclear genome (Supplementary Fig. S4a).

## Repeat identification and masking

Nine pseudochromosomes of the collapsed assembly were annotated for repeats and genes as nine chromosomes had the same number of BUSCOs as the whole genome (98.8%). Henceforth, we used the term “genome” for the 9 pseudochromosomes only without the small chloroplast contigs and any unplaced scaffolds. A large portion of the genome (52%) was comprised of interspersed repeats where most of them were unclassified. Among the classified transposable elements, LTR elements were the predominant type (21.0%). Among them, 21 192 regions were covered by LTR Gypsy elements accounting for the highest total size of the repetitive regions (30 530 861 bp) (9.8%) in the genome, followed by LTR Copia (23 997 689 bp) (7.7%). In addition to these dominant LTR elements, other LTR elements such as caulimovirus, ERV1, Ngaro, and pao were scattered throughout the genome with varied sizes. The total number of regions associated with Long Interspersed Nuclear Elements (LINEs) was 6205 accounting for a total of 5 269 998 bp (1.66%). There were no SINEs in the present genome. DNA transposons were present in smaller proportions (3.4%). The major types of DNA transposons were DNA/MULE-MuDR (1.3%) and DNA/hAT-Ac (0.8%). In addition to the transposable elements, a small proportion was composed of simple repeats (0.96), low complexity repeats (0.22%), small RNA repeats (0.1%) and satellites (0.01%). The highest number of repetitive regions were annotated in Chr5 24 034 718 (7.7%), which is the longest chromosome, and the least was recorded in Chr6, which is the shortest chromosome (3.92) (Table 01, Fig. 6, Supplementary Table S5; S6).

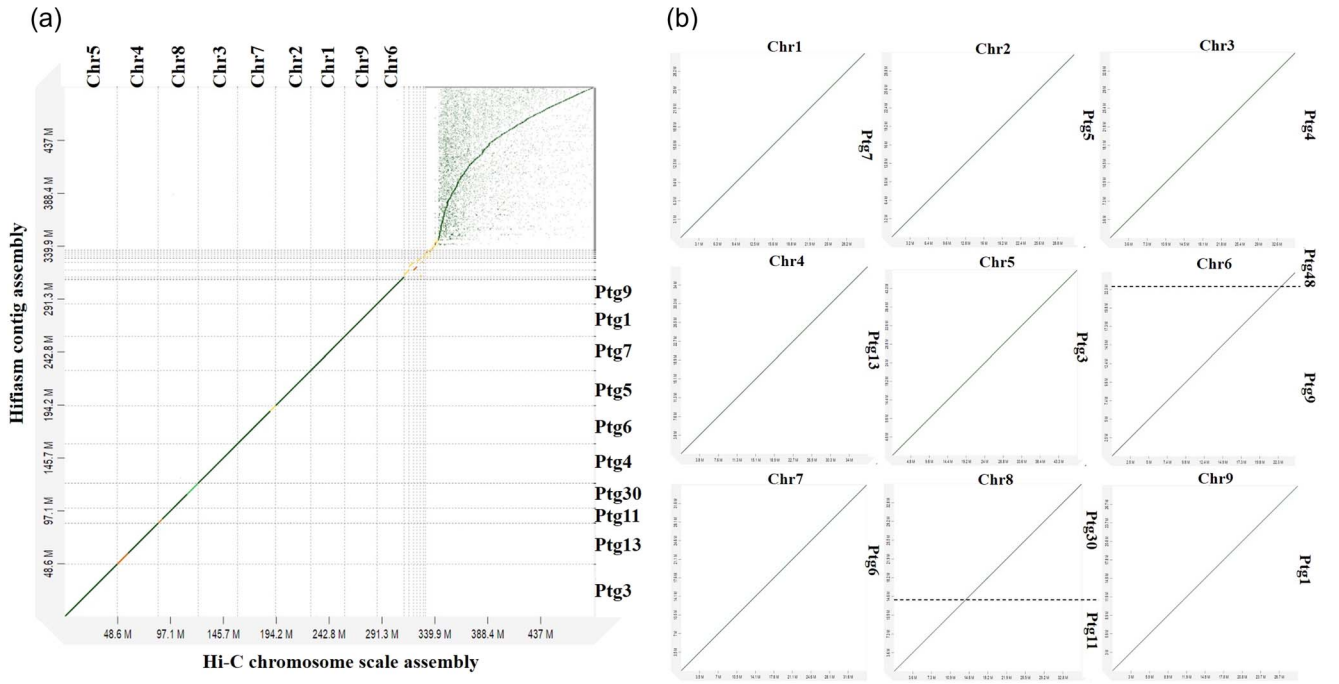


**Figure 2.** Hi-C chromosome scale pseudomolecules and the corresponding Hifiasm contigs. (a) The x-axis shows the chromosome scale assembly generated with Hi-C data. The y-axis indicates the contig assembly generated by Hifiasm with Hifi data and Hi-C paired end data. The lengthiest nine scaffolds are named with corresponding chromosome numbers based on the synteny analysis with previously published genomes. The middle set of scaffolds with the lengths of 1.5 Mb – 4.5 Mb contain large clusters of unknown repeats and are not assembled into 9 chromosomes (Total size = 19.6 Mb). Small scaffolds with lengths less than 0.1 Mb are chloroplast genome fragments inserted within the nuclear genome and are shown at the top right corner of the image. (b) Individual chromosomes and corresponding Hifiasm contigs. Chromosomes 6 and 8 are covered by two contigs whereas the other chromosomes are covered by one contig.

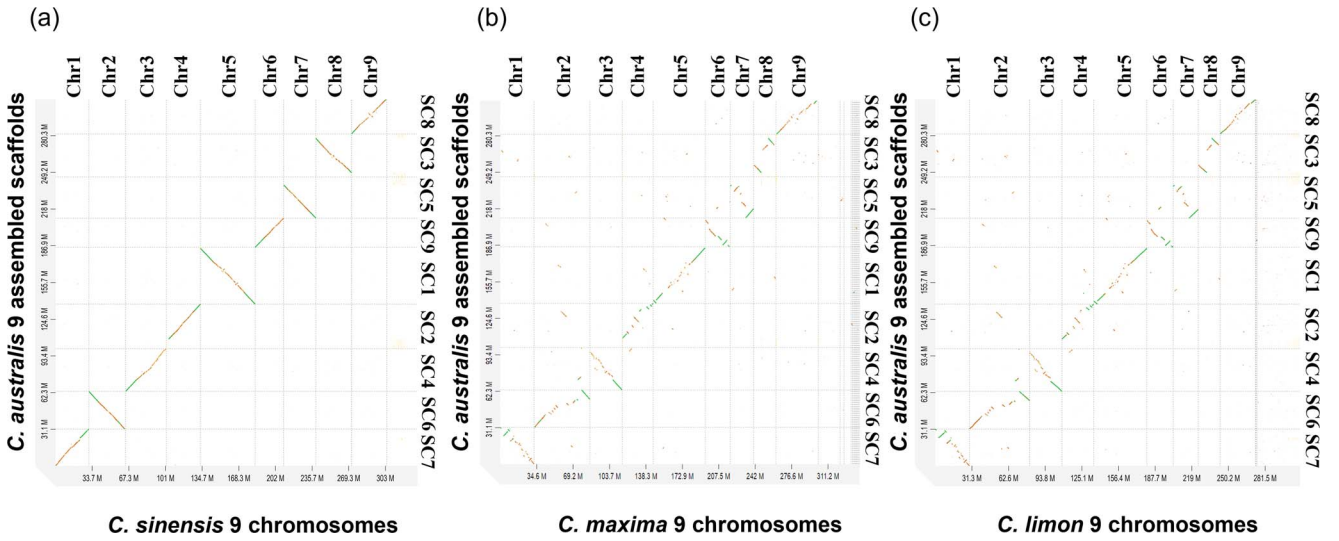


**Figure 3.** Dotplot analysis showing the synteny between *C. australis* assembled nine pseudomolecules and nine chromosomes of other three citrus genomes. a, *C. sinensis;* b, *C. maxima;* and c, *C. limon*. The sequence similarities between the corresponding pseudomolecules were used to rename *C. australis* scaffolds. Chr1, Chr2, Chr3, Chr4, Chr5, Chr6, Chr7, Chr8, and Chr9 correspond to scaffolds 7,6,4,2,1,9,5,3,8

A total of 165 Mbp (53.17%) of the genome was masked by repeat masking software. Hard masking and soft masking masked 165, 632, 936 bp (53.17%) of the genome. The hard masking with -nolow flag causes the software to hard mask all the repetitive regions excluding low complexity DNA such as Poly-purine or poly-pyrimidine stretches, or regions of extremely high AT or GC content and simple repeats accounting for 162, 287, 041 bp (52.10%) of the genome [18].

## RNA-seq read alignment

A total of 37.6 Gb (X110) in 250 million paired-end RNA-seq reads were mapped to the genome. Quality trimmed only RNA-seq reads and Quality and adapter trimmed RNA-seq reads were used for mapping to understand the effect of adapter trimming on overall alignment rates. The overall alignment rates for the quality trimmed only RNA-seq data with the unmasked, softmasked, hard masked with nolow option, and hard masked genomes were

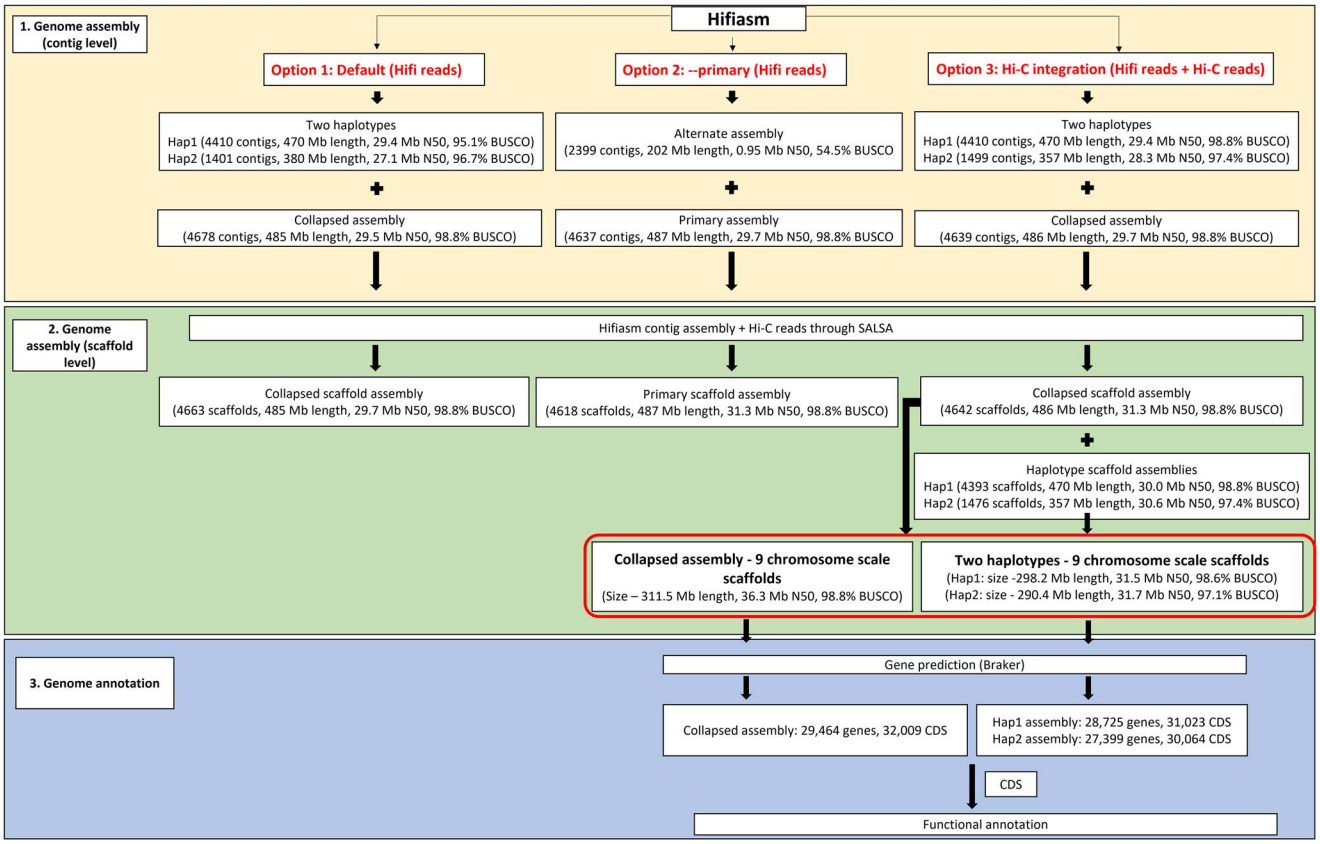


**Figure 4.** The summary of the genome assembly using Hifiasm De novo assembler and SALSA Hi-C scaffolder. The contig assemblies were generated in Hifiasm using three options (modes). In the first mode (default option), Hifi reads were used alone with built-in duplication parameters. This generated one collapsed assembly and two haplotypes. In the second mode (primary option), Hifi reads were used with –primary option generating a primary assembly and an alternate assembly. In the third mode (Hi-C integration mode), Hifi reads were used with Hi-C reads using Hi-C partition options in Hifiasm. This generated a collapsed assembly and two haplotypes. The assembly N50s were slightly improved in the second and third options. We used the collapsed contig assemblies generated by all 3 modes and the two haplotypes generated by Hi-C mode in Hifiasm for scaffolding. The scaffold assembly generated by the third option (Hi-C mode) was the best among all as those assemblies having a slightly improved N50. The assemblies had nine large chromosome scale scaffolds (24.8 Mb - 48.1 Mb for the collapsed genome, 24.4 Mb – 46.9 Mb for the haplotype 1 genome and 24.6 Mb – 47.4 Mb for the haplotype 2 genome) and they were considered as the final assemblies excluding the smaller contigs representing most of the chloroplast and mitochondrial genomes. The nine pseudochromosomes were subjected to structural and functional annotation.

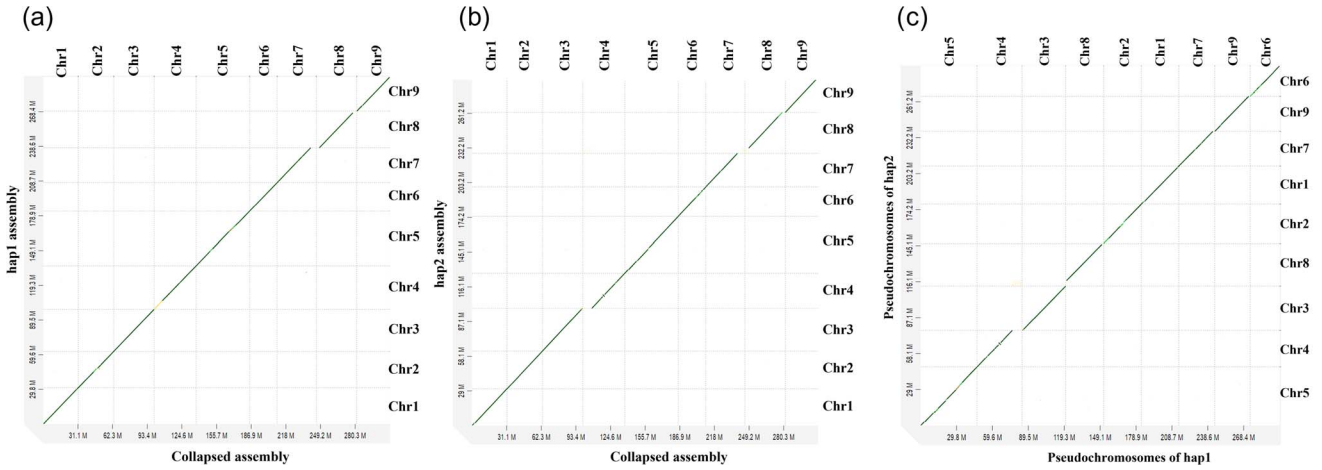


**Figure 5.** Pseudochromosome scale collapsed assembly vs pseudochromosome scale haplotype 1 and haplotype 2 assemblies of *C. australis*. (a) Collapsed assembly vs hap1 assembly (b) Collapsed assembly vs hap2 assembly (c) hap1 assembly vs hap2 assembly

60.6%, 60.6%, 52.2%, 50.8% respectively. The overall alignment rates of the quality and adapter trimmed RNA-seq data were improved in all the cases that were 79.1%, 79.1%, 68.1%, and 66.5% respectively.

## Gene prediction

Gene prediction was done for the unmasked and masked genomes to understand the impact of repeat masking on gene prediction. Quality trimmed only and quality and adapter

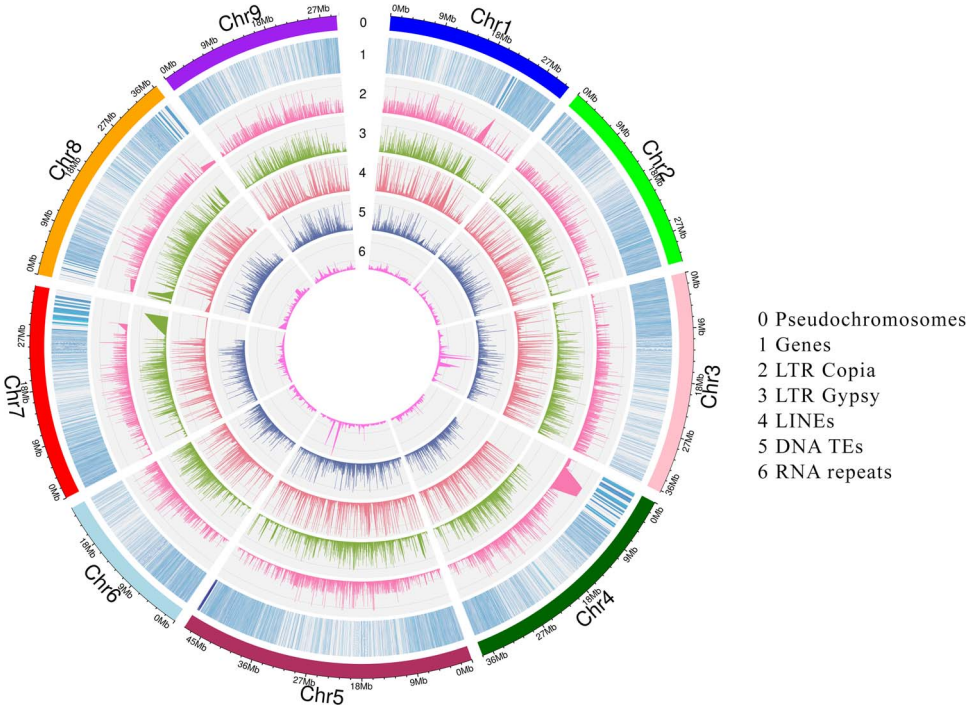


**Figure 6.** Characterization of repetitive regions and genes in *C. australis* genome. (0) Nine pseudochromosomes (Mb), (1) Regions of predicted genes (2) Regions of LTR Copia elements, (3) Regions of LTR Gypsy elements, (4) Regions of LINEs, (5) Regions of DNA TEs, (6) Regions of rRNA, tRNA and snRNA repeat regions.

trimmed RNA-seq data were independently used for gene prediction. Higher number of genes were predicted with quality and adapter trimmed RNA-seq data whereas low number of genes were predicted with quality trimmed only RNA-seq evidence (Supplementary Table S7). The higher number of genes is due to the higher alignment rates of quality and adapter trimmed RNA-seq data with the genome. The predicted number of genes with quality and adapter trimmed RNA-seq evidence for the soft masked genome was 29 464 (Fig. 6, Table 1). In the genome, some protein coding genes occupy repeat regions which will not be counted during the gene prediction with hard masking. Braker can still assess the gene sequences in repeat regions in a soft masked genome. Due to this, the softmasked genome is preferred over the hard masked genome for gene prediction. The highest number of genes was recorded in chromosome 5 (5240) while the lowest was recorded in chromosome 6 (2346) (Table 2). 98.6% of complete BUSCOs indicate the high completeness of the protein coding gene prediction (Table 1).

The total number of genes predicted for hap1 was 28 725 and for hap2 was 27 399. The total lengths of the haplotype assemblies (nine pseudochromosomes) differed by 7.8 Mb and the total number of genes differed by 1326. The size of the hap1 is 13.3 Mb shorter than the collapsed assembly and the total number of genes differed by 739 between the collapsed and hap1 genomes. The size of the hap2 was 21.1 Mb shorter than the collapsed assembly and the gene number differed by 2065 between the hap2 and collapsed genomes. (Table 2).

Synteny blocks were detected between the nine chromosomes of the two haplotypes (Fig. 7a). This analysis revealed that most of the genes were conserved in the two haplotypes across all the chromosomes. However, a few gene blocks of hap1 Chr4 and Chr7 also showed synteny with hap2 Chr8 indicating translocations of genes between the two sub-genomes. The Fig. 7b shows the synteny of the putative homologous genes between hap1 and hap2. The straight diagonal line with some interruptions along the nine chromosomes indicates that most of the genes in the two genomes are in a syntenic arrangement with some structural variations (inversions in Chr2 and Chr4) between the two genomes (Fig. 7b). The whole genome alignment at nucleotide level (Fig. 7c) further confirmed a syntenic pattern at the whole genome level, except for Chr4 and Chr8 with structural differences in the terminal regions. The highest length difference between the two haplotypes was recorded for Chr4 and Chr8 may be due to the structural variations in Ch4 and Ch8 (Table 2).

**Table 1.** Summary of BUSCO, repeat identification and gene prediction statistics for the collapsed genome

| Genome annotation | Annotation feature | Value |
|---|---|---|
| Annotation completeness | Complete BUSCO | 98.6% |
| Repeat identification | LINEs | 1.66% |
| | LTR elements | 21.01% |
| | DNA elements | 3.38% |
| | Unclassified | 25.95% |
| | Small RNA | 0.1% |
| | Satellites | 0.01% |
| | Simple repeats | 0.96% |
| | Low complexity | 0.22% |
| Gene prediction | Number of genes | 29 464 |
| | Number of CDS | 32 009 |
| | Number of single exons genes | 6700 |
| | Mean gene length | 2848 |
| | Mean CDS length | 1240 |
| | Mean exon length | 211 |
| | Number of exon in cds | 187 592 |
| | Longest gene size (bp) | 90 618 |
| | Shortest gene size (bp) | 201 |

**Table 2.** Sizes of the pseudochromosomes of collapsed genome and two haplotypes and the numbers of genes in each genome

| Chromosome number | Collapsed | | hap1 | | hap2 | |
|---|---|---|---|---|---|---|
| | Size (Mb) | Genes (total = 29 464) | Size (Mb) | Genes (Total = 28, 725) | Size (Mb) | Genes (Total = 27, 399) |
| 1 | 31.3 | 2876 | 31.1 | 2848 | 30.7 | 2864 |
| 2 | 32 | 3357 | 31.4 | 3382 | 31.7 | 3392 |
| 3 | 36.3 | 2974 | 36.1 | 2918 | 35.8 | 3018 |
| 4 | 37.8 | 3664 | 37.4 | 3864 | 29.4 | 2743 |
| 5 | 48.1 | 5240 | 46.9 | 5165 | 47.4 | 5077 |
| 6 | 24.8 | 2346 | 24.4 | 2405 | 24.6 | 2414 |
| 7 | 35.1 | 3436 | 30 | 3040 | 28.1 | 2723 |
| 8 | 36.4 | 3172 | 31.5 | 2709 | 34.4 | 2735 |
| 9 | 29.7 | 2399 | 29.4 | 2394 | 28.3 | 2433 |
| **Total size (Mb)** | **311.5** | | **298.2** | | **290.4** | |

## Non-coding genes prediction

Barrnap tool identified a large 5S ribosomal block in chromosome 6. A small number of 5S rRNA genes were predicted in Chr3, Chr4 and Chr7. No rRNA genes were found in chromosome 2. Chr1 only had 18S rRNA genes. 5S, 5.8S, 18S rRNA genes in Chr5, 5S, 5.8S, 18S and 28S rRNA genes in Chr6 and Chr8 and 5.8S, 18S and 28S rRNA genes in Chr9 were predicted by the tool.

## Functional annotation

Of the total number of predicted CDS by Genemark trained Augustus (32009), BLAST hits were obtained for 28 222 CDS (25 753 genes). The highest number of BLAST top hits (21, 642) were from *C. sinensis*. In addition, more than 8422 Top BLAST hits were from *Citrus clementina* and more than 5501 hits were from *Citrus unshiu*. A very small percentage of hits (537) were from other species (Supplementary Fig. S5A). The best hit with the smallest E value (below E $10^{-9}$) of all the annotations were used to describe the predicted genes. Among the transcripts with BLAST hits, 21 401 CDS (75.83%) were annotated with at least one GO term (Supplementary Fig. S5B). GO and Enzyme code distributions are given in Supplementary Fig. S5C; S5D. The coding potential assessment for BLAST hits with no BLAST hits [3808 CDS (3728 genes)] using *Arabidopsis thaliana* models revealed eight non-coding transcripts (8 genes) and 3800 coding transcripts. Based on citrus models, we identified 52 non-coding transcripts (48 genes) and 3756 coding transcripts. The predicted number of total non-coding transcripts was 54 corresponding to 50 genes from both models after removing the redundant transcripts.

## Important genes in citrus

### *Antimicrobial peptides*

A novel class of relatively short stable antimicrobial peptide (SAMP) with 67 amino acids (aa) has recently been detected from two wild Australian citrus species and a few citrus relatives [7]. The BLAST search identified two homologous genes in *C. australis* encoding a stress-response A/B barrel domain-containing protein HS1 (Supplementary Fig. S6). The first gene g9664 residing in chromosome 9, has two transcripts, one encoding 153 aa (g9664.t1) with 37% sequence similarity and the other transcript encoding 192 aa (g9664.t2) with 31% similarity with the 67 SAMP sequence previously characterized from resistant species. The sequence alignment showed that a substantial portion of 67 SAMP is present as a part of these two long peptides produced by *C. australis* with few amino acid substitutions and insertions (Supplementary Fig. S7a).

Another gene, g2059 on chromosome 8, encodes a 114 aa peptide (stress-response A/B barrel domain-containing protein HS1) with 21% sequence similarity with 67 SAMP (Supplementary Fig. S7b). HLB susceptible species such as *C. clementina* and *C. sinensis* have A/B barrel domain-containing protein HS1 with long amino acid sequences with different lengths (*C. clementina* – 118 aa, and 114 aa, *C. sinensis* – 109 aa, 114 aa, 126 aa, 175 aa). The *C. clementina* and *C. sinensis* proteins (114 aa) are identical to that of *C. australis* (114 aa) encoded by g2059 gene except for one SNP indicating that this is a common antimicrobial protein (AMP) present in these plants.

## Other defense related genes

Other defense-related genes that were annotated in *C. australis* genome, might potentially be involved in HLB resistance (Fig. 8). Genes encoding 17 guanine nucleotide-binding proteins (Supplementary Table S8), 13 pathogenesis-related proteins (Supplementary Table S9), and 76 Leucine rich repeat (LRR) proteins (Supplementary Table S10) were identified in the genome.

## Citrus acidity genes

We identified the homologs of two key acidity controlling genes; *PH1*(P-type Mg(2+) transporter) and *PH5* (P-type H(+)-exporting transporter) in *C. australis* genome by searching the sequence similarity with the corresponding protein sequences of *C. sinensis* for *PH1* (Cs1g20080) and *PH5* (Cs1g16150) [19] (Supplementary Table S11). The genes homologous to *PH1* and *PH5* were present on chromosomes 1 and 5 respectively. In addition to *PH5*, there were nine other genes encoding P-type H(+)-exporting transporters on chromosomes 4 and 6. PH1 protein of *C. australis* is distinguishable from other cultivated citrus species with a sweet flavor such as *C. sinensis*, *C. clementina* and *Citrus reticulata* by a few amino acid substitutions (Supplementary Fig. S8). The different expression levels of *PH1* and *PH5* are thought to be regulated by the transcription factors CitAN1 (basic helix–loop–helix transcription factor family protein), CitPH3 (WRKY transcription factor 44 isoform X1) and CitPH4 (R2R3-MYB family transcription factor). *C. australis* homologs of these genes were identified using the respective protein sequences of *C. sinensis* and they were present in one copy in the genome and encode only one transcript except for PH3 which had two transcripts.

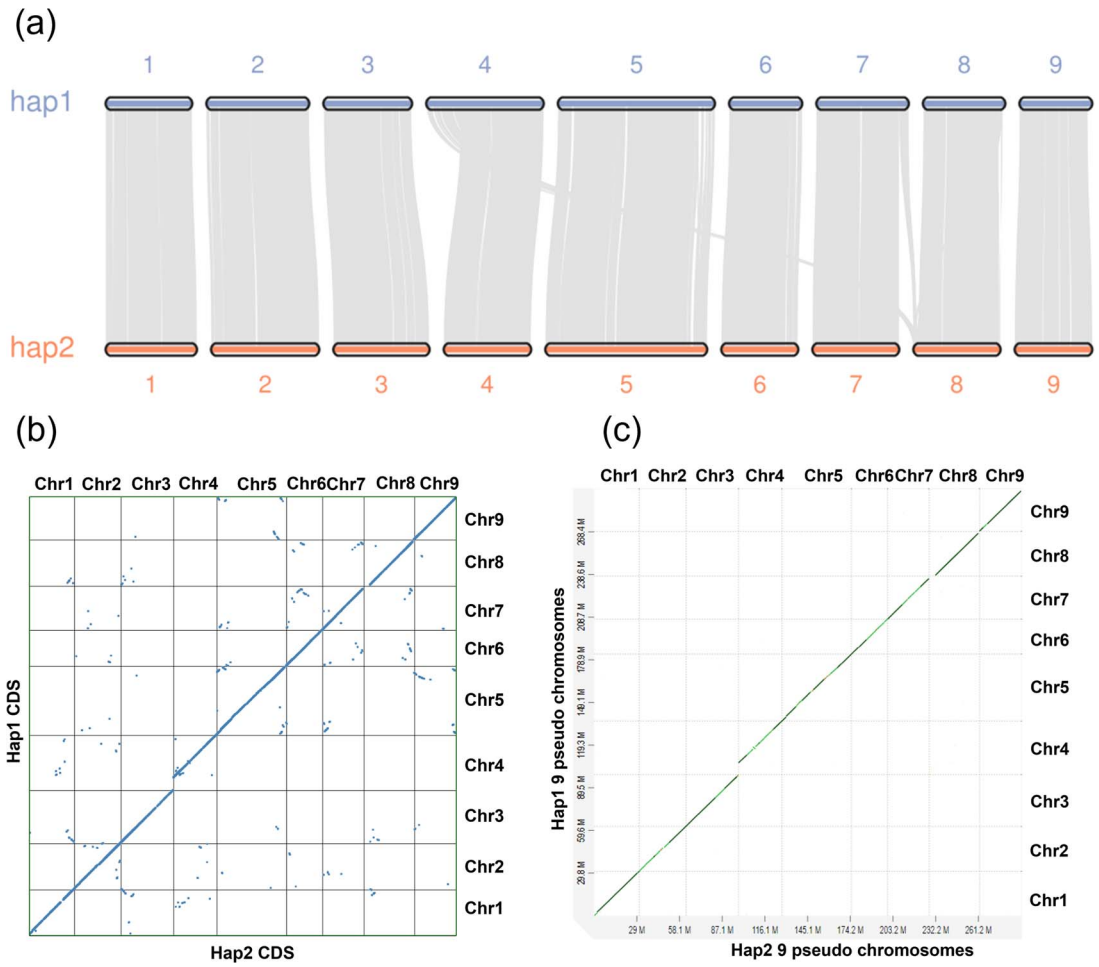


**Figure 7.** Synteny between hap1 and hap2 genomes. (a) Chromosomal synteny blocks of hap1 and hap2 genomes. Each colored block represents a chromosome. Grey lines extending from one region to another within or between chromosomes indicate the synteny blocks identified by MCScan. Many syntenic blocks are present within each chromosome of the two sub genomes revealing intragenomic similarity. Some syntenic blocks are also found between Chr8 of hap2 and Chr4 and Chr7 of hap1 which might represent translocations of homologous genes. (b) Whole protein coding genes syntenic dotplot. The black color vertical and horizontal lines delineate hap1 and hap2 CDS respectively. Each blue color dot represents a putative homolog. Each diagonal line per one chromosome shows the synteny between the two haplotypes at gene level. A few breaks at one terminal position of Chr4 and Chr8 indicate structural differences in the two chromosomes. Inversions at small regions of Chr2 and Chr4 of the two genomes could be seen. (c) Whole genome dotplot at nucleotide level. Nine chromosomes of the two haplotypes are in a syntenic arrangement with some structural differences between Chr4 and 8.

The citric acid level of juice sacs is also determined by the degradation of citric acids by a combination of enzymes including cytosol and mitochondrial aconitase (CitAco), NADP-isocitrate_dehydrogenase (CitIDH), and Glutamine synthetase (CitGS) and Glutamate decarboxylase (CitGAD). We identified three genes of aconitate hydratase protein CitAco3 (g19909 – Chr4, g22425 - Chr2, g17125 – Chr1) each encoding one transcript. Three CitIDH genes (g5386 – Chr3, g11983 – Chr9, g21560 – Chr2), three CitGS genes (g10022 – Chr9, g12494 – Chr6, g7198 – Chr7) and two GAD genes (g22832 – Chr2, g26170 – Chr5) in *C. australis* genome (Supplementary Table S11, Supplementary Fig. S6).

## Volatile compounds synthetic genes

We identified the key genes governing the synthesis of terpenoids in *C. australis* via the two main terpenoid biosynthetic pathways; the mevalonate (MVA) pathway and the non-mevalonate (MEP/-DOXP) pathway (Fig. 9, Supplementary Fig. S6, S9). There was one locus on Chr 2 encoding acetyl-CoA C-acetyltransferase, two loci for 3-hydroxy-3-methylglutaryl-CoAsynthase-2 (HMGS) on Chr 5 and 9, two loci for hydroxymethylglutaryl-CoA Reductase (HMGCR) on Chr 8 and 3, one locus for Mevalonate kinase (MVK) on Chr 3, three loci for phosphomevalonate kinase (PMK) on Chr 3, 2 and 5, one locus for diphosphomevalonate decarboxylase on Chr 7, 19 loci for geranyl-diphosphate:isopentenyl-diphosphate geranyltrans-transferase (GGPS) on Chr 3, 6, 4, 8, and 1, five loci for 1-deoxy-D-xylulose-5-phosphate synthase on Chr 4, 9, 1, and one locus for few other loci in non-mevolanate pathway (Supplementary Table S12).

In addition, we identified many terpene synthase (TPS) genes in the *C. australis* genome (Supplementary Fig. S9, S10). Of these, 37 genes were responsible for the synthesis of monoterpenoids (beta-myrcene/(E)-beta-ocimene synthase 2, S-(+)-linalool synthase, d-limonene synthase, (E,E)-geranyllinalool synthase, tricyclene synthase, alpha-terpineol synthase, gamma-terpinene synthase), 24 genes for sesquiterpenoids (alpha-humulene, alpha-copaene synthase-like) and nine genes were involved in the synthesis of diterpenoids (Ent-copalyl diphosphate synthase, cis-abienol synthase, Ent-kaur-16-ene synthase). Among monoterpenoids, eight genes encoding d-limonene synthase and ten genes encoding Beta-myrcene/(E)-beta-ocimene synthase 2 were annotated on Chr 2, 3 and 8. The highest number of genes (22) were annotated for alpha-humulene synthase (Supplementary Fig. S9).

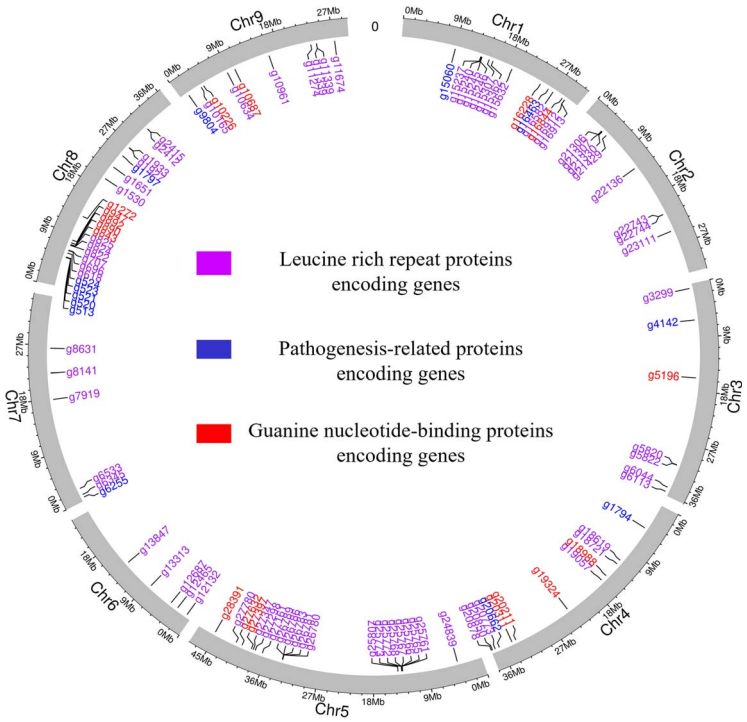


**Figure 8.** Circos plot showing the location of defense-related genes in the *C. australis* genome. Seventeen guanine nucleotide-binding proteins encoding genes (red), 13 pathogenesis-related proteins encoding genes (blue), and 76 Leucine rich repeat (LRR) proteins encoding genes (purple) were identified in the genome.

Farnesol is an acyclic sesquiterpene alcohol which is produced by *C. australis* (Fig. 9).

## Discussion

Here we present the first report of a high quality, complete, and haplotype resolved chromosome level genome assembly for *C. australis*. The size of the assembled nuclear genome was 331.1 Mb of which 311.5 Mb represents the nine pseudochromosomes and the remaining 19.6 Mb could not be anchored to the 9 chromosomes due to the presence of large clusters of uncharacterized repeat elements. With that the present assembly has anchored 94.1% of the total nuclear genome to the chromosome level. The N50 is higher for this genome than that for all other published citrus genomes [14, 15] suggesting that it could potentially be a high-quality reference genome for limes and their hybrids. We achieved the highest assembly BUSCO completeness (98.8%) thus far in citrus and 98.6% annotation completeness which is the third largest among published genomes. Assembly statistics showed that when Hi-C reads were integrated with Hifi reads, it generated phased assemblies with slightly improved contig and scaffold N50s for the collapsed genomes and the two haplotypes. Our results reveal that the combined use of Hifi reads and Hi-C reads leveraged the maximum potential of Hifi reads in assembling heterozygous, highly repetitive plant genomes which is in agreement with previous studies [20].

Hifiasm can run in 4 modes depending on the availability of sequence data. Hifiasm (trio) mode can generate fully haplotype resolved assemblies if maternal and paternal short reads are available. Hifiasm (primary/alternate) mode generates two assemblies; one is the primary assembly containing long contigs which are not haplotigs and an alternate assembly containing haplotigs which are fragmented. Hifiasm (dual) generates two hifiasm assemblies only with Hifi reads which are not fully haplotype resolved and are more likely a primary assembly. Hifiasm (Hi-C) mode maps Hi-C short reads on a Hifi assembly graph creating fully haplotype resolved assemblies containing haplotigs. Hi-C integrated Hifiasm has been applied on humans and other vertebrates [9] and plant genomes to create haplotype resolved genomes [20]. A previous study on a genome of a diploid African cassava cultivar has achieved a high accuracy and contiguity of two chromosome scale haplotypes with a low percentage of misjoined haplotigs from different chromosomes with hifiasm with Hi-C mode [20]. This reveals that the Hifiasm (Hi-C mode) works well for diploid genomes to generate haplotype resolved assemblies but requires further validation to detect haplotype specific variations with high confidence.

For citrus, haplotype resolved genomes have been generated for lemon using Falcon in combination with purge haplotig pipeline [14] and sweet orange using Falcon-unzip [15] to date. We produced two haplotypes for *C. australis* generated by Hi-C integrated mode of Hifiasm which facilitated the identification of variations in gene number and chromosomal lengths with respect to the collapsed assembly. Significant differences between haplotypes in terms of assembly lengths and annotated genes have previously been reported in other plant genomes with different assemblers [21, 22]. Genome and gene synteny results showed well conserved regions across the two sets of chromosomes with certain structural variations (inversions and translocations of genes) between the two haplotypes. Further deep analysis is required to figure out haplotype specific insertions and deletions between the two haplotypes. The differences between the haplotypes could possibly be due to the actual biological variations or due to assembly artifacts which should be verified by further analysis.

Citrus is threatened by a plethora of pests and diseases, with HLB being the most destructive disease in the world [23]. Recently,

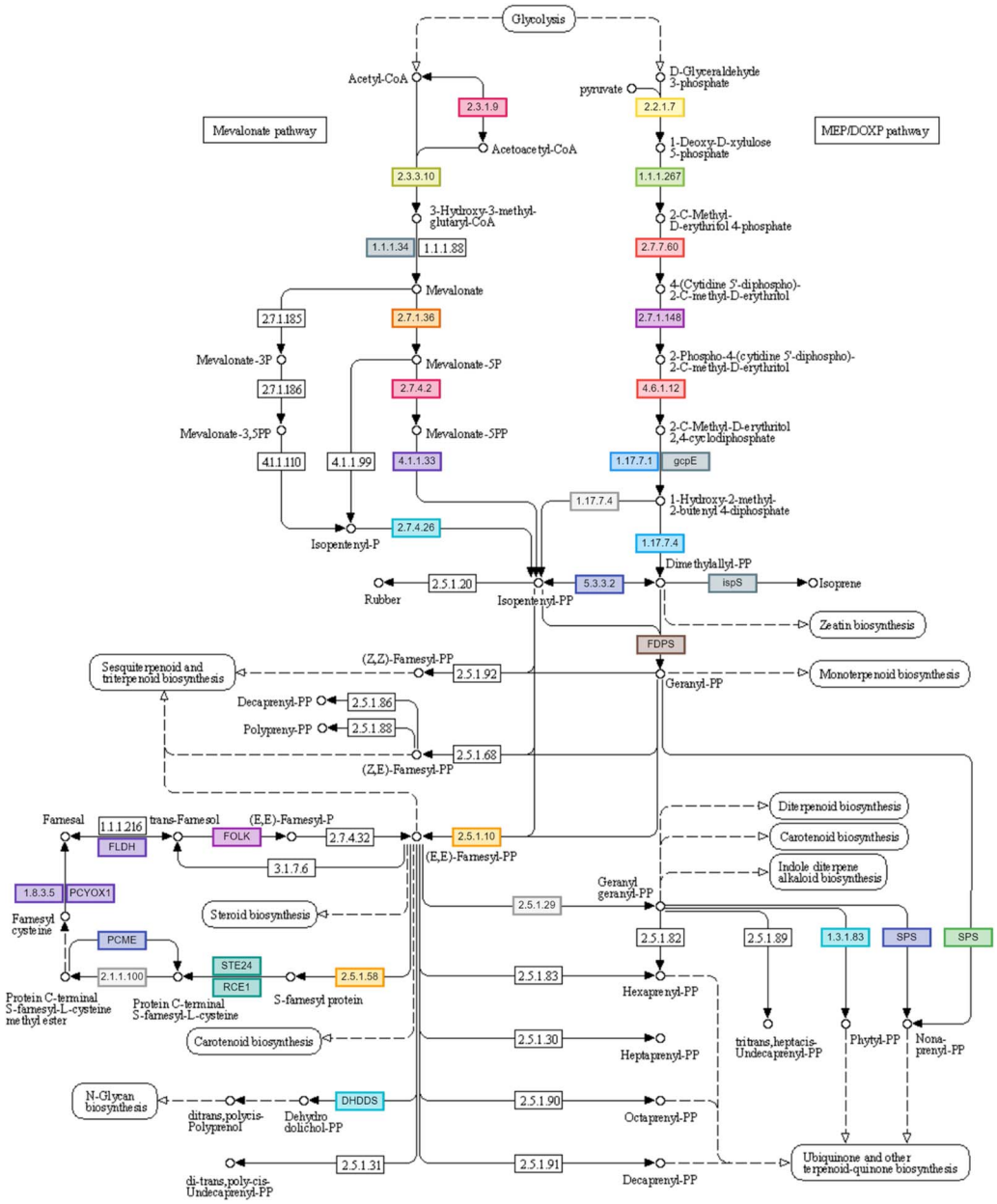


**Figure 9.** Two terpenoid biosynthetic pathways of *C. australis* reproduced with permission of Kanehisa Laboratories. Enzymes that were identified by the annotation are shown in a coloured box, therefore the associated end-products of them are thought to be synthesized in *C. australis*. The other pathways which are not coloured are not present in *C. australis*, however they might be present in other plants. The biosynthesis of terpenoids in *C. australis* is carried out by mevalonate (MVA) and non-mevalonate (MEP/DOXP) pathways in cytoplasm and plastids respectively. Sesquiterpenenoids and triterpenoids are derived from MVA pathway while monoterpenoids and diterpenoids are derived from MEP pathway. *C. australis* synthesizes Isopentenyl-PP (IPP), and its allelic isomer Dimethylallyl-PP (DMAPP) by one of the two pathways in the first phase of the process. The synthesis of IPP from acetyl coenzyme A (AcCoA) by MVA pathway is regulated by six reactions. Enzyme 2.3.1.9: Acetyl-CoA C-Acetyltransferase, 2.3.3.10: 3-hydroxy-3-methylglutaryl-coAsynthase-2 (HMGS), 1.1.1.34: hydroxymethylglutaryl-CoA reductase (HMGCR), 2.7.1.36: mevalonate kinase (MVK), 2.7.4.2: phosphomevalonate kinase (PMK), 4.1.1.33: diphosphomevalonate decarboxylase. The formation of dimethylallyl-PP from pyruvate and D-glyceraldehyde 3 phosphate via MEP pathway is regulated by seven reactions. Enzyme 2.2.1.7: 1-deoxy-D-xylulose-5-phosphate synthase, 1.1.1.267: 1-deoxy-D-xylulose-5-Phosphate Reductoisomerase, 2.7.7.60: 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase, 2.7.1.148: 4-(cytidine 5′-diphospho)-2-C-Methyl-D-Erythritol Kinase, 4.6.1.12: 2-C-Methyl-D-Erythritol 2,4-cyclodiphosphate synthase, 1.17.7.1: (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (ferredoxin), gcpE: GcpE, IspG ((E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase, 1.17.7.4: 4-hydroxy-3-methylbut-2-en-1-YI diphosphate reductase, 5.3.3.2: isopentenyl-diphosphate delta-isomerase, FDPS (GGPS): dimethylallyl-diphosphate:isopentenyl-diphosphate dimethylallyltranstransferase, 2.5.1.10 (GGPS): geranyl-diphosphate:isopentenyl-diphosphate geranyltrans-transferase, 2.5.1.29 (GGPS): trans, trans-farnesyl-diphosphate:isopentenyl-diphosphate farnesyltranstransferase. In the second phase, these C5 isoprene units are catalysed to form farnesyl-PP (C15), geranyl-PP (C10) and geranylgeranyl-PP (C20). In the third phase, geranyl-PP, farnesyl-PP and geranylgeranyl-PP are used to form the primary carbon skeletons of monoterpenes, sesquiterpenes/triterpenes and diterpenes respectively. In the final stage, the primary carbon skeletons of these terpene classes are used to form multiple different forms of terpenes through a variety of processes such as conjugation, oxidation, reductions, and transformations.

a novel class of stable amicrobial peptide has been isolated and characterized from HLB resistant species *C. australasica* and *Poncirus trifoliata* which differs mostly from other antimicrobial peptides produced by HLB susceptible species in terms of length. The short peptide having 67 aa and long peptides (109 aa) were detected in HLB resistant plants, however the susceptible species were only detected with a long aa (118 aa and 109 aa). This novel peptide containing two cystine residues and $\alpha$-helix2 domain can cause cytosol leakage and cell lysis of the disease causing *Candidatus liberibacter* bacteria, thus suppress their growth and induce the immunity in host plants, preventing further bacterial infections [7].

*C. australis* has also been characterized as a resistant species by previous extensive field experiments [4]. The present gene annotation identified the gene g9664 which is homologous to the short novel peptide of *C. australasica* with two transcripts, encoding long aa residues which are not present in susceptible species. The 67 aa sequence is present within these long peptides with only a few amino acid substitutions or insertions. Therefore, it is quite possible that these larger precursor proteins in *C. australis* may later be modified by proteolytic cleavage to produce the 67 SAMP versions of resistant cultivars. It has been reported that different accessions of some species such as *P. trifoliata* can have different degrees of resistance to HLB [4] and this reveals the importance of having a complete genetic picture of all the available accessions or varieties in a species to capture all the allelic variations among them. The gene g9664 could be a potential candidate for resistance against HLB and it's worth monitoring the expression of this gene in response to HLB infection to further validate the function of the gene.

We also annotated three other types of defence related genes which might play important roles against HLB. Leucine rich repeat containing proteins play pivotal roles providing innate immunity in plants by facilitating pathogen recognition [24]. A previous study on HLB resistance of *P. trifoliata* has identified NBS-LRR genes, a most common type of plant disease resistant genes, and a rapidly evolving gene family containing non-NBS type LRR genes which might play crucial roles in disease resistance [25]. Another study on an HLB resistant transgenic line identified enhanced expression levels of leucine-rich repeat receptor kinases (LRR-RKs) compared to susceptible plants revealing the importance of these genes for HLB resistant in citrus [26]. We identified 76 genes encoding LRR proteins in *C. australis* genome which might be crucial for HLB resistance. In addition, the annotation explored 17 guanine nucleotide binding proteins which confer resistance against biotic stresses [27] and 13 pathogenesis-related proteins which have previously been identified as highly upregulated genes in HLB infected *C. australasica* plants [28]. Currently no reference genome is available for Australian limes, therefore the high-quality genome presented in this study paves the way for comparative genomics with other HLB resistant citrus species to fully understand the resistance mechanisms for HLB.

Sugar and organic acids are major attributes of fruit flavour [29]. Some citrus species including limes are highly acidic compared to the cultivated citrus species and the reduced acidity is considered to be an important trait during citrus domestication [30]. Hyper acidification of vacuolar epidermal cells of highly acidic species including limes is controlled by two interacting P-ATPase, *CitPH1* (P3B-ATPase - Mg2+ pump) and *CitPH5* (P3A-ATPase - H+ pump). [19]. *C. australis* is a wild lime with an edible acidic pulp. A good flavour characterized by a proper balance of sourness and sweetness is vital in improving the citrus market value [31]. Australian wild limes are good genomic resources to be tested in breeding novel types of acidic fruits. We here identified the two key genes, *PH1* and *PH5* and their transcription regulators in *C. australis*. Besides the two key genes, nine other genes encoding P-type H(+)-exporting transporters reside in Chr 4 and 6 which might be the other vacuolar H + -ATPases (V-ATPase) driving the citrate import into vacuoles through H+ gradient. A few amino acid substitutions of PH1 protein identified in this study may be involved in the acidification of vacuoles in *C. australis* and may differentiate the acidic and sweet citrus species. The other sweet and acidic citrus species are required to validate these amino acid sequence variations with confidence; however, this is limited due to the lack of high-quality genomes for wild acidic citrus species. Aconitase (CitAco) would be a good target for developing reduced acidic cultivars through gene editing. The genetic loci identified in this study will provide valuable molecular markers for marker assisted breeding for selecting good flavours in *C. australis*. The annotated gene sequences will also be a good resource to understand the sequence variations of these genes in wild limes with compared to the cultivated species.

Leaf oils have previously been characterized in Australian wild limes where $\alpha$-pinene is the dominant compound in *C. australis* (68–79%) [2]. So far, none of the genes specific to the generation of terpenes have been identified and characterized in *C. australis*. Here we have identified and functionally characterized the genes encoding terpenoid bio-synthesis, scattered across *C. australis* genome. We couldn't identify the specific genes for $\alpha$-pinene in *C. australis* genome, however, probable terpene synthase 6 or probable terpene synthase 9 could be potential genes encoding $\alpha$-pinene. In addition to the principal component $\alpha$-pinene, other monoterpenoids including $\beta$-pinene, myrcene, limonene, $\beta$-phellandrene, linalool and sesquiterpenes such as bicyclogermacrene, globulol, and viridiflorol have been isolated from *C. australis* leaves previously [2]. A huge diversity of citrus essential oil components provide resistance for these plants against pest and pathogens [32]. Previous studies have shown that citrus plants rich in some monoterpenes ($\alpha$- and $\beta$-phellandrene, myrcen, d-limonene, and linalool) and sesquiterpenes such as $\gamma$-elemene, t-caryophyellene, germacrene D and $\beta$-elemene have antibacterial activities against *Candidatus liberibacter* bacteria and reduce its spread inside the phloem tissues and thereby suppress its growth and provide resistance for the plants [33]. The monoterpenes linalool, citronellal and citral are known to confer resistance against *Alternaria alternata* by suppressing the spore germination and hyphal development [34]. The function of volatile components as communication substances has also been widely studied in plants where they can attract pollinators though the emission of specific chemical signals. In addition to the roles of volatile components in plant defense and attracting pollinators, they also enhance the interactions among plants and adaptation to abiotic stresses [35, 36]. Diversity of the volatile compositions among different citrus species and different cultivars of the same species is also important in studying their origin and divergence from one another [32]. Identification and functional characterization of genes for specific volatile compounds would provide breeders with direction for developing cultivars with increased levels of beneficial compounds that would in turn improve the consumer demand.

The high-quality *C. australis* genome presented here provides a good resource to improve citrus quality attributed traits through genetic breeding. This genome provides unprecedented opportunities for comparative genomics with other Australian wild limes and commercial citrus species to further understand the species-specific traits, mechanisms underlining biotic stresses, and their

evolutionary relationships which will support the applied breeding efforts. The availability of good reference genomes for Australian native citrus species further facilitates the assembly of pangenomes to explore the existing genomic diversity among the species.

## Materials and methods

### Sample collection, DNA and RNA extraction and sequencing, hi-C sequencing and flow cytometry

Young fresh leaves of *Citrus australis* were collected from a plant grown in a glasshouse at The University of Queensland, Australia (−27.495859, 153.010139) which was sourced from Ross Evans Nursery in Kenmore, QLD, 4069. Total genomic DNA was extracted from pulverized leaf tissues using a CTAB (Cetyltrimethyl ammonium bromide) DNA extraction protocol [37]. PacBio Sequencing was performed on two PacBio Sequel II SMRT cells using the circular consensus sequence method to generate HiFi reads at The Australian Genome Research Facility (AGRF), University of Queensland. Total RNA was extracted from leaves using Trizol and Qiagen kit methods [38]. RNA was sequenced at the AGRF, The University of Queensland, Australia. Fresh young leaves were collected from the same plant for Hi-C sequencing and Flow cytometry. Hi-C sequencing was performed at The Ramaciotti Centre for Genomics, University of New South Wales, Australia. The Hi-C library preparation and analysis were done using Phase Genomics Proximo Plant Hi-C version 4.0. Flow cytometry was performed at the University of Queensland using the BD Biosciences LSR II Flow Cytometer and analysed with the FlowJo software package. Briefly, fresh *C. australis* was co-chopped with the reference standard, Macadamia tetraphylla (presumed size 796 Mb) in Arumuganathan and Earle buffer (https://doi.org/10.1007/BF02672073). Nuclei were gently filtered through a pre-soaked 40-$\mu$m nylon mesh and stained with 50 $\mu$g/mL of propidium iodide and 50 $\mu$g/ml of RNase A. Three biological replicates were performed on three different days.

### Genome assembly

PacBio high fidelity (HiFi) reads were assembled using the Hifiasm Denovo assembler [9] in three modes to produce contigs. In the first mode, Hifi reads were used alone with built-in duplication parameters. In the second mode, Hifi reads were used with —primary option in Hifiasm. In the third mode, Hifi reads were used with Hi-C reads using Hi-C partition options in hifiasm. The contig assemblies were scaffolded using SALSA tool [39]. Genome assembly completeness was assessed using Benchmarking Universal Single-Copy Orthologs against 425 single copy orthologs in viridiplantae lineage (BUSCO v5.2.2) [40] and the contiguity was assessed using QUAST (version 5.0.2) [41]. The longest nine scaffolds were assigned to chromosomes that correspond with previously published chromosome-scale genomes; *C. sinensis* v.03 [42], *C. maxima* [43] and *C. limon* [14]. The synteny among the genomes were visualized with D-Genies v.1.4 [44]. The pseudochromosomes were characterized in terms of telomere repeats [45] and Ribosomal RNA gene repeats [46]. Chloroplast genome which was assembled by GetOrganelle toolkit v.1.7.5 [47] was compared with three sets of scaffolds in dotplots using Dgenies to identify the proportion of the nuclear genome covered by the chloroplast genomic fragments. The K-mer analysis was performed using Jellyfish (v2.2.10) [48] and the histograms were further analyzed using genomescope web tool (http://qb.cshl.edu/genomescope/) [17] with the maximum k-mer coverage set to 1000,000. The genome size was also estimated using kmergenie with K = 97 [49] and Flow cytometry [50].

### Genome annotation

Repeat elements were de novo detected by Repeatmodeler2 version 2.0.1 [51] and masked by Repeatmasker version 4.0.9_p2 [18]. Repeat masking was performed in three options: soft masking, hard masking, and hard masking with "-nolow" option. Quality and adapter trimmed RNA-seq reads were aligned to the unmasked and masked genomes using HISAT2 [52]. Evidence based gene prediction was performed by Braker v2.1.6 [53]. BUSCO was used to assess the genome annotation completeness.

Functional annotation was performed in OmicsBox 2.2.4 [54]. CDS sequences were subjected to BLASTX program with viridiplantae taxonomy against the non-redundant protein sequences database with an e-value of 1.0E-10. CDS were ran through InterProScan and GO terms were retrieved for all the hits obtained by BLAST search using Gene Ontology mapping with Blast2GO annotation. InterProScan and Blast2go annotations were then combined, and the GO terms retrieved from InterProScan were merged with those retrieved from Blast2go annotations. CDS sequences with no BLAST hits obtained from Genemark trained Augustus were extracted and run through coding potential assessment using the prebuilt model of *A. thaliana* and by creating citrus specific models to distinguish the transcripts that were coding and non-coding.

Synteny analysis was performed using MCscan python version using CDS and bed files of two haplotypes as the input data. The comparison of CDS was performed using LAST algorithm followed by filtering the LAST output using the c-score 0.7 (default) to remove the tandem duplications and weak hits. The homologs identified by LAST (anchors) were clustered into synteny blocks using a single linkage clustering mechanism [55]. The syntenic dotplot of protein coding genes was created using SynMap tool in CoGe: comparative genomics database [56] and the whole chromosomal alignment was performed using D-GENIES.

### Identification of citrus specific genes

Genes involved in the production of antimicrobial peptides, defense, volatile compounds and acidity regulation were identified by BLAST homology search with an e-value of 1.0E-10, in CLC (Qiagen, USA). The homologs of *C. australis* and other citrus species were compared by sequence alignment using Clone Manager Ver. 9. The peptide sequences from citrus and other species were obtained from NCBI, citrus genome database (CGD) and Citrus Pan-genome to Breeding Database (CPBD) [57]. The genes involved in the biosynthesis of terpenes were identified by KEGG pathway analysis [58] using OmicsBox 2.2.4.

## Acknowledgements

This project was funded by the Hort Frontiers Advanced Production Systems Fund as part of the Hort Frontiers strategic partnership initiative developed by Hort Innovation, with co-investment from The University of Queensland, and contributions from the Australian Government and Bioplatforms Australia. UN was supported by a graduate scholarship from The University of Queensland. The authors acknowledge The University of Queensland Research Computing Centre (UQ-RCC) for providing all the computing resources for the study and the Flow Cytometry Facility at the Queensland Brain Institute.

## Author contributions

RH, AF, AKM supervised, managed the project, advised and supported data analysis, and data interpretation. AF advised on experiments and AKM supported the genome assembly work. UN and PM performed DNA extractions. UN conducted RNA extractions, data analysis and data interpretation. LC contributed flow cytometry analysis. The manuscript was organized and written by UN. LC contributed to the manuscript with data interpretation of flow cytometry analysis. All authors approved the submitted version.

## Data availability

Raw sequence data generated in this study have been deposited in NCBI Sequence Read Archive (SRA) under BioProject PRJNA910964 and BioSample SAMN32155198 with an accession ID of SRR22742835 for RNA-seq and SRR22793114 for whole genome short read data. The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center [59, 60], Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation, under accession number GWHBQDX00000000, BioProject [PRJCA013889], and Biosample [SAMC1020632] that is publicly accessible at https://ngdc.cncb.ac.cn/gwh. The whole genome and annotation data of Citrus australis have also been submitted to Citrus genome database (https://www.citrusgenomedb.org/).

## Conflict of interests statement

None declared.

## Supplementary Data

Supplementary data is available at *Horticulture Research* online.

## References


1. Gmitter FG, Chen C, Rao MN *et al.* Citrus fruits. In: Kole C, ed. *Fruits and Nuts*. Berlin Heidelberg New York: Springer, 2007, 265–79.
2. Brophy JJ, Goldsack RJ, Forster PI. The leaf oils of the Australian species of citrus (Rutaceae). *J Essent Oil Res*. 2001;**13**:264–8.
3. Lim T. Citrus australis. In: *Edible Medicinal and Non-Medicinal Plants*. Berlin Heidelberg New York: Springer, 2012, 629–30.
4. Ramadugu C, Keremane ML, Halbert SE *et al.* Long-term field evaluation reveals Huanglongbing resistance in citrus relatives. *Plant Dis*. 2016;**100**:1858–69.
5. Bové JM. Huanglongbing: a destructive, newly-emerging, century-old disease of citrus. *J Plant Pathol*. 2006;**88**:7–37.
6. Alquézar B, Carmona L, Bennici S *et al.* Engineering of citrus to obtain huanglongbing resistance. *Curr Opin Biotechnol*. 2021;**70**:196–203.
7. Huang C-Y, Araujo K, Sánchez JN *et al.* A stable antimicrobial peptide with dual functions of treating and preventing citrus Huanglongbing. *Proc Natl Acad Sci*. 2021;**118**:1–10.
8. Benevenuto J, Ferrão LFV, Amadeu RR *et al.* How can a high-quality genome assembly help plant breeders? *Gigascience*. 2019;**8**:giz068.
9. Cheng H, Concepcion GT, Feng X *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;**18**:170–5.
10. Hon T, Mars K, Young G *et al.* Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific data*. 2020;**7**:1–11.
11. Wenger AM, Peluso P, Rowell WJ *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;**37**:1155–62.
12. Guan D, McCarthy SA, Ning Z *et al.* Efficient iterative hi-C scaffolder based on N-best neighbors. *BMC Bioinformatics*. 2021;**22**:1–16.
13. Guk JY, Jang MJ, Choi JW *et al.* De novo phasing resolves haplotype sequences in complex plant genomes. *Plant Biotechnol J*. 2022;**20**:1031–41.
14. Guardo MD, Moretto M, Moser M *et al.* The haplotype-resolved reference genome of lemon (citrus Limon L.Burm f.). *Tree Genet Genom*. 2021;**17**:1–12.
15. Wu B, Yu Q, Deng Z *et al.* A chromosome-level phased Citrus sinensis genome facilitates understanding Huanglongbing tolerance mechanisms at the allelic level in an irradiation induced mutant. *Hortic Res*. 2023;**10**:1–30.
16. Shimizu T, Tanizawa Y, Mochizuki T *et al.* Draft sequencing of the heterozygous diploid genome of Satsuma (Citrus unshiu Marc.) using a hybrid assembly approach. *Front Genet*. 2017;**8**:180.
17. Vurture GW, Sedlazeck FJ, Nattestad M *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017;**33**:2202–4.
18. Chen N. Using repeat masker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinform*. 2004;**5**:4.10.
19. Strazzer P, Spelt CE, Li S *et al.* Hyperacidification of citrus fruits by a vacuolar proton-pumping P-ATPase complex. *Nat Commun*. 2019;**10**:1–11.
20. Qi W, Lim Y-W, Patrignani A *et al.* The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *GigaScience*. 2022;**11**:1–21.
21. Nashima K, Shirasawa K, Ghelfi A *et al.* Genome sequence of Hydrangea macrophylla and its application in analysis of the double flower phenotype. *DNA Res*. 2021;**28**:dsaa026.
22. Shirasawa K, Esumi T, Hirakawa H *et al.* Phased genome sequence of an interspecific hybrid flowering cherry, 'Somei-Yoshino' (Cerasus × yedoensis). *DNA Res*. 2019;**26**:379–89.
23. Duan Y, Zhou L, Hall DG *et al.* Complete genome sequence of citrus Huanglongbing bacterium, 'CandidatusLiberibacter asiaticus' obtained through metagenomics. *Mol Plant-Microbe Interact*. 2009;**22**:1011–20.
24. Padmanabhan M, Cournoyer P, Dinesh-Kumar S. The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. *Cell Microbiol*. 2009;**11**:191–8.
25. Peng Z, Bredeson JV, Wu GA *et al.* A chromosome-scale reference genome of trifoliate orange (Poncirus trifoliata) provides insights into disease resistance, cold tolerance and genome evolution in citrus. *Plant J*. 2020;**104**:1215–32.
26. Qiu W, Soares J, Pang Z *et al.* Potential mechanisms of AtNPR1 mediated resistance against Huanglongbing (HLB) in citrus. *Int J Mol Sci*. 2020;**21**:2009.
27. Patel JS, Selvaraj V, Gunupuru LR *et al.* Plant G-protein signaling cascade and host defense. *3 Biotech*. 2020;**10**:1–8.
28. Weber K, Mahmoud L, Stanton D *et al.* Insights into the mechanism of Huanglongbing tolerance in the Australian finger lime (Citrus australasica). *Front. Plant Sci*. 2022;**13**:1–23.
29. Chen M, Jiang Q, Yin X-R *et al.* Effect of hot air treatment on organic acid-and sugar-metabolism in Ponkan (Citrus reticulata) fruit. *Sci Hortic*. 2012;**147**:118–25.
30. Wang L, He F, Huang Y *et al.* Genome of wild mandarin and domestication history of mandarin. *Mol Plant*. 2018;**11**:1024–37.
31. Li L-J, Tan W-S, Li W-J *et al.* Citrus taste modification potentials by genetic engineering. *Int J Mol Sci*. 2019;**20**:6194.

32. Delort E, Jaquier A, Decorzant E *et al.* Comparative analysis of three Australian finger lime (Citrus australasica) cultivars: identification of unique citrus chemotypes and new volatile molecules. *Phytochemistry*. 2015;**109**:111–24.
33. Hijaz F, Nehela Y, Killiny N. Possible role of plant volatiles in tolerance against huanglongbing in citrus. *Plant Signaling Behavior*. 2016;**11**:1–12.
34. Yamasaki Y, Kunoh H, Yamamoto H *et al.* Biological roles of monoterpene volatiles derived from rough lemon (Citrus jambhiri lush) in citrus defense. *J Gen Plant Pathol*. 2007;**73**:168–79.
35. Bouwmeester H, Schuurink RC, Bleeker PM *et al.* The role of volatiles in plant communication. *Plant J*. 2019;**100**:892–907.
36. Vivaldo G, Masi E, Taiti C *et al.* The network of plants volatile organic compounds. *Sci Rep*. 2017, 2017;**7**:11050.
37. Furtado A. DNA extraction from vegetative tissue for next-generation sequencing. *Methods Mol Biol*. 2014;**1099**:1–5.
38. Furtado A. RNA extraction from developing or mature wheat seeds. *Methods Mol Biol*. 2014;**1099**:23–8.
39. Ghurye J, Pop M, Koren S *et al.* Scaffolding of long read assemblies using long range contact information. *BMC Genomics*. 2017;**18**:1–11.
40. Simão FA, Waterhouse RM, Ioannidis P *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;**31**:3210–2.
41. Gurevich A, Saveliev V, Vyahhi N *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;**29**:1072–5.
42. Xu Q, Chen L-L, Ruan X *et al.* The draft genome of sweet orange (Citrus sinensis). *Nat Genet*. 2013;**45**:59–66.
43. Wang X, Xu Y, Zhang S *et al.* Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nat Genet*. 2017;**49**:765–72.
44. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 2018;**6**:1–6.
45. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;**27**:573–80.
46. Seemann T. barrnap 0.9: rapid ribosomal RNA prediction (RRID:SCR_015995). 2013;
47. Jin J-J, Yu W-B, Yang J-B *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol*. 2020;**21**:1–31.
48. Manekar SC, Sathe SR. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*. 2018;**7**:giy125.
49. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. 2014;**30**:31–7.
50. Doležel J, Greilhuber J, Suda J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat Protoc*. 2007;**2**:2233–44.
51. Flynn JM, Hubley R, Goubert C *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci*. 2020;**117**:9451–7.
52. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;**37**:907–15.
53. Hoff KJ, Lomsadze A, Borodovsky M *et al.* Whole-genome annotation with BRAKER. *Methods Mol Biol*. 2019;**1962**:65–95.
54. OmicsBox – Bioinformatics made easy. BioBam Bioinformatics (version 2.2.4). 2019. https://www.biobam.com/omicsbox/.
55. Tang H, Bowers JE, Wang X *et al.* Synteny and collinearity in plant genomes. *Science*. 2008;**320**:486–8.
56. Lyons EH. *CoGe, a New Kind of Comparative Genomics Platform: Insights into the Evolution of Plant Genomes*. Berkeley: University of California; 2008.
57. Liu H, Wang X, Liu S *et al.* Citrus pan-genome to breeding database (CPBD): a comprehensive genome database for citrus breeding. *Mol Plant*. 2022;**15**:1503–5.
58. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;**28**:27–30.
59. Chen M, Ma Y, Wu S *et al.* Genome warehouse: a public repository housing genome-scale data. *Genomics Proteomics Bioinformatics*. 2021;**19**:584–9.
60. CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for bioinformation in 2022. *Nucleic Acids Res*. 2022;**50**:D27–38.