Comparative transcriptomics as support for gene predictions

Teuntje Peeters
Supervisors: Harm Nijveen en Marnix Medema
Date: 15-12-2017

Summary

Gene prediction is a technique used to identify genes in genomic DNA. Structural annotation of genes is the determination of the exact structure, e.g. the exact positions of introns and exons of genes. There are several tools available to generate high quality genome annotations and provide gene models with the annotations, like BRAKER1, Augustus, Genemark and SnowyOwl. These tools only use (RNA-Seq) data from the same organism, which is not always available. Here we provide a pipeline that tries to improve gene models, generated by Augustus, of one species (*Aspergillus oryzae*) based on gene models, generated by BRAKER1 with RNA-Seq, of a related species (*Aspergillus. terreus*). Validation of improved gene models did not show improvement compared to the reference genome and non-improved gene models of *A. oryzae*. The improved gene models in *A. oryzae* contained 22% alternative gene models. The compared gene models to the reference showed a maximal specificity of 14% and a gene sensitivity of 10% at most. Exon sensitivity and specificity show higher percentages, up to 50%. To summarize, our results show no improvement in gene annotation of *Aspergillus oryzae* based on *Aspergillus terreus* or *Aspergillus nidulans*. This pipeline needs an additional validation. The genome annotation of *A. oryzae* is incomplete since it is missing genes and the sampling algorithm of Augustus is not as good as the Viterbi algorithm.

Introduction

Background

Fungi produce a large number of natural products, referred to as specialized metabolites (Calvo, Wilson, Bok, & Keller, 2002). These metabolites mediate various interactions with bacteria, plants and other fungi. These metabolic products can have many advantages and can be valuable for different kinds of reasons, for both human and fungi. For humans this can be, for example, antibiotics like penicillin. Penicillin (Ligon, 2004), is an important and well-known example of a fungal antibiotic. However, only a fraction of these metabolic products has been identified and many valuable molecules are waiting to be discovered (Medema & Fischbach, 2015).

After determining which genes are responsible for previously mentioned metabolites in these organisms, gene prediction could be a method to determine where the genes coding for these metabolites are located on the genome of different species of fungi. One approach to identify genes in genomic DNA is Gene Prediction. (Haas, Zeng, Pearson, Cuomo, & Wortman, 2011; Yandell & Ence, 2012)

Structural annotation of genes is the determination of the exact structure e.g. the exact positions of introns and exons of genes. RNA-Seq data can be very helpful for gene prediction because the exons on the genome are covered by the reads and the introns are not while mapping RNA-Seq to the genome. Therefore, RNA-Seq data gives hints where the exons and introns are located on the genome, which leads to a reliable structural gene prediction (Z. Li et al., 2011) (L. Li et al., 2015). Several tools are available to generate high quality genome annotations (Yandell & Ence, 2012). One example is BRAKER1 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016; Supplementary et al., 2015). BRAKER1 uses RNA-Seq for genome annotation with GeneMark-ET (Lomsadze, Burns, & Borodovsky, 2014b) and Augustus (Stanke & Morgenstern, 2005). GeneMark-ET is an unsupervised gene prediction tool that takes RNA-Seq data into account in the prediction. Augustus is based on a generalised hidden Markov model (GHMM) that determines probability distributions for different parts on the genome, like introns, exons, intergenic regions etc. (Stanke & Morgenstern, 2005). Another example is a tool called SnowyOwl (Reid et al., 2014), that combines RNA-Seq data with homology information. SnowyOwl trains an HMM by assembling RNA-Seq reads into transcripts and uses these transcripts to predict genes. These genes are translated to proteins and (significant) homologues are found with BLASTP.

Another method to identify genes involved in biosynthesis of previously mentioned metabolites, is to look at biosynthetic gene clusters. Currently, *in silico* synthetic DNA based methods are being developed to discover products of unknown biosynthetic gene clusters. Unfortunately, gene cluster synthesises are limited by errors in intron and exon predictions (Wisecaver & Rokas, 2015). Consequently, this leads to non-functional polypeptide gene products. Improved gene prediction could be a solution to recover these errors in intron and exon predictions.

There are a lot of gene prediction tools available, but these tools only use (RNA-Seq) data from the same organism. However, RNA-Seq data is not always available for every organism. Previously described tools can only use data that is generated from the same organism. So, there is a need for a new prediction tool with accurate intron and exon prediction that takes (RNA-Seq) data from related species into account. One of the major benefits of RNA-Seq data is that RNA-Seq provides evidence for prediction of gene structure, the (exact) intron and exon locations of genes (Z. Li et al., 2011)(L. Li et al., 2015).

Project

The goal of this project was to improve gene annotation with the support of RNA-Seq data from a related genome. Here, we provide a pipeline that creates gene models for two different but closely related species (*A. oryzae and A. terreus*). This has been validated with the following three organisms: *Aspergillus oryzae, Aspergillus terreus* and *Aspergillus nidulans.* One species will be treated as it does not contain RNA-Seq data (*A. oryzae).* The other species will be treated as they do contain RNA-Seq data (*A. terreus, A. nidulans*). BRAKER1 is used to create gene models for the species with RNA-Seq data and Augustus is used for the other species. These gene models are being compared with BLASTP (Altschul, Gish, Miller, Myers, & Lipman, 1990) and homologous sequences are selected. Based on these homologous sequences the best gene model is selected as improved gene model. Validation will be a comparison of these gene models to the reference genome.

Methods

Genomes and RNA-Seq data

Three different (related) genomes were obtained from NCBI. *A. terreus* (NZ_AAJN00000000), *A. nidulans (GCF_000149205.1)*, *A. oryzae* (NC_008282). An example of their phylogenetic tree can be found in appendix Figure 11. RNA-Seq data was obtained from the sequence read archive (DRR059466, SRR2409424, SRR5740799). SRA-toolkit from NCBI (SRA_Handbook, 2010-) was used to get the SRA files and convert these files to FASTQ files. Quality check of the reads was done with FastQC (Andrews S., 2010).

Pre-processing data

Hisat2 (Kim, Langmead, & Salzberg, 2015) was used to create a genome index and was used for mapping all the reads to the genome. Samtools (H. Li et al., 2009) was used for the next steps:

Samtools View converted the SAM file to a BAM file. Samtools Sort sorted all reads so the resulting file was an aligned sorted file in BAM format.

Gene models

Augustus (Stanke & Morgenstern, 2005) and BRAKER1 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016a) were used to create gene models of the two species. Augustus and BRAKER1 ran with different parameters. Augustus was run with the sampling algorithm (parameter: alternatives from sampling) as well as with the Viterbi algorithm (Phys, Anderson, Ryon, & Forney, 1973). The Viterbi algorithm generates no alternative gene models for each gene. If the species has not been previously trained in Augustus, it should be trained before running Augustus. This can be done with our instruction file in the appendix below or with the written script for automatization on GitHub. One of the parameters used in Augustus is the species. Augustus created gene models for the organism with the corresponding genome. For the sampling algorithm: 'alternatives from sampling' with sample number 100, was used. This results in more alternative gene models for one gene. For running BRAKER1 the genome must be defined, which is the original genome FASTA file. Also, one of the parameters is the aligned and sorted BAM file to get hints from RNA-Seq data. BRAKER still uses Augustus, so a species has to be trained before BRAKER can create gene models. If the species are already trained, these species can be used by the parameter 'use existing'. Otherwise a new species should be trained.

Translation species

The translation between species is done with BLASTP (Altschul et al., 1990). The translation is done both ways, so one organism to the other organism and the other way around. The best bi-directional hits are selected by getting the highest BIT score. A cut-off of the alignment identity is set to 70% to get orthologues sequences. The Marnix index is calculated. This index is a sum of the length of both aligned sequences divided by the sum of the length of both sequences. For example, if one has two sequences, one of them is A amino acids long and the other is B amino acids long. The alignment length for the query sequence is C amino acids and for the subject sequence D. This would give the following formula:

$$Marnix\ index = \frac{\mathrm{C} + \mathrm{D}}{\mathrm{A} + \mathrm{B}}$$

Graphs and visualisation

All graphs and visualisations are made with R (R core team (2015)). The following packages are used for visualisations: gplots (G.R. Warnes et all, 2016), ggplot2 (H. Wickham, 2009).

Validation

Re-annotation

The results of the newly annotated organism was compared with a Python script to the annotation of NCBI. This python script was written by L. Schmitz (references). This script compares exact start and stop locations of genes. Also, this script compares exon locations but it is not that strict as with gene locations. If there is an overlap of 90% with this exon, this will be considered as a match.

Related organism

For validation, the same analysis as described above (*A. oryzae* and *A. terreus*) was done with another related organism described above, *A. nidulans*.

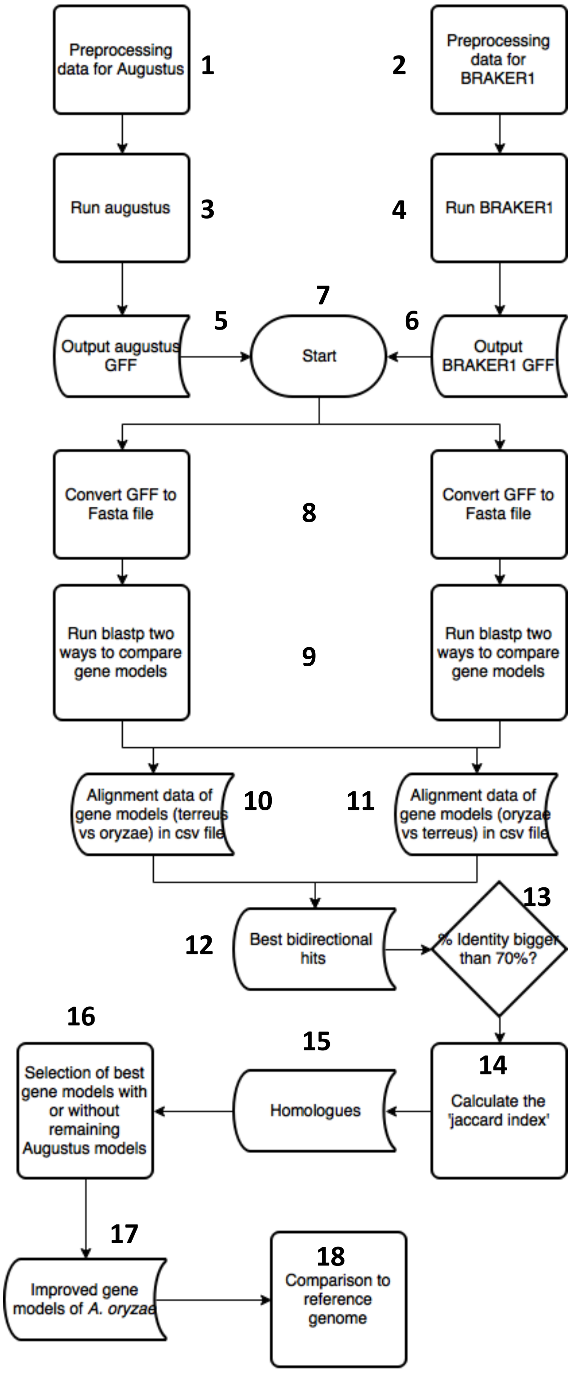

Preprocessing data for Augustus
1
2
Preprocessing data for BRAKER1
Run augustus
3
4
Run BRAKER1
7
Output augustus GFF
5
Start
6
Output BRAKER1 GFF
Convert GFF to Fasta file
8
Convert GFF to Fasta file
Run blastp two ways to compare gene models
9
Run blastp two ways to compare gene models
Alignment data of gene models (terreus vs oryzae) in csv file
10
11
Alignment data of gene models (oryzae vs terreus) in csv file
13
12
Best bidirectional hits
% Identity bigger than 70%?
16
15
14
Selection of best gene models with or without remaining Augustus models
Homologues
Calculate the 'jaccard index'
17
18
Improved gene models of A. oryzae
Comparison to reference genome

Figure 1, Analysis components of the computational workflow. Augustus and BRAKER1 both created gene models. These gene models are being compared with BLASTP. Two different alignments are made. The best bi-directional hits are selected and different selections are applied.

Results

Overview

Our analysis consisted of several computational steps, as seen in Figure 1. The exact pre-processing steps of the data for Augustus and BRAKER1 are described in the supplementary, Workflow. The pipeline uses different methods to improve structural annotations in fungal genomes. Transcriptomic data was used from related species to the query genome, with identification of homogues. A combination of Augustus (Stanke & Morgenstern, 2005) and BRAKER1 (Hoff et al., 2016a) was used to create gene models for both species Figure 1.3 and 1.4. Augustus was used for the genome without RNA-Seq data and BRAKER1 was used for the (related) genome with available RNA-Seq data. A combination of Augustus and BRAKER is used because RNA-Seq data is not available for all organisms. It is important that these species are closely related, because then the gene structure will be more similar. A phylogenetic tree of these organisms is seen in Figure 11 in Supplementary. BRAKER1 (Hoff et al., 2016a) takes sorted and mapped RNA-Seq data as an argument to create gene models based on evidence. Both tools have gene models as result Figure 1.5 and 1.6. The next step is to compare both models, Figure 1.7. First the homologues among the gene models have to be found. BLASTP (Altschul et al., 1990) creates alignments to find homologues, Figure 1.9. BLASTP is used both ways, so the query genome against the subject genome and the other way around. Second, the best hits are selected for each gene model, out of these best hits, the best bi-directional hits are selected Figure 1.12. Homologues were assigned through the identification of best bidirectional BLAST hits with >70% sequence identity, Figure 1.13. If the sequences are not orthologue it is possible that a wrong gene model is selected as 'correct gene model' which will lead to a bad gene model. The identity can be high (>70%), this does not necessarily mean that the sequences have a big overlap. That is why the Marnix index can give an indication about the overlap between the two sequences Figure 1.14. The Marnix index is calculated by adding up the overlap divided by the total length of both sequences. The tool was validated by comparing the re-annotation of *A. oryzae* with the gene models of *A. terreus* to the reference annotation of NCBI, Figure 1.18. Another validation will be done by re-annotating *A. oryzae* with *A. nidulans* again. This annotation was compared to the reference genome.

1. BRAKER1 outperforms Augustus

BRAKER1 (Hoff et al., 2016a) uses a combination of GeneMark-ET (Lomsadze, Burns, & Borodovsky, 2014) and Augustus (Stanke & Morgenstern, 2005). According to the makers of BRAKER1 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016b), BRAKER1 outperforms Augustus. RNA-Seq data from *A. terreus* [SRR2409424] was downloaded from the NCBI website. These datasets were aligned and mapped against the corresponding genome, 66% of the reads mapped against the genome. Gene models were created with Augustus and BRAKER1. Augustus was trained on the same organism, *A. terreus*. By visual inspection of approximately one hundred gene models, the gene models created by

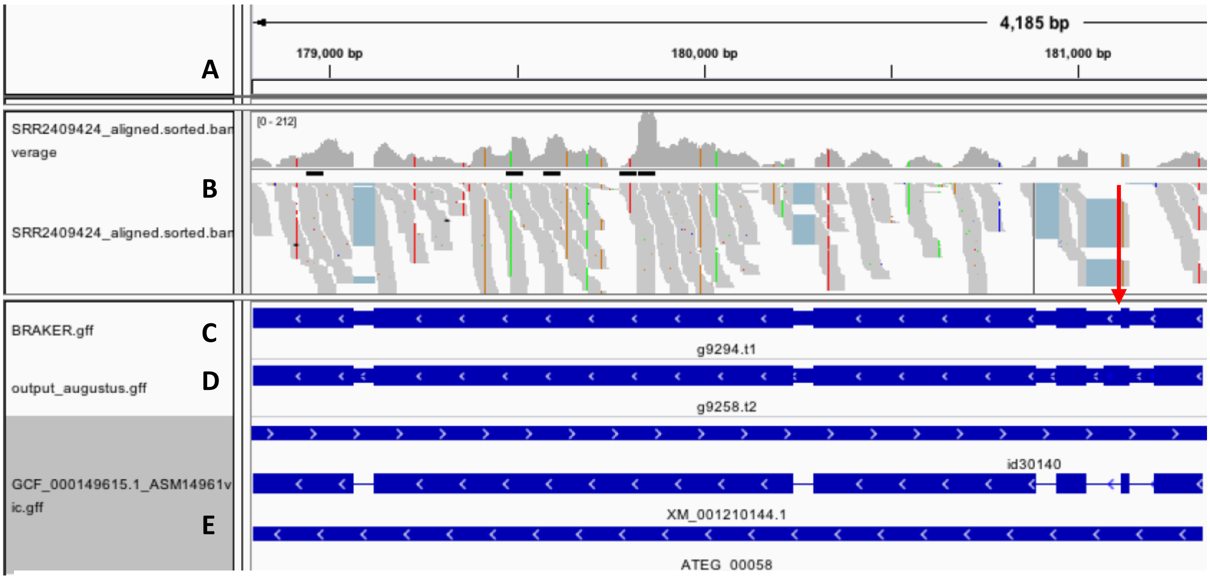
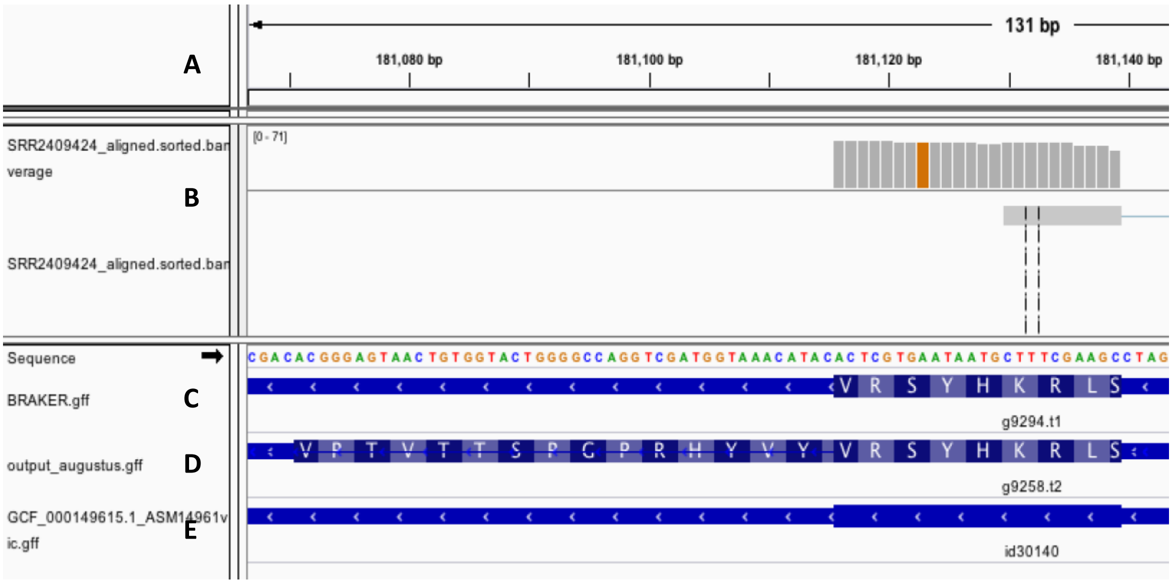


Figure 3, IGV viewer, gene ID: ATEG_00058 from A. terreus. Gene name: oligosaccharyl transferase stt3 subunit. A shows the positions on the genome. B represents the RNA-Seq reads mapped on the genome and the coverage. C represents the gene models created by BRAKER1. D shows the gene models created by Augustus. E represents the genome annotation from NCBI. The red arrow points to the difference in intron and exon position



131 bp
181,080 bp
181,100 bp
181,120 bp
181,140 bp
A
[0 - 71]
SRR2409424_aligned.sorted.bam Coverage
B
SRR2409424_aligned.sorted.bam
Sequence
CGACACGGGAGTAACTGTGGTACTGGGGCCAGGTCGATGGTAAACATACACTCGTGAATAATGCTTTCGAAGCCTAG
BRAKER.gff
C
V R S Y H K R L S
g9294.t1
output_augustus.gff
D
V P T V T T S P G P R H Y V Y V R S Y H K R L S
g9258.t2
GCF_000149615.1_ASM14961v1_genomic.gff
E
id30140

Figure 4, IGV viewer, gene ID: ATEG_00058 from A. terreus. Zoom on the different exon. Gene name: oligosaccharyl transferase stt3 subunit. A shows the positions on the genome. B represents the RNA-Seq reads mapped on the genome and the coverage. C represents the gene models created by BRAKER1. D shows the gene models created by Augustus. E represents the genome annotation from NCBI.

BRAKER1 corresponded better with the RNA-Seq data than the gene models created by Augustus. A few gene models from BRAKER1 and Augustus were manually checked with the genome annotation from NCBI (NZ_AAJN00000000), one example is seen in Figure 3. This is gene: ATEG_00058. There is one difference between the gene models of BRAKER1 and Augustus, the fifth exon (red arrow). Figure 4 zooms in on this problem. According to Augustus this exon should start at position 181,070 while BRAKER states that this exon should start at position 181,116. Based on the annotation on NCBI (Figure 3E, 4E) the annotation of BRAKER1 should be correct.

However, the gene models from BRAKER1 are not always correct. For example: neuronal calcium sensor 1 [ATEG_00933]. Figure 2 in supplementary, shows the RNA-Seq coverage and gene models created by BRAKER1 and Augustus for this gene. Figure 2E shows the gene annotation by NCBI. The gene model created by BRAKER1 is more similar to the gene annotation from NCBI. The gene model created by Augustus is too long. The start and stop position are incorrect, but the gene model created by BRAKER1 also misses an exon, the last exon, according to the annotation on NCBI.

2. Translation species

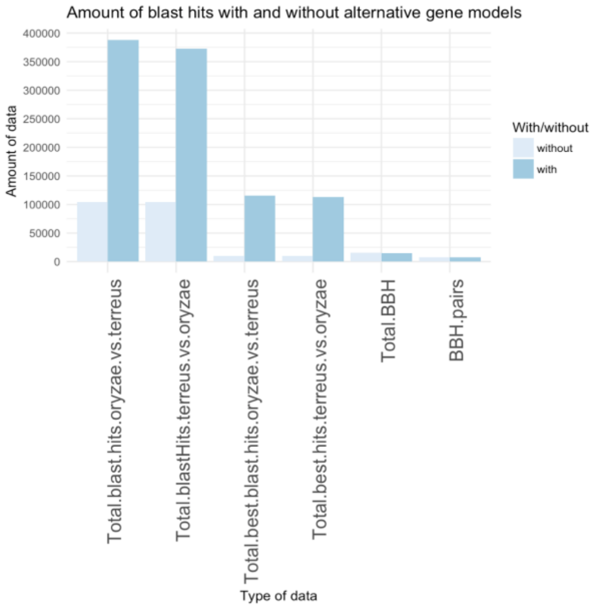


Figure 5, Number of blasts hits with and without alternative gene models. From left to right: the total number of blast hits of A. oryzae versus A. terreus, the total number of blast hits of A. terreus versus A. oryzae. The number of best (highest BIT score) hits for both comparisons and the last two elements are first, the best bi-directional hits (all) and the last column are the best bi-directional hits for the pairs of genes (BBH/2 = BBH.pairs)

Since we are improving gene models from one species with RNA-Seq data from the related species, the translation from one genome to another has to be made. BLASTP was used to find homologues in all gene models between the two related genomes. Each organism has two types of gene models. One dataset with gene models contains alternative gene models, these are predicted with the sampling algorithm of Augustus. The other dataset does not contain gene models with alternative models, so one gene model for each gene, generated by the Viterbi algorithm of Augustus. The total number of gene models with alternative gene models that were submitted in BLASTP was for *A. oryzae*: 28441 and for *A. terreus*: 10320. The number of gene models without alternative models was for *A. oryzae*: 11481 and *A. terreus*: 10550. Figure 5 shows the number of BLASTP hits with *A. oryzae* versus *A. terreus*, the other way around, the number of best hits both ways and the number of best bi-directional hits (unique and pairs). The number of gene models has a big difference between the 'with alternatives' and 'without alternatives'. Though, in the end, approximately the same number of best bi-directional hits are found, as seen in Figure 5.

First and alternative gene models

Both Augustus and BRAKER1 create alternative gene models based on evidence (BRAKER1) or sampling (Augustus). Previous results (Hoff et al., 2016b) show that BRAKER1 has a better performance than Augustus. So, assuming the gene models from BRAKER1 are more reliable, we assume that the highest scoring model from all alternative models is closer to the 'truth'. So preferably, the BLASTP results show that the best hit/alignment of the gene models from BRAKER1 are almost always the first model. The best hit is defined by the highest BIT score. Figure 6 shows that this is the case. From the Augustus hits 22% have approximately the alternative model as the best model while the of BRAKER1 hits less than 1% have the alternative model.

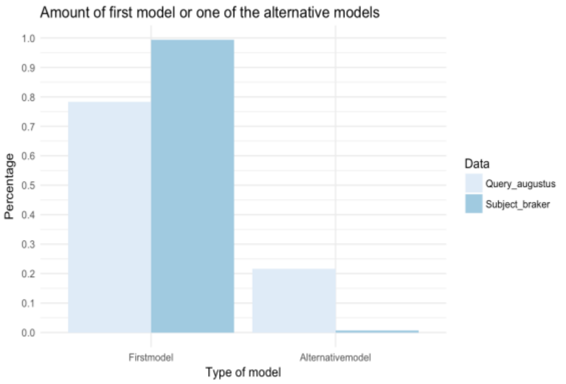


Figure 6. Graphical representation of the best Blast hits from A. oryzae (Augustus gene models) versus A. terreus (BRAKER1 gene models). How many of the first and alternative models are the best hits? The x-axis represents the type of model, first or alternative model. The y-axis shows the percentage of the total number of hits. Light

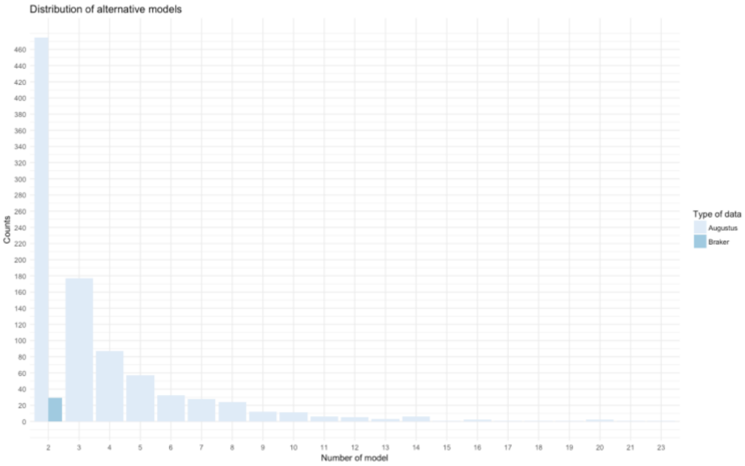


Figure 7, distribution of all alternative models from Augustus and BRAKER1. BRAKER1 only has the second alternative model while the number of Augustus models are up to model 23.

Figure 7 is a representation of the distribution of the alternative models of both Augustus and BRAKER1. So, these are the ranks of alternative models that had the best match with the other gene models. BRAKER1 only has one alternative model while the number of alternative models of Augustus are up to 23.

3. Validation

3.1 Comparison to reference

How well are these gene models compared to the reference genome annotation? Approximately 22% alternative models are used instead of the first and best model of the output of Augustus. These improved gene models are compared with a Python script to the golden standard, the official annotation of *A. oryzae*. Figure 8 shows this comparison in dark-blue. The gene models of *A. oryzae* are improved with the gene models of *A. terreus*. There are two categories, gene models that only had a BLASTP match with the *A. terreus* gene models and the second category are the gene models that had a BLASTP match with the gene models of *A. terreus* including the best Augustus model of the non-BLASTP matches. The gene sensitivity measures the proportion of positives that are truly positive, so the exact matches between our gene models of *A. oryzae* compared with the official annotation. The gene specificity measures the proportion of negatives that are truly negative, so the gene models that are absent in both models. The gene sensitivity and specificity are very low, the percentages are between 3% and 14%. For the exon positions, the values are higher. These values vary between 25% and 50%.

3.2 Another related organism

The analysis above was repeated with another related organism to *A. oryzae*, namely: *A. nidulans*. Figure 8, the colour, between dark blue and light blue, represents the gene models of *A. oryzae* supported by the gene models of *A. nidulans*. The results of the comparison are approximately the same as the gene models of *A. oryzae* supported by *A. terreus*.

3.3 More comparisons

To further investigate why the previous validation gave poor results, more comparisons were performed. Figure 8 shows all comparisons that are made for validation. All statistical details about Figure 8 and all comparisons are found in supplementary, Table1. The first element on the y–axis is: ‘ory all gene models’. This is the dataset of *A. oryzae* that contains also alternative models for each gene computed by Augustus with the sampling algorithm. The second element consists all gene models of *A. oryzae* computed by BRAKER1. So these gene models are supported by RNA-Seq. This organism was treated as it did not contain RNA-Seq data, but it does contain RNA-Seq data. This data was used as an extra comparison. ‘Ory single models’ indicates that Augustus created gene models for *A. oryzae* but it uses the Viterbi algorithm. So these gene models do not contain alternative models, just one model for each gene. The rest of the features is previously explained. When looking at Figure 8, what immediately stands out for the genes is that the gene sensitivity and specificity are very low. The percentages are not that different, but the datasets that have the biggest percentages are the gene models created by BRAKER1 and the gene models created by Augustus with the Viterbi algorithm. When comparing the exons there are bigger differences, also the percentages are higher.

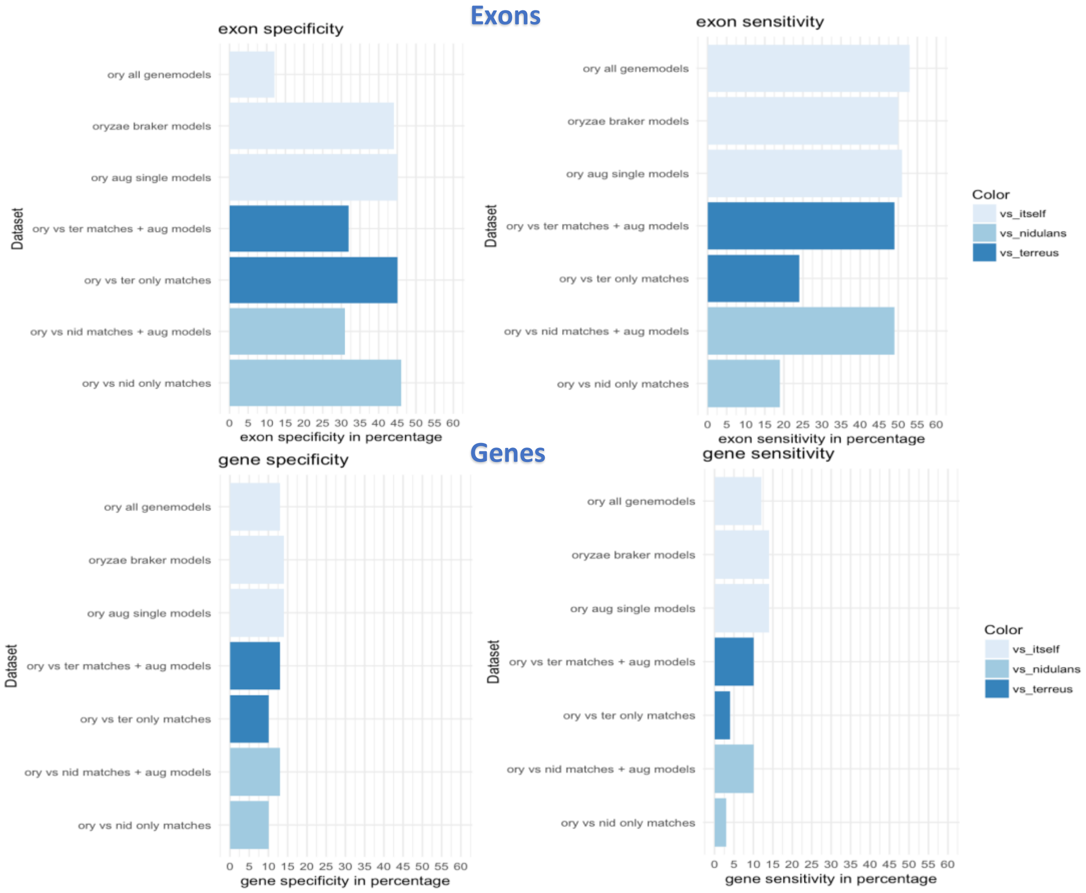


Figure 8, comparison of gene models of A. oryzae to the golden standard. The two graphs on top represent the exon location comparison, the two graphs on the bottom represent the gene location comparison. The two graphs on the left are the gene specificity and the two graphs on the right represent the sensitivity. Ory all gene models represent all predicted gene models for A. oryzae, so including all alternative models predicted by the sampling algorithm of Augustus. Oryzae braker models are the models predicted by BRAKER1, supported by RNA-Seq data. Ory aug single models are the models predicted by Augustus by the standard Viterbi algorithm. Ory vs ter matches + aug models represents the improved gene models of A. terreus supported by BRAKER1. But the first gene models predicted by Augustus that did not match with BLASTP are added to this dataset. The same for ory vs nic matches + aug models., but here the gene models of A. nidulans are used. The last category is ory vs ter only matches, this indicates that only the Blast matches are included. The same for ory vs nid only matches, only with A. nidulans as organism. All statistical details are found in appendix, Table1

The gene models: ‘ory all genemodels’ has the lowest exon specificity and the highest exon sensitivity. Apart from that, the ‘oryzae braker models’ and ‘ory aug single models’ have the highest exon sensitivity and specificity. Overall the gene models created with Augustus, Viterbi algorithm, all alternative gene models (sampling algorithm) and gene models predicted by BRAKER1 have the highest percentages.

When comparing the gene models: ‘ory vs the other organisms’ (*A. terreus or A. nidulans*) with each other. The ‘ory vs the other organisms’ with only blast matches have the highest specificity while ‘ory vs the other organisms’ with also non-blast matches (the best gene model of other Augustus models) have the highest sensitivity.

Discussion

The goal of this project was to improve gene annotation with the support of RNA-Seq data from a related genome. A pipeline was developed that creates gene models for two different but closely related species. BRAKER1 is used to create gene models for the species with RNA-Seq data and Augustus is used for the other species. These gene models are being compared with BLASTP (Altschul et al., 1990) the Blast matches are selected to get the best corresponding gene model. For validating the pipeline another related organism was used, *A. nidulans* (with RNA-Seq data).

We compared the annotation of *A. terreus* created by BRAKER1 with the annotation of *A. terreus* created by Augustus. According to BRAKER1 (Hoff et al., 2016b), BRAKER1 outperforms Augustus. By manual inspection it has been shown that BRAKER1 improves Augustus gene models (Hoff et al., 2016b), but these gene models are still not perfect, compared to the official annotation of *A. terreus*. So, these results show that the gene models from BRAKER1 are closer to the reference, but these gene models are still not perfect.

After creating gene models for *A. terreus* (BRAKER1) and for *A. oryzae* the translation between the gene models wasmade. The number of gene models has a big difference between the ‘with alternatives’ and ‘without alternatives’ parameter, as seen in Figure 5. But, in the end, around the same number of best-bidirectional hits are found. It could be interesting to take a closer look at the differences between these models. So comparing the gene models with alternatives from BRAKER (or Augustus) to gene models without alternatives from BRAKER (or Augustus). This was manually done by randomly choosing some gene models. When comparing the gene models with alternatives for each gene to the gene models with only one gene model for each gene, the first gene model of the genes with alternatives was not equal to the best gene model of the gene models with only one gene. The ‘best’ gene model with alternatives was not the same as the best (and only) of the gene models without alternatives. As mentioned before, BRAKER1 has a better performance than Augustus. We assume that, in general, the highest scoring model of BRAKER1 is more reliable than the highest scoring model of Augustus. We expect to have the first model of all alternative models of BRAKER1 to be the best hit with one of the alternative models of Augustus. This is the case, in 22% of the cases the alternative model of Augustus had a better hit with BRAKER1.

How well does this pipeline perform? Firstly, the pipeline was validated by comparing the gene models of *A. oryzae* supported by *A. terreus*, and the gene models of *A. oryzae* supported by *A. nidulans* with the reference, the official annotation of *A. oryzae*. The results show that the gene specificity and sensitivity is very low. This means the start and stop positions of the predicted genes differ from the start and stop positions of the reference genes. On the other hand, the exon specificity and sensitivity are quite high comparing to the gene sensitivity and specificity. Although we expected all of these percentages to be higher, because these gene models are supported by gene models from BRAKER1. We compared these results to the results from BRAKER1 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016). BRAKER1 shows a gene sensitivity and specificity between 52 and 77 percent on four different well annotated organisms. The exon sensitivity and specificity is even higher, it lies between 75 and 83 percent. After all, our gene models are not predicted by BRAKER1 but with Augustus. These gene models are improved with hints from gene models of a related organism generated with BRAKER1. So, we did not expect the high percentages of the results from BRAKER1. But we did expect a better sensitivity and specificity. Secondly, three extra comparisons were made to the reference, namely; gene models with all alternatives, gene models predicted by BRAKER1 (with RNA-Seq) and gene models predicted by Augustus without alternatives, so with the Viterbi algorithm. The results are not that different from previous results. The sensitivity and specificity of the genes are still low. Also for the exon sensitivity and specificity, the results are not that different. But, according to this comparison, the gene models generated by Augustus, Viterbi algorithm (single models), have the highest sensitivity and specificity in almost all comparisons. Even when comparing to the gene models generated by BRAKER1. While BRAKER1 uses RNA-Seq data in generating these models. Z

How is this possible? A few gene models were checked manually. Figure 8, Figure 9 and Figure 10 in supplementary are some examples that notable. Overall, a lot of gene models were correctly predicted at the start position, but the stop position was not correctly predicted. There were also some shifts in exons. These exons completely absent or too long or short (Figure 8 and Figure 10). Also, in some examples, Augustus predicted one gene model while it should have been two gene models, Figure 8. In this same example, the gene model of BRAKER1 was incorrectly prediced, even though the RNA-Seq data shows this should be two gene models. So, during prediction of these gene models, many mistakes were made. There were even examples (Figure 9) where the reference annotation did not have a gene model while Augustus and BRAKER do predict a gene model. These gene models are convincing when looking at the RNA-Seq coverage. The question is, how reliable is the reference genome? When comparing gene model predictions (with error) with the reference genome (with error) the chance of getting good results is minimal. So, how reliable is this validation? In Future work it will be continued.

What is the contribution to the current and already existing gene prediction tools? Our pipeline is unique compared to existing gene prediction tools because it uses RNA-Seq supported gene models from a related organism. Augustus is currently the most reliable tool for prediction of gene models, when there is no RNA-Seq data available. RNA-Seq data is not available for all organisms. This pipeline can be useful for the prediction of gene models, when RNA-Seq data is missing, supported by RNA-Seq data of a related organism.

Future work

For future prospect, how to improve this pipeline and how to improve the validation? First, the validation. The gene and exon sensitivity and specificity were low. The possible reasons for these results are a poor gene prediction by Augustus and/or a poor reference annotation. According to the results, the Augustus (Viterbi) models had the biggest overlap with the reference annotation. It would be an option to not use the first of the Augustus (sampling) models but to use the gene models of Augustus (Viterbi) in combination with the improved gene models (based on BLASTP). This could give a better overlap with the reference because the first model of the Augustus sampling models was not necessarily the same as the gene model generated by the Augustus Viterbi algorithm. Probably, the reference genome is poorly predicted (Machida et al., 2005). According to (Machida et al., 2005) the *A. oryzae* genome was annotated based on homologies to known genes in public databases of *A. oryzae* and *Aspergillus flavus*, in combination with statistical features of gene-finding software (Machida et al., 2005). This annotation is old and vaguely described. So, for validation in the future, running the pipeline on a well annotated organism could give better gene and exon sensitivity and specificity.

How to improve this pipeline for future prospect? The first recommendation is that this pipeline now only takes two species. One of these species does have accompanying RNA-Seq data while the other species does not contain RNA-Seq data. If the tools would use more than one organism supported by RNA-Seq data it could improve the annotation of the gene models without RNA-Seq support. To make the translation between the organisms it would be possible to make a multiple sequence alignment instead of BLASTP. For example, make a MSA and choose the sequence that has the highest score, so the sequence with the most similarity. Also, to determine what gene model of which organism is the most reliable, RNA-Seq coverage can be taken into account. The more RNA-Seq coverage, the more reliable the BRAKER1 gene models are. If the gene models created by BRAKER1 are very reliable, the gene models have more power to improve the gene models created by Augustus of the related species. But, how to combine these two improvements? It might be an idea to give a weight to gene models predicted by BRAKER1 based on the RNA-Seq coverage. If there is a higher coverage assign a higher weight and vice versa. When doing the MSA, first check the sequences with the most similarity. If this is also the sequence with the highest weight, choose this gene model as improved gene model. It could be an idea to create a model, by multiplying the distance scoring by the weight and choose the highest number as most reliable gene model.

The translation between the two organisms was done with BLASTP. As previously described, some gene models are missing when there is not a significant match with BLASTP. Maybe the parameters of BLASTP are too strict. For future research, it could be interesting use more flexible parameters. Due lack of time we did not test what would happen if the parameters are less strict or even more strict. The parameters should change when the distance between genomes change. The parameters should be more flexible when the genomes are more dissimilar.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

American Society for Microbiology., O., Trigos, Á., Ramos-Balderas, J. L., Viniegra-González, G., Deising, H. B., & Aguirre, J. (2002). *Eukaryotic cell.* American Society for Microbiology. Retrieved from http://ec.asm.org/content/6/4/710/F3.expansion.html

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables (2016). gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1. https://CRAN.R-project.org/package=gplots

Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., & Wortman, J. R. (2011). Approaches to Fungal Genome Annotation. *Mycology*, *2*(3), 118–141. https://doi.org/10.1080/21501203.2011.606851

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016a). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. *Bioinformatics*, *32*(5), 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016b). Genome analysis BRAKER1 : Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS, *32*(November 2015), 767–769. https://doi.org/10.1093/bioinformatics/btv661

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Chen, E., Yang, C., Zhu, J., Jayaraman, P., De Pons, J., … Lu, Y. (2015). Improved rat genome gene prediction by integration of ESTs with RNA-Seq information. *Bioinformatics*, *31*(1), 25–32.

https://doi.org/10.1093/bioinformatics/btu608

Li, Z., Zhang, Z., Yan, P., Huang, S., Fei, Z., & Lin, K. (2011). RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics*, *12*(1), 540. https://doi.org/10.1186/1471-2164-12-540

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, *42*(15), e119. https://doi.org/10.1093/nar/gku557

Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., … Kikuchi, H. (2005). Genome sequencing and analysis of Aspergillus oryzae. *Nature*, *438*(7071), 1157–1161. https://doi.org/10.1038/nature04300

Phys, A., Anderson, J. L., Ryon, J. W., & Forney, G. D. (1973). The Viterbi Algorithm. *PROCEEDINGS OF THE IEEE Phys. Phys. Rev. Phys. Reu. J. Math. Phys. Faraday, Experimental Researches in Electricity*, *6134*(167), 274–276. Retrieved from https://pdfs.semanticscholar.org/e926/9aa3243d9ef2a28a54f53551a6fafc29c333.pdf

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SRA Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Download Guide. 2009 Sep 9 [Updated 2016 Jan 14]. Available from https://www.ncbi.nlm.nih.gov/books/NBK242621/

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, *33*(Web Server issue), W465-7. https://doi.org/10.1093/nar/gki458

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329–42. https://doi.org/10.1038/nrg3174

All programming code can be found on:
A.G.R. Peeters, Master-thesis,(2017),GitHub respository: https://github.com/Teuntje/Master-thesis

Location of script made by Lucas for gene comparison:
/local/work/peete050/Thesis_final/scripts/comparison_genemodels/official_comparison/evaluate_accuracy_2.py

Supplement

Manual Augustus training organism

1. Create your own species
Go to [PATH to augustus]/config/species/

```
mkdir [myspecies]
cd [myspecies]
```

2. Copy file to your species
You should copy the generic files into the folder of your own species

```
cp -a [PATH to augustus]/config/species/generic/. ./
```

Edit all these files with my_species instead of generic in front of it with:

```
for f in *generic*; do mv "$f" "${f/generic/aspergillus_flavus}"; done
```

You should also edit the parameters file. Change the generic with the name of your organism with:

```
python /home/peete050/Thesis/scripts/change_generic_files.py generic_parameters.cfg [name organism]
```

3. Training augustus
You should train augustus with the parameters. But these parameters have to be optimised.

```
export PATH=$PATH:/home/peete050/Programs/augustus-3.2.3/config/
```

3.1 Optimising, run optimising the parameters

/home/peete050/Programs/augustus-3.2.3/scripts/optimize_augustus.pl —species=[myspecies] —-metapars=/home/peete050/Programs/augustus-3.2.3/config/species/[my species]/myspecies_metapars.cfg —aug_exec_dir=**‘Location + genbankfile’**

Example:

```
screen /home/peete050/Programs/augustus-3.2.3/scripts/optimize_augustus.pl --species=aspergillus_flavus --metapars=/home/peete050/Programs/augustus-3.2.3/config/species/aspergillus_flavus/aspergillus_flavus_metapars.cfg --aug_exec_dir=/home/peete050/Programs/augustus-3.2.3/bin/ --AUGUSTUS_CONFIG_PATH=/home/peete050/Programs/augustus-3.2.3/config/ GCF_000006275.2_JCVI-afl1-v2.0_genomic.gb
```

This made files myspecies_parameters.cfg.orig1, myspecies_parameters.cfg.orig2, myspecies_parameters.cfg.orig3 , myspecies_parameters.cfg.orig4, myspecies_parameters.cfg.orig5 in the AUGUSTUS_CONFIG_PATH directory. The final parameters are put into myspecies_parameters.cfg.

3.2 Now training augustus with the created parameters using etraining

```
/home/peete050/Programs/augustus-3.2.3/bin/etraining --species=myspecies [genbankfile]
```

Example:

```
/home/peete050/Programs/augustus-3.2.3/bin/etraining --species=aspergillus_flavus AUGUSTUS_CONFIG_PATH=/home/peete050/Programs/augustus-3.2.3/config/ GCF_000006275.2_JCVI-afl1-v2.0_genomic.gb
```

3.3 Check how accurate your prediction was

```
/home/peete050/Programs/augustus-3.2.3/bin/augustus --species=species [genbankfile]
```

Example:

```
/home/peete050/Programs/augustus-3.2.3/bin/augustus --species=aspergillus_terreus2 test.gb > testfile_results.gff
```

Figures

Workflow

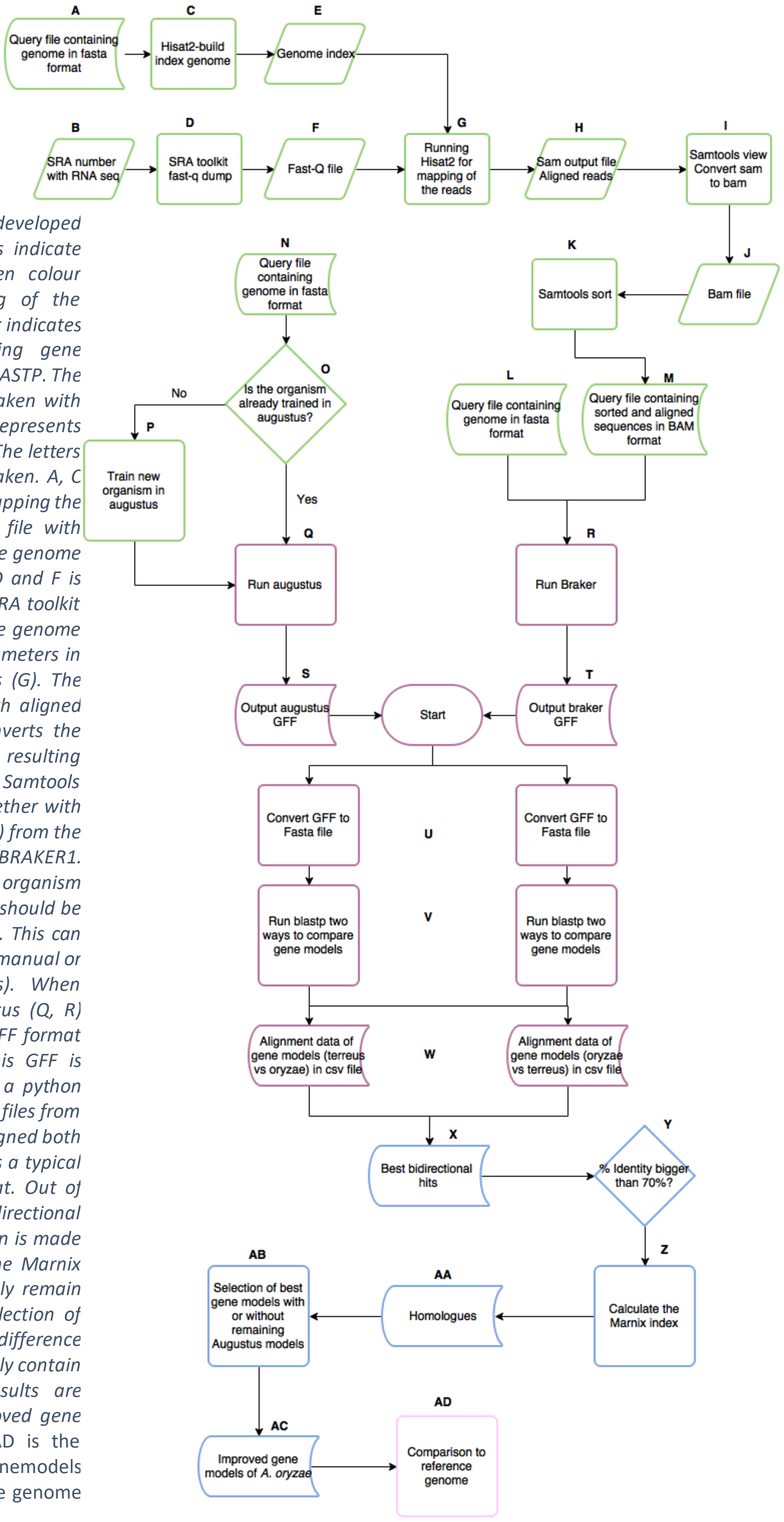

A
Query file containing genome in fasta format
C
Hisat2-build index genome
E
Genome index
B
SRA number with RNA seq
D
SRA toolkit fast-q dump
F
Fast-Q file
G
Running Hisat2 for mapping of the reads
H
Sam output file Aligned reads
I
Samtools view Convert sam to bam
J
Bam file
K
Samtools sort
N
Query file containing genome in fasta format
O
Is the organism already trained in augustus?
No
P
Train new organism in augustus
Yes
Q
Run augustus
L
Query file containing genome in fasta format
M
Query file containing sorted and aligned sequences in BAM format
R
Run Braker
S
Output augustus GFF
Start
T
Output braker GFF
Convert GFF to Fasta file
U
Convert GFF to Fasta file
Run blastp two ways to compare gene models
V
Run blastp two ways to compare gene models
Alignment data of gene models (terreus vs oryzae) in csv file
W
Alignment data of gene models (oryzae vs terreus) in csv file
X
Best bidirectional hits
Y
% Identity bigger than 70%?
Z
Calculate the Marnix index
AA
Homologues
AB
Selection of best gene models with or without remaining Augustus models
AC
Improved gene models of *A. oryzae*
AD
Comparison to reference genome

Figure 12, Overall workflow of developed pipeline. The different colours indicate different processes. The green colour represents the pre-processing of the desired data. The purple colour indicates generating the data, creating gene models and alignments with BLASTP. The blue colour shows the steps taken with the results. The pink colour represents the validation of the pipeline. The letters represent the steps that are taken. A, C and E are steps taken before mapping the reads. At first a query fasta file with genome. Hisat2-build builds the genome index from the fasta file. B, D and F is collecting the RNA-Seq data, SRA toolkit downloads the FAST-Q file. The genome index and FAST-Q file are parameters in Hisat2 for mapping the reads (G). The resulting file is a sam file with aligned reads (H). Samtools view converts the sam file to a bam file (I). The resulting bam file (J) is being sorted with Samtools (K). The resulting file (M), together with the genome in fasta format (M) from the same organism are used to run BRAKER1. For running Augustus (Q), the organism needs already to be trained or should be trained to run Augustus (O, P). This can be done with the handwritten manual or script on GitHub (references). When running BRAKER1 and Augustus (Q, R) the result is gene models in GFF format for both organism (S, T). This GFF is converted to a fasta file with a python script (GitHub) (U). These fasta files from two different organisms are aligned both ways in Blast (V). The results is a typical Blast output (W) in csv format. Out of these two csv files the best bi-directional hits are selected (X). A selection is made to remain homologues (Y). The Marnix index is calculated (Z). We only remain homologues (AA). Another selection of best gene models is done, the difference is that this data do or do not only contain Blast matches (AB). The results are different data sets with improved gene models of A. oryzae (AC). AD is the validation of the tool. These genemodels are compared to the reference genome to see how similar these are.

Figure 2

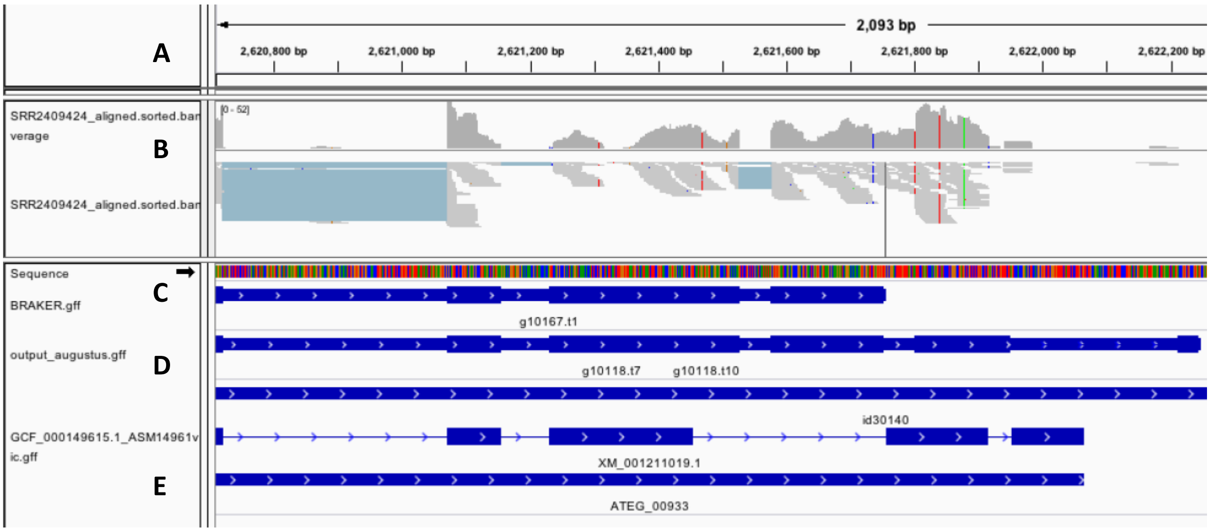


Figure 2, IGV viewer, gene is: ATEG_00933 neuronal calcium sensor 1. Organism: A. terreus. A shows the positions on the genome. B represents the RNA-Seq reads mapped on the genome and the coverage. C represents the gene models created by BRAKER1. D shows the gene models created by Augustus. E is the genome annotation from NCBI

Figure 8

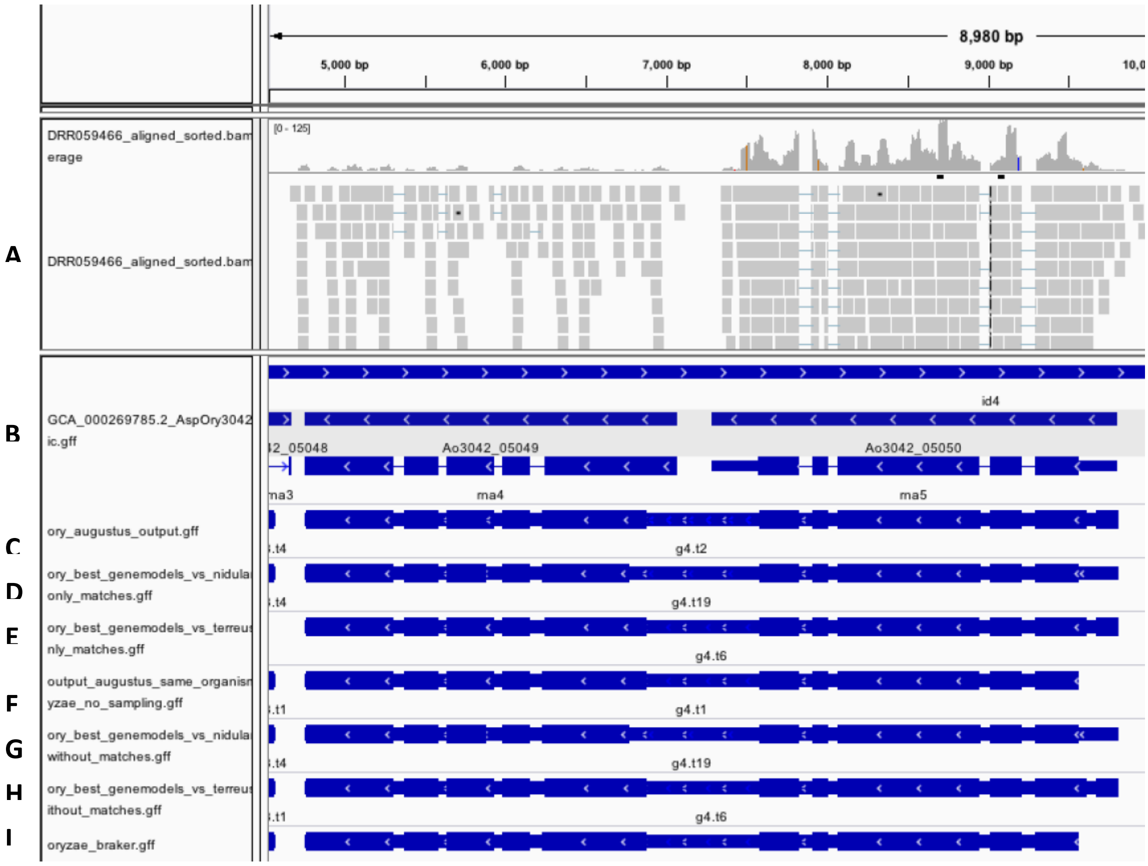


Figure 8, Comparison of golden standard to the different gene models. A is the RNA-Seq reads, B is the golden standard. It consists of three lines, the first is the region, the second the genes and the third are the intron, exon and UTR locations. C is the output of Augustus with all alternative models included. D are only the blast matches between A. nidulans and A. oryzae, so only the improved gene models. E is the same but with A. terreus instead of A. nidulans. F is the output of Augustus with the Viterbi algorithm, here are no alternatives present, one gene model for each gene. G are blast matches with A. nidulans plus the non-blast match gene models (first gene model) predicted by Augustus with the sampling algorithm. H is the same but for A. terreus instead of A. nidulans. I represents the results created by BRAKER1, with the back-up of RNA-Seq data.

Figure 9

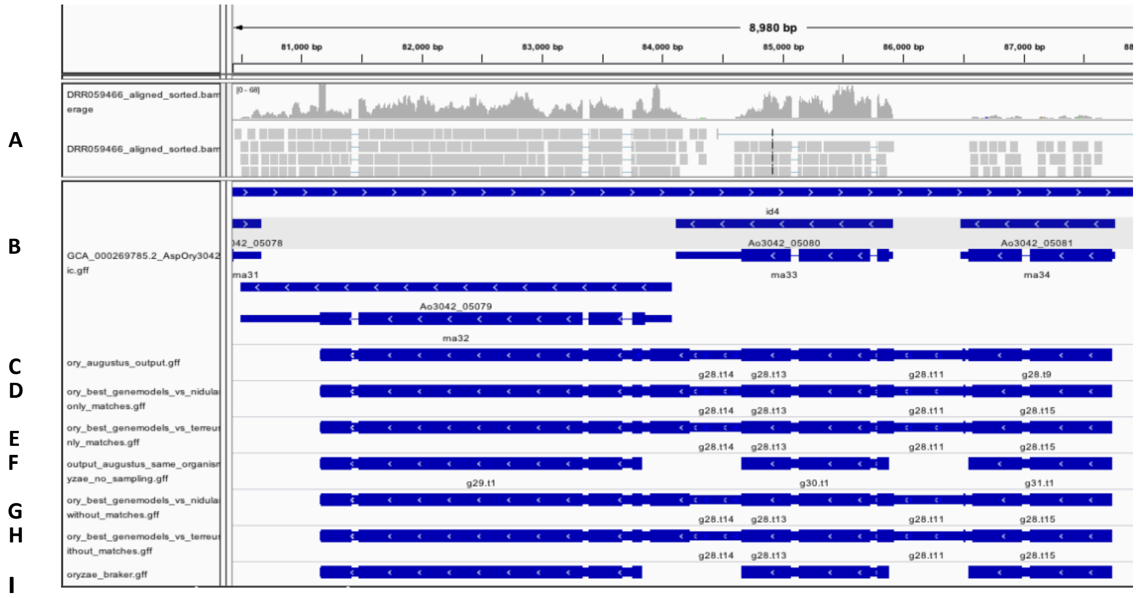


Figure 9, Comparison of golden standard to the different gene models. A is the RNA-Seq reads, B is the golden standard. It consists of three lines, the first is the region, the second the genes and the third are the intron, exon and UTR locations. C is the output of Augustus with all alternative models included. D are only the blast matches between A. nidulans and A. oryzae, so only the improved gene models. E is the same but with A. terreus instead of A. nidulans. F is the output of Augustus with the Viterbi algorithm, here are no alternatives present, one gene model for each gene. G are blast matches with A. nidulans plus the non-blast match gene models (first gene model) predicted by Augustus with the sampling algorithm. H is the same but for A. terreus instead of A. nidulans. I represents the results created by BRAKER1, with the back-up of RNA-Seq data.

Figure 10

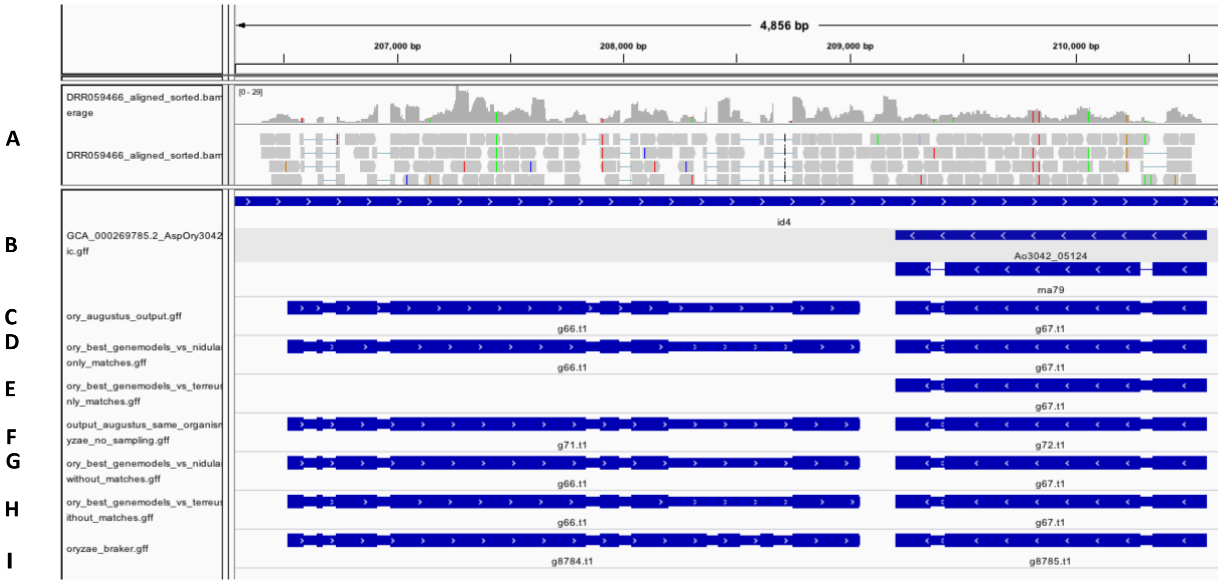


Figure 10, Comparison of golden standard to the different gene models. A is the RNA-Seq reads, B is the golden standard. It consists of three lines, the first is the region, the second the genes and the third are the intron, exon and UTR locations. C is the output of Augustus with all alternative models included. D are only the blast matches between A. nidulans and A. oryzae, so only the improved gene models. E is the same but with A. terreus instead of A. nidulans. F is the output of Augustus with the Viterbi algorithm, here are no alternatives present, one gene model for each gene. G are blast matches with A. nidulans plus the non-blast match gene models (first gene model) predicted by Augustus with the sampling algorithm. H is the same but for A. terreus instead of A. nidulans. I represents the results created by BRAKER1, with the back-up of RNA-Seq data.

Figure 11

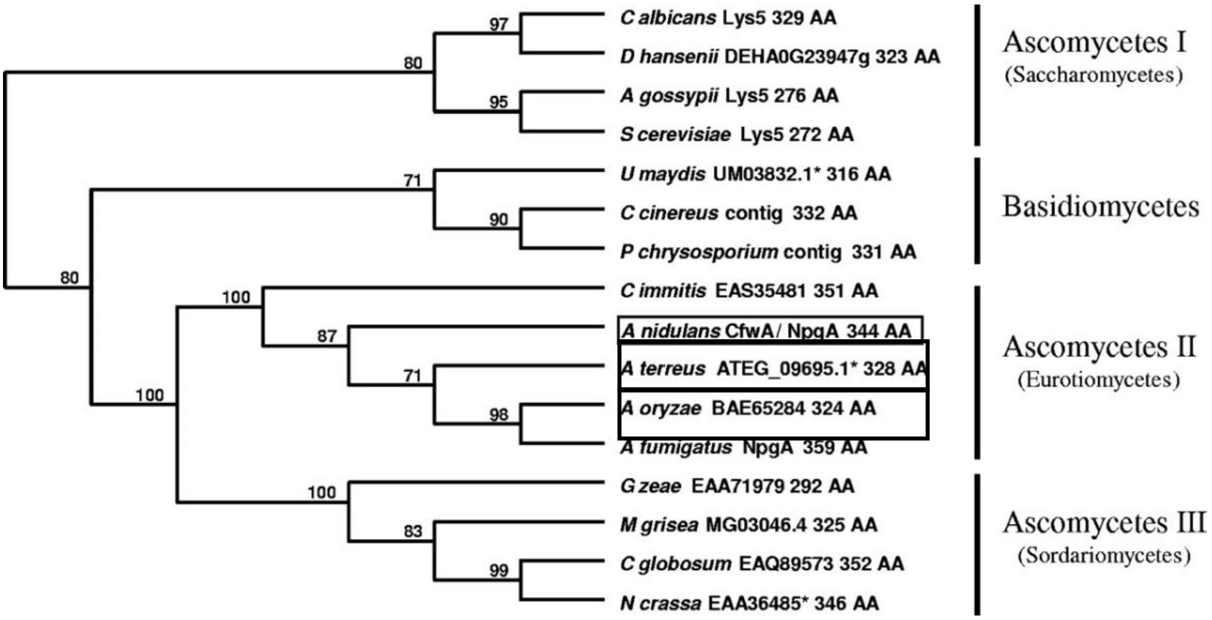


Figure 11, distance between A. nidulans, A. terreus and A. oryzae. (American Society for Microbiology. et al., 2002)

Table1

Table 1, comparisons of gene models

Name	actual genes	predicted genes	correc genes	gene sens	gene spec	actual exons	predicted exons	wrong exons	missing exons	correct exons	exon sens	exons spec	missing exon sens	wrong exon spec
oryzae vs nidulans only blast matches	11640	3232	32	0.03	0.1	35056	14470	7807	27935	6663	0.19	0.46	0.8	0.54
oryzae vs nidulans blast matches + augustus genes	11640	10739	1396	0.1	0.13	35056	54435	37321	17931	17114	0.49	0.31	0.51	0.69
oryzae vs terreus only blast matches	11640	4313	460	0.04	0.1	35056	18841	10433	26335	8408	0.24	0.45	0.75	0.55
oryzae vs terreus blast matches + augustus genes	11640	10739	1396	0.1	0.13	35056	52554	35474	17965	17080	0.49	0.32	0.51	0.68
oryzae augustus single models	11640	11481	1616	0.14	0.14	35056	39745	22017	17317	17728	0.51	0.45	0.49	0.55
oryzae braker models	11640	11421	1609	0.14	0.14	35056	40310	22753	17486	17557	0.5	0.44	0.5	0.56
oryzae all genemodels of augustus (with alternatives!)	11640	10739	1396	0.12	0.13	35056	157505	138811	16351	18694	0.53	0.12	0.47	0.88