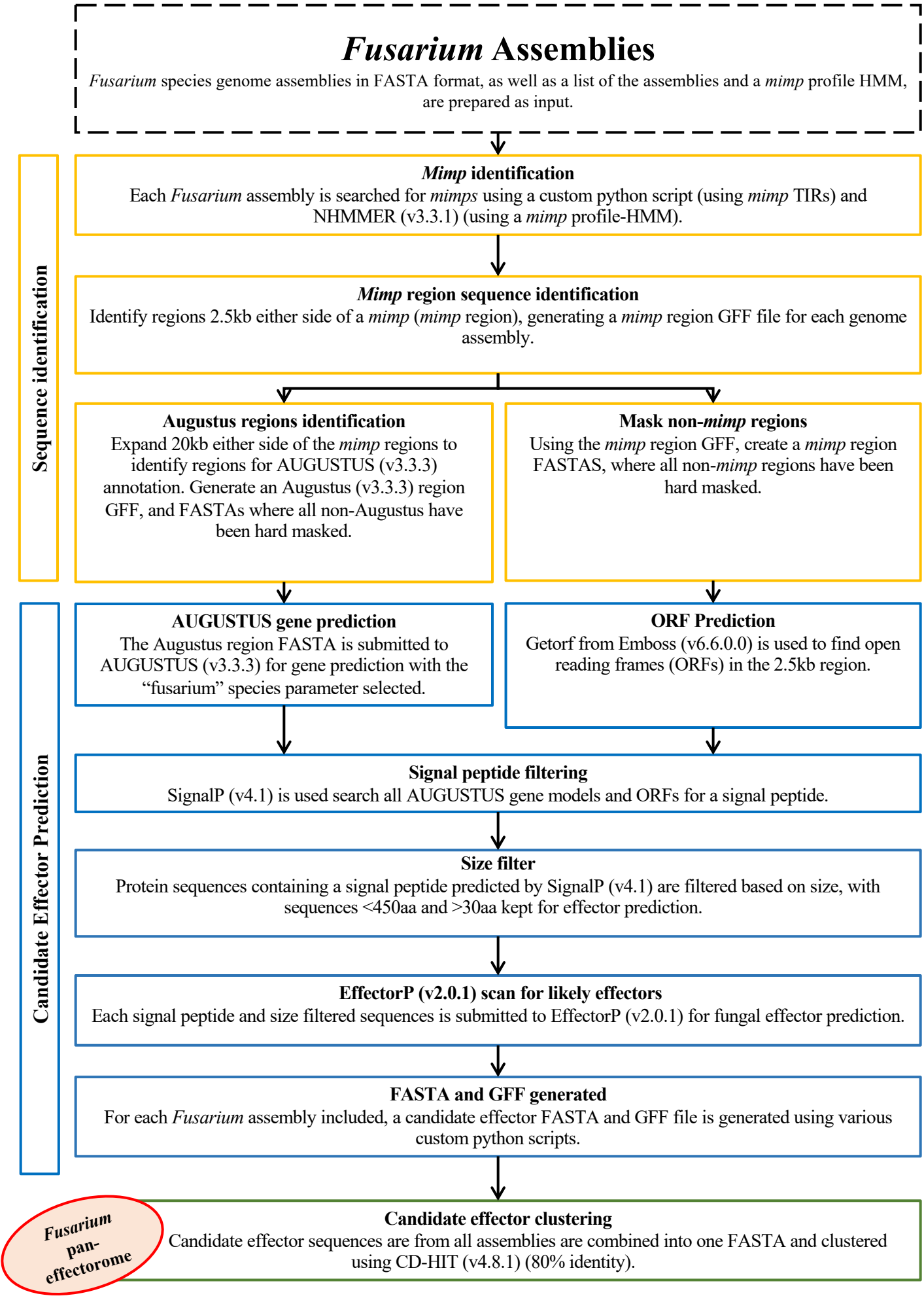
Fusarium Assemblies
Fusarium species genome assemblies in FASTA format, as well as a list of the assemblies and a mimp profile HMM, are prepared as input.
Sequence identification
Mimp identification
Each Fusarium assembly is searched for mimps using a custom python script (using mimp TIRs) and NHMMER (v3.3.1) (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.
Candidate Effector Prediction
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 (v4.1) 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 (v4.8.1) (80% identity).
Fusarium pan-effectorome