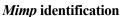
Fusarium Assemblies

Fusarium species genome assemblies in FASTA format, as well as a list of the assemblies and a mimp profile HMM, are prepared as input.



Each Fusarium assembly is searched for mimps using a custom python script (using mimp TIRs) and NHMMER (3.3.1) (using a mimp profile-HMM).

Mimp region sequence identification

Identify regions 2.5kb either side of a mimp (mimp region), generating a mimp region GFF file for each genome assembly.

Augustus regions identification

Expand 20kb either side of the *mimp* regions to identify regions for AUGUSTUS (v3.3.3) annotation. Generate an Augustus (v3.3.3) region GFF, and FASTAs where all non-Augustus have been hard masked.

Mask non-mimp regions

Using the *mimp* region GFF, create a *mimp* region FASTAS, where all non-mimp regions have been hard masked.

AUGUSTUS gene prediction

The Augustus region FASTA is submitted to AUGUSTUS (V3.3.3) for gene prediction with the "fusarium" species parameter selected.

ORF Prediction

Getorf from Emboss (v6.6.0.0) is used to find open reading frames (ORFs) in the 2.5kb region.

Signal peptide filtering

SignalP (4.1) is used search all AUGUSTUS gene models and ORFs for a signal peptide.

Size filter

Protein sequences containing a signal peptide predicted by SignalP (v4.1) are filtered based on size, with sequences <450aa and >30aa kept for effector prediction.

EffectorP (v2.0.1) scan for likely effectors

Each signal peptide and size filtered sequences is submitted to EffectorP (v2.0.1) for fungal effector prediction.

FASTA and **GFF** generated

For each Fusarium assembly included, a candidate effector FASTA and GFF file is generated using various custom python scripts..



Candidate effector clustering

Candidate effector sequences are from all assemblies are combined into one FASTA and clustered using CD-HIT (4.8.1) (80% identity (optional using command line input)).

Cluster analysis

A custom python script is used to generate a table of effector clusters, containing details on the sequences in each cluster, the assembly from which each sequence originated, and the percentage identity of each sequence within the cluster to the cluster reference.

Candidate effector profiling

An effector profile generated in R Studio (version 3.6.3), using the package Pheatmap (version 1.0.12). Data can be converted to binary at this stage to only generate a presence/absence heatmap rather than hit frequency heatmap.



