Tree Genetics & Genomes (2026) 22:8
https://doi.org/10.1007/s11295-026-01734-z

GENOME REPORT

Long-read Iso-Seq re-annotation of the Pyrus communis 'Bartlett' doubled haploid genome

Jon Eilers[1] · Huiting Zhang[2] · Christopher Gottschalk[3] · Melissa Williams[4] · Jane Grimwood[4] · Alex Harkess[4] · Loren Honaas[5] · Jessica M. Waite[5]

Received: 18 November 2025 / Revised: 27 February 2026 / Accepted: 9 March 2026 / Published online: 1 April 2026
This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2026

Abstract
As genome sequencing becomes more accessible to non-model species, it is key to ensure high-quality genome annotations exist. These efforts include the re-annotation of important genomes when newer data and software are available. Here, we present a re-annotation of the 'Bartlett' doubled haploid (DH) genome, which has served as the reference genome for European pear research. We employed a newly developed annotation pipeline that incorporated a polished version of the DH genome, an Iso-Seq RNA-seq dataset developed by sampling eleven different tissue types from adult trees and in vitro treelets, and RNA-seq evidence from nine publicly available transcriptomes encompassing ten tissue types. The resulting re-annotation exhibited improved BUSCO and OMArk scores, increased gene capture, and improved gene model accuracy. Further, we saw increases in the number of orthogroups represented, as well as improved completeness of those orthogroups, based on Core Orthogroup similarity to recent high-quality pome annotations. Together, the improvements generated from this annotation will enable efficient use for molecular biology, genetic evolution studies, and other downstream applications.

Keywords Pear · Isoform sequencing · RAGNAROK

Introduction

Genome annotation accuracy is integral to the development and use of molecular biological tools and comparative genomic analysis. Functional studies to understand protein function using genetic transformation require accurate gene models to determine a correct coding sequence and enable gene expression studies, such as quantitative RT-PCR, which rely on correct annotation of exon-intron structure, 3' Untranslated Regions (3' UTRs), and homologs to identify ideal target sequences. Furthermore, quality annotations are important for comparative studies, including gene family analysis across species and understanding how protein domains have evolved (Vaattovaara et al. 2019). However, annotation in non-model species is often plagued by poor identification and masking of repetitive elements, incomplete or missing evidential information on transcriptional products, error-prone gene predictions, and inadequate quality checks. Recent studies and tools have focused on addressing these challenges by outlining best practices and developing improvements in the annotation process (Freedman and Sackton 2025; Salzberg 2019; Vuruputoor et al. 2023).

While genome sequencing of non-model plants has expanded rapidly in recent years, quality annotation has lagged behind the assembly of new high-quality genomes. Non-model plant genomes have allowed for greater insights into comparative and functional analysis of genes, particularly in previously difficult-to-study crop plants (Kersey

Communicated by L. Bianco

✉ Jessica M. Waite
jessica.waite@usda.gov

1 Analytical Genomics LLC, Walla Walla, WA, USA

2 Dept. of Horticulture, Washington State University, Pullman, WA, USA

3 USDA ARS Appalachian Fruit Research Station, Kearneysville, WV, USA

4 HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA

5 USDA ARS Tree Fruit Research Lab, 1104 N. Western Ave, Wenatchee, WA 98801, USA

2019; Unamba et al. 2015). Annotations of high enough quality enable molecular, genome-wide, and evolutionary studies; however, this requires considerable time and effort, particularly with complex genomes (Vuruputoor et al. 2023). Annotation is influenced by the availability of high-quality transcriptome datasets that sample broadly across tissue types to inform gene model structure and reduce reliance on protein homology and prediction that often fail to capture novelty (Freedman and Sackton 2025; Vaattovaara et al. 2019). The process of eukaryotic genome annotation can be divided into several steps, including: identification and masking of repetitive sequences, so as not to over-annotate for genes associated with transposable elements such as transposases and pseudogenes; ab initio gene prediction, which relies on algorithmic identification of intrinsic gene signatures such as open reading frames; and evidence alignment, which uses RNA and protein sequences from the species of interest, as well as protein sequences of more distant species to inform gene model predictions (Yandell and Ence 2012). During the evidence alignment stage, RNA sequence quality and tissue type diversity strongly influence the gene model annotation (Vaattovaara et al. 2019). Publicly available transcriptome data using *Pyrus communis* has been relatively limited until recently and has largely focused on fruit and leaf tissues of mature trees (Supp Table 1). Thus, expression of genes that are not well represented in these tissues has less evidence to support accurate gene models.

The *Pyrus communis* 'Bartlett' doubled haploid genome (DHv2.0) has provided a highly useful resource for the pear and Rosaceae research communities and would benefit from genome re-annotation using newly available RNA-seq data. This genome was assembled using high-quality long-read PacBio DNA sequences and circumvented the problem of high heterozygosity in pear genomes by employing a doubled haploid (Linsmith et al. 2019). Updated Benchmarking Universal Single Copy Orthologue (BUSCO) analysis of the genome, using version 6.0.0 and the OrthoDB Embryophyta version 12 lineage, returned 99.4% complete (51.3% single and 48.0% duplicated), 0.1% fragmented, and 0.5% missing BUSCOs, compared to 94.3% complete (64.0% single, 30.3% duplicated), 1.9% fragmented, and 3.8% missing BUSCOs using version 3.0.2. However, analysis of the annotation-derived transcript set resulted in only 81.8% complete (59.8% single, 22.1 duplicated), 10.0% fragmented, and 8.2% missing BUSCOs (Fig. 1A)(Manni et al. 2021). Furthermore, our previous work identified problematic gene models in the DHv2.0a1 annotation for a targeted group of architecture-related genes, in part due to PacBio adapter contamination (Zhang et al. 2022). These lines of evidence led us to perform additional iterative polishing of the genome, which improved gene model accuracy for nearly all these genes (Zhang et al. 2022). Polishing decreased genome size and chromosome length slightly, while maintaining the same number of chromosomes, scaffolds, and contigs (Supp Table 2). We sought to re-annotate this polished version of the genome (referred to here as DHv2.1), including a long-read isoform sequencing (Iso-Seq from PacBio) dataset to improve gene model accuracy. Iso-Seq technology has allowed for the capture of full length RNA transcripts and isoforms (Gonzalez-Garay 2016). These transcripts can be aligned to the genome using splice-aware aligners that reveal accurate exon/intron structure of a gene, as well as capture untranslated regions (UTRs) of mRNA. In addition to the benefit of long reads, Iso-Seq data is also useful for predicting gene isoforms, something that is difficult to accurately do with short-read sequencing. To harness this information, we employ a developmental version of a new annotation pipeline, RAGANROK, and include recent, publicly available short-read transcriptome datasets representing a wider range of tissues, in addition to a new Iso-Seq dataset representing 11 different tissue types collected over two years (Kuster et al. 2025). Together, these efforts yielded improvements to gene model accuracy, gene capture, and orthogroup completeness, resulting in a more useful 'Bartlett' DH genome resource.

Materials and methods

Sample preparation and Iso-Seq sequencing

Eleven tissue types were collected from *P. communis* 'Bartlett' throughout the seasons of 2021–2022. Early (~0.25 inches in size) and late (~0.5 inches) buds, flowers, young and mature leaves, whole immature fruit, and mature fruit divided into peels, cortex, and seeds, were collected from 30 year old trees, grown on 'Old Home x Farmingdale 97' rootstocks at the Columbia View Research Orchard (47°33'51.4"N, 120°14'38.3"W, Orondo, WA), following standard management practices. Tissues from three trees were pooled and performed in triplicate for biological replication, and three replicates were repeated for technical replication. Stem and root tissues were collected from micropropagated and rooted plantlets, grown on Quoirin and Lepoivre media, under 16 h of 55 µmol light. Three in vitro plantlets were pooled for each of three biological replicates.

All collected tissues were flash-frozen in liquid nitrogen and ground using a mortar and pestle or SamplePrep Freezer/Mill (ColeParmer, Vernon Hills, IL) until a fine powder was yielded. RNA was extracted using a modified CTAB protocol (Hargarten et al. 2018), and quality was assessed using an Epoch spectrophotometer (Agilent Technologies, Santa Clara, CA), Qubit (ThermoFisher Scientific, Waltham, MA), and Tape Station (Agilent Technologies, Santa Clara, CA).

A.

	Genome	Transcriptome	Protein
Bartlett DHv2.0a1	C: 99.4% [S: 51.3%, D: 48.0%] F: 0.1% M: 0.5%	C: 81.8% [S: 59.8%, D: 22.1%] F: 10.0% M: 8.2%	C: 81.7% [S: 59.6%, D: 22.2%] F: 10.5% M: 7.7%
Bartlett DHv2.1a1	C: 99.4% [S: 51.2%, D: 48.1%] F: 0.1% M: 0.5%	C: 98.8% [S: 54.3%, D: 44.5%] F: 0.3% M: 0.9%	C: 98.2% [S: 56.3%, D: 42.0%] F: 0.4% M: 1.3%

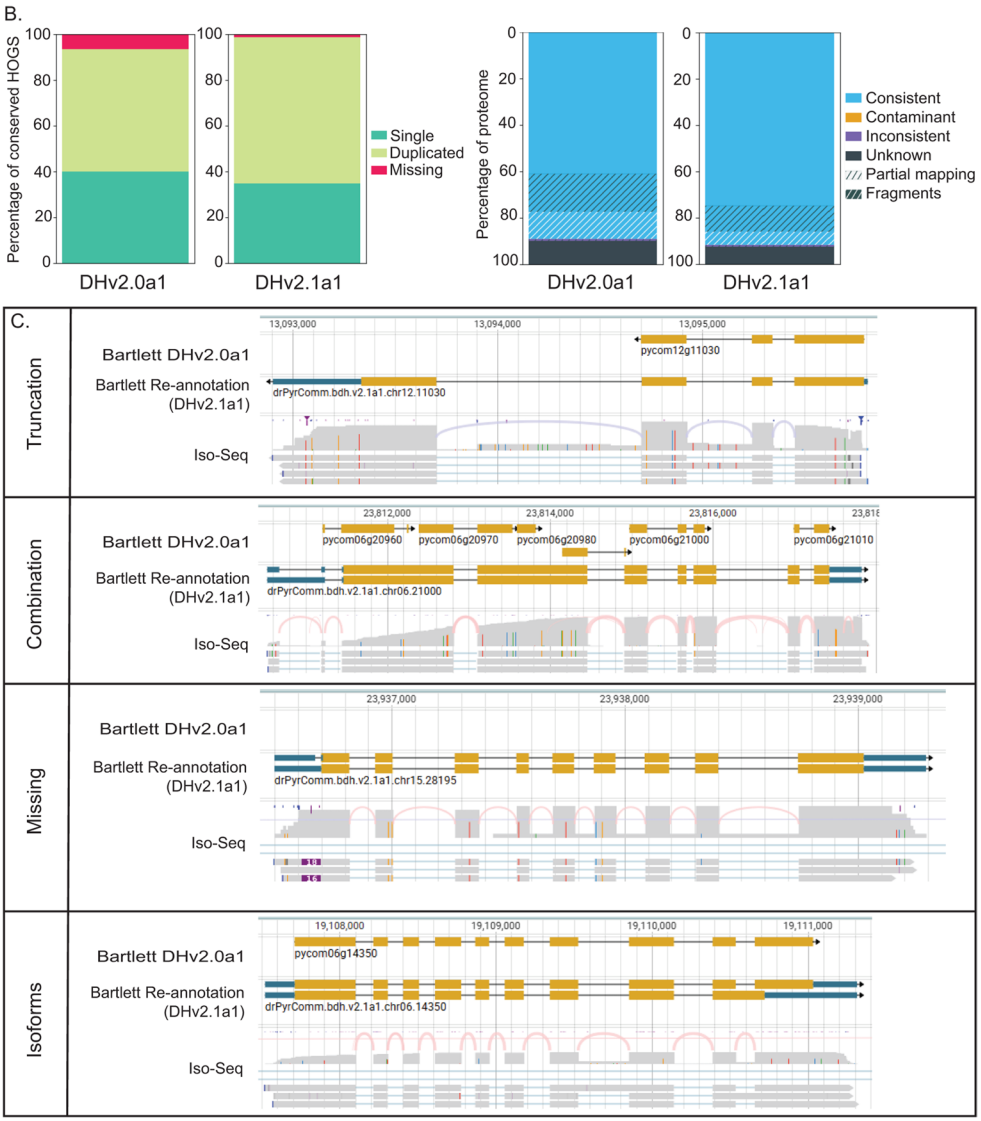


Fig. 1 Re-annotation of the 'Bartlett' DH genome improved gene capture and gene model accuracy (**A**) BUSCO analysis shows increased complete genes and decreased fragmented and missing genes in genome, transcript, and protein data after genome polishing and re-annotation. Of the complete genes, increased duplicates and decreased single genes were found. C - complete, S - single, D - duplicated, F - fragmented, M - missing. orthoDB embryophyta v12 was used for this analysis. (**B**) OMArk analysis indicates fewer missing and single genes and increased duplicated genes belonging to Hierarchical Orthologous Groups (HOGs) after re-annotation. This analysis also showed increased percentages of proteins consistent with known gene family lineages, and fewer fragments, partially mapped, and unknown proteins. Analysis was done using the Viridiplantae ancestral lineage. (**C**) Gene model visuals from JBrowse, representative of common improvements resulting from re-annotation. Visuals show a gene model from the DHv2.0a1 annotation (top row), the DHv2.1a1 re-annotation (middle row), and the top 3–4 Iso-Seq aligned reads (bottom row). Examples of truncated gene models, models that should be combined into one model, missing models, and putative isoforms are shown

The resulting RNAs were used as input for the PacBio Iso-Seq Express 2.0 kit, size selected with 0.86X beads, and sequenced on a PacBio Revio platform at HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA).

Transcriptome and annotation datasets used in this study

Publicly available RNA-seq datasets and existing genome annotations were downloaded from NCBI and decompressed to extract FASTQ files using prefetch and fasterq-dump from the SRA Toolkit (https://github.com/ncbi/sra-tools, v3.1.1). Details of transcriptomes and annotations can be found in Supp Table 3.

RNA-seq alignment and quality control

FastQC was used to assess the quality of each RNA-Seq dataset (Andrews 2010) (v12.1). Iso-Seq data was aligned to the assembly using minimap2 (Li 2021) (v2.28). Short-read data was mapped to the genome using STAR (Dobin et al. 2013) (v2.7.11b). STAR was configured to prioritize

specificity and accurate placement by retaining only high-confidence, uniquely mapped reads (outFilterMultimapNmax = 1) and enforcing strict mismatch and score thresholds (outFilterMismatchNmax = 5; outFilterMismatchNoverLmax = 0.05; outFilterScoreMinOverLread = 0.9; outFilterScoreMin = 30), and reads were not trimmed prior to input. The resulting BAM files were processed using samtools (Danecek et al. 2021) (v2.21). featureCounts was used to check mapping rates and quality (Liao et al. 2014) (v2.0.6). Multiqc was used to generate dataset-wide figures for comparison (Ewels et al. 2016) (v1.24.1).

Re-annotation of the polished *Pyrus communis* 'Bartlett' DHv2.1 genome.

Gene annotation was conducted using an early developmental version of the RAGNAROK (v1.0) pipeline (Kuster et al. 2025). Released and current versions can be found at https://github.com/ryandkuster/ragnarok. In brief, RAGNAROK relies on several lines of evidence for generating an annotation. First, we conducted annotation and masking of repetitive elements using the EDTA (v2.2.2) pipeline with the parameters -sensitive 1, -force 1, -cds using DH2.0 annotation, and a panMalus TE library previously developed for *Malus* but included *Pyrus* sequences (Mansfeld et al. 2023; Ou et al. 2019). 46.64% of the genome was found to be repetitive using EDTA (Supp Table 4). RAGNAROK then utilizes the GPU-accelerated Helixer gene prediction program to predict genes on the resulting masked genome sequence from EDTA (Holst et al. 2023). However, Helixer was not used for this annotation, as the version incorporated into RAGNAROK v1.0 predicted frequent false micro-exons. This issue has been solved for newer versions. RAGNAROK then mapped Iso-Seq and short-reads reads using Minimap2 (v2.28) and STAR (v2.7.11b) (Dobin et al. 2013; Li 2021). The resulting mappings were used as input into hybrid transcript assembly using StringTie2 (Shumate et al. 2022) (v2.2.3). TransDecoder was then used to identify and predict the longest coding sequences (https://github.com/TransDecoder/TransDecoder, v5.7.1). RAGNAROK then utilizes Mikado2 to select the best representative transcript models from all the lines of evidence (Venturini et al. 2018) (v2.3.4). As further support of the gene models, annotations from other *Pyrus* and *Malus* species that are considered to be high quality were lifted-over to the DHv2.1 assembly using LiftOn ((Chao et al. 2025), Supp Table 3). Accession numbers for both RNA-seq datasets and genome annotations can be found in Supp Table 3. To provide high quality protein support, Viridiplantae proteins from the UniProtKB/Swiss-Prot 2025-01 release were blasted against the genome assembly (UniProt 2025). The transcripts, open reading frames, lifted-over annotations, short-read alignments, and protein blast results were used as evidence and weighted according to our confidence in the dataset and analysis of the selected gene model set (Supp Tables 3 and 5).

Functional annotation was performed with EnTAP (v2.3.0) using an e-value cutoff of 1e−5. Similarity searches were run with DIAMOND (v2.1.8) using the default very-sensitive mode, minimum query coverage (qcov)≥50%, and target coverage (tcov)≥50%. Functional assignments were generated with eggNOG-mapper (v2.1.12) using the default more-sensitive mode. EnTAP homology searches used Plant RefSeq release 233 and UniProtKB/Swiss-Prot release 2026_01. Results from these searches can be found in Supp Table 6. InterProScan (v5.77-108.0; InterPro release 108.0) was used to assign InterPro (IPR) accessions and Gene Ontology terms based on Pfam (release 38.1) and PANTHER (release 19.0) signatures (with GO terms derived via InterPro integration/InterPro2GO).

Assessment of annotation quality

Summary statistics were generated using AGAT ((Dainat 2024), v1.4.2). Overall gene model quality was assessed using BUSCO (v6.0.0) and OMArk (v0.3.1) (Manni et al. 2021; Nevers et al. 2024). The database used for BUSCO was OrthoDB Embryophyta version 12, and for OMArk, the July 2023 release of the OMAmer database with the Viridiplantae lineage selected. For each of these, only the longest isoform for each gene was included in the analysis. Similarity between DHv2.0a1 and DHv2.1a1 annotations as assessed by lifting over gene models from DHv2.0a1 to DHv2.1a1 using LiftOn (v1.05), followed by comparison of annotations using Gffcompare (v0.12.6) (Chao et al. 2025; Pertea and Pertea 2020) https://github.com/gpertea/gffcompare). featureCounts provided gene and exon assignment percentages for Iso-Seq data (Liao et al. 2014). To check gene structure quality, Iso-Seq data was mapped to both DHv2.0 and DHv2.1 genome assemblies using minimap2 with their respective annotations as references. The resulting bam files were used as input into StringTie2 to generate transcripts that were then used by SQANTI3 (v0.12.6) for analyzing gene model structure relative to the Iso-Seq transcripts (Pardo-Palacios et al. 2024; Shumate et al. 2022). Because Iso-Seq data largely reflect full length transcripts, the resulting analysis was considered a reasonable ground truth for the gene model structure quality.

Orthogroup analysis

Orthogroup classification and Core Orthogroup (CROG) analysis was performed following (Zhang et al. 2024) and (Wafula et al. 2023). Briefly, DHv2.1a1 predicted protein

sequences were classified into pre-computed orthogroups (PlantTribes 26Gv2.0) using the GeneFamilyClassifier tool from PlantTribes2 (Wafula et al. 2023). To construct a CROG gene count matrix that includes DHv2.1a1, a CROG list was first obtained from the 'WA 38' genome assembly and annotation GitLab repository (https://gitlab.com/ficklinlab-public/wa-38-genome). The CROG list together with PlantTribes2 orthogroup classification summary table was then used to count the number of DHv2.1a1 genes classified into CROGs. Next, DHv2.1 CROG counts were integrated into the CROG gene count matrix from the same GitLab repository. Last, a CROG gene count clustermap was generated with a selected subset of genome annotations, including: *Pyrus communis* 'Bartlett' Doubled Haploid v2.0 (Linsmith et al. 2019), *P. betulaefolia* 'Shanxi Duli' (Dong et al. 2020), *P. communis* 'Bartlett' v1.0 (Chagne et al. 2014), *Malus sylvestris* (Ruhsam et al. 2022), *P. pyrifollia* 'Nijisseiki' (Shirasawa et al. 2021), *P. pyrifolia* 'Yunhong No.1' (Sun et al. 2023), *P. communis* 'd'Anjou' haplotypes A and B (Yocca et al. 2024), *M. domestica* 'Fuji' (Li et al. 2024), *M. domestica* 'M9' rootstock (Li et al. 2024), *M. domestica* 'MM106' rootstock (Li et al. 2024), *M. domestica* 'Honeycrisp' haplotype A (Khan et al. 2022), *M. domestica* 'WA38' haplotype A (Zhang et al. 2024). For genes where the annotation identified multiple isoforms, if different isoforms were classified into the same orthogroup, that gene was only counted once for CROG analysis. Only 7 genes in the annotation had isoforms that were classified into distinct orthogroups, in which case each orthogroup represented was counted.

Results

RAGNAROK annotation pipeline and inclusion of Iso-Seq RNA evidence improve 'Bartlett' genome annotation

To accomplish annotation of the polished version of the 'Bartlett' DH genome, DHv2.1, we used a newly developed annotation pipeline, RAGNAROK (Kuster et al. 2025). Briefly, RAGNAROK combines transcript assemblies from RNA-seq evidence, protein alignments, and Helixer predictions. The pipeline then employs Mikado to select the best models from those lines of evidence. As RNA-seq input, we used a combination of Iso-Seq data generated for this study, and publicly-available short-read data. We sought to include both technologies for obtaining high accuracy, as PacBio Iso-Seq long-read Circular Consensus Sequence technology allows for capture of high-accuracy, full-length transcripts, and short-read datasets have highly accurate sequences representing multiple cultivars, species, and tissue types. For Iso-Seq data, we sequenced pooled 'Bartlett' RNA extracted from early spring buds, flowers, leaves, fruit tissues, stems, and roots. Short-read data included publicly available short-read data from nine different studies, encompassing three *Pyrus* species (seven studies from *P. communis* cultivars, one from *P. ussuriensis*, and one from *P. betulafolia*) and ten different tissue types (leaf, stem, root, fruit cortex, fruit peel, immature fruit, flowers, buds, pedicels, and callus tissue) (Supp Table 3). As protein evidence, we used Viridiplantae (plant) proteins from the UniProtKB/Swiss-Prot 2025-01 release. Within the RAGNAROK pipeline, we chose not to include Helixer ab initio data, as we found this to introduce frequent false exons. Further, we altered the weights given to different RNA-seq and protein inputs in Mikado, which are summarized in Supp Table 3. For gene IDs, we followed the recommended AgBioData naming conventions (Cannon et al. 2025), while maintaining the gene numbers from the original DHv2.0a1. Maintenance of base ID numbers allows for maximum searchability of previously studied genes. In cases where multiple gene models were combined, the most appropriate ID number was selected to carry over.

Our resulting re-annotation, DHv2.1a1, showed an overall increase in gene and UTR number and length metrics (Table 1). Gene number increased from 37,445 in DHv2.0a1 to 41,611 in DHv2.1a1. We saw a larger increase in the number of mRNA (from 37,445 to 46,279), as this pipeline and the use of Iso-Seq data allowed for more confident identification of isoforms. This also enabled inclusion of 3' and 5'UTRs, which were not included in the DHv2.0a1 annotation. Mean

Table 1 Summary statistics compared between 'Bartlett' DH annotations.

	BartlettDH 2.0	BartlettDH 2.1
Number of genes	37445	41611
Number of mRNA	37445	46279
Number of single exon genes	6789	7303
Mean isoform number per gene	1	1.1
Mean exon number	5.1	5.5
Mean gene length	2327	3602
Mean transcript (mRNA) length	2327	3685
Mean CDS length	1121	1255
Mean exon length	222	345
Mean length of introns in CDS	297	375
Genes with a 5'UTR (excluding isoforms)	0	36251
Genes with a 3'UTR (excluding isoforms)	0	37004
Genes with both a 5'UTR and 3'UTR (excluding isoforms)	0	352921
Mean 5'UTR length (bp)	0	301
Mean 3'UTR length (bp)	0	412
Number of ncRNA	0	2552
Mono/multi-exotic ratio	18.1 (6789/37/445)	17.6 (3703/41,611)

transcript, CDS, exon, and intron length all increased in the re-annotation. BUSCO scores remained similar between the DHv2.0 and polished DHv2.1 assemblies, demonstrating that the additional polishing did not result in loss of BUSCO genes (Fig. 1A) (Manni et al. 2021; Zhang et al. 2022). Looking at the annotation-derived transcript sets, however, we found a large improvement in BUSCO defined complete genes (81.8% to 98.8%), including a slight decrease in single-copy genes (59.8% to 54.3%) and an increase in duplicated genes (22.1% to 44.5%), and major reductions in fragmented (10.0% to 0.3%) and missing genes (8.2% to 0.9%) (Fig. 1A). Similar changes were seen with the protein dataset (Fig. 1A). We found similar improvements using the OMArk tool, which assesses completeness and consistency of an annotation gene set relative to other related species (Nevers et al. 2024). To assess the DH annotations, we used the Viridiplantae ancestral lineage. In terms of completeness, the re-annotation showed a decreased percentage of missing conserved ancestral genes (represented as hierarchical orthologous groups, or HOGs), as well as a decreased percentage of single-copy HOGs, and an increased percentage of duplicated HOGs (Fig. 1B). In terms of consistency, the re-annotation showed an increased percentage of proteins found to be consistent with the lineages of known gene families. We found a decreased percentage of proteins falling into the partial mapping or fragmented categories, where proteins are partially matched with a gene family based on sharing *k*-mers. Further, we found a decrease in the percentage of proteins falling into the unknown category, for which no matches could be found. Together, these support an improvement in gene annotation for DHv2.1a1.

Re-annotation improved gene models

To assess improvement of gene models more closely, we first identified a list of genes whose gene models were erroneous or missing in the DHv2.0 annotation (Supp Table 5). These included 17 architecture, auxin, and light-related genes, some of which are commonly misannotated across genomes. Of these genes, 14 were annotated completely in DHv2.1a1, with support of multiple RNA-seq datasets, and 3 were improved, but with small mistakes noted. Among this set of genes, we noted four major categories of improvement: gene model missing in DHv2.0a1 and captured in DHv2.1a1; gene model truncated in DHv2.0a1 and completed in DHv2.1a1; two or more gene models in DHv2.0 combined into one model in DHv2.1a1; multiple isoforms now annotated in DHv2.1a1. We then scanned the genome for more examples falling into each of these categories. Representative examples are shown in Fig. 1C. For each example, gene models from both annotations are shown, as well as top Iso-Seq RNA-seq evidence.

To better understand how widespread these annotation differences were, we compared the two annotations using Gffcompare (Pertea and Pertea 2020). As the genome polishing changed chromosome coordinates, LiftOn (Chao et al. 2025) was used to lift over gene models from DHv2.0a1 to DHv2.1a1 in order to directly compare the annotations. A drawback of this approach is that not all gene models are found in the new version, and thus are not included in the lifted-over annotation. In this case 33 genes were not lifted over, half of which were from contigs or scaffolds not assigned to chromosomes. Additionally, the LiftOn algorithm attempts to find the closest match possible, but the exon/intron boundaries may be slightly different due to changes in sequence from genome polishing. Despite these caveats, 37412 genes were lifted over, and visual inspection using Jbrowse showed similar overall structure between versions. Gffcompare reports sensitivity, indicates the extent to which the original annotation was captured by the new annotation, and precision, which can be understood as the exact agreement between annotations, including exon/intron boundaries and 3' and 5' UTRs. Between the DH annotations, nucleotide base level sensitivity was 84.2% and precision was 47.5% (Supp Table 7). We hypothesized that the high sensitivity and low precision likely reflects the addition of UTRs, new genes, and changes in gene model structure. Exon level sensitivity and precision were 50.1% and 42% respectively, indicating about half of the original exons are reproduced, but a large number are new or shifted. While exon structure showed low percentages, intron sensitivity and precision were 74.3% and 61.1% respectively, which suggests the core splice junctions from the original annotation are largely conserved, but inclusion of UTRs and new gene models leads to lower precision in both introns and exons. To test this, we excluded UTRs and ran the analysis again. We found that precision increased at every level, showing the largest increase at the exon level (from 42% to 60.2%, Supp Table 7). Only 7921 transcripts matched perfectly between the lifted over DHv2.0a1 annotation and the DHv2.1a1 annotation. This is seen in the low transcript and locus level sensitivity and precision, at 23.8% and 19.4%. Missed elements (those found in DHv2.0a1 and not in DHv2.1a1) made up about 19% of exon, introns, and gene loci, while novel elements (found in DHv2.1a1 and not in DHv2.0a1) made up about 33% of exons, introns and gene loci (Supp Table 7).

While inclusion of Iso-Seq RNA evidence allowed for improved accuracy of gene models and identification of alternative isoforms, we noted gene models that had support from short-read datasets, but lacked Iso-Seq evidence. This led us to further investigate Iso-Seq mapping to both DHv2.0a1 and DHv2.1a1 annotations and alignment characterization using featureCounts and SQANTI3 analysis (Liao et al. 2014; Pardo-Palacios et al. 2024). Using featureCounts, we found

that of the 722,655 Iso-Seq reads that mapped to DHv2.1a1, 94.2% were successfully assigned to genes, while 86.3% of the 722,389 reads that mapped to DHv2.0a1 were successfully assigned. Similarly, 94.8% of Iso-Seq sequences were successfully assigned to exons in DHv2.1a1 and 85.9% in DHv2.0a1. The high assignment rate for both gene models and exons suggests the new annotation captures most of the genes present in the Iso-Seq data and that at least 10% of genes were missing from DHv2.0a1. Using SQANTI3, a tool that characterizes mapping, isoform, and splice junction agreement of Iso-Seq data with a reference, we found that our Iso-Seq sequences mapped to 27,937 unique genes in DHv2.1a1 and 26,878 in DHv2.0a1 (Supp Table 8) (Pardo-Palacios et al. 2024). This included 1,888 novel genes, or genes not represented in the annotation, in DHv2.1a1, and 6,010 novel genes in DHv2.0a1. Isoform structural characterization classified Iso-Seq reads into Full Splice Matches (FSM), indicating the reference and query isoforms (transcript) have the same number of exons and all splice sites agree, Incomplete Splice Matches (ISM), in which the Iso-Seq transcript has fewer 5' exons than the gene model, but internal splice junctions agree, Novel In Category (NIC), in which the query uses a combination of known splice sites, and Novel Not in Category (NNC), in which the query has at least one splice site that is not annotated. Re-annotation of the DH genome increased the percentage of FSMs from 15.33% to 39.58% and decreased the percentage of NNCs from 45.05% to 22.89%. NIC isoforms increased from 10.9% to 24.29% and ISM percentages remained similar (5.09% and 6.69%). Rarer isoform categories, which included Genic, Genomic Antisense, Fusion, Intergenic, and Genic Intron were largely decreased in DHv2.1a1 (Supp Table 8). These improvements suggest an increased incorporation of information from Iso-Seq sequences into the DHv2.1a1 annotation.

DHv2.1a1 captures more orthogroups and shows increased similarity to other recent pome genome annotations

To assess how re-annotation and gene model improvement affected gene family completeness across the genome, we used PlantTribes2 tools to analyze orthogroups (Wafula et al. 2023). Three 'Bartlett' annotations were compared, including both DH versions and the 'Bartlett' v1.0 genome annotation (Chagne et al. 2014). Comparing these, DHv2.1a1 was able to capture the most orthogroups (10,215, Fig. 2A). The majority of orthogroups (8,987) were represented in all three annotations, while 283 were unique to the DHv2.1a1 re-annotation. Only 534 were found in 'Bartlett' v1.0 and DHv2.0a1 that were not represented in DHv2.1a1, including 248 in 'Bartlett' v1.0 alone, 195 in DHv2.0a1 alone, and 91 shared between these (Fig. 2A).

We next compared orthogroup representation and completeness compared with other recent pome fruit genomes by performing a Core Orthogroup (CROG) analysis. The CROGs for Rosaceae were determined previously, consisting of 9,656 orthogroups determined through comparison of 26 genomes from six genera within the Rosaceae family (Wafula et al. 2023). We compared DHv2.1 with DHv2.0, 'Bartlett' v1, and 12 recent, high-quality *Pyrus* and *Malus* genomes (Fig. 2B). The generated heat maps show individual orthogroups as each row, with warm and cool colors representing higher or lower gene counts in that orthogroup from selected annotations, normalized by z-scores. Both orthogroups and genomes are clustered based on similarity. The DHv2.0 annotation clusters alone on the far left, and contains large regions with cooler colors, suggesting that it is missing many genes from these CROGs. We found that the DHv2.1 re-annotation clustered more closely with other recent genomes, such as the 'd'Anjou' and *P. pyrifolia* genomes, suggesting a recovery of shared Rosaceae CROGs (Fig. 2B). Re-annotation also improved the capture of selected architecture-related genes, beyond the improvements found in our previous targeted analysis of these gene families (Supp Table 9) (Zhang et al. 2022). After additional polishing of the DHv2.0 genome, but prior to full re-annotation, we performed a targeted analysis of 18 architecture- and hormone-related gene families, showing that the number of genes increased from 53 (including 18 fragmented or truncated sequences) to 103 (including 2 fragmented or truncated sequences) in the polished genome (Zhang et al. 2022). Re-annotation increased this number to 108 genes with no fragmented or truncated sequences (Supp Table 9).

Discussion

Continued and iterative improvement of genome annotations allows genomic data to become increasingly useful to a broader audience. We found value in re-annotating this important genome, given the emergence of an improved annotation pipeline, availability of more short-read transcriptomes representing a greater number of tissue types, and inclusion of Iso-Seq data from broadly sampled tissues. The original annotation of the 'Bartlett' DH genome relied on fruit tissues and use of annotation pipelines available at the time, and still contained some adapter sequences, resulting in a large number of erroneous gene models (Linsmith et al. 2019; Zhang et al. 2022). Polishing and re-annotation has improved gene capture, gene model accuracy, inclusion of UTRs, and incorporation of isoforms.

The inclusion of Iso-Seq data allowed for improved gene model annotations and the addition of isoforms, but we noted some interesting trends in the sequence data. First,

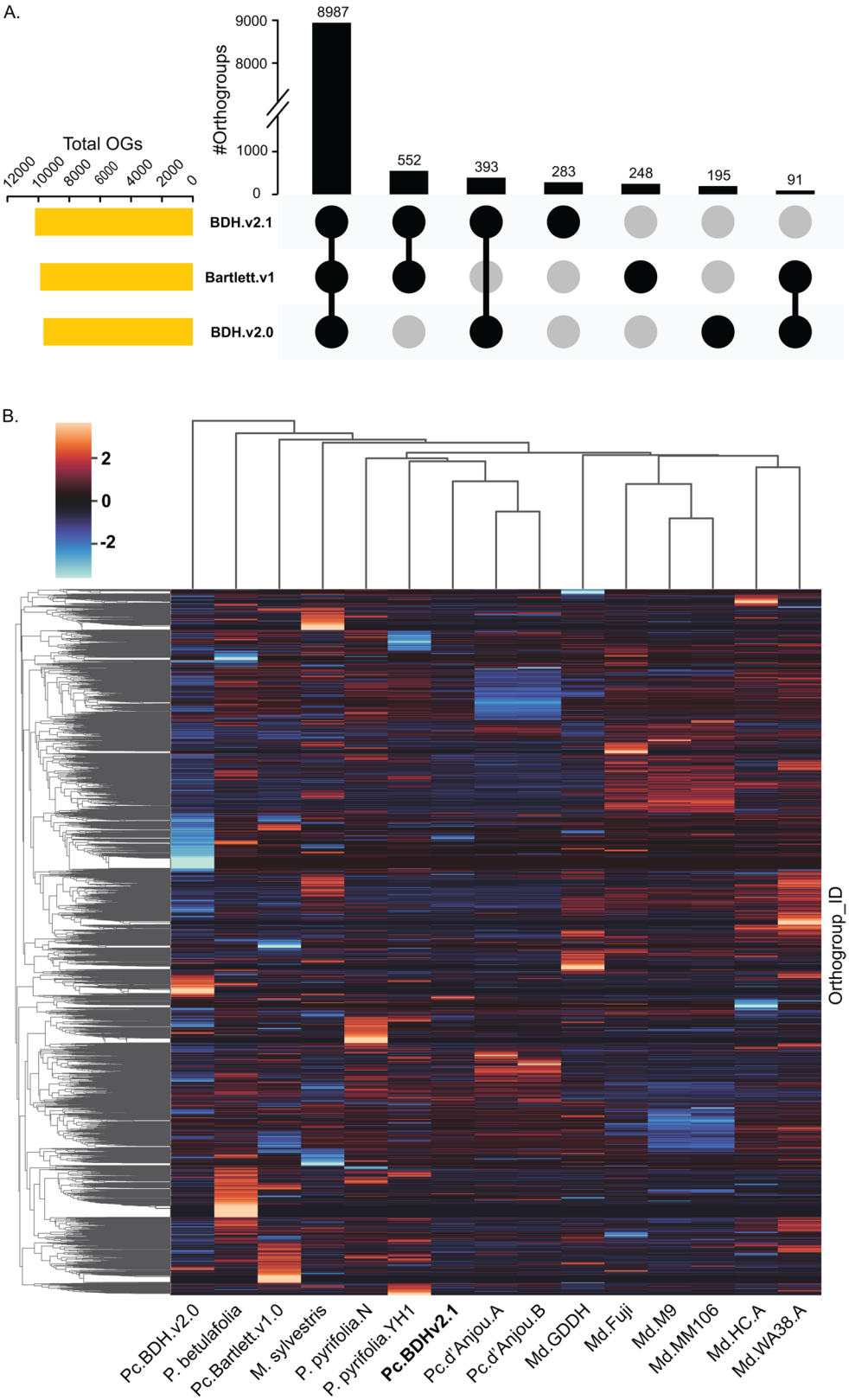


Fig. 2 Improved orthogroup representation resulting from re-annotation of the 'Bartlett' DH genome

(**A**) Upset plot indicating the number of orthogroups represented and shared between each 'Bartlett' genome, including the original annotation of the 'Bartlett' DHv2.0 genome, the current re-annotation, and the first version of the 'Bartlett' genome (Chagne et al. 2014). The black bars (upper right) show the number of orthogroups that are shared between the indicated genomes (lower right). The bar chart on the left shows the total number of represented orthogroups in each genome. (**B**) Heat map results from Core Orthogroups (CROG) analysis of 'Bartlett' and other recent pome genomes. Each row represents a CROG in Rosaceae, and color indicates gene counts in each CROG across the selected genomes, with z-score normalized by row. Warmer colors indicate a higher z-score and more genes in a CROG, cooler colors mean fewer genes. Rows and columns are clustered, with column clustering indicating genome similarity in terms of gene family contents. The 'Bartlett' DH v2.1a1 (bolded, current re-annotation) clusters with more recent genomes. Genome codes refer to: Pc.BDH.v2.0 - *Pyrus communis* 'Bartlett' Doubled Haploid v2.0, P. betulaefolia - *P. betulaefolia* 'Shanxi Duli', Pc.Bartlett.v1.0 - *P. communis* 'Bartlett' v1.0, M. sylvestris - *Malus sylvestris*, P.pyrifolia.N - *P. pyrifollia* 'Nijisseiki', P.pyrifolia.YH1 - *P. pyrifolia* 'Yunhong No.1', Pc.BDHv2.1 - *Pyrus communis* 'Bartlett' Doubled Haploid current reannotation, Pc.d'Anjou.A - *P. communis* 'd'Anjou' haplotype A, Pc.d'Anjou.B - *P. communis* 'd'Anjou' haplotype B, Md.Fuji - *M. domestica* 'Fuji', Md.M9 - *M. domestica* 'M9' rootstock, Md.MM106 - *M. domestica* 'MM106' rootstock, Md.HC.A - *M. domestica* 'Honeycrisp' haplotype A, Md.WA38.A - *M. domestica* 'WA 38' haplotype A

some genes had a large number of different reads, some of which may have been too short to constitute true isoforms. Some of these reads resulted from haplotypes, as the tissue used for Iso-Seq sequencing was not from a doubled haploid, and thus includes sequences from both haplotypes. Shorter reads may indicate the need for additional filtering of the data. Second, while the Iso-Seq data helped to capture additional genes, it only represented about 67% of annotated genes. Similar percentages have been reported in other genomes aligning broad transcriptomic data (Salojärvi et al. 2017), and is likely due to the inability to capture more specifically-expressed genes. This highlights the value of using multiple transcriptome datasets for RNA-seq evidence. Third, we noted examples in which Iso-Seq support was available, but gene models were pulled from other RNA or annotation inputs, which may reflect greater agreement between other input evidence (Supp Table 5).

Use of genome information for molecular biological applications or evolutionary genetics studies require confidence in gene model accuracy. For example, characterization of members of a gene family relies on the ability to find all family member sequences and have confidence that all domains and exons are accounted for, before drawing conclusions about their functions or developing clones (Vaattovaara et al. 2019). With non-model species' genomes, which often have less iterative corrections and annotations, this requires manual correction and curation. Commonly, transcriptome or isoform data is used to ensure sequences are correct, which not only takes a significant amount of time, but data may not be available for specific genes. When an annotation has numerous inaccurate gene models, this multiplies the time spent on manual curation. With this improved 'Bartlett' DH genome annotation, we hope to increase the efficiency of these downstream applications, including gene family investigation, cloning for genetic transformation, design of qRT-PCR primers, and evolutionary genetic studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11295-026-01734-z.

Author contributions Conceptualization – JMW, CG, LH, AH; Methodology – JE, HZ, CG, MW, JG, AH; Data Analysis – JE, HZ, CG, MW, JG; Writing – Original Draft Preparation – JMW, JE; Writing – Reviewing and Editing – All; Visualization – JE, HZ, JMW; Supervision and Project Administration – JMW, LH, AH. All authors have read and agreed to the submitted version of this manuscript.

Funding Funding was provided by USDA Agricultural Research Service in-house project 2094-21220-003-000-D.

Data availability The DHv2.1 polished genome version, the DHv2.1a1 annotation have been uploaded to the Genome Database for Rosaceae (GDR) (Jung et al. 2019) (www.rosaceae.org, publicly available upon publishing), and the Iso-Seq data submitted to NCBI (BioProject PRJNA1321962, SRR37599210).

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Andrews S (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. https://www.bioinformaticsbabraham.ac.uk/projects/fastqc/

Cannon EKS, Molik DC, Wright AJ, Zhang H, Honaas L, Chougule K, Dyer S (2025) Guidelines for gene and genome assembly nomenclature. Genetics 229:iyaf006. https://doi.org/10.1093/genetics/iyaf006

Chagne D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, Ireland H, Fiers M, Dzierzon H, Cestaro A, Fontana P et al (2014) The Draft Genome Sequence of European Pear (Pyrus communis L. 'Bartlett'). PLoS ONE 9:ARTNe92644. https://doi.org/10.1371/journal.pone.0092644

Chao KH, Heinz JM, Hoh C, Mao A, Shumate A, Pertea M, Salzberg SL (2025) Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Res 35:311–325. https://doi.org/10.1101/gr.279620.124

Dainat J (2024) AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. https://doi.org/10.5281/zenodo.3552717. Zenodo Version v1.4.2

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong XG, Wang Z, Tian LM, Zhang Y, Qi D, Huo HL, Xu JY, Li Z, Liao R, Shi M et al (2020) De novo assembly of a wild pear (Pyrus betuleafolia) genome. Plant Biotechnol J 18:581–595. https://doi.org/10.1111/pbi.13226

Ewels P, Magnusson M, Lundin S, Kaller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Freedman AH, Sackton TB (2025) Building better genome annotations across the tree of life. Genome Res 35:1261–1276. https://doi.org/10.1101/gr.280377.124

Gonzalez-Garay ML (2016) Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). In: Transcriptomics and Gene Regulation, Wu J. (ed.)Springer, Dordrecht

Hargarten H, Waliullah S, Kalcsits L, Honaas LA (2018) Leveraging Transcriptome Data for Enhanced Gene Expression Analysis in

Apple. J Am Soc Hortic Sci 143:333–346. https://doi.org/10.21273/jashs04424-18

Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, Kiel N, Saadat N, Ebenhöh O, Usadel B et al (2023) Helixer–*de novo* Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. https://doi.org/10.1101/2023.02.06.527280. bioRxiv

Jung S, Lee T, Cheng CH, Buble K, Zheng P, Yu J, Humann J, Ficklin SP, Gasic K, Scott K et al (2019) 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. Nucleic Acids Res 47:D1137–D1145. https://doi.org/10.1093/nar/gky1000

Kersey PJ (2019) Plant genome sequences: past, present, future. Curr Opin Plant Biol 48:1–8. https://doi.org/10.1016/j.pbi.2018.11.001

Khan A, Carey SB, Serrano A, Zhang H, Hargarten H, Hale H, Harkess A, Honaas L (2022) A phased, chromosome-scale genome of 'Honeycrisp' apple (Malus domestica). GigaByte 2022: gigabyte69 https://doi.org/10.46471/gigabyte.69

Kuster R, Smith Z, Whitt L, Staton M, Mansfeld B, Gottschalk C (2025) Ragnarok: a flexible and RApid GeNe Annotation (ROcKs) pipeline deployed through Nextflow. bioRxiv: 10.03.680343 https://doi.org/10.1101/2025.10.03.680343

Li H (2021) New strategies to improve minimap2 alignment accuracy. Bioinformatics 37:4572–4574. https://doi.org/10.1093/bioinformatics/btab705

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

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656

Linsmith G, Rombauts S, Montanari S, Deng CH, Celton JM, Guerif P, Liu C, Lohaus R, Zurn JD, Cestaro A et al (2019) Pseudo-chromosome-length genome assembly of a double haploid Bartlett pear (Pyrus communis L.). Gigascience 8: ARTN giz138 https://doi.org/10.1093/gigascience/giz138

Manni M, Berkeley MR, Seppey M, Zdobnov EM (2021) BUSCO: Assessing Genomic Data Quality and Beyond. Curr Protoc 1:e323. https://doi.org/10.1002/cpz1.323

Mansfeld BN, Yocca A, Ou S, Harkess A, Burchard E, Gutierrez B, van Nocker S, Gottschalk C (2023) 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. https://doi.org/10.1111/tpj.16433

Nevers Y, Warwick Vesztrocy A, Rossier V, Train CM, Altenhoff A, Dessimoz C, Glover NM (2024) Quality assessment of gene repertoire annotations with OMArk. Nat Biotechnol 124–133. https://doi.org/10.1038/s41587-024-02147-w

Ou C, Wang F, Wang J, Li S, Zhang Y, Fang M, Ma L, Zhao Y, Jiang S (2019) A de novo genome assembly of the dwarfing pear rootstock Zhongai 1. Sci Data 6:281. https://doi.org/10.1038/s41597-019-0291-3

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomas J, Amorin R, Estevan-Morio E, Liu T, Nanni A, McIntyre L et al (2024) SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods 21:793–797. https://doi.org/10.1038/s41592-024-02229-2

Pertea G, Pertea M (2020) GFF Utilities: GffRead and GffCompare. F1000Res 9: ISCB. Comm J-304 https://doi.org/10.12688/f1000research.23297.2

Ruhsam M, Bell D, Hart M, Hollingsworth P, Royal Botanic Garden Edinburgh Genome, Acquisition L Darwin Tree of Life Barcoding c, Wellcome Sanger Institute Tree of Life p, Wellcome Sanger Institute Scientific Operations DNAPc, Tree of Life Core Informatics c, Darwin Tree of Life C (2022) The genome sequence of the European crab apple, Malus sylvestris (L.) Mill., 1768. Wellcome Open Res 7: 296 https://doi.org/10.12688/wellcomeopenres.18645.1

Salojärvi J, Smolander O-P, Nieminen K, Rajaraman S, Safronov O, Safdari P, Lamminmäki A, Immanen J, Lan T, Tanskanen J et al (2017) Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nat Genet 49:904–912. https://doi.org/10.1038/ng.3862

Salzberg SL (2019) Next-generation genome annotation: we still struggle to get it right. Genome Biol 20:92. https://doi.org/10.1186/s13059-019-1715-2

Shirasawa K, Itai A, Isobe S (2021) Chromosome-scale genome assembly of Japanese pear (Pyrus pyrifolia) variety 'Nijisseiki'. DNA Res. https://doi.org/10.1093/dnares/dsab001. 28: ARTN dsab001

Shumate A, Wong B, Pertea G, Pertea M (2022) Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput Biol 18:e1009730. https://doi.org/10.1371/journal.pcbi.1009730

Sun M, Yao C, Shu Q, He Y, Chen G, Yang G, Xu S, Liu Y, Xue Z, Wu J (2023) Telomere-to-telomere pear (Pyrus pyrifolia) reference genome reveals segmental and whole genome duplication driving genome evolution. Hortic Res 10:uhad201. https://doi.org/10.1093/hr/uhad201

Unamba CI, Nag A, Sharma RK (2015) Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. Front Plant Sci 6:1074. https://doi.org/10.3389/fpls.2015.01074

UniProt C (2025) UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res 53. https://doi.org/10.1093/nar/gkae1010. D609-D617

Vaattovaara A, Leppala J, Salojarvi J, Wrzaczek M (2019) High-throughput sequencing data and the impact of plant gene annotation quality. J Exp Bot 70:1069–1076. https://doi.org/10.1093/jxb/ery434

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience 7:giy093. https://doi.org/10.1093/gigascience/giy093

Vuruputoor VS, Monyak D, Fetter KC, Webster C, Bhattarai A, Shrestha B, Zaman S, Bennett J, McEvoy SL, Caballero M, Wegrzyn JL (2023) Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. Appl Plant Sci 11:e11533. https://doi.org/10.1002/aps3.11533

Wafula EK, Zhang HT, Von Kuster G, Leebens-Mack JH, Honaas LA, dePamphilis CW (2023) PlantTribes2: Tools for comparative gene family analysis in plant genomics. Front Plant Sci 13:ARTN1011199. https://doi.org/10.3389/fpls.2022.1011199

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13:329–342. https://doi.org/10.1038/nrg3174

Yocca A, Akinyuwa M, Bailey N, Cliver B, Estes H, Guillemette A, Hasannin O, Hutchison J, Jenkins W, Kaur I et al (2024) A

chromosome-scale assembly for ‘d’Anjou’ pear. G3 (Bethesda) 14(jkae003). https://doi.org/10.1093/g3journal/jkae003

Zhang H, Ko I, Eaker A, Haney S, Khuu N, Ryan K, Appleby AB, Hoffmann B, Landis H, Pierro KA (eds) (2024) Jr. A Haplotype-resolved, Chromosome-scale Genome for Malus domestica Borkh. ‘WA 38’. G3 (Bethesda) 14: jkae222 https://doi.org/10.1093/g3journal/jkae222

Zhang HT, Wafula EK, Eilers J, Harkess AE, Ralph PE, Timilsena PR, dePamphilis CW, Waite JM, Honaas LA (2022) Building a foundation for gene family analysis in Rosaceae genomes with a novel workflow: A case study in Pyrus architecture genes. Front Plant Sci 13:ARTN975942. https://doi.org/10.3389/fpls.2022.975942

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