Application note

gFACs: Filtering, analysis, and conversion to unify genome annotations across alignment and gene prediction frameworks.

Madison Caballero[1]* and Jill Wegrzyn[1]*.

[1]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA

Abstract

Motivation: Published genome annotations are filled with erroneous gene models that represent issues associated with frame, start side identification, splice sites, and related structural features. The source of these inconsistencies can often be traced to translated text file formats designed to describe long read alignments and predicted gene structures. The majority of gene prediction frameworks do not provide downstream filtering to remove problematic gene annotations, nor do they represent these annotations in a format consistent with current file standards. In addition, these frameworks lack consideration for functional attributes, such as the presence or absence of protein domains which can be used for gene model validation.

Summary: To provide oversight to the increasing number of published genome annotations, we present gFACs as a software package to filter, analyze, and convert predicted gene models and alignments. gFACs operates across a wide range of alignment, analysis, and gene prediction software inputs with a flexible framework for defining gene models with reliable structural and functional attributes. gFACs supports common downstream applications, including genome browsers and generates extensive details on the filtering process, including distributions that can be visualized to further assess the proposed gene space.

Availability and Implementation: gFACs is freely available and implemented in Perl with support from BioPerl libraries: https://gitlab.com/PlantGenomicsLab/gFACs

Contact

Corresponding Authors: Madison.Caballero@uconn.edu and jill.wegrzyn@uconn.edu

Supplementary data

Supplemental table 1 and supplemental figure 1.

1. Introduction

In the era of high throughput sequencing, the size and complexity of the genomes assembled in recent years, has dramatically increased. Despite this, only a handful of the nearly 6,000 eukaryote genomes in Genbank are resolved at, or close to, chromosome level (Benson et al., 2017). In addition, over 85% of these genomes contain some type of gene annotation errors (Starmer et al., 2006; Poptsova et al., 2010; Denton et al., 2014). These challenges are unlikely to diminish since projects, such as the Earth BioGenome Project, intend to sequence 1.5M eukaryotic genomes in coming years. Initiatives such as these will assemble increasingly large and complex genomes to assess greater biodiversity.

The majority of genome annotations are semi-automated, derived from informatic approaches that involve a combination of sequence alignments and *ab initio* predictions (Cantarel et al., 2008; Haas et al., 2008; Hoff et al., 2015, Seeman T., 2014). The inputs may include pre-assembled transcripts, raw RNA-Seq reads, and closely related proteins. The resources considered will depend on the available evidence, as well as the complexity and size of the genome under investigation. The downstream genome annotations and upstream alignment files are represented in one of the more variable bioinformatic standard file formats, known as the Generic Feature Format (GFF). The GFF file provides structure for information rich annotations as compared to the reduced representation available through GTF (General Transfer Format). Generation of a final gene annotation requires filtering of incomplete or unlikely structural models and consideration of functional annotations at the full protein or protein domain level. The informatic packages that distill several sources of evidence into gene annotations frequently deliver these without tools to assess their validity.

gFACs represents a flexible annotation refining application that can accept standard annotations from primary gene annotation software as well as transcript/protein sequence aligners. This, in combination with the reference genome, can filter erroneous gene models, generate statistics/distributions, and provide outputs for standard downstream processing and/or visualization. This application does not replace the *ab initio* or similarity based prediction models but serves as a companion tool to resolve conflicting annotations and improve the quality of the final models.

2.1 Input

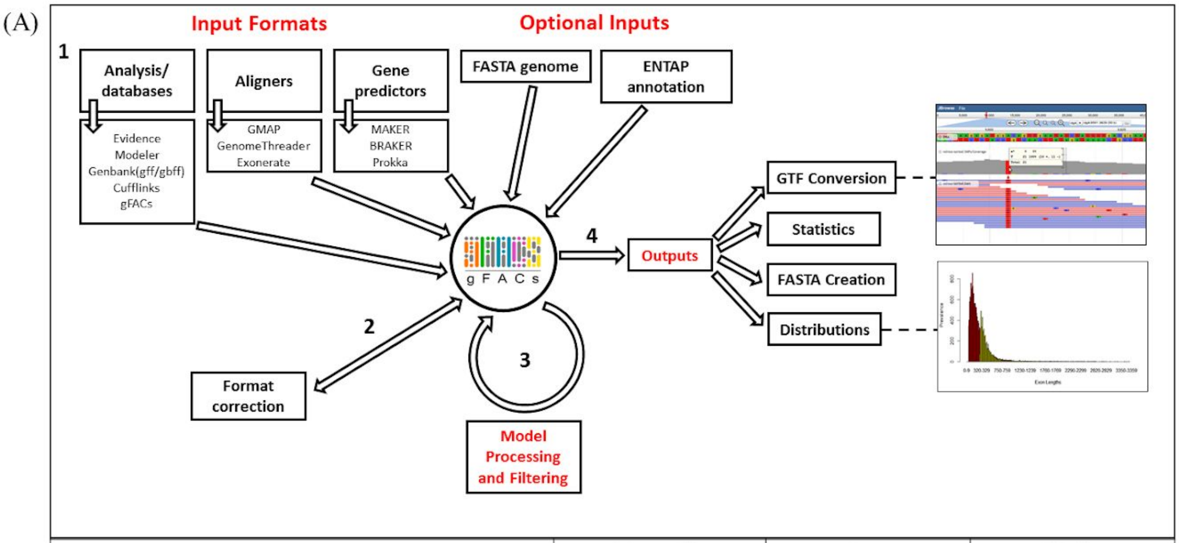
Accepted inputs span a range of aligners and gene predictors, which are presented in formats with similarities to GTF and GFF files. The user specifies the file source at runtime, which can be selected from an applicable set of flags. gFACs will optionally accept the reference genome in FASTA format to permit more refined filtering and analysis. The second optional file, includes the annotation flat file resulting from EnTAP which provides a functional annotation summary, including similarity search, protein domain, and gene family assignments for the proposed gene models or aligned sequences (Hart et al. 2018) (Figure 1A). The physical positions represented in these files are formatted into an intermediate text file to aid in processing and calculating the proposed gene space.

2.2 Model processing and filtering

gFACs removes erroneous models through a set of 14 user selected filtering options, optionally aided by a reference genome or functional annotation. A notable feature in gFACs is the ability to discern and separate isoforms and conflicting models. Each proposed model is subject to a predetermined set of filters as flagged by the user, many of which can be customized, such as setting minimum intron lengths or detection of in-frame stop codons. gFACs allows for independent isoforms to collapse into one model when full duplicates are observed. The addition of functional annotations will allow for the exclusion of models without a protein domain.

2.3 Output

gFACs provides a multitude of output options alongside a log detailing the gFACS process and each filtering effect. Additional options for output include gene/protein FASTA files, GTF represented models, comprehensive statistics on the selected gene models, and distribution tables of gene features.



(B)

	Borrelia burgdorferi B31	*Bos taurus*	*Malus domestica*
Reported protein coding genes	1,391	21,414	47,866
Monoexonic genes	1,391	9,599	4,079
Filter 1: Genes that lack a start codon	305	880	1,095
Filter 2: Genes that lack a stop codon	7	435	492
Filter 3: Genes with an inframe stop codon	0	810	834
Filter 4: Genes with CDS size < 360nt	182	863	2,479
Filter 5: Duplicate models	0	0	0
Models that pass gFACs Filters 1-5	953	18,680	42,864

Figure 1. (A) gFACs pipeline. (B) Quantitative evaluation of 5 of the 14 potential gFACs filtering options on three public NCBI RefSeq annotations. All models filtered are protein coding genes defined as having CDS feature in their respective GFF file. Rows 3-7 are filters on the original set and are non-additive. The additive effect of filters from Filters 1-5 is also represented.

3. Application

Examining public protein coding gene model annotations provides insight on some of the common issues associated with gene annotations (Figure 1B). Common problems noted in public annotations, include: completeness (lack of start/stop or in-frame stops), gene structure (splice sites, intron/exon lengths, and mono-exonic to multi-exonic model rations), fragmentation (incorrect start site assignment), and lack of functional validation (similarity searches, protein domains, gene family assignments). To demonstrate its utility, gFACs was applied to three public genomes (*Borrelia burgdorferi* B31*,* GCF_000008685.2; *Bos taurus*, GCF_000003055.6; *Malus domestica*, GCF_000148765.1) with existing annotations (Fraser et al. 1997; Zimin et al. 2009; Velasco et al. 2010). A total of five of the possible 14 filters were applied to the genomes, which represent unique sources: microbial, plant, and animal. Approximately 30% of gene models were removed for *B. burgdorferi* and at least 10% for *B. taurus and M. domestica*. By utilizing 4 more gFACs filtering options, a further 2,883 models were removed from *M. domestica* (Table S1). The filtering process generates a new set of gene models in FASTA format (nucleotide and amino acid) as well as a GTF representation of the new annotation. The distributions resulting from these filters can be easily imported in packages, such as R to view: gene lengths, CDS lengths, exon lengths, and exon size by order (Figure S1). gFACs represents a comprehensive framework for evaluating, filtering, and analyzing gene models from a range of input applications and preparing these annotations for formal publication or downstream analysis.

Acknowledgements

We thank members of the Ecology & Evolution Biology Department at the University of Connecticut, including: Elizabeth Jockusch, Janine Caira, Hannah Ralicki, and Kaitlin Gallagher for their valuable insight and testing. We also thank the members of the Plant Computational Genomics lab at the University of Connecticut: Sumaira Zaman, Nasim Rahmatpour, Taylor Falk, and Alex Trouern-Trend for the input on the software.

Funding

This work has been supported by the Science Foundation Plant Genome Research Program (1444573).

Conflict of Interest: none declared.

References

Benson D. et al. (2017). GenBank. Nucleic Acids Res. 45(D1), D37-D42.

Cantarel B. et al. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188–196.

Denton J. et al. (2014). Extensive error in the number of genes inferred from draft genome assemblies. PLOS 10, e1003998.

Fraser C et al. (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390, 580-586.

Haas B. et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology 9, R7.

Hart A. et al. (2018). EnTAP: Software to improve the quality and functional annotation of de novo assembled non model eukaryotic transcriptomes.

Hoff K. et al. (2015). BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32, 767-9.

Poptsova M. et al. (2010). Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiology 156, 1909-1917.

Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068-2069.

Starmer J. et al. (2006). Predicting Shine–Dalgarno sequence locations exposes genome annotation errors. PLOS 2, 454-466.

Velasco R. et al. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). Nat Genet 42, 833-839.

Zimin A. et al. (2009). A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10, R42.