BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS - **Supplementary**

Katharina J. Hoff, Simone Lange, Alexandre Lomsadze,
Mark Borodovsky and Mario Stanke

October 26, 2015

Contents

Chapter 1

Supplementary Results

Supplementary Tables 1.1 and 1.2 show for repeat masked an unmasked genomes respectively the values of accuracies of i/ GeneMark-ET (*ab initio*, first step in BRAKER1), ii/ AUGUSTUS (*ab initio*, trained on GeneMark-ET predictions) and iii/ AUGUSTUS using RNA-Seq read alignment information as extrinsic evidence (the last step of BRAKER1). Therefore, the last set of accuracy values is labeled as "BRAKER1".

On the fungus *S. pombe*, GeneMark-ET is superior to the current version of BRAKER1. This is likely to be related to special features in intron splicing mechanism in fungal genomes where the branch point site plays a larger role in splicing than in other species. On the other hand, the acceptor site in fungi plays a smaller role in intron recognition mechanism and carries less signal information. GeneMark-ET has special models to accomodate fungal gene organization, therefore, it delivers accuracy that is difficult to exceed even with use of RNA-Seq information in the prediction step. However, in future versions of BRAKER1 with AUGUSTUS adapting such a model we expect to see the same pattern of improving accuracy with respect to pure *ab initio* predictions made by GeneMark-ET.

	A. thaliana			*C. elegans*		
	GeneMark-ET	AUGUSTUS	BRAKER1	GeneMark-ET	AUGUSTUS	BRAKER1
Gene sensitivity	52.6	49.2	**64.4**	42.8	42.3	**55.0**
Gene specificity	51.0	48.2	**52.0**	42.4	46.2	**55.2**
Transcript sensitivity	44.5	41.6	**55.0**	32.7	32.7	**43.0**
Transcript specificity	51.0	48.2	**50.9**	42.4	46.2	**53.2**
Exon sensitivity	80.5	76.4	**82.9**	79.7	74.9	**80.2**
Exon specificity	77.9	**79.8**	79.0	78.7	83.4	**85.3**
	D. melanogaster			*S. pombe*		
	GeneMark-ET	AUGUSTUS	BRAKER1	GeneMark-ET	AUGUSTUS	BRAKER1
Gene sensitivity	54.6	56.1	**67.6**	**81.3**	70.5	77.4
Gene specificity	55.1	53.7	**61.1**	**83.9**	76.0	80.5
Transcript sensitivity	39.8	41.1	**50.2**	**81.3**	70.4	77.4
Transcript specificity	55.1	53.7	**59.9**	**83.9**	76.0	76.5
Exon sensitivity	66.6	65.0	**73.3**	**87.4**	73.2	83.2
Exon specificity	62.1	63.3	**67.3**	**88.2**	83.1	83.2

Table 1.1: Accuracy results of GeneMark-ET (*ab initio*), AUGUSTUS (*ab initio*) and BRAKER1 (AUGUSTUS trained on filtered GeneMark-ET predictions; run with RNA-Seq hints) on *softmasked* genomes.

	A. thaliana			*C. elegans*		
	GeneMark-ET	AUGUSTUS	BRAKER1	GeneMark-ET	AUGUSTUS	BRAKER1
Gene sensitivity	53.7	52.1	**63.9**	42.9	43.1	**55.0**
Gene specificity	46.1	42.9	**51.6**	41.3	44.7	**55.2**
Transcript sensitivity	45.3	43.8	**56.4**	32.8	33.3	**43.2**
Transcript specificity	46.1	42.9	**50.4**	41.3	44.7	**53.3**
Exon sensitivity	81.1	77.8	**82.5**	**79.9**	75.5	79.8
Exon specificity	72.3	73.8	**78.8**	77.7	82.3	**85.4**
	D. melanogaster			*S. pombe*		
	GeneMark-ET	AUGUSTUS	BRAKER1	GeneMark-ET	AUGUSTUS	BRAKER1
Gene sensitivity	56.1	57.6	**68.4**	**82.1**	71.5	77.0
Gene specificity	53.8	50.8	**60.6**	**84.0**	76.1	80.1
Transcript sensitivity	40.7	42.0	**50.6**	**82.1**	71.5	77.0
Transcript specificity	53.8	50.8	**59.3**	**84.9**	76.1	76.1
Exon sensitivity	67.3	66.0	**73.7**	**88.4**	75.4	83.0
Exon specificity	60.6	61.5	**67.1**	**88.2**	83.1	82.9

Table 1.2: The same as in Table 1.1 for unmasked genomes.

The MAKER2 annotation pipeline automatically repeat masks genomes, runs several gene finding tools

(such as SNAP, AUGUSTUS and GeneMark-ES), integrates extrinsic evidence when it is available and creates a "combined gene set" from all input sources.

We determined (trained) parameters of SNAP and AUGUSTUS on MAKER2 derived training sets generated from the Cufflinks assembled RNA-Seq mapped to unmasked genomes. GeneMark-ES self-training requires only a genome sequence.

Suppl. Table 1.3 shows *ab initio* prediction accuracy of SNAP, AUGUSTUS and GeneMark-ES on the four unmasked genomes with the above mentioned parameter sets (derived by MAKER2). We observed that for *A. thaliana* and *D. melanogaster*, MAKER2 was able to improve several measures of gene prediction accuracy in comparison with the stand alone tools. However, for *C. elegans* and *S. pombe* no improvement was observed.

	A. thaliana				*C. elegans*			
	SNAP	AUGUSTUS	GeneMark-ES	MAKER2	SNAP	AUGUSTUS	GeneMark-ES	MAKER2
Gene sensitivity	14.5	37.5	50.7	**51.3**	21.0	23.9	**42.4**	41.0
Gene specificity	11.2	32.1	44.3	**52.5**	15.3	25.7	**41.1**	30.8
Transcript sensitivity	11.9	31.6	42.7	**43.5**	16.4	18.3	**32.3**	31.3
Exon sensitivity	58.7	71.5	**80.3**	76.1	67.9	69.1	**80.0**	69.4
Exon specificity	44.3	66.5	68.8	**76.1**	53.1	72.1	**77.2**	62.3
	D. melanogaster				*S. pombe*			
	SNAP	AUGUSTUS	GeneMark-ES	MAKER2	SNAP	AUGUSTUS	GeneMark-ES	MAKER2
Gene sensitivity	40.8	41.3	52.6	**58.0**	50.3	61.8	**80.8**	42.8
Gene specificity	30.6	37.6	**50.3**	47.9	49.4	67.4	**84.2**	68.7
Transcript sensitivity	30.4	29.6	38.4	**42.3**	50.3	61.8	**80.8**	42.8
Exon sensitivity	57.4	56.3	**66.1**	64.9	66.4	64.8	**87.4**	50.1
Exon specificity	43.5	53.2	**56.9**	55.0	56.4	69.7	**88.4**	71.4

Table 1.3: The first three columns in each table show the *ab initio* prediction accuracy of SNAP, AUGUSTUS and GeneMark-ES (trained in accordance to the MAKER2 manual) on unmasked genomes. The fourth column in each cell shows accuracy of MAKER2 on the genome with masked repeats and SNAP and AUGUSTUS utilizing RNA-Seq hints. (Transcript specificity is not shown since it is in this case identical to Gene specificity; none of the methods predicted alternative transcripts.)

Ab initio gene prediction accuracy of AUGUSTUS depends on the parameters determined in training. We compare two cases: the parameter training for AUGUSTUS with BRAKER1 and the supervised parameter training on expert generated set of genes - Suppl. Table 1.4.

	A. thaliana		*C. elegans*		*D. melanogaster*		*S. pombe*	
	Mode 1	Mode 2	Mode 1	Mode 2	Mode 1	Mode 2	Mode 1	Mode 2
Gene sensitivity	**51.8**	49.2	**45.2**	42.3	53.2	**56.1**	**78.4**	70.5
Gene specificity	**56.6**	48.2	43.2	**46.2**	**63.4**	53.7	**84.8**	76.0
Transcript sensitivity	**44.2**	41.6	**35.2**	32.7	39.0	**41.1**	**78.4**	70.4
Exon sensitivity	**79.8**	76.4	**76.1**	74.9	**67.7**	65.0	**84.7**	73.2
Exon specificity	**81.7**	79.8	81.5	**83.4**	**67.4**	63.3	**89.4**	83.1

Table 1.4: *Ab initio* gene prediction accuracy of AUGUSTUS trained in two different modes. Mode 1: training on expert generated sets of genes; Mode 2: training on BRAKER1 generated set of genes. ("Transcript specificity" for *ab initio* gene prediction (when no alternative transcripts are predicted) is identical to "Gene specificity" and therefore is not shown in Table 1.4.)

1.1 Remark on Runtime

The main factor that dominates runtime is the genome size; the pipeline runtime scales linearly with the genome size. Main modules of BRAKER1 could be run in parallel. The time-consuming alignment of RNA-Seq reads against the genome was not counted to the runtime of BRAKER1; the alignment is an input to BRAKER1. It is highly likely that users would have already aligned the RNA-Seq data to genome by means of their favorite short read aligner. The runtime increases when more alternative splicing is represented within the RNA-Seq alignments by a factor that is roughly the number of predicted alternative transcripts per gene.

Chapter 2

Supplementary Methods

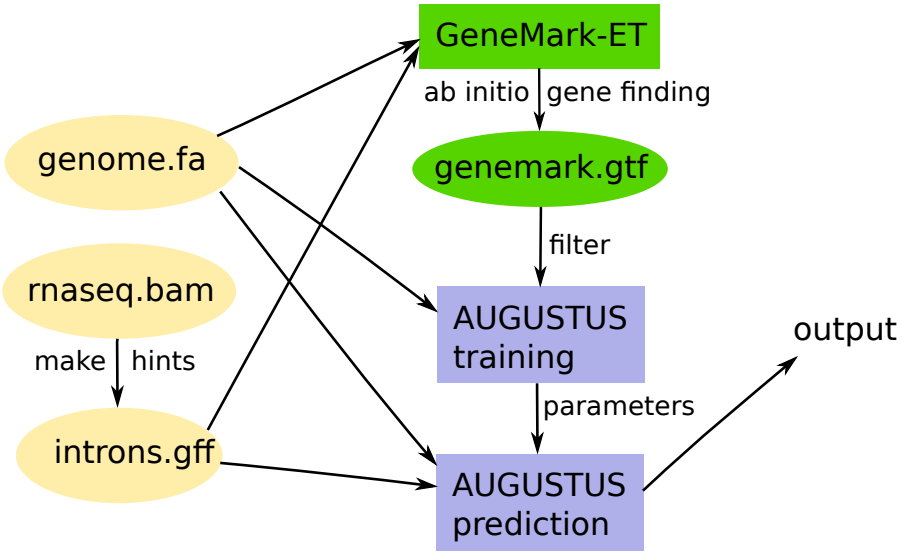

Figure 2.1: Schematic view of the BRAKER1 pipeline.

2.1 Test Data

In order to demonstrate prediction accuracy, nuclear genomes, reference annotations and RNA-Seq libraries were retrieved for four model organisms from the respective databases: for *Arabidopsis thaliana*, TAIR 10 from `http://arabidopsis.org`; for *Caenorhabditis elegans*, WS240 from `http://www.wormbase.org`; for *Drosophila melanogaster*, R5.55 from `http://flybase.org`; for *Schizosaccharomyces pombe*, ASM294v2.23 from `http://www.pombase.org`. The following RNA-Seq libraries were retrieved from the short read archive at NCBI: SRR934391 (for *A. thaliana*); SRR065719 (for *C. elegans*); SRR023505, SRR023546, SRR023608, SRR026433, SRR027108 (for *D. melanogaster*); SRR097898-SRR097900, SRR097902, SRR097903, SRR097905-SRR097909, SRR097912, SRR097915, SRR097917, SRR097921, SRR097922, SRR097925, SRR402833 (for *S. pombe*).

2.2 RNA-Seq Alignment

RNA-Seq libraries were aligned against the respective genomes with TopHat2 version 2.0.11 [Kim *et al.*, 2013] using Bowtie2 version 2.2.2
[Langmead and Salzberg, 2012].

In order to determine library specific characteristics such as insert size, TopHat2 was first run with standard parameters. Minimum and maximum Intron length parameters were adjusted based on the results from Tophat2 run with default parameters. Initial alignments were processed with Cufflinks [Mortazavi *et al.*, 2008] version 2.2.0. The Cufflinks run was interrupted after the output showed "Fragment Length Distribution" parameters. Subsequently, TopHat2 was run with the such determined mean insert size and standard deviation. Table 2.1 shows Tophat2 parameters that were used for all sets of RNA-Seq libraries in this publication.

Library	--mate-inner-dist	--mate-std-dev	--min-intron-length	--min-coverage-intron	--min-segment-intron	--microexon-search	--max-intron-length
A. thaliana: SRR934391	20	46	50	50	50	✓	100000
C. elegans: SRR065719	14	25	20	20	20	✓	100000
D. melanogaster: SRR023505, SRR023546, SRR023608, SRR026433, SRR027108	38	26	30	30	30	✓	50000
S. pombe: SRR097898, SRR097899, SRR097900, SRR097902, SRR097903, SRR097905, SRR097906, SRR097907, SRR097908, SRR097909, SRR097912, SRR097915, SRR097917, SRR097921, SRR097922, SRR097925, SRR402833	50	20	20	20	20	✓	100000

Table 2.1: TopHat2 parameters used for RNA-Seq alignment.

2.3 Cufflinks Assembly

MAKER2 [Holt and Yandell, 2011] and CodingQuarry [Testa *et al.*, 2015] require an RNA-Seq assembly. We used Cufflinks version 2.2.0 with standard parameters to assemble the libraries that had previously been aligned to the genome with TopHat2.

2.4 Repeat Masking

Genomes were softmasked for repeats using RepeatModeler 1.0.8 [Smit and Hubley, 2015].

2.5 Running BRAKER1

The command for running BRAKER1 version 1.6 with GeneMark-ET version 4.29 and AUGUSTUS version 3.1.0 on softmasked genomes was

```
braker.pl --genome=genome.fa --species=speciesname --bam=accepted_hits.bam --softmasked
```

where `accepted_hits.bam` is the TopHat2 alignment file and `speciesname` is a name for storing output parameters.
For unmasked genomes, the command flag `--softmasked` was removed.
For *S. pombe*, BRAKER1 was run with the flag `--fungus` which enable usage of the fungi-specific branchpoint model of GeneMark-ET.

2.6 Running MAKER2

For running MAKER2 (version 2.31.6), we followed the tutorial at `http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_GMOD_Online_Training_2014`, mostly.

A database of transposable elements and a RepBase repeat library were used for repeat masking with MAKER2.

MAKER2 was run with the gene finders SNAP (version 2006-07-28) [Korf, 2004], AUGUSTUS (version 3.1.0) [Stanke *et al.*, 2008], and GeneMark-ES [Lomsadze *et al.*, 2005] (version 4.29). MAKER2 was developed to run with protein database to support gene discovery. We did not enable protein database support because we were interested in gene finding performance using only RNA-seq data as an input. Thus, MAKER2 - in the way that we used it - is a method that calls and masks repeats, generates training gene structures on the basis of RNA-Seq data, and predicts genes with extrinsic evidence from RNA-Seq data. Running MAKER2 on a novel species consists of three steps:

1. Generating training gene structures for SNAP and AUGUSTUS with MAKER2,
2. training SNAP, AUGUSTUS and GeneMark-ES (outside of MAKER2), and
3. predicing genes with MAKER2 using GeneMark-ES, SNAP and AUGUSTUS.

Training gene structures can be improved an iterative fashion. We computed two iterations.

Training of the gene prediction tool GeneMark-ES depends on the genome sequence, only, and is therefore independent of MAKER2. GeneMark-ES was trained using the command

```
gm_es.pl --ES genome.fa
```

For *S. pombe*, GeneMark-ES was trained using the flag `--fungus` to enable the fungi-specific branch point model in GeneMark-ES.

2.6.1 Generating training gene structures for SNAP and AUGUSTUS (first iteration)

MAKER2 configuration files were generated

```
maker -CTL
```

The file `maker_opts.ctl` was edited to contain (besides default parameters):

```
genome=genome.fasta
est=transcripts.fa # Cufflinks assembly
est2genome=1
```

No HMMs for gene finders were configured. MAKER2 was run with the following command:

```
maker
```

Chromosome-specific `gff` files were converted to `ann/zff` and `dna` files (native format for training SNAP):

```
# maker2zff belongs to MAKER2
maker2zff Chr.gff
mv genome.ann genome.Chr.ann
mv genome.dna genome.Chr.dna
cat genome.Chr.ann ... > genome.ann
cat genome.Chr.dna ... > genome.dna
```

To obtain a training gene file for AUGUSTUS, `ann/zff` files were first converted to `gff3`, and from there to `gtf`:

```
# zff2gff3.pl belongs to SNAP
zff2gff3.pl genome.Chr.ann > Chr.gff3
cat Chr.gff3 | perl -ne '
  if(not(m/^\#/)){
    chomp; @t = split(/\t/);
    @t2 = split(/=/, $t[7]);
    print "$t[0]\t$t[1]\t$t[2]\t$t[3]\t$t[4]\t$t[5]\t$t[6]\t";
    print "\tgene_id \"$t2[1]Chr\"; transcript_id";
    print "  \"$t2[1]Chr\"\n";
  }'> Chr.gtf
# the last two lines makes gene/transcript IDs unique across
# different chromosomes

# join gtf files from different chromosomes:
cat Chr.gtf ... > all.gtf
```

AUGUSTUS training genes are excised with a flanking noncoding region. In BRAKER1, the average gene length divided by two is used as a flanking region length. The average gene length and resulting flanking region length were computed:

```
cat all.gtf | perl -ne '
  @t = split(/\t/);
  $seen{$t[8]} += ($t[4] - $t[3] + 1);
  if(eof()){
    $sum = 0; $c = 0;
    foreach my $key ( keys %seen ){
    $c=$c+1; $sum += $seen{$key};}
    print $sum."/".$c."=".($sum/$c);
    print "\n";
  }'
# the resulting number was divided by two
```

This resulted in the following flanking region lengths that were used for excising training genes from the genome for training AUGUSTUS for MAKER2 (first iteration):

Species	Flanking region length (nt)
Arabidopsis thaliana	644
Caenorhabditis elegans	616
Drosophila melanogaster	972
Schizosaccharomyces pombe	1050

The `gtf` file was converted to a `gb` file with above flanking region length:

```
# gff2gbSmallDNA.pl belongs to AUGUSTUS
gff2gbSmallDNA.pl all.gtf genome.fasta $flanking_region_length first.gb
```

Above described procedure lead to the generation of the following numbers of training genes:

Species	Number of training genes
Arabidopsis thaliana	13547
Caenorhabditis elegans	7660
Drosophila melanogaster	7049
Schizosaccharomyces pombe	307

2.6.2 Training SNAP and AUGUSTUS (first iteration)

Training SNAP

```
# fathom, forge and hmm-assembler.pl are part of SNAP
fathom -categorize 1000 genome.ann genome.dna
fathom -export 1000 -plus uni.ann uni.dna
forge export.ann export.dna
hmm-assembler.pl ${species} . > ${species}.hmm
```

Training AUGUSTUS

A new species for AUGUSTUS parameters was created:

```
# new_species.pl is part of AUGUSTUS
new_species.pl --species=${species}
```

The original training gene structure contained errors, such as occasionally missing start- or stop-codons. Such error containing genes were filtered out:

```
# etraining and filgerGenesOut_mRNAname.pl are part of AUGUSTUS
etraining --species=maker2_spomb1 first.gb 1> etrain-test.out
   2> etrain-test.err

fgrep "gene" etrain-test.err | cut -f 2 -d " " > bad.etraining-test.lst

filterGenesOut_mRNAname.pl bad.etraining-test.lst first.gb > second.gb

etraining --species=maker2_spomb1 second.gb
```

The training gene set was split into two sets, the second set was subsequently further split in another two sets, resulting in three different files:

1. A small "test set" of 200 (in case of *S. pombe* first iteration: 23) genes for measuring accuracy after `etraining` and `optimize_augustus.pl`,
2. a large gene set for `etraining`, that was further split into:
   (a) a large gene set for for the option `--onlytrain` of `optimize_augustus.pl`,
   (b) a small gene set for `optimize_augustus.pl`, the size was 1000 for all species except for `S. pombe`, where genes were not further split due to their small number.

```
# randomSplit.pl is part of AUGUSTUS
randomSplit.pl second.gb 200
randomSplit.pl second.gb.train 1000
# this results in the following files:
# 1) second.gb.test -> measuring accuracy
# 2) second.gb.train -> etraining
# 2a) second.gb.train.train -> --onlytrain in optimize_augustus.pl
# 2b) second.gb.train.test -> optimize_augustus.pl
```

Major AUGUSTUS parameters were ajusted with `etraining`:

```
etraining --species=${species} second.gb.train
```

Other parameters were optimized with `optimize_augustus.pl`:

```
optimize_augustus.pl --species=${species} --onlytrain=second.gb.train.train second.gb.train.test
```

2.6.3 Generating training gene structures (second iteration)

Generating training gene structures for SNAP

MAKER2 parameters were generated:

```
maker -CTL
```

The file `maker_opts.ctl` was edited to contain (besides default parameters):

```
genome=genome.fasta
est=transcripts.fa # Cufflinks assembly
snaphmm=${species}.hmm
```

MAKER2 was run with the following command:

```
maker
```

Training gene extraction for SNAP was performed as described for iteration 1 in section 2.6.1.

Generating training gene structures for AUGUSTUS

MAKER2 parameters were generated:

```
maker -CTL
```

The file `maker_opts.ctl` was edited to contain (besides default parameters):

```
genome=genome.fasta
est=transcripts.fa # Cufflinks assembly
augustus_species=${species}
```

MAKER2 was run with the following command:

```
maker
```

Training gene extraction for AUGUSTUS was performed as described for iteration 1 in section 2.6.1, leading to the following numbers of training genes:

Species	Number of training genes
Arabidopsis thaliana	11887
Caenorhabditis elegans	5711
Drosophila melanogaster	7568
Schizosaccharomyces pombe	1013

2.6.4 Training SNAP and AUGUSTUS (second iteration)

Training SNAP

Training SNAP was performed as described in section 2.6.2.

Training AUGUSTUS

Training AUGUSTUS was performed as described in section 2.6.2, except that no new species was created (parameters of iteration 1 were refined).

2.6.5 Predicting genes with MAKER2

Preparing `rnaseq.gff3`

Junctions generated by TopHat2 and Cufflinks transcripts were converted to `gff3` format:

```
# tophat2gff3 and cufflinks2gff3 are part of MAKER2
tophat2gff3 junctions.bed > tophat.gff3
cufflinks2gff3 transcripts.gtf > cufflinks.gff3
cat tophat.gff3 cufflinks.gff3 > rnaseq.gff3
```

Running MAKER2

MAKER2 parameters were generated:

```
maker -CTL
```

The file `maker_opts.ctl` was edited to contain (besides default parameters):

```
genome=genome.fasta
est=transcripts.fa # Cufflinks assembly
est_gff=rnaseq.gff3 # TopHat2 and Cufflinks
augustus_species=${species}
snaphmm=${species}.hmm #SNAP HMM file
gmhmm=${species}/gmhmm.mod #GeneMark HMM file
augustus_species=${species} # AUGUSTUS model
keep_preds=1
```

MAKER2 was run with the following command:

```
maker
```

2.7 Running CodingQuarry

As a tool specific for fungi, we ran CodingQuarry on *S. pombe*, only. First, Cufflinks assemblies were converted from `gtf` to `gff3`, then CodingQuarry was executed:

```
CufflinksGTF_to_CodingQuarryGFF3.py transcripts.gtf > transcripts.gff3
```

```
CodingQuarry -f genome.fasta -t transcripts.gff3 -d -p 8
```

2.8 *Ab initio* Gene Predictions

In order to show that using RNA-Seq data is beneficial to gene prediction accuracy as compared to *ab initio* prediction, genes were predicted *ab initio* with parameters that were used for running MAKER2 and BRAKER1, respectively.

2.8.1 For comparison with MAKER2

SNAP, AUGUSTUS and GeneMark-ET were used for running MAKER2. *Ab initio* predictions were produced with the following commands:

```
snap species_specific_maker2_snap_model unmasked_genome.fa > snap.zff
augustus species=species_specific_maker2_augustus_model unmasked_genome.fa > augustus.out
gmes_petap.pl  --ES  --sequence unmasked_genome.fa --pbs --v
```

2.8.2 For comparison with BRAKER1

AUGUSTUS was run with *expert trained* parameters as distributed with the AUGUSTUS release (parameter set `arabidopsis` for *A. thaliana*, `caenorhabditis` for *C. elegans*, `fly` for *D. melanogaster* and `schizosaccharomyces_pombe` for *S. pombe* in *ab initio* mode on softmasked genomes with the following command:

```
augustus --species=species --softmasking=on softmasked_genome.fa > augustus.out
```

AUGUSTUS was also run in *ab initio* mode using the exact same command with the parameters trained by BRAKER1 for each species.

2.9 Measuring Accuracy

Accuracy was measured using the Eval package [Keibler and Brent, (2003)].

Bibliography

[Kim *et al.*, 2013] Kim, D. and Pertea, G. and Trapnell, C. and Pimentel, H. and Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, *Genome Biology* **14**:R36.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2, *Nature Methods* **9**: 357-359.

[Mortazavi *et al.*, 2008] Mortazavi, A. and Williams, B.A. and McCue, K. and Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods* **5**: 621-628.

[Holt and Yandell, 2011] Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, *BMC Bioinformatics*, **12**:491.

[Testa *et al.*, 2015] Testa, A.C. and Hane, J.K. and Ellwood, S.R. and Oliver R.P. (2015) CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts, *BMC Genomics* **16**:170.

[Korf, 2004] Korf, I. (2004) Gene finding in novel genomes, *BMC Bioinformatics* **5**:59 S1-S9.

[Stanke *et al.*, 2008] Stanke, M. and Diekhans, M. and Baertsch, R. and Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics*, **24**(5), 637.

[Lomsadze *et al.*, 2005] Lomsadeze, A. and Ter-Hovhannisyan, V. and Chernoff, Y. and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm, *Nucleic Acids Research* **33**:20 6494-6506.

[Keibler and Brent, (2003)] Keibler, E. and Brent, M.R. (2003) Eval: a software package for analysis of genome annotations, *BMC Bioinformatics* **4**:50.

[Smit and Hubley, 2015] Smit, A.F.A. and Hubley, R. (2008-2015) RepeatModeler Open-1.0 *http://www.repeatmasker.org*.