scientific reports

Check for updates

OPEN

Improved genome assembly of double haploid *Prunus persica* siblings 'Lovell 2D' and 'Lovell 5D' and the peach NLRome

Christopher Gottschalk[1]✉, Jordan R. Brock[2], Ben N. Mansfeld[2], Dorrie Main[3], Sook Jung[3], Ping Zheng[3], Cheryl Vann[1], Mark Demuth[1], Dennis Bennett[1], Zongrang Liu[1] & Chris Dardick[1]✉

Prunus persica **(peach) has long served as a model fruit tree for studying phenological events. The peach variety 'Lovell 2D' was used to generate one of the first high-quality genome assemblies for a tree species, using Sanger sequencing of genetic-map ordered BAC clones. A key to the high quality of this early assembly was the use of a doubled haploid variety, which eliminates the challenges posed by mixed haplotypes. Here, we sequenced and re-assembled the 'Lovell 2D' genome along with a doubled haploid sibling 'Lovell 5D' using 3rd generation technologies. The resulting genomes were significantly more contiguous than the current 'Lovell 2D' reference genome (ver2.0 updated in 2017) and are closer to the estimated total genome size for peach (265 Mb). In addition, new gene, transposable element (TE), and Nucleotide-binding domain and Leucine-rich repeat receptor (NLR) annotations were performed to enhance the integrity and utility of the genome. These updated peach doubled-haploid reference assemblies will provide the research community with an improved reference genome for genomics-guided studies and breeding efforts.**

Peach (*Prunus persica* L. Batsch) is a temperate tree fruit crop adapted to diverse climates across the globe. U.S. production of peach ranges between 650,000 and 710,000 tons per year and routinely suffers from crop losses due to extreme weather events (i.e., spring frost events)[1]. Moreover, production of peach is threatened by numerous diseases such as Brown Rot (caused by *Monilinia spp.*), which are becoming more prevalent in response to the climacteric conditions[2]. Improved genomics tools can aid in the development of more secure, sustainable, and resilient peach varieties. The first peach genome released was of 'Lovell 2D', a doubled haploid variety that was assembled from first-generation paired-end Sanger sequencing of fosmid and BAC clones[3]. The resulting assembly had 234 scaffolds placed into eight pseudomolecules using a genetic map. The final assembly was 224.6 Mb in length with a scaffold N50 of 26.8 Mb[3]. A revision was performed four years later using Illumina next-generation sequencing data[4]. This revision effort resulted in the closing of 212 gaps in the 'Lovell 2D' assembly with a minor sequence gain of ~ 25 Kb and increased scaffold length to 27.4 Mb[4]. 2021 saw the release of four new peach genomes based on PacBio third-generation sequencing platform. The first was for the Chinese variety 'Longhua Shui Mi' with a reported Contig N50 of 5.17 Mb and a length of 257.2 Mb[5]. The second genome was from a Chinese flat peach variety '124 Pan'[6]. This assembly had a reported N50 of > 26 Mb and a total length of 206 Mb. The third genome was for another Chinese cling-stone variety with a reported scaffold N50 of 29.68 Mb and a length of 247.33 Mb[7]. That assembly represented 99.8% of the estimated genome size, the closest to date. The fourth genome was for a semi-dwarf breeding line variety 'Zhongyoutao 14' and the assembly had a reported scaffold N50 of 27.89 and a total length of 228.82 Mb[8].

Over the past decade, the peach 'Lovell 2D' reference genome has served as a critical resource to the research community. Although tremendous efforts have been made to improve the genome resources for peach, little has been done since 2017 to improve the reference genotype's assembly. To address this, we sequenced and assembled an improved peach reference genome using the original double haploid 'Lovell 2D' and a sibling doubled haploid 'Lovell 5D'. To achieve high-quality genome assemblies, we utilized the cutting-edge Oxford Nanopore PromethION sequencing platform to generate high accuracy simplex reads and combined them with PacBio HiFi reads. Combining these sequencing data types with the latest genome assembly software that utilizes mixed data types resulted in near-complete genome assemblies. These updated doubled haploid

[1]USDA-ARS Appalachian Fruit Research Station, 2217 Wiltshire Rd., Kearneysville, WV 25430, USA. [2]Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130, USA. [3]Washington State University, 255 E Main St, Pullman, WA 99163, USA. ✉email: christopher.gottschalk@usda.gov; chris.dardick@usda.gov

reference genomes will provide a valuable resource for molecular breeding strategies helping overcome apparent domestication bottlenecks that may limit breeding germplasm diversity for certain critical traits[3].

Materials and methods

Plant materials

Field grown trees of 'Lovell 2D' and 'Lovell 5D', of an estimated ~ 30 years of age, are maintained in the breeding germplasm at the Appalachian Fruit Research Station (AFRS) in Kearneysville, WV (Fig. 1A). The trees were grown using standard commercial practices for training, disease, and pest management. In 2017, young leaf material was collected and flash frozen. The tissues remained frozen at -80 °C until removed for DNA extraction. 'Lovell 2D' and 'Lovell 5D' are available through the National Plant Germplasm System as DPRU 263 – 'LOV-2-Haploid' and PI 673461 – 'LOV-5-Haploid', respectively (Fig. 1B, C).

DNA extraction

Frozen leaf tissues were ground into a fine powder using a mortar and pestle chilled with liquid N_2. A total of ~ 1 gram of tissue was used per extraction. High-molecular weight (HMW) DNA was extracted using the LeafGo (leaf to genome) extraction protocol that relies on gravity-based filter columns (Qiagen Genomics Tip Kit, Germantown, MD)[9]. There were no modifications to the protocol except during the precipitation of HMW DNA using 2-propanol. Following the addition of 2-propanol, the HMW DNA sample was centrifuged at 5,000 g for 15 min with the centrifuge chilled to 4 °C. Following pelleting, the HMW DNA pellet was washed using 80% ethanol and centrifuged at 5,000 g for 15 min with the centrifuge chilled to 4 °C to pellet the DNA. This washing step was repeated twice and after the second wash the pellets were left to air dry for 15 min. Following drying, HMW DNA was dissolved in TE buffer. The resulting HMW DNA samples were tested for fragment size, quality, and purity using an Agilent Tapestation (Santa Clara, CA) using the Genomic DNA screen tape and reagents according to manufacturer protocols.

Library preparation and sequencing

Oxford Nanopore Technologies (ONT, Oxford, UK) sequencing libraries were prepared for each variety using the Ligation Sequencing Kit V14 (SQK-LSK114) without modifications. The library DNA quality and quantity were measured using an Agilent Tapestation with the Genomic DNA screen tape and reagents. Libraries were sequenced on the ONT PromethION platform using independent R10.4.1 flow cells. Basecalling was conducted within the MinKnow GUI software (v. 23.11.7) using the Dorado basecaller (v.7.2.13) for real-time simplex basecalling. Following the completion of sequencing, command-line Dorado (v.0.7.2) was executed a second time with the *dna_r10.4.1_e8.2_400bps_sup@v5.0.0* model to generate highly accurate > Q20 reads. The resulting simplex reads were evaluated for PHRED scores and length using Nanoplot (v.1.42.0;[10]. The reads were then filtered for a length of 40 Kb for 'Lovell 2D', and a length of > 20 Kb and Q25 for 'Lovell 5D' using Filtlong (v.0.2.1; https://github.com/rrwick/Filtlong). For 'Lovell 2D', we additionally executed the command-li

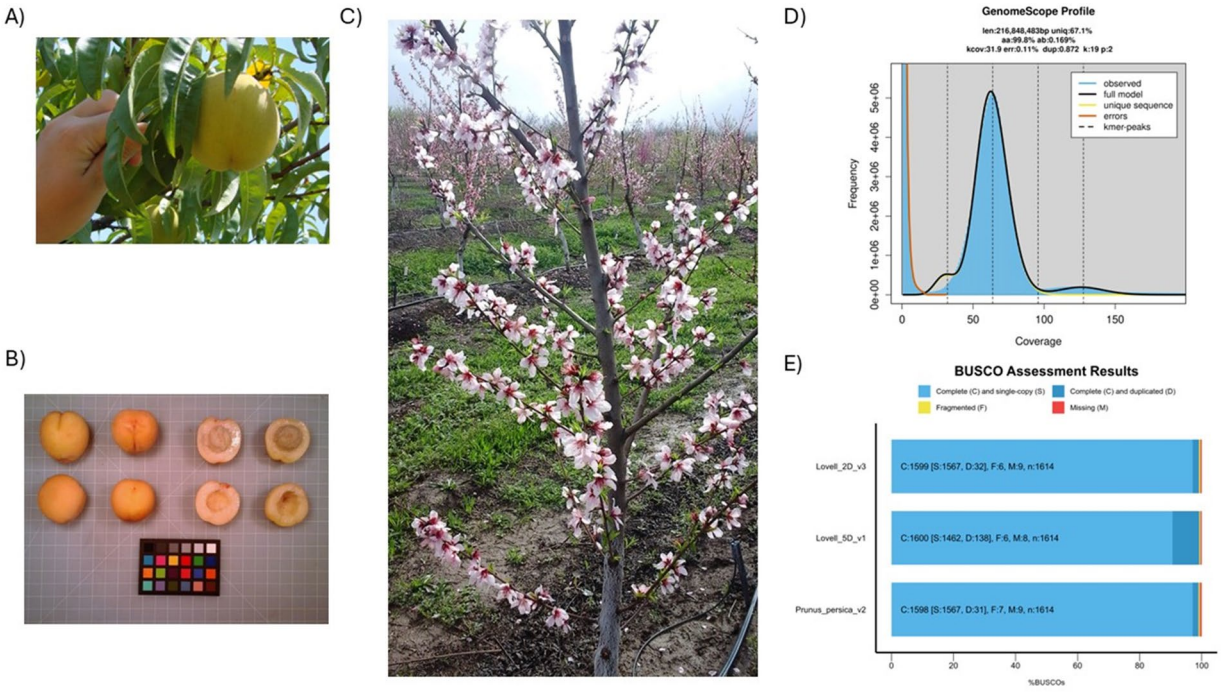

Fig. 1. (**A**) 'Lovell 5D' fruit grown at AFRS. (**B**) 'Lovell 5D' (PI 673461) fruit collected from the USDA germplasm in Parlier, CA. (**C**) 'Lovell 5D' (PI 673461) tree in bloom at the USDA germplasm in Parlier, CA. (**D**) GenomeScope 2.0 *k*-mer frequency plot and estimated genome statistics for 'Lovell 5D'. (**E**) BUSCO genome plots for the 'Lovell 2D v3.0', 'Lovell 5D v1.0', and 'Lovell 2D' v.2.0.a1 reference genome. Photos of 'Lovell 5D' (PI 673461) were retrieved from the USDA National Plant Germplasm System.

ne Dorado (v.0.4.1) a second time using the 'duplex' parameter with the *dna_r10.4.1_e8.2_400bps_hac@v4.2.0* model to generate duplex reads. For 'Lovell 5D', we used the same HMW DNA sample to sequence on a PacBio Revio (Menlo Park, CA) Flow Cell to generate a HiFi read dataset.

Genome assembly, scaffolding, and annotation

To estimate genome size and to confirm the doubled haploid state of 'Lovell 5D', we counted the *k-mers* within the HiFi dataset. The *k*-mer counting was performed using KMC (v.3.2.4) using the following parameters: -k19 -t20 -m60 -ci1 -cs10000 as recommended by GenomeScope 2.0 (http://genomescope.org/genomescope2.0/)[11,12]. The counted *k*-mers were then transformed into a histogram using the kmc_tools script in KMC package with the transform reads histogram reads.histo -cx10000 parameters[11,12]. The resulting histogram was then uploaded to GenomeScope 2.0 for *k*-mer plot generation (http://genomescope.org/genomescope2.0/)[12].

We assembled the genome of 'Lovell 2D' using two independent assembly programs. First, we used Verkko (v.2.0) with duplex Oxford Nnanopore data used as HiFi input and the simplex data as ultra-long reads[13]. All other parameters were set to default. Our second approach utilized HiFiasm (v.0.19.8) with the duplex data (Nanopore) as HiFi input and the simplex data as ultra-long reads[14]. PacBio HiFi data was not generated for 'Lovell 2D', as we aimed to evaluate the ability to generate a high-quality genome from the recently developed Nanopore duplex data that is similar in quality to PacBio HiFi. Again, all other parameters were set to default. The resulting assemblies were analyzed for completeness using GenomeTools statseq (v.1.6.5)[15]. Based on the statistics of each assembly, the most contiguous and longest 'Lovell 2D' assembly was selected to move forward with.

Our assembly approach for 'Lovell 5D', used HiFiasm (v.0.19.9) with HiFi sequence data used as HiFi input and the Oxford Nanopore simplex data as ultra-long reads[14]. We invoked the parameters --ul-cut 20,000, --telo-m TTTAGG and -l0 with HiFiasm. The resulting assemblies were analyzed for completeness using GenomeTools seqstat (v.1.6.5)[15]. The resulting primary assembly fastas for each genotype were scaffolded using RagTag (v2.1.0;[16]. This scaffolding method relied on mapping the genomes of each genotype to a reference, and in our case we used the 'Lovell 2D' v2.01.a1 scaffolded reference assembly[16,4]. The agp file generated by RagTag was then converted into a fasta using agptools assemble (https://github.com/WarrenLab/agptools). Final scaffolding stats were then obtained using gfastats (v.1.3.7)[17].

The scaffolded assemblies were then annotated for transposable elements (TE) and repeats using EDTA (v.2.0.1)[18]. Here, we used a highly curated TE library from the 'Texas' almond genome and CDS sequences from the 'Lovell 2D' v.2.0.a1 genome as input into EDTA for *de novo* TE/repeat annotation[19,4]. The resulting annotation and genome assembly were evaluated for quality using Long-Terminal Repeat (LTR) Assembly Index (LAI) within the EDTA suite[20]. EDTA also produced a masked genome file, which we used as input into the gene annotation pipeline.

For gene annotation, we utilized the MAKER2 pipeline, which incorporates transcriptional and protein homology evidence (Holt and Yandell, 2011[34]). For transcriptional evidence, the primary transcripts from 'Lovell 2D' v2.0.a1 genome were downloaded from Genome Database for Rosaceae (GDR)[4];[21]. For protein evidence based on homology, we downloaded and used the Araport11 protein sequence dataset[22]. We performed three rounds of MAKER (v.3.01.03) annotation as previously performed by the authors[23]. In short, the first round was the evidence-based round with ab initio gene prediction training. The following two rounds of annotation were conducted using retrained ab initio gene prediction and carrying over of evidence-based annotations. The gene predictors used were SNAP (v.2013_11_29) and Augustus (v.3.5.0)[24]; (Hoff and Stanke, 2019[35]). Noncoding genes were predicted using Infernal (v.1.1) (Nawrocki and Eddy, 2013[36]). Genome and transcriptome quality were additionally evaluated using BUSCO (v.5.8.0) with the embryophyta odb10[25].

Annotation of R-genes

We set out to annotate the *R*-gene space in our two new genome assemblies along with the four previously released *P. persica* genomes hosted on the GDR webpage[7,21,8,4,6]. These additional genomes included: `Lovell 2D` v2.0.a1 (Prup;[4], `Chinese Cling` (Ppcc;[7], `Zhongyoutao 14` (CN14;[8], `124 Pan` (124Pan;[6]. Each genome had its representative sequence processed to remove any soft masking. We then ran each genome separately through the FindPlantNLR snakemake pipeline (v.2.0) following the software's default methodology[26]. The resulting annotations of the complete catalog of *R*-genes (NBARC) for each genome was then compared to its respective gene annotation using gffcompare (v.0.12.6) to identify the number of novel annotations using the "u" flag[27].

Comparative genomics

Comparative genomics was conducted using the SyRI (v.1.7.0) package[28]. Default parameters were used when executing the SyRI commands as described in the manual. The plotting of the syntenic regions was generated using the plotsr (v.1.1.0) program as recommended in the SyRI manual[28].

Pan-NLR analyses

Nucleotide-binding domain and Leucine-rich repeat receptors (NLR) gene synteny of the primary (or longest) isoforms of NLRs were determined between the six accessions of *Prunus persica*, using the genome of 'Lovell 2D v3.0' as a reference genome for comparison. Pairwise analyses of syntenic genes and NLRs were conducted between all genomes. Coding sequences were extracted from genome annotations using gffread v0.12.7[27]. Macrosynteny and NLR gene synteny were assessed using the Python version of the MCScanX pipeline (v1.1.12)[29,30]. Macrosynteny was defined with a minimum syntenic block size (--minspan) of 10 genes for both sets of analyses. Syntenic blocks and collinear gene pairs were identified, and results were visualized using the jcvi.graphics.karyotype module.

Results and discussion

We generated a total of 69.44 Gb of ONT simplex sequence data for 'Lovell 2D'. This sequencing yield represents an estimated 262x coverage, assuming the estimated genome size is 265 Mb[31,3]. Following length and quality

filtering, the 'Lovell 2D' simplex data was reduced to 5.7 Gb with a N50 of 47.6 Kb and a median read quality of Q17.4. The ONT sequencing of 'Lovell 2D' additionally yielded 5.53 Gb of duplex reads that represent 21x coverage with a median read quality of Q24.8. We generated a total of 86.63 Gb of ONT simplex sequence data for 'Lovell 5D'. This sequencing yield represents an estimated 327x coverage. Following length and quality filtering, the 'Lovell 5D' simplex data was reduced to 24.02 Gb with a N50 of 20.7 Kb and a median read quality of Q26. The HiFi sequencing of 'Lovell 5D' yielded 10.51 Gb of HiFi reads with a 6.33 Kb mean length and a median read quality of Q39. This amount of sequence data was deemed sufficient to proceed with genome assembly for both genotypes. We counted *k*-mers in the HiFi data to confirm the double haploid state of 'Lovell 5D'. GenomeScope 2.0 reported a homozygosity minimum rate of 99.81% and an estimated genome size of 216 Mb (Fig. 1D). This lower estimated genome size than the previous estimate of 265 Mb could be an artifact of the homozygosity of 'Lovell 5D' and the relatively short PacBio read lengths[31,3].

An improved 'Lovell 2D' genome

Verkko's initial assembly of 'Lovell 2D' yielded a primary assembly of 267.6 Mb contained in 969 contigs (STable 1). The contig N50 was 16.5 Mb, which is 64-fold improvement in contig length and ~ 4-fold improvement in scaffold lengths over the 'Lovell 2D' v.2.0.a1 reference genome[4]. Using the 'Lovell 2D' v.2.0.a1 genome for homologous scaffolding with RagTag generated 904 scaffolds with a scaffold N50 of 28.2 Mb. The scaffold N50 is nearly 1 Mb longer than the 'Lovell 2D' v.2.0.a1 reference genome (Table 1)[4]. We also observed a small increase in complete BUSCO percentage of 0.1% over the 'Lovell 2D' v.2.0.a1 reference genome (Table 1)[4]. However, we observed an increase of 27.7% to the LAI score compared to the 'Lovell 2D' v.2.0.a1 reference genome[4]. Increased LAI indicates more contiguous assembly through the complex, repetitive regions of the genome.

A new genome for 'Lovell 5D'

HiFiasm generated a primary assembly of 257.8 Mb contained in 400 contigs (STable 1). The contig N50 was 23 Mb, which is 90-fold increase in contig length and > 5-fold increase in scaffold length over the 'Lovell 2D' v.2.0.a1 reference genome[4]. Using the 'Lovell 2D' v.2.0.a1 genome for homologous scaffolding with RagTag generated 193 scaffolds with a N50 of 29 Mb (STable 1). The scaffold N50 is nearly 2 Mb larger than the 'Lovell 2D' v.2.0.a1 reference genome[4]. Our assembly is the closest reported to the predicted length from Verde et al.,[3] of 265 Mb and exceeded the estimated length from our *k*-mer analysis (Fig. 1D). We observed no difference in complete BUSCO percentage over the 'Lovell 2D' v.2.0.a1 reference genome (Table 1)[4]. However, we observed a nearly 7% increase in the complete duplication rate of BUSCOs. Lastly, we observed an increase of 23.9% in the LAI score compared to the 'Lovell 2D' v.2.0.a1 reference genome[4](Table 1).

Our final assemblies of only the chromosome-scale scaffolds were 236 and 246 Mb in length for 'Lovell 2D' and 'Lovell 5D', respectively (Table 1). Each of these assemblies captured the expected eight chromosomes with N50s of > 28 Mb. In comparison to the 'Lovell 2D' v.2.0.a1 reference genome, each of our two assemblies were longer by 9–19 Mb and had N50 increases at the contig level by 64 to 90-fold[4]. As a result, the scaffold N50 were also increased by 0.9–1.8 Mb in our two assemblies compared to 'Lovell 2D' v.2.0.a1 (Table 1)[4]. These results together demonstrate a improvement over the 'Lovell 2D' v.2.0.a1 reference genome[4].

Comparison of gene annotations

The annotation of transposable elements and other repetitive features for our two genomes was found to be lower than the reported amount in the 'Lovell 2D' v.2.0.a1 reference genome[4](Table 2). Both of our genomes were found to have 82.9–85.1 Mb of repeat sequences compared to 84.4 Mb in the 'Lovell 2D' v.2.0.a1 reference genome. This lower content represents a decrease of ~ 4–7% over 'Lovell 2D' v.2.0.a1. However, the higher LAI

Assembly	Primary Contig 'Lovell 2D'	Scaffolded 'Lovell 2D'	Chromosome–only 'Lovell 2D'	Primary Contig 'Lovell 5D'	Scaffolded 'Lovell 5D'	Chromosome–only 'Lovell 5D'	Reference Scaffolded 'Lovell 2D'
Version	3.0	3.0	3.0	1.0	1.0	1.0	2.0.a1
Assembler	Verkko	Verkko	Verkko	HiFiasm	HiFiasm	HiFiasm	Arachne
Number of Contigs	969	969	65	400	400	193	2,525
Number of Scaffolds		904	8		193	8	191
Total Length	267,589,852	267,596,352	236,159,615	257,757,688	257,778,388	246,657,795	227,411,381
Contigs >1 M nt	17	8	8	16	8	8	8
Contig N50	16,505,057	16,505,057	16,505,057	23,125,875	23,125,875	23,125,875	255,400
Scaffold N50	–	28,261,597	28,261,597	–	29,122,444	29,122,444	27,368,013
L50	6	4	4	5	4	4	4
% BUSCO	–	99.10%	99.10%	–	99.00%	99.00%	99.00%
Complete single copy BUSCO	–	97.10%	97.10%	–	90.60%	90.60%	97.10%
Complete duplicate BUSCO	–	2.00%	2.00%	–	8.60%	8.60%	1.90%
Fragmented BUSCO	–	0.40%	0.40%	–	0.40%	0.40%	0.40%
Missing BUSCO	–	0.60%	0.60%	–	0.50%	0.50%	0.60%
LAI	–	27.46	–	–	26.65	–	21.5

Table 1. Genome assembly statistics for the 'Lovell 2D' v3.0, 'Lovell 5D', and the reference 'Lovell 2D' genome.

scores for our assemblies indicate improved contiguity and assembly quality of these repetitive regions. We observed 'Lovell 5D' exhibiting higher repeat space. This increase in repeat space was observed within all the annotated features except unknown LINE elements (Table 2).

We undertook a *de novo* gene annotation approach as opposed to lifting over the annotation from 'Lovell 2D' v.2.0.a1. We identified 26,834 and 26,764 genes in the new 'Lovell 2D' and 'Lovell 5D' genomes, respectively (Table 2). These results are lower than 'Lovell 2D' v.2.0.a1 by 37 and 99 genes, respectively. The resulting complete BUSCO scores reflect this, with the new 'Lovell 2D' being 97.00% and 'Lovell 5D' being 98.40%. Whereas 'Lovell 2D' v.2.0.a1 exhibited a 99.10% complete BUSCO score. Although lower by < 100 genes from the reference, these *de novo* gene annotations were conducted with no manual curation, unlike the reference. We also conducted a more modern approach to annotation for ncRNAs in our genomes, which resulted in a more thorough annotation compared to 'Lovell 2D' v.2.0.a1. These two new resources now include annotated ncRNA features such as rRNA, miRNA, and snoRNA (Table 2).

Synteny among 'Lovell' genomes

We conducted a syntenic analysis between our two new genomes and the 'Lovell 2D' v.2.0.a1 reference genome[4](Fig. 2). Overall, synteny was highly conserved as expected with > 96% of the reference genome being found in 593 syntenic regions with the v3.0 assembly of 'Lovell 2D' (Table 3). Within this comparison, we also identified > 7,000 insertion events corresponding to ~ 480 Kb of sequence, which is less than 2% of the total newly added sequence. We also identified several structural variations between the two assemblies. Of which, 10 were inversions, 65 were translocations, and 1,256 were annotated as duplications. The 'Lovell 5D' assembly also exhibited a high level of conserved sequence identity and organization with its sibling, the 'Lovell 2D' reference assembly and our third version (Fig. 2). Comparing the 'Lovell 2D' v3.0 with 'Lovell 5D', we identified 70 syntenic regions that span 98.8% of the total chromosomal length of 'Lovell 2D'. We also observed lower structural variation (inversions and translocations) and sequence annotated variation in this comparison than

Intergenic			
LTR - Copia	5.42%	5.91%	8.60%
LTR - Ty3/Gypsy	7.10%	7.59%	9.97%
LTR - Unknown	4.19%	4.64%	1.00%
TIR - CACTA	2.27%	2.46%	
TIR - Mutator	6.15%	6.29%	
TIR - PIF Harbinger	1.79%	1.93%	
TIR - Tc1 Mariner	0.23%	0.27%	
TIR - hat	1.34%	1.51%	
Undefined TIR			9.05%
LINE	0.70%	0.76%	0.63%
LINE - unknown	0.02%	0.02%	
Unknown Retroelements			0.35%
Helitron	1.52%	1.65%	0.24%
Unknown Repeats			7.54%
Total %	30.75%	33.03%	37.14%
Total bp	82,291,331	85,142,280	84,410,000
Genic			
Protein Coding	26,834	26,764	26,873
Average Exon per mRNA	5.3	5.1	5.2
tRNA	702	743	474
rRNA	2,299	2,414	-
miRNA	278	287	-
snoRNA	210	217	-
Total ncRNA	3,826	4,050	769
R-gene			
% BUSCO	97.00%	98.40%	99.10%
Complete single copy BUSCO	94.80%	92.60%	96.90%
Complete duplicate BUSCO	2.20%	5.80%	2.20%
Fragmented BUSCO	1.80%	0.90%	0.60%
Missing BUSCO	1.20%	0.70%	0.40%

Table 2. Genome annotation statistics for the 'Lovell 2D' v3.0, 'Lovell 5D', and the reference 'Lovell 2D' genome.

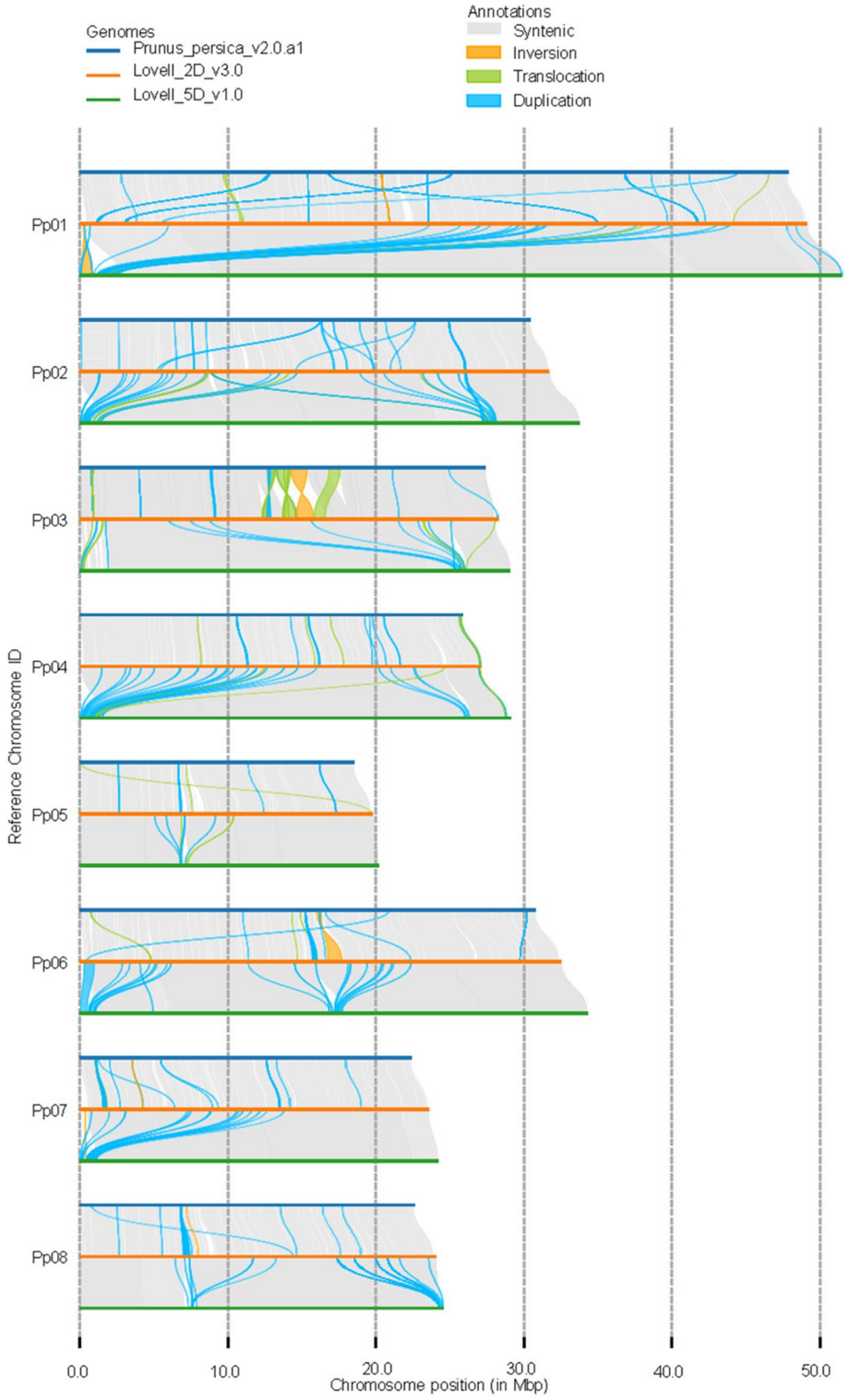

Fig. 2. Structural variation and synteny between 'Lovell 2D' v3.0, 'Lovell 5D', and the reference 'Lovell 2D' genime.

within the intra-synteny of the 'Lovell 2D' assemblies. However, the total duplicated sequence was 1.5 Mb more than the intra-synteny of the 'Lovell 2D' assemblies (Table 3; Fig. 2). This result was also observed in the higher rate of complete and duplicated BUSCO genes in the 'Lovell 5D' assembly.

To explore the duplication events in the 'Lovell 5D' assembly, we undertook two methods to verify if these regions could be the result of biological variation. First, we reassembled the 'Lovell 5D' genome using only the PacBio HiFi reads to exclude errors potentially introduced by the lower quality Nanopore reads. We realigned

Assembly Comparison	Reference 'Lovell 2D' v.2.0.a1 vs. 'Lovell 2D' v3.0			Lovell 2D' v3.0 vs. 'Lovell 5D' v1.0		
Structural Variation	Count	Length in Reference	Length in 'Lovell 2D' v3.0	Count	Length in 'Lovell 2D' v3.0	Length in 'Lovell 5D' v1.0
Syntenic regions	593	219,121,542	219,682,496	70	233,263,902	233,249,832
Inversions	10	1,589,177	2,235,701	2	93,046	796,373
Translocations	65	3,452,533	3,384,659	38	1,638,787	1,601,927
Duplications (reference)	51	745,435	–	33	917,300	–
Duplications (query)	1,205	–	8,161,825	391	–	9,681,140
Not aligned (reference)	385	2,564,223	–	88	1,624,068	–
Not aligned (query)	1,060	–	5,235,988	390	–	1,943,551
Sequence annotations						
SNPs	73,076	73,076	73,076	18,765	18,765	18,765
Insertions	7,301	–	486,478	26,194	–	154,579
Deletions	24,582	258,741	–	3,761	126,528	–
Copy gains	86	–	399,330	4	–	80,640
Copy losses	34	154,436	–	7	110,166	–
Highly diverged	1,419	1,973,272	2,587,501	73	236,516	906,086
Tandem repeats	38	229,062	355,692	9	418,240	407,214

Table 3. Genomic variation statistics for the 'Lovell 2D' v3.0, 'Lovell 5D', and the reference 'Lovell 2D' genome.

the HiFi-only 'Lovell 5D' assembly with 'Lovell 2D' v.2.0.a1 reference genome using Syri (SFig. 2)[4]. The resulting alignments demonstrated similar high duplication and structural variation as previously observed (SFig 2). Our second approach was to map the HiFi reads of 'Lovell 5D' onto the 'Lovell 2D' genome using minimap2 to evaluate coverage around these duplicate rich hot spots. We observed high coverage surrounding these regions (SFigs. 3 and 4). This resulting high read coverage suggests that the 'Lovell 5D' possesses these duplications but they are absent in the 'Lovell 2D' leading to the piling up of reads under the regions of similarity. However, we cannot completely exclude that these duplications in 'Lovell 5D' are sequencing artifacts, assembly or scaffolding errors.

R-gene space within *P. persica* genomes

Like other crops, peach production can suffer from several plant pathogens, including but not limited to peach leaf curl, brown rot, and peach scab[2]. Understanding the repertoire and diversity of pathogen resistance loci in the currently sequenced peach genomes is crucial for mobilizing these genes in future germplasm improvement. Using the FindPlantNLR pipeline, we identified between 343 and 448 Nucleotide Binding Leucine Rich repeat domain (NLRs or *R*-genes)-genes (primary transcripts) in the *P. persica* genomes[26]. The Chinese Cling genome exhibited the highest *R*-gene count with 448, of which 131 were not overlapping with the reference gene annotation[7](Table 4). The 'Lovell' double haploid clones 2D and 5D contained 382 and 401 *R*-genes, respectively, for the latest versions. The number of novel *R*-gene loci in the *R*-gene annotation was between 43 and 53 for

Classifications	Lovell 2D' v3.0	Lovell 5D' v1.0	Reference 'Lovell 2D' v.2.0.a1	Chinese Cling	124 Pan	CN14
CNB	33	35	29	6	38	35
CNL	28	29	26	6	33	32
JNB	–	–	–	–	–	–
JNL	–	–	–	–	–	–
NLR	367	390	396	467	364	383
RNB	13	13	13	13	11	13
RNL	12	12	12	12	10	12
TNB	173	198	179	242	180	171
TNL	146	168	163	208	155	153
RxNB	172	175	162	214	156	165
RxNL	153	146	190	178	137	145
NBARC (Total)	445	485	471	579	439	452
NBARC (primary transcript/longest isoform)	382	401	N/A*	448	343	378
Unique Loci	43	53	28	241	119	32

Table 4. *R*-gene classifications that were annotated using the FindPlantNLR pipeline in *P. persica* genomes. Unique loci were determined by comparing gene annotation vs. NBARC annotation. *Primary transcripts and unique loci of the reference 'Lovell 2D' v2.0.a1 were not calculated to avoid redunacy with 'Lovel 2D' v3.0 in the synteny analyses.

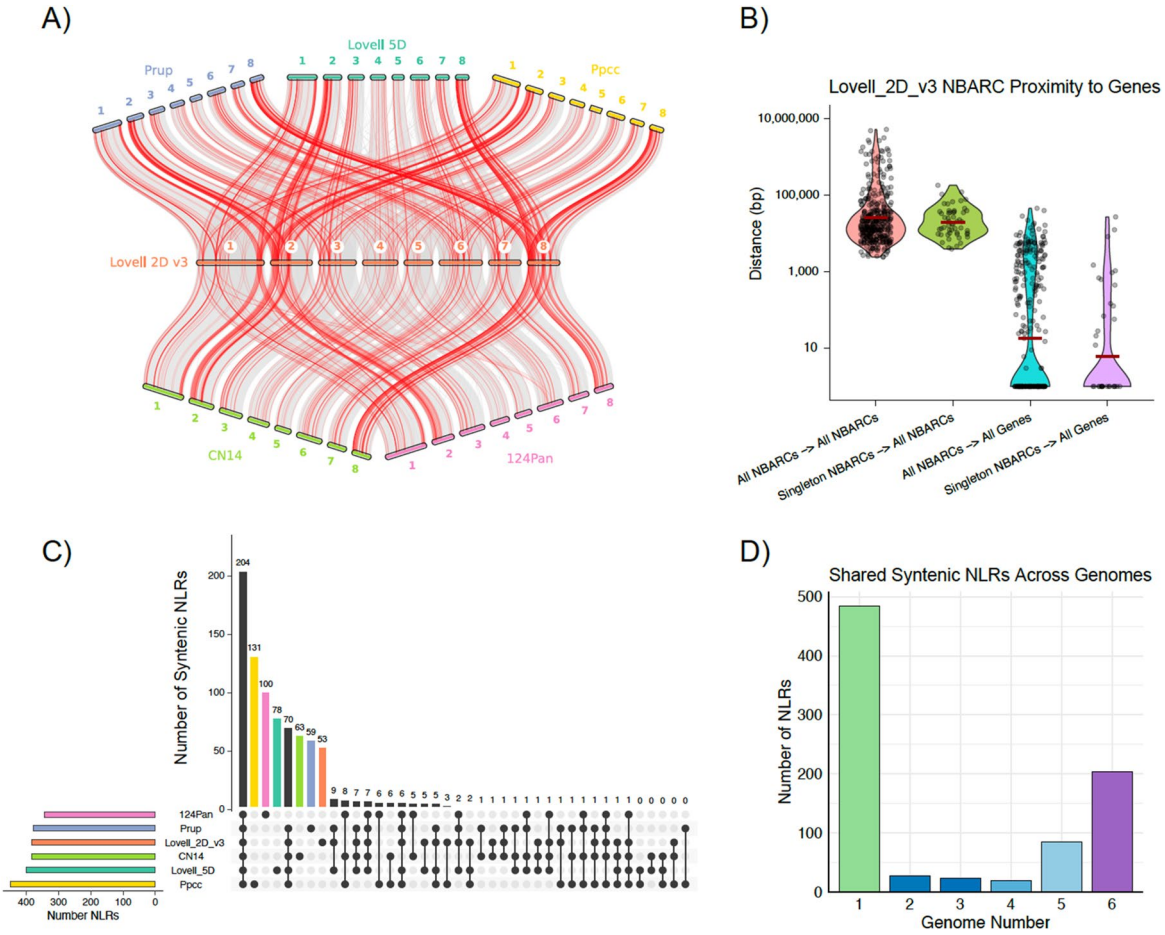

Fig. 3. Pan-NLR analyses for the six genomes of *P. persica*. (**A**) Macrosynteny (gray) of five *P. persica* genomes against the reference, Lovell 2D v3, along with syntenic orthologous NLR genes (red). (**B**) Nearest NLR neighbors in Lovell 2D v3, with the distance to the nearest NBARC shown in red, the distance to the nearest NBARC for singletons (those NBARCs unique to Lovell 2D v3) in green, the distance of NBARCs to the nearest coding gene in blue, and the distance of singleton NBARCs to the nearest coding gene in purple. (**C**) Upset plot of NBARCs in the pan-NLRome with those NBARCs unique to a single genome being colored according to that genome. (**D**) Number of NBARCs found to be unique to a single genome (green), shared among 2–5 genomes (shades of blue), and those found to be core (found in all six genomes, purple).

the 'Lovell' genomes. In comparison, the 'Lovell 2D' v.2.0.a1 genome had 471 *R*-genes annotated with the FindPlantNLR pipeline, and only 28 were novel compared to the reference gene annotation (Table 4)[4].

Syntenic analysis of NLR genes revealed that NBARC-containing loci are largely colinear with broader genome-scale macrosynteny, indicating minimal NLR-specific genome rearrangement across accessions (Fig. 3A). However, some exceptions were observed. For example, several NLRs located on chromosome 1 and chromosome 6 of the '124Pan' genome were syntenic with chromosomes 2 and 1, respectively, in the 'Lovell 2D' v3.0 genome, suggesting instances of NLR translocation. Interchromosomal translocations of R-genes are often mediated by transposable element movement and can facilitate the introduction of novel integrated domains in NLRs that have an impact on function and resistance[32,33].

We next examined the spatial distribution of NLR genes within the genome, focusing on proximity to other NLRs and annotated coding sequences. To do so, we calculated distances between each NBARC domain and its nearest neighboring NBARC, and also to the closest gene model from the structural annotation (Fig. 3B). In general, NLR genes were found to be more closely positioned to annotated coding genes than to other NLRs. This trend was also observed among singleton NLRs (defined as NLRs present in only one genome and lacking syntenic orthologs in the remaining accessions) suggesting that these loci are not preferentially clustered and may not arise from recent tandem duplication events.

We compared the conservation of NBARC loci across the pan-NLRome and showed that the majority of NBARC loci are shared among two to five genomes, with a substantial subset ($n = 204$) classified as core, being conserved across all six accessions (Fig. 3C and D). Singleton NLRs (those unique to a single genome) comprised 484 of the total NLRs identified, ranging from 53 in 'Lovell 2D' v3.0 to 131 in the 'Chinese Cling' genome (Fig. 3D). These data indicate a large reservoir of genome-specific NLR diversity across *P. persica* accessions, underscoring the dynamic nature of resistance gene evolution even within a single species. Together, these

insights provide a valuable framework for identifying candidate loci in future studies and for breeding programs aiming to harness the diversity of disease resistance genes for crop improvement.

Conclusion

In this study, we present improved chromosome-level genome assemblies for the doubled haploid peach genotypes 'Lovell 2D' and 'Lovell 5D', generated with Oxford Nanopore PromethION and PacBio HiFi. The resulting assemblies exhibit increased contiguity and completeness, allowing more accurate representation of repetitive genomic regions compared to previously published peach genome resources. Additionally, the new annotations generated here provide improved gene models and detailed characterization of transposable elements and non-coding RNAs. Our comparative genomic analyses confirm high synteny across peach genomes but also revealed structural variation that may underlie phenotypic and agronomic differences. Further, in-depth characterization of the NLR gene repertoire across multiple peach accessions highlights the diversity within disease resistance loci, providing valuable insights for future breeding strategies aimed at developing resilient peach cultivars. Collectively, these improved genomic resources establish a foundation for future genetic, functional, and evolutionary studies, and will accelerate marker-assisted selection and breeding programs focused on sustainable peach production. Upon acceptance of the manuscript, the genome data will be integrated into GDR, allowing users to access marker, GWAS, and QTL data aligned to the genome sequences, along with additional functional annotations and syntenic regions across key Rosaceae genomes through search pages and graphical interfaces.

Data availability

Nanopore and HiFi sequence data has been deposited on to the NCBI SRA database under BioProject PRJNA1274434 ([https://dataview.ncbi.nlm.nih.gov/object/PRJNA1274434?reviewer=kk7sackpvep1p5pgo4vqiha2l2](https:/dataview.ncbi.nlm.nih.gov/object/PRJNA1274434?reviewer=kk7sackpvep1p5pgo4vqiha2l2)). The genome assembly and gene annotations have been deposited on the GDR under accession number tfGDR1086 ([https://www.rosaceae.org/Analysis/24764266](https:/www.rosaceae.org/Analysis/24764266) and [https://www.rosaceae.org/Analysis/24764267](https:/www.rosaceae.org/Analysis/24764267)) (Jung *et al.* , 2019). Additionally, the genome and annotations for genes and NLRs have been deposited in a Zenodo Repository DOI: 10.5281/zenodo.15576776 ([https://zenodo.org/records/15576776](https:/zenodo.org/records/15576776))). Inquires on data availability can be made to the corresponding author CG.

Received: 24 June 2025; Accepted: 28 March 2026
Published online: 21 April 2026

References

1. Foreign Agricultural Service USDA. *Fresh Peaches and Cherries: World Markets and Trade* (Foreign Agricultural Service USDA, 2023).
2. Luo, C. X., Schnabel, G., Hu, M. & De Cal, A. Global distribution and management of peach diseases. *Phytopathol. Res.* **4**, 30 (2022).
3. Verde, I. et al. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494 (2013).
4. Verde, I. et al. The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genom.* **18**, 225 (2017).
5. Yu, Y. et al. Population-scale peach genome analyses unravel selection patterns and biochemical basis underlying fruit flavor. *Nat. Commun.* **12**, 3604 (2021).
6. Zhang, A., Zhou, H., Jiang, X., Han, Y. & Zhang, X. The Draft Genome of a Flat Peach (Prunus persica L. cv. '124 Pan') Provides Insights into Its Good Fruit Flavor Traits. *Plants (Basel)*. **10**, 538 (2021).
7. Cao, K. et al. New high-quality peach (Prunus persica L. Batsch) genome assembly to analyze the molecular evolutionary mechanism of volatile compounds in peach fruits. *Plant J.* **108**, 281–295 (2021).
8. Lian, X. et al. De novo chromosome-level genome of a semi-dwarf cultivar of Prunus persica identifies the aquaporin PpTIP2 as responsible for temperature-sensitive semi-dwarf trait and PpB3-1 for flower type and size. *Plant Biotechnol. J.* **20**, 886–902 (2022).
9. Driguez, P. et al. LeafGo: Leaf to Genome, a quick workflow to produce high-quality de novo plant genomes using long-read sequencing technology. *Genome Biol.* **22**, 256 (2021).
10. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
11. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
12. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
13. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
14. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
15. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10**, 645–656 (2013).
16. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
17. Formenti, G. et al. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* **38**, 4214–4216 (2022).
18. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
19. Castanera, R. et al. A phased genome of the highly heterozygous 'Texas' almond uncovers patterns of allele-specific expression linked to heterozygous structural variants. *Hortic. Res.* **11**, uhae106 (2024).
20. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
21. Jung, S. et al. 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res.* **47**, D1137–D1145 (2019).
22. Cheng, C. Y. et al. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* **89**, 789–804 (2017).

23. Mansfeld, B. N. et al. A haplotype resolved chromosome-scale assembly of North American wild apple Malus fusca and comparative genomics of the fire blight Mfu10 locus. *Plant J.* **116**, 989–1002 (2023). GottschalkC.
24. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
25. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protocols.* **1**, e323 (2021).
26. Chen, S. H. et al. A high-quality pseudo-phased genome for Melaleuca quinquenervia shows allelic diversity of NLR-type resistance genes. *GigaScience* **12**, giad102 (2023).
27. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. (2020). Available at: https://f1000research.com/articles/9-304 [Accessed July 10, 2023].
28. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
29. Tang, H. et al. Synteny and Collinearity in Plant Genomes. *Science* **320**, 486–488 (2008).
30. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49. (2012).
31. Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species. *Plant. Mol. Biol. Rep.* **9**, 208–218 (1991).
32. Bailey, P. C. et al. Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biol.* **19**, 23 (2018).
33. Krasileva, K. V. The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. *Curr. Opin. Plant. Biol.* **48**, 18–25 (2019).
34. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genomeprojects. *BMC Bioinformatics.* **12**, 491, PMID: 22192575; PMCID: PMC3280279, (2011).
35. Hoff KJ, Stanke M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics.* **65**(1), e57, 2018 Nov 22. PMID: 30466165, (2019).
36. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29**(22), 2933-5, Epub 2013 Sep 4. PMID: 24008419; PMCID: PMC3810854 (2013).

Author contributions

C.G. and C.D. supervised and conceptualized the study. C.G., J.R.B, B.N.M., and C.D. conducted the analyses, wrote the main manuscript text, and prepared the figures and tables. D.M., S. J., and P.Z. integrated the resulting genomes into public databases and transferred existing information from previous genome versions onto the new revision. C.V., M.D., D.B., and Z.L., assisted in material collection and conducting experiments. All authors edited and reviewed the manuscript.

Funding

Funding for the sequencing, assembly, and annotation were provided by appropriated funds award to CG and CD under USDA-ARS National Program 301 Plant Genetic Resource, Genomics and Genetic Improvement, project number 8080-21000-033-000D. Indexing, hosting, and further analysis of the genomic resources provide by GDR was made possible from funding awarded to SJ, PZ, and DM under SCRI-NIFA Award 2022-51181-38449.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-46952-6.

Correspondence and requests for materials should be addressed to C.G. or C.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.