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 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 inputs GGF requires are alignment files produced by GMAP alignment of CDS which form the basis of gene prediction; it is expected that these alignments are present as GFF3 (General Feature Format 3) files which can be generated by GMAP with the "-f 2" argument. Secondly, the CDS FASTA file used for GMAP alignment is also required; 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 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 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) alignment of a potential gene model against the original CDS from which GMAP's alignment 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 as determined by maximising the alignment score).

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 alignments 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 (e.g., CTG), 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 codon with a canonical 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. Moreoever, 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 greater than 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 this would represent extension into (what should be) non-coding sequence which is 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 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 criteria. The first check involves comparison of gene models whereby if only one gene model contains a microexon (exon less than 30 bp in length), it is removed from consideration. Although microexons are a biologically valid occurrence (Ustianenko, Weyn-Vanhentenryck, & Zhang, 2017), the presence of such a feature in only one of the sequences may indicate that sequence alignment has attempted 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; an additional rationale for this is that longer sequences are less likely to be 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 (data not shown) 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 possible in such species that non-canonical splice sites are more frequent occurrences than that observed in highly-studied species. Moreoever, non-canonical and unknown splice sites may instead represent chance errors resulting from long-read genome assemblies' propensity towards substitution, insertion, and deletion errors; nonetheless, this approach allows the discovery of additional real gene models beyond what PASA might find while still acting to limit false positive annotations.

Final checks include the removal of models that contain only short introns less than 50 bp in length or gene models that partially overlap existing genes and contain introns greater than 10,000 bp in length. Gene models that only consist of short introns have been observed by us to occur when frameshift mutations are present in the genomic sequence which are avoided by alignment and annotation algorithms via introduction of false introns which span these indels. An example scenario where this is expected to occur is 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.

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

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

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

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

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

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

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

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