Comparative genome analysis indicates rapid evolution of pathogenicity genes in *Colletotrichum tanaceti*

Ruvini V. Lelwala[1*], Pasi K. Korhonen[1], Neil D. Young[1], Jason B. Scott[2], Peter K. Ades[3], Robin B. Gasser[1], Paul W. J. Taylor[1]

[1] Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, Victoria, Australia

[2]Tasmanian Institute of Agriculture, Burnie, Tasmania, Australia

[3]Faculty of Science, The University of Melbourne, Parkville, Victoria, Australia

*Corresponding author

Emails: llelwala@student.unimelb.edu.au

ABSTRACT

Colletotrichum tanaceti is an emerging foliar fungal pathogen of pyrethrum (*Tanacetum cinerariifolium*), posing a threat to the global pyrethrum industry. Despite being reported consistently from field surveys in Australia, the molecular basis of pathogenicity of *C. tanaceti* on pyrethrum is unknown. Herein, the genome of *C. tanaceti* (isolate BRIP57314) was assembled *de novo* and annotated using transcriptomic evidence. The inferred pathogenicity gene suite of *C. tanaceti* comprised a large array of genes encoding secreted effectors, proteases, CAZymes and secondary metabolites. Comparative analysis of its CAZyme pathogenicity profiles with those of closely related species suggested that *C. tanaceti* had additional hosts to pyrethrum. The genome of *C. tanaceti* had a high repeat content and repetitive elements were located significantly closer to genes inferred to influence pathogenicity than other genes. These repeats are likely to have accelerated mutational and transposition rates in the genome, resulting in a rapid evolution of certain CAZyme families in this species. The *C. tanaceti* genome consisted of a gene-sparse, A-T rich region facilitating a "two-speed" genome. Pathogenicity genes within this region were likely to have a higher evolutionary rate than the 'core' genome. This "two-speed" genome phenomenon in certain *Colletotrichum* spp. was hypothesized to have caused the clustering of species based on the pathogenicity genes, to deviate from taxonomy. With the large repertoire of pathogenicity factors that can potentially evolve rapidly in response to control measures, *C. tanaceti* may pose a high-risk to global pyrethrum production. Knowledge of the pathogenicity genes will facilitate future research in disease management of *C. tanaceti* and other *Colletotrichum* spp..

Key words: *Colletotrichum*, Genome, Pyrethrum, Pathogenicity gene suite, Repeats, Evolution

INTRODUCTION

Plant pathogens cause diseases world-wide that have devastating economic, social and ecological consequences [1]. Fungi are among the dominant causal agents of plant diseases [2] and the genus *Colletotrichum* has been ranked among the top-ten most important fungal plant pathogens [3]. Many *Colletotrichum* species are known to cause major economic losses globally, and have been extensively used in the study of the molecular and cellular bases of fungal pathogenicity [4]. The publication of 25 whole genome sequences of *Colletotrichum* species has significantly improved understanding of the biology, genetics and evolution of this genus [5-11]. However, a large research gap still exists with this ever-expanding genus consisting of more than 200 accepted species [12] and 14 major species complexes [13, 14]. The availability of only one genome of a member of the destructivum complex, *C. higginsianum,* [5, 15] has constrained comparative studies within and among species complexes. Insights into the genomic organization and the pathogenicity gene repertoire of other *Colletotrichum* species in the destructivum complex therefore, will significantly expand the knowledge base of this important genus.

Colletotrichum tanaceti, a member of the destructivum complex [16], is an emerging foliar fungal pathogen [17] of Dalmatian pyrethrum (*Tanacetum cinerariifolium*). Pyrethrum is commercially cultivated as a source of the natural insecticide pyrethrin [18]. *Colletotrichum tanaceti* has been consistently reported in Australian field surveys of the crop [19] since 2013 [17] and causes leaf anthracnose, with black, water-soaked, sunken lesions [17]. Due to its hemibiotrophic lifestyle, characteristic symptoms of *C. tanaceti* are not evident on leaves until around 120 hours after infection [17, 20], when it switches from biotrophy to necrotrophy. A significant reduction in green leaf area occurs usually 10 days after infection [17]. This suggests a rapid disease cycle for *C. tanaceti* in pyrethrum and, given its aggressiveness, the potential for serious crop damage.

The molecular basis of pathogenicity of *C. tanaceti,* which includes the pathogenicity genes and their evolution, has not been studied. *Colletotrichum tanaceti* has only been reported from pyrethrum in Australia, but may have crossed over from another plant host species. However, cross-host pathogenicity has not yet been assessed and the pathogen's origin and the potential host range are currently unknown. Therefore, the threat posed by *C. tanaceti* to the local and global pyrethrum industry remains largely unknown.

The genome sequence of an emerging plant pathogen such as *C. tanaceti* can provide a foundation for identifying genes associated with the pathogen life cycle, pathogenicity and virulence. Effectors [21], proteases [22], and carbohydrate active enzymes (CAZymes) that [23] are important gene categories in fungal pathogenesis. Furthermore secondary metabolites and transporters, *P450s* and transcription factors [24] associated with biosynthesis of secondary metabolites are also important pathogenicity factors. Fungal mitogen activated protein (MAP) kinase pathways regulate the cascade of reactions that respond to various environmental stresses and are also important factors determining pathogenicity and virulence [25]. Draft genomes of many fungal pathogens have been used to infer genes involved in pathogenicity with a high accuracy [26, 27] using homology searches against curated databases [28, 29] and *de novo* inference using bioinformatics tools [21, 30]. Therefore, characterization of the genome of *C. tanaceti*, followed by inference and quantification of these important pathogenicity gene categories will be beneficial for future functional and pathogenicity studies of this and related pathogens.

Comparative genomics has enabled inference of patterns of speciation, pathogenesis and host determination within *Colletotrichum* lineages [31]. These studies have indicated that the gain and loss of putative pathogenicity gene families in *Colletotrichum* genomes are important determinants of host specificity and pathogenic adaptation of these species [7, 11].

Comparison of putative pathogenicity gene repertoires of *Colletotrichum* species from different species complexes and species closely related to the genus *Colletotrichum* would provide insights into the evolutionary rates of these genes. Comparative genomics will also enable the quantification of pathogenicity at the molecular level and identification of the host range of *C. tanaceti* with respect to other *Colletotrichum* species. Therefore, combined genomics and comparative genomics analyses can provide sound means of assessing the current and future risks posed by *C. tanaceti*.

In order to achieve the major goal of evaluating the potential threat to the pyrethrum industry form *C. tanaceti*, the aims of this study were to: 1) infer the pathogenicity gene suite of *C. tanaceti*; 2) quantify the molecular basis of pathogenicity; 3) infer the host range of *C. tanaceti*; and 4) quantify the rate of evolution of pathogenicity genes in *C. tanaceti*.

MATERIALS AND METHODS

Sequencing and *de novo*-assembly of the genome of *C. tanaceti*

Fungal strain

The ex-holotype of *C. tanaceti* strain BRIP57314 (CBS 132693=UM01) [17] was acquired from the culture collection of BRIP (Plant Pathology Herbarium, Department of Primary Industries, Queensland, Australia). This isolate was propagated on potato dextrose agar (PDA; Sigma Aldrich, St. Louis, USA) and incubated at 24°C using a 12 h:12 h light:dark photoperiod. Genomic DNA was isolated using a modified CTAB protocol [32]. The integrity and quantity of DNA was confirmed by 1.5% agarose gel electrophoresis and a nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, USA)

Genome sequencing and assembly

Genomic DNA was fragmented using a Covaris ultrasonicator (Covaris Inc., Massachusetts, USA) to achieve an average fragment length of 532 base pairs (bp). A genomic DNA library

with an average insert size of 420 bp was constructed using the KAPA Hyper Prep Library Preparation Kit [33] and was paired-end sequenced (2×300 bp reads) using the Illumina Miseq platform (San Diego, USA). The raw reads were filtered for low quality nucleotides and adapters using Trimmomatic [34] (Phred score-33, leading-3, trailing-6, slidingwindow-4:15, minlen-36) to retain 22,871,341 sequences and were profiled using KAT [35]. Filtered reads were then assembled using DISCOVAR *de novo* [36]. The completeness of the assembly was assessed with the Sordaromyceta_*odb9* gene set [37] using the program Benchmarking Universal Single-Copy Orthologs (BUSCO v2) [37] in the Genomics Virtual Laboratory platform [38]. The GC-bias of the genome was detected using OcculterCut version 1.1 with default settings [39].

Prediction of repetitive elements

Species-specific repeats were first inferred using the program RepeatModeler [40], in which the programs RECON [41] and RepeatScout [42] were used. Long terminal repeats (LTRs) were predicted using the program LTR_Finder [43]. The program RepeatMasker v4.0.5 [44] was employed to mask resulting species-specific repeats and LTRs; and applied the program Tandem Repeat Finder (TRF) [45] and the database Repbase v.17.02 [46] to predict and mask interspersed and simple repeats. All repeats predicted were combined using ProcessRepeats command in RepeatMasker.

RNA sequencing

Inoculation of Pyrethrum leaves

Pyrethrum leaves were inoculated using the leaf-sandwich method [47, 48] by placing a fungal ‘mat’ between two pyrethrum leaves in a petridish. Each petri dish was sealed with parafilm and incubated at 24°C with a 12 h-photoperiod. Induced mycelia were harvested at 6, 24 and 48 h after inoculation, and total RNA was extracted using the RNeasy Plant Mini

kit (Qiagen, Australia) following the manufacturer's instructions. Total RNA was extracted from the saprobic stage (1-week-old cultures growing on potato dextrose agar). Contaminating genomic DNA was removed from RNA samples by Ambion™ DNase I (Thermo Fisher Scientific, USA) treatment; the integrity and quantity of total RNA was confirmed by 1% agarose gel electrophoresis and the Experion™ automated electrophoresis system (Biorad Laboratories, Australia).

RNA libaries were prepared using both E7530L and E&335L NEBNext® Ultra™ RNA Library Prep Kits (New England Biolabs, USA) to generate fragment sizes of 351-371 bp. The transcriptome was paired-end sequenced (2 × 150 bp reads) on the Illumina Hiseq 2500 platform (San Diego, USA). Raw reads were trimmed for quality using Trimmomatic [34] (leading-25, trailing-25, slidingwindow-4:25, minlen-40) to retain between 17,935,938 – 18,761,773 sequences for each library and profiled using FastQC [49].

Gene prediction

Genes were first predicted using the MAKER3 v3.0.0-beta [50], in which both the transcriptomic data from *C. tanaceti* and the proteomic and *ab initio* gene predictions from *C. graminicola;* [51] and *C. higginsianum;* [51] were combined into a consensus prediction. In brief, transcriptomic RNAseq reads of *C. tanaceti* were assembled into transcripts in both *de novo* and genome-guided modes of the program Trinity v2.2.0 [52]. In genome guided assembly, reads were mapped onto the genome using the program TopHat2 v2.1.0 [53]. Genome guided and *de novo* transcriptomic assemblies were combined, redundancy (99% similarity) was removed using the program cd-hit-est [54, 55] and resulting transcripts were filtered for full-length open reading frames (ORFs) using the program Transdecoder [52]. Resulting full-length transcripts were further reduced to 80% similarity using the program cd-hit-est and checked for splicing sites. These high quality transcripts were then used as a

168 training set for *ab initio* gene prediction programs AUGUSTUS v3.1 [56] and SNAP v6.7
169 [57] and GENEMARK v4.2.9 [58]. Evidence data from assembled transcriptomes (with 99%
170 redundancy using cd-hit-est) and the proteomes were provided to Maker3. The predicted
171 genes (length of conceptually translated protein ≥ 30 amino acids) were further clustered
172 using the *k*-means clustering algorithm [59] with following metrics: 1) Maker3 annotation edit
173 distance (AED); 2) number of exons in the mRNA; 3) length of translated protein sequence;
174 4) fraction of exons that overlap transcript alignment; 5) fraction of exons that overlap
175 transcript and protein alignment; 6) fraction of splice sites confirmed by a SNAP prediction
176 from Maker3; 7) percentage for repeat overlap with gene-, exon- and CDS-sequence; 8) size
177 of the inferred orthologous group the gene belongs to using OrthoMclv2.0.9 [60]; and 9)
178 presence of functional annotation (see Functional annotation of the *C. tanaceti* genome
179 section below). Resulting clusters with transposons and *ab initio* gene predictions with no
180 transcriptome or proteome support were removed.

181 **Functional annotation of the *C. tanaceti* genome**

182 Putative coding regions were subjected to protein homology searches against the NCBI (nr)
183 and Swiss-Prot database using BLAST v 2.7.1 (E-value of ≤ 1e-8) [61]. Conserved protein
184 domains and gene ontology (GO) terms were assigned to predicted proteins using
185 InterProScan 5 [62]. Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG)
186 Orthology (KO) terms were assigned to predicted proteins using the Blastkoala search engine
187 [63]. Assigned KO terms were used to generate *C. tanaceti* pathway maps using KEGG
188 mapper [64]. Putative genes of *C. tanaceti* with functional annotations were subjected to
189 species-specific gene enrichment analysis on the DAVID functional annotation database tool
190 [65, 66] and using *C. graminicola* as the reference species.

191 **Comparison to related taxa**

192 The genome and proteome of *C. tanaceti* was compared to genomes of related taxa using
193 genome alignment, synteny and orthology analyses as following.

194 **Genome alignment and synteny analysis**

195 *Colletotrichum tanaceti* genome contigs were aligned to 13 other publicly available genomes
196 (Table 1) of *Colletotrichum* species using nucmer in Mummer v 3.9.4 [67]. Contig-alignments
197 were then filtered for a minimum 30% nucleotide identity and 200 bp in aligned length. The
198 global coverage of each of the genomes by contigs of *C. tanaceti* was computed. The program
199 'Synteny Mapping and Analysis Program', SyMAP v 4.2 [68] was used to map *C. tanaceti*
200 contigs (>150 kb) to the genome with the highest coverage to identify the syntenic regions
201 which are the regions that are in preserved order in chromosomes of *C. higginsianum*
202 IMI349063 reference genome [5].

203 **Table 1.** Genomes used in the comparative genomic analyses

Organism	Identifier[a]	Taxonomy ID[b]	Genbank accession number[c]	Bio project ID[d]	Strain[e]	Assembly[f]	Referenc
Colletotrichum chlorophyti	CCh	708187	MPGH00000000.1	PRJNA350752	NTL11	ASM193710v1	[69]
Colletotrichum fioriniae	CFi	1445577	JARH00000000.1	PRJNA233987	PJ7	GCA_000582985.1	[70]
Colletotrichum fructicola	CFr	1213859	ANPB00000000.1	PRJNA225509	Nara gc5	GCA_000319635.1	[6]
Colletotrichum gloeosporioides	CGl	1237896	AMYD00000000.1	PRJNA176412	Cg-14	GCA_00446055.1	[71]
Colletotrichum graminicola	CGr	645133	ACOD00000000.1	PRJNA37879	M1.001	M1_0001_v1	[5]
Colletotrichum higginsinum	CHi	759273	LTAN00000000.1	PRJNA47061	IMI 349063	GCA 001672515.1	[5]
Colletotrichum incanum	CIn	1573173	LFIW00000000.1	PRJNA286717	MAFF 238704	GCA_001189835.1	[9]
Colletotrichum nymphaeae	CNy	1460502	JEMN00000000.1	PRJNA237763	IMI 504889	GCA_001563115.1	[7]
Colletotrichum orchidophilum	COc	1209926	MJBS00000000.1	PRJNA411788	IMI 309357	GCF_001831195.1	[72]
Colletotrichum orbiculare	COr	1213857	AMCV00000000.1	PRJNA171217	MAFF 240422	Corbiculare240422v01	[6]
Colletotrichum salicis	CSa	1209931	JFFI00000000.1	PRJNA238477	CBS 607.94	GCA_001563125.1	[7]

Colletotrichum simmondsii	CSi	703756	JFBX00000000.1	PRJNA239224	CBS 122122	GCA_001563135	[7]
Colletotrichum sublineola	CSu	1173701	JMSE00000000.1	PRJNA246670	TX430BB	GCA_000696135.1	[73]
Colletotrichum tanaceti	CT1	1306861	PJEX00000000	PRJNA421029	BRIP57314		
Verticillium dahliae	VDh	498257	ABJE00000000.1	PRJNA225532	VdLs.17	GCF_000150675.1	[74]
Botrytis cinerea	BCi	332648	AAID00000000.2	PRJNA15632	B05.10	GCF_000143535.2	[75]
Sordaria macrospora	SMa	771870	CABT00000000.2	PRJNA51569	k-hell	GCF_000182805.2	[76]
Fusarium oxysporum	FOx	426428	AAXH00000000.1	PRJNA18813	CBS 123668	GCF_00149955.1	[77]

[a] Short identifier used in pace of the species name in supplementary information
[b] Taxonomy ID of each species according to the NCBI taxonomy database
[c] Genbank accession number for the deposited nucleotide sequence
[d] NCBI bioproject ID
[e] version of the genome assembly

Orthology search and phylogenomics analysis

The proteomes of *C. tanaceti* and the publicly available 17 other species (Table 1) were subjected to ortholog searching using OrthoMCL v2.0.9 [60] and MCL [78] with an inflation value of 1.5. The orthoMCL output was used to determine the percent orthology among the species and to determine the core gene set for *Colletotrichum.* The ortho-groups with pathogenicity genes (inferred as below) of *C. tanaceti* were extracted and used to determine the percent conservation of those gene categories within the genus. Furthermore the single copy orthologs were extracted from the orthoMCL output and aligned using MAFFT v.7 [79]. These alignments were then trimmed using trimAl v.1.3 [80] to remove all positions in the alignment with gaps in 20% or more of the sequences, unless this leaves less than 60% of the sequence remaining. The trimmed reads were concatenated using FASconCAT-G [81]. The concatenated alignment was partitioned and amino acid substitution models were predicted for each partition using ProtTest 3 [82] in FASconCAT-G. The partitioned, concatenated alignment was subjected to maximum likelihood phylogenetic analysis using RAxML v8.2.10 [83] to find the best tree from 20 maximum likelihood searches and using

100 bootstrap replicates. Evolutionary distance in number of substitutions per site was computed using the *ape* package [84] in the R statistical language framework v 3.5.1. [85] from the maximum likelihood tree.

Estimation of divergence dates

The phylogram developed from above was utilized to estimate the divergence dates of the species considered as following. The final RAxML phylogenetic tree was used to generate an ultrametric tree in r8s v1.81 [86] applying the penalized likelihood method [87] and the truncated Newton (TN) algorithm [88]. Divergence times were estimated using previously derived estimates [8, 11, 89] of 267-430 million years (Myr) for the Leotiomycetes-Sordaromycetes crown, 207-339 Myr for the Sordaromycete crown and 45-75 Myr for the *Colletotrichum* crown as calibrations. An optimal smoothing factor which was deduced using the cross validation process [86] among 50 values across 1 to 6.3e+09 was used in the divergence time estimation.

Prediction of secretome and database searches for identifying other virulence factors

Predicted proteins of *C. tanaceti* were used in downstream prediction of the secretome [90]. A combination of three software tools: SignalPv4.1 [91], Phobius [92] and WoLFPSORT [93] was used to predict the signal peptides. Proteins with transmembrane domains were identified using TMHMM v.2.0 [94] and were excluded as secreted proteins. Proteins with signals targeting the endoplasmic reticulum and GPI anchors were identified and excluded using Ps-SCAN [95] and Pred-GPI [96] respectively. NLStradamus [97] was used to identify proteins with nuclear localization signals. Curated secretome was subjected to homology search against the CDD database to identify the conserved domains (E-value $\leq$ 1e-10). The candidate secreted effector proteins were identified by passing the secretome through the program EffectorP [21]. Predicted effector candidates were manually inspected and candidates with

known plant cell wall degrading catalytic domains, such as cutinases (PF01083.21), short-chain dehydrogenases (PF00106.24), glycosyl hydrolases (PF00457), peptidases (PF04117.11) and lipases (PF13472.5) were excluded. The candidates with no detectable conserved domains and no homology (E-value ≤ 1e-3) to any other proteins in NCBI– non-redundant protein sequence database were defined as species-specific. Putative secreted peptidases and inhibitors were predicted by stand-alone blastp (E-value ≤ 1e-10) homology searches of the domain database of MEROPS release 12.0 [98]. Furthermore, potential virulence factors of *C. tanaceti* were identified by blastp searches (E-value ≤ 1e-10) against PHI-base v 4.4 [28]. The online analysis tools, Antibiotics and Secondary Metabolite Analysis Shell (antiSMASHV.4) [30] with default parameters and SMURF [99] were used to predict potential secondary metabolite backbone genes and clusters using the default parameters. Cytochrome P450s and transporters were described based on blastp (E-value ≤ 1e-10) homology searches against the Fungal Cytochrome P450 database [100] and the Transported Classification Database [101]. The functional annotations for *C. tanaceti* were compared across 17 other closely related taxa (Table1). The family specific Hidden Markov Model profiles of dbCan database v6 [102] were employed using the program HMMScan in HMMER v31.b2 [103] in order to identify the carbohydrate active enzymes (CAZymes) and the CAZyme families in the proteome of *C. tanaceti*. Fungi specific cut-off E-value of 1e-17 and a coverage cut-off of 0.45 [102] were used in the analysis which was repeated for seventeen related species (Table 1). The identified CAZymes were run though InterProScan 5 [62] to check for false positives. The member counts of each CAZyme family for each taxon were corrected accordingly.

Evolution of CAZyme gene families

CAFE v4.0 [104, 105] was used to estimate the number of CAZyme gene family expansions, contractions and the number of rapidly evolving gene families upon divergence of different

lineages. Error-models [105] were estimated to account for the genome assembly errors and were incorporated into computations. A universal lambda value (maximum likelihood value of the birth-death parameter) was assumed and gene families with significant size variance were identified using a probability value cut-off of 0.01. The branches responsible for significant evolution, were further identified using the Viterbi algorithm [104] with a probability value cutoff of 0.05. Sizes of plant pathogenicity-related gene families from CAZomes of each of the species; the 'CAZyme pathogenicity profiles' were retrieved and compared using the online tool ClustVis [106]. The 'CAZyme pathogenicity profile' of a particular species included the gene families that have activities in binding to or degradation of plant cell wall components such as cellulose, hemicelluloses, lignin, pectin and cutin.

Relationship of pathogenicity related genes and repeat elements

The mean distances between repetitive elements and pathogenicity related genes were analyzed using permutation tests implemented in the package regioneR [107] in the R. Repetitive element categories incorporated in this analysis included: 1) tandem and interspersed repeats combined; 2) tandem repeats; and 3) interspersed repeats. These were compared to the pathogenicity related gene classes: 1) CAZymes; 2) peptidases; 3) secondary metabolite biosynthetic gene clusters; and 4) effectors. The mean distance between each gene in above categories and the nearest repetitive element was compared against a distribution of distances of random samples from the whole genome. Ten thousand random iterations were conducted, from which a *Z*-statistic estimate, and its associated probability, were computed for each gene category.

RESULTS

Colletotrichum tanaceti genome and gene content

296 The genome of isolate BRIP57314 was assembled into 5,242 contigs with an N50 value of
297 103,135 bp and assembly size of 57.91Mb. The average GC content was 49.3% (Table 2).
298 The genome size and GC content of *C. tanaceti* was within the range previously reported to
299 other *Colletotrichum* spp. (S1 Fig). Draft genome assembly and the raw unassembled
300 sequences are available under the accession no PJEX00000000 in Genbank. The genome
301 contained 12,172 coding genes with an average gene length of 2,575bp. Mean exon count per
302 gene was 3, and 54.1% of the genome sequence contained protein-encoding genes. In the
303 BUSCO analysis, out of the 3,725 benchmarking genes in the Sordaromyceta group, the
304 genome was reported to contain 3,656 complete BUSCOs (98.2%), of which two were
305 duplicated and the rest were single copy genes (98.1%). A total of 30 (0.8%) BUSCOs were
306 fragmented and 39 were missing (1.0%). The repeat content of *C. tanaceti* was 24.6% of the
307 total genome of which 85.2% was interspersed repeats (Table 3).

308 **Table 1.** Features of the *Colletotrichum tanaceti* BRIP57314 genome

Feature	Statistics
GC_content (%)	49.3
N50 (bp)	103,135
Maximum sequence length (bp)	945,015
Mean length (bp)	11,047
Number of base pairs	57,912,474
Number of contigs	5,242
Number of genes	12,172
Number of exons	35,792
Number of introns	23,620
Number of CDS	12,172
Overlapping genes	3,983
Contained genes	1,586
Mean gene length (bp)	2,575
Mean exon length (bp)	787
Mean intron length (bp)	137
Mean CDS length (bp)	1,440
% of genome covered by genes	54.1
% of genome covered by CDS	30.3
Mean mRNAs per gene	1
Mean exons per mRNA	3
Mean introns per mRNA	2

309

310 **Table 3.** Repetitive elements of the *C. tanaceti* genome

Repetitive element	Number of elements	Length occupied (bp)	Percentage of sequence
SINEs:	49	4,123	0.01
ALUs	0	0	0
MIRs	11	869	0
LINEs:	612	251,619	0.43
LINE1	207	48,554	0.08
LINE2	35	2,588	0
L3/CR1	82	5,928	0.01
LTR elements:	7,299	4,825,086	8.33
ERVL	2	120	0
ERVL-MaLRs	1	39	0
ERV_classI	3	209	0
ERV_classII	1	32	0
DNA elements:	1,436	905,846	1.56
hAT-Charlie	3	140	0
TcMar-Tigger	6	529	0
Unclassified:	8,863	6,153,241	10.62
Total interspersed repeats		12,139,915	20.96
Small RNA:	754	210,370	0.36
Satellites:	0	0	0
Simple repeats:	9,941	1,757,918	3.04
Low complexity:	3,064	147,883	0.26
Total repeat content			24 .62%

311 Of the 12,172 predicted proteins, 11,352 had an annotation edit distance (AED) value of less
312 than 1.0, and 2962 genes had an AED value of zero. The number of genes without putative
313 annotation from the public database searches was only 958. A total of 8,945 proteins (73.5%
314 of proteome) had InterProScan annotations of which 6,911 contained 9,647 Pfam domain
315 annotations and 5,452 had GO term ontology annotation. The most abundant (*n*=129) Pfam
316 domain was the cytochrome P450 family (PF00067) followed by the protein kinase domain
317 (*n*=127; PF00069). Gene enrichment analysis suggested enrichment of many GO terms
318 including those associated with translation and chromosome telomeric region (S1 Table).
319 Putative proteins of *C. tanaceti* were subjected to KEGG pathway analysis which returned

assignment of 5,883 proteins to known pathways (S2 Table). The highest number of KO identifiers was among the metabolic pathway assignments (*n*=693) of which the majority (*n*=363) were for amino acid metabolism followed by carbohydrate metabolism (*n*=290) (S3 Table). Among the environmental information processing pathways, 81 *C. tanaceti* genes were assigned into 47 KO identifiers belonging to MAPK pathway (S4 Table). Furthermore, 24 *C. tanaceti* proteins were annotated with 10 aflatoxin biosynthesis pathway KO assignments (S5 Table) and 56 proteins were assigned KOs for ABC transporters (S6 Table).

Genome alignment and synteny

The global alignment coverage of 13 other *Colletotrichum* genomes from *C. tanaceti* contigs was proportionate to the evolutionary proximity to *C. tanaceti* (Fig 1a). The highest coverage was in *C. higginsianum* (63.8%) and the least was in *C. orbiculare* 4.26%. Among the *C. tanaceti* contigs aligned to the chromosomes of *C. higginsianum*, the best alignment coverage was to chromosome NC_030961.1 (chromosome 9) (S7 Table). *Colletotrichum tanaceti* contigs (n=155 of size≥10 kb) were mapped in SyMAP synteny analysis to form 142 synteny blocks which covered 44.0% of the *C. higginsianum* and 80.0% of the *C. tanaceti* sequences that were used (S2 Fig). Genes were present in 92.0% of the syntenic regions in *C. tanaceti* and in 77.0% of *C. higginsianum*. No inverted synteny blocks were reported. Despite the highest coverage in *C. higginsianum* chromosome 9, the largest synteny block was identified between the complete *C. tanaceti* contig 4 (945.01 kb of length) and *C. higginsianum* chromosome NC_030954 (Chromosome 1). A total of 38 effector candidates of *C. tanaceti* were within these syntenic regions between *C. tanaceti* and *C. higginsianum*. No synteny blocks were detected to the two mini chromosomes (NC_030963.1 and NC_030964.1) of *C. higginsianum*.

Fig 1. Comparison of the *C. tanaceti* genome to previously published *Colletotrichum* spp. genomes. (a) Percentage global alignment (y axis) of 13 *Colletotrichum* draft genomes to contigs representing the *C. tanaceti* draft genome, plotted against evolutionary distance with reference to *C. tanaceti* (x axis), (b) Number of orthologs shared by 13 *Colletotrichum* draft genomes and *C. tanaceti* (y axis) plotted against the evolutionary distance with reference to *C. tanaceti* (x axis); evolutionary distance given in number of substitutions per site, computed using the ape package [98] in R from a maximum likelihood tree.

Orthology search

Of 221,456 total genes from 18 genomes, the number of core genes reported for all ascomycetes in the orthology analysis was 3,944. A total of 10,695 putative proteins from *C. tanaceti* were assigned to 10,074 groups containing orthologs and/or recent paralogs and/or co-orthologs across all species. A total of 6,002 genes were conserved in the genus *Colletotrichum*. *Colletotrichum tanaceti* had 9,679 orthologs with *C. higginsianum* which was the highest ortholog count among *Colletotrichum* spp. followed by 8,855 orthologs with *C. nymphaea* (Fig 1b). Twenty of these groups, with 48 genes among them were exclusive to *C. tanaceti* and were defined as recent paralogs (*in-paralogs*) of *C. tanaceti* with no homology to the 16 other species tested.

Divergence time in *Colletotrichum* lineages

A total of 2,214 single copy ortholog (SCO) genes identified among the *C. tanaceti* and 17 closely related genomes (Table 1) were used to generate a maximum likelihood (ML) evolutionary tree in which all branches achieved bootstrap support of 100%. *Colletotrichum tanaceti* formed a clade with *C. higginsianum*, a member of the destructivum complex and the two destructivum complex members formed a sister clade with the graminicola complex members and *C. incanum.* A smoothing factor value of 1 was reported as the optimal value

367 for divergence time predictions in r8s. *Colletotrichum tanaceti* and *C. higginsianum* were
368 reported to have diverged ~ 9.97 million years ago (mya). The most recent common ancestor
369 (MRCA) of gloeosporioides, graminicola, and acutatum clades were reported to be 6.12,
370 10.98 and 15.78 mya, respectively (Fig 2).

371 **Fig 2.Chronogram showing divergence time estimations (in million years) for**
372 ***Colletotrichum* spp. and related taxa**.

373 **Identification of pathogenicity related genes in *C. tanaceti***

374 **Secretome of *C. tanaceti***

375 Of the 12,172 predicted proteins, 1,024 (8.41%) were predicted to be secreted. A total of
376 2,702 Conserved Domain Database (CDD) domains were found in the secretome. Of these,
377 287 were specific features with NCBI curated models, 124 were generic features with only
378 the superfamily annotations [108]. Only 433 queries had no known domain hits. The
379 secretome was rich in alpha beta hydrolase superfamily (cl21494) containing enzymes,
380 glycosyl hydrolases and proteolytic enzymes and cytochrome P450 monoxygenases (*P450*)
381 (S8 Table). A total of 100 secreted proteins had nuclear-localization signals (S8 Table).

382 A total of 233 effector candidates were predicted by EffectorP. Following manual inspection
383 and censoring for candidates with known plant cell wall degrading catalytic domains, a total
384 of 168 candidates were selected as *C. tanaceti* effector candidates for further analysis (S9
385 Table). The secreted candidate effector repertoire of *C. tanaceti* contained homologs of
386 known effectors, such as the *Ecp6* of *Cladosporium fulvum* [109], *MC69* of *Magnoporthe*
387 *oryzae* [110], *ToxB* of *Pyrenophora tritici repentis* [111] and *Magnoporthe oryzae Bas3*
388 [112]. Furthermore, among the effector candidates, there were proteins with conserved
389 domains of known virulence factors. Most effector candidates were small (average length of
390 155 amino acids) and rich in cysteine (average cysteine composition was 3.3%) which are the

391 hallmarks of effectors. A total of 78 conserved motifs of fungal effectors [113] were present
392 in 62 effector candidates which had at least one motif each. Twenty-two effector candidates
393 that did not cluster in ortholog search among the 14 *Colletotrichum* and three related species,
394 and also did not show detectable homology to the NCBI-nr and swissprot databases were
395 defined as *C. tanaceti*-specific. Only 24% of the effectors of *C. tanaceti* were conserved
396 among all 14 *Colletotrichum* spp.

397 A total of 98 secreted peptidases were predicted with the majority (*n*=64) being serine
398 peptidases largely comprising the S08 and S09 subfamilies. The second most abundant class
399 was the metallo peptidases (*n*=19) (S10 Table). All six aspartic peptidases belonged to
400 subfamily A01. A total of 20 secreted peptidase inhibitors were reported in *C. tanaceti*
401 comprising two carboxypeptide-y inhibitors, five family-19 inhibitors and 13 family-14
402 inhibitors (S11 Table). Forty nine percent of the proteases of *C. tanaceti* were among the
403 "core" set of proteases of *Colletotrichum*.

404 **Secondary metabolite-related genes and clusters**

405 Forty-one secondary metabolite backbone genes were predicted in *C. tanaceti* using SMURF
406 and the majority were polyketide synthases (PKs, *n*=13) with four PKs-like proteins.
407 Furthermore, nine non-ribosomal peptide synthases (NRPS), eight NRPs–like proteins, two
408 hybrid PKs-NRPS enzymes and five dimethylallyltryptophans (DMATS) were also predicted
409 as backbone genes (S12 Table). A total of 52% of these backbone genes were within the core
410 set of genes in *Colletotrichum*. A total of 33 secondary metabolite gene-clusters were
411 predicted surrounding the backbone genes. However, the program antiSMASH predicted a
412 total of 50 clusters. Among the clusters, there were 12 type1-PKS, two type3- PKs, thirteen
413 terpenes, eleven NRPS, four indoles, three T1pks-nrps, one T1PKs-indole and four other
414 proteins. Cluster 10 of T1PKS showed 100% similarity to the genes in LL-Z1272 beta

biosynthetic gene cluster (BGC0001390_cl). Furthermore, a homolog to the melanin biosynthetic gene *SCD1* was also reported in *C. tanaceti* (CTA1_6632). When predictions from the two tools were compared, putative SMB clusters on 31 contigs of *C. tanaceti* were predicted by both tools and 19 of the backbone genes from SMURF were also predicted in antiSMASH (Supplementary Table 12). A total of 37 SM clusters were within the syntenic blocks of *C. higginsianum*. The conserved SM domains identified in each cluster were reported (S13 Table). Predictions from antiSMASH were compared across taxa and majority of the clusters were type1-PKs like followed by NRPS in all ascomycetes compared (Fig 3a). The highest number of clusters were reported from *C. fructicola* (*n*=84) followed by *C. higginsianum* (*n*=74) and *C. gloeosporioides* (*n*=73). The composition of the SMB gene cluster composition of *C. tanaceti* was most similar to *C. orchidophilum,* the acutatum complex members and *C. orbiculare* (S3 Fig).

Fig 3.Composition of different pathogenicity gene categories predicted for *Colletotrichum tanaceti* and related species. The number of genes in each gene category (x axis) plotted for each species (y axis). (a) secondary metabolite biosynthetic gene clusters- (gene clusters producing polyketides, terpenes, non-ribosomal peptides (NRPs), indoles and the hybrids of above); (b) number of homologs in the fungal cytochrome P450 database and the transporter classification database (TCDB); (c) homologs in the pathogen-host interaction database; homologs to entries in the "unaffected pathogenicity" database were excluded; (d) CAZyme classes; glycoside hydrolases (GH), polysaccharide lyases (PL), glycosyltransferases (GT), carbohydrate esterases (CE), molecules with auxiliary activities (AA), and carbohydrate binding molecules (CBM).

Cytochrome P450 monoxygenases (P450s) and transporters

In the *C. tanaceti* genome, 1,457genes had homologs in the fungal cytochrome P450 database (S14 Table) and 911 out of that had >30% identity. There were 1,824 homologs (S15 Table) in the transport classification database for *C. tanaceti* with 1,276 genes with >30% identity. The majority (*n*=430) of the homologs were genes of the major facilitator superfamily (MFS, 2.A.1) followed by 129 genes of the ABC transporter family (3.A.1) and 123 of N.P.C 1.I.1. Within *Colletotrichum* genus, members of the gloeosporioides complex had the highest number of homologs for both P450s and transporters (Fig 3b).

Homologs in PHI-base

A total of 3,497 homologs were recorded in *C. tanaceti* from the pathogen-host interaction database (PHI), of which 1,592 represented mutated phenotypes with reduced virulence (S16 Table). The second most common (*n*=1,514) were the unaffected pathogenicity category, 382 homologs were for loss of pathogenicity and 42 were in the effector category. Notably, the mutant phenotype of 141 homologs was lethal to this particular pathogen, and 103 homologs had increased virulence after mutation (Fig 3c). The two gloeosporioides complex members had the highest number of homologs in the database among the *Colletotrichum* spp., followed by the acutatum complex species, *C. simmondsii*, *C. fioriniae* and *C. nymphaea.* Despite *C. higginsianum* having a large number of homologs, *C. tanaceti* had a below average number for all the categories among the *Colletotrichum* spp., with a profile similar to *C. orchidophilum*, *C. chlorophyti* and *C. graminicola* (S4 Fig).

CAZymes

A total of 608 *C. tanaceti* proteins were assigned to 121 CAZyme families of which 43% was glycosyl hydrolases followed by 18 % of redox enzymes (auxiliary activities) and 14% carbohydrate esterases (S17 Table). Carbohydrate binding molecules and polysaccharide lyases both formed 7% each of the *C. tanaceti* CAZome whereas 11% was

462 glycosyltransferases. Members of the gloeosporioides and acutatum complexes had the
463 largest CAZomes among *Colletotrichum* spp. The CAZyme repertoires of the graminicola
464 complex members were relatively small (Fig 3d).

465 **Evolution of CAZyme families upon divergence of *Colletotrichum* lineages**

466 A total of 152 CAZyme families, predicted at the node of MRCA for *S. macrospora* and *B.*
467 *cinerea*, were used in gene family evolution analyses in CAFÉ. A uniform birth-death
468 parameter (λ) of 0.0023 was computed. Thirty gene families were reported to be significantly
469 evolving (family-wide p value ≥ 0.05), of which 21 were rapidly evolving (family-wide $p \geq$
470 0.01 and *Viterbi* $p \geq 0.01$ in any lineage) (S18 Table).

471 At the divergence of *Colletotrichum* spp., 39 expansions and 12 contractions were predicted
472 with respect to its MRCA with *Verticillium* species (S19 Table). Expansions included the
473 lignin hydrolase family AA2, pectin degrading polysaccharide lyase families (PL1, 3, 4, 9
474 and GH78), lignocellulose degrading families (AA3, AA9, GH131, GH5, GH6, GH7),
475 hemicelluloses degrading families (CE1, CE4, CE5, CE12, GH3, GH16, GH30, GH43,
476 GH51, GH67, and GH10), Lys M domain containing family CBM50 and cutinase family
477 CE5. The cellulose degrading family GH131 was the only rapidly evolving CAZyme family
478 (family-wide $p \geq 0.01$ and *Viterbi* $p \geq 0.01$) which expanded upon the divergence of
479 *Colletotrichum* spp. Within the genus, the highest number of expansions (n=38) was reported
480 at the divergence of the gloeosporioides-complex clade with only 4 contractions. Notably, the
481 CBM18 and GH10 families were contracted and many families with plant cell wall degrading
482 enzyme activity were expanded. The rapidly and significantly expanded families, (family-
483 wide $p \geq 0.01$ and *Viterbi* $p \geq 0.01$) upon the divergence of the gloeosporioides-complex clade
484 include GH43, GH106, CBM50 and AA7. At the divergence of the acutatum-complex clade,
485 there were 22 expansions, of which expansions in GH78, GH43 families were rapid and

significant and there was only one contraction. The divergence event of the graminicola-complex clade involved contractions in many CAZyme families with pectin degradation activity showing significant, rapid contractions (family-wide $p \geq 0.01$ and *Viterbi* $p \geq 0.01$)in families AA7, CBM50, CE8, GH28, GH78, PL1, and PL3. Divergence of the destructivum complex-clade was associated with 11 expansions and 21 contractions, of which expansion in AA7, GH74 and CE10 was significant and rapid.

Among the other species considered, *Fusarium oxysporum* had the highest number of genes (*n*=344) that were gained, with 75 expanded CAZyme with respect to its MRCA (S20 Table). *Colletotrichum incanum* had the second highest number of gene family expansions (*n*=35) and genes gained (*n*=69) followed by *C. higginsianum* (31 and 68 respectively). The highest number of contracted CAZyme families was identified in *Sordaria macrospora* (*n*=87) with a loss of 219 genes compared to the ancestral node. Forty CAZyme families contracted and only nine expanded in *C. tanaceti* with respect to the MRCA with *C. higginsianum*. The AA2 family with lignin peroxidase activity and the hemicellulose degrading GH12, GH74 families were among the expanded families, but many families with pathogenicity and plant cell wall degrading activity had contracted in *C. tanaceti*. However, the highest number of significant, rapidly evolving gene families was reported from *C. tanaceti* (*n*=9) followed by *F. oxysporum* and *C. higginsianum,* both which had seven rapidly evolving gene families each. In *C. tanaceti,* rapidly evolving CAZyme families included AA9, GH131 with lignocellulose degrading activity, chitin binding molecule families CBM18 and CBM50, GH18 with chitinase activity, GH3 and GH74 with hemicelluloses degrading activity, GH78 with pectinase activity and GT1 with glucuronosyltransferase activity. However, CBM18 and GH74 were the only families that expanded among those above with the rest contracting in *C. tanaceti* with respective to their MRCA. Gloeosporioides complex species had the largest 'CAZyme pathogenicity profiles' among all *Colletotrichum* species considered. The

CAZyme pathogenicity profile of *C. tanaceti* was most similar to those of *Colletotrichum* species known to have an intermediate host range, infecting many hosts within a single plant family or few hosts across several plant families (Fig 4). When compared the overall pathogenicity gene profiles of all *Colletotrichum* spp., which included the numbers of the SMB clusters, transporters, P450s, CAZymes and the homologs to the PHI database, the profile of *C. tanaceti* was most similar to *C. orchidophilum* and *C. chlorophyti* (Fig 5).

Fig 4. Comparison of CAZyme pathogenicity profiles predicted for *Colletotrichum* species. Hierarchical clustering performed with Euclidean distance and Ward linkage.

Fig 5. Comparison of the overall pathogenicity profiles predicted for *Colletotrichum* species. The numbers of CAZymes, secondary metabolite biosynthetic gene clusters (SMB), homologs in the transporter classification database (transporters), homologs in the fungal cytochrome P450 database (P450) and the number of homologs in the PHI database, excluding the homologs to entries in the "unaffected pathogenicity" database were used in the analysis. Hierarchical clustering was performed using Euclidean distance and Ward linkage methods.

Relationship of pathogenicity-related gene categories and repeat elements

The permutation tests confirmed that genes in all the tested pathogenicity-related gene categories are located significantly closer to tandem repeats than expected in a random sample (Table 4). The negative *Z*-scores confirmed the mean distance between those genes and the nearest repetitive element was less than mean of a random sample of the genome. Furthermore, all gene categories except the CAZymes were located significantly closer to the interspersed repeats. However, the expanded and the contacted subgroups of the total CAZome were significantly associated with interspersed repeats (Table 4).

536 **Table 4.** Permutation tests for association of repetitive elements with pathogenicity gene
537 categories

Gene categories of interest	All repeats [a]		Tandem repeats			Interspersed repeats	
	Z score [b]	*P* value [c]	Number of overlaps	*Z* score [b]	*P* value [c]	*Z* score [b]	*P* value [c]
CAZymes	-5.97	≤0.001	554	-3.914	≤0.001	-0.443	0.334
Expanded CAZymes	-3.514	≤0.001	96	-4.553	≤0.001	-3.050	≤0.001
Contracted CAZymes	-4.413	≤0.001	240	-3.237	≤0.001	-5.883	≤0.001
Effectors	-5.631	≤0.001	98	-4.725	≤0.001	-3.861	≤0.001
Peptidases	-5.787	≤0.001	82	-4.679	≤0.001	-3.895	≤0.001
SMB clusters	-7.901	≤0.001	171	-8.490	≤0.001	-2.610	0.003

538 [a] tandem and interspersed repeats
539 [b] Z-statistic estimate and its [c] associated probability computed based on 10,000 random
540 iterations.

541 **Accumulation of Pathogenicity genes in the A-T rich regions of *C. tanaceti* genome**

542 Distinct A-T rich regions and G-C equilibrated regions were identified in the genome of *C.*
543 *tanaceti* (Fig 6). A total of 24.3% of the genome which had an average length of 3.77 Kb was
544 rich in A-T and had a maximum G-C of 29%. A total of 85 genes were reported in these
545 regions which had a gene density of 6.04 genes per Mb but the majority (68.25%) of these
546 genes was hypothetical. Two secondary metabolite biosynthetic genes, 3 CAZymes, 2
547 cytochrome P450s, 2 lipases, 4 transporters, one transcription factor and one DNA
548 polymerase were also among the genes in the A-T rich regions (S21 Table). The G-C
549 equilibrated regions accounted for 75.7% of the genome and the average length was 14.6 Kb.
550 The maximum G-C percentage in these regions was 55.6 and 12,087 genes were reported
551 with a gene density of 276 genes per Mb.

552

553 **Fig 6. Plot of GC-content in the draft genome of *Colletotrichum tanaceti* against**
554 **proportion of the genome.** Genome segments were classified into A-T rich (24.3%) and G-
555 C equilibrated (75.7%) using a GC content threshold of 40% (vertical blue line).

DISCUSSION

Genome and the repeat content of *Colletotrichum tanaceti*

This study reports the first draft genome sequence and annotations of the emerging plant pathogen, *C. tanaceti*. The high N50 value and BUSCO completeness indicates the high quality of the assembly and AED scores of less than one for the majority of predicted genes (93.3%) suggested that these genes had at least partial congruence with the transcriptomic evidence [114]. These good quality gene predictions and annotations will provide a solid foundation for downstream genetic, population genomic and evolutionary studies.

The genome of *C. tanaceti* had a larger repeat content (25%) than the typical 3-10% in fungi [115]. Simple sequence repeats comprised 3.03% of the genome of *C. tanaceti* which itself was unusually high for fungi (generally 0.08-0.67%) [116]. However, the majority of repeats were interspersed transposable elements (TE) (21%). TE content of *C. tanaceti* was higher than in six previously studied *Colletotrichum* species, including *C. higginsianum* which is in the same species complex, but lower than in *C. orbiculare* (44.8%). The majority of TE were retro-transposons, similar to other *Colletotrichum* spp. [117]. Proliferation of repetitive elements especially transposons, is known to be a major mechanism driving expansion of eukaryote genomes [118, 119] and therefore, could be the reason the *C. tanaceti* genome is larger than average for fungi in the phylum Ascomycota (36.91 Mb) [120].

Repeat-induced-point mutation (RIP) is a fungal-specific mechanism for limiting transposon proliferation below destructive levels [39]. RIP is known to generate A-T rich regions with lower gene densities and higher evolutionary rates than the core genome, thus generating "two-speed" genomes in several fungi [117, 121-123]. The presence of A-T rich, gene sparse regions in the *C. tanaceti* genome could therefore, be a byproduct of the RIP due to TE proliferation. Accumulating repeats followed by expanding genome size with respect to the

non-pathogenic strains is a trend observed in many plant pathogenic fungi and can provide an evolutionary advantage in terms of pathogenesis [124]. The high repeat content of *C. tanaceti* may have an important role in generating genome plasticity [125].

Pathogenicity genes of *C. tanaceti*

A large array of putative genes related to pathogenicity was inferred from the sequenced genome of *C. tanaceti*. Apart from many plant cell wall-degrading enzymes, effectors, *P450*s and the proteolytic enzymes, there were proteins with CFEM domain (pfam05730) [126] with roles in conidial production and stress tolerance [127] among the secreted proteins. The average cysteine composition, length and proportion of specificity of the candidate secreted effectors of *C. tanaceti* were similar to those hemibiotrophic pathogens [128]. However, a minority of effector candidates was neither small (<300bp) nor rich in cysteine (>3%), similar to previous reports of atypical effectors [129]. Effector candidates with a nuclear localization signal might translocate to the host nucleus and reprogram the transcription of genes related to host immune responses. Homologs to known effectors, and effectors with conserved domains of virulence factors may have similar functions in *C. tanaceti*, for example, in penetration peg formation (cyclophllin) [130], phytotoxity induction (cerato-platanin) [131] and adherence of the fungal structures to other organisms (hydrophobin) [132].

Most secreted proteases of *C. tanaceti* were serine proteases predicted to evade plant immune responses by degrading plant chitinases [22]. Subtilisins (S08) were the most abundant of these in *C. tanaceti*, similar to reports in other fungi [22]. Subtilisins, with their alkaline optima, and the proteases in other subfamilies with acidic optima, such as A01, C13, G01, M20 and S10 [133], might enable *C. tanaceti* to degrade plant proteins across a wide pH

603 range. Also, the protease inhibitors of *C. tanaceti* might have effector-like roles via inhibition
604 of plant defense proteases [134].

605 The SMB gene clusters and the candidate proteins of MAPKs pathways identified in the
606 genome of *C. tanaceti* are also believed to play an important role in pathogenesis. The
607 majority of the secondary metabolite clusters of *C. tanaceti* were type 1 PKs-like which are
608 usually associated with synthesizing fungal toxins. Melanin, another important secondary
609 metabolite aids penetration via increasing turgor pressure [135]. Even though the gene cluster
610 associated with melanin biosynthesis was not identified, the homolog of the melanin
611 biosynthetic gene *SCD1* encoding Scytalone dehydratase [136] in *C. tanaceti* is worth
612 investigating further since *SCD1* has been successfully used as a target for fungicides to
613 control other pathogens [137]. Apart from their function in SM biosynthesis, the candidate
614 P450s of *C. tanaceti* could be involved in housekeeping roles and therefore, could be good
615 targets for fungicide development, as in the case of azoles targeting CYP51 [138].
616 Furthermore, the candidate proteins of MAPKs pathway in *C. tanaceti* could play a crucial
617 role in appressorium formation [25, 139], penetration [140], conidiation [141] and
618 pathogenesis-related morphogenesis [142], as reported for *C. higginsianum* and *C. lagenaria*.

619 Of the CAZyme families identified to be expanded in *C. tanaceti*, the chitin binding family
620 CBM18 could play a role in protecting the *C. tanaceti* cell wall from exogenous chitinases, as
621 is the case in *Trichoderma reesei* [143]. The expansion of the hemicellulose-degrading GH74
622 family could promote rapid degradation of host tissues by *C. tanaceti* during the necrotrophic
623 phase. The expansion of the lignin-degrading AA2 family in *C. tanaceti* has the potential to
624 assist infection of xylem vessels and thereby aid translocation of propagules to different parts
625 of the plant and establishing secondary infections.

The conserved nature of certain pathogenicity genes, such as the secondary metabolite clusters within the destructivum complex, was evident with their presence within the syntenic blocks with *C. higginsianum*. However, only a minority of the effectors, proteases and SM backbone genes of *C. tanaceti* were among the core gene set for *Colletotrichum* overall, therefore emphasizing their role in adaptation to new hosts. The species-specific effectors, singletons from the orthology analysis and the genes exclusive to *C. tanaceti* might have been horizontally transferred or be related to the host affiliation and niche specialization of *C. tanaceti*. Taken together, this inferred pathogenicity gene suite of *C. tanaceti* could be targeted in future resistance breeding and other disease management strategies for *C. tanaceti*.

Host range of *Colletotrichum tanaceti*

The proposed pathogenicity gene repertoire of *C. tanaceti* was most similar to that of pathogens with intermediate host ranges. The number of pathogenicity genes inferred from *C. tanaceti* was either similar to or less than the average for all *Colletotrichum spp.* investigated but the overall composition was similar to *Colletotrichum* spp. which either were able to infect many species within a plant family or few species across families. The comparison of CAZyme pathogenicity profiles among *Colletotrichum* spp., with both expansions and contractions with respective to its MRCA clearly suggested an intermediate host range for *C. tanaceti*.

The pathogenicity profile of *C. tanaceti* was very distinct from that of the other destructivum complex member, *C. higginsianum,* despite the two species sharing the highest number of orthologs and having the shortest evolutionary distance. Contractions in many pathogenicity gene families in *C. tanaceti* compared to *C. higggginsianum* indicated more restricted pathogenicity in *C. tanaceti*. The most similar CAZyme pathogenicity profile to that of *C.*

650 *tanaceti* was from *C. chlorophyti* which has been reported to infect herbaceous hosts such as
651 tomato (plant family Solanaceae) and soybean (plant family Fabaceae) [69]. The similarity to
652 *C. chlorophyti* was consistent for other gene categories such as the P450s, transporters and
653 the overall pathogenicity profile. A homolog to the demethylase (*PDA*), which provides
654 tolerance to the phytoalexin pisatin synthesised by *Pisum sativum* [144], was inferred in *C.*
655 *tanaceti* (CTA1_6324s) which could be an indicator of the ability of *C. tanaceti* to infect
656 Fabaceae. The composition of the SMB cluster was however, more similar to *C.*
657 *orchidophilum*, another pathogen reported to infect the herbaceous, monocot plant family of
658 Orchidaceae [145]. The similarity of the pathogenicity profile of *C. tanaceti* to two pathogens
659 infecting multiple herbaceous plant species was notable as the only known host of *C. tanaceti*
660 is also herbaceous. Both *C. chlorophyti* and *C. orchidophilum* have been reported from
661 multiple host species. Therefore, the pathogenicity gene suite of *C. tanaceti* suggests that *C.*
662 *tanaceti* has the genetic ability to infect more hosts than currently recognized. If *C. tanaceti*
663 can infect other hosts, such crops could also provide an external gene pool of inoculum for
664 infection of pyrethrum crops increasing the evolutionary potential of the pathogen
665 populations. Based on results of comparative analysis of pathogenicity profiles, a further
666 hypothesis is that these alternative hosts are likely to be herbaceous plants. Future studies
667 investigating the cross-host infectivity and pathogenicity of *C. tanaceti* are recommended.

668 **Evolution of pathogenicity genes**

669 Pathogenicity genes of *C. tanaceti* appear to be capable of evolving relatively rapidly.
670 Tandem repeats such as simple sequence repeats have high mutation rates [146] and could
671 promote frameshift mutations in adjacent genes by slipped misalignment during replication.
672 Therefore, the significant overlap between the tandem repeats and the pathogenicity genes
673 suggested high potential to mutate and create different pathotypes. Transposons promote
674 insertional mutations that can either cause disruption or modification of gene expression or

675 generate new proteins and also are major drivers of gene duplication [147]. Transposons were
676 in close proximity to gene categories of pathogenicity in *C. tanaceti* such as the SMB
677 clusters, peptidases and effectors. The significant association of TE with pathogenicity genes
678 were previously reported in *C. truncatum* [117] and *C. higginsianum* [15].

679 *Colletotrichum tanaceti* had the highest number of rapidly evolving CAZyme families among
680 the 17 species studied which also was indicative of the rapid evolutionary rate in these
681 pathogenicity genes. Interspersed repeats were not in close proximity to the total CAZome.
682 They were however, located significantly closer to the expanded or contacted families
683 indicating that interspersed repeats were a major contributor to CAZyme family
684 expansions/contractions in *C. tanaceti* by causing gene duplication (in expansions) or gene
685 disruptions (in contractions) [118, 148]. Although gene sparse, the A-T rich regions of the
686 genome contained several (n=18) known pathogenicity and virulence factors and many
687 hypothetical proteins which could be facilitating adaptive evolution. According to the two-
688 gene hypothesis the genes in the A-T rich regions can evolve faster than the ‘core‘ genome
689 [124]. Duplication of pathogenicity and virulence genes and a higher mutation rate may allow
690 more rapid pathogen responses to evolution of resistance in existing hosts or adaptation to
691 new host species.

692 **Genus *Colletotrichum***

693 Phylogenetic relationship throughout the genus was consistent with previous observations,
694 with gloeosporioides complex members and *C. orbiculare* forming a clade separately from
695 the destructivum, graminicola and acutatum clades [9-11]. One notable difference was in the
696 divergence time estimates for the divergence of *Colletotrichum* species complexes which
697 were more ancient than reported by Liang et al [11], despite using the same calibration times.
698 This could have been due to this study using cross-validation across 50 smoothing factors in

699 CAFÉ as opposed to using 12 different constraints and smoothing factor combinations
700 differences, as the use of the different data sets.

701 Comparative genomic analyses emphasized the rapid evolutionary rate and the high diversity
702 within the genus. The short time for speciation within the acutatum complex, and the fourteen
703 *Colletotrichum* species in general, was suggestive of the high evolutionary rate within the
704 genus with respective to the typical evolutionary rate of the fungal kingdom (0.0085 species
705 units per Myr) [149]. The sequence similarity between *C. tanaceti* and other species of
706 *Colletotrichum* varied widely and dropped drastically with evolutionary distance, suggesting
707 high diversity within the genus. However, the drop in orthology was less dramatic,
708 emphasizing the contribution of non-coding regions in generating diversity within the genus.
709 The extent of synteny between *C. tanaceti* and *C. higginsianum* was high and very similar to
710 the percentage synteny previously reported for the two graminicola complex species, *C.*
711 *sublineola* and *C. graminicola* [150]. This suggested that even though there was high
712 diversity within the genus, the species in the same species complex tend to share more
713 synteny than the between species complexes.

714 Evolutionary analysis of CAZyme families of different *Colletotrichum* lineages revealed an
715 association between CAZyme families and host range. The GH131 with cellulose degrading
716 activity was the only rapidly evolving gene family at the MRCA of *Colletotrichum* spp.
717 suggesting a possible association of this family with speciation and host determination within
718 the genus. Families GH43, with hemicellulose degrading activity and AA7, with gluco-
719 oligosaccharide activity significantly expanded upon divergence of both the gloeosporioides
720 and acutatum-complex clades, which could have broadened the host ranges of members of
721 these two complexes. The significant expansions in pectin degrading enzyme families GH106

722 in gloeosporioides and GH78 in the acutatum clades could also have enabled degradation of
723 pectin rich cell walls of young fruits [151] of these fruit-rotting species.

724 The most significant contractions were reported in pectin degrading families upon the
725 divergence of the graminicola complex clade. This could have been the reason for species in
726 this complex exclusively infecting monocot plant species considering that the pectin content
727 of monocot cell walls is generally less than in dicots [152]. Even though this was a similar
728 result to previous studies [7], *C. orchidophilum* which is known to infect plants from
729 monocot family Orchidaceae [153], deviated from this pattern. Gene family AA7 was rapidly
730 evolving in many *Colletotrichum* species and could have been involved in biotransformation
731 or detoxification of the lignocellulosic compounds [154].

732 In general, the overall CAZyme pathogenicity profiles of *Colletotrichum* spp. followed host
733 range of those species rather than the taxonomy. The gloeoporioides and acutaum complex
734 members which have broad host ranges, but are evolutionary distant, were clustered together.
735 This could be a byproduct of the "two-speed" genome scenario in certain *Colletotrichum* spp.
736 such as *C. orbiculare, C. chlorophyti, C. graminicola* [117] and as suggested, also in *C.*
737 *tanaceti*. In this scenario, the pathogenicity genes are located in repeat-rich regions, allowing
738 them to evolve at a higher rate than the rest of the genome. This was also evident by the
739 significant association of TE with pathogenicity genes in *C. tanaceti* and in *C. truncatum*
740 [117] and *C. higginsianum* [15]. This scenario would cause the species with similar
741 pathogenicity gene profiles to cluster together, despite their evolutionary distance.

742 **CONCLUSION**

743 In conclusion, a draft genome of *C. tanaceti* was used to quantify the molecular basis of
744 pathogenicity of the species and to improve the knowledge of the evolution of the fungal
745 genus *Colletotrichum. Colletotrichum tanaceti* is likely to have alternative hosts and is a

potential threat to the crops grown in rotation with pyrethrum. The genome of *Colletotrichum tanaceti* contains a large component of repetitive elements that may result in genome expansion and rapid generation of novel genotypes. The tendency of the pathogenicity genes to evolve rapidly was evident in genomic signals of the RIP and association of repeats with the pathogenicity genes. Therefore, with a large array of pathogenicity genes that potentially can evolve rapidly, *C. tanaceti* is likely to become a high-risk pathogen to global pyrethrum production. Complexity of the *Colletotrichum* genus was evident with its high diversity and evolutionary rate. The significant expansions and contractions of gene families upon divergence of different lineages within the genus could be important determinants in species complex diversification in *Colletotrichum*. The reason for pathogenicity genes to have different clustering than the phylogeny in *Colletotrichum* could be the occurrence of "two-speed" genomes in certain species. These findings will facilitate future research in genomics and disease management of *Colletotrichum*.

ACKNOWLEDGEMENTS

We thank Dr Kym Pham and Dr Arthur Hsu for performing library preparation, sequencing and genome assembly which were conducted at the Melbourne Translational Genomics Platform (Department of Pathology, University of Melbourne). We also thank Botanical Resources Australia - Agricultural Services, Pty. Ltd for supporting the project. RVL received Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship from the University of Melbourne, Australia. NDY was supported by a Career Development Fellowship (CDF) from NHMRC. PKK was supported by an Early Career Fellowship (CDF) from NHMRC.

768 **REFERENCES**

769

770 1. Velásquez AC, Castroverde CDM, He SY. Plant-Pathogen Warfare under Changing
771 Climate Conditions. Curr Biol. 2018;28(10):R619-R34. doi: 10.1016/j.cub.2018.03.054.
772 2. Doehlemann G, Ökmen B, Zhu W, Sharon A. Plant Pathogenic Fungi. Microbiology
773 Spectrum. 2017;5(1). doi: doi:10.1128/microbiolspec.FUNK-0023-2016.
774 3. Dean R, Van K A L, Pretorius Z A, Hammond-Kosack K E, Di P A, Spanu P D, et al.
775 The Top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol.
776 2012;13(4):414-30. doi: doi:10.1111/j.1364-3703.2011.00783.x.
777 4. Cannon PF, Damm U, Johnston PR, Weir BS. *Colletotrichum* – current status and
778 future directions. Studies in Mycology. 2012;73(1):181-213. doi: 10.3114/sim0014. PubMed
779 PMID: PMC3458418.
780 5. O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, et al.
781 Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and
782 transcriptome analyses. Nat Genet. 2012;44(9):1060-5. doi:
783 http://www.nature.com/ng/journal/v44/n9/abs/ng.2372.html#supplementary-information.
784 6. Gan P, Ikeda K, Irieda H, Narusaka M, O'Connell RJ, Narusaka Y, et al. Comparative
785 genomic and transcriptomic analyses reveal the hemibiotrophic stage shift of *Colletotrichum*
786 fungi. New Phytol. 2013;197(4):1236-49. doi: 10.1111/nph.12085.
787 7. Baroncelli R, Amby DB, Zapparata A, Sarrocco S, Vannacci G, Le Floch G, et al.
788 Gene family expansions and contractions are associated with host range in plant pathogens of
789 the genus *Colletotrichum*. BMC Genomics. 2016;17(1):555. doi: 10.1186/s12864-016-2917-
790 6.
791 8. Hacquard S, Kracher B, Hiruma K, Münch PC, Garrido-Oter R, Thon MR, et al.
792 Survival trade-offs in plant roots during colonization by closely related beneficial and
793 pathogenic fungi. Nature Communications. 2016;7:11362. doi: 10.1038/ncomms11362

794 https://www.nature.com/articles/ncomms11362#supplementary-information.
795 9. Gan P, Narusaka M, Kumakura N, Tsushima A, Takano Y, Narusaka Y, et al. Genus-
796 Wide Comparative Genome Analyses of *Colletotrichum* Species Reveal Specific Gene
797 Family Losses and Gains during Adaptation to Specific Infection Lifestyles. *Genome Biology*
798 *and Evolution*. 2016;8(5):1467-81. doi: 10.1093/gbe/evw089.
799 10. Rao S, Nandineni MR. Genome sequencing and comparative genomics reveal a
800 repertoire of putative pathogenicity genes in chilli anthracnose fungus *Colletotrichum*
801 *truncatum*. PLOS ONE. 2017;12(8):e0183567. doi: 10.1371/journal.pone.0183567.
802 11. Liang X, Wang B, Dong Q, Li L, Rollins JA, Zhang R, et al. Pathogenic adaptations
803 of *Colletotrichum* fungi revealed by genome wide gene family evolutionary analyses. PLOS
804 ONE. 2018;13(4):e0196303. doi: 10.1371/journal.pone.0196303.
805 12. Mongkolporn O, Taylor PWJ. Chili anthracnose: *Colletotrichum* taxonomy and
806 pathogenicity. Plant Pathol. 2018;67(6):1255-63. doi: doi:10.1111/ppa.12850.
807 13. Marin-Felix Y, Hernández-Restrepo M, Wingfield MJ, Akulov A, Carnegie AJ,
808 Cheewangkoon R, et al. Genera of phytopathogenic fungi: GOPHY 2. Studies in Mycology.
809 2019;92:47-133. doi: https://doi.org/10.1016/j.simyco.2018.04.002.
810 14. Damm U, Sato T, Alizadeh A, Groenewald JZ, Crous PW. The *Colletotrichum*
811 *dracaenophilum*, *C. magnum* and *C. orchidearum* species complexes. Studies in Mycology.
812 2019;92:1-46. doi: https://doi.org/10.1016/j.simyco.2018.04.001.
813 15. Dallery J-F, Lapalu N, Zampounis A, Pigné S, Luyten I, Amselem J, et al. Gapless
814 genome assembly of *Colletotrichum higginsianum* reveals chromosome structure and

association of transposable elements with secondary metabolite gene clusters. BMC Genomics. 2017;18(1):667. doi: 10.1186/s12864-017-4083-x.

16. Damm U, O'Connell RJ, Groenewald JZ, Crous PW. The *Colletotrichum* destructivum species complex - hemibiotrophic pathogens of forage and field crops. Stud Mycol. 2014;79:49-84. Epub 2014/12/11. doi: 10.1016/j.simyco.2014.09.003. PubMed PMID: 25492986; PubMed Central PMCID: PMCPmc4255528.

17. Barimani M, Pethybridge SJ, Vaghefi N, Hay FS, Taylor PWJ. A new anthracnose disease of pyrethrum caused by *Colletotrichum tanaceti* sp. nov. Plant Pathol. 2013;62(6):1248-57. doi: 10.1111/ppa.12054. PubMed PMID: 91824730.

18. Duke SO, Cantrell CL, Meepagala KM, Wedge DE, Tabanca N, Schrader KK. Natural toxins for use in pest management. Toxins. 2010;2(8):1943-62. doi: 10.3390/toxins2081943. PubMed PMID: 22069667.

19. Hay FS, Gent DH, Pilkington SJ, Pearce TL, Scott JB, Pethybridge SJ. Changes in distribution and frequency of fungi associated with a foliar disease complex of pyrethrum in Australia. Plant Dis. 2015;99(9):1227-35. doi: 10.1094/PDIS-12-14-1357-RE.

20. De Silva DD, Crous PW, Ades PK, Hyde KD, Taylor PWJ. Life styles of *Colletotrichum* species and implications for plant biosecurity. Fungal Biology Reviews. 2017;31(3):155-68. doi: https://doi.org/10.1016/j.fbr.2017.05.001.

21. Sperschneider J, Gardiner D, M., Dodds PN, Tini F, Covarelli L, Singh KB, et al. EffectorP: predicting fungal effector proteins from secretomes using machine learning. New Phytol. 2015;210(2):743-61. doi: 10.1111/nph.13794.

22. Muszewska A, Stepniewska-Dziubinska MM, Steczkiewicz K, Pawlowska J, Dziedzic A, Ginalski K. Fungal lifestyle reflected in serine protease repertoire. *Scientific Reports*. 2017;7(1):9147. doi: 10.1038/s41598-017-09644-w.

23. Zhao Z, Liu H, Wang C, Xu J-R. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics. 2013;14(1):274. doi: 10.1186/1471-2164-14-274.

24. Howlett BJ. Secondary metabolite toxins and nutrition of plant pathogenic fungi. Curr Opin Plant Biol. 2006;9(4):371-5. doi: https://doi.org/10.1016/j.pbi.2006.05.004.

25. Zhao X, Mehrabi R, Xu J-R. Mitogen-Activated Protein Kinase Pathways and Fungal Pathogenesis. Eukaryot Cell. 2007;6(10):1701-14. doi: 10.1128/ec.00216-07.

26. Sharma KK. Fungal genome sequencing: basic biology to biotechnology. Crit Rev Biotechnol. 2016;36(4):743-59. doi: 10.3109/07388551.2015.1015959.

27. Yoder OC, Turgeon BG. Fungal genomics and pathogenicity. Curr Opin Plant Biol. 2001;4(4):315-21. doi: https://doi.org/10.1016/S1369-5266(00)00179-5.

28. Urban M, Cuzick A, Rutherford K, Irvine A, Pedro H, Pant R, et al. PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database. Nucleic Acids Res. 2017;45(D1):D604-D10. doi: 10.1093/nar/gkw1089.

29. Lu T, Yao B, Zhang C. DFVF: database of fungal virulence factors. Database: The Journal of Biological Databases and Curation. 2012;2012:bas032. doi: 10.1093/database/bas032. PubMed PMID: PMC3478563.

30. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 2015;43(W1):W237-W43. doi: 10.1093/nar/gkv437.

31. Crouch JA, O'Connell R, Gan P, Buiate E, Torres M, Beirn L, et al. The Genomics of *Colletotrichum*2014.

32. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin. 1987;19:11-5. doi: citeulike-article-id:678648.

33. Kapabiosystems. Kapa Biosystems | Enzyme Solutions | Next Generation PCR 2015 [cited 2015 08/11]. Available from: https://www.kapabiosystems.com/.

34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. Epub 2014/04/04. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PubMed Central PMCID: PMCPmc4103590.
35. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017;33(4):574-6. doi: 10.1093/bioinformatics/btw663.
36. Broad Institute. software.broadinstitute.org 2018 [cited 2016 12/12]. Available from: https://software.broadinstitute.org/software/discovar/blog/?page_id=98.
37. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2. doi: 10.1093/bioinformatics/btv351.
38. Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, et al. Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud. *PLOS ONE*. 2015;10(10):e0140829. doi: 10.1371/journal.pone.0140829.
39. Testa AC, Oliver RP, Hane JK. OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. Genome biology and evolution. 2016;8(6):2044-64. doi: 10.1093/gbe/evw121. PubMed PMID: 27289099.
40. Smit AFA, Robert H, Kas A, Siegel A, Gish W, Price A, et al. RepeatModeler. 1.0.5 ed. http://www.repeatmasker.org: Institute of Systems Biology; 2011.
41. Bao Z, Eddy SR. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12(8):1269-76. Epub 2002/08/15. doi: 10.1101/gr.88502. PubMed PMID: 12176934; PubMed Central PMCID: PMC186642.
42. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005;21 Suppl 1:i351-8. Epub 2005/06/18. doi: 10.1093/bioinformatics/bti1018. PubMed PMID: 15961478.
43. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(suppl 2):W265-W8.
44. Smit AFA, Hubley R, Green P. RepeatMasker. http://www.repeatmasker.org: Institute of Systems Biology; 1996-2010.
45. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573-80. Epub 1998/12/24. PubMed PMID: 9862982; PubMed Central PMCID: PMC148217.
46. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1-4):462-7. Epub 2005/08/12. doi: 10.1159/000084979. PubMed PMID: 16093699.
47. Auyong ASM, Ford R, Taylor PWJ. Genetic transformation of *Colletotrichum truncatum* associated with anthracnose disease of chili by random insertional mutagenesis. J Basic Microbiol. 2012;52(4):372-82. doi: doi:10.1002/jobm.201100250.
48. Chen X, Shen G, Wang Y, Zheng X, Wang Y. Identification of *Phytophthora sojae* genes upregulated during the early stage of soybean infection. FEMS Microbiol Lett. 2007;269(2):280-8. doi: 10.1111/j.1574-6968.2007.00639.x.
49. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data 2017 [cited 2017]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
50. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics. 2011;12:491. Epub 2011/12/24. doi: 10.1186/1471-2105-12-491. PubMed PMID: 22192575; PubMed Central PMCID: PMC3280279.

913 51. O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, et al.
914 Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and
915 transcriptome analyses. Nat Genet. 2012;44(9):1060-5.
916 52. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De*
917 *novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for
918 reference generation and analysis. Nat Protoc. 2013;8(8):1494-512. Epub 2013/07/13. doi:
919 10.1038/nprot.2013.084. PubMed PMID: 23845962; PubMed Central PMCID:
920 PMC3875132.
921 53. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate
922 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
923 Genome biology. 2013;14(4):R36.
924 54. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-
925 generation sequencing data. Bioinformatics. 2012;28(23):3150-2. doi:
926 10.1093/bioinformatics/bts565. PubMed PMID: PMC3516142.
927 55. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of
928 protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9. doi:
929 10.1093/bioinformatics/btl158.
930 56. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab
931 initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Web Server
932 issue):W435-9. Epub 2006/07/18. doi: 10.1093/nar/gkl200. PubMed PMID: 16845043;
933 PubMed Central PMCID: PMCPmc1538822.
934 57. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004;5:59. Epub
935 2004/05/18. doi: 10.1186/1471-2105-5-59. PubMed PMID: 15144565; PubMed Central
936 PMCID: PMC421630.
937 58. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding.
938 Nucleic Acids Res. 1998;26(4):1107-15. Epub 1998/03/21. PubMed PMID: 9461475;
939 PubMed Central PMCID: PMC147337.
940 59. R Core Team. R: A language and environment for statistical computing.: R
941 Foundation for Statistical Computing, Vienna, Austria.; 2015.
942 60. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for
943 eukaryotic genomes. Genome Res. 2003;13(9):2178-89. PubMed PMID: 12952885.
944 61. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
945 BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. Epub
946 2009/12/17. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PubMed Central
947 PMCID: PMCPmc2803857.
948 62. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5:
949 genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-40. doi:
950 10.1093/bioinformatics/btu031. PubMed PMID: PMC3998142.
951 63. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools
952 for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol.
953 2016;428(4):726-31. doi: https://doi.org/10.1016/j.jmb.2015.11.006.
954 64. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new
955 perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res.
956 2017;45(Database issue):D353-D61. doi: 10.1093/nar/gkw1092. PubMed PMID:
957 PMC5210567.
958 65. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large
959 gene lists using DAVID bioinformatics resources. *Nature Protocols*. 2008;4:44. doi:
960 10.1038/nprot.2008.211

961 https://www.nature.com/articles/nprot.2008.211#supplementary-information.

962 66. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths
963 toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res.
964 2009;37(1):1-13. doi: 10.1093/nar/gkn923. PubMed PMID: PMC2615629.
965 67. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al.
966 Versatile and open software for comparing large genomes. *Genome Biology*. 2004;5(2):R12-
967 R. doi: 10.1186/gb-2004-5-2-r12. PubMed PMID: PMC395750.
968 68. Soderlund C, Nelson W, Shoemaker A, Paterson A. SyMAP: A system for
969 discovering and viewing syntenic regions of FPC maps. Genome Res. 2006;16(9):1159-68.
970 doi: 10.1101/gr.5396706. PubMed PMID: PMC1557773.
971 69. Gan P, Narusaka M, Tsushima A, Narusaka Y, Takano Y, Shirasu K. Draft Genome
972 Assembly of *Colletotrichum chlorophyti*, a Pathogen of Herbaceous Plants. Genome
973 Announc. 2017;5(10). Epub 2017/03/11. doi: 10.1128/genomeA.01733-16. PubMed PMID:
974 28280027; PubMed Central PMCID: PMCPMC5347247.
975 70. Baroncelli R, Sreenivasaprasad S, Sukno SA, Thon MR, Holub E. Draft Genome
976 Sequence of *Colletotrichum acutatum* Sensu Lato (*Colletotrichum fioriniae*). Genome
977 announcements. 2014;2(2):e00112-14. doi: 10.1128/genomeA.00112-14. PubMed PMID:
978 24723700.
979 71. Alkan N, Meng X, Friedlander G, Reuveni E, Sukno S, Sherman A, et al. Global
980 Aspects of pacC Regulation of Pathogenicity Genes in *Colletotrichum gloeosporioides* as
981 Revealed by Transcriptome Analysis. Mol Plant-Microbe Interact. 2013;26(11):1345-58. doi:
982 10.1094/MPMI-03-13-0080-R.
983 72. Baroncelli R, Sukno SA, Sarrocco S, Cafa G, Le Floch G, Thon MR. Whole-Genome
984 Sequence of the Orchid Anthracnose Pathogen *Colletotrichum orchidophilum*. Mol Plant
985 Microbe Interact. 2018;31(10):979-81. Epub 2018/04/14. doi: 10.1094/mpmi-03-18-0055-a.
986 PubMed PMID: 29649963.
987 73. Baroncelli R, Sanz-Martín JM, Rech GE, Sukno SA, Thon MR. Draft Genome
988 Sequence of *Colletotrichum sublineola*, a Destructive Pathogen of Cultivated Sorghum.
989 Genome Announcements. 2014;2(3):e00540-14. doi: 10.1128/genomeA.00540-14. PubMed
990 PMID: PMC4056296.
991 74. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BPHJ, et al.,
992 editors. Comparative genomics of the plant vascular wilt pathogens, *Verticillium dahliae* and
993 *Verticillium albo-atrum*2010.
994 75. Staats M, van Kan JAL. Genome update of *Botrytis cinerea* strains B05.10 and T4.
995 Eukaryot Cell. 2012;11(11):1413-4. doi: 10.1128/EC.00164-12. PubMed PMID: 23104368.
996 76. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, et al. De novo
997 assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a
998 model organism for fungal morphogenesis. PLoS Genet. 2010;6(4):e1000891. Epub
999 2010/04/14. doi: 10.1371/journal.pgen.1000891. PubMed PMID: 20386741; PubMed Central
1000 PMCID: PMCPMC2851567.
1001 77. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, et
1002 al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature.
1003 2010;464:367. doi: 10.1038/nature08850

1004 https://www.nature.com/articles/nature08850#supplementary-information.
1005 78. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale
1006 detection of protein families. Nucleic Acids Res. 2002;30(7):1575-84. PubMed PMID:
1007 PMC101833.
1008 79. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
1009 Improvements in Performance and Usability. Mol Biol Evol. 2013;30(4):772-80. doi:
1010 10.1093/molbev/mst010. PubMed PMID: PMC3603318.

80. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3. doi: 10.1093/bioinformatics/btp348. PubMed PMID: PMC2712344.
81. Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology*. 2014;11(1):81. doi: 10.1186/s12983-014-0081-x.
82. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27(8):1164-5. doi: 10.1093/bioinformatics/btr088.
83. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3. doi: 10.1093/bioinformatics/btu033. PubMed PMID: PMC3998144.
84. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289.
85. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2018.
86. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 2003;19(2):301-2. doi: 10.1093/bioinformatics/19.2.301.
87. Sanderson MJ. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. Mol Biol Evol. 2002;19(1):101-9. doi: 10.1093/oxfordjournals.molbev.a003974.
88. Nash SG. A survey of truncated-Newton methods. *Journal of Computational and Applied Mathematics*. 2000;124(1):45-59. doi: https://doi.org/10.1016/S0377-0427(00)00426-X.
89. Beimforde C, Feldberg K, Nylinder S, Rikkinen J, Tuovila H, Dörfelt H, et al. Estimating the Phanerozoic history of the Ascomycota lineages: Combining fossil and molecular data. Mol Phylogen Evol. 2014;78:386-98. doi: https://doi.org/10.1016/j.ympev.2014.04.024.
90. Meinken J, Asch DK, Neizer-Ashun KA, Chang GH, Cooper CRJ, Min XJ. FunSecKB2: A fungal protein subcellular location knowledgebase2014. 1-17 p.
91. Nielsen H. Predicting Secretory Proteins with SignalP. In: Kihara D, editor. Protein Function Prediction: Methods and Protocols. New York, NY: Springer New York; 2017. p. 59-73.
92. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 2007;35(Web Server issue):W429-W32. doi: 10.1093/nar/gkm256. PubMed PMID: PMC1933244.
93. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35(Web Server issue):W585-W7. doi: 10.1093/nar/gkm259. PubMed PMID: PMC1933216.
94. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J Mol Biol. 2001;305(3):567-80. doi: https://doi.org/10.1006/jmbi.2000.4315.
95. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006;34(Web Server issue):W362-W5. doi: 10.1093/nar/gkl124. PubMed PMID: PMC1538847.
96. Pierleoni A, Martelli PL, Casadio R. PredGPI: a GPI-anchor predictor. BMC Bioinformatics. 2008;9(1):392. doi: 10.1186/1471-2105-9-392.

97. Nguyen Ba AN, Pogoutse A, Provart N, Moses AM. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics. 2009;10:202-. doi: 10.1186/1471-2105-10-202. PubMed PMID: PMC2711084.

98. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 2018;46(D1):D624-D32. doi: 10.1093/nar/gkx1134.

99. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010;47(9):736-41. doi: https://doi.org/10.1016/j.fgb.2010.06.003.

100. Park J, Lee S, Choi J, Ahn K, Park B, Park J, et al. Fungal cytochrome P450 database. BMC Genomics. 2008;9:402-. doi: 10.1186/1471-2164-9-402. PubMed PMID: PMC2542383.

101. Saier MH, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res. 2016;44(Database issue):D372-D9. doi: 10.1093/nar/gkv1103. PubMed PMID: PMC4702804.

102. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012;40(Web Server issue):W445-W51. doi: 10.1093/nar/gks479. PubMed PMID: PMC3394287.

103. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2010;11(1):431. doi: 10.1186/1471-2105-11-431.

104. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22(10):1269-71. doi: 10.1093/bioinformatics/btl097.

105. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. Mol Biol Evol. 2013;30(8):1987-97. doi: 10.1093/molbev/mst100.

106. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 2015;43(Web Server issue):W566-W70. doi: 10.1093/nar/gkv468. PubMed PMID: PMC4489295.

107. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2016;32(2):289-91. doi: 10.1093/bioinformatics/btv562.

108. Derbyshire MK, Lanczycki CJ, Bryant SH, Marchler-Bauer A. Annotation of functional sites with the Conserved Domain Database. Database. 2012;2012:bar058-bar. doi: 10.1093/database/bar058.

109. de Jonge R, Peter van Esse H, Kombrink A, Shinya T, Desaki Y, Bours R, et al. Conserved Fungal LysM Effector Ecp6 Prevents Chitin-Triggered Immunity in Plants. *Science*. 2010;329(5994):953-5. doi: 10.1126/science.1190859.

110. Saitoh H, Fujisawa S, Mitsuoka C, Ito A, Hirabuchi A, Ikeda K, et al. Large-Scale Gene Disruption in Magnaporthe oryzae Identifies MC69, a Secreted Protein Required for Infection by Monocot and Dicot Fungal Pathogens. PLoS Path. 2012;8(5):e1002711. doi: 10.1371/journal.ppat.1002711.

111. Aboukhaddour R, Kim YM, Strelkov SE. RNA-mediated gene silencing of ToxB in Pyrenophora tritici-repentis. Mol Plant Pathol. 2012;13(3):318-26. doi: doi:10.1111/j.1364-3703.2011.00748.x.

112. Mosquera G, Giraldo MC, Khang CH, Coughlan S, Valent B. Interaction Transcriptome Analysis Identifies Magnaporthe oryzae BAS1-4 as Biotrophy-Associated Secreted Proteins in Rice Blast Disease. *The Plant Cell*. 2009;21(4):1273-90. doi: 10.1105/tpc.107.055228. PubMed PMID: PMC2685627.
113. Sonah H, Deshmukh RK, Bélanger RR. Computational Prediction of Effector Proteins in Fungi: Opportunities and Challenges. *Frontiers in Plant Science*. 2016;7:126. doi: 10.3389/fpls.2016.00126. PubMed PMID: PMC4751359.
114. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12(1):491. doi: 10.1186/1471-2105-12-491.
115. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B. Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res. 2005;15(12):1620-31. Epub 2005/12/13. doi: 10.1101/gr.3767105. PubMed PMID: 16339359.
116. Karaoglu H, Lee CMY, Meyer W. Survey of Simple Sequence Repeats in Completed Fungal Genomes. Mol Biol Evol. 2005;22(3):639-49. doi: 10.1093/molbev/msi057.
117. Rao S, Sharda S, Oddi V, Nandineni MR. The Landscape of Repetitive Elements in the Refined Genome of Chilli Anthracnose Fungus *Colletotrichum truncatum*. Frontiers in Microbiology. 2018;9(2367). doi: 10.3389/fmicb.2018.02367.
118. Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, et al. Transposable elements versus the fungal Genome: impact on whole-genome architecture and transcriptional profiles. PLoS Genet. 2016;12(6):e1006108. doi: 10.1371/journal.pgen.1006108.
119. Pellicer J, Hidalgo O, Dodsworth S, Leitch JI. Genome Size Diversity and Its Impact on the Evolution of Land Plants. Genes. 2018;9(2). doi: 10.3390/genes9020088.
120. Mohanta TK, Bae H. The diversity of fungal genome. *Biological Procedures Online*. 2015;17:8. doi: 10.1186/s12575-015-0020-z. PubMed PMID: PMC4392786.
121. Cambareri EB, Jensen BC, Schabtach E, Selker EU. Repeat-induced GC to AT mutations in Neurospora. Science. 1989;244(4912):1571-5.
122. Santana MF, Silva JC, Mizubuti ES, Araújo EF, Condon BJ, Turgeon BG, et al. Characterization and potential evolutionary impact of transposable elements in the genome of Cochliobolus heterostrophus. BMC Genomics. 2014;15(1):536.
123. Li W-C, Huang C-H, Chen C-L, Chuang Y-C, Tung S-Y, Wang T-F. Trichoderma reesei complete genome sequence, repeat-induced point mutation, and partitioning of CAZyme gene clusters. Biotechnology for biofuels. 2017;10(1):170.
124. Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 2015;35:57-65. doi: https://doi.org/10.1016/j.gde.2015.09.001.
125. Wostemeyer J, Kreibich A. Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. Curr Genet. 2002;41(4):189-98. Epub 2002/08/13. doi: 10.1007/s00294-002-0306-y. PubMed PMID: 12172959.
126. Kulkarni RD, Kelkar HS, Dean RA. An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. Trends Biochem Sci. 2003;28(3):118-21. doi: 10.1016/S0968-0004(03)00025-2.
127. Zhu W, Wei W, Wu Y, Zhou Y, Peng F, Zhang S, et al. BcCFEM1, a CFEM Domain-Containing Protein with Putative GPI-Anchored Site, Is Involved in Pathogenicity, Conidial Production, and Stress Tolerance in Botrytis cinerea. *Frontiers in Microbiology*. 2017;8(1807). doi: 10.3389/fmicb.2017.01807.
128. Kim K-T, Jeon J, Choi J, Cheong K, Song H, Choi G, et al. Kingdom-Wide Analysis of Fungal Small Secreted Proteins (SSPs) Reveals their Potential Role in Host Association.

Frontiers in Plant Science. 2016;7:186. doi: 10.3389/fpls.2016.00186. PubMed PMID: PMC4759460.

129. Selin C, de Kievit TR, Belmonte MF, Fernando WGD. Elucidating the Role of Effectors in Plant-Fungal Interactions: Progress and Challenges. Frontiers in Microbiology. 2016;7:600. doi: 10.3389/fmicb.2016.00600. PubMed PMID: PMC4846801.

130. Viaud MC, Balhadère PV, Talbot NJ. A *Magnaporthe grisea* Cyclophilin Acts as a Virulence Determinant during Plant Infection. *The Plant Cell*. 2002;14(4):917-30. doi: 10.1105/tpc.010389. PubMed PMID: PMC150692.

131. Baccelli I. Cerato-platanin family proteins: one function for multiple biological roles? *Frontiers in Plant Science*. 2014;5:769. doi: 10.3389/fpls.2014.00769. PubMed PMID: PMC4284994.

132. Bayry J, Aimanianda V, Guijarro JI, Sunde M, Latgé J-P. Hydrophobins—Unique Fungal Proteins. PLoS Path. 2012;8(5):e1002700. doi: 10.1371/journal.ppat.1002700. PubMed PMID: PMC3364958.

133. Rao MB, Tanksale AM, Ghatge MS, Deshpande VV. Molecular and biotechnological aspects of microbial proteases. Microbiology and molecular biology reviews : MMBR. 1998;62(3):597-635. PubMed PMID: 9729602.

134. Jashni MK, Mehrabi R, Collemare J, Mesarich CH, de Wit PJGM. The battle in the apoplast: further insights into the roles of proteases and their inhibitors in plant–pathogen interactions. *Frontiers in Plant Science*. 2015;6:584. doi: 10.3389/fpls.2015.00584. PubMed PMID: PMC4522555.

135. Kubo Y, Furusawa I. Melanin Biosynthesis. In: Cole GT, Hoch HC, editors. The fungal spore and disease initiation in plants and animals. Boston, MA: Springer US; 1991. p. 205-18.

136. Kubo Y, Takano Y, Endo N, Yasuda N, Tajima S, Furusawa I. Cloning and structural analysis of the melanin biosynthesis gene *SCD1* encoding scytalone dehydratase in *Colletotrichum lagenarium*1997. 4340-4 p.

137. Yamada N, Motoyama T, Nakasako M, Kagabu S, Kudo T, Yamaguchi I. Enzymatic Characterization of Scytalone Dehydratase Val75Met Variant Found in Melanin Biosynthesis Dehydratase Inhibitor (MBI-D) Resistant Strains of the Rice Blast Fungus. Biosci, Biotechnol, Biochem. 2004;68(3):615-21. doi: 10.1271/bbb.68.615.

138. Becher R, Wirsel SGR. Fungal cytochrome P450 sterol 14α-demethylase (CYP51) and azole resistance in plant and human pathogens. Appl Microbiol Biotechnol. 2012;95(4):825-40. doi: 10.1007/s00253-012-4195-9.

139. Takano Y, Kikuchi T, Kubo Y, Hamer JE, Mise K, Furusawa I. The *Colletotrichum lagenarium* MAP Kinase Gene *CMK1* regulates diverse aspects of fungal pathogenesis. Mol Plant-Microbe Interact. 2000;13(4):374-83. doi: 10.1094/MPMI.2000.13.4.374.

140. Tsuji G, Fujii S, Tsuge S, Shiraishi T, Kubo Y. The *Colletotrichum lagenarium* Ste12-Like Gene *CST1* is essential for appressorium penetration. Mol Plant-Microbe Interact. 2003;16(4):315-25. doi: 10.1094/MPMI.2003.16.4.315.

141. Kojima K, Kikuchi T, Takano Y, Oshiro E, Okuno T. The mitogen-activated protein kinase gene *MAF1* is essential for the early differentiation phase of appressorium formation in *Colletotrichum lagenarium*. Mol Plant-Microbe Interact. 2002;15(12):1268-76.

142. Sakaguchi A, Tsuji G, Kubo Y. A Yeast STE11 Homologue *CoMEKK1* is essential for pathogenesis-related morphogenesis in *Colletotrichum orbiculare*. Mol Plant-Microbe Interact. 2010;23(12):1563-72. doi: 10.1094/MPMI-03-10-0051.

143. Liu P, Stajich JE. Characterization of the carbohydrate Binding Module 18 gene family in the amphibian pathogen *Batrachochytrium dendrobatidis*. Fungal Genet Biol. 2015;77:31-9. doi: https://doi.org/10.1016/j.fgb.2015.03.003.

144. Delserone LM, McCluskey K, Matthews DE, Vanetten HD. Pisatin demethylation by fungal pathogens and nonpathogens of pea: association with pisatin tolerance and virulence. Physiol Mol Plant Pathol. 1999;55(6):317-26. doi: https://doi.org/10.1006/pmpp.1999.0237.
145. Baroncelli R, Sukno S, Sarrocco S, Cafà G, Le Floch G, Thon MR. Whole-genome sequence of the orchid anthracnose pathogen Colletotrichum orchidophilum. Mol Plant-Microbe Interact. 2018. doi: 10.1094/MPMI-03-18-0055-A.
146. Xu X, Peng M, Fang Z, Xu X. The direction of microsatellite mutations is dependent upon allele length. Nat Genet. 2000;24:396. doi: 10.1038/74238.
147. Chia N, Goldenfeld N. Dynamics of gene duplication and transposons in microbial genomes following a sudden environmental change. Physical Review E. 2011;83(2):021906. doi: 10.1103/PhysRevE.83.021906.
148. Seidl MF, Thomma BPHJ. Transposable elements direct the coevolution between plants and microbes. Trends Genet. 2017;33(11):842-51. doi: 10.1016/j.tig.2017.07.003.
149. Scholl JP, Wiens JJ. Diversification rates and species richness across the Tree of Life. *Proceedings of the Royal Society B: Biological Sciences*. 2016;283(1838). doi: 10.1098/rspb.2016.1334.
150. Buiate EAS, Xavier KV, Moore N, Torres MF, Farman ML, Schardl CL, et al. A comparative genomic analysis of putative pathogenicity genes in the host-specific sibling species *Colletotrichum graminicola* and *Colletotrichum sublineola*. BMC Genomics. 2017;18:1-24. doi: 10.1186/s12864-016-3457-9. PubMed PMID: 120657703.
151. Brummell DA. Cell wall disassembly in ripening fruit. Funct Plant Biol. 2006;33(2):103-19. doi: https://doi.org/10.1071/FP05234.
152. Pattathil S, Hahn MG, Dale BE, Chundawat SPS. Insights into plant cell wall structure, architecture, and integrity using glycome profiling of native and AFEXTM-pre-treated biomass. J Exp Bot. 2015;66(14):4279-94. doi: 10.1093/jxb/erv107.
153. Damm U, Cannon PF, Woudenberg JHC, Crous PW. The *Colletotrichum* acutatum species complex. *Studies in Mycology*. 2012;73(1):37-113. doi: 10.3114/sim0010. PubMed PMID: PMC3458416.
154. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels*. 2013;6(1):41. doi: 10.1186/1754-6834-6-41.

Supplementary information

S1 Table. GO term enrichment analysis in *C. tanaceti*

S2 Table. KEGG orthology annotations of *C. tanaceti*

S3 Table. KEGG pathway map IDs of *C. tanaceti*

S4 Table. KEGG orthology assignments of Map kinase pathway in *C. tanaceti*

S5 Table. KEGG orthology assignments of Aflatoxin biosynthesis pathway in *C. tanaceti*

S6 Table. KEGG orthology assignments of ABC transporters in *C. tanaceti*

S7 Table. Global alignment of *C. tanaceti* contigs to the *C. higginsianum* chromosomes

S8 Table. Secreted proteins of *C. tanaceti* and their conserved domains

S9 Table. Secreted effector candidates of *C. tanaceti* with homology to known effectors and conserved motifs

S10 Table. Secreted peptidases of *C. tanaceti*

S11 Table. Secreted peptidase inhibitors of *C. tanaceti*

S12 Table. Secondary metabolite bioysnthetic gene cluster predtions of *C. tanaceti*

S13 Table. Conserved domains of secondary metabolite biosynthetic genes of *C. tanaceti*

S14 Table. Homologs in *C. tanaceti* to fungal cytochrome P450 database

S15 Table. Homologs in *C. tanaceti* to transporter classification database

S16 Table. Homologs to pathogen host interaction database in *C. tanaceti*

1258 **S17 Tabel**. CAZyme family assignment of *C. tanaceti*

1259 **S18 Table**. Family-wide probability values and viterbi probability values of CAZyme
1260 families across taxa

1261 **S19 Table**. Expansions and contractions of CAZyme familes upon divergence of different
1262 lineages

1263 **S20 Table**. Statistics of CAZyme gene family evolution across taxa

1264 **S21 Table**. Genes in the Atrich region of *C. tanaceti* genome

1265 **S1 Fig. Median GC percentage and median total length (Mb) (x axis) of publicly**
1266 **available draft genomes representing *Colletotrichum* species (y axis).**
1267 **S2 Fig. Circular plot showing synteny between *Colletotrichum tanaceti* contigs**
1268 **(numbers) mapped to the 12 individual chromosomes (NC_codes) of the *C.***
1269 ***higginsianum* genome.**

1270 **S3 Fig. Comparison of type of secondary metabolite biosynthetic gene clusters** (gene
1271 clusters producing Terpenes, Indoles, Polyketides (PKs), Non-ribosomal peptides (NRPs_and
1272 hybrids of the above categories) in *Colletotrichum* species; hierarchical clustering performed
1273 using Euclidean distance and Ward linkage.

1274 **S4 Fig. Comparison of composition of pathogen host interaction database (PHIbase)**
1275 **homolog profiles** (number homologs to entries in "reduced virulence", "unaffected
1276 pathogenicity", loss of pathogenicity", "effector", "lethal" and "increased virulence"
1277 categories in the PHIbase) in *Colletotrichum* and related species; hierarchical clustering
1278 performed with Euclidean distance and Ward linkage.
1279

1280

1281

1282

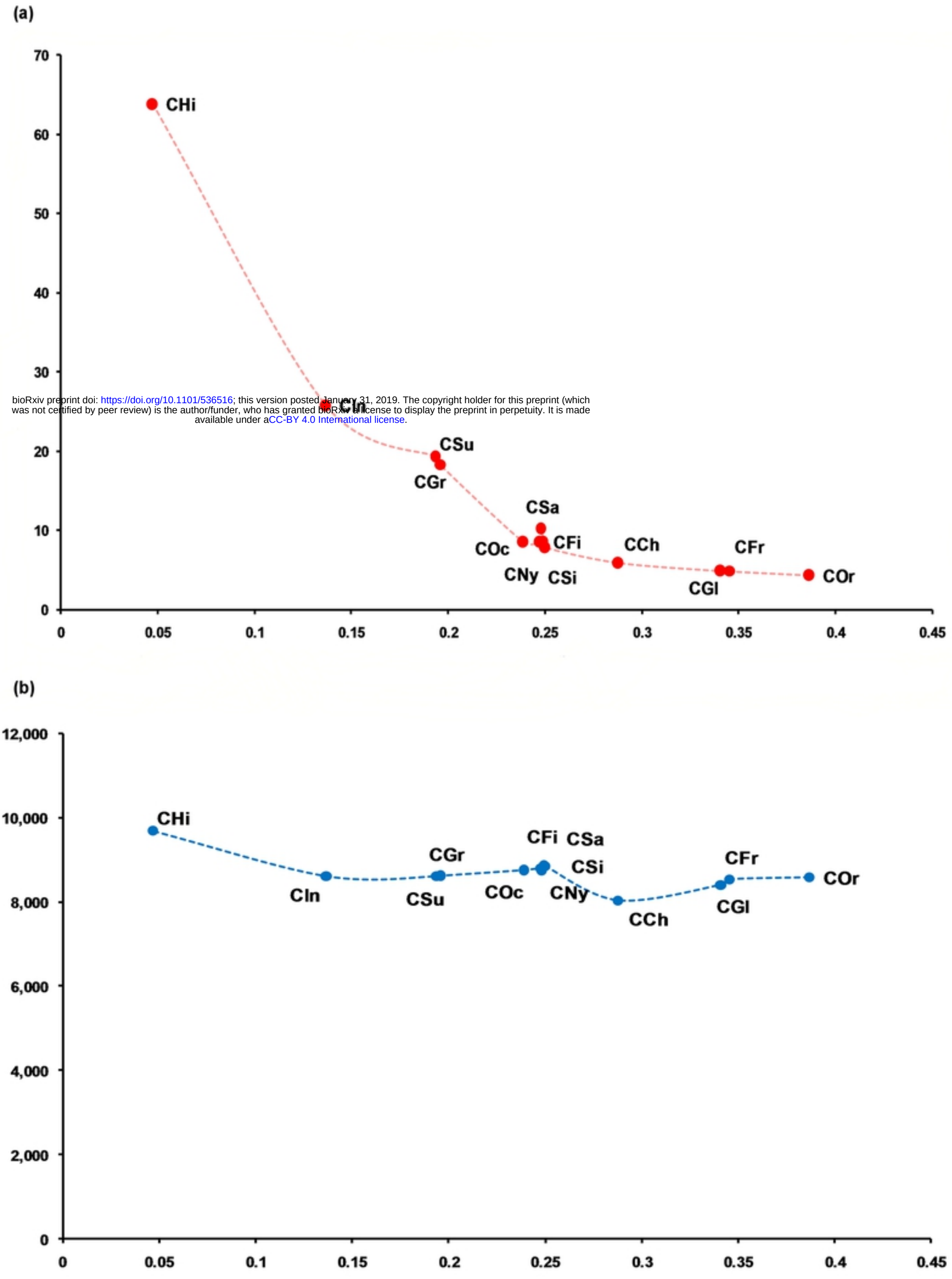

Fig 1

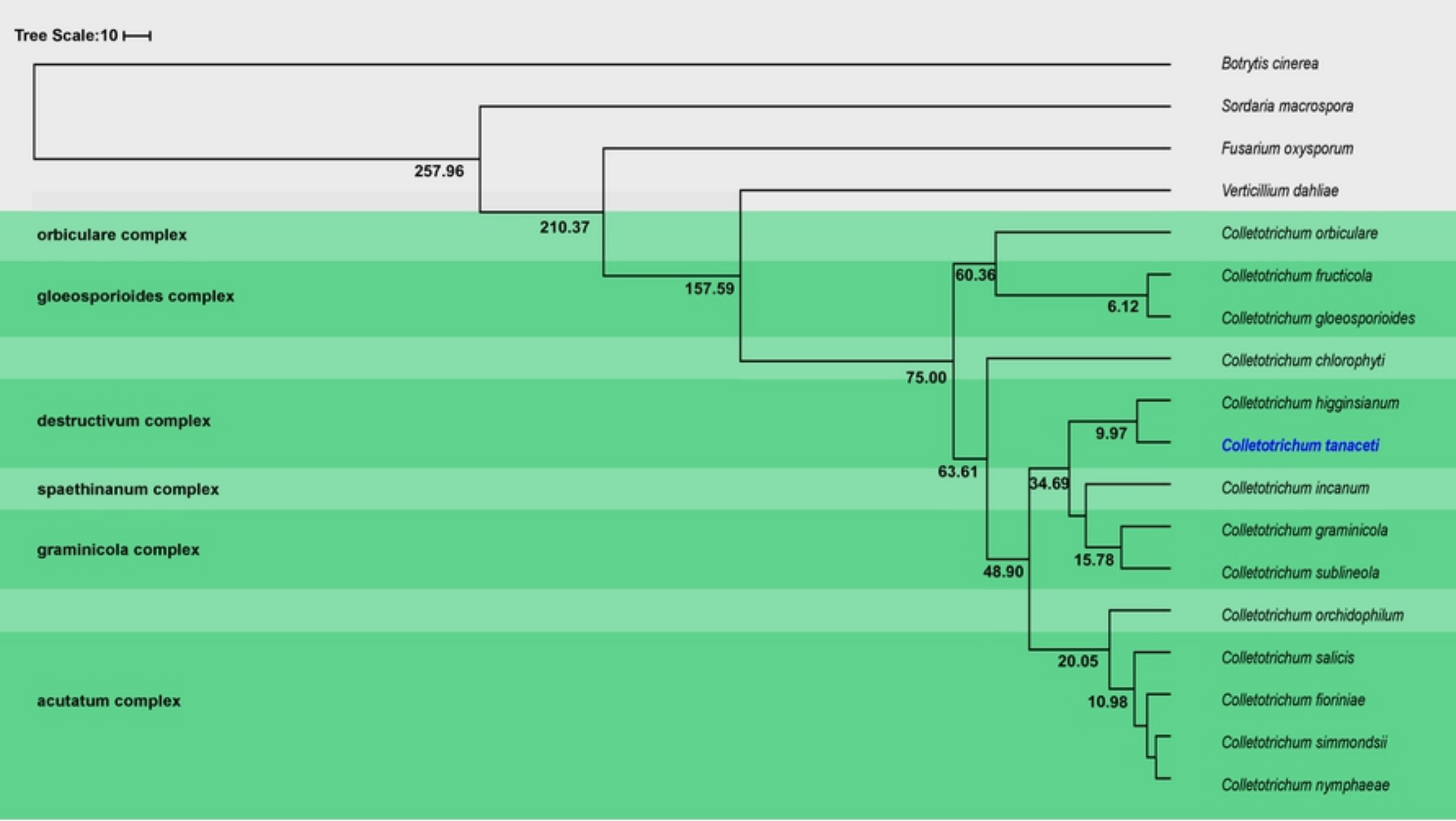
Tree Scale:10
Botrytis cinerea
Sordaria macrospora
Fusarium oxysporum
Verticillium dahliae
257.96
210.37
157.59
75.00
60.36
6.12
63.61
9.97
34.69
15.78
48.90
20.05
10.98
orbiculare complex
gloeosporioides complex
destructivum complex
spaethinanum complex
graminicola complex
acutatum complex
Colletotrichum orbiculare
Colletotrichum fructicola
Colletotrichum gloeosporioides
Colletotrichum chlorophyti
Colletotrichum higginsianum
Colletotrichum tanaceti
Colletotrichum incanum
Colletotrichum graminicola
Colletotrichum sublineola
Colletotrichum orchidophilum
Colletotrichum salicis
Colletotrichum fioriniae
Colletotrichum simmondsii
Colletotrichum nymphaeae

Fig 2

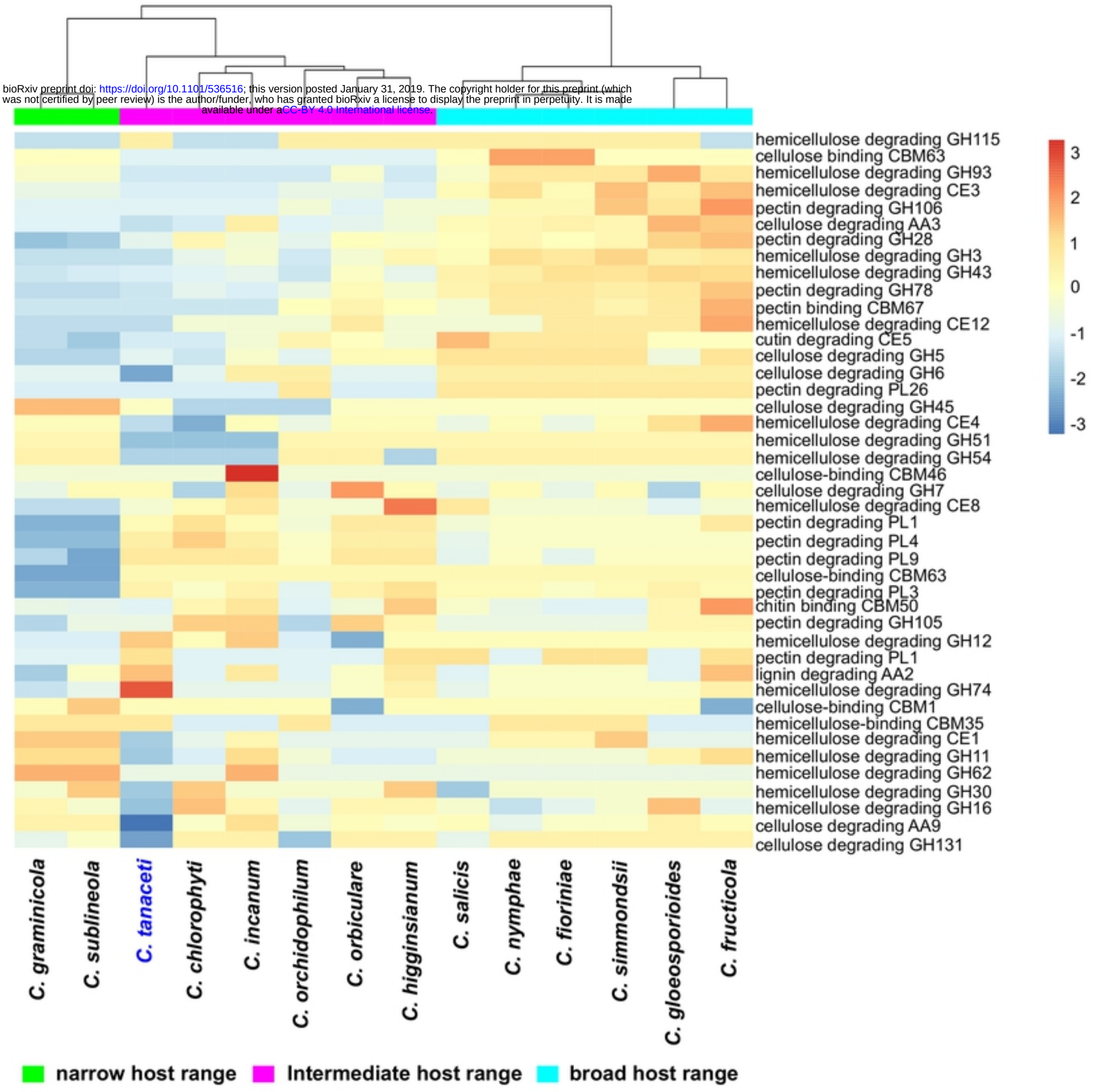

Fig 4

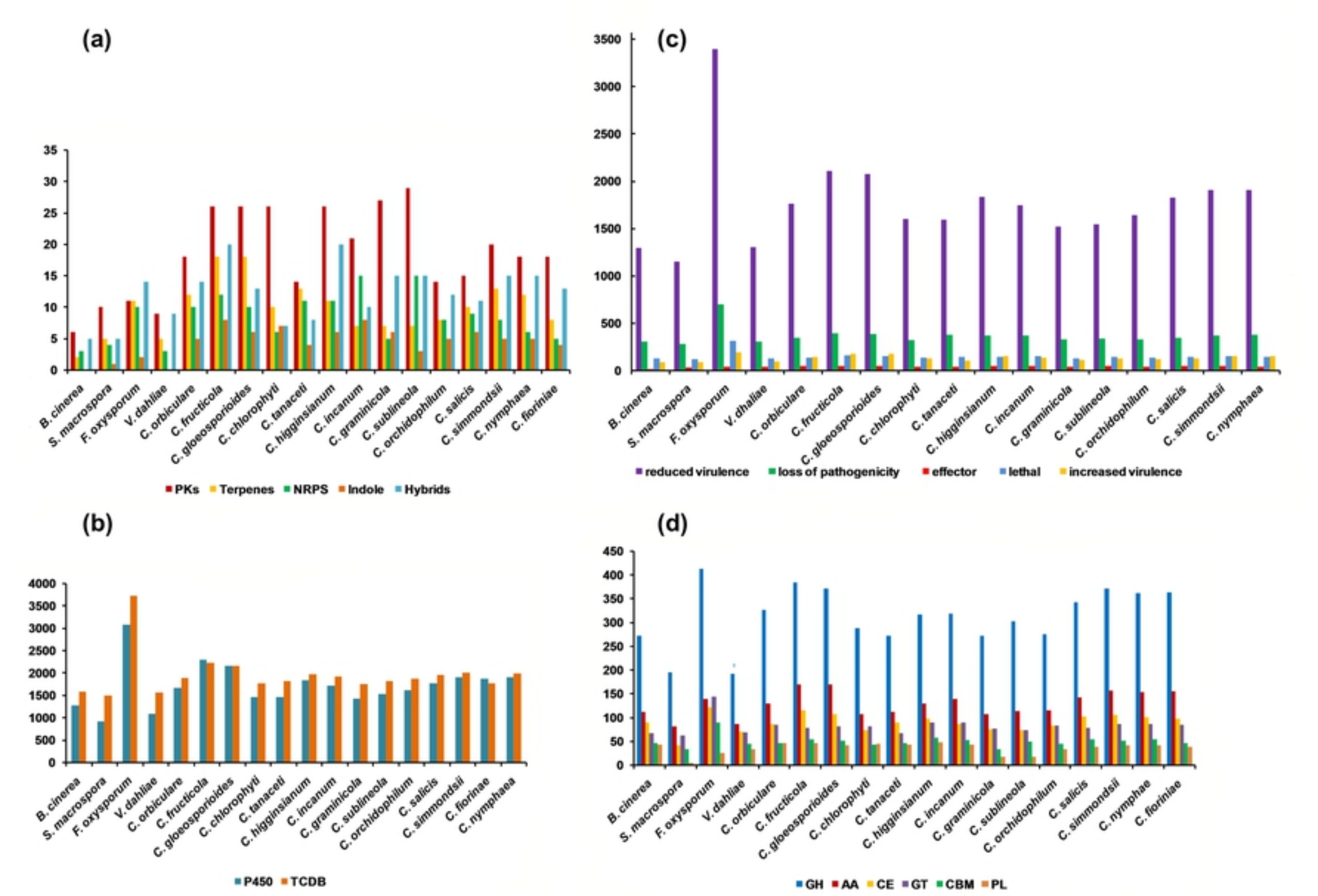
(a)
35
30
25
20
15
10
5
0
B. cinerea
S. macrospora
F. oxysporum
V. dahliae
C. orbiculare
C. fructicola
C. gloeosporioides
C. chlorophyti
C. tanaceti
C. higginsianum
C. incanum
C. graminicola
C. sublineola
C. orchidophilum
C. salicis
C. simmondsii
C. nymphaea
C. fiorniae
PKs
Terpenes
NRPS
Indole
Hybrids
(c)
3500
3000
2500
2000
1500
1000
500
0
B. cinerea
S. macrospora
F. oxysporum
V. dhaliae
C. orbiculare
C. fructicola
C. gloeosporioides
C. chlorophyti
C. tanaceti
C. higginsianum
C. incanum
C. graminicola
C. sublineola
C. orchidophilum
C. salicis
C. simmondsii
C. nymphaea
reduced virulence
loss of pathogenicity
effector
lethal
increased virulence
(b)
4000
3500
3000
2500
2000
1500
1000
500
0
B. cinerea
S. macrospora
F. oxysporum
V. dahliae
C. orbiculare
C. fructicola
C. gloeosporioides
C. chlorophyti
C. tanaceti
C. higginsianum
C. incanum
C. graminicola
C. sublineola
C. orchidophilum
C. salicis
C. simmondsii
C. fiorinae
C. nymphaea
P450
TCDB
(d)
450
400
350
300
250
200
150
100
50
0
B. cinerea
S. macrospora
F. oxysporum
V. dahliae
C. orbiculare
C. fructicola
C. gloeosporioides
C. chlorophyti
C. tanaceti
C. higginsianum
C. incanum
C. graminicola
C. sublineola
C. orchidophilum
C. salicis
C. simmondsii
C. nymphae
C. fiorimiae
GH
AA
CE
GT
CBM
PL

Fig 3

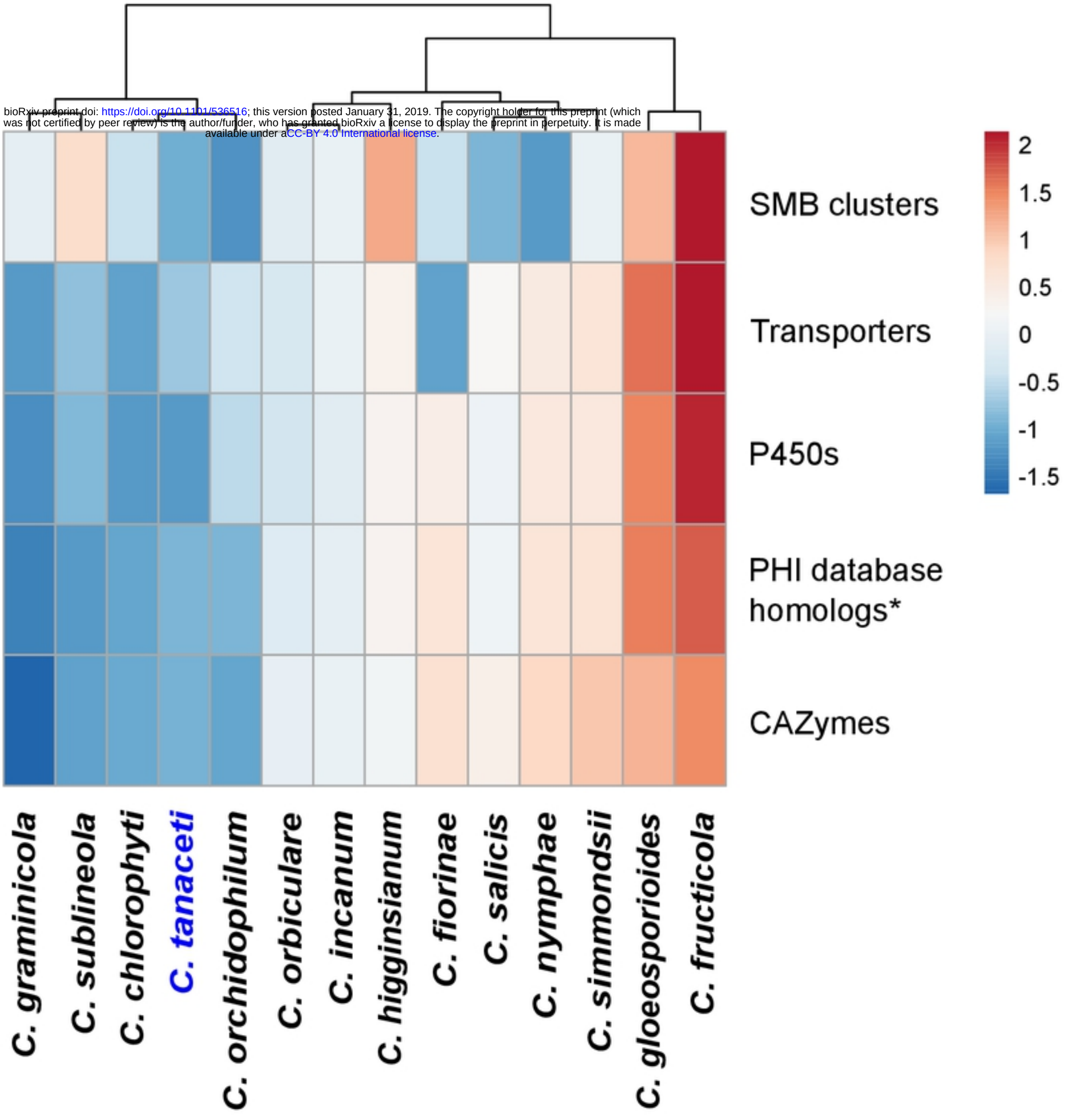

Fig 5

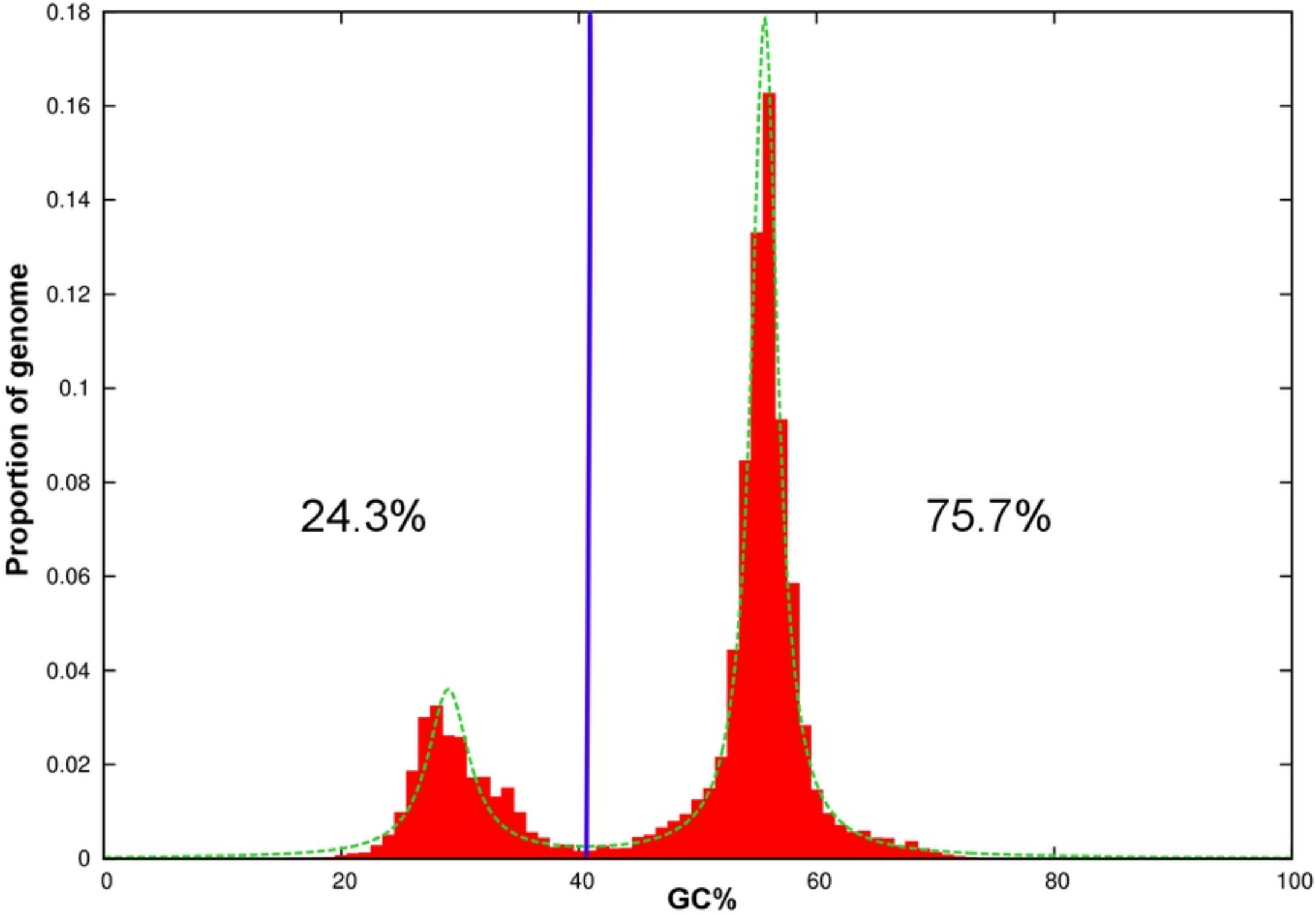
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Proportion of genome
24.3%
75.7%
0
20
40
60
80
100
GC%

Fig 6