A Haplotype-resolved, Chromosome-scale Genome for *Malus domestica* 'WA 38'

Huiting Zhang[1,2,*], Abigail Eaker[3,*], Itsuhiro Ko[3,*] ,Sabrina Haney[4], Ninh Khuu[3], Kara Ryan[5], Aaron B. Appleby[6], Brendan Hoffmann[7], Henry Landis[5], Kenneth Pierro[7], Noah Willsea[8], Heidi Hargarten[2], Alan Yocca[2,9], Alex Harkess[9], Loren Honaas[2], Stephen Ficklin[1,**]

[1] Department of Horticulture, Washington State University, Pullman, WA, 99164
[2] USDA Agricultural Research Service, Wenatchee, WA, USA 98801
[3] Department of Plant Pathology, Washington State University. Pullman, WA, 99164
[4] Department of Animal Science, Washington State University. Pullman, WA, 99164
[5] School of Biological Sciences, Washington State University. Pullman, WA, 99164
[6] Department of Crop and Soil Science, Washington State University. Pullman, WA, 99164
[7] Integrated Plant Sciences Program, Washington State University. Pullman, WA, 99164
[8] Department of Horticulture, WSU Tree Fruit Research and Extension Center, Wenatchee, WA, 98801
[9] HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806

* Co-first author.
** Corresponding author.

Abstract

Genome sequencing for agriculturally important Rosaceous crops has made rapid progress both in completeness and annotation quality. Whole genome sequence and annotation gives breeders, researchers, and growers information about cultivar specific traits such as fruit quality, disease resistance, and informs strategies to enhance postharvest storage. Here we present a haplotype-phased, chromosomal level genome of *Malus domestica*, 'WA 38', a new apple cultivar released to market in 2017 as Cosmic Crisp ®. Using both short and long read sequencing data and a k-mer based approach, chromosomes originating from each parent were assembled and segregated. This is the ***first*** haplotype-resolved assembly in pome fruit genomes, a milestone in the new genomic era. The two haplome assemblies, 'Honeycrisp' originated HapA and 'Enterprise' originated HapB, are about 650 Megabases each, and both have a BUSCO score of 98.7% complete. A total of 53,028 and 54,235 genes were annotated from HapA and HapB, respectively. Additionally, we provide genome-scale comparisons to 'Gala', 'Honeycrisp', and other relevant cultivars highlighting major differences in genome structure and gene family circumscription. This assembly and annotation was done in collaboration with the American Campus Tree Genomes project that included 'WA 38' (Washington State University), 'd'Anjou' pear (Auburn University), and many more. To ensure transparency, reproducibility, and applicability for any genome project, our genome assembly and annotation workflow is recorded in detail and shared under a public GitLab repository. All software is containerized, offering a simple implementation of the workflow.

Keywords

Apple genomics, *Malus domestica* 'WA 38', genome sequence, comparative genomics, plant genomics, haplotype-resolved assembly, genome annotation

Background

For economically important crop species, having full-resolution reference genomes aids in the understanding of traits associated with commodity quality, disease resistance, long-term storage, and shelf life. Apple (*Malus domestica*) is the number one consumed fruit in the United States, with a Farm-Gate Revenue of $3.2 billion in the U.S.[1], and $78 billion globally[2]. There are over 7,000 apple varieties grown world wide[3], each with unique colors, flavors, and textures[4]. Therefore, a single genome is unlikely to capture the complexity of all cultivars within this highly heterozygous species (H. Zhang et al. 2022; W. Li et al. 2022). One such cultivar is 'WA 38', commercially released as Cosmic Crisp® in 2017 by the Pome Fruit Breeding Program at Washington State University's (WSU) Tree Fruit Research and Extension Center (Figure 1). 'WA 38' is a cross between 'Honeycrisp' and 'Enterprise', made using classical breeding methods in 1997. One parent, 'Honeycrisp', is well-known for its crisp texture, firmness retention in storage, disease resistance, and cold hardiness, but is highly susceptible to production and postharvest issues (Khan et al. 2022). The other parent, 'Enterprise', is an easy-to-grow cultivar that has extended postharvest storage capabilities, however it is not widely cultivated commercially due to its less desirable eating quality (Crosby et al. 1994). Their resulting cross has been met with favorable reviews for its appealing color, texture, flavor, cold hardiness, long-term storage capabilities (>1 year), and scab resistance (Evans et al. 2012). However, it inherited undesirable traits as well, such as a propensity for physiological symptoms that may be related to mineral imbalances (Sheick et al. 2023; Sallato, Whiting, and Munguia 2021), maturity at harvest[5], and an 'off' flavor that has been brought up by consumers that may be the result of improper picking times, crop load management, handling/packing practices or other post harvest processes[6]. Most concerning is green spot (Figure 1 C, D, and F), a corking disorder that seems to be unique to 'WA 38', but with etiology similar to disorders associated with mineral imbalances such as bitter pit and drought spot (Sheick et al. 2022, 2023). The propensity for and cause of physiological disorders often differs on a cultivar-by-cultivar basis (Pareek 2019), and a genetic basis for such predispositions is likely (Di Guardo et al. 2013; Lum

---

[1] https://usapple.org/industry-at-a-glance
[2] https://fruitgrowersnews.com/news/global-apple-market-reached-78m-set-to-continue-moderate-growth/
[3] https://waapple.org/did-you-know/
[4] Retrieved December 27, 2022, from https://plants.ces.ncsu.edu/plants/malus-domestica/common-name/apples/
[5] https://www.actahort.org/members/showpdf?session=1982017
[6] https://treefruit.wsu.edu/article/2022-update-additional-wa-38-harvest-and-storage-considerations/

et al. 2016; Johnston and Brookfield 2012; Liebhard et al. 2003). Thus, improved resolution of cultivar-specific genomic differences is critical for advancing our understanding of how economically important traits, such as physiological disorders, are inherited and how they can be managed more efficiently.

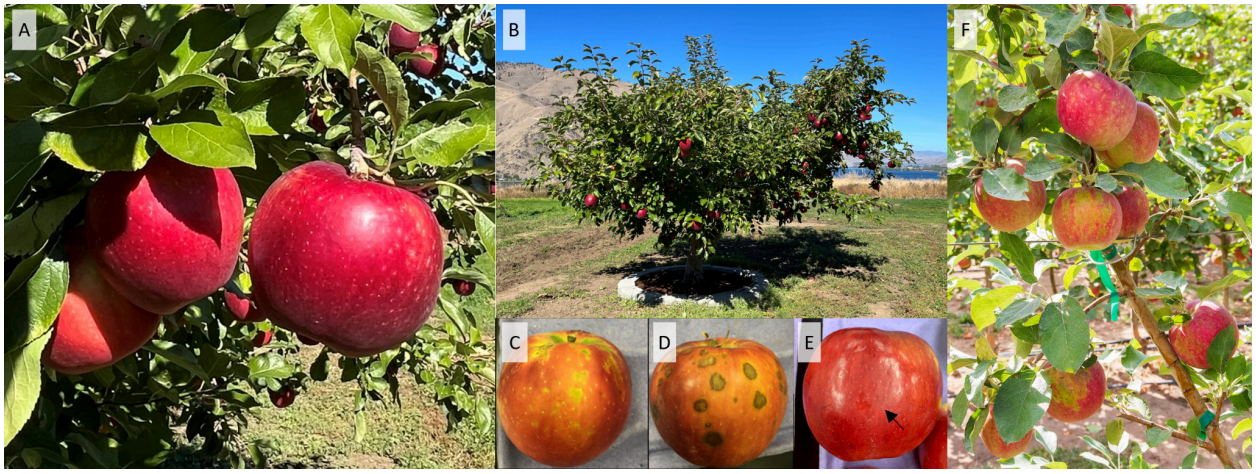

Figure. 1. 'WA 38', a cultivar of apple developed by the Washington State University Apple Breeding Program (a cross between 'Honeycrisp' and 'Enterprise'), marketed as Cosmic Crisp®. ***A)*** *'WA 38' apples ready for harvest on the mother tree, located at the WSU and USDA-ARS Columbia View Research Orchard near Orondo, WA, USA. B) The 'WA 38' mother tree. C & D) Green spot, a corking disorder which results in green blemishes on the fruit peel and brown, corky cortex tissue. Symptom severity generally increases during fruit maturation and time in storage, resulting in cullage. E) Natural peel greasiness as a result of more advanced maturity at harvest can interfere with artificial waxes applied in the packinghouse after removal from postharvest storage, creating unappealing, dull spots. F) Green Spot symptoms can begin to appear while fruit is still developing on the tree. Photo Credits: A&B: Heidi Hargarten/ USDA-ARS; C&D: Bernardita Sallato/WSU; E: Carolina Torres/WSU; F: Ross Courtney/Good Fruit Grower.*

To develop full resolution reference genomes of superior quality, having skilled bioinformaticians is required. To train the next generation of bioinformaticians for agricultural genomic research, a national effort spearheaded by Auburn University, HudsonAlpha Institute for Biotechnology, and Washington State University was started in 2021 - The American Campus Tree Genomes project (ACTG). ACTG aims to break through institutional barriers that have traditionally prevented many students from accessing valuable, hands-on research projects and experience

in bioinformatics[7]. To accomplish this goal, a course has been developed to involve students in genome projects from inception, through analysis, to publication[8]. During the course, students learn genome assembly and annotation workflows using the raw sequence data from genomes of beloved trees (e.g., Toomer's oak and ‘d’Anjou’ pear (Yocca et al. 2024) at Auburn University, Sabal palm at University of South Carolina) and are listed as authors on the final publication. The ‘WA 38’ genome introduced here was developed through ACTG by students from Washington State University, presenting three major outcomes: 1) a fully annotated, chromosomal level, haplotype-phased genome of ‘WA 38’ utilizing PacBio HiFi, Dovetail Omni-C, and Illumina DNA and RNA sequencing data, 2) identification of unique regions of interest using a comparative genomics approach from the other economically important *M. domestica* cultivars including ‘Gala’ and ’Honeycrisp’, and 3) establishment of a containerized, reproducible, flexible, high performance computing workflow for complete genome assembly and annotation (Figure 2, Supplemental Figure 1).

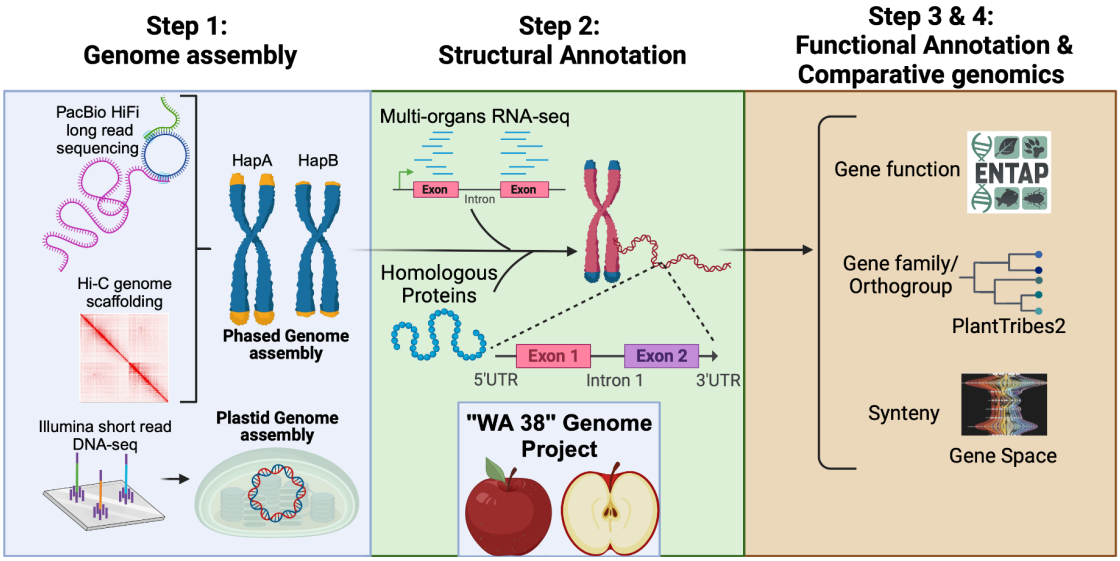

Figure. 2 Schematic chart of “WA 38” genome project.

7 https://hudsonalpha.shorthandstories.com/learning-from-the-trees/&sa=D&source=docs&ust=1704831742820941&usg=AOvVaw1rRBKfG02GkZH8ciIAxNwn

8 https://actg-wgaa.readthedocs.io/en/latest/index.html

Results

A Complete, Reproducible, Publicly-available Workflow

To ensure transparency and reproducibility, the 'WA 38' Whole Genome Assembly and Annotation (WA 38 WGAA) project workflow was made publicly accessible through a public GitHub repository[9]. This workflow includes detailed method documentation of scripts executed for each analysis with parameter settings, as well as any command line manipulation of the files generated. Summary diagrams for the workflow are available in Supplemental Figure S1. To enhance ease of use and reproducibility across diverse computational systems, each software utilized in the project has been containerized and shared on Docker Hub[10]. Any user that follows the workflow can successfully complete the analysis and reproduce the results. Furthermore, this workflow is designed to provide training that is applicable for a wide range of species. In fact, we are currently adapting this versatile methodology into a generic, species-agnostic workflow that will serve the broader American Campus Tree Genome (ACTG) community. As part of our commitment to knowledge sharing, we have initiated an ACTG course GitHub repository[11], leveraging resources from the WA 38 WGAA project. This repository comprises three main sub-repositories: 1) wgaa-compute: A generic whole genome assembly and annotation workflow template, derived from the WA 38 WGAA project, that can be adapted for other species; 2) wgaa-docker: The Docker recipes for all the software employed in the project; and 3) wgaa-doc: An interactive documentation repository containing teaching materials for current and future ACTG instructors, providing a collaborative space for instructors to learn from and contribute to the enhancement of the course materials.

Nuclear Genome Assembly

Sequence Quality Assessment

To begin, raw sequencing data (Table 1) was assessed for read quality. The Illumina shotgun short read data consisted of 807.2 million total reads with a mean length of 151bp for a total of 121.9 Gigabytes (GB) of data after adapter trimming. Filtered Illumina data GC content is 38.3% and has 91.8% Q20 bases and 83.4% Q30 bases. Duplication rates ranged from 23.3% to

---

[9] https://gitlab.com/ficklinlab-public/wa-38-genome
[10] Docker images can be accessed at https://hub.docker.com/u/systemsgenetics
[11] https://github.com/actg-course/

27.8%. PacBio long read raw data consisted of 3.9 million reads from 499-49,283bp in length for a total of 60 GB. Sequence duplication rates ranged from 2.2% to 2.4%. PacBio sequence GC content is 38%. These metrics indicate high quality sequences from both short and long read technologies.

Table 1. Yield of Illumina DNA short reads (Shotgun and Omni-C) and PacBio HiFi sequencing reads from young leaf tissues of 'WA 38'.

	Long Read	Short Read			
	PacBio HiFi	Shotgun DNA seq		OmniC-Seq	
		Read1	Read2	Read1	Read2
Total read	3,870,263	403,610,448	403,610,448	865,134,180	865,134,180
Number of bases (Gb)	60	121.89	121.89	130.6	130.6
Average length (bp)	15,495	151	151	151	151

Genome complexity

Using a *k*-mer frequency approach, genome characteristics such as heterozygosity and genome size were estimated (Figure 3). Analysis of both short and long reads resulted in an estimated heterozygosity of ~1.35%, similar to estimates from the 'Honeycrisp' cultivar (1.27% (Khan et al. 2022)). Estimate genome size ranged from 466.5Mb from the short reads and 606Mb from the long reads. These estimates are lower than expected from other apple genomes ('Honeycrisp': 660-674Mb, (Khan et al. 2022) and 'Golden Delicious': ~701Mb, (Xuewei Li et al. 2016)). Additionally, the percent of unique sequence was estimated at 69.5% for the short reads and 53.4% for the long reads, with the long read estimate being more consistent with what is expected from the 'Honeycrisp' (51.7%) (Khan et al. 2022) and of wild apple species *Malus baccata* (58.6%) (X. Chen et al. 2019).

Genome assembly, scaffolding, and curation

Two phased haplomes, hereafter named Hap1 and Hap2, were assembled and scaffolded using both PacBio long reads and Omni-C short reads. Hi-C maps of the haplome assemblies show no mis-assemblies (Supplemental Figure S2). For Hap1 and Hap2, a total of 22 joins and 20 joins, respectively, were made in the scaffolding step to build the final assemblies into 17 chromosomes each, with the remaining scaffolds representing unincorporated contigs.

Unincorporated contigs were investigated and found to be bacterial or other contamination and were removed. After removing contaminants, Hap1 is 645.41 Mbp in length with an N50 of 36.1 Mbp, while Hap2 is 651.07 Mbp in length with an N50 of 37.2 Mbp. Additional assembly statistics for both haplomes are included in Supplemental Table S1. 'WA 38' has a comparable genome size to other previously sequenced apple cultivars, including its parent 'Honeycrisp' (Khan et al. 2022). Notably, the 'WA 38' scaffold N50 is among the longest across all published apple genomes, indicating high levels of assembly contiguity (Supplemental Table S2).

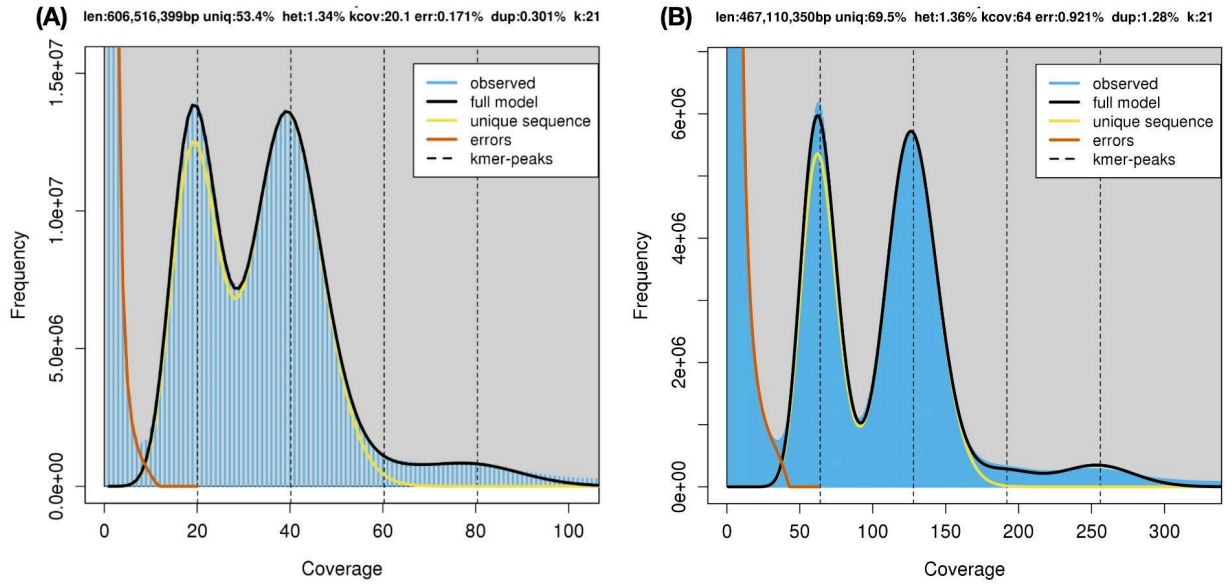

Figure.3 Genome complexity of 'WA 38' genome using PacBio long read data (A) and Illumina short read (B). The output figure was generated by GenomeScope (k=21).

Haplotype-binning, Structural Comparison and Completeness Assessment

The K-mer based binning method identified the origin of chromosomes in each haplome assembly. Ten out of the 17 chromosomes in Hap1 originated from 'Honeycrisp', while the other seven were from 'Enterprise'. After reorganizing the chromosomes, the haplome containing all the 'Honeycrisp' origin chromosomes is designated as HapA, whereas the 'Enterprise' originated haplome is designated as HapB. HapA and HapB are structurally similar; a total of ~44Mb are affected by structural variants and are mainly contributed by indels and repeat expansion and contractions (Supplemental Table S3 and Supplemental Figure S3). Additionally, three large inversions are observed on chromosomes 1, 11, and 13 (Supplemental Figure S4). Based on the BUSCO analysis, both the HapA and HapB assemblies were ~98.7% complete, with only ~19 BUSCOs missing and ~12 partially detected (Supplemental Table S4). This BUSCO score suggests high genome completeness for both haplomes, comparable to the

‘Honeycrisp’ and ‘Fuji’ apple genome assemblies which is the most contiguous of all apple genomes to date (Table 2 and Supplemental Table S2) (Khan, et al., 2022, Li et al., 2024).

Table 2. Comparison of genomic features and assembly statistics of current assembly of ‘WA-38’ genome and previously published genomes of apples.

	‘WA 38’		‘Honeycrisp’		‘Antonovka’	‘Gala’	GDDH13	‘Fuji’
	HapA	HapB	HapA	HapB				
Number Scaffold	17	17	473	215	168	812	1,081	1,358
Haploid genome size (Mbp)	645.41	651.07	674	660	643.5	652.4	709.6	736.9
N50 (Mbp)	36.1	37.2	31.6	32.8	35.85	23.9	5.5	36.8
L50	8	8	8	8	8	8	NA	9
Number of protein-coding gene	53,028	54,235	47,563	48,655	45,085	45,352	40,175	49,972
Complete BUSCO (%) Assembly	98.7	98.7	98.6	98.7	97.6	97.9	98	98.8
Complete BUSCO (%) Annotation	98	98.4	96.8	97.4	97.25	95.5	96.1	97.2
Number of orthogroups in 26Gv2	10,494	10,511	10,350	10,366	10,293	10,095	10,117	10,243
Reference	This paper		Khan et al. 2022		Švara et al., 2023	Sun et al. 2021	Daccord et al. 2017	Li et al., 2024

*NA: Data not available. For consistency, genome statistics and BUSCO analyses were performed on the publicly available genomes using the same methods for ‘WA 38’, except for N50 and L50 of GDDH13 as the scaffold assembly is not publicly available. ‘*Antonovka*’ data is the average of the two haplomes. The unphased version of ‘Fuji’ was used. A more in-depth comparison is available in Supplemental Table S2.*

Nuclear Genome Structural Annotation

Repeat annotation

In both haplomes, approximately 58.7% of the assembly was predicted to be repetitive regions by EDTA (Cite, Table 3). RepeatMasker identified an additional 4% repeat elements, resulting in

a total of 62.7% repeat regions in both HapA and HapB, comparable to the 'Honeycrisp' genome (Khan et al. 2022). In both haplomes, the most dominant type of repeat element is long terminal repeat (LTR), followed by Terminal Inverted Repeat (TIR) (Table 3, Supplemental Table S5), consistent with that in 'Honeycrisp'. We also compared the repeat landscape of 'WA 38' with 'd'Anjou' pear which was annotated with the same methodology. While they share the major repeat classes, 'd'Anjou' pear has a much lower percentage of repeat elements (Table 3).

Table 3. Summary of repetitive element annotation in the 'WA 38' and other apple genomes.

	Class	'WA 38' HapA (%)	'WA 38' HapB (%)	Honeycrisp HapA (%)	Honeycrisp HapB (%)	'd'Anjou' pear Hap1 (%)	'd'Anjou' pear Hap2 (%)
LTR	Copia	9.37	10.22	9.73	9.6	5.6	5.73
	Gypsy	17.19	18.32	20.29	17.8	12.32	12.88
	unknown	16.37	14.52	14.89	16.86	8.46	10
TIR	CACTA	1.94	2.15	2.21	1.95	1.4	1.4
	Mutator	3.96	4.18	4.16	4.25	3.47	3.41
	PIF Harbinger	2.4	2.52	2.43	2.6	1.81	1.81
	Tc1_Mariner	0.16	0.24	0.15	0.27	0.13	0.11
	hAT	2.15	2.37	2.3	2.31	0.58	0.84
	polinton	0	0	0	0.01	0	0
nonLTR	LINE_elemnet	0.14	0.16	0.18	0.17	0.14	0.14
	unknown	0.11	0.09	0.09	0.18	0.06	0.06
nonTIR	helitron	3.41	2.20	2.95	3.18	1.56	1.92
Repeat region		1.52	1.74	2.91	2.78	3.98	4.22
RM*		3.98	3.99	NA	NA	NA	NA
Total		62.71	62.71	62.43	61.97	39.78	42.52
Reference		This paper		Khan et al. 2022		Yocca et al., 2023	

** repeat regions annotated by RepeatMakser*

Gene space annotation

To annotate the gene space, we utilized a combination of *ab initio* prediction and evidence-based prediction with transcriptome and homologous protein, functions implemented in BRAKER2. However, BRAKER2 was unable to annotate UTR regions and yielded erroneous splice variants (Supplemental Figure S5). Therefore, the gene models were further processed with PASA (Haas et al. 2003) and a custom script. A total of 53,028 and 54,235 genes were annotated from HapA and HapB, respectively, more than most published apple genomes (Table 2, Supplemental Table S6). The complete BUSCO scores for HapA and HapB annotations are 98.5% and 98.4%, respectively, the highest score among all *M. domestica* genomes sequenced to date (Supplemental Table S2). The average protein annotated from HapA and HapB contains 361.3 and 356.4 amino acids, respectively, very similar to protein lengths from other *M. domestica* annotations (Supplemental Table S7). On average, ~1.3 splice variants were identified for each gene in both HapA and HapB annotations. The only other apple genome with splice variant annotation is 'Honeycrisp', and on average, 1.05 splice variants were annotated per gene (Supplemental Table S7). Additionally, 53.5% and 52.2% of the annotated transcripts from HapA and HapB, respectively, contain untranslated regions (UTRs). Notably, 'WA 38' is the only other apple genome besides the 'GDDH13' genome that has more than half of the genes annotated with UTRs.

The 'WA 38' genes were named in accordance with the convention following guidance from the Genome Database for Rosaceae (GDR). This convention was first proposed by our group for the 'Honeycrisp' genome and was later adopted with modification by GDR (Gene name example: *drMalDome.wa38.v1a1.ch10A.g00001.t1*). This convention meets recommendations proposed by the AgBioData consortium to reduce gene ID ambiguity and improve reproducibility.

Nuclear Genome Functional Annotation

EnTAP (Hart et al. 2020) functional annotation assigned functional terms to 89.5% and 88.8% of proteins annotated from HapA and HapB, respectively. Specifically, an average of 83% and 55% of all proteins (including both HapA and HapB) have strongly supported hits in the NCBI RefSeq (O'Leary et al. 2016) and UniProt database, respectively, 75% were annotated with an InterPro term, and 88% have functional annotations from at least one of the databases included in InterProScan. Additionally, EggNOG (O'Leary et al. 2016; Huerta-Cepas et al. 2019) search provided extra function information: 90% of the annotated proteins were assigned into EggNOG

orthogroups, 84% were annotated with protein domains, 21% were classified into KEGG pathways, and 63%, 53%, and 61% proteins were annotated with GO biological process, cellular component, and molecular function terms, respectively (Supplemental Table S8).

Comparative Analyses

Synteny and gene family analyses were performed to investigate the similarity and uniqueness of ‘WA 38’ genome to other closely related species and cultivars.

Synteny analysis was performed with genomes of ‘WA 38’, one of its parents, ‘Honeycrisp’, and the most commonly referenced apple genome, ‘GDDH13’, using GeneSpace (Figure 4).

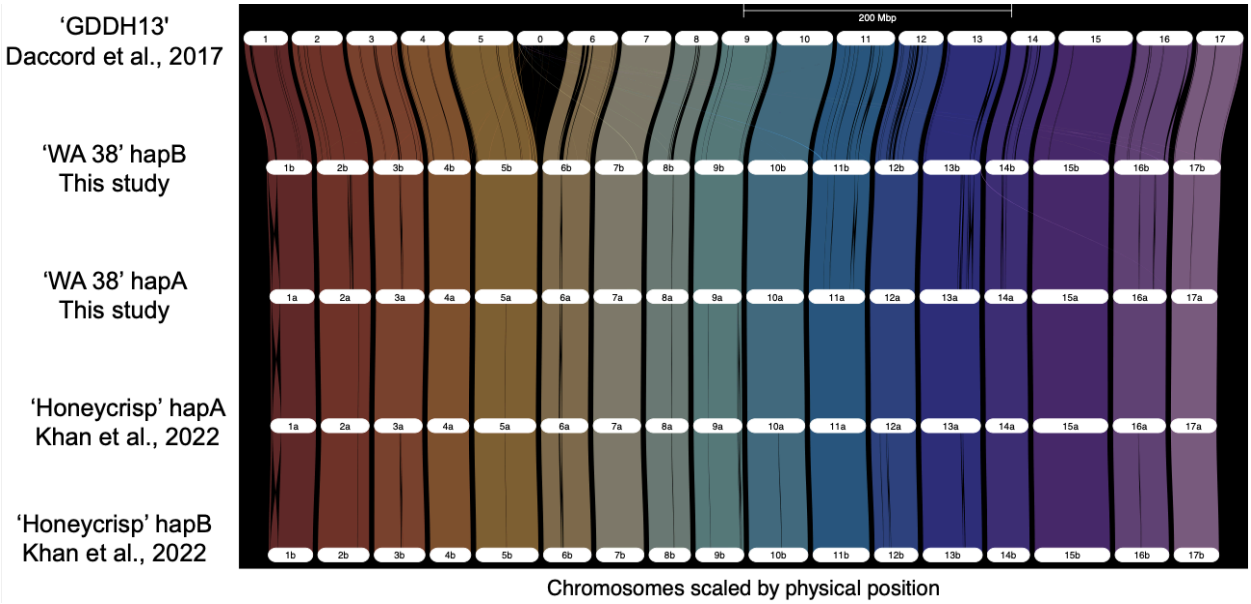

Figure 4. Riparian plot comparing ‘WA 38’ Haplotype A and B with ‘Honeycrisp’ Haplotype A and B and ‘Golden Delicious’ (GDDH13) genomes.

Gene family analysis is performed using PlantTribes2 and the pre-constructed 26Gv2-scaffold orthogroup database (Wafula et al. 2022). Out of the 18,110 pre-constructed orthogroups, proteins from all apple annotations (including 6 published scion cultivar genomes, 2 rootstock genomes, and the ‘WA 38’ genome from this work) are found in 11,698 orthogroups. ‘Golden Delicious’ Genome v1.0 (Velasco et al. 2010) was omitted from this analysis due to poor annotation quality. Proteins from HapA and HapB of ‘WA 38’ were classified into 10,494 and 10,511 orthogroups, respectively, similar or slightly higher in number compared to previously published *M. domestica* genomes, including ‘Honeycrisp’, ‘Gala’, and ‘GDDH13’ (Table 2, Figure

5). An investigation into shared and unique orthogroups across all the scion genomes showed that most orthogroups (8,800 or 75%) are shared by all six apple genomes considered. Additionally, 824 orthogroups are shared by both ‘WA 38’ haplomes and the nine other annotations (the two haplomes from ‘Honeycrisp’ and ‘Antonovka 172670-B’ are counted as two annotations). Annotations that shared the largest number of orthogroups with ‘WA 38’ is ‘Honeycrisp’, as expected due to being one parent of ‘WA 38’ (Supplemental Table S9). These results indicate that the ‘WA 38’ annotations capture genes in virtually all *M. domestica* orthogroups. Additionally, 39 orthogroups were unique to ‘WA 38’ (i.e. present only in the two ‘WA 38’ haplomes) and each haplome of ‘WA 38’ contains 44 unique orthogroups (Figure 5).

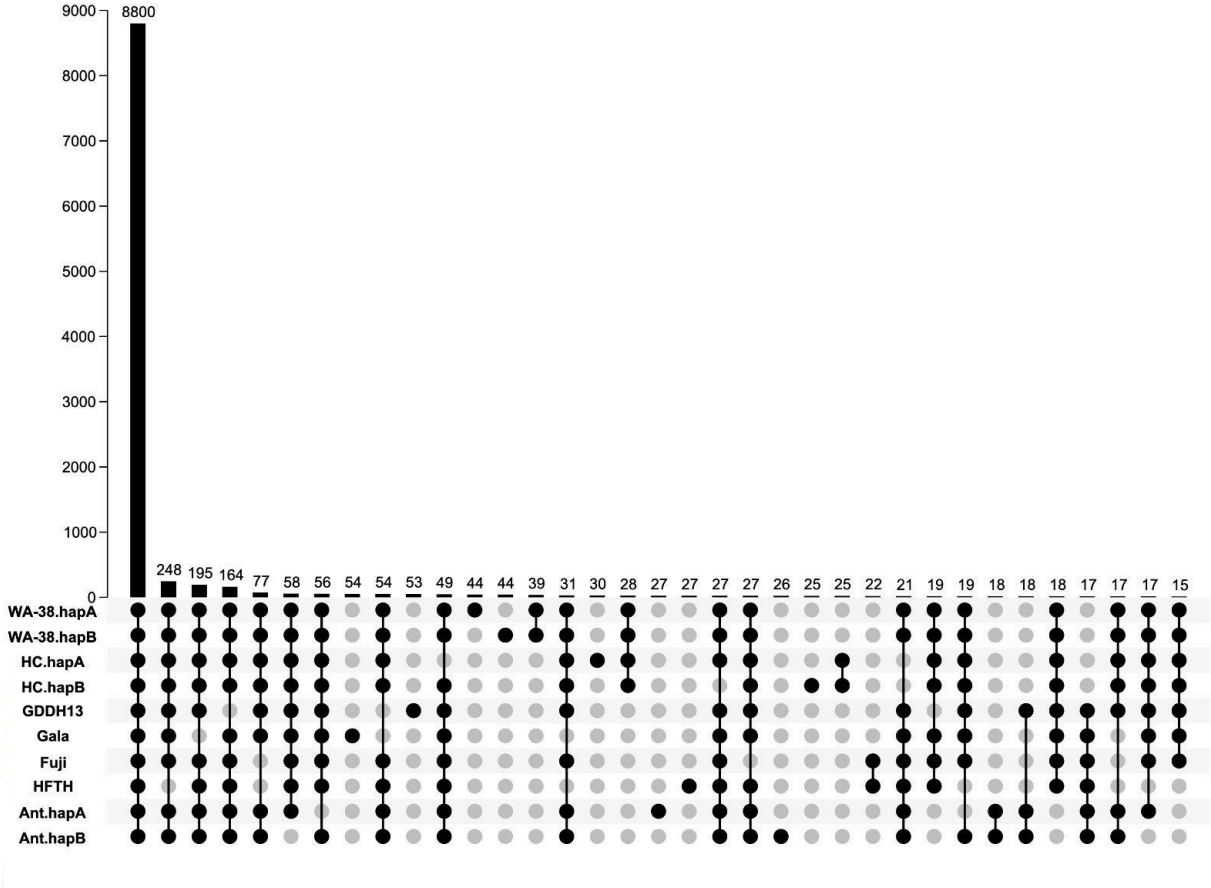

Figure. 5 Upset plot of shared and unique orthogroups among Malus domestica genomes. Rows in the bottom of the figure are genomes used for the comparison. Columns (categories, x-axis of the bar graph) are annotated with black or gray dots where black is present and gray is absent. The height of the black bars (y-axis of the bar graph) is scaled to match the number of orthogroup in each category, which are also printed above the bars.

In addition to identifying the shared and unique orthogroup, a CoRe OrthoGroup (CROG) - Rosaceae analysis is performed to further investigate orthogroup contents. As expected, in the CROG gene count clustermap (Figure 6), ‘WA 38’ clustered closely with ‘Honeycrisp’. The ‘WA 38’ + ‘Honeycrisp’ group is clustered with ‘GDDH13’, as expected based on pedigree. Interestingly, a strong ‘publication’ bias, first mentioned by Wafula et al., 2023, is observed: genomes released in the same publication or annotated by the same person clustered together. Such groups are: ‘Gala’, *Malus sieversii*, and *M. sylvestris*; ‘Fuji’, ‘M19’, and ‘MM106’; *M. fusca* and *Pyrus communis* ‘d’Anjou’; ‘Honeycrisp’ and ‘WA 38’. The CROG gene count *z*-score box plot shows (Figure 7) that the average *z*-score of ‘WA 38’ gene counts are slightly higher than expected (with 0 as the perfect score), indicating that there are a number of CROGs containing more genes from the ‘WA 38’ annotations compared to other apples.

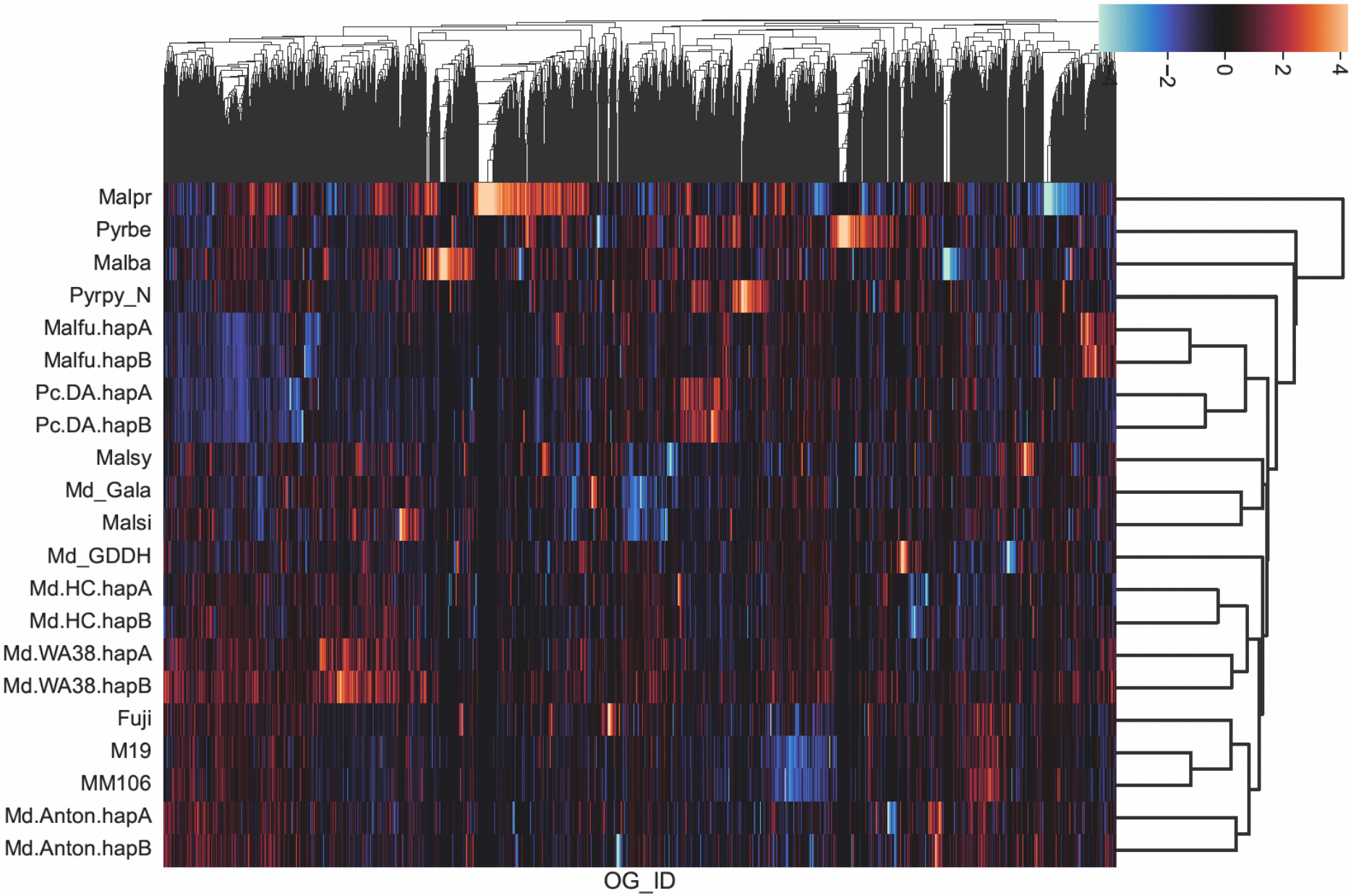
Malpr
Pyrbe
Malba
Pyrpy_N
Malfu.hapA
Malfu.hapB
Pc.DA.hapA
Pc.DA.hapB
Malsy
Md_Gala
Malsi
Md_GDDH
Md.HC.hapA
Md.HC.hapB
Md.WA38.hapA
Md.WA38.hapB
Fuji
M19
MM106
Md.Anton.hapA
Md.Anton.hapB
OG_ID
−4
−2
0
2
4

Figure 6. CoRe OrthoGroup (CROG) - Rosaceae gene count clustermap. Each row represents a CROG and each column represents a genomes. Color indicates the number of genes in each cell relative to the row average (z-score). Warmer color means more genes. Cooler color means fewer genes. The darker a color, the closer the value is to the row average.

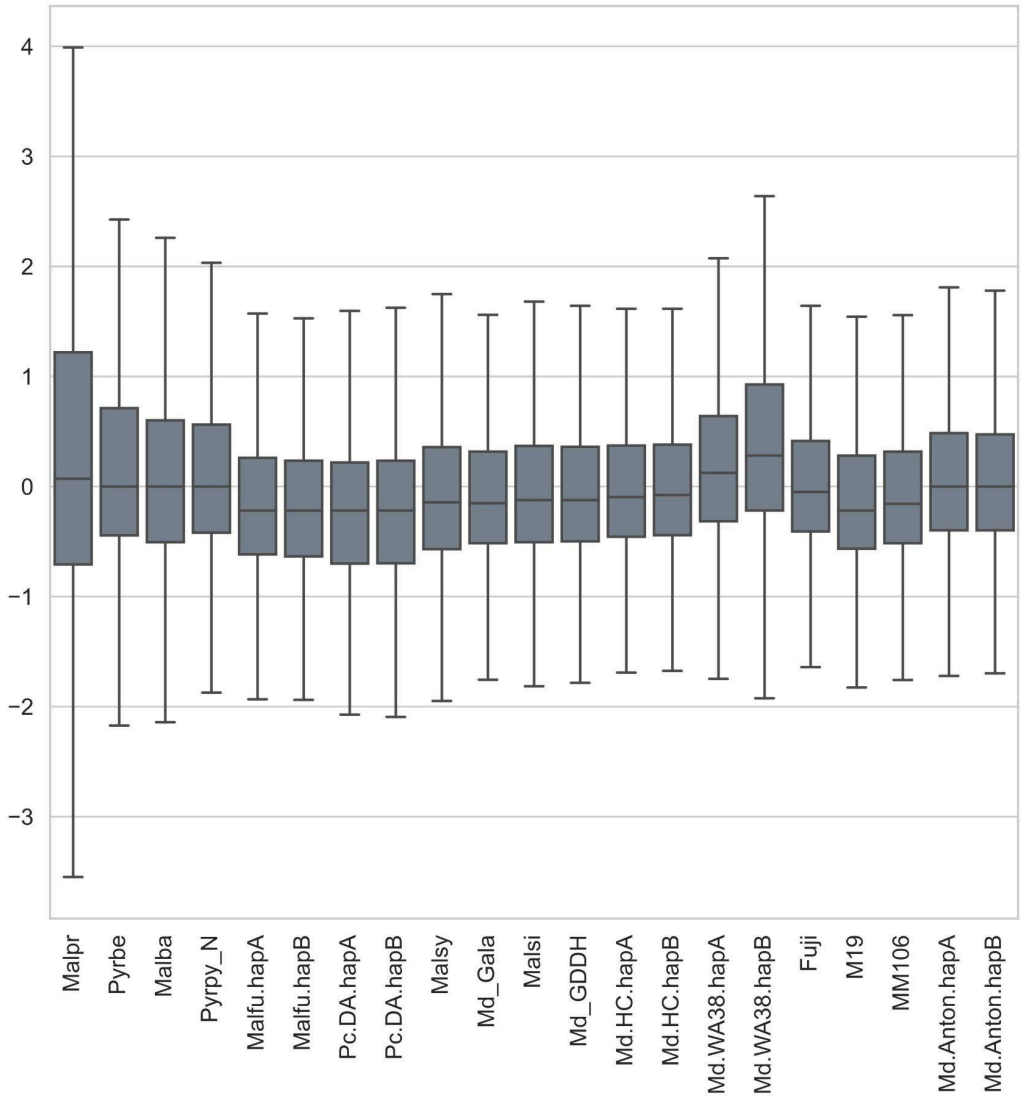

Figure 7. Boxplot summarizing z-score distribution of CROG gene counts in pome fruit genomes.

Plastid Genomes Assembly and Annotation

The chloroplast genome of the ‘WA 38’ apple is 159,915 bp in length, which is smaller than most assembled *M. sierversii* chloroplast genomes (Xueli Li et al. 2022; Miao et al. 2022; Naizaier et al. 2019; Yan et al. 2019; Zhao et al. 2019; Ha et al. 2020). The plastome consisted of a typical

quadripartite structure with a pair of inverted repeat regions (IR) of the same length (26,352 bp) separated by a large -single copy region (LSC) (88,052 bp) and a small -single copy region (SSC) (19,159 bp). The inverted repeat regions and the SSC regions were all similar in length to that of other *M. sierversii* cultivars (Xueli Li et al. 2022; Miao et al. 2022; Naizaier et al. 2019; Yan et al. 2019; Zhao et al. 2019; Ha et al. 2020). A total of 134 unique genes were annotated, including 86 protein-coding genes, 42 tRNA genes, and 7 rRNA genes. Moreover, eight protein-coding genes (*ycf1, ycf2, rpl2, rpl23, ndhB, rps7, rps12, rps19-fragment*), ten tRNA genes (*trnE-UUC, trnI-GAU, trnA-UGC, trnL-CAA, trnM-CAU, trnN-GUU, trnR-ACG, trnI-CAU, trnN-GUU, trnV-GAC*), all four rRNA genes (*rrn16, rrn23, rrn4.5, rrn5*) were located wholly within the IR regions (Figure 7). Twelve protein-coding genes, eight tRNA genes, and one rRNA gene (*rrn16*) contain introns. The majority of which contained one intron (19 genes), with only two genes (*pafl* and *clpP1*) containing two introns.

The mitochondrial genome of the ‘WA 38’ apple is 451,423bp long and contains 64 annotated genes. This annotation includes 4 rRNA genes (two copies of 26S, and one copy of both 18S and 5S), 20 tRNA genes (including two copies of *trnaA-FME* and three copies of *trnaF-GAA*), and 40 protein-coding genes (including two copies of *atp1, apt8, cox3, nad6, nad7, rnaseH, rps12*, *rps3*, and *sdh4*). (Figure 8)

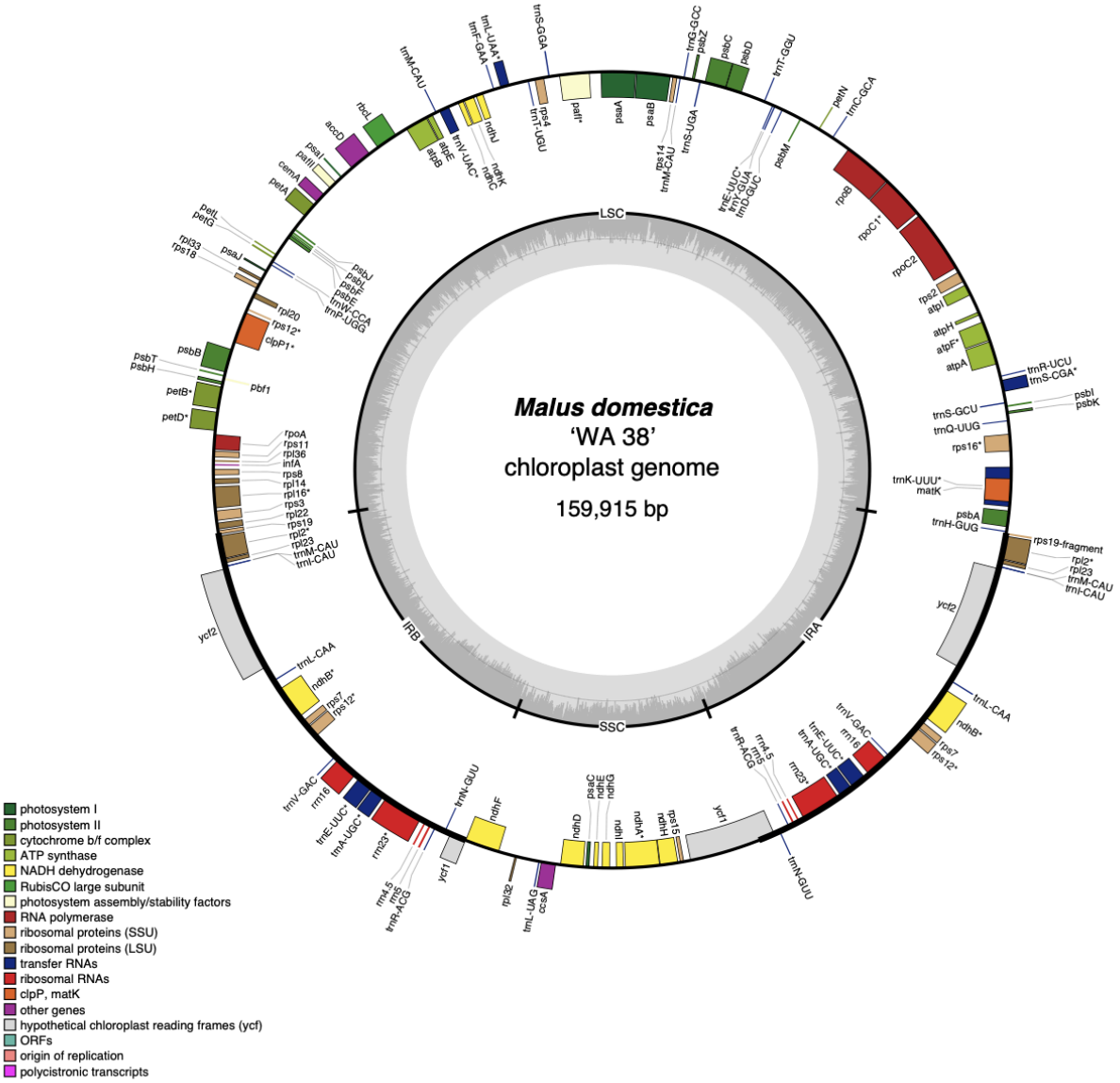

Figure.8 Chloroplast genome map of 'WA 38' with annotation.

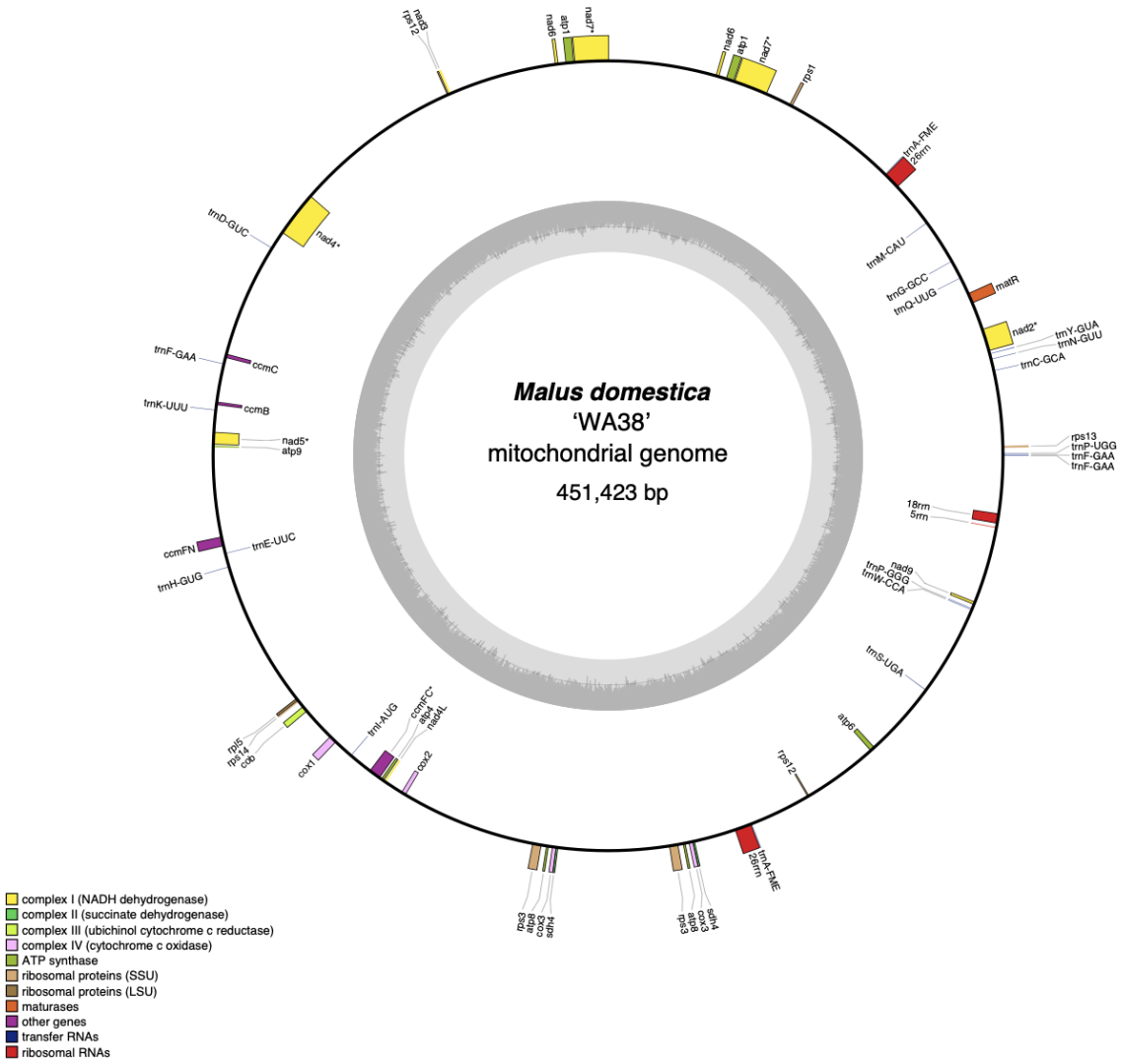

Figure.9 Mitochondrial genome map of 'WA 38' with annotation.

Methods

Workflows developed for each stage of the project and the summary workflow of the whole project are available in Supplemental Figure S1. Scripts with parameters for each computation step and methods in markdown format are available in GitHub at: https://gitlab.com/ficklinlab-public/wa-38-genome. Please refer to the GitHub material for step-by-step description of the workflow.

Sample Collection

Approximately 20 grams of young leaf material was harvested from the *Malus domestica* 'WA 38' mother tree at the Washington State University and USDA-ARS Columbia View Research Orchard near Orondo, WA, USA and flash-frozen in liquid nitrogen. Tissue was sent to the HudsonAlpha Institute for Biotechnology in Huntsville, AL, USA for DNA extraction, sequencing library preparation, and sequencing, following the same protocol (detailed below) used to generate the 'd'Anjou' pear genome (Yocca et al. 2024).

To assess heterozygosity and genome size of 'WA 38', DNA was isolated using a standard CTAB isolation (Doyle and Doyle 1987). Illumina TruSeq DNA PCR-free libraries were constructed from 3 ug of input DNA following the manufacturer's instruction, and sequenced on an Illumina NovaSeq6000.

For PacBio HiFi sequencing, high molecular weight DNA was isolated using a Nanobind Plant Nuclei Big DNA kit (Circulomics), with 4 g of input tissue and a 2-hour lysis. DNA purity, quantity, and fragment sizes were measured via spectrophotometry, Qubit™ dsDNA Broad Range assay (Invitrogen™), and Femto Pulse system (Agilent), respectively. DNA that passed quality control was sheared with a Megaruptor (Diagenode) and size-selected to roughly 25 kb on a BluePippin. The SMRTbell Express Template Prep Kit 2.0 (PacBio) was used to construct the PacBio sequencing library, and HiFi reads were produced using circular consensus sequencing (CCS) mode with two 8M flow cells on a PacBio Sequel II long-read system.
To scaffold PacBio HiFi contigs into chromosome pseudomolecules, a Dovetail Genomics Omni-C library was generated using 1 g of flash-frozen young leaf material as input following the manufacturer's instruction (Dovetail Genomics), and sequenced on an Illumina NovaSeq6000 S4 PE150 flow cell.

Sequence quality assessment and genome complexity analysis

Adapter sequences were trimmed from the raw Illumina shot-gun DNA reads using fastp (v0.23.2) (S. Chen et al. 2018) with all the other trimming functions disabled. Both the raw and trimmed Illumina reads, PacBio HiFi reads, and Omni-C Illumina reads were assessed for quality with FastQC (v0.11.9) (Andrews, 2010). Genome complexity, *i.e.* nuclear genome size and ploidy, was estimated using Jellyfish (v2.2.10) (Marçais and Kingsford 2011). With trimmed paired-end Illumina reads as input and a k-mer size set to 21, a k-mer count file was generated by Jellyfish. The k-mer histogram, also created by Jellyfish, was visualized in GenomeScope

(v1.0) (Vurture et al. 2017) with the following parameters: k-mer size = 21, Read length = 151, and Max k-mer coverage = 1000. A summary statistic report of the sequence quality and complexity analysis was generated with MultiQC (v1.13a).

Genome Assembly

Genome assembly and scaffolding

Phased haplomes were assembled by Hifiasm (v0.16.1) (Cheng et al. 2021) with default parameters, using both the Omni-C data and the PacBio HiFi long reads. The statistical summary of the assembly was produced using the assimilation_stats Perl scripts described in (Earl et al. 2011). Both hifiasm-assembled haplotype unitigs were then sorted by MUMmer (v3.23) (Kurtz et al. 2004) using the 'nucmer' function with flag -maxmatch. The resulting files were uploaded to the Assemblytics web server[12] (Nattestad and Schatz 2016) to visualize structural variations in two haplotype unitigs with the default settings.
Following the initial assembly step, bwa (H. Li and Durbin 2009) was used to index the draft contigs, and subsequently to align the Hi-C reads to the indexed contigs. The sorted files were input into Phase Genomics hic_qc[13] to validate the overall quality of the library. Both assembled haplomes were scaffolded into chromosomes with YaHS (Danecek et al. 2021; Zhou, McCarthy, and Durbin 2022), using default parameters.

Assembly curation and completeness assessment

Hi-C files were generated using YaHS Juicer Pre (v1.2a.2-0) with flag -a allowing manual curation. The resulting files were used as input for Juicer Tools Pre (v 1.22.01) to generate Hi-C contact maps (Zhou, McCarthy, and Durbin 2022; Durand et al. 2016). Juicebox Assembly Tools (v1.11.08) was used to explore the Hi-C maps for miss-assemblies (Robinson et al. 2018) . After manual examination of the Hi-C maps, the final genome assembly was generated by linking remaining files from YaHS Juicer Pre and original HiFi scaffold, using YaHS Juicer Post (v 1.2a.2-0, (Durand et al. 2016)).

For consistency and reproducibility, 'WA 38' chromosomes were renamed and reorientated to match published genomes. First, MUMmer (v3.23) was used to align the 'WA 38' assembly to the 'Gala' v1 HapA assembly using the –maxmatch parameter for unique matches (Kurtz et al.

---

[12] http://assemblytics.com,
[13] Available at https://github.com/phasegenomics/hic_qc

2004; Sun et al. 2020). Next, Assemblytics dotplot was used to identify 'WA 38' scaffolds that aligned with the 'Gala' v1 chromosomes and 'WA 38' scaffolds were renamed accordingly. To determine orientation, each 'WA 38' chromosome was aligned to the corresponding 'Gala' v1 chromosomes using LASTZ (v 1.02.00) implemented in Geneious (v9.0.5, Harris . 2007) with the 'search both strands' option. The chromosomes on the reverse stand were reoriented with the Reverse Complement (RC) function in Geneious (Supplemental Figure 6). The resulting assembly was searched against NCBI's RefSeq Plastid database[14] using megablast and a custom virus and bacteria database using kraken (v2.1.3) to identify contaminants. Scaffolds identified as plastid or microbe contaminants were removed in the assembly.

The cleaned assembly was compared to the 'Honeycrisp' genome assembly with a kmer approach using meryl (v1.4.1, Rhie et al., 2020). Chromosomes with a 'Honeyrisp' origin were placed in HapA, whereas the others were placed in HapB.

The two final haplome assemblies were compared to each other using MUMmer and Assemblytics as described above to identify structural variants. Benchmarking universal single-copy gene orthologs (BUSCO, v5.4.3_cv1) analysis was performed in genome mode with the eudictos_odb10 database to assess the completeness of gene content (Manni et al. 2021).

Structural and functional annotation

Repeat annotation

Repetitive elements from both haplotypes were annotated using EDTA (v2.0.0, (Ou et al. 2019)), with flags 'sensitive=1' and 'anno=1'. The full coding sequence from 'Gala' HapA, obtained from the Genome Database for Rosaceae (GDR, (Jung et al. 2019)), was used as reference to aid repeat finding. The custom transposable element library generated by EDTA was then imported to RepeatMasker (Smit et al. 2015) to further identify potentially overlooked repetitive elements. Repeatmasker was used to generate masked versions of the genome, where repetitive elements are converted to N to preserve the spacing but will not be included in the downstream annotation.

[14] https://www.ncbi.nlm.nih.gov/genome/organelle/

Gene Annotation

To annotate gene space, a combination of *ab initio* prediction and evidence-based prediction were performed with two rounds of BRAKER using transcriptome and homologous protein evidence. PASA (v2.5.2) (Haas et al. 2003) was then used to refine gene models and add UTR annotation. Lastly, a custom script was used for filtering. The detailed methods are described below.

BRAKER1 - annotation with transcriptome evidence

To perform transcriptome guided annotation, same data RNA-seq data from Khan et al. (2022) (eight tissue types from six pome fruit cultivars including 'WA 38') were first aligned to the assembled 'WA 38' masked haplomes using the STAR aligner implemented in GEMmaker (v2.1.0) (Hadish et al. 2022) Nextflow workflow. The resulting read alignments were used as extrinsic evidence in BRAKER1 (Hoff et al. 2016) to predict gene models in each soft-masked haplome with the following parameters: --softmasking, --UTR=off, --species=malus_domestica.

BRAKER2 - annotation with homologous protein evidence

To provide protein evidence data for BRAKER2 (Brůna et al. 2021), protein sequences from three sources were used: 1) Predicted protein sequences of 13 Rosaceae genomes retrieved from GDR (*Fragaria vesca* v4a2(Y. Li et al. 2019), *Malus baccata* v1.0(X. Chen et al. 2019), *M. domestica* var.Gala v1(Sun et al. 2020), *M. domestica* var.GDDH13 v1.1 (Daccord et al. 2017), *M. domestica* var. Honeycrisp v1.0(Khan et al. 2022), *M. sieversii* v1(Sun et al. 2020), *M. sylvestris* v1(Sun et al. 2020), *Prunus persica* v2.0.a1(Verde et al. 2017), *Pyrus betulifolia v1.0(Dong et al. 2020)*, *P. communis* var.d'Anjou v1(Yocca et al. 2023), *P. pyrifolia* var. Nijisseikiv v1.0(Shirasawa, Itai, and Isobe 2021)*, Rosa chinensis* var.Old Blush v2.0.a1(Raymond et al. 2018), and *Rubus occidentalis* v3 (VanBuren et al. 2018)); 2) Peptide sequences predicted from *de novo* transcriptome assemblies from the 'Honeycrisp' genome annotation(Khan et al. 2022); and 3) Viridiplantae OrthoDBv11 (Kuznetsov et al. 2022). In the same manner as BRAKER1, the masked haplome assemblies were used as input.

TSEBRA - transcript selection

The gene annotation results from BRAKER1 and BRAKER2 were merged and filtered based on the supporting evidence using TSEBRA (v.1.0.3) (Gabriel et al. 2021) with the default configuration (file obtained in August 2022) provided by TSEBRA developers.

PASA - gene model curation and UTR annotation

Two sources of transcriptome assembly evidence were obtained to facilitate PASA annotation: 1) Transcript sequences predicted from *de novo* transcriptome assemblies used by 'Honeycrisp' genome annotation; and 2) Reference guided assemblies were created with read alignment files from GEMmaker (see the BRAKER1 section for details) using Trinity with max intron size set to 10,000. Four rounds of PASA (v2.5.2) curation were performed using the aforementioned evidence and a starting annotation. The first round of PASA curation used TSEBRA annotation as the starting annotation, and annotations from the previous round were used as the starting annotation for rounds two through four. The curation results from each round were manually inspected using the PASA web portal. No significant improvement was observed after the fourth round of curation, therefore no further rounds were performed.

Gene model filtering and gene renaming

Repeat and gene model annotations were loaded to IGV (v2.15.1, (Robinson et al. 2011)) for manual inspection. Three types of erroneous gene models were observed consistently throughout the annotations. Type 1: Genes overlapping with repeat regions (e.g. transposon was wrongly annotated as a gene), Type 2: Gene models overlapping with each other on the same strand (e.g. single gene was wrongly annotated with multiple gene models), and Type 3: Gene models with splice variants that had no overlap (e.g. different genes were wrongly annotated as the single gene's splice variants). A custom script was used to address these errors. The Type 1 error was resolved by removing genes with 90% of its coding region overlapped with repeat regions. The Type 2 error was resolved by removing the shorter gene of a pair that overlaps on the same strand. The Type 3 error was resolved by splitting splice variant models with no overlap into two separate gene models. Finally, custom scripts were used to generate the final annotation files (gene, mRNA, cds, protein, gff3) and rename genes to match the naming convention proposed by GDR (https://www.rosaceae.org/nomenclature/genome).

Functional Annotation

The final gene sets from both 'WA 38' haplomes were annotated with EnTAPnf (Hart et al. 2020) with Interproscan, Panther, RefSeq, and uniprot_sprot databases that are automatically downloaded using the download.py script provided by EnTAPnf.

Comparative Analysis

Synteny analysis

A synteny comparison was performed using GENESPACE (Lovell et al. 2022) with five *Malus domestica* assemblies and annotations, (GDDH13 (Daccord et al. 2017), both haplomes from 'Honeycrisp' (Khan et al. 2022), and both haplomes from 'WA 38').

Gene family analysis

Gene family or orthogroup analyses were carried out to identify shared and unique gene families in 'WA 38' and other pome fruit genomes (i.e., *Malus* sp. and *Pyrus* sp. A full list of genomes analyzed can be found in Supplemental Table S10) following the method described by (Khan et al. 2022). Briefly, predicted protein sequences from the selected pome fruit genomes were classified into a pre-computed orthogroup database (26Gv2.0) using the 'both HMMscan and BLASTp' option implemented in the GeneFamilyClassifier tool from PlantTribes2 (Wafula et al. 2023). Overlapping orthogroups among *M. domestica* genomes were calculated and visualized with the UpSet plot function implemented in TBtools v2.030 (Chen C., et al. 2023).

A Core OrthoGroup (CROG) - Rosaceae gene count analysis was carried out following the method described by (Wafula et al. 2023). First, a CROG gene count matrix was created by counting genes classified into CORGs from each pome fruit genome. Next, the matrix was visualized as a clustermap using the Seaborn clustermap package (CROGs with standard deviation of 0 were removed prior to plotting) with rows normalized by *z*-score. Finally, the derived *z*-score of CROGs in each genome was summarized into a boxplot to illustrate z-score distribution using the boxplot function in Seaborn.

Chloroplast & Mitochondria Assembly and Annotation

The chloroplast genome was assembled from trimmed Illumina shot-gun DNA reads using NOVOplasty (v4.3.1, (Dierckxsens, Mardulyn, and Smits 2017)) with the *Malus sierversii* chloroplast genome (NCBI accesion ID: MH890570.1, (Naizaier et al. 2019)) as the reference sequence and the NOVOplasty *Zea mays RUBP* gene as the seed sequence. The assembled chloroplast was annotated using GeSeq[15] (website accessed on Dec. 19th, 2023) (Tillich et al. 2017) with default settings for 'circular plastid genomes for land plants' and the following

[15] https://chlorobox.mpimp-golm.mpg.de/geseq.html

parameters: annotating plastid inverted repeats and plastid trans-spliced *rps12*. Additionally, annotations from third party softwares Chloë (v0.1.0) and ARAGORN (v1.2.38), as well as a BLAT (v.35×1) search against all land plant chloroplast reference sequences (CDS and rRNA), were integrated with the GeSeq results(Tillich et al. 2017). Genes identified by multiple tools were manually reviewed to produce the final, curated annotation. The curated chloroplast annotation was visualized by OGDRAW (v1.3.1) (Greiner, Lehwark, and Bock 2019).

The mitochondrial genome sequence was isolated from the Hifiasm assembled contigs using MitoHifi (v3.2) (Uliano-Silva et al. 2023). The *Malus domestica* mitochondria sequencing from NCBI (NC_018554.1, (Goremykin et al. 2012)), which contained 57 genes consisting of 4 rRNAs, 10 tRNAS, and 33 protein-coding genes, was used as the closely related reference sequence. Briefly, MitoHifi compares the assembled contigs to the reference mitogenome using the BLAST algorithm. The resulting contigs were manually filtered by size and redundancy and then are circulated. To increase the annotation quality, GeSeq was deployed in mitochondrial mode with the *M. domestica* NCBI rRefsSeq sequence to annotate the 'WA 38' mitochondria assembly. Fragmented genes from the annotation were manually removed prior to visualization in OGDRAW (v1.3.1).

Availability of source code and requirements

Project name: 'WA 38' whole genome assembly and annotation
Project home page: https://gitlab.com/ficklinlab-public/wa-38-genome
Operating system(s): Platform independent
Programming language: bash, python, perl
Other requirements: singularity, nextflow, java, python
License: Not applicable
Any restrictions to use by non-academics: No restrictions
RRID: Not applicable

Data Availability

Raw reads generated for this project are publicly available at NCBI under BioProject: PRJNA1072127. Genome assembly and annotation are available on GDR: https://www.rosaceae.org/Analysis/20220983

Competing interests

The author(s) declare that they have no competing interests.

Funding

This work was supported by the Washington Tree Fruit Research Commission (WTFRC) project #AP-19-103 and USDA ARS internal appropriation funds.

Authors' contributions

L.H and S.F acquired funding for this project. All authors contributed to data analysis, data interpretation, and manuscript writing.

References

Brůna, Tomáš, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. 2021. “BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database.” *NAR Genomics and Bioinformatics* 3 (1): lqaa108.

Chen C., Wu Y., Li J., Wang X., Zeng Z., Xu J., Liu Y., Feng J., Chen H., He Y., and Xia R. 2023. “TBtools-II: A ‘one for All, All for One’ Bioinformatics Platform for Biological Big-Data Mining.” *Molecular Plant* 16 (11): 1733–42.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. “Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm.” *Nature Methods* 18 (2): 170–75.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” *Bioinformatics* 34 (17): i884–90.

Chen, Xilong, Shiming Li, Dong Zhang, Mingyu Han, Xin Jin, Caipin Zhao, Songbo Wang, et al. 2019. “Sequencing of a Wild Apple (Malus Baccata) Genome Unravels the Differences Between Cultivated and Wild Apple Species Regarding Disease Resistance and Cold Tolerance.” *G3* 9 (7): 2051–60.

Crosby, Jeffrey A., Jules Janick, Paul C. Pecknold, Joseph C. Goffreda, and Schuyler S. Korban. 1994. “`Enterprise’ Apple.” *HortScience: A Publication of the American Society for Horticultural Science* 29 (7): 825–26.

Daccord, Nicolas, Jean-Marc Celton, Gareth Linsmith, Claude Becker, Nathalie Choisne, Elio Schijlen, Henri van de Geest, et al. 2017. “High-Quality de Novo Assembly of the Apple Genome and Methylome Dynamics of Early Fruit Development.” *Nature Genetics* 49 (7): 1099–1106.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2): giab008.

Dierckxsens, Nicolas, Patrick Mardulyn, and Guillaume Smits. 2017. “NOVOPlasty: De Novo

Assembly of Organelle Genomes from Whole Genome Data." *Nucleic Acids Research* 45 (4): e18.

Di Guardo, Mario, Alice Tadiello, Brian Farneti, Giorgia Lorenz, Domenico Masuero, Urska Vrhovsek, Guglielmo Costa, Riccardo Velasco, and Fabrizio Costa. 2013. "A Multidisciplinary Approach Providing New Insight into Fruit Flesh Browning Physiology in Apple (Malus X Domestica Borkh.)." *PloS One* 8 (10): e78004.

Dong, Xingguang, Zheng Wang, Luming Tian, Ying Zhang, Dan Qi, Hongliang Huo, Jiayu Xu, et al. 2020. "De Novo Assembly of a Wild Pear (Pyrus Betuleafolia) Genome." *Plant Biotechnology Journal* 18 (2): 581–95.

Doyle, J. J., and J. L. Doyle. 1987. "A Rapid DNA Isolation Procedure for Small Quantities of Fresh Leaf Tissue." RESEARCH. https://worldveg.tind.io/record/33886/.

Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3 (1): 95–98.

Earl, Dent, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. "Assemblathon 1: A Competitive Assessment of de Novo Short Read Assembly Methods." *Genome Research* 21 (12): 2224–41.

Evans, Kate M., Bruce H. Barritt, Bonnie S. Konishi, Lisa J. Brutcher, and Carolyn F. Ross. 2012. "'WA 38' Apple." *HortScience: A Publication of the American Society for Horticultural Science* 47 (8): 1177–79.

Gabriel, Lars, Katharina J. Hoff, Tomáš Brůna, Mark Borodovsky, and Mario Stanke. 2021. "TSEBRA: Transcript Selector for BRAKER." *BMC Bioinformatics* 22 (1): 566.

Goremykin, Vadim V., Peter J. Lockhart, Roberto Viola, and Riccardo Velasco. 2012. "The Mitochondrial Genome of Malus Domestica and the Import-Driven Hypothesis of Mitochondrial Genome Expansion in Seed Plants." *The Plant Journal: For Cell and Molecular Biology* 71 (4): 615–26.

Greiner, Stephan, Pascal Lehwark, and Ralph Bock. 2019. "OrganellarGenomeDRAW (OGDRAW) Version 1.3.1: Expanded Toolkit for the Graphical Visualization of Organellar Genomes." *Nucleic Acids Research* 47 (W1): W59–64.

Haas, Brian J., Arthur L. Delcher, Stephen M. Mount, Jennifer R. Wortman, Roger K. Smith Jr, Linda I. Hannick, Rama Maiti, et al. 2003. "Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies." *Nucleic Acids Research* 31 (19): 5654–66.

Hadish, John A., Tyler D. Biggs, Benjamin T. Shealy, M. Reed Bender, Coleman B. McKnight, Connor Wytko, Melissa C. Smith, F. Alex Feltus, Loren Honaas, and Stephen P. Ficklin. 2022. "GEMmaker: Process Massive RNA-Seq Datasets on Heterogeneous Computational Infrastructure." *BMC Bioinformatics* 23 (1): 1–11.

Hart, Alexander J., Samuel Ginzburg, Muyang (sam) Xu, Cera R. Fisher, Nasim Rahmatpour, Jeffry B. Mitton, Robin Paul, and Jill L. Wegrzyn. 2020. "EnTAP: Bringing Faster and Smarter Functional Annotation to Non-Model Eukaryotic Transcriptomes." *Molecular Ecology Resources* 20 (2): 591–604.

Ha, Young-Ho, Bagila Maisupova, Kyung Choi, Hyun-Jun Kim, Daniyar Dosmanvetov, Bulkair Mambetov, Ainur Utebekova, et al. 2020. "Report on a Complete Chloroplast Genome Sequence of Wild Apple Tree, Malus Sieversii (Lebed.) M. Roem." *Mitochondrial DNA. Part B, Resources* 5 (2): 1504–5.

Hoff, Katharina J., Simone Lange, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2016. "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS." *Bioinformatics* 32 (5): 767–69.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. "eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090

Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–14.
Johnston, J. W., and P. Brookfield. 2012. "Delivering Postharvest Handling Protocols for Apples and Pears Faster: Integrating 'Omics' and Physiology Approaches." *Acta Horticulturae*, no. 945 (April): 23–28.
Jung, Sook, Taein Lee, Chun-Huai Cheng, Katheryn Buble, Ping Zheng, Jing Yu, Jodi Humann, et al. 2019. "15 Years of GDR: New Data and Functionality in the Genome Database for Rosaceae." *Nucleic Acids Research* 47 (D1): D1137–45.
Khan, Awais, Sarah B. Carey, Alicia Serrano, Huiting Zhang, Heidi Hargarten, Haley Hale, Alex Harkess, and Loren Honaas. 2022. "A Phased, Chromosome-Scale Genome of 'Honeycrisp' Apple ()." *GigaByte (Hong Kong, China)* 2022 (September): gigabyte69.
Kurtz, Stefan, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. 2004. "Versatile and Open Software for Comparing Large Genomes." *Genome Biology* 5 (2): R12.
Kuznetsov, Dmitry, Fredrik Tegenfeldt, Mosè Manni, Mathieu Seppey, Matthew Berkeley, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2022. "OrthoDB v11: Annotation of Orthologs in the Widest Sampling of Organismal Diversity." *Nucleic Acids Research* 51 (D1): D445–51.
Liebhard, R., M. Kellerhals, W. Pfammatter, M. Jertmini, and C. Gessler. 2003. "Mapping Quantitative Physiological Traits in Apple (Malus X Domestica Borkh.)." *Plant Molecular Biology* 52 (3): 511–26.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
Li, Wei, Jianan Liu, Hongyu Zhang, Ze Liu, Yu Wang, Longsheng Xing, Qiang He, and Huilong Du. 2022. "Plant Pan-Genomics: Recent Advances, New Challenges, and Roads Ahead." *Journal of Genetics and Genomics = Yi Chuan Xue Bao* 49 (9): 833–46.
Li, Xueli, Zhijie Ding, Haoyu Miao, Jinbo Bao, and Xinmin Tian. 2022. "Complete Chloroplast Genome Studies of Different Apple Varieties Indicated the Origin of Modern Cultivated Apples from and." *PeerJ* 10 (March): e13107.
Li, Xuewei, Ling Kui, Jing Zhang, Yinpeng Xie, Liping Wang, Yan Yan, Na Wang, et al. 2016. "Improved Hybrid de Novo Genome Assembly of Domesticated Apple (Malus X Domestica)." *GigaScience* 5 (1): 35.
Li, Yongping, Mengting Pi, Qi Gao, Zhongchi Liu, and Chunying Kang. 2019. "Updated Annotation of the Wild Strawberry Fragaria Vesca V4 Genome." *Horticulture Research* 6 (1): 1–9.
Lovell, John T., Avinash Sreedasyam, M. Eric Schranz, Melissa Wilson, Joseph W. Carlson, Alex Harkess, David Emms, David M. Goodstein, and Jeremy Schmutz. 2022. "GENESPACE Tracks Regions of Interest and Gene Copy Number Variation across Multiple Genomes," September. https://doi.org/10.7554/eLife.78526.
Lum, Geoffrey B., Barry J. Shelp, Jennifer R. DeEll, and Gale G. Bozzo. 2016. "Oxidative Metabolism Is Associated with Physiological Disorders in Fruits Stored under Multiple Environmental Stresses." *Plant Science: An International Journal of Experimental Plant Biology* 245 (April): 143–52.
Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021. "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes." *Molecular Biology and Evolution* 38 (10): 4647–54.
Marçais, Guillaume, and Carl Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers." *Bioinformatics* 27 (6): 764–70.
Miao, Haoyu, Jinbo Bao, Xueli Li, Zhijie Ding, and Xinmin Tian. 2022. "Comparative Analyses of Chloroplast Genomes in 'Red Fuji' Apples: Low Rate of Chloroplast Genome Mutations." *PeerJ* 10 (February): e12927.

Naizaier, Rouxianguli, Zezhong Qu, Songlin Wu, and Xinmin Tian. 2019. “The Complete Chloroplast Genome of Malus Sieversii (Rosaceae), a Wild Apple Tree in Xinjiang, China.” *Mitochondrial DNA. Part B, Resources* 4 (1): 983–84.

Nattestad, Maria, and Michael C. Schatz. 2016. “Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly.” *Bioinformatics* 32 (19): 3021–23.

O’Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic Acids Research* 44 (D1): D733–45.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. “Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *Genome Biology* 20 (1): 1–18.

Pareek, Sergio Tonetto de Freitas Sunil. 2019. *Postharvest Physiological Disorders in Fruits and Vegetables*. Edited by Sergio Tonetto de Freitas and Sunil Pareek. Boca Raton : Taylor & Francis, 2018.: CRC Press.

Raymond, Olivier, Jérôme Gouzy, Jérémy Just, Hélène Badouin, Marion Verdenaud, Arnaud Lemainque, Philippe Vergne, et al. 2018. “The Rosa Genome Provides New Insights into the Domestication of Modern Roses.” *Nature Genetics* 50 (6): 772–77.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.

Robinson, James T., Douglass Turner, Neva C. Durand, Helga Thorvaldsdóttir, Jill P. Mesirov, and Erez Lieberman Aiden. 2018. “Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data.” *Cell Systems* 6 (2): 256–58.e1.

Sallato, Bernardita, Matthew D. Whiting, and Juan Munguia. 2021. “Rootstock and Nutrient Imbalance Leads to ‘“Green Spot”’ Development in ‘WA 38’ Apples.” *HortScience: A Publication of the American Society for Horticultural Science* 56 (12): 1542–48.

Sheick, Ryan, Sara Serra, Stefano Musacchi, and David Rudell. 2023. “Metabolic Fingerprint of ‘WA 38’ Green Spot Symptoms Reveals Increased Production of Epicuticular Metabolites by Parenchyma.” *Scientia Horticulturae* 321 (112257): 112257.

Sheick, Ryan, Sara Serra, David Rudell, and Stefano Musacchi. 2022. “Investigations of Multiple Approaches to Reduce Green Spot Incidence in ‘WA 38’ Apple.” *Agronomy (Basel, Switzerland)* 12 (11): 2822.

Shirasawa, Kenta, Akihiro Itai, and Sachiko Isobe. 2021. “Chromosome-Scale Genome Assembly of Japanese Pear (Pyrus Pyrifolia) Variety ‘Nijisseiki.’” *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 28 (2). https://doi.org/10.1093/dnares/dsab001.

Sun, Xuepeng, Chen Jiao, Heidi Schwaninger, C. Thomas Chao, Yumin Ma, Naibin Duan, Awais Khan, et al. 2020. “Phased Diploid Genome Assemblies and Pan-Genomes Provide Insights into the Genetic History of Apple Domestication.” *Nature Genetics* 52 (12): 1423–32.

Švara, Anže, Honghe Sun, Zhangjun Fei, and Awais Khan. 2023. “Chromosome-Level Phased Genome Assembly of ‘Antonovka’ Identified Candidate Apple Scab-Resistance Genes Highly Homologous to HcrVf2 and HcrVf1 on Linkage Group 1.” *G3 Genes|Genomes|Genetics*, November, jkad253.

Tillich, Michael, Pascal Lehwark, Tommaso Pellizzer, Elena S. Ulbricht-Jones, Axel Fischer, Ralph Bock, and Stephan Greiner. 2017. “GeSeq - Versatile and Accurate Annotation of Organelle Genomes.” *Nucleic Acids Research* 45 (W1): W6–11.

Uliano-Silva, Marcela, João Gabriel R. N. Ferreira, Ksenia Krasheninnikova, Giulio Formenti, Linelle Abueg, James Torrance, Eugene W. Myers, Richard Durbin, Mark Blaxter, and

Shane A. McCarthy. 2023. “MitoHiFi: A Python Pipeline for Mitochondrial Genome Assembly from PacBio High Fidelity Reads.” *BMC Bioinformatics* 24 (1): 1–13.
VanBuren, Robert, Ching Man Wai, Marivi Colle, Jie Wang, Shawn Sullivan, Jill M. Bushakra, Ivan Liachko, et al. 2018. “A near Complete, Chromosome-Scale Assembly of the Black Raspberry (Rubus Occidentalis) Genome.” *GigaScience* 7 (8). https://doi.org/10.1093/gigascience/giy094.
Velasco, Riccardo, Andrey Zharkikh, Jason Affourtit, Amit Dhingra, Alessandro Cestaro, Ananth Kalyanaraman, Paolo Fontana, et al. 2010. “The Genome of the Domesticated Apple (Malus × Domestica Borkh.).” *Nature Genetics* 42 (10): 833–39.
Verde, Ignazio, Jerry Jenkins, Luca Dondini, Sabrina Micali, Giulia Pagliarani, Elisa Vendramin, Roberta Paris, et al. 2017. “The Peach v2.0 Release: High-Resolution Linkage Mapping and Deep Resequencing Improve Chromosome-Scale Assembly and Contiguity.” *BMC Genomics* 18 (1): 225.
Vurture, Gregory W., Fritz J. Sedlazeck, Maria Nattestad, Charles J. Underwood, Han Fang, James Gurtowski, and Michael C. Schatz. 2017. “GenomeScope: Fast Reference-Free Genome Profiling from Short Reads.” *Bioinformatics* 33 (14): 2202–4.
Wafula, Eric K., Huiting Zhang, Gregory Von Kuster, James H. Leebens-Mack, Loren A. Honaas, and Claude W. dePamphilis. 2022. “PlantTribes2: Tools for Comparative Gene Family Analysis in Plant Genomics.” *Frontiers in Plant Science* 13: 1011199.
Yan, Ming, Xueqing Zhao, Jianqing Zhou, Yan Huo, Yu Ding, and Zhaohe Yuan. 2019. “The Complete Chloroplast Genome of Cultivated Apple (Malus Domestica Cv. ‘Yantai Fuji 8’).” *Mitochondrial DNA. Part B, Resources* 4 (1): 1213–16.
Yocca, Alan, Mary Akinyuwa, Nick Bailey, Brannan Cliver, Harrison Estes, Abigail Guillemette, Omar Hasannin, et al. 2023. “A Chromosome-Scale Assembly for ‘d’Anjou’ Pear.” *bioRxiv*. https://doi.org/10.1101/2023.08.22.554305.
———. 2024. “A Chromosome-Scale Assembly for ‘d’Anjou’ Pear.” *G3* 14 (3). https://doi.org/10.1093/g3journal/jkae003.
Zhang, Huiting, Eric K. Wafula, Jon Eilers, Alex E. Harkess, Paula E. Ralph, Prakash Raj Timilsena, Claude W. dePamphilis, Jessica M. Waite, and Loren A. Honaas. 2022. “Building a Foundation for Gene Family Analysis in Rosaceae Genomes with a Novel Workflow: A Case Study in Pyrus Architecture Genes.” *Frontiers in Plant Science* 13 (November): 975942.
Zhang, Liyi, Jiang Hu, Xiaolei Han, Jingjing Li, Yuan Gao, Christopher M. Richards, Caixia Zhang, et al. 2019. “A High-Quality Apple Genome Assembly Reveals the Association of a Retrotransposon and Red Fruit Colour.” *Nature Communications* 10 (1): 1494.
Zhao, Xueqing, Ming Yan, Yu Ding, Xuesen Chen, and Zhaohe Yuan. 2019. “The Complete Chloroplast Genome of Apple Rootstock ‘M9.’” *Mitochondrial DNA. Part B, Resources* 4 (2): 2187–88.
Zhou, Chenxi, Shane A. McCarthy, and Richard Durbin. 2022. “YaHS: Yet Another Hi-C Scaffolding Tool.” *Bioinformatics* 39 (1): btac808.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0.2013-2015 <http://www.repeatmasker.org>.
Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Rhie, A., Walenz, B.P., Koren, S. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21, 245 (2020). https://doi.org/10.1186/s13059-020-02134-9

Li, W., Chu, C., Li, H. et al. Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing. Nat Genet 56, 505–516 (2024). https://doi.org/10.1038/s41588-024-01657-2