Sequence identification

Effector Prediction

Fusarium Assemblies

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

Mimp identification

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.

Augustus regions identification

Expand 20kb either side of the *mimp* regions to identify regions for Augustus annotation. Generate an Augustus 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 (3.3.3) for gene prediction with the "fusarium" species parameter selected.

ORF Prediction

Getorf from Emboss (6.6.0.0) 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 are filtered based on size, with sequences <450aa and >30aa kept for effector prediction.

EffectorP Scan for Likely Effectors

Each signal peptide and size filtered sequences is submitted to EffectorP (2.0.1) 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..

Effector Profiling

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)).

***Fusarium* pan-effectorome**

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.

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.

Fusarium Effector Cluster Profiles