A MAKER pipeline for prediction of protein-coding genes in parasitic worm genomes

Matthew Berriman (✉ mb4@sanger.ac.uk)  
Berriman Lab Group, Sanger Institute

Eleanor Stanley  
Berriman Lab Group, Sanger Institute

Avril Coghlan  
Berriman Lab Group, Sanger Institute

Method Article

Keywords: MAKER, Augustus, GeneMark, SNAP, RATT, genBlastG, gene-finding, annotation, gene prediction, gene models, gene set, parasites, genomes, assemblies, parasitic worms, helminths, nematodes, flatworms, platyhelminths, Nematoda, Platyhelminthes

Posted Date: May 15th, 2018

DOI: https://doi.org/10.1038/protex.2018.056

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Once a genome has been sequenced, a challenging task is to predict protein-coding genes in the genome assembly. If several genomes of related species have been sequenced, it is desirable that the pipeline is scalable to multiple species, and will give consistent results across species. Here we describe a consistent, scalable and automated computational pipeline, based on the MAKER software, for prediction of protein-coding genes in parasitic worm (nematode and platyhelminth) genome assemblies. This protocol can be used to generate a relatively accurate and complete gene set, for a broad range of nematode and platyhelminth species. Furthermore, it does not require RNAseq data and does not require training using a curated set of genes of known structure.

Introduction

Once a genome has been sequenced, a challenging task is to predict protein-coding genes in the genome assembly. If several genomes of related species have been sequenced, it is desirable that the pipeline is scalable to multiple species, and will give consistent results across species. Here we describe a consistent, scalable and automated computational pipeline, based on the MAKER software, for prediction of protein-coding genes in parasitic worm (nematode and platyhelminth) genome assemblies. In our pipeline, gene predictions are generated using a pipeline that uses MAKER version 2.2.28 (Holt et al. 2011). Our MAKER annotation pipeline consists of four steps, taking into account evidence from multiple sources (Figure 1). First, repetitive elements in the genome are identified and masked using RepeatMasker (www.repeatmasker.org) by scanning scaffolds for matches to repeats from a repeat library generated using RepeatModeler (www.repeatmasker.org/RepeatModeler.html). Second, ab initio gene models to be used as evidence within MAKER are generated using Augustus 2.5.5 (Stanke et al. 2006), GeneMark-ES 2.3a (self-trained) (Ter-Hovhannisyan et al. 2008), and SNAP 2013-02-16 (Korf 2004). Further gene models to use as MAKER input are generated using comparative algorithms genBlastG (She et al. 2011) (using comparisons to C. elegans gene models from WormBase; Yook et al. 2012) and RATT (Otto et al. 2011; transferring gene models from the taxonomically nearest published ‘reference’ genome from the list: Haemonchus contortus for clade V parasites; Ascaris suum for Ascaridomorpha; Brugia malayi (and Onchocerca volvulus) for Spiruromorpha; Trichuris muris for clade I; Strongyloides ratti for clade IV; Hymenolepis microstoma for cestodes except Echinococcus multilocularis for Taenia species; Schistosoma mansoni for trematodes). Third, species-specific ESTs and cDNAs from INSDC (Cochrane et al. 2016), and proteins from related species (see below), are aligned against the genome using BLASTN and BLASTX (Altschul et al. 1997), respectively, and these alignments are further refined with respect to splice sites using exonerate (Slater & Birney 2005). Last, the EST and protein homology alignments, comparative gene models, and ab initio gene predictions are integrated and filtered by MAKER to produce a gene set for the species, with just one transcript for each gene. The four-step MAKER pipeline is run three consecutive times (Figure 1). The first run is performed using the est2genome option with species-specific ESTs and cDNAs and the protein2genome for nematode protein sequences from UniProt’s UniRef 90 clusters for nematodes \[...\]
For this first MAKER run, Augustus and SNAP are trained using CEGMA \(\text{Parra } et al\_2007) \) gene models for KOGs, as well as ‘nematode orthologous groups’ \(\text{(NOGs)} \) \(\text{(Martin } \& \text{ Mitreva }2018)\), ‘trematode orthologous groups’ \(\text{(TROGs)}\), or ‘cestode orthologous groups’ \(\text{(CEOGs)}\) as appropriate. Gene models obtained from the first MAKER run are used to train SNAP, and MAKER is run a second time, using the same nematode proteins as in the first run. Gene models from the second run are then used to train Augustus. Using the trained versions of SNAP and Augustus, MAKER is run a third time, using a taxonomically broader protein set that includes proteins from metazoans with complete proteomes from UniProt and proteins from helminths from GeneDB \(\text{(Logan-Klumpler } et al\_2012)\). The resulting MAKER gene set is then filtered to remove less reliable gene models. This protocol can be used to generate a relatively accurate and complete gene set, for a broad range of nematode and platyhelminth species. Unlike gene-finding approaches that require extensive use of RNAseq data \(\text{(for example, use of Augustus in the } _{Strongyloides} \text{ project; Hunt } et al\_2016)\), this protocol does not require RNAseq data, so is suitable when RNAseq data is not available. Furthermore, unlike many gene-finding approaches, it does not require training using a curated set of genes of known structure \(\text{(e.g. confirmed using cDNAs or RNAseq)}\), as it uses conserved genes from KOGs/NOGs/TROGs/CEOGs for an initial round of training, and then its own gene MAKER models for later rounds of training.

**Reagents**

WormBase database \(\text{(www.wormbase.org)}\) INSDC database \(\text{(www.insdc.org)}\) UniProt database \(\text{(www.uniprot.org)}\) GeneDB database \(\text{(www.genedb.org)}\)

**Equipment**

Computer cluster.

**Procedure**

Gene predictions are generated using MAKER version 2.2.28 \(\text{(Holt } et al\_2011)\). The MAKER annotation pipeline consists of four steps, taking into account evidence from multiple sources. The four-step MAKER pipeline is run three consecutive times \(\text{‘round 1’, ‘round 2’, ‘round 3’ in Figure 1)}\). **Step 1** 1. Generate a repeat library for the genome of interest using RepeatModeler \(\text{(www.repeatmasker.org/RepeatModeler.html)}\). 2. Identity and mask repetitive elements in the genome using RepeatMasker \(\text{(www.repeatmasker.org)}\) by scanning scaffolds for matches to the repeats in the repeat library. **Step 2** 1. For the first MAKER run \(\text{‘round 1’ in Figure 1)}\), train SNAP 2013-02-16 \(\text{(Korf 2004)}\) and Augustus 2.5.5 \(\text{(Stanke } et al\_2006)\) using CEGMA \(\text{(Parra } et al\_2007)\) gene models for KOGs, as well as ‘nematode orthologous groups’ \(\text{(NOGs)} \) \(\text{(Martin } \& \text{ Mitreva }2018)\), ‘trematode orthologous groups’ \(\text{(TROGs)}\), or ‘cestode orthologous groups’ \(\text{(CEOGs)}\) as appropriate. For the second MAKER run \(\text{‘round 2’)}\), first train SNAP using the MAKER gene models obtained from the first MAKER run \(\text{‘round 1’)}\). For the third MAKER run \(\text{‘round 3’)}\), first train Augustus using the MAKER gene models obtained from the second MAKER run \(\text{‘round 2’)}\). 2. Generate _ab initio_ gene models to be used as
evidence within MAKER, using Augustus, GeneMark-ES 2.3a (self-trained) (Ter-Hovhannisyan _et al_ 2008), and SNAP. 3. Generate further gene models to use as MAKER input, using comparative algorithms genBlastG (She _et al_ 2011) (using comparisons to _C. elegans_ gene models from WormBase; Yook _et al_ 2012) and RATT (Otto _et al_ 2011; transferring gene models from the taxonomically nearest published ‘reference’ genome from the list: _Haemonchus contortus_ for clade V parasites; _Ascaris suum_ for Ascaridomorpha; _Brugia malayi_ (and _Onchocerca volvulus_) for Spiruromorpha; _Trichuris muris_ for clade I; _Strongyloides ratti_ for clade IV; _Hymenolepis microstoma_ for cestodes except _Echinococcus multilocularis_ for _Taenia_ species; _Schistosoma mansoni_ for trematodes). **Step 3**

1. Align species-specific ESTs and cDNAs from INSDC (Cochrane _et al_ 2016), and proteins from related species (see below), against the genomes using BLASTN and BLASTX (Altschul _et al_ 1997), respectively. The first and second MAKER runs (‘round 1’ and ‘round 2’ in Figure 1) are performed using the est2genome option with species-specific ESTs and cDNAs, and the protein2genome for nematode protein sequences from UniProt’s UniRef 90 clusters for nematodes (UniProt Consortium, 2015). The third MAKER run (‘round 3’) is performed using species-specific ESTs and cDNAs, and a taxonomically broader protein set that includes proteins from metazoans with complete proteomes from UniProt and proteins from helminths from GeneDB (Logan-Klumpler _et al_ 2012). 2. Refine these alignments further with respect to splice sites using exonerate (Slater & Birney 2005). **Step 4**

1. The EST and protein homology alignments, comparative gene models, and _ab initio_ gene predictions are integrated and filtered by MAKER to produce a gene set for the species, with just one transcript for each gene. **Step 5**

After running the MAKER pipeline above (Steps 1-4) three times (rounds 1-3), filter the resulting MAKER gene set to remove less reliable gene models, as follows: 1. Discard any MAKER gene models that were based on exonerate or BLASTX alignments, and do not overlap any Augustus, genBlastG or RATT gene model (as they were probably due to spurious alignments). 2. Discard MAKER gene models that encode proteins of shorter than 30 amino acids. 3. If two different MAKER gene models overlap in their coding sequence, discard the gene model with the worst MAKER score (i.e. AED score).

**Troubleshooting**

It is likely that in a highly fragmented genome assembly, many genes will be split into multiple partial gene predictions. To correct for the effect of assembly fragmentation on the gene count, the gene count in your assembly can be normalised, by dividing the total proteome length by the mean protein length for _C. elegans_ (409.82 amino acids). This should give a more accurate estimate of the true gene count in your species.

**Anticipated Results**

The output from the protocol is a set of protein-coding gene predictions for one or more genome assemblies of interest.

**References**
Altschul, S. F. _et al._ Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. _Nucleic acids research_ **25**, 3389-3402 (1997). Cochrane, G. _et al._ The International Nucleotide Sequence Database Collaboration. _Nucleic acids research_ **44**, D48-50 (2016). Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. _BMC bioinformatics_ **12**, 491 (2011). Hunt, V. L. _et al._ The genomic basis of parasitism in the _Strongyloides_ clade of nematodes. _Nat Genet._ **48**, 299-307 (2016).

Korf, I. Gene finding in novel genomes. _BMC bioinformatics_ **5**, 59 (2004). Logan-Klumpler, F. J. _et al._ GeneDB—an annotation database for pathogens. _Nucleic acids research_ **40**, D98-108 (2012). Martin, J. & Mitreva, M. Finding set of "core genes" for Nematoda, Trematoda and Cestoda. _Protocol Exchange_ (2018). Otto, T. D.,  _et al._ RATT: Rapid Annotation Transfer Tool. _Nucleic acids research_ **39**, e57 (2011). Parra, G. _et al._ CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. _Bioinformatics_ **23**, 1061-1067 (2007). She, R. _et al._ genBlastG: using BLAST searches to build homologous gene models. _Bioinformatics_ **27**, 2141-2143 (2011). Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. _BMC bioinformatics_ **6**, 31 (2005). Stanke, M. _et al._ AUGUSTUS: _ab initio_ prediction of alternative transcripts. _Nucleic acids research_ **34**, W435-439 (2006). Ter-Hovhannisyan, V. _et al._ Gene prediction in novel fungal genomes using an _ab initio_ algorithm with unsupervised training. _Genome Res._ **18**, 1979-1990 (2008). UniProt Consortium. UniProt: a hub for protein information. _Nucleic acids research_ **43**, D204-212 (2015). Yook, K. _et al._ WormBase 2012: more genomes, more data, new website. _Nucleic acids research_ **40**, D735-741 (2012).

**Acknowledgements**

We would like to thank the WSI Pathogen Informatics team, especially Jacqueline Keane and Andrew Page.

**Figures**
Figure 1

Flowchart of protocol Four-step MAKER pipeline.