Accurate gene copy number prediction in genome sequences using gmap_gene_find (GGF)

Alternatives: Accurate prediction of gene model copies for genome sequences using GGF; Automatic genome annotation of gene copies using gmap_gene_find (GGF)

Zachary K. Stewart[1*], David R. Lovell[2], Peter J. Prentis[1,3]

[1]School of Earth, Environmental and Biological Sciences, Queensland University of Technology, 2 George St, Brisbane, Australia

[2]School of Electrical Engineering and Computer Science, Queensland University of Technology, 2 George St, Brisbane, Australia

[2]Institute for Future Environments, Queensland University of Technology, 2 George St, Brisbane, Australia

*Corresponding author; Stewart, ZK: zkstewart1@gmail.com

Abstract

Automated gene model prediction is an essential component of virtually all new genome annotation projects. The bioinformatic tools used to accomplish this, despite their importance and utility, can suffer from an inability to annotate all gene copies in a genome, especially when near-identical genes have been duplicated numerous times. Accurate annotation of all gene copies is important for studies assessing the evolution of gene families and is essential for generating a high-quality genome annotation. This study describes gmap_gene_find (GGF), a program designed to augment existing gene annotations by identifying additional gene copies through alignment of coding sequence to genome assemblies. To ascertain how this program may improve existing gene annotations we generated new annotations for four model species and assessed how the inclusion of GGF in the annotation pipeline affects the results via comparison to official annotations available for these species. GGF improved genome annotations in all four species as measured by sensitivity, precision, and their combined F1 score. Moreover, the annotation of additional gene copies resulted in gene family sizes that corresponded more closely with primary annotations of these species. Consequently, GGF is a valuable tool for the annotation of genomes and is available from https://github.com/zkstewart/Genome_analysis_scripts/tree/master/ggf.

Keywords: genome annotation, gene models, gene prediction, bioinformatics, genomics, PASA

Introduction

Recent advancements in long-read sequencing has increased the number of new genome assembly projects being undertaken. While many projects may be considered "reassemblies" for model species e.g., *Aedes aegypti*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, human, chimpanzee, and orangutan (Jain et al., 2018; Kronenberg et al., 2018; Matthews et al., 2018; Michael et al., 2018; Pendleton et al., 2015; Salazar et al., 2017; Seo et al., 2016; Solares et al., 2018; Tyson et al., 2018), the relative ease with which long reads can be used for genome assembly has enabled researchers to sequence species that receive less scientific attention. As an example, genome sequencing may now be used as a means to draw public and scientific attention to endangered and threatened species including the European eel (Jansen et al., 2017), the Hawaiian crow (Sutton et al., 2018), the Murray cod (Austin et al., 2017) and the African wild dog (Armstrong et al., 2017).

Once an assembly has been generated, gene prediction (also referred to as "annotation") takes place (Dominguez Del Angel et al., 2018). New genome projects use many tools to annotate gene models. These bioinformatic programs can be broadly categorised according to their methodological approach to gene prediction; specifically, these refer to homology-based prediction, *ab initio* prediction, and RNA-based prediction. Homology-based prediction relies upon sequence alignment of known genes against the genome of interest. An example of this is GeMoMa (Keilwagen, Hartung, Paulini, Twardziok, & Grau, 2018) which, as a core part of its pipeline, utilises the tblastn tool (Altschul, Gish, Miller, Myers, & Lipman, 1990) for sequence alignment. This approach is highly successful for reassembled genomes or if a genome sequence for a closely related species is available. Such methods, however, may not identify many species-specific genes – that is, genes not found in other species – and these genes may make up 10 – 20 % of the total gene repertoire of a species (Khalturin, Hemmrich, Fraune, Augustin, & Bosch, 2009). *Ab initio* methods involve the prediction of gene models directly from genomic sequence by building a comprehensive statistical model that can differentiate coding from non-coding sequence (Wang, Chen, & Li, 2004). Some notable examples of *ab initio* prediction programs include AUGUSTUS (Stanke, Steinkamp, Waack, & Morgenstern, 2004) and GlimmerHMM (Majoros, Pertea, & Salzberg, 2004). Although the sensitivity of *ab initio* prediction can approach 100 % (Yandell & Ence, 2012), these methods have comparatively poor accuracy when compared to homology or RNA-based prediction. For many newly sequenced genomes, RNA sequencing data offers the greatest potential for accurately predicting gene models while also capturing species-specific genes (Yandell & Ence, 2012). Programs such as TopHat and Cufflinks (Trapnell et al., 2012) can be used to directly produce gene models after RNA-seq read alignment to the genome; however, depending on genome attributes such as gene density, this approach can lead to incorrect merging of adjacent gene models. In such cases one can assemble these reads into a transcriptome using *de novo* approaches (i.e., without using the genome as a guide) such as Trinity (Haas et al., 2013) before alignment of these sequences to the genome. For this reason, the Program to Assemble Spliced Alignments (PASA; Haas et al., 2008) still sees use in genome projects.

PASA automates a pipeline of utilities that includes alignment of assembled transcripts to the genome using GMAP (Wu & Watanabe, 2005) and/or BLAT (Kent, 2002) with subsequent processing to derive gene models including alternatively spliced isoforms. Although PASA provides a powerful means to annotate gene models, it is not without weakness. Specifically, PASA's alignment is performed with a "best hit" procedure whereby transcripts will be aligned to the single best position in the genome. While this may result in more accurate gene model annotation, it can also result in a reduced copy number discovered for genes that are duplicated numerous times in a genome and conserved extensively. A case where this might occur is in duplicated toxin genes, wherein genes with little to no sequence divergence may be present at extremely high copy numbers. An example is the myotoxin gene of certain viper species which may be present in greater than twenty copies (Margres, Bigelow, Lemmon, Lemmon, & Rokyta, 2017). Identical and near-identical transcripts derived from genes such as these will likely be “hidden” in transcriptome assemblies (i.e., multiple real, biological transcripts may be represented by only one assembled transcript; Macrander, Broe, & Daly, 2015) and, subsequently, one can expect PASA to potentially miss a significant number of real gene predictions due to its “best hit” alignment procedure.

This observation led to the creation of the bioinformatics software described herein: gmap_gene_find (GGF), a program coded in the Python language that aims to address the aforementioned issues. GGF relies upon GMAP alignment of transcripts to the genome similarly to PASA but with modified GMAP parameters to capture gene models present at higher copy numbers. To demonstrate the utility of the program and assess its performance, a study using the model organisms *Saccharomyces cerevisiae* (baker’s yeast), *Caenorhabditis elegans* (nematode worm), *Arabidopsis thaliana* (thale cress), and *Mus musculus* (house mouse) was performed. These species represent eukaryote taxa from different phyla or kingdoms with varying genome sizes and, importantly, have existing high-quality gene set annotations. For this study, we used publicly available RNA-seq data to annotate the gene models of these organisms using PASA alone and a combination of PASA with GGF and compared these gene models to official annotations to assess the merits and/or drawbacks of using GGF in a genome project. The code for GGF is available from https://github.com/zkstewart/Genome_analysis_scripts/tree/master/ggf under GNU General Public License v3.0 agreement.

Materials and Methods

gmap_gene_find code overview

The gmap_gene_find (GGF) program was written in the Python 3.6 coding language. Third-party Python prerequisites include biopython (Cock et al., 2009), skbio, pandas, and ncls (https://github.com/hunt-genes/ncls; Alekseyenko & Lee, 2007). Currently, the skbio and ncls packages are not compatible with Windows operating systems, which limits the operation of GGF; in the future, GGF may be Windows-compatible if these underlying packages are modified. All code testing and operation was conducted on Linux-based operating systems.

A simplified overview of GGF's internal operations are depicted in Figure 1 with a complete description of the program and its parameters available at the GGF repository. Prior to program start, certain inputs are required to be generated. Firstly, a FASTA file consisting solely of coding DNA sequence (CDS) should be produced. These sequences should originate from the species in which gene models are being predicted, and while it is suggested that these sequences be derived from computational prediction of CDS from transcriptomes using programs like TransDecoder (Haas et al., 2013) or EvidentialGene (Gilbert, 2016), GGF will work with the alignments of any CDS. GMAP alignment of CDS to genome sequence should be performed with the "-f 2" argument to produce a GFF3 (General Feature Format 3) file. Furthermore, GMAP should be configured to allow each sequence to align to multiple sites by specifying a value greater than one to the "-n" parameter. The GMAP file and its corresponding CDS in FASTA format should be provided to GGF in addition to the genome sequence (as FASTA) and a previously existing gene model annotation file (as GFF3).

Once the necessary inputs have been provided as command-line arguments, GGF parses one or more input GMAP GFF3 file(s) and extracts alignments that meet minimum identity and coverage cut-off and which also contain two or more exons. Next, CDS alignment boundaries are refined and any exons that do not contain CDS are removed from the model. After this has occurred, a final CDS extension is performed within the boundaries of upstream and downstream stop codons to maximally extend the CDS region up to an earlier in-frame ATG or, if the original CDS queried against the genome by GMAP had an alternative codon (e.g., CTG), it may be extended to this codon. The potential CDS model derived from this process is then subjected to curation checks to ensure that the model is of sufficient quality for further consideration. Sequences which pass these checks are compared to the existing annotation GFF3 and any models which have greater than 35 % of their length overlapping previously annotated genes will be removed. At this stage there might be some redundancy in the gene models if multiple highly similar transcripts exist or multiple input GFF3 files were provided. All potential gene models which overlap are compared to each other in a pairwise fashion to pick the highest quality model, wherein quality is assessed by a ranking system of criteria relating to sequence lengths (exon, intron, and total) and splice sites (canonical, non-canonical and rare, and unknown). Non-redundant gene models are then subjected to several final curation checks intended to remove gene models that are not likely to represent real predictions i.e., models which may be artefacts of CDS alignment to regions of the genome that code for genes with similar but non-identical sequence. Sequences that are not removed during this or any of the previously mentioned curation and redundancy removal steps are considered as genuine genes and an output GFF3 file styled to resemble PASA's output is produced.

RNA-seq datasets and transcriptome assembly

Publicly available RNA-seq datasets were downloaded from the NCBI short read archive (SRA) for each of the four species. Datasets were non-systematically selected from recently uploaded projects which had approximately 7.5 – 8.0 Gbp of paired-end sequencing performed on Illumina HiSeq instruments. Specifically, we chose the following; *S. cerevisiae*: SRR5963435 (Nielsen et al., 2017), *C. elegans*: SRR5849945 and SRR5849946 (unpublished), *A. thaliana*: SRR6814509 (Nallu et al., 2018), and for *M. musculus* we obtained a consistently sized dataset with the others (7.8 Gbp; SRR7828327; unpublished) and a high-coverage dataset (15 Gbp; ERR2540221; unpublished) as transcriptome assembly of the 7.8 Gbp dataset did not appear to obtain high completeness using BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) (detailed later).

Comprehensive transcriptomes were built using the datasets collected by use of multiple transcriptome assembly programs which were combined by the EvidentialGene tr2aacds pipeline following the methodology of (Visser, Wegrzyn, Steenkmap, Myburg, & Naidoo, 2015) with some modifications. To summarise, each dataset was *de novo* assembled using SOAPdenovo-Trans v.1.03 (Xie et al., 2014) and Velvet/Oases (Schulz, Zerbino, Vingron, & Birney, 2012; Zerbino & Birney, 2008) with k-mer lengths 23, 25, 31, 39, 47, 55, and 63 for both plus 71 for SOAPdenovo-Trans only, with *de novo* Trinity v.2.5.1 (Haas et al., 2013) assembly also being performed. We differ from (Visser et al., 2015) during Trinity assembly by not specifying "min_contig_length" and by providing the arguments "--min_kmer_cov 2" and "--SS_lib_type RF". *De novo* assemblies were concatenated and nucleotide sequences shorter than 350 bp were removed. Alongside *de novo* transcriptome assembly we also performed genome-guided assembly using Trinity and scallop v.0.10.2 (Shao & Kingsford, 2017) with RNA-seq read alignments in BAM format being generated by STAR v.2.5.4b (commit 5dbd58c) (Dobin et al., 2013) using the two-pass procedure to identify intron splice sites ("--twopassMode Basic"). Genome-guided Trinity was performed with "--genome_guided_max_intron 21000" to limit false positives, and scallop was run with defaults excepting the provided parameter "--library_type first". Genome-guided and *de novo* transcriptomes were concatenated into a single file and then a non-redundant transcriptome including alternative isoforms was created with the EvidentialGene tr2aacds pipeline which predicted transcript CDS regions as part of its operations. Outputs are organised into "okay" and "okalt" groups, the former containing representative isoforms of loci and the latter containing alternative isoforms. BUSCO was used to assess the completeness of each transcriptome with reference to databases of single-copy orthologs provided by BUSCO's authors which are as follows; *S. cerevisiae*: saccharomycetales_odb9, *C. elegans*: nematoda_odb9, *A. thaliana*: eudicotyledons_odb10, *M. musculus*: mammalia_odb9.

PASA and GGF gene model annotations

The PASA v.2.3.3 gene prediction pipeline was used to align transcripts to their respective species' genome using BLAT and GMAP and subsequently render gene model predictions. Parameters specified include the following: "--MIN_PERCENT_ALIGNED=75", "--MIN_AVG_PER_ID=95", "--NUM_BP_PERFECT_SPLICE_BOUNDARY=0", and "--stringent_alignment_overlap 30". After this, TransDecoder v.5.3.0 was used to extract open reading frames (ORFs) with TransDecoder.LongOrfs and the locations of these ORFs were mapped to genomic coordinates using the provided cdna_alignment_orf_to_genome_orf.pl utility script. Resulting GFF3 files were updated with two iterations of the PASA annotation comparison pipeline to refine CDS regions and add alternative isoforms that were not annotated by the initial prediction pipeline. Default parameters were used except that we specified "--stringent_alignment_overlap 30".

Prior to GGF, GMAP alignments of transcriptome-predicted and PASA-predicted CDS regions to their species' genome was performed with parameters "-f 2 -n 12 -x 50 -B 5 –max-intron-length-middle=500000 --max-intronlength-ends=500000". These parameters are specified by PASA during its GMAP alignment with differences being "-f 2" to produce GFF3 output and "-n 12" to align each CDS up to twelve times in the genome to capture increased copy numbers. The GGF software (commit 1fa2505) was run with these two GFF3 inputs, their corresponding FASTA CDS files, the relevant species' genome FASTA file, and the updated PASA GFF3 file. The output GFF3 file was merged into the PASA GFF3 file with a custom utility program to perform this operation (gff3_merge.py; https://github.com/zkstewart/Genome_analysis_scripts); henceforth, this merged file will be referred to as PASA+GGF.

Both PASA and PASA+GGF were processed to remove putative transposable elements, fragmented genes, and rRNAs erroneously annotated as genes with a custom pipeline (processing_pipeline.sh; https://github.com/zkstewart/Genome_analysis_scripts) which, as part of its operations, utilises HMMER v.3.1b2 (Eddy, 2011) and RNAmmer v.1.2 (Lagesen et al., 2007) to predict transposon-associated domains and rRNAs, respectively. After these processing steps we considered the gene catalogues to be finalised and downstream comparison to primary genome annotations could then take place.

Comparison to primary annotations

Primary genome annotations for each of the four species were obtained from Ensembl and The Arabidopsis Information Resource (TAIR), and the versions of these are as follows; *S. cerevisiae*: R64-1-1.94, *C. elegans*: WBcel235.94, *A. thaliana*: TAIR9, *M. musculus*: GRCm38.94. Genes present on organellar DNA were removed from primary, PASA, and PASA+GGF annotations prior to all analyses described henceforth as the annotation of mitochondrial and chloroplast genomes is outside the expected scope of GGF.

Comparison of PASA and PASA+GGF annotations to primary annotations was facilitated by mikado compare (commit 2b984c2; Venturini, Caim, Kaithakottil, Mapleson, & Swarbreck, 2018) which, using the primary annotations as the "reference" file, computes statistics for predicted loci including sensitivity (sensitivity = true positives / [true positives + false negatives]), precision (precision = true positives / [true positives + false positive]), and F1 score (harmonic average of sensitivity and precision) with novel loci also reported. In this context, true positives refer to genes present in the primary annotation and in PASA or PASA+GGF annotations, false negatives refer to features in the primary annotation that have no match in PASA or PASA+GGF annotations, and novel loci refer to those gene models present in the PASA or PASA+GGF annotations and not in the primary annotation (also referred to as "false positives" when computing statistics). A custom script (mikadocompare_novel_analyser.py; https://github.com/zkstewart/Genome_analysis_scripts) was created to gain further insight into the nature of these putative novel loci i.e., whether they were potentially true novel genes or if they were instead fragmentary genes or false positive annotations such as pseudogenes.

Homologous gene grouping and gene copy number assessment

To validate that GGF accomplishes its intended goal of annotating gene copies missed by PASA, we used OrthoFinder v.2.2.7 (commit 165808f) to predict groups of genes that include orthologs and paralogs (called "orthogroups" by OrthoFinder). Gene model CDS' translated as amino acids from primary, PASA, and PASA+GGF annotations were provided to OrthoFinder with default parameters except that MMseqs2 (Steinegger & Söding, 2017, p. 2) search was utilised. Orthogroup results of PASA and PASA+GGF were compared to the primary annotation to assess similarities and differences in copy number with a custom script (orthofinder_group_statistics.py; https://github.com/zkstewart/Various_scripts/tree/master/Orthofinder).

Results

Transcriptome metrics

Metrics obtained from *de novo* and genome-guided transcriptome assembly with EvidentialGene for each species are presented in Table 1. There is some variation in transcriptome assembly quality as seen in BUSCO completeness scores for *S. cerevisiae* and *C. elegans* (98.3% and 96.9 %, respectively) and lower scores for the remaining species (< 85 %). Quality can also be determined by comparing the number of "okay" EvidentialGene contigs (refer to the methods for a description of this term) to the actual number of coding genes annotated within the primary annotations as these should be comparable. The number of "okay" contigs assembled for *S. cerevisiae* (7,631) is comparable to the R64-1-1.94 annotation (6,600), as is the number of "okay" contigs for *C. elegans* (okay = 18,069, WBcel235.9 = 20,222) and for *A. thaliana* (okay = 25,674, TAIR9 = 27,379). There were some quality issues for *M. musculus*, with both datasets obtaining poor N50 values and incomparable numbers of "okay" contigs (7.8 Gbp = 170,118, 15 Gbp = 96,686) to the GRCm38.94 annotation (22,619 coding genes). It is probable that more depth of RNA-seq sequencing was required to assemble a high-quality mouse transcriptome; nonetheless, it may still be used to assess the performance of GGF. Since *ab initio* gene prediction was not involved in PASA or PASA+GGF annotations, these statistics represent the theoretical upper limits of our annotation completeness.

PASA gene models

PASA gene model prediction metrics are presented in Table 2. Variable quality of transcriptomes are propagated to the genome as expected, with BUSCO scores slightly lower than the upper limits imposed by the transcriptomes. We do note that the number of loci and BUSCO completeness scores in *S. cerevisiae* were reduced by the automated processing system of processing_pipeline.sh substantially more than occurred in other species (PASA's BUSCO completeness prior to curation = C:93.4%[S:91.4%,D:2.0%],F:2.6%,M:4.0%,n:1711; a 3 % completeness decrease). This is likely because the genome of this species has a higher gene density than the pipeline can handle. Otherwise, a notable observation is that PASA has improved CDS N50 values relative to transcriptomes for all genomes indicating that some transcripts were fragmented, but that the genes coding for these transcripts have been annotated without being similarly truncated.

GGF gene models

PASA+GGF gene model annotations are also presented in Table 2. The number of loci is increased in all annotations, with this change being modest in *S. cerevisiae* (0.5 % increase), moderate in *M. musculus* (1.1 % and 1.6 % increases in 7.8 Gbp and 15 Gbp datasets, respectively), and relatively high in *C. elegans* (2.4 % increase) and *A. thaliana* (3.3 % increase). The number of alternative isoforms rarely increases, with only 2 new isoforms in *A. thaliana* and 3 new isoforms in the *M. musculus* 15 Gbp dataset. This result is expected since GGF prevents new genes from overlapping existing genes by > 35 % of the new gene length. N50 values after GGF merge decrease slightly which indicates that these gene models are smaller on average than those annotated by PASA. Importantly, BUSCO scores increase for all species; a 0.1 % increase in completeness is seen for *S. cerevisiae*, with a 0.4 % increase in *M. musculus* 7.8 Gbp, a 0.6 % increase in *M. musculus* 15 Gbp, a 1.5 % increase in *C. elegans*, and a 2.4 % increase in *A. thaliana*.

Primary annotation comparison statistics

Statistics obtained from comparison of PASA and PASA+GGF annotations to the primary annotations of each species using mikado compare are presented in Table 3. Sensitivity and precision scores for *S. cerevisiae* are relatively high compared to other species. Comparing the values of PASA and PASA+GGF indicates that sensitivity has increased for all species (*S. cerevisiae*: 0.18 % increase, *M. musculus* 15 Gbp: 0.38 % increase, *M. musculus* 7.8 Gbp: 0.39 % increase, *A. thaliana*: 1.19 % increase, *C. elegans*: 1.24 % increase) with precision scores also increasing for most (*S. cerevisiae*: 0.17 % decrease, *M. musculus* 15 Gbp: 0.02 % decrease, *M. musculus* 7.8 Gbp: 0.05 % increase, *A. thaliana*: 0.18 % increase, *C. elegans*: 0.65 % increase), with these values combined resulting in an increased F1 score for all species. Use of GGF resulted in additional novel loci being discovered when compared to PASA alone (*S. cerevisiae*: +3 genes, *A. thaliana*: +61 genes, *C. elegans*: +66 genes, *M. musculus* 7.8 Gbp: +246 genes, *M. musculus* 15 Gbp: +419 genes).

Further assessment of novel loci using mikadocompare_novel_analyser.py provided limited insight into the nature of these novel genes (see Supplementary Table 1). Overall, most novel genes cannot be unambiguously defined as being true or false positives, with between 44.94 % and 85.92 % (average 72.93 % across each dataset) of the novel genes identified in an annotation being classified as “novel or flawed”. The next most populous category includes novel genes which overlap pseudogene predictions within primary annotations, which make up between 0 % and 38.99 % of the novel genes predicted (average 9.75 %); while GGF did predict some genes in this category, most were originally predicted by PASA (*S. cerevisiae*: PASA=0; PASA+GGF=0, *A. thaliana*: PASA=61;

PASA+GGF=68, *C. elegans*: PASA=320; PASA+GGF=347, *M. musculus* 7.8 Gbp: PASA=105; PASA+GGF=111, *M. musculus* 15 Gbp: PASA=255; PASA+GGF=264). Most other categories of genes are of relatively low abundance and the PASA+GGF annotation does not differ substantially from PASA.

OrthoFinder gene copy assessment

The results of OrthoFinder assessment of gene copy number within orthogroups as interpreted by orthofinder_group_statistics.py are presented in Supplementary Table 2. Key results include that, for all species, the number of orthogroups which have an identical number of members when compared to the primary annotation is increased in PASA+GGF relative to PASA (*S. cerevisiae*: PASA=4,759 out of 5,553 orthogroups; PASA+GGF=4,779 out of 5,553 (+20), *A. thaliana*: PASA=11,883 out of 20,778; PASA+GGF=12,329 out of 20,778 (+446), *C. elegans*: PASA=11,199 out of 18,651; PASA+GGF=11,559 out of 18,651 (+360), *M. musculus* 7.8 Gbp: PASA=8,557 out of 49,542; PASA+GGF=8,676 out of 49,542 (+119), *M. musculus* 15 Gbp: PASA=9,394 out of 47,739; PASA+GGF=9,522 out of 47,739 (+128). This equates to an increase in the number of PASA+GGF orthogroups with equivalent membership to that of the primary annotation when compared to PASA by 0.42 % in *S. cerevisiae*, 3.75 % in *A. thaliana*, 3.21 % in *C. elegans*, 1.39 % in *M. musculus* 7.8 Gbp, and 1.36 % in *M. musculus* 15 Gbp. A similar increase is not seen in the number of orthogroups which contain memberships exceeding the amount of the primary annotation, with such groups decreasing by 0.14 % in *S. cerevisiae* and increasing by 0.28 % in *A. thaliana*, 0.18 % in *C. elegans*, 0.04 % in *M. musculus* 7.8 Gbp, and 0.06 % in *M. musculus* 15 Gbp.

To further understand what changes GGF had introduced to orthogroup membership, we looked at a single group with the most increased membership count in PASA+GGF relative to PASA for each species (predicted automatically by orthofinder_group_statistics.py). For *A. thaliana* we found that, for an orthogroup with three additional identical members predicted by GGF, two of these models had perfect CDS matches in the TAIR9 annotation, with the third overlapping an existing gene but with divergence at 5’ and 3’ ends. BLASTX search to NCBI’s non-redundant (nr) database performed 22nd November 2018 indicated that these genes were “*UDP-3-O-acyl N-acetylglycosamine deacetylase family protein*” and were full-length (i.e., full end-to-end alignments occurred for all three gene models). Three additional members in PASA+GGF relative to PASA were discovered for a *C. elegans* orthogroup, with none of these being represented by models in the primary annotations. Significant matches to sequences annotated as “*hypothetical protein W03G1.3*” (E-value = 0) were found, with one of the predicted models’ CDS being identical to this database sequence and the two remaining sequences showing an internal 9 bp deletion at the 5’ end of the gene. Five additional sequences were found in a *M. musculus* 7.8 Gbp orthogroup which were not represented by primary annotations, but they did obtain significant matches (E-value = from $2e^{-68}$ to $6e^{-70}$) to a sequence annotated as “*mCG4448*” with some small 5’ and 3’ extensions relative to this database sequence. There were thirteen additional sequences within an orthogroup identified for *M. musculus* 15 Gbp, with 6/13 of these sequence being represented by a primary annotation with some small extensions to the 5’ and 3’ ends in the GGF annotated sequences relative to the primary sequences; 3/10 were not represented by a primary model, 3/10 overlapped existing models but had large truncations to the 5’ and/or 3’ ends, and 1/10 of these sequences shared two of its seven exons with a primary annotation and had similar overall coding sequence with 5’ extension but the remaining five exons were not annotated by the primary annotation. From the thirteen total additional sequences, eight were near-identical (i.e., minor substitutions), were 783 bp in length, and obtained significant matches (E-value = 0) to “*PREDICTED: uncharacterized protein Gm3173 isoform X1*”, with the remaining five sequences being only 390 bp in length, also near-identical, and obtained significant matches (E-value = from $2e^{-89}$ to $5e^{-90}$) to the following sequences: “*PREDICTED: uncharacterized protein Gm3005 isoform X3*”, “*alpha6-takusan-like*”, “*predicted gene 3264*”, and “*mCG129267, isoform CRA_a*”. Finally, *S. cerevisiae* had an orthogroup with two additional GGF-annotated members which were represented by primary annotation models but had large truncations to the annotated 3’ end with one of these also having a small extension to its 5’ end. Both sequences obtained significant matches (E-value = 0) to “*Y' element ATP-dependent helicase protein 1 copy 7*”.

Discussion

Automated gene annotation efforts are an essential first step towards producing a high-quality genome annotation. Software used to perform this currently face several limitations which lead to false positives, false negatives, and gene models that are not entirely accurate, with these outcomes being caused by the complexity involved in real biological data (Fawal, Li, Mathé, & Dunand, 2014). While manual annotation of genes is necessary to ensure correct gene model annotations, the development of tools to address current weaknesses in automated approaches is necessary. As our results show, GGF is a tool that can improve a gene annotation via the annotation of additional gene copies and by the annotation of gene models originally missed by PASA (or, potentially, any other annotation program). Although not detailed here in-depth, program execution on the assessed datasets took between 1.5 minutes for *S. cerevisiae* and 2.75 hours for *M. musculus* 15 Gbp on a computer using a single core with peak memory consumption being approximately 7 GB for *M. musculus* 15 Gbp (larger datasets result in more memory consumption).

Quality of GGF annotations

Our results indicate that PASA+GGF annotations represent an improvement over the original PASA annotation, with F1 scores increasing in all species. In all datasets except *S. cerevisiae* and *M. musculus* 15 Gbp, this includes increases to both sensitivity and precision. This indicates that GGF-annotated gene models tend to correspond well to gene models that are present in the primary annotations but were missed by PASA originally.

While these results provide a positive indication of GGF's use in genome annotation projects, it is important to acknowledge certain limitations that might affect GGF's outputs. Specifically, GGF solely relies upon GMAP's ability to convert transcript alignments into GFF3-formatted gene predictions and enacts a series of curation checks to remove alignments that are suspect or do not accurately reflect the transcript originally used for alignment. This approach is greatly simplified when compared to PASA which, like most other gene annotation programs, utilises sophisticated algorithms to predict genes. PASA relies upon a concept referred to as "maximal alignment assembly" (Haas et al., 2003) which is capable of handling fragmented transcripts and expressed sequence tags (ESTs). The benefit of this is shown in the present study, where the transcriptomes for both *M. musculus* datasets appear to be highly fragmented based upon their poor N50 values, whereas PASA annotation N50 values are substantially improved. This highlights an important limitation to the use of GGF, namely being that the quality of the input transcriptome will determine the quality of GGF's results in a very direct manner. It is recommended that a thorough approach like that detailed in this study is performed when assembling a transcriptome for use with GGF, and that this assembly be assessed to ensure its quality. Although we did not find readily apparent reason to suggest that GGF should not be used in our *M. musculus* datasets, caution should still be applied in cases similar to this.

Annotation of new and novel genes

Although we did not originally intend GGF to identify new gene predictions (i.e., not simply additional copies of genes) that PASA missed, BUSCO scoring shows that adding the results of GGF to an annotation produced by PASA may result in a more complete genome annotation, with completeness scores increasing by 0.1 % to 2.4 % in the four species assessed.

GGF may also serve to assist in discovery efforts for novel genes not currently part of primary genome annotations. While it is difficult to validate whether most of the novel genes predicted by GGF have been annotated correctly and in full-length or if they are fragmented or false positive predictions, we don’t find reason to believe that GGF has an error rate that compares poorly to that of PASA as discussed previously. Thus, the main sources of error would involve fragmented transcripts producing fragmented gene models or the incorrect annotation of non-coding RNAs (ncRNAs) and pseudogenes, the former issue requiring quality inspection of the transcription as mentioned above and these latter two problems also being faced by PASA and similar programs. Preventing spurious annotation of ncRNAs would require the detection and removal of their corresponding transcripts from the transcriptome prior to use of GGF, a task that may be assisted by a program such as FEELnc (Wucher et al., 2017). Due to GGF’s goal of annotating many copies of genes rather than just the single best, incorrect annotation of pseudogenes would appear to represent a specific point of weakness for the program. However, the requirement that the annotated gene model must contain more than one exon reduces the likelihood of annotating processed pseudogenes, and the requirement that the model be similar in overall length and optimal alignment length to the original transcript means that pseudogenes which are no longer transcribed or translated should, theoretically, accumulate premature stop codons or frameshift mutations which prevent these requirements from being fulfilled. Our results show that this system is not perfect and that PASA’s system is similarly imperfect, and as such downstream processing would be required to identify and annotate these models correctly.

Annotation of additional gene copies

OrthoFinder results show that the PASA+GGF methodology results in gene families being more completely annotated relative to PASA alone, which was measured by there being more orthogroups that contain the same number of members as the primary annotations. An underlying assumption to this is that the primary annotation has all gene copies correctly annotated. While this assumption will not always be met as these primary annotations continue to receive improvements, it should prove true for most gene families since the four species were chosen due to how well-established and developed their genome annotations are.

Correct annotation of all gene copies is essential for studies of the evolutionary relationship of gene families, as this will directly impact any conclusions drawn. This study attempted to provide some examples of these impacts by assessing gene families which saw increased membership after GGF annotation. In *A. thaliana*, we found that at least two of the three additional gene copies were directly supported by the primary annotation, with a third overlapping the coordinates of a similar gene but with some differences. Beyond providing some assurance of the quality of GGF's models, it is of interest to note that all three of these sequences are 100% identical which is unlikely to occur by chance if the third model (which only partially overlaps an existing gene model) is incorrectly annotated. Perfectly conserved gene copies may be of biological importance for reasons relating to genetic redundancy (Gu et al., 2003; Nowak, Boerlijst, Cooke, & Smith, 1997) or increased dosage effects (Kahlem et al., 2004), with an alternative explanation being that these genes might instead point towards recent duplication, a phenomenon suggested to provide the raw genetic material necessary for adaptive change (Farslow et al., 2015).

In other species, the additional copies annotated by GGF are commonly unsupported by primary annotations but typically have sequence matches within the NCBI nr database albeit with some extension or truncation in certain cases. It is difficult to ascertain the likelihood that these models are biologically real or are mostly correct (i.e., that they have roughly accurate start and stop sites for the CDS region with all internal exons represented), however, due to these sequences all being highly similar, a consequence of GGF's default algorithmic requirements, we believe that there is sufficient evidence to suggest that most of these predictions are likely to correspond to genomic regions that are undergoing purifying selection or have been recently duplicated.

Regardless, it is easy to imagine how improving the annotation of gene copies as highlighted in this study might affect interpretation of certain gene families. For example, the *mCG4448* gene of *M. musculus* 7.8 Gb increases from being a single-copy gene to being present in six copies. These genes are present in endogenous retrovirus-K (ERVK) long terminal repeat (LTR) masked DNA sequence, with the ERVK family of LTRs known to be involved in enabling novel gene transcription (Veselovska et al., 2015); while we cannot be sure what the biological relevance of these gene features are, their similarity in coding sequence suggests that these are genuine genes that are being conserved or that these retrotransposons have been recently active. Finally, the PASA+GGF annotation results for *M. musculus* 15 Gbp would appear to add at least three additional copies of the uncharacterised *Gm3173* gene to a family that, according to OrthoFinder, contains thirty-eight current members in the primary annotation.

Conclusions

This study demonstrates that GGF is a tool which may be used in the context of automatic genome annotation to improve the completeness of an existing genome annotation, both by annotation of new gene models and through the annotation of additional gene copies that may be missed by PASA or other programs which use a "best hit" approach to sequence alignment. Recommended program operation consisting of transcript and existing gene model CDS alignment is demonstrated here, with the alignment of existing gene models being especially useful if high-quality *ab initio* prediction has occurred since these sequences may not be represented in a transcriptome. There are certain limitations associated with the use of GGF, namely being that the quality of output GGF models will directly correspond to the quality of input sequences. Through the accurate annotation of gene copies, GGF can ensure more correct interpretation for studies of gene families in a variety of species.

Acknowledgements

Computational resources and services used in this work were provided by the HPC and Research Support Group, Queensland University of Technology (QUT), Brisbane, Australia.

References

Alekseyenko, A. V., & Lee, C. J. (2007). Nested Containment List (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases. *Bioinformatics*, *23*(11), 1386–1393. https://doi.org/10.1093/bioinformatics/btl647

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Armstrong, E. E., Taylor, R. W., Prost, S., Blinston, P., Meer, E. van der, Madzikanda, H., … Petrov, D. (2017). Entering the era of conservation genomics: Cost-effective assembly of the African wild dog genome using linked long reads. *BioRxiv*, 195180. https://doi.org/10.1101/195180

Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., … Gan, H. M. (2017). De novo genome assembly and annotation of Australia’s largest freshwater fish, the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read. *GigaScience*, *6*(8). https://doi.org/10.1093/gigascience/gix063

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., … Lantz, H. (2018). Ten steps to get started in genome assembly and annotation. *F1000Research*, *7*, 148. https://doi.org/10.12688/f1000research.13598.1

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology, 7*(10). https://doi.org/10.1371/journal.pcbi.1002195

Farrar, M. (2007). Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics, 23*(2), 156–161. https://doi.org/10.1093/bioinformatics/btl582

Farslow, J. C., Lipinski, K. J., Packard, L. B., Edgley, M. L., Taylor, J., Flibotte, S., … Bergthorsson, U. (2015). Rapid increase in frequency of gene copy-number variants during experimental evolution in *Caenorhabditis elegans*. *BMC Genomics, 16*(1), 1044. https://doi.org/10.1186/s12864-015-2253-2

Fawal, N., Li, Q., Mathé, C., & Dunand, C. (2014). Automatic multigenic family annotation: risks and solutions. *Trends in Genetics: TIG, 30*(8), 323–325. https://doi.org/10.1016/j.tig.2014.06.004

Gilbert, D. (2016). Accurate & complete gene construction with EvidentialGene. *F1000Research, 5*. https://doi.org/10.7490/f1000research.1112467.1

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., & Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature, 421*(6918), 63–66. https://doi.org/10.1038/nature01198

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., … White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research, 31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols, 8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., … Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to

Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., & Shiu, S.-H. (2010). sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics (Oxford, England)*, *26*(3), 399–400. https://doi.org/10.1093/bioinformatics/btp688

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F.-A., Swinkels, W., … Henkel, C. V. (2017). Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports*, *7*(1), 7213. https://doi.org/10.1038/s41598-017-07650-6

Kahlem, P., Sultan, M., Herwig, R., Steinfath, M., Balzereit, D., Eppens, B., … Yaspo, M.-L. (2004). Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of Down syndrome. *Genome Research*, *14*(7), 1258–1267. https://doi.org/10.1101/gr.1951304

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*, *19*(1), 189. https://doi.org/10.1186/s12859-018-2203-5

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., & Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*, *25*(9), 404–413. https://doi.org/10.1016/j.tig.2009.07.006

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., … Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science*, *360*(6393), eaar6343. https://doi.org/10.1126/science.aar6343

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/nar/gkm160

Macrander, J., Broe, M., & Daly, M. (2015). Multi-copy venom genes hidden in *de novo* transcriptome assemblies, a cautionary tale with the snakelocks sea anemone *Anemonia sulcata* (Pennant, 1977). *Toxicon*, *108*, 184–188. https://doi.org/10.1016/j.toxicon.2015.09.038

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*(16), 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

Margres, M. J., Bigelow, A. T., Lemmon, E. M., Lemmon, A. R., & Rokyta, D. R. (2017). Selection to increase expression, not sequence diversity, precedes gene family origin and expansion in rattlesnake venom. *Genetics*, *206*(3), 1569–1580. https://doi.org/10.1534/genetics.117.202655

Matthews, B. J., Dudchenko, O., Kingan, S. B., Koren, S., Antoshechkin, I., Crawford, J. E., … Vosshall, L. B. (2018). Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature*, *563*(7732), 501–507. https://doi.org/10.1038/s41586-018-0692-z

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., … Ecker, J. R. (2018). High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications, 9*. https://doi.org/10.1038/s41467-018-03016-2

Nallu, S., Hill, J. A., Don, K., Sahagun, C., Zhang, W., Meslin, C., … Kronforst, M. R. (2018). The molecular genetic basis of herbivory between butterflies and their host plants. *Nature Ecology & Evolution*, *2*(9), 1418–1427. https://doi.org/10.1038/s41559-018-0629-9

Nielsen, J. C., Senne de Oliveira Lino, F., Rasmussen, T. G., Thykær, J., Workman, C. T., & Basso, T. O. (2017). Industrial antifoam agents impair ethanol fermentation and induce stress responses in yeast cells. *Applied Microbiology and Biotechnology*, *101*(22), 8237–8248. https://doi.org/10.1007/s00253-017-8548-2

Nowak, M. A., Boerlijst, M. C., Cooke, J., & Smith, J. M. (1997). Evolution of genetic redundancy. *Nature*, *388*(6638), 167–171. https://doi.org/10.1038/40618

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., … Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, *12*(8), 780–786. https://doi.org/10.1038/nmeth.3454

Salazar, A. N., Gorter de Vries, A. R., van den Broek, M., Wijsman, M., de la Torre Cortés, P., Brickwedde, A., … Abeel, T. (2017). Nanopore sequencing enables near-complete de novo assembly of Saccharomyces cerevisiae reference strain CEN.PK113-7D. *FEMS Yeast Research*, *17*(7). https://doi.org/10.1093/femsyr/fox074

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., … Kim, C. (2016). *De novo* assembly and phasing of a Korean human genome. *Nature*, *538*(7624), 243–247. https://doi.org/10.1038/nature20098

Shao, M., & Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*, *35*(12), 1167–1169. https://doi.org/10.1038/nbt.4020

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., & Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research*, *34*(14), 3955–3967. https://doi.org/10.1093/nar/gkl556

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Solares, E. A., Chakraborty, M., Miller, D. E., Kalsow, S., Hall, K., Perera, A. G., … Hawley, R. S. (2018). Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3: Genes, Genomes, Genetics, 8*(10), 3143–3154. https://doi.org/10.1534/g3.118.200162

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research, 32*(suppl_2), W309–W312. https://doi.org/10.1093/nar/gkh379

Steinegger, M., & Söding, J. (2017, October 16). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets [Comments and Opinion]. https://doi.org/10.1038/nbt.3988

Sutton, J. T., Helmkampf, M., Steiner, C. C., Bellinger, M. R., Korlach, J., Hall, R., … Ryder, O. A. (2018). A high-quality, long-read *de novo* genome assembly to aid conservation of Hawaii’s last remaining crow species. *Genes, 9*(8). https://doi.org/10.3390/genes9080393

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols, 7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

Tyson, J. R., O’Neil, N. J., Jain, M., Olsen, H. E., Hieter, P., & Snutch, T. P. (2018). MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Research, 28*(2), 266–274. https://doi.org/10.1101/gr.221184.117

Ustianenko, D., Weyn-Vanhentenryck, S. M., & Zhang, C. (2017). Microexons: discovery, regulation, and function. *Wiley Interdisciplinary Reviews. RNA, 8*(4). https://doi.org/10.1002/wrna.1418

Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annual Review of Genetics*, *19*, 253–272. https://doi.org/10.1146/annurev.ge.19.120185.001345

Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L., & Swarbreck, D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*, *7*(8). https://doi.org/10.1093/gigascience/giy093

Veselovska, L., Smallwood, S. A., Saadeh, H., Stewart, K. R., Krueger, F., Maupetit-Méhouas, S., … Kelsey, G. (2015). Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome Biology*, *16*(1), 209. https://doi.org/10.1186/s13059-015-0769-z

Visser, E. A., Wegrzyn, J. L., Steenkmap, E. T., Myburg, A. A., & Naidoo, S. (2015). Combined de novo and genome guided assembly and annotation of the Pinus patula juvenile shoot transcriptome. *BMC Genomics*, *16*, 1057. https://doi.org/10.1186/s12864-015-2277-7

Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics*, *2*(4), 216–221. https://doi.org/10.1016/S1672-0229(04)02028-5

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., … Derrien, T. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, *45*(8), e57–e57. https://doi.org/10.1093/nar/gkw1306

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., … Wang, J. (2014). SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Data accessibility statement

The program described is available from Github at https://github.com/zkstewart/Genome_analysis_scripts/tree/master/ggf and is provided under a GNU General Public License v3.0 agreement.

Competing interests statement

Declarations of interest: none.

Author contributions

ZKS conceived of and designed the software. ZKS and PJP contributed to writing the manuscript; both authors approve of the final manuscript.

Figures and Tables

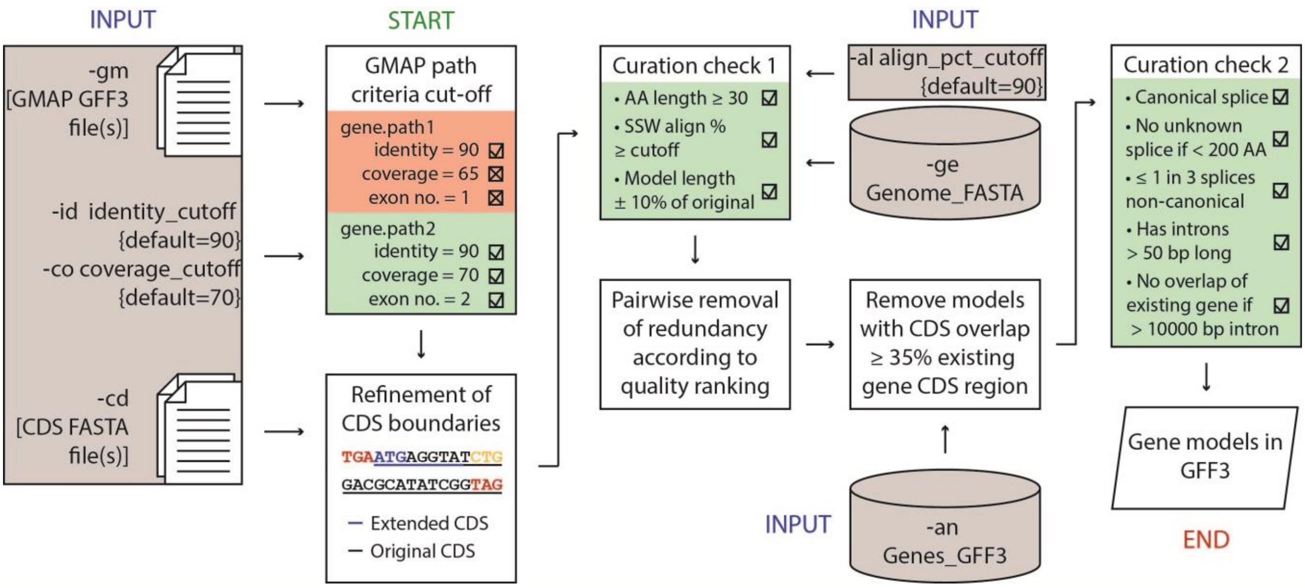

Figure 1. Simplified overview of the programmatic process underlying gene model prediction with gmap_gene_find (GGF). As depicted, GGF makes only minor modifications to sequence alignments which are subject to a variety of curation systems to remove low-quality alignments and return alignments likely to represent real genes. The "GMAP path criteria cut-off" box uses red colouration to show an example of a GMAP alignment ("path") which has been rejected due to the coverage value and exon number (no.) not meeting minimum requirements; green coloured sections in this and other boxes show paths that have been accepted. In "Refinement of CDS boundaries" box, blue colouration is used to depict a canonical ATG (methionine) start codon which is upstream of the yellow coloured non-canonical CTG (leucine) start codon, with red coloured codons indicating stop codons. Abbreviations; AA: amino acid, CDS: coding DNA sequence, GFF3: General Feature Format 3, GMAP: Genomic Mapping and Alignment Program, pct: percent, SSW: Striped Smith-Waterman (Farrar, 2007).

Table 1. Transcriptome assembly results from EvidentialGene combination of multiple *de novo* and genome-guided assemblies. Two datasets are presented for *Mus musculus* where datasets with a different amount of RNA-seq coverage were used. “Okay” contigs should refer to representative isoforms of loci and “okalt” should refer to alternative isoforms. N50 statistics for predicted coding DNA sequence (CDS) are provided; sequence length is measured in nucleotide bases. BUSCO short summary notation indicates the proportion of complete orthologs (C) that were identified within the transcriptome when compared to a database of genes present as single-copies in related species, with this value being broken down into those present in single copy (S) and in duplicate (D); fragmented (F) and missing (M) genes are also indicated, with the number of genes present in the BUSCO database (n) also depicted.

Species (dataset)	"Okay" contig number	"Okalt" contig number	N50 of CDS prediction	BUSCO short summary
Saccharomyces cerevisiae	7,631	7,287	1,692	C:98.3%[S:76.7%,D:21.6%],F:1.6%,M:0.1%,n:1711
Caenorhabditis elegans	18,069	24,770	1,329	C:96.9%[S:62.9%,D:34.0%],F:0.7%,M:2.4%,n:982
Arabidopsis thaliana	25,674	29,527	1,185	C:79.2%[S:66.1%,D:13.1%],F:7.6%,M:13.2%,n:2121
Mus musculus (7.8 Gbp)	170,118	47,347	441	C:72.9%[S:57.3%,D:15.6%],F:5.8%,M:21.3%,n:4104
Mus musculus (15 Gbp)	96,686	93,256	897	C:83.2%[S:51.6%,D:31.6%],F:4.2%,M:12.6%,n:4104

Table 2. Genome annotation results from PASA (P) and PASA combined with GGF (P+G). The number of loci are presented including the number of alternative isoforms. Coding DNA sequence (CDS) N50 values are presented; these values were obtained from all CDS including alternative isoforms. BUSCO short summary notation indicates the proportion of complete orthologs (C) that were identified within the genome when compared to a database of genes present as single-copies in related species, with this value being broken down into those present in single copy (S) and in duplicate (D); fragmented (F) and missing (M) genes are also indicated, with the number of genes present in the BUSCO database (n) also depicted.

	Number of loci		Number of alternative isoforms		N50 of CDS prediction		BUSCO short summary	
Species (dataset)	P	P+G	P	P+G	P	P+G	P	P+G
Saccharomyces cerevisiae	5,923	5,953	37	37	1,857	1,857	C:90.4%[S:88.5%, D:1.9%],F:2.2%, M:7.4%,n:1711	C:90.5%[S:88.6%, D:1.9%],F:2.2%, M:7.3%,n:1711
Caenorhabditis elegans	22,089	22,616	1,541	1,541	1,512	1,509	C:93.1%[S:77.3%, D:15.8%],F:2.4%, M:4.5%,n:982	C:94.6%[S:78.8%, D:15.8%],F:2.3%, M:3.1%,n:982
Arabidopsis thaliana	24,279	25,075	2,768	2,770	1,389	1,383	C:76.1%[S:67.2%, D:8.9%],F:7.5%, M:16.4%,n:2121	C:78.5%[S:69.7%, D:8.8%],F:7.6%, M:13.9%,n:2121
Mus musculus (7.8 Gbp)	56,648	57,312	1,784	1,784	1,272	1,269	C:72.0%[S:57.7%, D:14.3%],F:6.0%, M:22.0%,n:4104	C:72.4%[S:58.0%, D:14.4%],F:6.0%, M:21.6%,n:4104
Mus musculus (15 Gbp)	60,280	61,243	2,979	2,982	1,449	1,443	C:81.2%[S:57.3%, D:23.9%],F:5.4%, M:13.4%,n:4104	C:81.8%[S:57.9%, D:23.9%],F:5.4%, M:12.8%,n:4104

Table 3. Metrics including sensitivity, precision, and F1 score (harmonic average of sensitivity and precision values) were generated by using mikado compare to perform comparison of annotations generated by PASA (P) and PASA combined with GGF (P+G) to the primary annotations of the listed species. Presented are those statistics obtained for the "Gene level (100% base F1)" category of comparison; sensitivity refers to genes in the primary annotation whose coding DNA sequence has a perfect match in the P or P+G annotations (i.e., are true positives) and is calculated as "sensitivity = true positives / (true positives + false negatives)" wherein false negatives refer to features in the primary annotation that have no match in either P or P+G annotations; precision is calculated by "precision = true positives / (true positives + false positives)" wherein false positives refer to features in the P or P+G annotation which have no match in the primary annotation and are referred to as novel loci.

	Matching loci						Unmatched loci	
	P			P+G			P	P+G
Species (dataset)	Sensitivity	Precision	F1	Sensitivity	Precision	F1	Number of novel loci	Number of novel loci
Saccharomyces cerevisiae	65.66	73.88	69.53	65.84	73.71	69.55	421	424
Caenorhabditis elegans	21.08	19.76	20.4	22.32	20.41	21.32	824	890
Arabidopsis thaliana	31.07	34.76	32.81	32.26	34.94	33.55	1151	1212
Mus musculus (7.8 Gbp)	21.44	8.79	12.47	21.83	8.84	12.59	6515	6761
Mus musculus (15 Gbp)	25.96	10.3	14.75	26.34	10.28	14.78	9224	9643