ReadMe - TGAC_Chalara_fraxinea_ass_s1v1_ann_v1.1

Repeat Masking

RepeatModeler v1-0-7 was used to generate a species specific repeat library based on the TGAC s1v1 Chalara assembly. Interspersed repeats were identified using the Chalara fraxinea repeat library and RepeatMasker. Low complexity repeats were identified with RepeatMasker and tantan http://www.cbrc.jp/tantan/.

GFF files:
/Repeats/RM_int_repeats.gff
/Repeats/RM_loc_complex_repeats.gff
/Repeats/tantan_low_complex_repeats.gff

Evidence alignments:

Proteins

Protein sequences from 4 species Amorphotheca resinae v1.0, Botrytis cinerea v1.0, Oidiodendron maius Zn v1.0 and Sclerotinia sclerotiorum v1.0 (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=leotiomycetes) were soft masked for low complexity (segmasker) and aligned to the soft masked TGAC s1v1 Chalara assembly with exonerate protein2genome, alignments were filtered at a minimum 50% identity and 50% coverage.

GFF files:
Source/Amore1_GeneCatalog_proteins_20110719.aa.mask.exonerate.50id-50cov.gff
Source/Botci1_GeneCatalog_proteins_20110903.aa.mask.exonerate.50id-50cov.gff
Source/Oidma1_GeneCatalog_proteins_20110606.aa.mask.exonerate.50id-50cov.gff
Source/Sclsc1_GeneCatalog_proteins_20110903.aa.mask.exonerate.50id-50cov.gff

RNA-Seq

The public RNA-Seq reads were aligned to the unmasked TGAC s1v1 Chalara assembly with Tophat v2.0.4 (--min-anchor-length 12 --max-multihits 20 --min-intron-length 50 --max-intron-length 10000 --min-coverage-intron 50 --max-coverage-intron 5000 --min-segment-intron 50 --max-segment-intron 10000).

reads:
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R1_001.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R1_002.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R1_003.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R1_004.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R1_005.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R1_006.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R2_001.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R2_002.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R2_003.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R2_004.fastq.gz

ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R2_005.fastq.gz
ftp-oadb.tsl.ac.uk/mixed_material/ashwellthorpe_AT2/RNA_seq/lane5_NoIndex_L005_R2_006.fastq.gz

RNA-Seq data contains 23,096,021 PE reads of 76 bp of these 3,623,719 (16%) read pairs have at least 1 read of the pair mapping to the s1v1 Chalara assembly assembly subset (6,255,701 positions mapped 67% properly paired)

Splice junctions identified with Tophat were filtered requiring support from a minimum of 4 reads.

Junction file (unfiltered):
Source/jasmbl_filtered.gbrowse.gff

Bigwig file of RNA-Seq read density:
Source/accepted_hits.sort.bam.update.bw

RNA-Seq alignments were assembled with Cufflinks v2.0.2 (-I 10000 -A 0.15) into 11804 contigs.

stats_info	bases	21858299
stats_info	contigs	11804
stats_len	max	17863
stats_len	mean	1851.77
stats_len	median	1492
stats_len	min	86
stats_len	mode	769
stats_len	modeval	13
stats_len	range	17778
stats_len	stddev	1430.24

Longest ORFs were selected:
Source/cufflinks_transcripts_cds.gff

The public Trinity Assembly of the same RNA-Seq data (AT2_trinity_version2) was softmasked (dustmasker) and aligned to the soft masked TGAC s1v1 Chalara assembly with exonerate est2genome and filtered at a minimum of 95% identity, 50% coverage.

GFF file:
Source/AT2_trinity_version2.mask.fasta.exonerate.95id-50cov.gff

Augustus Training and genebuild

Augustus v2.5.5 was trained for Chalara_fraxinea using a subset of cufflinks assemblies. Briefly, blast searches against the 4 protein datasets (see above) were used to identify predicted cufflinks CDS features with support from cross species alignments, these were further filtered to remove features showing >80% sequence similarity within the selected subset and/or any genomic overlap. Augustus was trained based on the filtered set of 859 cufflinks models with 100 models reserved for testing. Based on this the ab initio predictions

achieved sensitivity results of 0.96 nucleotide, 0.80 exon and 0.65 at the gene level.

Augustus gene models were predicted using the trained ab initio model with the 4 cross species protein alignments, RNA-Seq junctions, Cufflinks alignments, and Aligned Trinity contigs as evidence hints. RNA-Seq read density was provided as exon hints and repeat information as nonexonpart hints.
Augustus gene models were filtered to remove redundant genes giving a total of 10252 predicted genes with 10405 predicted mRNAs.

GFF file:
Gene_predictions/TGAC_Chalara_fraxinea_ass_s1v1_ann_v1.1
/Chalara_fraxinea_ass_s1v1_ann_v1.1.gene.gff

Genemodels are assigned a uniq identifier of the following form

CHAFR746836.1.1_0000010.1
(uniq species code incorporating NCBI taxonomy id CHAFR746836, 1.1. annotation version, 0000010 gene identifier, .1 transcript identifier)

Gene Stats	Total (bp) 23087543	Total Count 10252	Mean size (span) 2252
mRNA Stats	Total (bp) 23462930	Total Count 10405	Mean size (span) 2254.97
Exon Stats (distinct)	Total (bp) 20614305	Total Count 34735	Mean size 593.474
Intron Stats (distinct)	Total (bp) 1849527	Total Count 22347	Mean size 82.764
CDS Exon Stats (distinct)	Total (bp) 15661318	Total Count 32568	Mean size 480.881
5UTR Exon Stats	Total (bp) 2402584	Total Count 11954	Mean size 200.986
3UTR Exon Stats	Total (bp) 2598484	Total Count 10758	Mean size 241.54

cDNA transcripts stats
Mean length 2000.53983661701 Median 1725 Min 204 Max 23327 Total Length 20815617

CDS transcript stats
Mean length 1519.89082172033 Median 1272 Min 84 Max 22749 Total Length 15814464

Maker genebuild

An alternative genebuild was generated with maker-2.10 using the above evidence files and trained augustus model.

GFF file:
Gene_predictions/TGAC_Chalara_fraxinea_ass_s1v1_ann_v1.1
/maker.genes.utr.gff

A Gbrowse instance is available at
http://browser.tgac.bbsrc.ac.uk/cgi-bin/gb2/gbrowse/Chalara_fraxinea_ass_s1v1_ann_v1.1