New annotations for three pea aphid genome assemblies allow comparative analyses of duplication and gene family evolution

Kevin D. Deem[1,*], Jennifer A. Brisson[1]

1. Department of Biology, University of Rochester, Rochester, NY, 14627
*corresponding author: kdeem@ur.rochester.edu

Abstract: Reliable genome annotation is crucial for analyses of gene function, conservation, duplication, and evolution. Factors such as the sequencing technology used to create the assembly, as well as duplication and rearrangements within the genome of interest, can have a large impact on the quality of gene annotations. In particular, short-read-based assemblies tend to mis-assemble duplicated genes as single loci, a problem that requires additional long-read sequencing to resolve. Pea aphids exhibit a high level of gene duplication from frequent genomic rearrangements, which has led to the mis-assembly and mis-annotation of genes. Here, we re-annotate the pea aphid reference genome, along with two long-read pea aphid genomes, to facilitate future analyses of gene duplication and function in pea aphids. We use an integrated approach, consolidating both *ab initio* and RNA-Seq-based annotations into unified gene models. The new annotations contain genes that were missing, mis-annotated, or mis-assembled in the reference genome, and are generally consistent across assemblies, showing very good agreement between the long-read assemblies. Our annotation method is sensitive enough to refine existing gene models, uncovering alternatively used promoters and isoforms, and aids in finding gene duplications. These data provide a useful supplement to the existing reference annotations and a new comparative framework for discovery and analysis of gene function and duplication in this important emerging model insect.

Keywords: gene, annotation, discovery, duplication, paralogs, RNA-Seq, pea aphids

Introduction

Proper assembly and annotation of genomes is crucial for bioinformatic and molecular biology investigations into gene function and evolution. As arguably one of the most important processes in gene family evolution, gene duplication presents a unique challenge for the proper assembly and annotation of genomes. The method(s) by which a genome assembly is produced have a substantial impact on the accuracy of gene duplicate assembly and annotation. One major deciding factor in accuracy of assembly is the choice of read type. Illumina short reads (50-300bp) are readily obtained in high numbers in a cost-effective manner. However, short-read-based genome assemblies are the most prone to mis-assemble duplicated genes, especially for recent duplications (Alkan et al., 2011; Mastretta-Yanes et al., 2014; Wang et al., 2021). Long-read technologies such as Pacific Bioscience (Eid et al., 2009) or Oxford Nanopore Technologies (Jain et al., 2016), provide the most accurate assembly of overall gene structure, even in the trickiest cases of recent or partial gene duplication (Fadaie et al., 2021; Logsdon et al., 2020; Nguyen, 2024; Showpnil et al., 2024).

Pea aphids (*Acyrthosiphon pisum*) are an important insect model for studying gene duplication (Fernández et al., 2020), the molecular genetic basis of morphological evolution (Li et al., 2020), and aphid-bacterial symbiont interactions (Price et al., 2011). However, these types of studies can be hindered by mis-assembly within the short-read pea aphid reference genome (Li et al., 2019). For instance, new long-read genome assemblies from winged and wingless males were required to properly assemble a region on the X chromosome controlling male wing dimorphism called *api* (Braendle et al., 2005; Li et al., 2020). This region was mis-assembled in the short-read reference genome (Li et al., 2019), but the assembled long-read genome revealed the presence of a duplicated gene (*follistatin-3*) as part of an insertion specific to wingless males (Li et al., 2020). More recently, we investigated aphid-specific duplications of wing gene regulatory network components and found that a number of these newly identified paralogs were incorrectly annotated in the reference genome (Saleh Ziabari et al., 2025a). These instances of mis-assembly and mis-annotation highlight the need for higher quality genome assemblies and annotations in this species.

Multiple methods exist for gene annotation, each with their own advantages and drawbacks. *Ab initio* methods such as AUGUSTUS (Stanke and Morgenstern, 2005) allow annotation of genes from raw genome sequence by predicting groups of exons that code for proteins. While these

methods require no additional data beyond the assembled genome to perform gene annotation, the accuracy and completeness of the annotation may suffer as a result. This accuracy can be increased by providing RNA-Seq data to better train models of splice sites using BRAKER1 (Hoff et al., 2016). Other RNA-Seq based methods of transcript discovery such as genome-guided Trinity (Grabherr et al., 2011) or StringTie (Pertea et al., 2015) utilize read sequence and/or coverage to identify exons and intron-spanning reads to determine how they are spliced into transcripts. These methods can be highly accurate and provide more complete gene models, but they can only identify genes expressed in RNA-Seq input (*i.e.* tissue-specific data will provide fewer genes than whole-body data). In addition, Trinity can also assemble transcriptomes *de novo* purely from RNA-Seq reads (the initial intended use of the software), potentially uncovering genes that are missing or mis-assembled in a given genome (Grabherr et al., 2011).

The software PASA takes a more unified approach, integrating Trinity *de novo*, Trinity genome-guided, BRAKER1, and StringTie transcript predictions into clustered gene models (Haas et al., 2008). These transcripts are first filtered for protein coding potential, quality of alignment to the genome, and intron size before being clustered into genes based on overlap (Haas et al., 2008). The advantage conferred by using PASA is that the complementary strengths of *ab initio* and RNA-Seq-based methods work synergistically when combined into a unified annotation.

Here we worked to develop a comprehensive reference annotation for the pea aphid genome. To do this, we integrated RNA-Seq data from many morphs (asexual females, sexual females, males; winged and wingless when available) and developmental stages (embryos, nymphal instars, and adults) with the reference genome (Li et al., 2019), as well as our own long-read assemblies (Li et al., 2020). We then used Trinity (genome-guided and *de novo*), StringTie, and BRAKER1 to individually predict transcripts, followed by PASA clustering of transcripts into unified gene models. We have made our long-read assemblies, a version of the reference assembly with a corrected *api* region, and their new annotations publicly available. These new gene annotations provide a rich source of information for future studies of duplication, function, and gene family evolution in pea aphids and other insects.

Methods

A summary diagram of the workflow described below is provided (**Supplemental Figure S1**).

98

99 **Generation of winged and wingless male embryo transcriptomes**

100 Pea aphid mothers producing winged and wingless males were generated as previously
101 described (Saleh Ziabari et al., 2025b). Briefly, asexual females from lines producing exclusively
102 winged (412) or wingless (409) males were reared in short day conditions to induce production
103 of males and sexual females. Once mothers stop producing females, they begin producing only
104 males, which was confirmed via PCR as in Saleh Ziabari et al. (2025b). Embryos were collected
105 from the ovaries of male-producing mothers and sorted into young (germaria through onset of
106 eye pigmentation) and old (onset of eye-pigmentation through birth) samples. Total RNA was
107 extracted from embryos via TRIzol-chloroform extraction (Invitrogen, Carlsbad, CA) and treated
108 with DNaseI. RNA clean-up was performed on 3ug of total RNA per replicate with the Zymo
109 Clean & Concentrator-5 kit (Zymo Research, Irvine, CA). Three biological replicates of each of
110 the four conditions (young and old wingless, young and old winged) were then sequenced on an
111 Illumina HiSeq 3000 producing 20-30M 2x150bp paired-end reads per replicate (Azenta Life
112 Sciences, Burlington, MA). Raw reads were submitted to NCBI under BioProject
113 PRJNA1380473.

114

115 **Trinity *de novo* transcriptome assembly.**

116 Only paired-end reads were used to generate the Trinity *de novo* assembly (Grabherr et al.,
117 2011). Reads were first adapter- and quality-trimmed with Trimmomatic v0.39 (Bolger et al.,
118 2014) in paired-end mode using the parameters “ILLUMINACLIP: TruSeq3-PE-2.fa:2:30:10
119 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25”. Reads were then randomly
120 subsampled to 2% using the view method in Sambamba v0.6.6 (Tarasov et al., 2015) with the
121 parameter “-s 0.02”. Trinity (v2.15.1) *de novo* assembly of transcripts was then performed in
122 singularity-ce v3.11.4 (Kurtzer et al., 2017) using the image file trinityrnaseq.v2.15.1.simg and
123 the parameters “--normalize_by_read_set --no_parallel_norm_stats --min_kmer_cov 2”.

124

125 **Read mapping and RNA-Seq-based transcript assembly.**

126 Both single-end and paired-end samples were used for RNA-Seq-based gene annotation
127 methods. Read accessions (Grantham et al., 2020; Purandare et al., 2014; Saleh Ziabari et al.,
128 2025b) are found in **Supplemental Table S1**. All of the following steps were performed on the
129 Winged Male Long-Read, Wingless Male Long-Read (Li et al., 2020), and what we call the
130 Modified Reference (Li et al., 2019; Saleh Ziabari et al., 2025b) genome assemblies. For the
131 latter, we modified the reference genome of (Li et al., 2019) to include an insertion that is

present solely in wingless males (Li et al., 2020). Specifically, we soft-masekd positions 62,565,000-62,660,000 on the X chromosome of that genome (Li et al., 2019) and added a new scaffold containing the properly assembled *api* insertion region from the Wingless Long-Read assembly (Li et al., 2020) named “NC_WLinsert_pilon10_start62575000_to_62660000” (Saleh Ziabari et al., 2025b).

Reads were aligned to each assembly with Hisat2 v2.2.1 (Kim et al., 2019) using default mapping parameters. The .sam files for each replicate from each sample were converted to .bam, sorted, merged, and indexed using SAMtools v1.13 (Li et al., 2009). Duplicate reads were removed from the merged replicate .bam files using the markdup function of Sambamba v0.6.6 with the parameter “--remove-duplicates”. These files were then each randomly subsampled at a rate of 20% using the view method of Sambamba with the parameter “-s 0.2”. These merged-replicate, duplicate-less, subsampled .bam files were then merged using SAMtools and used for all RNA-Seq-based annotation methods. Braker was run in ET-mode (Hoff et al., 2016) on the respective processed .bam file for each assembly in singularity-ce v3.11.4 (Kurtzer et al., 2017) using the image braker3.sif (Gabriel et al., 2024). Genome-guided transcriptome assembly was performed with these files for each assembly in Trinity v2.15.1 (Grabherr et al., 2011), also run in singularity-ce v3.11.4 with the image trinityrnaseq.v2.15.1.simg, using the flag “--genome_guided_bam” and the parameter “--genome_guided_max_intron 200000”. Transcripts were mapped to the genome using GMAP v2021-08-25 (Wu and Watanabe, 2005). StringTie v2.2.3 (Pertea et al., 2015) was run using default parameters and transcript sequences were extracted from genomic intervals using GffRead v0.12.7-2 (Pertea and Pertea, 2020).

Integrated gene clustering of transcripts using PASA.

Transcript files from Trinity *de novo* (Trinity-DN) and genome-guided (Trinity-GG) assembly, and genomic coordinate files output from StringTie and Braker, were filtered and clustered into genes using PASA v2.5.3 (Haas et al., 2008) for all three assemblies. Following extraction of Trinity-DN accessions, clustering was run in singularity-ce v3.11.4 using the image file pasapipeline.v2.5.3.simg with the following parameters: “--create --run -I 100000 --ALIGNERS gmap --ALT_SPLICE –TRANSDECODER”. The final output included nucleotide sequences and genomic intervals for high-confidence protein-coding transcript isoforms with a maximum intron size of 100kb clustered into ‘genes’.

Protein prediction, BUSCO analysis, and ortholog clustering.

Protein sequences coded by transcripts from Trinity-DN, and from Trinity-GG, StringTie -> GffRead, Braker, and PASA for each genome assembly, were predicted separately using the "Longest-ORFs.fasta" output of TransDecoder (Haas, n.d.) run either locally or on the Galaxy webserver (The Galaxy Community et al., 2024). Gene-transcript maps output from Trinity-DN and Trinity-GG were provided using the "–gene_trans_map parameter". For all transcript files, TransDecoder was run using the Universal Genetic Code with a minimum protein length of 100 amino acids "-m 100" in stranded mode "-S".

Reference protein sequences were obtained from NCBI (O'Leary et al., 2016) from the RefSeq genomes for pea aphids and twenty other invertebrate and vertebrate species [*Aedes aegypti* (GCF_002204515.2), *Aphis gossypii* (GCF_020184175.1), *Apis mellifera* (GCF_003254395.2), *Acyrthosiphon pisum* (GCF_005508785.2), *Bombyx mori* (GCF_030269925.1), *Caenorhabditis elegans* (GCF_000002985.6), *Drosophila melanogaster* (GCF_000001215.4), *Daphnia pulex* (GCF_021134715.1), *Daktulosphaira vitifoliae* (GCF_025091365.1), *Homarus americanus* (GCF_018991925.1), *Homo sapiens* (GCF_000001405.40), *Homalodisca vitripennis* (GCF_021130785.1), *Mus musculus* (GCF_000001635.27), *Macrobrachium nipponense* (GCF_015104395.2), *Myzus persicae* (GCF_001856785.1), *Nilaparvata lugens* (GCF_014356525.2), *Neocloeon triangulifer* (GCF_031216515.1), *Rhopalosiphum maidis* (GCF_003676215.2), *Rhopalosiphum padi* (GCF_020882245.1), *Schistocerca gregaria* (GCF_023897955.1), and *Tribolium castaneum* (GCF_031307605.1)].

Protein files from gene annotation methods of the three assemblies of interest, Trinity-DN, and the RefSeq genomes above were used as input proteomes for ortholog clustering using OrthoFinder v2.5.5 (Emms and Kelly, 2019). OrthoFinder clustered these proteins into so-called orthogroups, which are groups of orthologs inferred to be descended from a single gene in the last common ancestor of all species present in the analysis. This clustering was output in the OrthoGroups.tsv file (see Dryad dataset).

BUSCO v5.2.2 (Manni et al., 2021; Tegenfeldt et al., 2025) was run on the TransDecoder Longest ORFs protein output for PASA, Trinity-GG, StringTie, and Braker for each assembly, as well as Trinity-DN, and the RefSeq proteins. OrthoDB version 10 protein databases for the Eukaryota, Metazoa, Arthropoda, and Insecta levels were used (Tegenfeldt et al., 2025).

Annotation of genomic interval files.

The genomic intervals for mapped transcripts in PASA clusters for each genome assembly were annotated using the OrthoGroups.tsv file output from OrthoFinder, along with header information from the RefSeq protein fasta files for the twenty additional species obtained from NCBI, using

200 custom python scripts. For each potential protein encoded in a transcript in an orthogroup, the
201 first protein listed for a species within the same orthogroup had its header information appended
202 to the corresponding transcript entry in the genomic intervals file as “Species_ortholog:header
203 information”. In cases where a species had no proteins in that orthogroup, the ortholog is listed
204 as ‘none’ in the genomic intervals file. Annotated genomic interval files can be found in the
205 PASA subfolder for each assembly in the Dryad dataset (See **Data Availability Statement**).
206
207 **Estimating gene number.**
208 The following was performed using custom python scripts. A new orthogroup table was
209 generated replacing the protein(s) (often multiple per locus) from PASA annotations of the three
210 pea aphid assemblies with a single entry for each PASA gene cluster present for each assembly
211 in the orthogroup, along with the proteins present for all other species (and the RefSeq pea
212 aphid proteins). Orthogroups were then analyzed for the number of PASA clusters present for
213 each pea aphid assembly, as well as the number of orthogroup members from the following
214 groups: other pea aphid assemblies (including RefSeq), all non-pea aphid members, as well as
215 all non-aphid members. The total number of PASA clusters for each assembly was then
216 summed based on presence in an orthogroup containing: >1 pea aphid member, ≥1 non-pea
217 aphid member, ≥1 non-aphid member, ≥10 non-pea aphid members, or ≥10 non-aphid
218 members.
219
220 **Extracting novel annotations.**
221 Orthogroups containing PASA clusters for all three newly annotated assemblies, as well as ≥10
222 non-aphid members including *Drosophila melanogaster*, but no RefSeq pea aphid member,
223 were extracted from the orthogroup table generated above for further examination using custom
224 python scripts.
225
226 **Visualization of results.**
227 Plots were generated from comma-separated value files output from custom python scripts
228 visualized in Microsoft Office 365 Excel v2502 and edited in Adobe Photoshop v24.5.0. Gene
229 diagrams were first visualized in Integrative Genomics Viewer v2.16.1 using annotated genomic
230 interval files for PASA clusters in each assembly using a fixed window size across assemblies.
231 Gene annotations were then isolated and compiled in Adobe Photoshop v24.5.0.
232
233

Results

The numbers of transcripts, predicted proteins, proteins in orthogroups, and unassigned proteins, for PASA, Trinity genome-guided, StringTie and Braker for each assembly, as well as for Trinity *de novo* and pea aphid RefSeq annotations, are shown in **Figure 1**. The raw data for this figure can be found in **Supplemental Table S2**. Integration of transcripts from the other four methods into PASA clusters consistently yielded the highest number of transcripts coding for proteins in orthogroups by clustering orthologs of 20 other species with OrthoFinder (**Figure 1 A-C; see Supplemental Table S3 for ortholog clustering**). Braker consistently produced the highest percentage of unassigned proteins of the four genome + RNA-Seq-based methods (1.8%-3.2%) although Trinity *de novo* produced the highest percentage of unassigned proteins overall (5.6%, **Figure 1 A-D, Supplemental Table S2**). Trinity *de novo* also produced the lowest percentage of coding transcripts overall (34.8%), with Trinity genome-guided producing only slightly higher percentages of coding transcripts for each assembly (WD Male Long-Read: 35.8%, WL Male-Long Read: 35.9%, Modified Reference: 36.1%, **Figure 1 A-C, Supplemental Table S2**). Still, both Trinity methods produced a large number of total assembled transcripts relative to other methods, and thus still identified a high number of protein coding transcripts in orthogroups despite the low percentages (**Figure 1 A-D**, **Supplemental Table S2**). StringTie proved more conservative, consistently predicting much lower numbers of transcripts across each assembly (~55.9k-58.6k) compared to Trinity genome guided, but with a much higher percentage of coding transcripts (81.6%-87.9%) and the lowest number and percentage of unassigned proteins of any of the methods (144-196, 0.3%-0.4%, **Figure 1 A-C**, **Supplemental Table S2**).

To assess completeness of the proteomes produced by the various methods outlined above, we employed Benchmarking Universal Single-Copy Orthologs (BUSCO, Manni et al., 2021; Tegenfeldt et al., 2025). We evaluated proteome completeness for all methods, including the RefSeq annotations, at the Eukaryote, Metazoan, Arthropod, and Insect levels using version 10 databases from OrthoDB (Tegenfeldt et al., 2025). The results of this analysis are shown in **Figure 2** (see **Supplemental Table S4** for exact percentages). Integration of transcripts from the four other methods into PASA clusters yielded highly complete predicted proteomes comparable to RefSeq (**Figure 2 A-C, E**). Focusing on the Insecta level, the methods that produced the most complete predicted proteomes were Braker (94.0%-95.6%) and PASA (93.7%-94.9%). The proteomes produced by those two methods were only slightly less

268 complete than the RefSeq annotation (96.7%). This is despite Braker producing the highest
269 percentages of unassigned proteins, likely through detection of Universal Single Copy Orthologs
270 not expressed in our RNA-Seq dataset, which were then incorporated into PASA annotations.
271 StringTie proteomes were less complete than those produced by Braker and PASA (89.6%-
272 91.1%), but still considerably more complete than those produced by Trinity genome-guided
273 (54.0%-55.2%) or Trinity *de novo* (49.7%).
274
275 Given widespread gene duplication known to occur in pea aphids (Li et al., 2019; The
276 International Aphid Genomics Consortium, 2010), we next asked how many genes are predicted
277 within each of these genomes by our novel annotations. A recent gene number estimate, based
278 on ESTs, cDNA libraries, and WQ-maker predictions, made during assembly of the
279 chromosome-level reference genome, came to 40,090 (Li et al., 2019). We find that our
280 annotation methods produced comparatively high numbers of PASA clusters (141,397 [WD
281 Male LR], 145,815 [WL Male LR], 115,952 [Modified Reference]), likely containing
282 pseudogenes, transposons, and potentially, endosymbiont contamination.
283
284 We assessed more stringent gene estimates based on the number of PASA clusters (genes)
285 that code for a protein in an orthogroup with members from other pea aphid assemblies and/or
286 twenty other eukaryotic species (**Figure 3**). Requiring that a PASA cluster from a given
287 assembly codes for a protein in an orthogroup with at least one other protein from another pea
288 aphid assembly produced gene count estimates more similar to the 40,090 estimate of Li et al.,
289 2019 (48,932 [WD Male LR], 47,898 [WL Male LR], 31,888 [Modified Reference]) (**Figure 3,**
290 **Supplemental Table S5**). The number of genes in orthogroups drops sharply as requirements
291 on the number of orthogroup members from more distantly related species increases (**Figure**
292 **3**), suggesting that lineage-specific gene duplicates are rapidly diverging.
293
294 While a large number of newly annotated genes appear to be lineage-specific, there are also
295 ~1,000-1,500 more genes with ten or more non-aphid orthologs in the long-read genomes
296 versus the short-read reference (**Figure 3, Supplemental Table S5**). We searched for genes in
297 orthogroups with no RefSeq member but ten or more non-aphid members, including *Drosophila*
298 *melanogaster*, to identify newly annotated pea aphid orthologs of conserved fly genes. Using
299 this method, we found genes that are missing from the RefSeq annotations (**Figure 4A**), as well
300 as genes whose mis-annotation or mis-assembly disrupted protein clustering (**Figure 4B**).
301 These include missing annotations for *Centrocortin* (*Cen*), *Nad-kinase 2* (*Nadk2*), *Rad17*

302 *checkpoint clamp loader component* (*Rad17*), *Cytochrome-c oxidase assembly factor 5* (*COA5*)
303 (**Figure 4A**), and mis-annotation of the genes *trithorax* (*trx*), *Valette* (*Vlet*), and *snail* (*sna*)
304 (**Figure 4B**).
305
306 Based on our ability to detect and correct previously unannotated and mis-annotated genes, we
307 checked if these data support the previous identification of four paralogs of the BMP-branch
308 TGF-β ligand *dpp* (Brisson et al., 2010; Saleh Ziabari et al., 2025a). We were able to confirm the
309 presence and gene structure of all four paralogs across all three assemblies (**Figure 5A**).
310 Predicted gene structures, however, did differ between RefSeq features (red) and PASA clusters
311 (black). Our PASA clusters contain novel isoforms using alternative promoters for *dpp-1* and
312 *dpp-3* and predict that the promoter for *dpp-4* is upstream of the RefSeq annotations (**Figure**
313 **5A**). These refined promoter annotations for the *dpp* paralogs are consistent across all three
314 genome assemblies.
315
316 Recent findings have shown that paralogs of the gene *follistatin* are important for pea aphid
317 wing dimorphisms (Li et al., 2020; Saleh Ziabari et al., 2025b). While we have speculated that
318 morph-specific transcripts from *dpp* duplicates may play a role in these dimorphisms (Saleh
319 Ziabari et al., 2025a), potentially interacting with *follistatin* duplicates, Dpp is not known to
320 interact with Follistatin *in vivo* (Bickel et al., 2008; Pentek et al., 2009). For this reason, we set
321 out to identify potential gene duplication for the Activin branch TGF-β ligand and primary
322 Follistatin target, *myoglianin* (Bickel et al., 2008; Pentek et al., 2009). We identified three
323 paralogs of *myoglianin* present in all three assemblies (**Figure 5B**). Two fragmented RefSeq
324 annotations for *myo-1* are actually a single gene with a large duplicated exon containing much
325 of the coding sequence for the separate protein isoforms. PASA clusters in the Long-Read
326 assemblies agree on additional upstream promoters for *myo-1* that are absent from the RefSeq
327 features and PASA clusters in the Modified Reference, indicating potential mis-assembly of this
328 region in the refence (**Figure 5B**). The RefSeq annotation of *myo-2* indicates an alternative
329 downstream exon not supported by RNAseq-based annotation in any of the three genomes
330 (**Figure 5B**). *myo-3* is a pseudogene, with different intact ORFs present in each of the three
331 genome assemblies, responsible for differences in ortholog clustering between them (**Figure**
332 **5B**). That is, PASA clusters from the Modified Reference and WD male Long-Read assemblies
333 code for proteins that cluster with other *myoglianin* paralogs and orthologs, but the RefSeq
334 annotation and the PASA cluster in the WL Male Long-Read assembly do not.
335

Discussion

Here we present new gene annotations leveraging RNA-Seq data from across life stages, morphs, and sex, for three pea aphid genome assemblies. We integrated annotations from multiple methods (Trinity *de novo*, Trinity genome-guided, StringTie, Braker-ET) into PASA gene clusters, maximizing the number of proteins in orthogroups (**Figure 1**) and overall completeness as determined by BUSCO (**Figure 2**).

Gene count estimates from our annotations of the long-read (48,932 / 47,898) and short-read (31,888) genomes lie above and below, respectively, the previous estimate of 40,090 for the reference genome (**Figure 3, Supplemental Table S5**) (Li et al., 2019). For the long-read genomes, >20k of these genes are predicted to be lineage-specific, while only half as many lineage-specific genes are annotated in the short-read reference (**Figure 3**). This large disparity in gene count is not observed for more broadly conserved genes. New annotations on all three assemblies and RefSeq agree on 9,283-10,854 genes which cluster with orthologs in at least ten non-aphid species (**Figure 3, Supplemental Table S5**). There are likely a number of factors contributing to this large disparity in count estimates for new genes.

The difference in pea aphid gene counts for the reference genome (31,888 in this study, 40,090 in Li et al., 2019) may stem from the numerous differences in estimation methods, as well as our requirement that proteins cluster with those from another pea aphid assembly or the RefSeq proteins. One likely major contributor is that most of the 40,090 predicted genes from Li et al. (2019) do not have associated RefSeq annotations or reference proteins. Additionally, for rapidly evolving gene duplicates that are not part of the reference protein set, mis-assembly in the short-read reference genome may lead to predicted protein sequences that do not cluster with proteins from the properly assembled loci in the long-read genomes. Thus, these mis-assembled genes would not show up in our gene count estimate. Further, even when mis-assembly of duplicates into a single locus in the reference does not disrupt clustering with properly assembled paralogs in the long-read genomes, the gene count estimate for the reference only increases by one, while it increases by more than one for the long-read assemblies.

The two long-read genomes (WD and WL Male) were derived from closely related, inbred lines. Thus, when paralogs are properly assembled in one long-read genome, it is more likely that

370 they will be similarly assembled, annotated and clustered, keeping gene estimates similar.
371 However, there is also the possibility that the gene counts in the long-read genomes (48,932 /
372 47,898) are artificially inflated by redundant contig fragments. Given these factors, we find it
373 reasonable that we obtained gene counts above and below the 40,090 predicted by Li et al.
374 (2019) for our long-read and reference genomes, respectively. Still, further analyses are
375 required to determine where the true gene count lies within the range shown in **Figure 3**.
376
377 The new annotations provided by integrating multiple methods into PASA gene clusters have
378 identified previously missing or mis-annotated genes in the reference assembly (**Figure 4**).
379 These annotations have also helped in confirming previous reports on *dpp* gene duplication
380 (**Figure 5A**) as well as identified a previously unknown *myo* duplicate pseudogene (**Figure 5B**).
381 The new annotations also identify novel promoters for many of these gene duplicates, and
382 consistency of the gene models across the three assemblies strongly suggests that they are
383 accurate (**Figure 5**).
384
385 Differences were observed between the assemblies in the annotated gene structures and
386 protein clustering for the *myo3* pseudogene (**Figure 5B**). This seems to be due to differential
387 fragmentation of the ancestral ORF by different premature stop codons in each assembly,
388 disrupting transcript discovery, gene structure annotation, and protein clustering. This highlights
389 an additional challenge in initial pseudogene annotation and detection with a genome assembly
390 generated from a single genetic line. However, such disagreement in structure and coding
391 potential across assemblies may be a hallmark of pseudogenes. Detection of such
392 discrepancies requires a comparative approach, which the data presented here help to facilitate
393 for pea aphids.
394
395 These data provide a useful supplement to existing RefSeq annotations by identifying and
396 correcting missing and mis-annotated genes. The annotations of the two long-read genomes
397 provide a platform for investigating mis-assembly of multiple loci into a single locus in the short-
398 read reference genome. These annotations will aid in future investigations into the structure,
399 function, and evolution of genes and gene families.

Data availability statement:

The resources produced in this work have been made available through deposition to public repositories as well as inclusion in the Supplementary data files. The genome assemblies annotated in this work have been deposited to NCBI under BioProjects PRJNA1378164 (Modified Reference), PRJNA1378146 (Wingless Male Long-read), and PRJNA1378148 (Winged Male Long-Read). The male embryo RNA-Seq reads generated in this study were deposited under BioProject PRJNA1380473. Additionally, the predicted transcripts, proteins, and genomic coordinates produced by all methods for all three assemblies, annotated genomic coordinate files for PASA genes, genome sequences with original chromosome names, and the ortholog clustering information for all predicted proteins and those from 20 other species have been deposited to Dryad (https://doi.org/10.5061/dryad.s1rn8pknd).

Acknowledgements:

This research is supported by the National Institute of General Medical Sciences within the National Institutes of Health under award number R35GM144001 to JAB and by the National Science Foundation, Division of Biological Infrastructure, Postdoctoral Research Fellowships in Biology Program under award number 2305817 to KDD.

References:

Alkan, C., Sajjadian, S., Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. Nat. Methods 8, 61–65. https://doi.org/10.1038/nmeth.1527

Bickel, D., Shah, R., Gesualdi, S.C., Haerry, T.E., 2008. Drosophila Follistatin exhibits unique structural modifications and interacts with several TGF-β family members. Mech. Dev. 125, 117–129. https://doi.org/10.1016/j.mod.2007.09.013

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

Braendle, C., Caillaud, M.C., Stern, D.L., 2005. Genetic mapping of aphicarus – a sex-linked locus controlling a wing polymorphism in the pea aphid (Acyrthosiphon pisum). Heredity 94, 435–442. https://doi.org/10.1038/sj.hdy.6800633

Brisson, J.A., Ishikawa, A., Miura, T., 2010. Wing development genes of the pea aphid and differential gene expression between winged and unwinged morphs. Insect Mol. Biol. 19, 63–73. https://doi.org/10.1111/j.1365-2583.2009.00935.x

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323. https://doi.org/10.1126/science.1162986

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238. https://doi.org/10.1186/s13059-019-1832-y

Fadaie, Z., Neveling, K., Mantere, T., Derks, R., Haer-Wigman, L., Den Ouden, A., Kwint, M., O'Gorman, L., Valkenburg, D., Hoyng, C.B., Gilissen, C., Vissers, L.E.L.M., Nelen, M., Cremers, F.P.M., Hoischen, A., Roosing, S., 2021. Long-read technologies identify a

hidden inverted duplication in a family with choroideremia. Hum. Genet. Genomics Adv. 2, 100046. https://doi.org/10.1016/j.xhgg.2021.100046

Fernández, R., Marcet-Houben, M., Legeai, F., Richard, G., Robin, S., Wucher, V., Pegueroles, C., Gabaldón, T., Tagu, D., 2020. Selection following Gene Duplication Shapes Recent Genome Evolution in the Pea Aphid Acyrthosiphon pisum. Mol. Biol. Evol. 37, 2601–2615. https://doi.org/10.1093/molbev/msaa110

Gabriel, L., Brůna, T., Hoff, K.J., Ebel, M., Lomsadze, A., Borodovsky, M., Stanke, M., 2024. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 34, 769–777. https://doi.org/10.1101/gr.278090.123

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883

Grantham, M.E., Shingleton, A.W., Dudley, E., Brisson, J.A., 2020. Expression profiling of winged- and wingless-destined pea aphid embryos implicates insulin/insulin growth factor signaling in morph differences. Evol. Dev. 22, 257–268. https://doi.org/10.1111/ede.12326

Haas, B.J., n.d. https://github.com/TransDecoder/TransDecoder.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., Wortman, J.R., 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9, R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., Stanke, M., 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32, 767–769. https://doi.org/10.1093/bioinformatics/btv661

Jain, M., Olsen, H.E., Paten, B., Akeson, M., 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17, 239. https://doi.org/10.1186/s13059-016-1103-0

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kurtzer, G.M., Sochat, V., Bauer, M.W., 2017. Singularity: Scientific containers for mobility of compute. PLOS ONE 12, e0177459. https://doi.org/10.1371/journal.pone.0177459

Li, B., Bickel, R.D., Parker, B.J., Saleh Ziabari, O., Liu, F., Vellichirammal, N.N., Simon, J.-C., Stern, D.L., Brisson, J.A., 2020. A large genomic insertion containing a duplicated follistatin gene is linked to the pea aphid male wing dimorphism. eLife 9, e50608. https://doi.org/10.7554/eLife.50608

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y., Park, H., Smith, T.E., Moran, N.A., 2019. Gene Family Evolution in the Pea Aphid Based on Chromosome-Level Genome Assembly. Mol. Biol. Evol. 36, 2143–2156. https://doi.org/10.1093/molbev/msz138

Logsdon, G.A., Vollger, M.R., Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nat. Rev. Genet. 21, 597–614. https://doi.org/10.1038/s41576-020-0236-x

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M., 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic

Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. 38, 4647–4654. https://doi.org/10.1093/molbev/msab199

Mastretta-Yanes, A., Zamudio, S., Jorgensen, T.H., Arrigo, N., Alvarez, N., Piñero, D., Emerson, B.C., 2014. Gene Duplication, Population Genomics, and Species-Level Differentiation within a Tropical Mountain Shrub. Genome Biol. Evol. 6, 2611–2624. https://doi.org/10.1093/gbe/evu205

Nguyen, A.K., 2024. Duplications and Retrogenes Are Numerous and Widespread in Modern Canine Genomic Assemblies. Genome Biol. Evol. 16, 1–19. https://doi.org/10.1093/gbe/evae142

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. https://doi.org/10.1093/nar/gkv1189

Pentek, J., Parker, L., Wu, A., Arora, K., 2009. Follistatin preferentially antagonizes activin rather than BMP signaling in *Drosophila*. genesis 47, 261–273. https://doi.org/10.1002/dvg.20486

Pertea, G.M., Pertea, M., 2020. GFF Utilities: GffRead and GffCompare. F1000Research 9, 304. https://doi.org/10.12688/f1000research.23297.2

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295. https://doi.org/10.1038/nbt.3122

Price, D.R.G., Duncan, R.P., Shigenobu, S., Wilson, A.C.C., 2011. Genome Expansion and Differential Expression of Amino Acid Transporters at the Aphid/Buchnera Symbiotic Interface. Mol. Biol. Evol. 28, 3113–3126. https://doi.org/10.1093/molbev/msr140

Purandare, S.R., Bickel, R.D., Jaquiery, J., Rispe, C., Brisson, J.A., 2014. Accelerated Evolution of Morph-Biased Genes in Pea Aphids. Mol. Biol. Evol. 31, 2073–2083. https://doi.org/10.1093/molbev/msu149

Saleh Ziabari, O., Deem, K.D., Zhong, Q., Brisson, J.A., 2025a. Analysis of Duplication and Potential Functional Divergence of Wing Gene Network Components in Pea Aphids. Evol. Dev. 27, e70021. https://doi.org/10.1111/ede.70021

Saleh Ziabari, O., Liu, F., Deem, K.D., Liu, X., Kholwadwala, A., Brisson, J.A., 2025b. Gene duplication captures morph-specific promoter usage in the evolution of aphid wing dimorphisms. Proc. Natl. Acad. Sci. 122, e2420893122. https://doi.org/10.1073/pnas.2420893122

Showpnil, I.A., E. Hernandez Gonzalez, M., Ramadesikan, S., Marhabaie, M., Daley, A., Dublin-Ryan, L., Pastore, M.T., Gurusamy, U., Hunter, J.M., Stone, B.S., Bartholomew, D.W., Manickam, K., Miller, A.R., Wilson, R.K., Stottmann, R.W., Koboldt, D.C., 2024. Long-read genome sequencing resolves complex genomic rearrangements in rare genetic syndromes. Npj Genomic Med. 9, 66. https://doi.org/10.1038/s41525-024-00454-4

Stanke, M., Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33, W465–W467. https://doi.org/10.1093/nar/gki458

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., Prins, P., 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E.M., Kriventseva, E.V., 2025. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. Nucleic Acids Res. 53, D516–D522. https://doi.org/10.1093/nar/gkae987

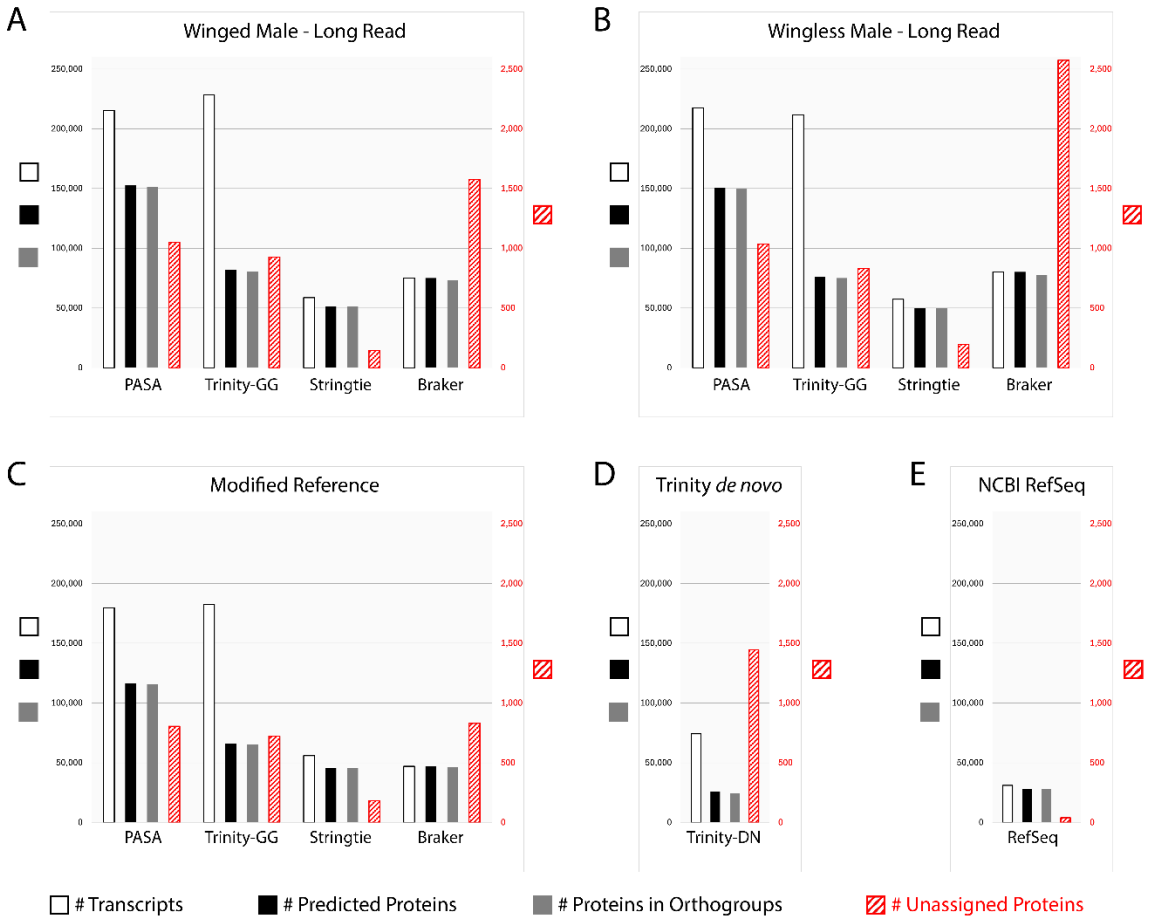
The Galaxy Community, Abueg, L.A.L., Afgan, E., Allart, O., Awan, A.H., Bacon, W.A., Baker, D., Bassetti, M., Batut, B., Bernt, M., Blankenberg, D., Bombarely, A., Bretaudeau, A., Bromhead, C.J., Burke, M.L., Capon, P.K., Čech, M., Chavero-Díez, M., Chilton, J.M., Collins, T.J., Coppens, F., Coraor, N., Cuccuru, G., Cumbo, F., Davis, J., De Geest, P.F., De Koning, W., Demko, M., DeSanto, A., Begines, J.M.D., Doyle, M.A., Droesbeke, B., Erxleben-Eggenhofer, A., Föll, M.C., Formenti, G., Fouilloux, A., Gangazhe, R., Genthon, T., Goecks, J., Beltran, A.N.G., Goonasekera, N.A., Goué, N., Griffin, T.J., Grüning, B.A., Guerler, A., Gundersen, S., Gustafsson, O.J.R., Hall, C., Harrop, T.W., Hecht, H., Heidari, A., Heisner, T., Heyl, F., Hiltemann, S., Hotz, H.-R., Hyde, C.J., Jagtap, P.D., Jakiela, J., Johnson, J.E., Joshi, J., Jossé, M., Jum'ah, K., Kalaš, M., Kamieniecka, K., Kayikcioglu, T., Konkol, M., Kostrykin, L., Kucher, N., Kumar, A., Kuntz, M., Lariviere, D., Lazarus, R., Bras, Y.L., Corguillé, G.L., Lee, J., Leo, S., Liborio, L., Libouban, R., Tabernero, D.L., Lopez-Delisle, L., Los, L.S., Mahmoud, A., Makunin, I., Marin, P., Mehta, S., Mok, W., Moreno, P.A., Morier-Genoud, F., Mosher, S., Müller, T., Nasr, E., Nekrutenko, A., Nelson, T.M., Oba, A.J., Ostrovsky, A., Polunina, P.V., Poterlowicz, K., Price, E.J., Price, G.R., Rasche, H., Raubenolt, B., Royaux, C., Sargent, L., Savage, M.T., Savchenko, V., Savchenko, D., Schatz, M.C., Seguineau, P., Serrano-Solano, B., Soranzo, N., Srikakulam, S.K., Suderman, K., Syme, A.E., Tangaro, M.A., Tedds, J.A., Tekman, M., Cheng (Mike) Thang, W., Thanki, A.S., Uhl, M., Van Den Beek, M., Varshney, D., Vessio, J., Videm, P., Von Kuster, G., Watson, G.R., Whitaker-Allen, N., Winter, U., Wolstencroft, M., Zambelli, F., Zierep, P., Zoabi, R., 2024. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Res. 52, W83–W94. https://doi.org/10.1093/nar/gkae410

The International Aphid Genomics Consortium, 2010. Genome Sequence of the Pea Aphid Acyrthosiphon pisum. PLoS Biol. 8, e1000313. https://doi.org/10.1371/journal.pbio.1000313

Wang, P., Meng, F., Moore, B.M., Shiu, S.-H., 2021. Impact of short-read sequencing on the misassembly of a plant genome. BMC Genomics 22, 99. https://doi.org/10.1186/s12864-021-07397-5

Wu, T.D., Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

579 **Figures:**



580

581 **Figure 1. Relative amounts of coding and non-coding transcripts annotated by various**
582 **methods.**

583 Number of predicted transcripts (white), protein-coding transcripts (black), and protein-coding
584 transcripts in orthogroups (gray) are shown on the left axis. Number of predicted transcripts
585 coding for proteins that were not assigned to any orthogroup (red hatched) are indicated on the
586 right axis. All plots share the same scales for both axes. (A-C) Numbers are shown for genome
587 + RNA-Seq-based methods for each assembly. (D) Numbers for Trinity *de novo* assembly of
588 transcripts directly from RNA-Seq reads, integrated with predicted transcripts from the other
589 three methods to produce PASA annotations for each assembly. (E) Numbers for RefSeq
590 annotations of transcripts and proteins in the reference genome.

591

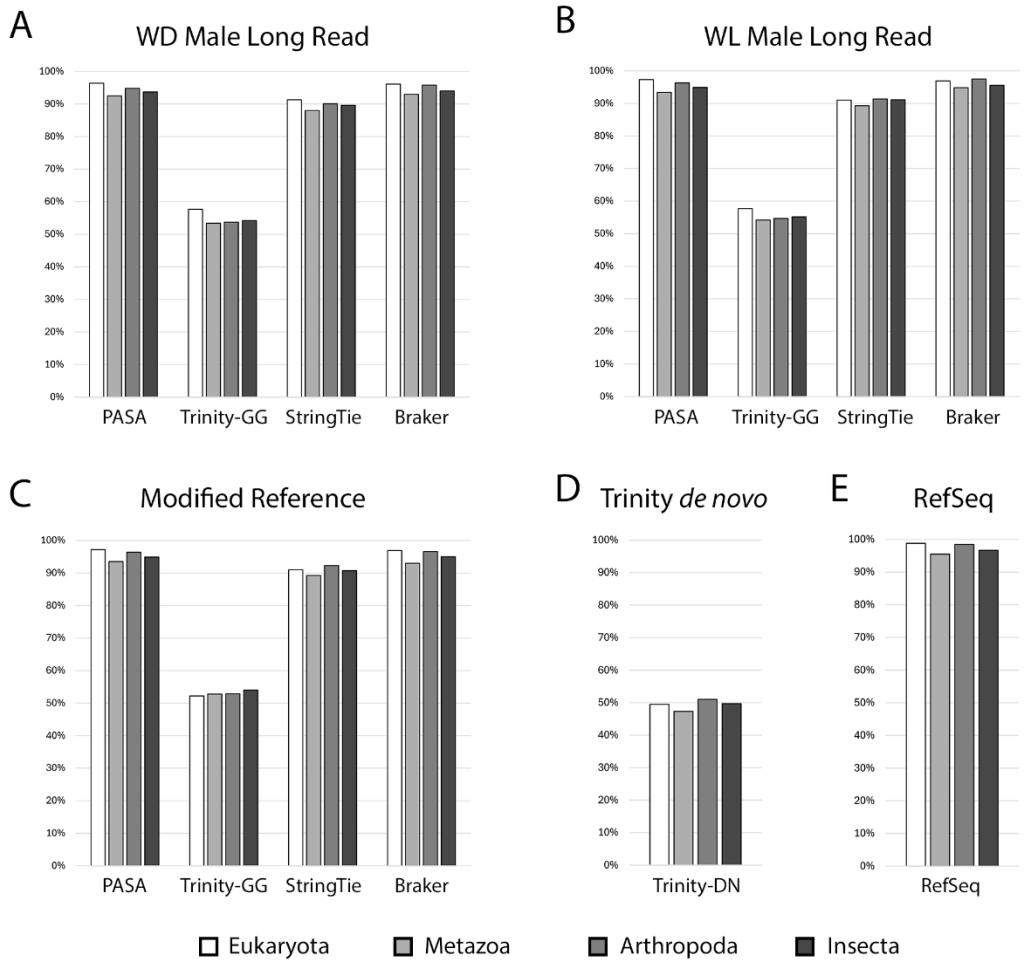


Figure 2. BUSCO analysis of proteomes

The percent of complete BUSCO's contained in PASA, Trinity genome-guided (Trinity-GG), StringTie, and Braker proteomes for the WD Male Long Read (A), WL Male Long Read (B), and Modified Reference (C) assemblies, as well as Trinity *de novo* (Trinity-DN, D) and RefSeq proteomes (E). Bar shade corresponds to the OrthoDB v10 dataset used: Eukaryota (white), Metazoa (light gray), Arthropoda (gray), or Insecta (dark gray).

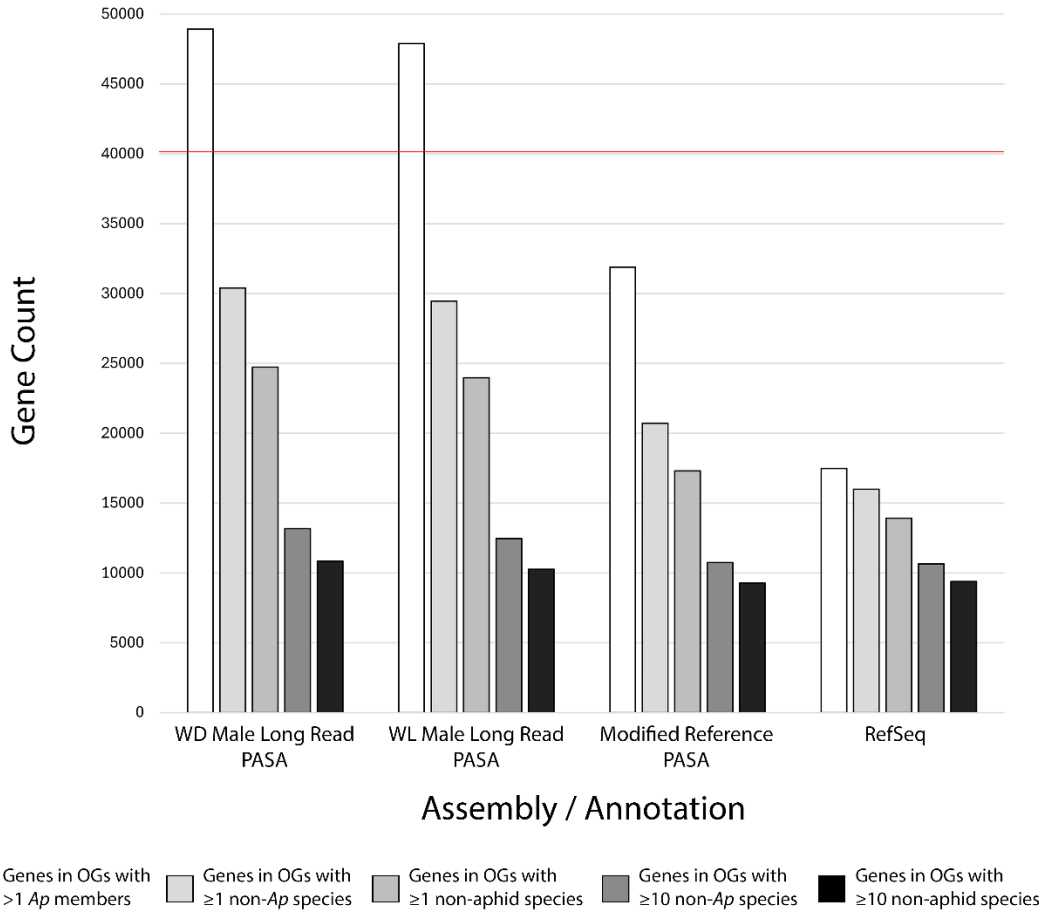


Figure 3. Gene count estimates in pea aphid assemblies with orthologs of varying evolutionary distance.

Numbers of PASA gene clusters for each assembly, or RefSeq genes, producing protein-coding transcripts assigned to orthogroups containing members of varying evolutionary distance. The categories are as follows: at least one other member from another pea aphid assembly (white), at least 1 non-pea aphid species (light grey), at least one non-aphid species (medium grey), at least 10 non-pea aphid species (dark gray), or at least 10 non-aphid species (black). The red line indicates the 40,090 gene estimate from the most recent genome assembly (Li et al., 2019).

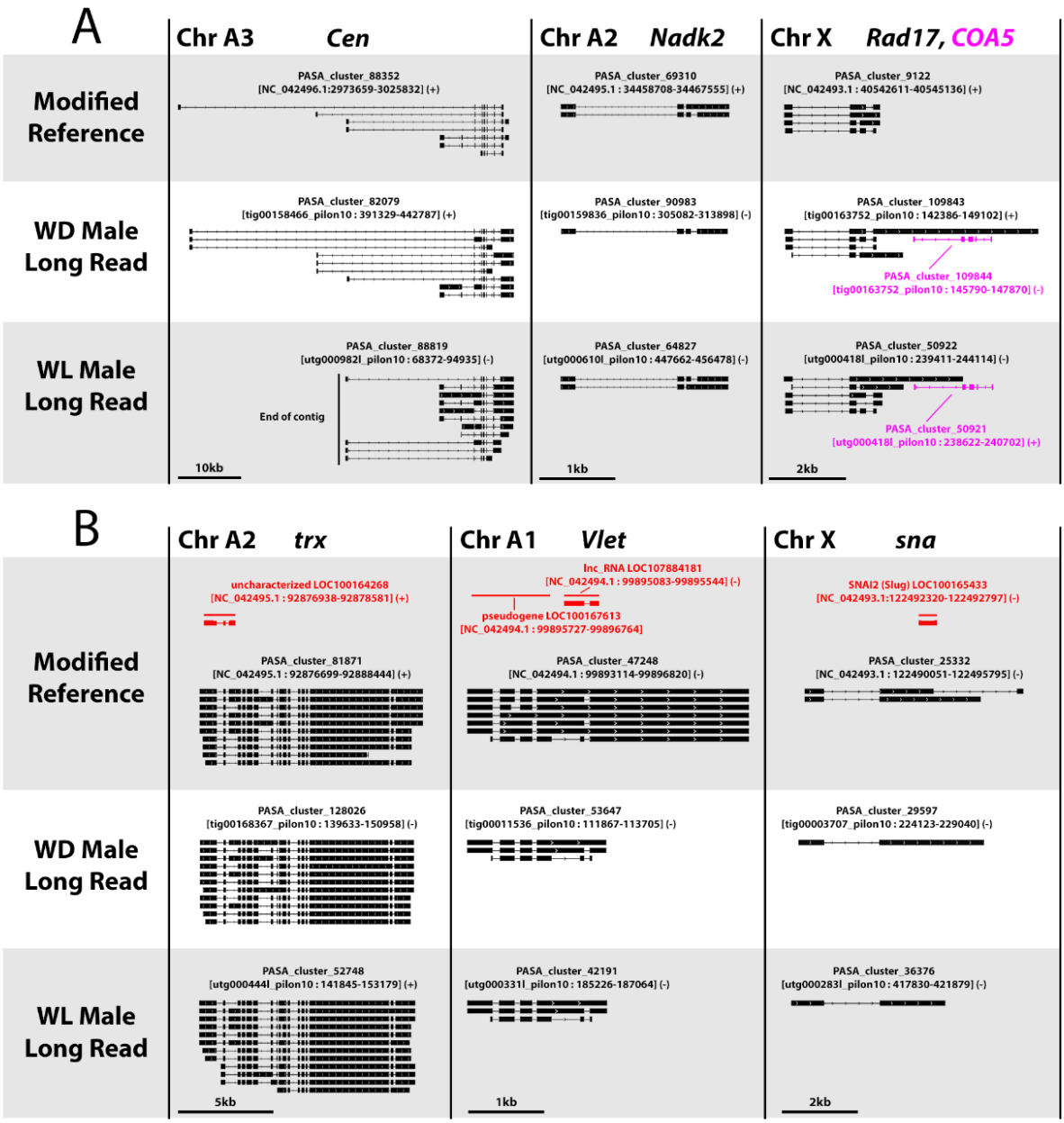


Figure 4. PASA assemblies of missing and mis-annotated genes

Gene structure diagrams for PASA-assembled genes that are either missing or incorrectly annotated in the RefSeq features. RefSeq annotations are shown in red, while newly annotated PASA clusters in the Modified Reference, WD Male Long Read, and WL Male Long Read assemblies are shown in black or magenta. The chromosomal location within the Modified Reference genome is indicated next to the gene name(s). Genes are shown at different scales, indicated by scale bars at the bottom. For clarity, overlapping annotations of other genes are not shown. (A) PASA-assembled gene structures for the genes *Centrocortin* (*Cen*), *Nad-kinase 2*

618 (*Nadk2*), *Rad17 checkpoint clamp loader component* (*Rad17*) (all black), and *Cytochrome-c*
619 *oxidase assembly factor 5* (*COA5*, magenta) that are absent from the RefSeq annotation. The
620 edge of the contig adjacent to the WL Male Long Read annotation of *Cen* is marked by a black
621 line. (B) PASA assemblies for the genes *trithorax* (*trx*), *Valette* (*Vlet*) and *snail* (*sna*) that are
622 incorrectly annotated in the RefSeq features.
623

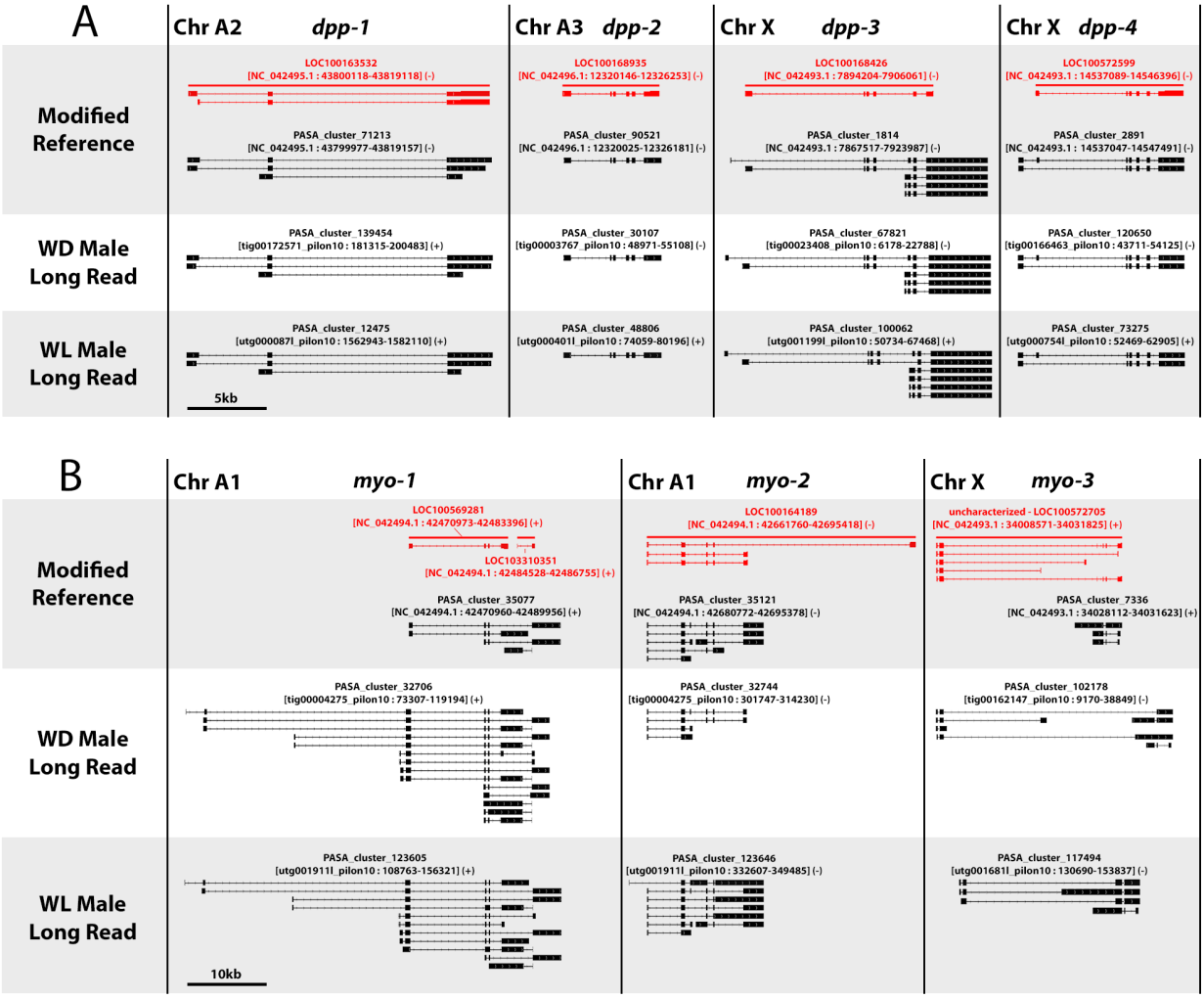


Figure 5. Annotations of duplicated TGF-β signaling ligands

Gene structure diagrams for paralogs of the TGF-β signaling ligands *decapentaplegic* (*dpp*) (A) and *myoglianin* (*myo*) (B) across pea aphid genome assemblies. RefSeq annotations are shown in red, PASA annotations are shown in black. For clarity, overlapping annotations of other genes are not shown. A) Annotations for all four *dpp* paralogs, shown at the same scale. Scale bar: 5kb B) Annotations for the three paralogs of the gene *myo*, shown at the same scale. Scale bar: 10kb