Accurate copy number annotation for genomes using gmap_gene_find (GGF)

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

Zachary K. Stewart[1*], Peter J. Prentis[1,2]

[1]School of Earth, Environmental and Biological Sciences, 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

-

Keywords: genome annotation, gene models, gene prediction,

Introduction

Recent advancements in DNA sequencing with long-read technologies has catalysed an upsurge in the amount of new genome assembly projects being undertaken. This is especially true of plants and animals which commonly have genomes exceeding 1 Gb in size (Gregory et al., 2007) whose repeat content has proven problematic for short-read technologies (Myers, 2016). While many such projects may be considered “reassemblies” for model species with existing genomes e.g., X, Y, Z, the relative ease with which long reads can be used to assemble novel and non-model species has enabled researchers to ask fundamental questions in a range of species that receive less scientific attention e.g., A, B, C. The researchers working on these new species' genomes may themselves be new to the field of genome assembly and will face many challenges in the downstream handling of genomes.

Once an assembly has been obtained the annotation process then takes place (Dominguez Del Angel et al., 2018). The choices and decisions made here will have important impacts on all subsequent analysis. The first step in this process involves the annotation of repeat sequences, which consists of simple repeats e.g., satellite DNA, or transposable elements e.g., transposons and retrotransposons, which in most eukaryotes makes up a major (or majority) component of total genome length with these sequences contributing to genomic evolution in various ways (Shapiro & von Sternberg, 2005). Although repeats are of great scientific interest due to these contributions, for many genome projects repeat annotation primarily serves as a precursor to the intended project outcome and perhaps most difficult step of the annotation process, namely being the annotation of gene models.

New genome projects commonly utilise a variety of different tools to annotate gene models. These bioinformatic programs, despite their differences, can be broadly categorised according to their mechanism of action; specifically, these refer to homology-based prediction, *ab initio* prediction, or transcript-based prediction. Homology-based prediction relies upon sequence alignment of known genes against the genome of interest, a modern example of this process being evinced by 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 in cases where a genome is being reassembled or a close evolutionary relative's genome exists but can be expected to perform worse when this is not true. Moreover, these methods may not identify many or most species-specific genes – that is, genes not found in other species – and these genes may make up 10-20% of a species' total gene repertoire (Khalturin, Hemmrich, Fraune, Augustin, & Bosch, 2009). *Ab initio* prediction, also known as intrinsic prediction, is a means of annotating gene models directly from genomic sequence by building a comprehensive statistical model that can differentiate coding from non-coding sequence (Wang, Chen, & Li, 2004).

Many of the *ab initio* prediction programs used over a decade ago (Wang et al., 2004) are still in common use, albeit with some updates; notable examples include AUGUSTUS (Stanke, Steinkamp, Waack, & Morgenstern, 2004) and GlimmerHMM (Majoros, Pertea, & Salzberg, 2004). Although the sensitivity of these programs can approach 100%, the specificity and accuracy of these programs can suffer in achieving such a goal (Yandell & Ence, 2012). For many newly sequenced species genomes, RNA sequencing data offers the greatest potential for accurately predicting gene models while also capturing species-specific genes (Yandell & Ence, 2012). Although long-read RNA sequencing is also revolutionising the prediction of gene models, due to resource limitations for many species there is still a reliance upon short RNA-seq reads. Aligning RNA-seq reads to the genome and directly producing gene models from this can be accomplished with programs like TopHat and Cufflinks (Trapnell et al., 2012); however, depending on genome attributes such as the 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 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 venom toxin genes, wherein genes with little to no sequence divergence may be present at extremely high copy numbers; an extreme example is the myotoxin gene of certain viper species which may be present in up to 47 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 weakness of PASA. GGF accomplishes this by using a complementary approach to PASA which also relies upon GMAP alignment of transcripts to the genome but with modified GMAP parameters to capture gene models present at higher copy numbers. To demonstrate the program’s use and assess its performance a study involving 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 a variety of eukaryotes with varying genome sizes and, importantly, have existing high-quality annotations. For this study, we used publicly available RNA-seq data to annotate the gene models of these organisms using PASA alone and PASA with GGF’s results merged into the annotation and compared these gene models to official annotations to assess the merits and/or drawbacks of GGF’s use in a genome project. The code for GGF is available from https://github.com/zkstewart/Genome_analysis_scripts/tree/master/ggf with XX licence.

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.

There are four inputs required for program operation. Prior to GMAP alignment, a FASTA file consisting solely of gene 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. As described in this study we align transcript CDS as well as coding sequences predicted by PASA; gene models predicted by *ab initio* software would also be recommended as transcripts for these models may not be represented in an assembled transcriptome. The first major input are alignment files produced by GMAP alignment of CDS form the basis of gene prediction, and it is expected that these files were produced with GFF3 (General Feature Format 3) output format specified by providing GMAP the "-f 2" argument. Secondly, the CDS FASTA file used for GMAP alignment is also a required input; multiple GMAP files can be provided so long as their partner CDS files are also provided in the same order on the command-line. Thirdly, the genome sequence in which gene models are being predicted is required. Finally, GGF also expects that the genome has an existing gene annotation in GFF3 format such as from PASA or any other annotation software. Although default parameters are advised, users of GGF can optionally modify three parameters to alter program behaviour. GMAP provides indication of alignment coverage i.e., what proportion of the query sequence is aligned against the genome, and the identity of this alignment i.e., the proportion of aligned positions that share identical nucleotide sequence; the first two optional parameters, coverageCutoff and identityCutoff, control which GMAP alignment "paths" (this term will henceforth be used to refer to GMAP alignments) will be considered in downstream analyses as potential gene models by enforcing minimum coverage and identity values. The third parameter, alignPctCutoff, is used when performing Striped Smith-Waterman (SSW; Farrar, 2007) of a potential gene model against the original CDS from which GMAP’s path was derived (both nucleotide sequences are translated to protein sequences first), and this parameter specifies the minimum proportion of the potential gene model which must optimally align against the original CDS (optimal alignment refers to the “best” start and end positions of an alignment).

A simplified overview of GGF’s internal operations are depicted in Figure 1. Once the necessary inputs have been provided as command-line arguments, GGF parses the input GMAP file(s) and extracts paths that meet minimum identity and coverage cut-off and which also contain two or more exons. Strict identity and coverage cut-offs are important for limiting discovery to only very similar gene copies which reduces the risk of annotating fragmented genes. Additionally, single-exon genes are excluded from consideration as it is difficult to distinguish these from processed pseudogenes which result from insertion of processed mRNAs into the genome sequence (Vanin, 1985).

Next, CDS boundaries are refined by extending the sequence by 100 bp at the 5’ and 3’ ends of the gene model to assist in finding appropriate start and stop codons for the longest ORF in the region. During this process, any exons that do not contain CDS are removed from the model, and if the model becomes a single-exon gene at this stage it is not considered further. 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, it may be extended to this codon. This is done because it was observed that the CDS regions predicted by programs like EvidentialGene or TransDecoder may often predict a non-canonical start codon when an in-frame ATG is present upstream of the genomic sequence; it is assumed that, in most cases, relatively short extensions to correct a non-canonical start with an ATG should be biologically correct. An additional benefit to this procedure is detailed below during length comparison.

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 whose amino acid (AA) translation is less than 30 AA in length will be removed from consideration in the interest of reducing false positive rates; programs specifically designed for predicting small open reading frames (sORFs) should be used instead, such as sORF finder (Hanada et al., 2010). Additionally, the sequence is compared to the original CDS used for GMAP alignment using SSW alignment as described previously to ensure that the potential gene model is highly similar to the original sequence, and the length of the potential gene model must be of a similar overall length to the original sequence i.e., ±10 % of the nucleotide length. A benefit of this is that sequences which, during CDS extension, were able to be extended and become > 10 % longer than the original CDS will be removed from consideration. Extension of an exon beyond the originally predicted boundaries should not be possible for any considerable length of nucleotides since we assume that these nucleotides represent non-coding sequence which are not subject to selection to eliminate in-frame stop codons. Thus, this would indicate that the genomic location being assessed only represents a fragment of a gene, and that the input CDS may itself be fragmented.

Potential gene models are then compared to the existing annotation GFF3 and any models which have greater than 35 % of their length overlapping previously annotated genes will be removed. The assumption made here is that the existing annotation is of high quality, and it additionally reduces the chance of annotating trans-spliced transcripts as legitimate gene models.

At this stage there might be some redundancy in the gene models if multiple highly similar transcripts exist or multiple input GMAP GFF3 files were provided. All potential gene models are compared to each other in a pairwise fashion and any that overlap are directly compared to pick the highest quality model, wherein quality is assessed by a ranking system of the certain criteria. The first check involves comparison of gene models whereby if only one gene model contains a microexon (exon < 30 bp in length), it is removed from consideration. Although microexons are a biologically valid occurrence (Ustianenko, Weyn-Vanhentenryck, & Zhang, 2017) the occurrence of these features in only one of the sequence alignments may indicate that sequence alignment is attempting to predict an exon in non-coding sequence and has truncated the exon in an attempt to avoid in-frame stop codons. At the second step, the longest gene model is selected as this ensures that, in cases where a fragmented transcript and full-length transcript align to the same position, we will annotate the full-length feature; a valid alternative rationale is also that longer sequences are less likely to identified by chance in non-coding sequence. Thirdly, the gene model with a higher proportion of canonical splice sites (i.e., GT-AG) or, if equivalent, a higher proportion of non-canonical and rare splice sites (i.e., GC-AG, AT-AC) relative to unknown splices (i.e., anything not mentioned previously) is selected. Finally, the gene model with the highest minimum exon length is selected for reasons similar to that described during the first check.

Non-redundant gene models are then curated by assessing their splice sites with certain criteria which includes the following: models which entirely lack canonical splices are not considered further; models with CDS shorter than 200 AA which contain unknown splices (i.e., not any splices defined above) are not considered further; if greater than 33 % of a model's splice sites are unknown it is not considered further. These checks are guided by the principle that real gene models typically evince canonical splices with unknown splices being exceptionally rare (Sheth et al., 2006). Compared to many other gene annotation programs, however, GGF does take a somewhat relaxed approach to splice site rules since we have observed in some non-model species that PASA will not annotate certain gene models due to the occurrence of one or more non-canonical and/or unknown splice sites. It is uncertain if these splice sites are real or are substitution errors resulting from long-read genome assemblies' propensity towards substitution errors; nonetheless, this approach allows the discovery of additional real gene models beyond what PASA might find while acting to limit false positive annotations.

Final checks include the removal of models that only contain short introns < 50 bp in length or gene models that partially overlap existing genes and contain introns > 10000bp in length. Gene models that only consist of short introns have been observed by us to occur when indel errors were present in the genomic sequence which are avoided by alignment and annotation algorithms by introducing false introns which span these indels. Such cases are expected to occur in recently pseudogenised genes where sequence conservation deteriorates. Gene models that overlap existing genes and contain large introns may occur where an alignment is "borrowing" a protein domain from existing gene models. 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 loosely 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. In tandem with *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 with 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"; "--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 excepting for "--stringent_alignment_overlap 30".

GMAP alignments of transcript 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 formatted 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, precision, and F1 score (harmonic average of sensitivity and precision) with novel loci also reported (novel loci being those present in the comparison GFF3 and not in the reference GFF3). 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 sequences 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 the results of *de novo* and genome-guided transcriptome assembly combination with EvidentialGene for each species are presented in Table 1. There is some variation in transcriptome assembly quality which can be seen by the high BUSCO completeness scores achieved for *S. cerevisiae* and *C. elegans* (98.3% and 96.9 %, respectively) and comparatively mediocre scores for the remaining species (< 85 %). Quality can also be determined by comparing the number of “okay” contigs 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 suggested for *Mus 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 and provides the ability to ascertain how the quality of the input transcriptome might impact results. Since *ab initio* gene prediction was not involved in PASA or PASA+GGF annotations, these statistics represent the theoretical upper limits of how complete our annotations could be.

PASA gene models

PASA gene model prediction metrics are presented in Table 2. Quality variability in transcriptomes are propagated to the genome as expected, with BUSCO scores being slightly lower than the upper limits imposed by the transcriptomes in most cases. We do note that *S. cerevisiae*’s number of loci and BUSCO completeness scores were reduced by the automatic processing system of processing_pipeline.sh substantially more than occurred in other species (data not shown), and this is likely because this species' genome has a higher gene density than the pipeline is designed to 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 is expected since GGF prevents new genes from overlapping existing genes > 35 % of the new gene's 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 P and P+G annotations to each species' primary annotations 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 P and P+G 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 being "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*: P=0; P+G=0, *A. thaliana*: P=61; P+G=68, *C. elegans*: P=320; P+G=347, *M. musculus* 7.8 Gbp: P=105; P+G=111, *M. musculus* 15 Gbp: P=255; P+G=264). Most other categories of genes are of relatively low abundance and the P+G annotation does not differ substantially from P.

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 P+G relative to P (*S. cerevisiae*: P=4,759 out of 5,553 orthogroups; P+G=4,779 out of 5,553, *A. thaliana*: P=11,883 out of 20,778; P+G=12,329 out of 20,778, *C. elegans*: P=11,199 out of 18,651; P+G=11,559 out of 18,651, *M. musculus* 7.8 Gbp: P=8,557 out of 49,542; P+G=8,676 out of 49,542, *M. musculus* 15 Gbp: P=9,394 out of 47,739; P+G=9,522 out of 47,739). This equates to an increase in the number of orthogroups with equivalent membership to that of the primary annotation (assumed to be the result of correctly annotating all members of a gene family) 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.

Look at the orthogroup that saw the most expansion in a bit more detail…

Discussion

Pseudogenes are often ambiguously annotated in genomes which makes it difficult to ascertain whether these values indicate an “error” on the part of GGF.

Note somewhere that aligning gene models back to the genome as CDS is useful if you used ab initio prediction during the initial gene annotation, as these genes may not be represented in your transcriptome but might have multiple copies throughout the genome.

Fundamentally, GGF is designed to supplement existing annotations by identifying genes that were missed by prior annotation efforts and, by providing an existing gene annotation, we can ensure that…

Detail the running time of program for each species (1 sentence, no need for figures, just to let people know that it is relatively quick and not memory-intensive).

Subheading

-

Role of duplicate genes in genetic robustness against null mutations: S cerevisiae doesn’t have many “gene families”?

There are a lot of pseudogenes in C. elegans genome https://genome.cshlp.org/content/12/5/770.full.

Limitations

Quality of input transcriptome is major factor. PASA uses special handling for fragmented transcripts which, while not immune to errors itself, does help to alleviate any issues that might occur due to low-quality inputs. Thus, it is recommended that GGF only be used in cases where the input transcriptome was generated with high-depth sequencing and assembled using a thorough approach like that detailed in this study.

Conclusions

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

-

Data accessibility statement

Github link **

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 / Tables

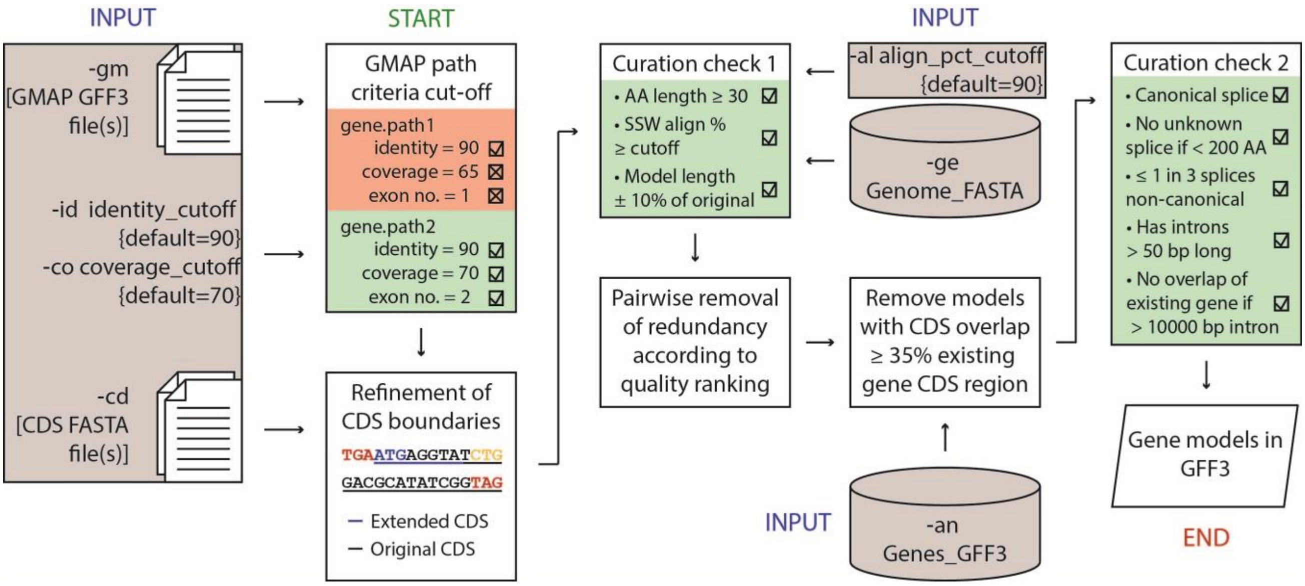

Figure 1. Simplified overview of the programmatic process underlying gene model prediction with gmap_gene_find (GGF). As depicted, GGF makes only minor modifications to sequence alignments which are subject to a variety of curation systems to remove low-quality alignments and return alignments likely 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.

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 refers to these same genes, but 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.

	Matching loci						Unmatched loci	
	P			P+G			P	P+G
Species (dataset)	Sensitivity	Precision	F1	Sensitivity	Precision	F1	Novel loci	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