Bioinformatics, 2017, 1–2
doi: 10.1093/bioinformatics/btx353
Advance Access Publication Date: 5 June 2017
Applications Note

Genome analysis

FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation

Byoungnam Min,[1] Igor V. Grigoriev[2,3] and In-Geol Choi[1,*]

[1]Department of Biotechnology, College of Life Sciences and Biotechnology, Korea University, Seoul 02841, Korea, [2]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA and [3]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on February 22, 2017; revised on April 27, 2017; editorial decision on May 29, 2017; accepted on May 31, 2017

Abstract

Motivation: Successful genome analysis depends on the quality of gene prediction. Although fungal genome sequencing and assembly have become trivial, its annotation procedure has not been standardized yet.

Results: FunGAP predicts protein-coding genes in a fungal genome assembly. To attain high-quality gene models, this program runs multiple gene predictors, evaluates all predicted genes, and assembles gene models that are highly supported by homology to known sequences. To do this, we built a scoring function to estimate the congruency of each gene model based on known protein or domain homology.

Availability and implementation: FunGAP is written in Python script and is available in GitHub (https://github.com/CompSynBioLab-KoreaUniv/FunGAP). This software is freely available only for noncommercial users.

Contact: igchoi@korea.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene prediction is the most fundamental process in eukaryotic genome annotation. Several eukaryotic gene prediction programs, such as Augustus (Stanke *et al.*, 2006), Maker (Cantarel *et al.*, 2008), GeneMark (Lukashin and Borodovsky, 1998) and Braker1 (Hoff *et al.*, 2016), are available. Because these programs often output different predictions for the same genomic region, evaluation of predicted gene models is essential when using multiple gene predictors.

We developed a fully automated fungal Genome Annotation Pipeline designated 'FunGAP', which employs three publicly available programs, Augustus, Braker and Maker, to generate all plausible predicted gene models. FunGAP is a powerful tool for evaluating and filtering preliminary gene models that are highly supported by homology to known sequences. We benchmarked the pipeline with five representative fungal genomes: *Saccharomyces cerevisiae*, *Neurospora crassa*, *Schizophyllum commune*, *Rhizopus oryzae* and *Gonapodya prolifera*, which represent different branches of the fungal tree of life. The outputs were compared to those of other eukaryotic gene prediction programs.

2 Software description

2.1 Input files

FunGAP takes two inputs: genome assembly in FASTA format and Illumina-generated mRNA sequencing reads in FASTQ format (two paired-end files). Users also need to provide their own protein database in FASTA for homology searching, which can be facilitated by *download_sister_orgs.py* script packaged with FunGAP.

2.2 Annotation procedure

FunGAP comprises three major steps: (i) preprocessing, (ii) gene prediction and (iii) evaluation and filtration. In preprocessing, the repeats in the input assembly are masked, and mRNA sequencing reads are assembled. FunGAP uses Augustus, Braker and Maker for gene prediction. The evidence scores for all predicted gene models are calculated based on the alignment of translated protein sequences with Pfam (Finn *et al.*, 2016), Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao *et al.*, 2015) and BLAST (Boratyn *et al.*, 2013). The tools embedded in FunGAP are listed in Supplementary Table S1. The detailed pipeline description is explained in the Supplementary Notes.

The most prominent feature of FunGAP is its evaluation and filtration function. FunGAP produces 'non-overlapping' coding sequences by evaluating all gene models and retaining only best-scored models. The evaluation is performed by three tools: BLASTp, BUSCO and InterProScan (Jones *et al.*, 2014). The assumption is that gene models with higher similarity to known sequences are more likely to be actual genes. Every gene model is translated and is assigned an evidence score by summing alignment bit scores for Pfam, BUSCO, and BLAST against the user-provided protein database. In the BLAST alignment, matching sequence length coverages (*matched length/query length* and *matched length/target length*) are multiplied by the BLAST bit score because longer gene models have more chances of getting higher bit scores. The evaluation score can be represented by the following equation:

$$\begin{aligned}\text{Evidence score} &= (\text{BLAST score} * \text{coverage}) + \text{BUSCO score} \\ &\quad + \sum \text{Pfam scores}.\end{aligned}$$

To aggregate the related gene models, FunGAP finds a set of gene models with at least one base pair overlap and tagged as a 'gene block' (Supplementary Fig. S2). In a gene block, the best combination of gene models with the highest evidence score sum is selected as the final gene models. Short coding sequence overlap (<10% of coding sequence length) is allowed.

2.3 Output

FunGAP organizes outcomes into the various public formats, such as GFF3 and FASTA.

3 Results

We compared FunGAP annotation results with those of three known programs (Augustus, Maker, and Braker) and representative predictions (RefSeq in NCBI). The tested fungal genomes include *S. cerevisiae*, *N. crassa*, *S. commune*, *R. oryzae* and *G. prolifera* (Supplementary Table S2). Evaluation criteria established for checking the quality of gene predictions include BLAST hits against the provided protein database and SwissProt, all Pfam domains and gene models containing them, complete and missing BUSCOs, all transcriptome alignments and those with >90% coverage, and the number of genes with the same exon–intron structure to the reference genes. As a whole, FunGAP showed more reliable predictions than those shown by the sole usage of other programs. The benchmark results are summarized in Table 1 and Supplementary Tables S3–S7. For all the five genomes, integrating three programs using our evaluation and filtration method made best matches to the references. This indicates that a reliable algorithm was used to combine multiple programs.

Table 1. Comparison of gene predictions in the five benchmarked genomes

Genomes	Augustus	Braker	Maker	FunGAP	References
S. cerevisiae	6,144	5488	5523	5654	5913
	4776	4793	4695	**5222**	5913
N. crassa	12 211	9009	10 241	10 130	10 785
	5276	6503	5789	**6642**	10 785
S. commune	22 428	13 718	13 115	15 337	13 194
	1395	3448	1896	**3542**	13 194
R. oryzae	13 146	13 218	15 971	15 548	17 459
	3734	3826	3536	**4462**	17 459
G. prolifera	38 867	15 502	21 551	21 784	13 831
	554	4770	2180	**4980**	13 831

Note: Predicted genes (up) and reference matches (down) are shown. We counted the matches when predicted genes have the same exon–intron coordinates with the corresponding reference genes.

Benchmarking was performed on a server with 48 cores [Intel(R) Xeon(R) CPU E5-2670 v3] and 64 GB of RAM. For multithread jobs, 40 cores were used. The total running time ranged from 9 to 36 h (Supplementary Fig. S4). FunGAP requires 14 GB hard disk space for installation, which the majority is from InterProScan (11 GB).

Funding

This work was supported by the Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ01044003), Rural Development Administration, Republic of Korea; and School of Life Sciences and Biotechnology for BK21 PLUS, Korea University.

Conflict of Interest: B.M. and I.C. receive royalties from commercial users of FunGAP. Code patent is pending approval.

References

Boratyn,G.M. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.

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

Finn,R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

Hoff,K.J. *et al.* (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.

Jones,P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Lukashin,A.V., and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

Simao,F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Stanke,M. *et al.* (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.*, **34**, W435–W439.