Long title:

Genome mining of the citrus pathogen *Elsinoë fawcettii*; prediction and prioritisation of candidate effectors, cell wall degrading enzymes and secondary metabolite gene clusters

Short title:

Genome mining of *Elsinoë fawcettii*; prediction and prioritisation of candidate virulence genes

Sarah Jeffress\(^1\), Kiruba Arun-Chinnappa\(^1\), Ben Stodart\(^2\), Niloofar Vaghefi\(^1\), Yu Pei Tan\(^3\), Gavin Ash\(^{1,2}\)

\(^1\) Centre for Crop Health, Institute for Life Sciences and the Environment, Research and Innovation Division, University of Southern Queensland, Toowoomba QLD 4350, Australia

\(^2\) Graham Centre for Agricultural Innovation, (Charles Sturt University and NSW Department of Primary Industries), School of Agricultural and Wine Sciences, Charles Sturt University, Wagga Wagga NSW 2650, Australia

\(^3\) Department of Agriculture and Fisheries, Queensland Government, Brisbane QLD 4000 Australia

Email addresses:

Sarah Jeffress: sarah.jeffress@usq.edu.au
Abstract:

_Elsinoë fawcettii_, a necrotrophic fungal pathogen, causes citrus scab on numerous citrus varieties around the world. Known pathotypes of _E. fawcettii_ are based on host range; additionally, cryptic pathotypes have been reported and more novel pathotypes are thought to exist. _E. fawcettii_ produces elsinochrome, a non-host selective toxin which contributes to virulence. However, the mechanisms involved in potential pathogen-host interactions occurring prior to the production of elsinochrome are unknown, yet the host-specificity observed among pathotypes suggests a reliance upon such mechanisms. In this study we have generated a whole genome sequencing project for _E. fawcettii_, producing an annotated draft assembly 26.01 Mb in size, with 10,080 predicted gene models and low (0.37%) coverage of transposable elements. The assembly showed evidence of AT-rich regions, potentially indicating genomic regions with increased plasticity. Using a variety of computational tools, we mined the _E. fawcettii_ genome for potential virulence genes as candidates for future investigation. A total of 1,280 secreted proteins and 203 candidate...
45 effectors were predicted and compared to those of other necrotrophic (Botrytis cinerea,
46 Parastagonospora nodorum, Pyrenophora tritici-repentis, Sclerotinia sclerotiorum and
47 Zymoseptoria tritici), hemibiotrophic (Leptosphaeria maculans, Magnaporthe oryzae,
48 Rhynchosporium commune and Verticillium dahliae) and biotrophic (Ustilago maydis) plant
49 pathogens. Genomic and proteomic features of known fungal effectors were analysed and
50 used to guide the prioritisation of 77 candidate effectors of E. fawcettii. Additionally, 378
51 carbohydrate-active enzymes were predicted and analysed for likely secretion and sequence
52 similarity with known virulence genes. Furthermore, secondary metabolite prediction
53 indicated nine additional genes potentially involved in the elsinochrome biosynthesis gene
54 cluster than previously described. A further 21 secondary metabolite clusters were
55 predicted, some with similarity to known toxin producing gene clusters. The candidate
56 virulence genes predicted in this study provide a comprehensive resource for future
57 experimental investigation into the pathogenesis of E. fawcettii.

58

59 Introduction:

60 Elsinoë fawcettii Bitancourt & Jenkins, a necrotrophic fungal species within the Ascomycota
61 phylum (class Dothideomycetes, subclass Dothideomycetidae, order Myriangiales), is a
62 filamentous phytopathogen which causes a necrotic disease, known as citrus scab, to the
63 leaves and fruit of a variety of citrus crops around the world. Susceptible citrus varieties
64 include lemon (Citrus limon), rough lemon (C. jambhiri), sour orange (C. aurantium),
65 Rangpur lime (C. limonia), Temple and Murcott tangors (C. sinensis x C. reticulata), Satsuma
66 mandarin (C. unshiu), grapefruit (C. paradisi), Cleopatra mandarin (C. reshni), clementine
67 (C. clementina), yuzu (C. junos), kinkoji (C. obovoidea), pomelo (C. grandis) and
Numerous pathotypes of *E. fawcettii* are defined by host range, including the Florida Broad Host Range (FBHR), Florida Narrow Host Range (FNHR), Tyron’s, Lemon, Jinguel, SRGC and SM, while cryptic and novel pathotypes are also reported \[1, 3, 10\]. Only the Tyron’s pathotype (which infects Eureka lemon, Rough lemon, clementine, Rangpur lime and Cleopatra mandarin) and the Lemon pathotype (which only infects Eureka lemon, Rough lemon, Rangpur lime) have been described in Australia \[2, 3, 7\], however *E. fawcettii* has reportedly been isolated from kumquat (*Fortunella* sp.), tea plant (*Camellia sinensis*) and mango (*Mangifera indica*) \[11\], indicating a wider range of pathotypes to be present in Australia. Additional species of *Elsinoë* found causing disease in Australia include *E. ampelina*, which causes anthracnose to grapes \[12\] and two *E. australis* pathotypes; one which causes scab disease to jojoba (*Simmondsia chinensis*) \[13\] and a second found on rare occasions on finger lime (*C. australasica*) in Queensland forest areas \[14\]. Species of *Elsinoë* causing crop disease in countries neighbouring Australia include *E. batatas*, which causes large yield losses in sweet potato crops in Papua New Guinea \[15, 16\] and *E. pyri*, which infects apples in organic orchards in New Zealand \[17\]. Around the world there are reportedly 75 *Elsinoë* species, the majority of which appear to be host specific \[18\]. While citrus scab is not thought to affect yield, it reduces the value of affected fruit on the fresh market. Australia is known for producing high quality citrus fruits for local consumption and export, and so understandably, there is great interest in protecting this valuable commodity from disease.

*E. fawcettii* is commonly described as an anamorph, reproducing asexually. Hyaline and spindle shaped conidia are produced from the centre of necrotic citrus scab lesions \[19, 20\]. Conidia are dispersed by water splash, requiring temperatures between 23.5-27 °C with four...
hours of water contact for effective host infection. Therefore, disease is favoured by warm
weather with overhead watering systems or rain [21]. Only young plant tissues are
vulnerable to infection; leaves are susceptible from first shoots through to half expanded
and similarly fruit for 6 to 8 weeks after petal fall, while mature plants are resistant to
disease [19]. Cuticle, epidermal cells and mesophyll tissue are degraded within 1 to 2 days
of inoculation, hyphal colonisation proceeds and within 3 to 4 days symptoms are visible
[20, 22]. After formation of necrotic scab lesions on fruit, twigs and leaves, conidia are
produced from the scab pustules providing inoculum for further spread. Within 5 days, host
cell walls become lignified separating infected regions from healthy cells, which is thought
to limit internal spread of the pathogen [20]. The necrosis that occurs during infection is
produced in response to elsinochrome, a well-known secondary metabolite (SM) of species
of *Elsinoë*. Elsinochromes are red or orange pigments which can be produced in culture [23,
24]. In aerobic and light-activated conditions, reactive oxygen species are produced in
response to elsinochromes in a non-host selective manner, generating an environment of
cellular toxicity [25]. Elsinochrome production is required for full virulence of *E. fawcettii*,
specifically the *EfPKS1* and *TSF1* genes are vital within the elsinochrome gene cluster [26,
27]. However, two points indicate that *E. fawcettii* pathogenesis is more complex than
simply the result of necrotic toxin production: (I) the production of elsinochrome appears to
be variable and does not correlate with virulence [28]; and (II) elsinochrome is a non-host
selective toxin, yet *Elsinoë* species and *E. fawcettii* pathotypes cause disease in a host
specific manner. Host-specific virulence factors targeted for interaction with distinct host
proteins to overcome immune defences, prior to elsinochrome production, could explain
the observed host specificity. Candidate virulence genes may include effectors and cell wall
degrading enzymes. Effectors are secreted pathogen proteins, targeted to either the host
cytoplasm or apoplast, which enable the pathogen to evade recognition receptor activities of the host’s defence system and, if successful, infection proceeds. Resistant hosts, however, recognise pathogen effectors using resistant (R) genes which elicit plant effector-triggered immunity and pathogenesis is unsuccessful [29, 30]. While it was previously thought that necrotrophic fungal pathogens would use only a repertoire of carbohydrate-active enzymes (CAZymes) or SM’s to infect host plants [31], there is increased awareness of their utilisation of secreted protein effectors [32-37], highlighting the importance of protein effector identification in all fungal pathogens. Frequently shared features of effectors include; a signal peptide at the N-terminal and no transmembrane helices or glycosylphosphatidylinositol (GPI) anchors. Other features less frequently shared include; small size, cysteine rich, amino acid polymorphism, repetitive regions, gene duplication, no conserved protein domains, coding sequence found nearby to transposable elements, and absence in non-pathogenic strains [38-45]. Furthermore, some appear to be unique to a species for example the necrosis-inducing protein effectors NIP1, NIP2 and NIP3 of *Rhynchosporium commune* [46] and three avirulence effectors AvrLm1, AvrLm6 and AvrLm4-7 of *Leptosphaeria maculans* [47]. Others have orthologous genes or similar domains in numerous species for example the chorismate mutase effector, Cmu1, of *Ustilago maydis* [48] and the cell death-inducing effector, MoCDIP4, of *Magnaporthe oryzae* [49]. Understandably, with such a large variety of potential features, effector identification remains challenging. Effectors are found in biotrophs, for example *U. maydis* [50-53], hemibiotrophs, such as *L. maculans* [54-56], *M. oryzae* [57, 58], *R. commune* [46] and *Verticillium dahliae* [59-61], necrotrophs, for example *Botrytis cinerea* [62, 63], *Parastagonospora nodorum* [34, 42, 64], *Pyrenophora tritici-repentis* [65], *Sclerotinia sclerotiorum* [32] and also the hemibiotroph/latent necrotroph *Zymoseptoria tritici* [66].
Genomic location has potential to be an identifying feature of virulence genes in some species, for example pathogenicity-related genes of *L. maculans*, including those coding for secreted proteins and genes potentially involved in SM biosynthesis, are found at higher rates in AT-rich genomic regions in comparison to GC-equilibrated blocks [47]. It is thought that effectors and their target host proteins co-evolve, in a constant arms race [67], presenting genomic regions with higher levels of plasticity as potential niches which harbour effector genes.

Another group of virulence factors likely to play a role in *E. fawcettii* pathogenesis are cell wall degrading enzymes (CWDE), these are CAZymes, including glycoside hydrolases, polysaccharide lyases and carbohydrate esterases, which can be secreted from fungal pathogens and promote cleavage of plant cell wall components [68-70]. Cell wall components, such as cellulose, hemicelluloses (xyloglucan and arabinoxylan) and pectin (rhamnogalacturonan I, homogalacturonan, xylogalacturonan, arabinan and rhamnogalacturonan II) [71], are targets for pathogens to degrade for nutrients and/or to overcome the physical barrier to their host. CWDE’s can include polygalacturonases, pectate lyases, and pectinesterases which promote pectin degradation [72-78], glucanases (also known as cellulase) which breaks links between glucose residues [79] and xylanases which cleave links in the xylosyl backbone of xyloglucan [80-82].

*E. fawcettii* effectors and/or CWDE’s which interact with certain host plant cell wall components could explain the observed host specificity of pathotypes. Computational prediction of genes coding for such virulence factors can lead to many candidate effectors (CE) and potential CWDE’s, leading to an overabundance of candidates which require...
prioritisation. This study aimed to generate an assembly of the *E. fawcettii* isolate, BRIP 53147a, through whole genome shotgun (WGS) sequencing, to identify candidate virulence genes and appropriately shortlist these predictions to improve the focus of future experimental validation procedures. Computational methods involving genomic, proteomic and comparative analyses enabled the prediction and prioritisation of CE’s and CWDE’s which may be interacting with the host plant and overcoming immune defences prior to the biosynthesis of elsinochrome. Additional genes potentially involved in the elsinochrome gene cluster were also predicted, as were additional SM clusters which may be impacting virulence of *E. fawcettii*.

**Materials and Methods:**

**Sequencing, assembly, gene prediction, annotation and genomic analyses:**

*E. fawcettii* (BRIP 53147a), collected from *C. limon* (L.) Burm.f. in Montville, Queensland, Australia, was obtained from DAF Biological Collections [11]. The isolate was cultured on potato dextrose agar (Difco) and incubated at 23 to 25 °C for two months. Whole genomic DNA was extracted using the DNeasy Plant Mini kit (QIAGEN) according to the manufacturer’s protocol. Paired-end libraries were prepared according to Illumina Nextera™ DNA Flex Library Prep Reference Guide using a Nextera™ DNA Flex Library Prep Kit and Nextera™ DNA CD Indexes. WGS sequencing was performed on Illumina MiSeq platform (600-cycles) at the molecular laboratories of the Centre for Crop Health, USQ. Assembly was performed on the Galaxy-Melbourne/GVL 4.0.0 webserver [83]. Raw reads were quality checked using FastQC (v0.11.5) [84] and trimmed using Trimmomatic (v0.36) [85] with the following parameters: TruSeq3 adapter sequences were removed using default
settings, reads were cropped to remove 20 bases from the leading end and 65 bases from
the trailing end of each read, minimum quality of leading and trailing bases was set to 30, a
sliding window of four bases was used to retain those with an average quality of 30 and the
minimum length read retained was 31 bases. *De novo* assembly was performed in two
steps, first using Velvet (v1.2.10) [86] and VelvetOptimiser (v2.2.5) [87] with input k-mer size
range of 81-101 (step size of 2). Secondly, SPAdes (v3.11.1) [88] was run on trimmed reads
with the following parameters: read error correction, careful correction, automatic k-mer
values, automatic coverage cutoff and Velvet contigs (>500 bp in length), from the previous
step, included as trusted contigs. Contigs >500 bp in length were retained. Reads were
mapped back to the assembly using Bowtie2 (v2.2.4) [89] and Picard toolkit (v2.7.1) [90] and
visualised using IGV (v2.3.92) [91]. The genome assembly was checked for completeness
with BUSCO (v2.0) [92] using the Ascomycota orthoDB (v9) dataset [93]. The extent and
location of AT-rich regions was determined using OcculterCut (v1.1) [94] with default
parameters and mitochondrial contigs.

The prediction of genes and transposable elements (TE) was performed on the GenSAS
(v6.0) web platform [95], using GeneMarkES (v4.33) [96] for gene prediction and
RepeatMasker (v4.0.7) [97], using the NCBI search engine and slow speed sensitivity, for the
prediction of TE’s. Predicted gene models containing short exons, missing a start or stop
codon or which overlapped a TE region were removed from the predicted proteome. The
genome was searched for Short Simple Repeats (SSR) using the Microsatellite Identification
tool (MISA) [98], with the SSR motif minimum length parameters being 10 for mono, 6 for
di, and 5 for tri, tetra, penta and hexa motifs.
Annotation was performed using BLASTP (v2.7.1+) [99] to query the *E. fawcettii* predicted proteome against the Swiss-Prot Ascomycota database (release 2018_08) [100] with an e-value of 1e-06 and word size of 3. BLAST results were loaded into Blast2GO Basic (v5.2.1) [101], with InterProScan, mapping and annotation steps being performed with default parameters, except HSP-hit coverage cutoff was set to 50% to increase stringency during annotation. Further annotation was achieved using HmmScan in HMMER (v3.2.1) [102] to query the predicted proteome against the Protein Family Database (Pfam) (release 32) [103]. GC% content of the coding DNA sequence (CDS) of each gene was determined using nucBed from Bedtools (v2.27.1) [104]. Predicted proteins were searched for polyamino acid (polyAA) repeats of at least five consecutive amino acid residues using the FIMO motif search tool [105] within the Meme suite (v5.0.2) [106]. The Whole Genome Shotgun project was deposited at DDBJ/ENA/GenBank under the accession SDJM00000000. The version described in this paper is version SDJM01000000. Raw reads were deposited under the SRA accession PRJNA496356.

**Phylogenetic Analysis:**

ITS and partial TEF1α sequences of 12 *E. fawcettii* pathotypes, 11 closely related *Elsinoë* species and *Myriangium hispanicum* were obtained from GenBank (accessions provided in S1) for phylogenetic analysis with *E. fawcettii* (BRIP 53147a). Sequences for each locus were aligned using MUSCLE [107] with a gap open penalty of -400, concatenated and used to perform maximum likelihood analysis in MEGA7 [108] based on the General Time Reversible model [109] with partial deletion of 90% and 1000 bootstrap replicates. The initial tree for the maximum likelihood analysis was automatically selected using Neighbor-Join and BioNJ on the matrix of pairwise distances estimated using the Maximum Composite Likelihood
A discrete Gamma distribution utilising 4 categories (+G, parameter = 0.4095) was used and the rate variation model allowed some sites to be invariable (+I, 26.6862% sites).

The character matrix and tree were combined and converted to nexus format using Mesquite (v3.6) [110] prior to TreeBASE submission (TreeBASE reviewer access: http://purl.org/phylo/treebase/phylows/study/TB2:S25460?x-access-code=f3c2b3e55c147986b2a24b44407d9e48&format=html).

E. fawcettii (BRIP 53147a) ITS and partial TEF1α sequences (accessions MN784182 and MN787508) were submitted to GenBank.

Sequence Information:

Genome assemblies and predicted proteomes included in the comparative analysis were obtained from GenBank. These included *U. maydis* (accession GCF_000328475.2, no. of scaffolds = 27) [111], *L. maculans* (accession GCF_000230375.1, no. of scaffolds = 76) [112], *M. oryzae* (accession GCF_000002495.2, no. of scaffolds = 53) [113], *R. commune* (accession GCA_900074885.1, no. of scaffolds = 164) [114], *V. dahliae* (accession GCF_000150675.1, no. of scaffolds = 55) [115], *B. cinerea* (accession GCF_000143535.2, no. of scaffolds = 18) [116], *Parastagonospora nodorum* (accession GCF_000146915.1, no. of scaffolds = 108) [117], *Pyrenophora tritici-repentis* (accession GCA_003231415.1, no. of scaffolds = 3964) [118], *S. sclerotiorum* (accession GCF_000146945.2, no. of scaffolds = 37) [119] and *Z. tritici* (accession GCA_900184115.1, no. of scaffolds = 20) [120]. Sequences of experimentally verified effector proteins were obtained from EffectorP 2.0 [121]. TE’s were identified in each assembly, as described above for *E. fawcettii*, and predicted genes which overlapped them were similarly removed from predicted proteomes.
Prediction of secretome and effectors:

Secretome and effector prediction was performed on the predicted proteomes of *E. fawcettii* and 10 fungal species known to contain effector proteins. Secretome prediction for each species began with a set of proteins predicted as secreted by either SignalP (v4.1) [122], Phobius [123] or ProtComp-AN (v6) [124]. This set was run through both the TMHMM Server (v2.0) [125] and PredGPI [126] to predict proteins with transmembrane helices and GPI-anchors, respectively. Those proteins with >1 helix or with 1 helix beyond the first 60 amino acids were removed, as were those with “highly probable” or “probable” GPI anchors. Remaining proteins formed the predicted secretome and were subjected to candidate effector prediction using EffectorP (v2.0) [121].

Genomic, proteomic and known effector analyses:

Sequences of 42 experimentally verified effector proteins, which showed >98% similarity to proteins from the 10 species included in this study, and which appeared in both the predicted secretome and candidate effector list for the respective species, were utilised in the known effector analysis. The following analyses were performed on the proteome/genome of each species. Results relating to the 42 known effectors were compared to results of all proteins from each species. Length of the intergenic flanking region (IFR) was determined as the number of bases between the CDS of two adjacent genes. Median IFR values were determined in R (v3.5.1) [127]. Genes were labelled as gene-dense if the IFR on each side was less than the median IFR length for that particular species, genes on a contig edge were not included among gene-dense labelled genes. Genes with IFR greater than the median on both sides were labelled as gene-sparse. SM clusters were predicted by passing genome assemblies and annotation files through antiSMASH fungal
(v4.2.0) [128] using the Known Cluster Blast setting. Core, accessory and unique genes for each species were determined by mapping proteins into ortholog groups using the orthoMCL algorithm [129] followed by ProteinOrtho (v5.16b) [130] on remaining unclassified genes. Core genes were those shared by all comparative species, accessory genes were shared by at least two species, but not all, and unique genes were found in only one species. GC% content of the CDS of each gene was determined as described above, \( Q_1 \) and \( Q_3 \) values were determined for each species using R [127]. HmmScan [102] of all protein sequences against the Pfam database [103] was performed as described above. Genomic AT-rich region identification was performed using OcculterCut (v1.1) [94] as described above. For genomes with identified AT-rich regions, the distance between genes and their closest AT-rich region edge was determined using Bedtools closestBed [104], as was the distance between genes and the closest TE.

Prioritisation of candidate effectors:

CE’s of each species were prioritised using an optimised scoring system based on the analysis of known effectors in 10 fungal species. All were scored out of at least five points, corresponding to one point allocated for each of the following conditions: (I) not labelled as gene-dense; (II) no involvement in predicted SM clusters; (III) labelled as either unique to the species or allocated to the same orthoMCL group as a known effector; (IV) GC% of CDS was either below the \( Q_1 \) value or above the \( Q_3 \) value of the respective species; and (V) within 10 genes upstream or downstream was at least one gene coding for a protein with a top Pfam ID hit from the following list: p450, Mito carr, FAD binding 3, FAD binding 4, Ras, DUF3328, BTB, Peptidase M28, AA permease or AA permease 2. For species with genomes which had >2% TE coverage or >25% AT-rich region coverage, CE’s were scored out of six.
points. Those genomes which had both >2% TE and >25% AT-rich region coverage, CE’s were scored out of seven points. Hence, all candidate effectors were scored out of \( n \) (five, six or seven) points, those CE’s which obtained a score of \( n \) or \( n-1 \) points were labelled as prioritised CE’s.

Prediction of other virulence genes:

SM clusters were predicted using antiSMASH fungal (v4.2.0) [128] as described above. CAZy\( n \)ymes were predicted by passing the predicted proteomes through the dbCAN2 meta server [131] and selecting three tools including HMMER scan against the dbCAN HMM database [132], Diamond [133] search against the Carbohydrate-Active enZymes (CAZy) database [134] and Hotpep query against the Peptide Pattern Recognition library [135]. Predicted CAZymes were taken as those with positive results for at least two out of the three tools. Potential pathogenesis-related proteins were identified by querying the predicted proteomes against the Pathogen Host Interactions Database (PHI-base) (v4.6, release Oct 2018) [136] using BlastP (v2.7.1) [99] analyses with an e-value of 1e-06 and a query coverage hsp of 70%, those results with >40% similarity were retained. Prioritised candidate CWDE’s were shortlisted from the predicted CAZymes to those which were predicted as secreted and obtained hits to plant associated fungal pathogenicity-related genes in PHI-base which showed evidence of reduced virulence in knockout or mutant experiments.
Results and Discussion

Genome assembly and features:

The genome assembly of *E. fawcettii* (BRIP 53147a), deposited at DDBJ/ENA/GenBank (accession SDJM00000000), was sequenced using paired-end Illumina WGS sequencing technology. Assembly of reads produced a draft genome 26.01 Mb in size with a coverage of 193x (Table 1) and consisted of 286 contigs greater than 500 bp in length, with an N50 of 662,293 bp, a mean contig length of 90,948 bp and an overall GC content of 52.3%. Running the assembly against the Ascomycota orthoDB (v9) [93] showed 97.6% of complete single copy genes were found in the *E. fawcettii* assembly, indicating a high degree of coding DNA sequence completeness. The genome of *E. fawcettii* is comparable in size to other fungal genomes including *Eurotium rubrum* (26.21 Mb) [137], *Xylona heveae* (24.34 Mb) [138] and *Acidomyces richmondensis* (29.3 Mb) [139], however it is smaller than the average Ascomycota genome size of 36.91 Mb [140]. When analysed against the 10 fungal species included in this comparative analysis (*B. cinerea*, *L. maculans*, *M. oryzae*, *Parastagonospora nodorum*, *Pyrenophora tritici-repentis*, *R. commune*, *V. dahliae*, *S. sclerotiorum*, *U. maydis* and *Z. tritici*), the *E. fawcettii* assembly is the second smallest, after *U. maydis* at 19.6 Kb. TE identification, by analysis against Repbase (release 18.02) [141], showed a coverage of only 0.37%, indicating a low proportion of the *E. fawcettii* genome is represented by currently known TE’s, this is a likely contributor to its comparatively small genome size. This low TE coverage may also be the result of a fragmented genome [142]. It is possible, should long read sequencing of this isolate be completed in the future, TE coverage may appear higher.
Table 1. Features of *Elsinoë fawcettii* (BRIP 53147a) genome assembly

<table>
<thead>
<tr>
<th>General Features</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly length (bp)</td>
<td>26,011,141</td>
</tr>
<tr>
<td>Coverage</td>
<td>193x</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>286</td>
</tr>
<tr>
<td>Mean GC content (%)</td>
<td>52.3</td>
</tr>
<tr>
<td>N50 (bp)</td>
<td>694,004</td>
</tr>
<tr>
<td>Mean contig length (bp)</td>
<td>90,948</td>
</tr>
<tr>
<td>Minimum contig length (bp)</td>
<td>501</td>
</tr>
<tr>
<td>Maximum contig length (bp)</td>
<td>2,345,732</td>
</tr>
<tr>
<td>Coverage of interspersed repeats (bp)</td>
<td>95,654 (0.37%)</td>
</tr>
<tr>
<td>Coverage of short simple repeats (bp)</td>
<td>6868 (0.026%)</td>
</tr>
<tr>
<td>Number of predicted gene models</td>
<td>10,080</td>
</tr>
<tr>
<td>Number of contigs containing predicted genes</td>
<td>141</td>
</tr>
<tr>
<td>Mean gene length (bp)</td>
<td>1,573</td>
</tr>
<tr>
<td>Mean number of exons per gene</td>
<td>2.35</td>
</tr>
<tr>
<td>Number of genes containing a polyAA repeat</td>
<td>1,073</td>
</tr>
<tr>
<td>Mean GC content of CDS (%)</td>
<td>54.7</td>
</tr>
</tbody>
</table>

The *E. fawcettii* genome has less predicted gene models than the average Ascomycota genome of 11129.45 [140]. Gene prediction produced 10,080 gene models, 5,636 (55.91%) of which were annotated, while 4,444 (44.09%) were labelled as coding for hypothetical proteins. The average gene length was 1,573 bp with an average of 2.35 exons per gene, there were 3,280 single exon genes. The mean GC content of CDS was 54.7%, which was 2.4% higher than the overall GC content and showed a wide variation in range, with the lowest scoring gene at 44.29% GC and the highest being 71.53%, thus exposing a spectrum on which genes may be differentiated. Hmmscan [102] analysis of the predicted proteome against the Pfam database [103] revealed a high proportion (70.1% = 7,069) of genes with at least one hit to a Pfam model. The same analysis performed on the proteomes of the 10 fungal species included in the comparative analysis gave results ranging from 48.6% for
S. sclerotiorum, with the lowest proportion of Pfam hits, to 74.9% for U. maydis with the highest, and a mean of 62.1% over the 11 species (S2).

Analysis of orthologous genes among E. fawcettii and the 10 comparative species indicated 3,077 (30.5%) of the predicted genes of E. fawcettii were core genes, finding hits through OrthoMCL or ProteinOrtho in all 11 species (S2). There were 4,874 (48.4%) E. fawcettii genes found in at least one other species but not all and were therefore considered accessory genes. Lastly, the remaining 2,129 (21.1%) were found in only the E. fawcettii proteome, 140 of these, however, obtained a hit to an orthoMCL group and were therefore set aside and not considered as unique proteins in subsequent analyses, leaving 1,989 (19.7%) genes presumed to be Elsinoë-specific and therefore potentially involved in either Elsinoë- or E. fawcettii-specific pathogenesis pathways. A comparative analysis among the core, accessory and unique genes of the 11 species (S2) (Figure 1) indicated that U. maydis was set apart from the other species by showing the lowest proportion of accessory genes, this was expected as U. maydis was the only biotroph and Basidiomycete among the group.

E. fawcettii showed a below average percentage of unique genes which may be expected due its smaller than average sized genome and proteome. A lower number of unique genes may place a limitation on the ability of E. fawcettii to infect a larger range of host plants.

Figure 1. Comparison of gene classifications among the proteomes of 11 fungal pathogens.

Genes were categorised using orthoMCL group IDs, or proteinortho if no group was assigned. Genes were considered; (I) core if they were shared by all 11 species; (II) accessory
388 if they were shared by at least two species, but not all; (III) unique if they were found in only
389 one of the 11 species.

390 While the overall GC content of \textit{E. fawcettii} was 52.3%, when taking AT-rich regions into
391 consideration, the average GC content of 98.97% of the genome was 52.8%, while the AT-
392 rich regions had an average GC content of 33.8%. AT-rich regions are sections of DNA that
393 are scattered throughout the genome and have a significantly higher AT content compared
394 to adjacent GC equilibrated blocks \cite{94}. The presence of AT-rich regions in genomes varies
395 widely, for example \textit{S. sclerotiorum} does not show evidence of AT-rich regions \cite{143}, while
396 36\% of the \textit{L. maculans} genome is covered by AT-rich regions which have an average GC
397 content of 33.9\% \cite{47}. AT-rich regions are thought to develop in, and nearby to, regions
398 containing TE repeats, through Repeat-Induced Point mutation (RIP), a mechanism used to
399 inhibit the destructive actions of TE’s against an organism’s genome. Through a fungal
400 genome defence mechanism causing cytosine to thymine polymorphisms, a TE repeat
401 sequence is inhibited from further movement and potential destruction of necessary genes.
402 This same type of polymorphism can also occur in genes nearby to TE regions \cite{144-147},
403 potentially providing numerous genomic locations with increased plasticity scattered
404 throughout the genome. While RIP occurs during the sexual phase it has also been observed
405 in asexual fungi and is thought to indicate a species reproductive history or potential \cite{148}.
406 AT-rich regions are present within the \textit{E. fawcettii} genome, however the extent of their
407 coverage in the present assembly is low, 59 regions with an average GC content of 33.8%
408 cover only 1.03\% of the genome. Sixteen regions are found overlapping TE’s, while four are
409 found within 2 Kb of a TE region, meaning 33.9\% of the AT-rich regions potentially represent
410 RIP-affected regions. The remaining 66.1\%, found either >2Kb away or on a contig that does
not contain a predicted TE region, are potentially RIP-affected regions where the TE is no longer recognisable. The AT-rich regions of *E. fawcettii* are not scattered evenly throughout the genome, instead 29/59 (49.2%) are situated at the edge of a contig and 15/59 (25.4%) cover the entire length of a contig, specifically contigs not containing genes. Two further AT-rich regions were located between the edge of a contig and the beginning of the first gene and so were grouped with those located at the edge of a contig. The remaining 13 regions (22.0%) were situated within a contig with genes residing on both sides. Hence, the majority either made up the edge of a contig which contained genes or filled entire contigs which did not contain genes, meaning it is likely that the sequence of many *E. fawcettii* AT-rich regions contain sections of such low complexity that contig breaks result, a hypothesis which could be tested in the future using long read sequencing technology. Eight predicted genes at least partially overlap these regions and 57 are located within 2 Kb, a finding which has potential significance as AT-rich regions have been known to harbour effector genes in fungal pathogens [149, 150]. There was a large range of diversity of AT-rich region coverage among the fungal pathogens analysed in the current study; *S. sclerotiorum*, *Pyrenophora tritici-repentis*, *M. oryzae* and *U. maydis* showed no AT-rich regions; *V. dahliae* (1.5%), *B. cinerea* (4.9%), *Parastagonospora nodorum* (6.6%) and *Z. tritici* (17.3%) showed lower degrees of AT-rich coverage; while *R. commune* (29.5%) and *L. maculans* (37%) showed the greatest extent. These levels of AT-rich coverage did not appear to corelate with pathogen classification as necrotrophic, hemibiotrophic or biotrophic, nor as host-specific or broad-host range pathogens. The genomic location of AT-rich regions was, however, further included in the known effectors and candidate effectors analyses.
Identification and analysis of SSR’s in the *E. fawcettii* genome located 400 regions covering 6,868 bp (0.026%), 164 (41%) of which were contained within a predicted gene. Furthermore, polyAA repeats, of at least five identical and adjacent residues, were identified within 1,073 predicted protein sequences. The presence of repetitive sequences has been noted in fungal effectors [33, 45, 151] and implicated in the function and evolution of pathogenicity-related genes of other plant-associated microorganisms [152]. Analysis of the 1,105 proteins which obtained either an SSR or polyAA hit indicated 237 (21.45%) were categorised as *E. fawcettii*-specific and did not obtain a Pfam hit, highlighting potentially novel genus- or species-specific genes involved in host pathogenesis.

Phylogenetic analysis of partial ITS and TEF1-α regions of *E. fawcettii* (BRIP 53147a) in comparison with other *E. fawcettii* isolates and closely related *Elsinoë* species (Figure 2) indicates *E. fawcettii* (BRIP 53147a) closely aligns with the *E. fawcettii* clade. Substitutions appearing in the Jingeul pathotype isolates are not seen in isolate BRIP 53147a. One G to A substitution in the TEF1-α region sets isolate BRIP 53147a apart from the other *E. fawcettii* isolates (S3), a base which is at the 3rd position of a Glu codon and hence does not result in a translational difference. This substitution in the BRIP 52147a isolate appeared with a high degree of confidence, 100% of sequence reads aligned back to the assembly and a coverage of 241x, at this point, agreed with the substitution. While it is thought that isolate BRIP 53147a belongs to either the Lemon or Tyron’s pathotype, it is yet to be determined which or if it constitutes a new pathotype of its own. Aside from the one base substitution in the TEF1-α region, there would be some expected differences throughout the genomes of the *E. fawcettii* BRIP 53147a isolate and the other *E. fawcettii* isolates due to differences in collection details, such as geographical location, year and host specificity. Specifically,
isolate BRIP 53147a was collected in Montville, Queensland in 2009, while the other
Australian isolates, DAR 70187 and DAR 70024, belonging to the Lemon and Tyron’s
pathotypes, were collected 15 years earlier in Somersby and Narara in NSW, respectively [7], both a distance of almost 1000 km away. Several isolates from Figure 2 have been
tested for host pathogenicity leading to the designation of specific pathotypes [3], as
opposed to relying on only sequence data and thus illustrating the importance of
experimental validation prior to pathotype or species classification. For example, Jin-1 and
Jin-6 are classified as the Jingeul pathotype, SM3-1 as FBHR, S38162 as FNHR, CC-132 as
SRGC, DAR 70187 and CC-3 as the Lemon pathotype, and DAR 70024 as Tyron’s pathotype [3]. Host specificity experimentation for the *E. fawcettii* BRIP 53147a isolate is a suggested
future step, as is the whole genome sequencing and analysis of further *E. fawcettii* isolates
for comparison. The comprehensive host pathogenicity testing of 61 *E. fawcettii* isolates and
their subsequent classification into six pathotypes [3] coupled with genomic sequencing
data analysis would provide a wealth of knowledge of potential host-specific pathogenicity-
related genes and mutations.

Figure 2. **Maximum likelihood phylogenetic tree of *E. fawcettii* isolates and closely related species.** The phylogenetic tree was inferred from a concatenated dataset including ITS and
partial TEF1-α regions. *Myriangium hispanicum* was used as the outgroup. The branch
length indicates the number of nucleotide substitutions per site, bootstrap values are shown
at nodes, host in parentheses, new isolate described in the current study denoted with
asterisk (*) and type strains are in bold.
Prediction of secretome and effectors:

A total of 1,280 genes (12.7% of the proteome) were predicted to code for secreted proteins (SP) in the *E. fawcettii* genome (Table 2). Using the discovery pipeline outlined in Figure 3, classically secreted proteins with a detectable signal peptide were predicted by either SignalP and/or Phobius providing 1,449 proteins, while ProtComp identified a further 120 as potential non-classically secreted proteins. Of these 1,569 proteins 186 were removed as they were predicted to contain transmembrane helices, an indication that while targeted for secretion the protein likely functions while situated in the cell membrane. A further 103 were removed as they contained a predictable GPI anchor, also suggesting they associate with the cell membrane to perform their function, leaving a total of 1,280 proteins identified as likely SP’s. To enable comparison of the species’ predicted secretomes and CE’s, the same prediction pipeline (Figure 3) was used on the proteomes of 10 further fungal species included in the analysis (Table 2), essentially utilising genomes which contain known protein effectors for comparison. The proportion of predicted SP’s in the *E. fawcettii* proteome was similar to that of other necrotrophic fungal pathogens, which ranged from *B. cinerea* at 11.3% to *Parastagonospora nodorum* at 13.9%. It was, however, lower in comparison to the hemibiotrophs; *R. commune* showed a low of 12.5% SP’s while *M. oryzae* was the highest scoring at 18.5%, demonstrating a small increase in proportion of SP’s for the hemibiotrophs compared to the necrotrophs. This potentially provides them with a larger array of secreted proteins compared to biotrophs and necrotrophs, to first support a biotrophic, and secondly a necrotrophic, host interaction.
<table>
<thead>
<tr>
<th>Species</th>
<th>Total proteins*</th>
<th>SP (% of total)</th>
<th>CE (% of SP)</th>
<th>Known effectors correctly predicted as SP’s and CE’s</th>
<th>Known effectors not predicted as SP’s and CE’s</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Necrotrophs:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Elsinoë fawcettii</em></td>
<td>10,080</td>
<td>1,280 (12.7%)</td>
<td>203 (15.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Botrytis cinerea</em></td>
<td>11,481</td>
<td>1,294 (11.3%)</td>
<td>214 (16.5%)</td>
<td>NEP1</td>
<td></td>
</tr>
<tr>
<td><em>Parastagonospora nodorum</em></td>
<td>15,878</td>
<td>2,206 (13.9%)</td>
<td>614 (27.8%)</td>
<td>Tox1, ToxA</td>
<td></td>
</tr>
<tr>
<td><em>Pyrenophora tritici-repens</em></td>
<td>10,771</td>
<td>1,298 (12.1%)</td>
<td>284 (21.9%)</td>
<td>ToxB</td>
<td></td>
</tr>
<tr>
<td><em>Sclerotinia sclerotiorum</em></td>
<td>13,770</td>
<td>1,707 (12.4%)</td>
<td>490 (28.7%)</td>
<td>SsSSVP1</td>
<td></td>
</tr>
<tr>
<td><em>Zymoseptoria tritici</em></td>
<td>11,936</td>
<td>1,514 (12.7%)</td>
<td>480 (31.7%)</td>
<td>Zt6, AvrStb6</td>
<td></td>
</tr>
<tr>
<td><strong>Hemibiotrophs:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Leptosphaeria maculans</em></td>
<td>12,337</td>
<td>1,883 (15.3%)</td>
<td>495 (26.3%)</td>
<td>AvrLM6, AvrLM11, AvrLM4-7</td>
<td></td>
</tr>
<tr>
<td><em>Magnaporthe oryzae</em></td>
<td>12,236</td>
<td>2,263 (18.5%)</td>
<td>742 (32.8%)</td>
<td>SPD10, Msp1, BAS1, SPD4, SPD2, MoCDIP3, MoCDIP4, AVR-Pik, MoCDIP1, Bas107, BAS2, BAS3, BAS4, Avr-Pita1, Bas162, MoHEG13, SPD7, MC69, AvrPi9, AvrPiz-t, SPD9, MoCDIP5</td>
<td>MoCDIP2</td>
</tr>
<tr>
<td><em>Rhynchosporium commune</em></td>
<td>12,100</td>
<td>1,510 (12.5%)</td>
<td>387 (25.6%)</td>
<td>NIP1, NIP2, NIP3</td>
<td></td>
</tr>
<tr>
<td><em>Verticillium dahliae</em></td>
<td>10,441</td>
<td>1,407 (13.5%)</td>
<td>270 (19.2%)</td>
<td>PevD1, VdSCP7</td>
<td>Vdlscl</td>
</tr>
<tr>
<td><strong>Biotroph:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Ustilago maydis</em></td>
<td>6,692</td>
<td>856 (12.8%)</td>
<td>178 (20.8%)</td>
<td>Pit2, Pep1, See1, Cmu1, Tin2</td>
<td>Eff1-1</td>
</tr>
</tbody>
</table>

*Not including gene models which overlap a predicted TE region
Figure 3. **Pipeline for the discovery of the predicted secretome and candidate effectors.**

The secretome search started with the predicted proteins of a species, proteins were predicted as secreted using at least one of three tools, proteins with predicted transmembrane helices or GPI-anchors were removed. Candidate effectors were predicted using EffectorP. The number of proteins shown for the predicted proteome, secretome and effectome refers to the *Elsinoë fawcettii* BRIP 53147a genome.

Known effectors were frequently identified by the CE pipeline (Figure 3), with 43/45 (95.6%) correctly predicted as being secreted and 42/45 (93.3%) also predicted as effectors (Table 2), highlighting the effectiveness of the pipeline among these fungal species. Those known effectors which were tested but not identified as SP’s included Vdlsc1 (*V. dahliae*) and MoCDIP2 (*M. oryzae*). Vdlsc1 lacks an N-terminal signal peptide and is unconventionally secreted [153], however it was not identified as a non-classically secreted protein. MoCDIP2 was removed as it obtained a GPI-anchor hit. Additionally, Eff1-1 (*U. maydis*) was predicted as secreted but not as a candidate effector, Eff1-1, along with MoCDIP2, are both known false negatives of EffectorP 2.0 [121].

The total number of CE’s identified for *E. fawcettii* was 203, meaning only 15.9% of SP’s gained CE classification, this was the lowest proportion out of all 11 species analysed (Table 2). This may be explained by the potential favouring of EffectorP towards SP’s of species on which it was trained. To further investigate this potential, results of EffectorP for the 11
species were compared to the results of an alternate candidate effector search; SP’s with a protein length less than the species’ median and with no Pfam hit other than to that of a known effector (S4). While this second method resulted in the identification of a higher number of CE’s for each species, *E. fawcettii* still obtained the lowest proportion of CE’s out of predicted SP’s, indicating *E. fawcettii* may have a lighter dependence, compared to other fungal pathogens, on protein effectors. It also highlighted the advantage of using EffectorP to narrow down an extensive catalogue of SP’s, as opposed to identifying CE’s based on arbitrary features. However, the CE’s predicted by EffectorP still range in the hundreds (Table 2), it was therefore beneficial to further shortlist candidates for prioritisation. To achieve this, known effectors which were correctly predicted as both SP’s and as CE’s (Table 2) were retained for further analysis to generate an optimised prioritisation scoring system.

**Known effector analysis:**

A total of 42 known effectors from 10 fungal species were analysed for; (I) gene density; (II) GC content; (III) involvement in SM clusters; (IV) uniqueness; (V) Pfam hits of surrounding genes; (VI) distance to the closest TE; and (VII) distance to the closest AT-rich regions (Table 3). Results were compared to those of all predicted genes from each of the same 10 species (S5). Features observed at a higher rate among the known effector group compared with each species’ proteome were used to generate a prioritisation pathway using a point allocation system. (I) Genes were labelled as gene-dense if the IFR’s on both sides were less than the IFR median value for that specific species, allowing an analysis relative to each organism. The proportions of gene-dense genes ranged from 21.6% (*Pyrenophora tritici-repentis*) to 28.0% (*B. cinerea*) (S5), in contrast to 3/42 (7.1%) known effectors (Table 3). This provided grounds to allocate one point to each known effector which was not labelled
as gene-dense. (II) GC content of the CDS of each gene was determined and median values calculated for each species, revealing the GC percentage of 32/42 (76.2%) known effectors fell either below the $Q_1$ value or above the $Q_3$ value for the respective species. When compared to an expected 50% in the upper and lower quartiles, this provided reason for the allocation of one point to known effectors should they fall in these two quartiles. (III) No overlap was observed between known effectors and the predicted SM clusters within each species, giving strong reason for the allocation of one point to known effector’s that were not included in SM clusters. (IV) Analysis of gene classification (core, accessory or unique) for each known effector highlighted that 41/42 (97.6%) were either unique to the species (31/42) or were assigned an orthoMCL group ID of a known effector (10/42). In contrast, the proportion of unique genes for each species was much lower, ranging from 11.9% (B. cinerea) to 33.7% (S. sclerotiorum), with an average of 25.4%. The proportion of genes allocated an orthoMCL of a known effector was similarly low at less than 0.3% for all species. Thus, a point was allocated to known effectors that were either unique to the species or obtained the same orthoMCL ID of a known effector. (V) Pfam hits of genes surrounding known effectors were also compared to the rates of Pfam hits from all 10 proteomes together. Analysis of the 10 genes upstream and downstream of a known effector revealed 10 Pfam hits which appeared at a rate at least double to that seen among the concatenated proteomes. For example, Pfam hits to cytochrome P450 accounted for 2.82% of all hits among the 10 genes up and downstream of a known effector, compared to only 1.26% of Pfam hits from the predicted proteins of all 10 species. Aside from cytochrome P450, further Pfam hits overrepresented among the genes surrounding known effectors included mitochondrial carrier protein, FAD binding domains 3 and 4, Ras family, domain of unknown function (DUF3328), BTB/POZ domain, peptidase family M28, and
580 amino acid permease 1 and 2. At least one of these Pfam hits was found within 10 genes of
581 66.7% of the known effectors, which was higher when compared to all genes of each of the
582 10 species. Proportions ranged from only 28.7% (Pyrenophora tritici-repentis) up to 43.9%
583 (B. cinerea), with an average of 34.1%, over the 10 species, of genes being within 10 genes
584 of an overrepresented Pfam hit. A point was therefore allocated to known effectors which
585 lay within 10 genes of a gene with one of the above mentioned Pfam hits. (VI) Those
586 genomes with >2% TE coverage also showed a high proportion of known effectors in the
587 close vicinity of TE’s. Specifically, 29/32 (90.6%) known effectors from Z. tritici,
588 S. sclerotiorum, B. cinerea, R. commune, L. maculans and M. oryzae were within seven genes
589 of a TE region, compared to an average of 47.8% of genes within seven genes of a TE for the
590 same six species. This led to the allocation of one point for known effectors within seven
591 genes of a TE for species with >2% TE coverage. (VII) Lastly, of the genomes analysed, only
592 those consisting of >25% AT-rich regions, being R. commune and L. maculans, were found to
593 have a noticeable association between the location of known effectors and AT-rich regions.
594 The distance of all known effectors to the closest AT-rich region, of these two species, were
595 found to be less than the Q₁ value for each species. Hence, known effectors with these
596 specifications, in species with >25% AT-rich region coverage, were allocated one point. It
597 can be seen that depending on the degree of TE and AT-rich region coverage, each species
598 known effectors may be scored out of five, six or seven points, henceforth referred to as “n
599 points”. Over the 10 species with known effectors which were analysed, Table 3 illustrates a
600 total of 36/42 (85.7%) known effectors obtained n or n-1 points, revealing a process which
601 could be used to prioritise the many CE’s predicted for the E. fawcettii genome.
### Table 3. Features of known fungal effectors used to guide candidate effector prioritisation

<table>
<thead>
<tr>
<th>Effector</th>
<th>Gene density class</th>
<th>CDS GC%</th>
<th>Within SM gene cluster</th>
<th>Ortholog class</th>
<th>Within 10 genes of a specified Pfam hit&lt;sup&gt;A&lt;/sup&gt;</th>
<th># of genes from a TE</th>
<th>Distance to AT-rich region</th>
<th>Total possible points (&lt;i&gt;n&lt;/i&gt; points)</th>
<th>Points scored</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Necrotrophic:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Botrytis cinerea:</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEP1</td>
<td>Sparse</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>Yes</td>
<td>3</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td><strong>Parastagonospora nodorum:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tox1</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>ToxA</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>N/A</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Pyrenophora tritici-repentis:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ToxB</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>Sclerotinia sclerotiorum:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SsSSVP1</td>
<td>Sparse</td>
<td>Q&lt;sub&gt;2&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>2</td>
<td>N/A</td>
<td>6</td>
<td>4&lt;sup&gt;D&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Zymoseptoria tritici:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zt6</td>
<td>Sparse</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Core&lt;sup&gt;B&lt;/sup&gt;</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>7</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>AvrStb6</td>
<td>Sparse</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td><strong>Hemibiotrophic:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Leptosphaeria maculans:</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AvrLM6</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>AvrLM11</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>AvrLM4-7</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td><strong>Magnaporthe oryzae:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD10</td>
<td>Dense&lt;sup&gt;C&lt;/sup&gt;</td>
<td>Q&lt;sub&gt;2&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>4</td>
<td>N/A</td>
<td>6</td>
<td>4&lt;sup&gt;D&lt;/sup&gt;</td>
</tr>
<tr>
<td>Msp1</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>Yes</td>
<td>15&lt;sup&gt;C&lt;/sup&gt;</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>BAS1</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD4</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>SPD2</td>
<td>Dense&lt;sup&gt;c&lt;/sup&gt;</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>6</td>
<td>N/A</td>
<td>6</td>
<td>4&lt;sup&gt;D&lt;/sup&gt;</td>
</tr>
<tr>
<td>MoCDIP3</td>
<td>Sparse</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>MoCDIP4</td>
<td>Sparse</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>Yes</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>AVR-Pik</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>MoCDIP1</td>
<td>Sparse</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>Yes</td>
<td>29&lt;sup&gt;C&lt;/sup&gt;</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Bas107</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>4</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>BAS2</td>
<td>Q&lt;sub&gt;3&lt;/sub&gt; &lt;sup&gt;C&lt;/sup&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;C&lt;/sup&gt;</td>
<td>Yes</td>
<td>3</td>
<td>N/A</td>
<td>6</td>
<td>4&lt;sup&gt;D&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>BAS4</td>
<td>Sparse</td>
<td>Q&lt;sub&gt;3&lt;/sub&gt; &lt;sup&gt;C&lt;/sup&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>2</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>BAS3</td>
<td>Q&lt;sub&gt;3&lt;/sub&gt; &lt;sup&gt;C&lt;/sup&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>6</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Avr-Pita1</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>Yes</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Bas162</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>4</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>MoHEG13</td>
<td>Sparse</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>3</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>SPD7</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>3</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>MC69</td>
<td>Sparse</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>6</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>AvrP!9</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>3</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>AvrPiz-t</td>
<td>Q&lt;sub&gt;3&lt;/sub&gt; &lt;sup&gt;C&lt;/sup&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>1</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>SPD9</td>
<td>Sparse</td>
<td>Q&lt;sub&gt;3&lt;/sub&gt; &lt;sup&gt;C&lt;/sup&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>2</td>
<td>N/A</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>MoCDIP5</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Accessory&lt;sup&gt;B&lt;/sup&gt;</td>
<td>Yes</td>
<td>2</td>
<td>N/A</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

**Rhynchosporium commune:**

| NIP3   | Q<sub>3</sub> <sup>C</sup> | No | Unique | Yes | N/A | N/A | 11<sup>C</sup> | 1368 | 7 | 5<sup>D</sup> |
| NIP1   | Sparse | >Q<sub>3</sub> | No | Unique | Yes | 1 | 1814 | 7 | 7 |
| NIP2   | Sparse | >Q<sub>3</sub> | No | Unique | No<sup>C</sup> | 1 | N/A | 6572 | 7 | 6 |

**Verticillium dahliae:**

| PevD1   | >Q<sub>3</sub> | No | Accessory<sup>B</sup> | No<sup>C</sup> | N/A | N/A | 5 | 4 |
| VdSCP7  | Q<sub>3</sub> <sup>C</sup> | No | Unique | Yes | N/A | N/A | 5 | 4 |

**Biotrophic:**

**Ustilago maydis:**
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pit2</td>
<td>Sparse</td>
<td>Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Pep1</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>See1</td>
<td>Dense&lt;sup&gt;C&lt;/sup&gt;</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No&lt;sup&gt;C&lt;/sup&gt;</td>
<td>N/A</td>
</tr>
<tr>
<td>Cmu1</td>
<td>&gt;Q&lt;sub&gt;3&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Tin2</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>&lt;Q&lt;sub&gt;1&lt;/sub&gt;</td>
<td>No</td>
<td>Unique</td>
<td>No</td>
<td>N/A</td>
</tr>
</tbody>
</table>

604<sup>A</sup> Specified Pfam hits: p450, Mito carr, FAD binding 3, FAD binding 4, Ras, DUF3328, BTB, Peptidase M28, AA permease or AA permease 2.

605<sup>B</sup> Allocated the same orthoMCL group ID as a known effector

606<sup>C</sup> Possible point not allocated

607<sup>D</sup> Less than n-1 points scored
Prioritisation of candidate effectors:

While EffectorP correctly determined most known effectors, it also identified a large number of additional CE’s. While it is likely some of these candidates are unknown effectors being utilised by the pathogen to infect its host, it would be worthwhile to shortlist this group, to a list of the more likely candidates, prior to expensive and time-consuming experimental validation procedures. A points-based process was developed, based on the analysis of known effectors, to prioritise CE’s based on several features including: their distance to neighbouring genes, lack of involvement in predictable SM clusters, GC% of CDS, proximity to genes obtaining certain Pfam hits and potential uniqueness (Figure 4). For species with genome assemblies containing >2% TE coverage the number of genes a CE was from a TE was taken into consideration. Similarly, the distance between genes and AT-rich regions was acknowledged if AT-rich regions covered >25% of the species’ assembly. For each CE gene, one point was available for each of the above features, hence, as described for the known effector analysis, CE’s of each species were allocated a possible five, six or seven points (n points). E. fawcettii, Parastagonospora nodorum, Pyrenophora tritici-repentis, V. dahlia and U. maydis each had <2% TE coverage and <25% coverage of AT-rich regions, their CE’s were therefore scored out of five points. Z. tritici, S. sclerotiorum, B. cinerea and M. oryzae had >2% TE coverage but <25% coverage of AT-rich regions and so were scored out of six points. Only the assemblies of R. commune and L. maculans showed >2% TE’s and >25% AT-rich regions, and as such their CE’s were scored out of seven points. By using n or n-1 points as an acceptable score for CE prioritisation, revealed that CE’s of the 11 species could be reduced, by 51.1% - 83.6% (average 66.2%) (S6), with species that were scored out of more points achieving higher reductions.
Candidate effector prioritisation features and points. The candidate effectors (CE’s) of all genomes analysed were scored using features shown in the blue box. Additional features were considered for CE’s from genomes with >2% TE coverage (red box) and >25% AT-rich region coverage (green box).

Applying the method outlined in Figure 4 to the CE’s of E. fawcettii led to the prioritisation of 77 CE’s, a reduction of 62%, for future experimental validation. This is a comparable reduction to that of the other necrotrophic pathogens (Figure 5, S6), for which six out of seven known effectors were retained within the shortlisted CE’s. Features of the 77 CE’s of E. fawcettii (S7) indicated many were small in size, had a high GC content, had a high proportion of cysteine residues and were more likely to be classified as gene-sparse. The median protein length was 181 aa, compared to 409 aa for all E. fawcettii predicted genes. The mean GC content was 55.82% and the mean cysteine content was 3.4%, compared to 54.69% and 1.2%, respectively for all predicted genes of E. fawcettii. The high proportion (44.2%) of gene-sparse genes among prioritised CE’s was expected, as CE’s which were not classified as gene-dense were favoured during the prioritisation process, however high proportions of gene-sparse genes were also observed among the SP’s and CE’s (Table 4).

Specifically, 26.8% of all E. fawcettii predicted genes were classed as gene-sparse, 26.3% as gene-dense and the remaining 46.9% classed as neither. In comparison, 31.9% of SP’s and 35.7% of CE’s were classed as gene-sparse and only 15.9% and 16.25%, respectively, were classed as gene-dense, indicating a preference for gene-sparse locations by proteins likely secreted by the pathogen. PolyAA repeat-containing proteins were not overrepresented among the prioritised CE’s, two were found to contain five consecutive Ala residues and one

Figure 4.
other contained five consecutive Arg residues. Additionally, no CE’s were found to contain SSR’s suggesting that diversity of *E. fawcettii* effector sequences is not being generated through an increased mutational rate related to short repetitive sequences. Furthermore, the prioritised CE’s were found scattered throughout the genome over 34 of the 141 gene-containing contigs and did not appear to cluster together. While AT-rich regions were not taken into consideration during the prioritisation of *E. fawcettii* CE’s, due to a low AT-rich coverage of 1.03%, it should be noted that higher proportions of SP’s and CE’s were found among genes on the edge of a contig and those within 2 Kb of an AT-rich region than expected. Out of the 252 genes found at the edge of a contig, 36 (14.2%) were SP’s and 11 (4.3%) were CE’s, compared to 12.7% and 2.0%, respectively, out of all *E. fawcettii* proteins. Similarly, of the 57 genes found within 2 Kb of an AT-rich region, 12 (21.1%) were SP’s and four (7.0%) were CE’s (S7). This suggests that genomic regions near contig breaks, such as sequences of low complexity or regions under-represented by short read sequencing technology, and AT-rich regions may be indicators within the *E. fawcettii* genome of nearby SP’s and effector genes. Interestingly, SP’s and CE’s were not overrepresented among genes found within 2 Kb of a predicted TE region, of the 120 genes found in these regions 12 (10%) were SP’s and 2 (1.7%) were CE’s, both slightly less than their proportions across the whole genome. This suggested while potential effector genes are more likely to be found near AT-rich regions, a nearby predictable TE region was not necessary. Thus, *E. fawcettii*, a necrotrophic pathogen not considered at first thought to utilise protein effectors to increase virulence, shows a subtle, yet intriguing, pattern of SP’s and CE’s near AT-rich regions, at contig edges and in more gene-sparse locations. This potentially points towards a set of virulence-related genes being maintained in specific genomic locations and therefore suggesting their potential significance.
Figure 5 **Comparison of numbers of secreted proteins, candidate effectors and prioritised candidate effectors among 11 fungal pathogens.** Secreted proteins and candidate effectors were predicted using the pipeline in Figure 3. Prioritised candidate effectors were determined using features shown in Figure 4.

Table 4. Gene density classification of *Elsinoë fawcettii* predicted proteins

<table>
<thead>
<tr>
<th>Classification</th>
<th>All predicted proteins</th>
<th>Secreted proteins</th>
<th>Candidate effectors</th>
<th>Prioritised candidate effectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene-sparse</td>
<td>26.8%</td>
<td>31.9%</td>
<td>35.7%</td>
<td>44.2%</td>
</tr>
<tr>
<td>Gene-dense</td>
<td>26.3%</td>
<td>15.9%</td>
<td>16.3%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Neither</td>
<td>46.9%</td>
<td>52.1%</td>
<td>48.3%</td>
<td>54.5%</td>
</tr>
</tbody>
</table>

While analysing proteins using the features mentioned above can shortlist CE’s, awareness of limitations should be considered. For example, only prioritising CE’s which are unique to a species, or obtain the same orthoMCL hit as a known effector, limits the identification of novel effectors which may be utilised by multiple species. Hence, a blast search of *E. fawcettii* CE’s against CE’s of the 10 other fungal pathogens was conducted and indicated 12 (5.9%) *E. fawcettii* CE’s had >70% similarity to at least one candidate effector of another species (S7). Four of these 12 proteins were prioritised CE’s, one of which had 72.9% similarity to MoCDIP1 (*M. oryzae*), a known effector which is expressed in planta and induces host cell death [49], thus highlighting this CE for further investigation.
Prediction and prioritisation of cell wall degrading enzymes:

Further potential pathogenicity-related genes of *E. fawcettii* which deserve attention include CWDE’s. The *E. fawcettii* proteome showed 378 (3.75%) predicted CAZymes (S8), comparable to the proportion of CAZymes seen in the other 10 pathogen genomes, which ranged from 2.8% (*S. sclerotiorum*) to 4.3% (*V. dahliae*) (S2). Of the total *E. fawcettii* CAZymes, 203 (53.7%) were also predicted as secreted, highlighting numerous potential CWDE’s secreted by the pathogen and targeted for interaction with host carbohydrates. It would be beneficial to compare these potential CWDE’s with transcriptomic data once available, however, currently they can be cross-referenced against the Pfam database.

Analysis of the 203 potential CWDE’s revealed frequently appearing Pfam hits to pectate lyase and pectinesterase (19 hits), the glycosyl hydrolases family 28 of pectin-degrading polygalacturonases (11 hits) and the glycosyl hydrolases family 43 of hemicellulose-degrading beta-xylosidases (10 hits). Hemicellulose- and pectin-degrading enzymes target plant cell wall components including xyloglucans and pectin’s, respectively [68], both found in high proportions in the primary cell wall, potentially revealing an arsenal of CWDE’s of *E. fawcettii* which are targeted towards young plant tissues. Polygalacturonases break bonds between polygalacturonic acid residues, thereby degrading pectin, while beta-xylosidases hydrolyse xylan, a hemicellulose component of the cell wall. It is possible that the CWDE’s of *E. fawcettii* have the ability to degrade components of a growing cell wall, however as the host cell wall matures, the *E. fawcettii* CWDE repertoire becomes less effective, perhaps explaining why only young plant tissues are susceptible to citrus scab. The 203 potential CWDE’s were also cross-referenced against PHI-base, resulting in the prioritisation of 21 proteins which had similarity to known virulence factors of plant pathogens (Table 5, S8).
thus highlighting candidate virulence genes of *E. fawcettii* for future experimental investigation. Among these 21 proteins were 14 predicted pectin-degrading enzymes, including two with similarity to pectin-degrading enzymes, specifically *pg1* (53.7%) and *pgx6* (66.4%) of *Fusarium oxysporum* which have been shown to reduce pathogen virulence when both are mutated simultaneously [74]; two showed similarity (61.6% and 41.8%) to the *PecA* polygalacturonase gene of *Aspergillus flavus*, a CWDE which primarily degrades pectin, and has been shown to improve pathogen invasion and increase spread during infection [73]; one with similarity to the pectin methylesterase *Bcpme1* gene of *B. cinerea* [78]; four with similarity (45.7% - 63.5%) to *PelA* and *PelD*, two pectate lyase virulence factors of *Nectria haematococca* [75]; and a further five obtained a pectate lyase Pfam hit, of which four showed similarity (40.3% - 53.5%) to the *Pnl1* pectin lyase gene of citrus pathogen *Penicillium digitatum* [76] and one with 58.4% similarity to *PelB* pectate lyase B gene of *Colletotrichum gloeosporioides*, seen to affect virulence on avocado [77]. A further five prioritised candidate CWDE’s, classed as hemicellulose-degrading enzymes, showed similarity (46.7% - 61.6%) to the endo-1,4-beta-xylanases (glycosyl hydrolase families 10 and 11) of *M. oryzae*, the knockdown of which is seen to reduce pathogenicity [80]. The remaining two prioritised CWDE’s, classed as cellulose-degrading enzymes, showed 51.9% and 52.9% similarity to the *Glu1* glucanase gene, a known virulence factor of wheat pathogen *Pyrenophora tritici-repentis* [79]. The similarities seen between these predicted secreted CAZymes and known virulence factors provides a collection of likely CWDE’s of *E. fawcettii* for future investigation. Unlike SP’s or CE’s, predicted CWDE’s of *E. fawcettii* were not overrepresented among genes found at the contig edge or within 2 Kb of an AT-rich region (S8). There was some crossover between CE’s and CWDE’s, with five *E. fawcettii*
proteins being labelled as both prioritised CE’s and prioritised CWDE’s, thus providing some CE’s with potential carbohydrate-interacting functions.

Table 5. Predicted function of prioritised candidate cell wall degrading enzymes of *Elsinoë fawcettii*

<table>
<thead>
<tr>
<th>Gene accession</th>
<th>PHI-base hit</th>
<th>Similarity (%)</th>
<th>Top Pfam hit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicted pectin-degrading enzymes:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D9617_30g011650</td>
<td>PGX6 <em>Fusarium oxysporum</em> (PHI:4880)</td>
<td>66.39</td>
<td>Glycosyl hydrolases family 28 (GH28)</td>
</tr>
<tr>
<td>D9617_1g083410</td>
<td>PG1 <em>F. oxysporum</em> (PHI:4879)</td>
<td>53.69</td>
<td>GH28</td>
</tr>
<tr>
<td>D9617_17g047910</td>
<td>PECA <em>Aspergillus flavus</em> (PHI:88)</td>
<td>61.64</td>
<td>GH28</td>
</tr>
<tr>
<td>D9617_61g013180</td>
<td>PECA <em>A. flavus</em> (PHI:88)</td>
<td>41.80</td>
<td>GH28</td>
</tr>
<tr>
<td>D9617_36g063380</td>
<td>BCPME1 <em>Botrytis cinerea</em> (PHI:278)</td>
<td>47.97</td>
<td>Pectinesterase</td>
</tr>
<tr>
<td>D9617_23g005810</td>
<td>PelD <em>Nectria haematococca</em> (PHI:180)</td>
<td>47.27</td>
<td>Pectate lyase (PL)</td>
</tr>
<tr>
<td>D9617_10g074300</td>
<td>PelD <em>N. haematococca</em> (PHI:180)</td>
<td>63.45</td>
<td>PL</td>
</tr>
<tr>
<td>D9617_18g032830</td>
<td>PelA <em>N. haematococca</em> (PHI:179)</td>
<td>46.38</td>
<td>PL</td>
</tr>
<tr>
<td>D9617_32g092100</td>
<td>PelA <em>N. haematococca</em> (PHI:179)</td>
<td>45.69</td>
<td>PL</td>
</tr>
<tr>
<td>D9617_22g066030</td>
<td>PNL1 <em>Penicillium digitatum</em> (PHI:3226)</td>
<td>53.46</td>
<td>PL</td>
</tr>
<tr>
<td>D9617_1g083530</td>
<td>PNL1 <em>P. digitatum</em> (PHI:3226)</td>
<td>44.74</td>
<td>PL</td>
</tr>
<tr>
<td>D9617_2g054490</td>
<td>PNL1 <em>P. digitatum</em> (PHI:3226)</td>
<td>41.70</td>
<td>PL</td>
</tr>
<tr>
<td>D9617_1g083610</td>
<td>PNL1 <em>P. digitatum</em> (PHI:3226)</td>
<td>40.33</td>
<td>PL</td>
</tr>
<tr>
<td>D9617_23g006380</td>
<td>PELB <em>Colletotrichum gloeosporioides</em> (PHI:222)</td>
<td>58.40</td>
<td>PL</td>
</tr>
<tr>
<td><strong>Predicted Hemicellulose-degrading enzymes:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D9617_9g026290</td>
<td>Endo-1,4-beta-xylanase <em>Magnaporthe oryzae</em> (PHI:2204)</td>
<td>61.56</td>
<td>Glycosyl hydrolase family 10 (GH10)</td>
</tr>
<tr>
<td>D9617_18g032910</td>
<td>Endo-1,4-beta-xylanase <em>M. oryzae</em> (PHI:2204)</td>
<td>57.69</td>
<td>GH10</td>
</tr>
<tr>
<td>D9617_3g022390</td>
<td>Endo-1,4-beta-xylanase <em>M. oryzae</em> (PHI:2208)</td>
<td>46.67</td>
<td>GH10</td>
</tr>
<tr>
<td>D9617_36g063160</td>
<td>Endo-1,4-beta-xylanase <em>M. oryzae</em> (PHI:2214)</td>
<td>58.87</td>
<td>Glycosyl hydrolases family 11 (GH11)</td>
</tr>
<tr>
<td>D9617_1g082440</td>
<td>Endo-1,4-beta-xylanase <em>M. oryzae</em> (PHI:2213)</td>
<td>56.72</td>
<td>GH11</td>
</tr>
<tr>
<td><strong>Predicted Cellulose-degrading enzymes:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D9617_40g012710</td>
<td>GLU1 <em>Pyrenophora tritici-repentis</em> (PHI:3859)</td>
<td>52.89</td>
<td>Cellulase - glycosyl hydrolase family 5 (GHS)</td>
</tr>
<tr>
<td>D9617_8g049020</td>
<td>GLU1 <em>P. tritici-repentis</em> (PHI:3859)</td>
<td>51.93</td>
<td>Cellulase – GHS</td>
</tr>
</tbody>
</table>
Prediction of secondary metabolite clusters

Much research surrounding *E. fawcettii* has focused on the SM elsinochrome, which contributes to the formation of necrotic lesions [25-28]. Analysis of the *E. fawcettii* genome assembly enabled the prediction of further genes potentially involved in the elsinochrome gene cluster than previously described, as well as the prediction of additional SM clusters throughout the assembly. In total, there were 22 predicted SM clusters, involving 404 (4.0%) genes (Table 6, S9). Comparing this to the results of the 10 comparative species showed that the number of predicted SM clusters varies widely among the pathogens, from 13 clusters (*U. maydis*) to 53 clusters (*M. oryzae*) (Figure 6). This wide variety among fungal species, in particular an overrepresentation of SM clusters among hemibiotrophs and necrotrophs has been seen before [154]. From the comparative analysis, it appears *E. fawcettii* has a lighter dependence upon the variety of secondary metabolite clusters compared to the other necrotrophs and hemibiotrophs, particularly for T1PKS clusters. Blast analysis of the previously determined *E. fawcettii* elsinochrome cluster [27] against the *E. fawcettii* proteome indicated high similarities in amino acid sequence for six genes of the predicted Type I Polyketide synthase (T1PKS) SM cluster 1 (S9). Specifically, the predicted core biosynthetic gene of cluster 1 (accession D9617_1g081920) showed 98.6% similarity to the *E. fawcettii* polyketide synthase (*EfPKS1*) gene (accession ABU63483.1). An additional predicted biosynthetic gene (accession D9617_1g081900) had 99.6% similarity to the *E. fawcettii* ESC reductase (*RDT1*) gene (accession ABZ01830) and the predicted transport-related gene (accession D9617_1g081940) showed 70.3% similarity to the *E. fawcettii* ECT1 transporter (*ECT1*) gene (accession ABZ82008). Additional genes within the *E. fawcettii* SM cluster 1 obtained hits to the *E. fawcettii* elsinochrome cluster [27], specifically D9617_1g081930, D9617_1g081910 and D9617_1g081890 had high (97.4% - 100%)
similarity to PRF1 prefoldin protein subunit 3 (accession ABZ01833.1), TSF1 transcription factor (accession ABZ01831.1) and EfHP1 coding a hypothetical protein (accession ABZ82009.1). Hence, SM cluster 1 contains the two genes, EFPKS1 and TSF1, which have been shown to be essential in elsinochrome production, as well as four genes (RDT1, PRF1, ECT1 and EfHP1) also thought to be involved in elsinochrome biosynthesis [26, 27]. SM cluster 1 appears to lack four genes, being OXR1, EfHP2, EfHP3 and EfHP4, which have all been reported to code for hypothetical proteins and not thought to be involved in biosynthesis [27]. However, to further investigate these omissions, BLAST analysis querying the nucleotide sequences of the elsinochrome cluster [27] against the contigs of the
E. fawcettii genome assembly indicated regions with high similarities (99.3% - 99.7%) consistent with the location of predicted SM cluster 1 on contig 1. This suggests that these unnecessary nearby genes may have become slightly degraded in the E. fawcettii BRIP 53147a isolate and were therefore not recognised during gene prediction. The use of alternate gene model prediction programs between the studies may also be a contributing factor. These differences may be further investigated through future transcriptomics analyses of E. fawcettii. Interestingly, SM cluster 1 consisted of an additional nine genes to the elsinochrome cluster previously described [27], all of which lay in a cluster adjacent to ECT1. Several of these additional genes obtained Pfam hits such as the THUMP domain, peptidase M3, Apolipoprotein O, Gar1/Naf1 RNA binding region and Endonuclease/Exonuclease/phosphatase family, suggesting these additional neighbouring proteins may perform functions such as RNA binding and modification, peptide cleavage, lipid binding and intracellular signalling, thus providing further genes for future investigation into the elsinochrome biosynthesis pathway.
<table>
<thead>
<tr>
<th>Cluster #</th>
<th>SM class</th>
<th>Genomic location (number of genes involved)</th>
<th>Similarity to known SM biosynthetic gene clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T1PKS</td>
<td>Contig_1, 641093:686753 (15 genes)</td>
<td>Elsinochrome A/B/C:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EfHP1 hypothetical protein (ABZ82009.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ESC reductase (ABZ01830.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Transcription factor (ABZ01831.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Polyketide synthase (ABU63483.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ESC prefoldin protein subunit 3 (ABZ01833.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ECT1 transporter (ABZ82008.1)</td>
</tr>
<tr>
<td>2</td>
<td>terpene-T1PKS</td>
<td>Contig_1, 1100227:1205433 (43 genes)</td>
<td>PR toxin:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Short-chain dehydrogenase/reductase SDR (CDM31317.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Aristolochene synthase (CDM31315.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FAD-binding, type 2 (CDM31316.1)</td>
</tr>
<tr>
<td>3</td>
<td>other</td>
<td>Contig_2, 204508:248496 (18 genes)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>other</td>
<td>Contig_2, 1497538:1541073 (22 genes)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>terpene</td>
<td>Contig_3, 564086:586459 (10 genes)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>terpene</td>
<td>Contig_3, 907579:930486 (11 genes)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>other</td>
<td>Contig_4, 582204:627436 (23 genes)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>other</td>
<td>Contig_6, 282237:328303 (19 genes)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>other</td>
<td>Contig_6, 329430:373960</td>
<td></td>
</tr>
<tr>
<td>Contig</td>
<td>Accession</td>
<td>Description</td>
<td>Genes</td>
</tr>
<tr>
<td>--------</td>
<td>-----------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>10</td>
<td>other</td>
<td>Contig_6, 783514:830534</td>
<td>(19 genes)</td>
</tr>
<tr>
<td>11</td>
<td>terpene</td>
<td>Contig_7, 20929:44027</td>
<td>(17 genes)</td>
</tr>
<tr>
<td>12</td>
<td>T1PKS</td>
<td>Contig_7, 199413:248702</td>
<td>(11 genes)</td>
</tr>
<tr>
<td>13</td>
<td>NRPS</td>
<td>Contig_8, 153859:208507</td>
<td>(25 genes)</td>
</tr>
<tr>
<td>14</td>
<td>other</td>
<td>Contig_9, 468080:512558</td>
<td>(21 genes)</td>
</tr>
<tr>
<td>15</td>
<td>NRPS</td>
<td>Contig_15, 163571:217225</td>
<td>(15 genes)</td>
</tr>
<tr>
<td>16</td>
<td>terpene</td>
<td>Contig_20, 268495:289017</td>
<td>(18 genes)</td>
</tr>
<tr>
<td>17</td>
<td>T1PKS</td>
<td>Contig_25, 66786:116804</td>
<td>(9 genes)</td>
</tr>
<tr>
<td>18</td>
<td>T1PKS</td>
<td>Contig_28, 107087:155682</td>
<td>(18 genes)</td>
</tr>
<tr>
<td>19</td>
<td>T3PKS</td>
<td>Contig_34, 15555:58185</td>
<td>(21 genes)</td>
</tr>
</tbody>
</table>

**Trypacidin:**
- Putative toxin biosynthesis regulatory protein AfJ (EAL89340.1) (43 genes)
- Hypothetical protein (EAL89347.1) (72 genes)
- Putative metallo-beta-lactamase domain protein (EAL89338.1) (57 genes)
- Putative polyketide synthase (EAL89339.1) (59 genes)

**Pestheic acid:**
- PtaD (AGO59044.1) (57 genes)
- PtaB (AGO59041.1) (63 genes)
- PtaA (AGO59040.1) (59 genes)

**Cercosporin:**
- Polyketide synthase (AAT69682.1) (53 genes)
- Cercosporin toxin biosynthesis protein (ABC79591.2) (52 genes)
- Oxidoreductase (ABK64184.1) (41 genes)
- O-methyltransferase (ABK64180.1) (61 genes)
- Oxidoreductase (ABK64182.1) (60 genes)
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>NRPS</td>
<td>Contig_35, 41518:95090 (22 genes)</td>
</tr>
<tr>
<td>21</td>
<td>NRPS</td>
<td>Contig_37, 59764:106480 (19 genes)</td>
</tr>
<tr>
<td>22</td>
<td>other</td>
<td>Contig_59, 16530:45727 (15 genes)</td>
</tr>
</tbody>
</table>
Comparison of numbers of predicted secondary metabolite gene clusters among 11 fungal species. Numbers of SM gene clusters, shown on the x axis, are divided into SM types; (I) Type I Polyketide synthase (T1PKS); (II) terpene; (III) non-ribosomal peptide synthetase (NRPS); and (IV) other, which contains all clusters identified by antiSMASH as either Type 3 Polyketide synthase (T3PKS), terpene-T1PKS, indole-T1PKS-NRPS, T1PKS-NRPS, indole-T1PKS, T1PKS-terpene-NRPS, indole, siderophore, lantipeptide, T3PKS-T1PKS or other.

An additional predicted SM cluster deserving of further investigation was SM cluster 2, a terpene-T1PKS, located 415,394 bp from the elsinochrome SM cluster 1 on contig 1. This cluster shows sequence similarity to three proteins within the PR toxin biosynthetic gene cluster, namely aristolochene synthase (accession CDM31315.1) with 60% similarity to D9617_1g083910, short-chain dehydrogenase/reductase (accession CDM31317.1) with 54% similarity to D9617_1g083830 and the type 2 FAD-binding protein (accession CDM31316.1) with 42% similarity to D9617_1g083960. The PR toxin is produced by the saprobe Penicillium roqueforti, a known contaminant of silages [155], while the mechanisms of its likely role in plant degeneration are unknown [156], PR toxin is seen to induce necrosis in human intestinal epithelial cells and monocytic immune cells [157] and exhibits mutagenic activity towards rats [158]. Thus, indicating the potential production of a toxin by E. fawcettii with DNA-binding capabilities. Another predicted SM gene cluster of interest was the T1PKS SM cluster 12. Three genes of cluster 12 (D9617_7g030040, D9617_7g030060 and D9617_7g030070) showed similarity to multiple known biosynthetic
genes clusters; including the pestheic acid biosynthetic gene cluster of *Pestalotiopsis fici* [159] thought to function as a plant growth regulator [160] and the Trypacidin biosynthetic gene cluster of *Aspergillus fumigatus*, which produces a SM toxic to human lung cells [161].

Lastly, SM cluster 18 is predicted to code for five proteins with sequence similarity to those of the cercosporin biosynthetic gene cluster of *Cercospora nicotianae* [162]. Specifically, D9617_28g065380 (53% similarity to polyketide synthase, accession AAT69682.1), D9617_28g065390 (52% similarity to cercosporin toxin biosynthesis protein, accession ABC79591.2), D9617_28g065400 (41% similarity to oxidoreductase, accession ABK64184.1), D9617_28g065420 (61% similarity to O-methyltransferase, accession ABK64180.1) and D9617_28g065450 (60% similarity to oxidoreductase, accession ABK64182.1). Cercosporin, similar to elsinochrome, is a fungal toxin which promotes the generation of reactive oxygen species in the presence of light, killing plant cells [163]. Cercosporin produced by *C. nicotianae* has been shown to cause necrotic lesions on tobacco leaves [164] and is also produced by the apple pathogen *Colletotrichum fioriniae* [165]. While it has been shown that elsinochrome production is important for full virulence by *E. fawcettii* [26, 27], biosynthesis of further SM’s, such as cluster 2, 12 or 18, may be beneficial to pathogenesis by potentially disrupting host plant signalling, causing additional necrosis or inhibiting competing microbes.

Analysis of the distances between predicted SM genes and TE’s indicated no TE’s were in the close vicinity of SM cluster 1 (elsinochrome), the closest TE to the edge of the cluster was 199,748 bp or 77 genes away. This lack of association was seen among all *E. fawcettii* predicted SM clusters, with seven clusters predicted on contigs without identified TE’s (S9). Of those clusters which did lie on contigs with TE’s, genes were an average distance of
236,556 bp away, suggesting recent activity of known TE’s was unlikely to be involved in the formation of *E. fawcettii* SM clusters. The closest AT-rich region to SM cluster 1 was a distance of 90,363 bp, while this was less than the mean distance (257,863 bp), this indication of potential TE degradation by RIP is still quite distant. In contrast to multiple SP’s and CE’s seen in the close vicinity of AT-rich regions, there were no genes from predicted SM clusters within 2 Kb of an AT-rich region, suggesting genes involved in SM production may benefit from residing in more stable genomic regions.

**Conclusion:**

The WGS sequencing, genome mining and comparative analyses conducted in this study illustrates the potential that exists within the genome of *E. fawcettii* for virulence factors such as protein effectors and CWDE’s. The identification of these potential pathogenicity-related genes is a first step in determining further mechanisms utilised by *E. fawcettii* in addition to elsinochrome production, thus enabling this pathogen to defeat plant immune strategies in a host-specific manner. This study provides predicted virulence genes for future experimental investigation of *E. fawcettii* pathogenesis pathways, as well as establishing a comprehensive genomic resource for use in future studies to determine improved methods of control and screening of this pathogen.

**Acknowledgments:**

We would like to thank members of the Centre for Crop Health at the University of Southern Queensland for their time and work. In particular, Lauren Huth and Katelynn Hadzi for
providing laboratory and organisational support for the project, and Levente Kiss for
providing guidance throughout this study.

References:


897 8. Whiteside J. Biological characteristics of *Elsinoe fawcettii* pertaining to the

899 9. Whiteside J. Pathogenicity of two biotypes of *Elsinoë fawcettii* to sweet orange and
some other cultivars. Phytopathology. 1978;68(1).

901 10. Wang LY, Liao HL, Bau HJ, Chung KR. Characterization of pathogenic variants of
*Elsinoë fawcettii* of citrus implies the presence of new pathotypes and cryptic species

904 11. Queensland Government. Department of Agriculture and Fisheries. DAF Biological
Collections [cited 2019 May 10]. Available from:

907 12. dos Santos RF, Spósito MB, Ayres MR, Sosnowski MR. Phylogeny, morphology and
pathogenicity of *Elsinoë ampelina*, the causal agent of grapevine anthracnose in

caused by a putative new pathotype of *Elsinoë australis*. Plant disease.
2012;96(5):629-34.

913 14. Miles AK, Tan YP, Shivas RG, Drenth A. Novel pathotypes of *Elsinoë australis*
associated with *Citrus australasica* and *Simmondsia chinensis* in Australia. Tropical

916 15. Kokoa P. Review of sweet potato diseases in PNG. Food Security for Papua New
Guinea Proceedings of the Papua New Guinea Food and Nutrition 2000 Conference
16. Mau YS. Resistance response of fifteen sweet potato genotypes to scab disease
(Sphaceloma batatas) in two growing sites in East Nusa Tenggara, Indonesia. Tropical Drylands. 2018;2(1):5-11.


Shieh MT, Brown RL, Whitehead MP, Cary JW, Cotty PJ, Cleveland TE, et al. Molecular genetic evidence for the involvement of a specific polygalacturonase, P2c,


1265 130. Steiner, L., Findeiß, S., Lechner, M., Marz, M., Stadler, Peter F., Prohaska, Sonja J.


1295 141. Bao W, Kojima K, Kohany O. Repbase Update, a database of repetitive elements in
the gap; seven reasons to close fragmented genome assemblies. Fungal Genetics and
analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis
1303 144. Cambareri EB, Jensen BC, Schabtach E, Selker EU. Repeat-induced G-C to A-T
1306 145. Selker EU, Cambareri EB, Jensen BC, Haack KR. Rearrangement of duplicated DNA in
1308 8674(87)90097-3.
1309 146. Selker EU. Premeiotic Instability of Repeated Sequences in Neurospora Crassa.
1312 147. Gladyshev E. Repeat-Induced Point Mutation (RIP) and Other Genome Defense
1314 148. Braumann I, Berg M, Kempken F. Repeat induced point mutation in two asexual
1317 149. Gout L, Fudal I, Kuhn ML, Blaise F, Eckert M, Cattolico L, et al. Lost in the middle of
nowhere: the AvrLm1 avirulence gene of the Dothideomycete Leptosphaeria

1319  maculans. Molecular Microbiology. 2006;60(1):67-80. doi: 10.1111/j.1365-
1320  2958.2006.05076.x.
1321  150. Fudal I, Ross S, Brun H, Besnard AL, Ermel M, Kuhn ML, et al. Repeat-Induced Point
1322  Mutation (RIP) as an alternative mechanism of evolution toward virulence in
1325  151. Kloppholz S, Kuhn H, Requena N. A Secreted Fungal Effector of Glomus intraradices
1326  Promotes Symbiotic Biotrophy. Current Biology. 2011;21(14):1204-9. doi:
1327  10.1016/j.cub.2011.06.044.
1328  152. Mesarich CH, Bowen JK, Hamiaux C, Templeton MD. Repeat-containing protein
1329  effectors of plant-associated organisms. Frontiers in Plant Science. 2015;6. doi:
1330  10.3389/fpls.2015.00872.
1332  of two filamentous pathogens target plant salicylate biosynthesis. Nature
1333  Communications. 2014;5(1). doi: 10.1038/ncomms5686.
1337  analysis of maize silage by LC-MS/MS. Analytical and Bioanalytical Chemistry.
1339  156. Dubey MK, Aamir M, Kaushik MS, Khare S, Meena M, Singh S, et al. PR Toxin -
1340  Biosynthesis, Genetic Regulation, Toxicological Potential, Prevention and Control
1341  Measures: Overview and Challenges. Frontiers in Pharmacology. 2018;9. doi:


Supporting information captions:

S1 Table. GenBank accessions for ITS and TEF1-α sequences included in the phylogenetic analysis with *Elsinoë fawcettii* isolate (BRIP 53147a).

S2 Table. Comparison of predicted gene classifications among *Elsinoë fawcettii* and 10 other species; Pfam hits, predicted CAZymes and core/acc/unique genes.

S3 Text. Sequence alignment of partial ITS and TEF1-α regions of *Elsinoë fawcettii* (BRIP 53147a) in comparison with other *E. fawcettii* isolates and closely related *Elsinoë* species.

S4 Table. Comparison of results of EffectorP predicted candidate effectors and alternate candidate effector search among 11 species.

S5 Table. Genomic and proteomic analyses of 11 species for use in known effector analysis and candidate effector prioritisation.
S6 Table. Comparison of numbers of predicted secreted proteins, candidate effectors and prioritised candidate effectors among 11 species.

S7 Table. Features and GenBank accessions of 203 *Elsinoë fawcettii* candidate effectors.

S8 Table. Features and GenBank accessions of 378 *Elsinoë fawcettii* predicted CAZymes.

S9 Table. Features and GenBank accessions of 404 *Elsinoë fawcettii* genes with predicted involvement in secondary metabolite clusters.
Fig 1

This figure shows the number of predicted proteins for various species, classified into core, accessory, and unique categories. The species listed are:

- Elsinoë fawcettii
- Botrytis cinerea
- Parasagognosphora nodorum
- Pyrenophora tritici-repentis
- Sclerotinia sclerotiorum
- Zymoseptoria tritici
- Leptosphaeria maculans
- Magnaporthe oryzae
- Rhynchosporium commune
- Verticillium dahliae
- Ustilago maydis

The x-axis represents the number of predicted proteins, ranging from 0 to 15,000. The y-axis lists the species, with each species represented by a bar that is divided into three sections corresponding to core, accessory, and unique proteins.
Fig 2

This phylogenetic tree shows the relationships among various species of the genus *Elsinoë*. The tree includes species such as *Elsinoë fawcettii Jin-6* (Citrus sunki), *Elsinoë fawcettii* BRIP 53147a (Citrus limon)*, and others, each associated with their respective plant hosts. The tree illustrates the evolutionary connections and branchings among these species, providing insights into their taxonomic relationships. The scale bar at the bottom indicates the evolutionary distance, with 0.050 units representing the change in genetic distance.
One point for each feature:
- IFR on at least one side of gene is >median
- No involvement in predicted SM gene clusters
- GC content of CDS is $<Q_1$ or $>Q_3$
- Within 10 genes of gene with a specified Pfam hit
- Unique or obtained same orthoMCL ID as a known effector

CE’s scored out of a possible 5 points:
- *Elsinoë fawcettii*, *Parastagonospora nodorum*, *Pyrenophora triticic-repentis*, *Verticillium dahliae* & *Ustilago maydis*

Additional points:

<table>
<thead>
<tr>
<th>Genomes with &gt;2% TE coverage:</th>
<th>Additional points:</th>
<th>Genomes with &gt;2% TE coverage &amp; &gt;25% AT-rich regions:</th>
<th>Genomes with &gt;2% TE coverage &amp; &gt;25% AT-rich regions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Within 7 genes of a TE region</td>
<td>CE’s scored out of a possible 6 points:</td>
<td>- Distance from gene to closest AT-rich region is $&lt;Q_1$ value</td>
<td>CE’s scored out of a possible 7 points:</td>
</tr>
<tr>
<td></td>
<td>- <em>Zymoseptoria tritici</em>, <em>Sclerotinia sclerotiorum</em>, <em>Botrytis cinerea &amp; Magnaporthe oryzae</em></td>
<td></td>
<td>- <em>Rhynchosporium commune &amp; Leptosphaeria maculans</em></td>
</tr>
</tbody>
</table>

Fig 4
Fig 5

Bar chart showing the number of proteins for different species, categorized as:
- Predicted secreted proteins
- Candidate effectors
- Prioritised candidate effectors
Fig 6

The bar chart illustrates the number of secondary metabolite (SM) gene clusters for various species. Each species is represented by a bar divided into segments, each color-coded as follows:
- **Blue**: T1pks
- **Teal**: Terpene
- **Light Green**: NRPS
- **Yellow**: Other

Species included in the chart are:
- Elsinoë fawcettii
- Botrytis cinerea
- Parastagonospora nodorum
- Pyrenophora tritici-repentis
- Sclerotinia sclerotiorum
- Zymoseptoria tritici
- Leptosphaeria maculans
- Magnaporthe oryzae
- Rhynchosporium commune
- Verticillium dahliae
- Ustilago maydis

The x-axis represents the number of SM gene clusters, while the y-axis lists the species.