Fungal Biology 123 (2019) 351–363

Contents lists available at ScienceDirect

Fungal Biology

journal homepage: www.elsevier.com/locate/funbio

Genomic analysis of the aggressive tree pathogen *Ceratocystis albifundus*

Magriet A. van der Nest*, Emma T. Steenkamp, Danielle Roodt, Nicole C. Soal, Marike Palmer, Wai-Yin Chan, P. Markus Wilken, Tuan A. Duong, Kershney Naidoo, Quentin C. Santana, Conrad Trollip, Lieschen De Vos, Stephanie van Wyk, Alistair R. McTaggart, Michael J. Wingfield, Brenda D. Wingfield

Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

ARTICLE INFO

Article history:
Received 4 June 2018
Received in revised form
8 February 2019
Accepted 11 February 2019
Available online 16 February 2019

Corresponding Editor: Laszlo Nagy

Keywords:
Adaptation
Ceratocystis wilt
Core and accessory genomes
Host jumps
Pathogenomics
Transposable elements

ABSTRACT

The overall goal of this study was to determine whether the genome of an important plant pathogen in Africa, *Ceratocystis albifundus*, is structured into subgenomic compartments, and if so, to establish how these compartments are distributed across the genome. For this purpose, the publicly available genome of *C. albifundus* was complemented with the genome sequences for four additional isolates using the Illumina HiSeq platform. In addition, a reference genome for one of the individuals was assembled using both PacBio and Illumina HiSeq technologies. Our results showed a high degree of synteny between the five genomes, although several regions lacked detectable long-range synteny. These regions were associated with the presence of accessory genes, lower genetic similarity, variation in read-map depth, as well as transposable elements and genes associated with host-pathogen interactions (e.g. effectors and CAZymes). Such patterns are regarded as hallmarks of accelerated evolution, particularly of accessory subgenomic compartments in fungal pathogens. Our findings thus showed that the genome of *C. albifundus* is made-up of core and accessory subgenomic compartments, which is an important step towards characterizing its pangenome. This study also highlights the value of comparative genomics for understanding mechanisms that may underly and influence the biology and evolution of pathogens.

1. Introduction

Genome comparisons provide a wealth of knowledge on the relationship among organisms and the mechanisms that shape their biology (Hardison, 2003; Wittenberg et al., 2009; Goodwin et al., 2011). From a fungal perspective, genome comparisons have provided insights into the molecular basis of speciation, host-specificity and pathogenicity mechanisms, as well as lineage-specific innovations (Plissonneau et al., 2017; Steenkamp et al., 2018). For example, comparative genomic studies have revealed various genomic features that are directly or indirectly responsible for species-specific lifestyle traits. These include genome size and predicted gene products, the pathways and processes encoded by the genome, genome architecture, gain/loss of dispensable chromosomes, repetitive genomic islands and genomic plasticity (Croll and McDonald, 2012; Ma et al., 2013; Grandaubert et al., 2014; Shi et al., 2018).

Comparative genomics has revealed that many fungi have “two-speed genomes” made-up of fast and slow-evolving subgenomic compartments (Croll and McDonald, 2012; Dong et al., 2015; Raffaele and Kamoun, 2012). The slow-evolving compartment typically contains core genes (i.e., those that are shared by all members of a species), the products of which mediate general physiology and housekeeping functions. It is usually not very rich in repetitive elements. This is in contrast to the fast-evolving compartment, which is typically repeat-rich, architecturally dynamic and contains accessory genes (i.e., those that are absent from certain members of a species). In pathogens, accessory genes are often enriched for those involved in virulence and host interactions (Ma et al., 2013; Faino et al., 2016; Plissonneau et al., 2016, 2018). Structurally, the fast-evolving subgenomic compartment may be distributed across all or most chromosomes of an individual and/or reside on specific chromosomes

[*] Corresponding author.
E-mail address: magriet.vandernest@fabi.up.ac.za (M.A. van der Nest).

https://doi.org/10.1016/j.funbio.2019.02.002
1878-6146/

that may be conditionally dispensable (Ma et al., 2010, 2013; Goodwin et al., 2011; Leclair et al., 1996; Tzeng et al., 1992; Hatta et al., 2002; Plissonneau et al., 2018).

Previous genomic comparisons of diverse fungi have shown that transposable elements (TEs) play important roles in evolution and adaptation (Grandaubert et al., 2014; Gladieux et al., 2014; Chiapello et al., 2015). TEs are mobile DNA segments capable of movement ("jumping") within a specific genome (Wicker et al., 2007; Amselem et al., 2015). This allows for the insertion of novel sequences within or close to existing genes, which may result in gene duplications, gene loss or gene inactivation (Daboussi, 1996; Amselem et al., 2015; Biémont, 2010). TEs are most prevalent in the fast-evolving subgenomic compartment (Dong et al., 2015), and their activity affects the size, structure and dynamics of the genomes harbouring them (Kidwell and Lisch, 2000; Böhne et al., 2008). TE activity has been linked, for instance, to accelerated evolution of genes involved in fungal pathogenicity and host-specificity (Fudal et al., 2009; Manning et al., 2013). This is often due to the development of gene repertoires implicated in niche expansion (Casacuberta and Santiago, 2003) or whole chromosomes enriched for TEs and genes associated with pathogenicity and virulence (Ma et al., 2010; Goodwin et al., 2011).

In this study, we investigated the genomic substructure of *Ceratocystis albifundus* (Phylum: Ascomycota; Order: Microascales; Family: Ceratocystidaceae) (De Beer et al., 2014). This fungus is an aggressive pathogen of exotic *Acacia mearnsii* (Roux et al., 1999, 2001, 2005; Heath et al., 2009) and commercially propagated *Protea cynaroides* in South Africa (Lee et al., 2016; Aylward et al., 2017). It has also been isolated from a wide range of native tree species without causing obvious signs of disease (Roux et al., 2007). Even though *C. albifundus* is homothallic with much of its reproduction occurring through selfing, populations of this fungus have high levels of genetic diversity with intermediate levels of gene flow (Roux et al., 2001, 2007; Barnes et al., 2005; Lee et al., 2016). This high diversity, together with its wide host range and absence from other continents, suggests that *C. albifundus* is native in southern Africa (Roux et al., 2001, 2007; Barnes et al., 2005). Its pathogenic niche on cultivated tree crops in South Africa may have resulted from a recent host jump and subsequent invasion (Roux et al., 2007).

Despite the importance of *C. albifundus* very little is known about the mechanisms underlying its behaviour as an aggressive tree pathogen. Knowledge regarding its genomic make-up and substructure would inform our understanding of the molecular basis of its biology, diversity and evolution. The overall goal of this study was, therefore, to determine whether the genome of *C. albifundus* is structured into core and accessory compartments, and if so, to establish how these compartments are distributed across the genome. We compared genes (especially those commonly associated with pathogenicity and interactions with the plant host), TEs and repetitive elements, and analysed synteny across five genomes of *C. albifundus* from different hosts and geographic locations. For this purpose, the publicly available genome of *C. albifundus* (van der Nest et al., 2014a) was complemented with sequenced genomes of four additional isolates using Illumina HiSeq. In addition, a high-quality reference genome for one of the individuals was assembled using a combination of PacBio and Illumina HiSeq data.

2. Materials and methods

2.1. Genome assemblies and annotation

Five isolates of *C. albifundus* originating from geographically diverse regions were included in this study (Table 1). Three of the isolates were collected in South Africa (CMW 4068, CMW 17274, CMW 17620), one in Zambia (CMW 13980) and one in Kenya (CMW 24685). Three of the isolates were from native trees (i.e., CMW 13980, CMW 17274 and CMW 17620) and two were from *A. mearnsii* (i.e., CMW 4068 and CMW 24685). Cultures were obtained from the CMW Culture Collection at the Forestry and Agricultural Biotechnology Institute at the University of Pretoria and maintained on 2 % malt extract agar (MEA, 20 gL-1 Agar, 20 gL-1 malt extract) at 25 °C (Lee et al., 2015).

The genome of isolate CMW 17620 was previously sequenced using the Illumina platform (van der Nest et al., 2014a). A total of 5 μg of DNA was prepared for the remaining isolates (CMW 4068, CMW 17274, CMW 13980 and CMW 17274) using previously described methods (Barnes et al., 2001) and sequenced using the Illumina Genome Analyser IIx platform (Genome Centre, University of California, Davis, California, USA). CLC Genomics Workbench v. 6.0.1 (CLC bio, Aarhus, Denmark) was used to trim the Illumina reads of low quality (*P* error limit of 0.05). The remaining reads were assembled using the Velvet *de novo* assembler with an optimal k-mer as determined with VelvetOptimiser (Zerbino, 2010; Zerbino and Birney, 2008). Thereafter, the pre-assemblies were scaffolded using SSPACE v. 2.0 and gaps reduced using GapFiller v. 2.2.1 (Boetzer et al., 2011; Boetzer and Pirovano, 2012). After assembly, the completeness of each genome was evaluated through the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool (BUSCO v. 1.1b1) by determining the percentage of the most highly conserved fungal gene orthologs present in the respective genomes (Simão et al., 2015). The genes or open reading frames (ORFs) for each assembly were predicted using the *de novo* prediction software AUGUSTUS with *Fusarium graminearum* gene models (Stanke et al., 2004) and then annotated using Blast2GO (Conesa et al., 2005).

For isolate CMW 4068, we also assembled a high-quality hybrid reference genome using the PacBioRS II and Illumina HiSeq sequencing platforms. A total of 30 μg of genomic DNA was prepared using the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) and sequenced by Macrogen (Seoul, Korea) using PacBio's Single Molecule Real Time (SMRT) sequencing technology. PacBio's SMRT Portal (v. 2.0.0) was used for read corrections with the default PacBio parameters, after which the genome was assembled *de novo* with Canu v. 1.7 (Berlin et al., 2015). Here, a range of values for the master errorRate parameter was evaluated and a final errorRate of 0.0075 was used. Scaffolding was done using SSPACE-LongRead v.

Table 1
Information about the *Ceratocystis albifundus* isolates used in this study[a].

Isolate number[b]	Geographic origin	Host (Family)	Collector(s)
CMW 4068	KwaZulu Natal, RSA	*Acacia mearnsii* (Fabaceae)	J. Roux
CMW 24685	Kenya	*Acacia mearnsii* (Fabaceae)	R.N. Heath & J. Roux
CMW 13980	Zambia	*Parinari curatellifolia* (Chrysobalanaceae)	J. Roux
CMW 17274	Gauteng, RSA	*Faurea saligna* (Proteaceae)	J. Roux
CMW 17620	Kruger National Park, RSA	*Terminalia serecia* (Combretaceae)	J. Roux

[a] The genome assembly for isolate CMW17620 was available from a previous study (GenBank accession number: JSSU00000000; van der Nest et al., 2014a), while those for the remaining isolates were determined here.

[b] CMW: Culture collection of the Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria.

1.1 (Boetzer and Pirovano, 2014). By including the quality-filtered Illumina data for isolate CMW 4068, the assembly was polished using Pilon v. 2 (Walker et al., 2014). For this hybrid assembly, completeness was estimated and ORFs were predicted as before.

2.2. Identification and analysis of core and accessory compartments

To determine whether the genome of *C. albifundus* is separated into core and accessory subgenomic compartments, nucleotide sequence and gene content was compared. For the gene content-based analysis, we used the reciprocal protein Basic Local Alignment Search Tool (BLASTp) to determine whether individual genes were included in all or only some of the Illumina genome assemblies. For this purpose, we used all the genes predicted from the five Illumina assemblies. Those occurring in all five of the genomes were designated as "core genes", while those predicted to be present only in some of the isolates were designated as "accessory genes". The latter also included the "unique genes" that were identified in only one of the isolates examined. All reciprocal BLASTp searches were done using a custom python script (available from the authors). The script considered gene sequences with a BLAST result Expect (E)-value cut-off of ≤0.00001 as shared between the pairs of genomes. To control for annotation inconsistencies, individual translated nucleotide BLAST (tBLASTn) searches (E-value cut-off of ≤0.00001) against the genomes were used to verify that genes designated as accessory were indeed absent in one or more of the five assemblies using BioEdit v. 7.2.5 (Hall, 2011).

JSpecies (Richter and Rosselló-Móra, 2009; Goris et al., 2007) was used to compare genome-wide (coding plus non-coding) nucleotide similarity between pairs of the five Illumina genome assemblies. For each pairwise analysis, JSpecies artificially sectioned one of the genomes into fragments ranging between 100 and 1020 nucleotides in length, which were then compared using the BLAST algorithm against the other genome, and vice versa (Altschul et al., 1997). In these comparisons, only those fragments that aligned over more than 70 % of their entire lengths and shared more than 30 % identity with the reference were considered as homologous (Goris et al., 2007). In addition to calculating the conserved fraction among the genomes, this software was also used to determine the average sequence similarity between genome pairs (Richter and Rosselló-Móra, 2009; Goris et al., 2007).

2.3. Predicted pathways and processes for the core and accessory genes

The predicted proteins in the core and accessory datasets were mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases using the GhostKoala mapping tool (Kanehisa et al., 2016). GhostKoala assigned KEGG ORTHOLOGY identifiers (K numbers) to each gene, which were used to reconstruct pathways on the KEGG web server (http://www.genome.jp/kegg/). For the accessory gene set, ClustVis (Metsalu and Vilo, 2015) was used to construct a copy number-based heatmap for each of the KEGG ORTHOLOGY identifiers in the five examined *C. albifundus* genomes. A Fisher's exact test (two-sided), implemented in Blast2GO (Conesa et al., 2005), was employed to detect Gene Ontology (GO) terms that were significantly enriched ($P < 0.05$) in the accessory set using the whole genome as reference. The REVIGO web server (Supek et al., 2011) was used to summarize these Blast2GO results.

The predicted proteins included in the core and accessory sets were also examined for the presence of those known to be involved in interactions with their host, as well as virulence and pathogenicity related proteins (Ghannoum, 2000; Paris et al., 2003; Tanabe et al., 2011; Ohm et al., 2012; Zerillo et al., 2013; Li et al., 2015). These included Carbohydrate-Active EnZymes (CAZymes), peptidases, catalases, superoxide dismutases, phospholipases and effectors. Within the two datasets, we also identified genes whose products are potentially involved in the movement and activity of TEs. Chi-squared tests were used to determine whether the frequency of these genes differed significantly ($P < 0.05$) between the core and accessory gene sets (the null hypothesis was that the frequencies did not differ between the two sets).

Putative CAZymes were identified and annotated with HMMER v. 3.0 (hmmer.org; Finn et al., 2011) using the family-specific hidden Markov model profiles in the dbCAN database (DataBase for automated Carbohydrate-active enzyme Annotation; Yin et al., 2012). Putative peptidases were identified using the MEROPS database and BLASTp searches (E-value cut-off of ≤0.00001) (http://merops.sanger.ac.uk; Rawlings et al., 2016). A similar approach was also used to compare genes across the five genomes to those in the Pathogen-Host Interactions database (PHI-base) v. 4.2 (http://www-phi4.phibase.org/), which includes a collection of experimentally verified pathogenicity, virulence and effector genes from fungi, oomycetes and bacteria (Winnenburg et al., 2008). To identify putative effectors, the two datasets were first filtered for putative secreted proteins using SignalP v. 4.1 (Petersen et al., 2011), from which we then identified those with lengths, net charge and amino acid content typical of fungal effectors using EffectorP v. 1.0 (Sperschneider et al., 2016). Putative proteins involved in the movement and activity of TEs were identified by BLASTp searches (E-value cut-off of ≤0.00001) against 2636 known reference sequences from the "Core" set of the Gypsy DataBase (GyDB) v. 2.0 (Llorens et al., 2011).

In each of the five genomes, putative catalases, superoxide dismutases and phospholipases were identified using BLASTp searches (E-value cut-off of ≤0.00001) with previously characterized protein sequences. These included catalases CATA, CATB, KATG1 and KATG2 (NCBI accession numbers MG100061, MG06442, A4R5S9 and A4QUT2, respectively), superoxide dismutases SOD1 to SOD5 (NCBI accession numbers EFZ03762, EFY99820, EFY99375, EFZ00595 and EFZ00365, respectively) and phospholipases PLA2, PLB2 and PLC2 (NCBI accession numbers KEY75421, KEY77760 and XP_011319931, respectively). These sequences were aligned using MAFFT v. 7 (Multiple Alignment using Fast Fourier Transform; http://mafft.cbrc.jp/alignment/server/) and then subjected to phylogenetic analysis using the distance-based neighbour-joining method in MEGA v. 7 (Molecular Evolutionary Genetic Analysis; http://www.megasoftware.net).

All putative host-associated genes included in the accessory gene set were subjected to selection analysis with CODEML, as implemented in PAML v. 4.9 h (Phylogenetic Analysis by Maximum Likelihood; Yang and Nielson, 2002). For this purpose, gene sequences from isolate CMW 4068 were used in local tBLASTx searches (i.e., translated nucleotide BLAST searches against a translated nucleotide database) to identify homologues in the other four isolates. The identified sequences were then aligned with MAFFT and the phylogenetic trees required by CODEML were inferred with MEGA as described above. Positive selection was evaluated in each dataset by calculating the ratio (w) of non-synonymous (dN) versus synonymous (dS) substitution rates across all sites (Yang et al., 2000). To test for variation of selective pressures across the codons, goodness of fit was calculated for the different site-specific models using likelihood ratio tests (Yang et al., 2000).

2.4. Genome conservation and synteny

We compared the Illumina assemblies for isolates CMW 17620, CMW 17274, CMW 13980 and CMW 17274 against the PacBio-

Illumina hybrid assembly for CMW 4068 using the LASTZ (Large-Scale Genome Alignment Tool) (Harris, 2007) and Mauve (Multiple Alignment of Conserved Genomic Sequence With Rearrangements) (Darling et al., 2004) plugins implemented in Geneious v. 7 (Kearse et al., 2012). LASTZ aligned the Illumina assemblies to the sequences of the 10 largest contigs in the hybrid assembly (contigs >1.1 Mb), by making use of “seed-and-extend” and “iterative refinement” strategies to allow alignment of both the conserved and more variable regions (Harris, 2007). LASTZ was also used for calculating the similarity between sequences. Mauve was used to plot sequence similarity and to identify regions of local collinearity (Darling et al., 2004). The latter are known as Locally Collinear Blocks (LCBs), which are defined as homologous sequence regions shared by the reference (i.e., the CMW 4068 hybrid assembly) and query (i.e., one of the Illumina assemblies), and that lack rearrangements. Synteny break points between the reference and query genomes were further determined using SynChro (Drillon et al., 2014), which employs Reciprocal Best-Hits (RBH) between coding sequences to identify conserved syntenic blocks. The pairwise alignments extracted from LASTZ were annotated with AUGUSTUS and used as input for the SynChro analysis. SynChro then computed RBH to reconstruct synteny block backbones, after which it automatically completed these blocks with non-RBH syntenic homologues.

CLC Genomics Workbench was used to determine the genomic distribution of single nucleotide polymorphisms (SNPs). The quality-filtered Illumina reads for each isolate were mapped (using default parameters) to the sequences for the 10 largest contigs in the CMW 4068 hybrid assembly. The R/Bioconductor package KaryoplotEr (Gel and Serra, 2017) was used to calculate and plot SNP distribution in 5000 bp, non-overlapping intervals across the sequence by expressing SNP content as the number of SNPs per interval. This package was also used to examine coverage or read depth using a sliding window of 5000 bp. The latter was calculated following the removal of duplicate reads (to avoid over-representation of specific reads due to sample preparation artefacts), with the maximum representation of a minority sequence set to 20 %.

2.5. Genomic distribution of host-associated and accessory genes in the CMW 4068 hybrid assembly

Synteny and sequence similarity were used to identify genes in the CMW 4068 hybrid assembly that potentially forms part of the accessory genome of *C. albifundus*. For this purpose, we aligned the individual Illumina assemblies to the hybrid genome assembly using LASTZ (see above). In these alignments, genes that were missing in one or more of the Illumina genome sequences were regarded as accessory genes in the CMW 4068 assembly. The locations of these accessory genes were plotted across the 10 largest contigs in the hybrid assembly using KaryoplotEr. Genes potentially involved in host interactions (CAZymes, peptidases and effectors) were identified as described before and their locations also plotted with KaryoplotEr. Differences in distribution of these accessory genes among the different genomic regions were evaluated using Chi-squared tests as described above.

2.6. TE identification, annotation and distribution in the CMW 4068 hybrid assembly

The TEdenovo and TEannot pipelines in the REPET v. 2.3 package (Flutre et al., 2011) were used to identify and annotate TEs in the CMW 4068 hybrid assembly. TEs were detected *de novo* with tBLASTx using E-value thresholds ($E < 10^{-10}$) (Gish and States, 1993), the BLASTER suite using similarity thresholds ($E = 10^{-300}$, minimum identity = 90 %) (Quesneville et al., 2003), and LTRHarvest using TE structure (Ellinghaus et al., 2008). The TEdenovo pipeline was then used to cluster the identified TEs, and to reconstruct a consensus for each group of matches using the programs Piler (Edgar and Myers, 2003), GROUPER (Quesneville et al., 2003) and RECON (Bao and Eddy, 2002). Cut-offs for clustering individual TEs were set at 90 % identity over 95 % of the length (Flutre et al., 2011). Consensus TE sequences were classified using the Repbase Update database (http://www.girinst.org/repbase/update/index.html) and named according to the classification proposed by Wicker et al. (2007).

Genomic locations of the identified TEs were plotted using KaryoploteR (https://bioconductor.org/packages/release/bioc/html/karyoploteR.html) across the 10 largest contigs of the hybrid assembly. This software was also used to plot the location of genes potentially involved in the movement and activity of TEs in isolate CMW 4068, which were identified using BLASTp searches against the “Cores” set of GyDB as described above. For the latter, differences in distribution among specific genomic regions were evaluated with Chi-squared tests.

3. Results

3.1. Genome assemblies and annotation

After data filtering, we generated high-quality raw Illumina sequence data for *C. albifundus* isolate CMW 4068 (mean read length of 94.6 bases), isolate CMW 13980 (mean read length of 95.7 bases), isolate CMW 17274 (mean read length of 88.8 bases) and isolate CMW 24685 (mean read length of 95.7 bases) (Supplementary Table 1). The filtered data were used to generate four genome assemblies that consisted of 818–2122 contigs and that were 26.7–27.2 Mb in size (Table 2), which is similar to the 27.3 Mb-assembly published for isolate CMW 17620 (van der Nest et al., 2014a). Based on the BUSCO results (Simão et al., 2015), all five of the genomes were more than 98.0 % complete (Supplementary Table 2), which is congruent with the general trend observed for *Ceratocystis* genomes (van der Nest et al., 2014a, 2014b, 2015; Wilken et al., 2013; Wingfield et al., 2015, 2016a, 2016b).

A total of 443 868 quality-filtered PacBio sequence reads (with an average length of 8317 bases) was generated for CMW 4068 (Supplementary Table 1). Assembly, scaffolding and polishing yielded 16 contigs larger than 200 000 bases, spanning a total of 28.36 Mb and containing 7103 genes (Table 2 and Supplementary Table 3). Compared to the Illumina assemblies, the hybrid assembly had a much higher N50-value (2.31 Mb compared to 0.02–0.07 Mb) and size for its largest contig (4.1 Mb compared to 0.15–0.36 Mb). The hybrid PacBio-Illumina assembly and the Illumina assemblies were similar regarding gene density (250 genes/Mb) and BUSCO completeness (98 %) (Table 2 and Supplementary Table 2), which allows for meaningful gene-based genomic comparisons.

The CMW 4068 hybrid assembly was less fragmented than the Illumina assemblies (Table 2). This is because PacBio long-read sequencing allowed for the closing of gaps and sequencing through repetitive regions (Yue et al., 2017). However, the assembly included an additional 27 contigs, each of which consisted of less than 200 000 bases. In total, the 27 small contigs spanned 1.36 Mb and contained 341 genes, which were overrepresented (Chi-squared test, $P > 0.05$) for genes showing similarity to those in the GyDB database with potential roles in TE activity and movement (Llorens et al., 2011). Another 21 % shared similarity to known genes, while the remaining genes (37 %) did not show similarity to known genes in any public database.

Table 2
Genome statistics for the five *Ceratocystis albifundus* isolates used in this study.

Isolate number[a]	Size (Mb)	Nr. of large contigs	N50	N80	Largest contig	Nr. of genes[d]	Gene density (genes/Mb)
Illumina[b]							
CMW 4068	27.05	1003	50 666	23 438	224 223	6695	248
CMW 24685	26.97	1072	48 267	21 815	289 214	6710	249
CMW 13980	27.20	818	68 699	30 671	357 115	6759	249
CMW 17274	26.68	2122	23 089	9324	153 182	6699	251
CMW 17620	27.33	939	42 183	18 306	274 425	6544	239
PacBio[c]							
CMW 4068	28.36	16	2 308 174	1 271 756	4 074 369	7103	250

[a] CMW: Culture collection of the Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria.
[b] The genome assembly for isolate CMW17620 was available from a previous study (GenBank accession number: JSSU00000000; van der Nest et al., 2014a), while those for the remaining isolates were determined here.
[c] A high-quality reference genome was produced for isolate CMW 4068 using the PacBio and Illumina HiSeq sequencing platforms. The reference genome consisted of 16 contigs (>200 000 bases).
[d] ORFs was predicted with the *de novo* gene prediction software AUGUSTUS, using the gene models of *Fusarium graminearum* (Stanke et al., 2004).

3.2. Identification and analysis of core and accessory compartments

The Illumina assemblies for the five isolates of *C. albifundus* were predicted to share 6241 genes, which were included in the core set. This represented a large portion (92.4–95.4 %) of the total number of genes predicted in each assembly (Fig. 1, Supplementary Table 4). The accessory genes (i.e., those missing from one or more of the genomes) numbered between 300 (representing 4.6 % of the CMW 17620 genome) and 515 (representing 7.6 % of the CMW 24685 genome) (Supplementary Tables 5 and 6). A small proportion of accessory genes were identified in only a single isolate (i.e., “unique genes”). These ranged from 17 (0.3 % of the genes encoded on the genomes of isolates CMW 4068 and CMW 24685) to 51 genes (0.8 % of genes encoded on the CMW 17274 genome) (Fig. 1, Supplementary Table 6).

Genome-wide nucleotide comparisons of the 5 Illumina assemblies revealed that the conserved fraction of the five *C. albifundus* genomes was very similar (Fig. 2). Based on the simple

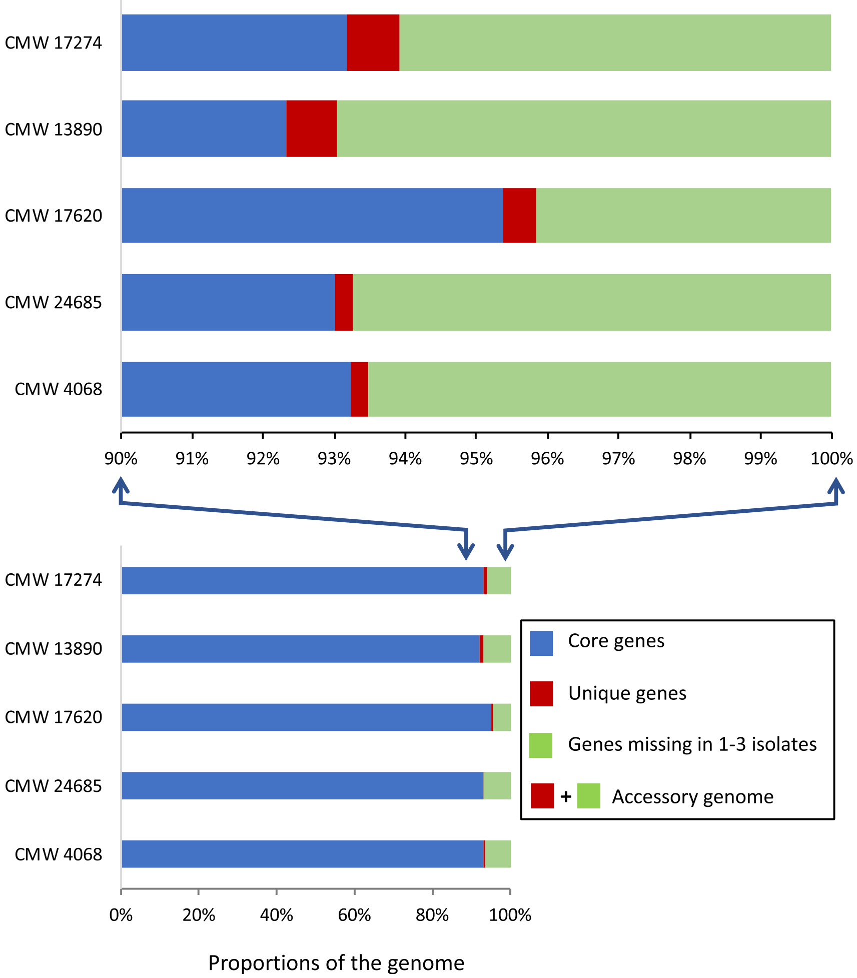
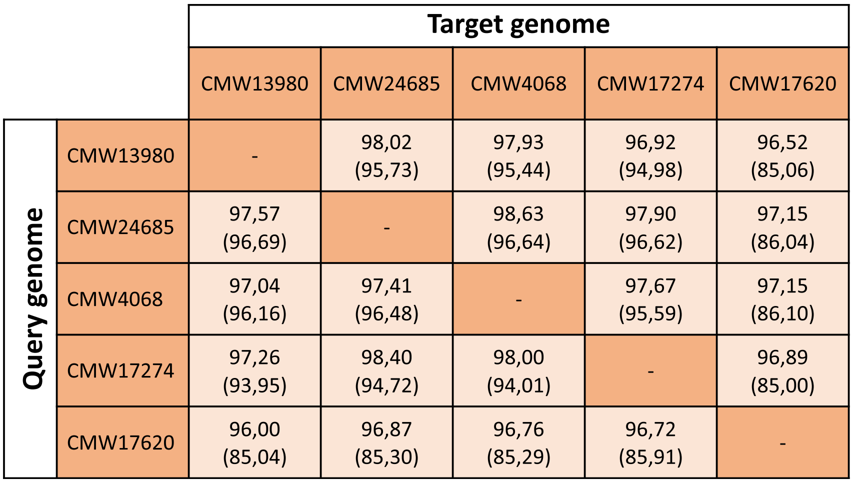


Fig. 1. The proportion of genes predicted to be encoded by the examined *Ceratocystis albifundus* genomes relative to the proportion of genes common to all five (i.e., the core genes), as well as the genes unique to a specific isolate and the genes absent from one or more of the isolates (i.e., the accessory genes).



		Target genome				
		CMW13980	CMW24685	CMW4068	CMW17274	CMW17620
Query genome	CMW13980	-	98,02 (95,73)	97,93 (95,44)	96,92 (94,98)	96,52 (85,06)
	CMW24685	97,57 (96,69)	-	98,63 (96,64)	97,90 (96,62)	97,15 (86,04)
	CMW4068	97,04 (96,16)	97,41 (96,48)	-	97,67 (95,59)	97,15 (86,10)
	CMW17274	97,26 (93,95)	98,40 (94,72)	98,00 (94,01)	-	96,89 (85,00)
	CMW17620	96,00 (85,04)	96,87 (85,30)	96,76 (85,29)	96,72 (85,91)	-

Fig. 2. Average sequence similarity (%) values for the various pairwise comparisons of the five *Ceratocystis albifundus* genomes calculated using JSpecies (Richter and Rosselló-Móra, 2009). The conserved proportions (%) of the genomes used for these comparisons are indicated in parentheses. JSpecies artificially sectioned individual genomes into fragments consisting of 100–1020 nucleotides, followed by pairwise comparisons using BLAST. Average sequence similarity was estimated only for those fragments that aligned over >70 % of their entire lengths and were >30 % similar.

BLAST-based alignment strategy implemented in JSpecies, the average sequence similarity values for the various pairwise comparisons ranged from 96.00 % (for isolates CMW 17620 and CMW 13980) to 98.63 % (for isolates CMW 24685 and CMW 4068). The non-conserved or variable regions (i.e., where JSpecies fragments that did not align across 70 % or more of their lengths and/or that were not 30 % or more similar) represented between 3.31 % and 14.96 % of the total genomes of these fungi. These variable regions may represent parts of the genome that were either too repetitive to be included in our assemblies with Velvet and/or that represent parts of the genome that are missing or variable among the genomes compared.

3.3. *Predicted pathways and processes for the core and accessory genes*

Only a portion of core (46.1 %, Table 3, Supplementary Table 7) and accessory (12–21 %; Supplementary Table 8; Supplementary Fig. S1) genes could be assigned to specific orthologs in the KEGG database. Among the various KEGG categories identified for the core genes, some are likely involved in housekeeping functions (e.g., “Transcription” and “Replication and repair”). However, it is possible that some categories are involved in functions related to niche utilization or host-pathogen interactions (e.g., “Biosynthesis of other secondary metabolites” and “Metabolism of terpenoids and polyketides”). As was expected, several of the accessory genes may be linked with host-pathogen interactions. Of the KEGG categories identified for the accessory genes, some are likely also related to niche utilization or host-pathogen interactions (e.g., “Carbon metabolism”, “Fructose and mannose metabolism”, “Fatty acid metabolism”, “Histidine metabolism”, “MAPK signalling pathway”, “mTOR signalling pathway”, and “Ras signalling pathway”) (Supplementary Table 8).

Fisher's exact test indicated that numerous GO terms were overrepresented in the accessory gene set ($P < 0.05$). These included GO terms associated with processes potentially involved in niche utilization and/or host-pathogen interactions (Supplementary Table 9; Supplementary Fig. S5). The enriched GO terms were involved in biological functions (e.g., “Establishment or maintenance of actin cytoskeleton polarity”, “Establishment or maintenance of cell polarity”, “Establishment or maintenance of cytoskeleton polarity”) and cellular components associated with host penetration (e.g., “New growing cell tip” and “Old growing cell tip”). We also found enrichment in GO terms associated with secondary metabolism (e.g., “Aromatic compound biosynthetic process” and “Cellular aromatic compound metabolic process”), cellular transport (e.g., “Nitrogen compound transport” and “Import into cell”) and signalling (e.g., “Cdc42 protein signal transduction”).

Further detailed analyses showed that several of the core genes may be linked with host-pathogen interactions. These included genes (1095) that shared significant similarity with previously

Table 3
KEGG functional classifications[a] for the core genes shared by the five *Ceratocystis albifundus* isolates.

KEGG categories	Number of shared genes associated with pathway[b]
Metabolism	
Carbohydrate metabolism	229
Energy metabolism	122
Lipid metabolism	128
Nucleotide metabolism	130
Amino acid metabolism	241
Glycan biosynthesis and metabolism	76
Metabolism of cofactors and vitamins	106
Metabolism of terpenoids and polyketides	27
Biosynthesis of other secondary metabolites	31
Xenobiotics biodegradation and metabolism	46
Genetic Information Processing	
Transcription	130
Translation	298
Folding, sorting and degradation	226
Replication and repair	143
Environmental Information Processing	
Membrane transport	5
Signal transduction	385
Cellular Processes	
Transport and catabolism	183
Cell growth and death	230
Cellular community	45

[a] For these classifications, the predicted proteins were assigned to pathways using the Kyoto Encyclopaedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/) database and the GhostKoala mapping tool (Kanehisa et al., 2016).

[b] The analysis was done using the protein sequences for the 6241 genes shared by all five of the examined isolates.

characterized pathogenicity associated genes in the PHI-database (e.g., “fungal development, secondary metabolism and virulence” and “fungal development and pathogenicity”) (Supplementary Table 10), as well as genes (56) that encoded putative effectors (Supplementary Table 4) that may modulate the host immune system and promote infection (Stergiopoulos and de Wit, 2009). Among the core genes, the MEROPS-based analyses identified various putative peptidases (260) involved in protein degradation and modification (Supplementary Table 11). CAZymes (243) potentially involved in the degradation of plant polysaccharide materials (Supplementary Table 12) and phospholipases potentially capable of hydrolysing plant phospholipids (Supplementary Fig. S2) for facilitating infection and/or gaining nutrition were also identified among the core genes (Cantarel et al., 2009; Ghannoum, 2000; Zhao et al., 2013). The core genes further included putative catalases (Supplementary Fig. S3) and superoxide dismutases (Supplementary Fig. S4) that are known to scavenge reactive oxygen species to protect fungi from the host defence responses (Tanabe et al., 2011; Li et al., 2015).

Many accessory genes (222) also shared significant similarity with previously characterized pathogenicity associated genes in the PHI-database, including genes predicted to be involved in “Cell wall adhesion”, “Establishment of turgor in appressoria” and “Appressorial penetration” (Supplementary Table 10). The accessory gene set also included those encoding putative CAZymes (Supplementary Table 12). These included enzymes that catalyse the degradation of chitin (families CBM18 and GH18) (Hartl et al., 2012) and lignin (family AA7) (Levasseur et al., 2013), as well as putative carbohydrate esterases responsible for deacetylating plant polysaccharides (families CE4 and GH10) (Biely, 2012). The accessory gene set further contained genes encoding putative peptidases (Supplementary Table 11) with possible roles in fungus-plant interactions (e.g., two putative proteases in the sedolisin family [Serine peptidase family S53], and eleven putative subtilisin peptidases [Serine peptidase family S08]) (Rawlings et al., 2016). Another 18 genes encoded putative M13 peptidases (Metallo peptidase family M13) that likely play a role in regulation of peptide signalling (Bland et al., 2008). Even though, the core and the accessory compartments of *C. albifundus* encoded significantly different gene repertoires (Chi-squared test, $P > 0.05$), the frequency of specific pathogenicity-related genes (i.e., peptidase, CAZyme and PHI-base genes) did not differ significantly between the core and accessory gene sets (Chi-squared test, $P > 0.05$).

The accessory gene set included a large number of genes encoding putative effectors (Supplementary Tables 5 and 6). These differed greatly among the isolates examined, ranging from 11 putative effector genes in isolate CMW 17620 to 36 in isolate CMW 13890. The frequency of these in the accessory gene set differed significantly from that in the core set (Chi-squared test, P-value < 0.05). Furthermore, a number of these genes also appeared to evolve under diversifying selection (Supplementary Table 16), because CODEML indicated that the positive-selection models (M5, M6 and M8) provided a better fit compared to those that assume no positive selection (M1 and M7) (Yang et al., 2000). This was also true for other host-associated genes present in the accessory set, i.e., putative peptidases (M13 peptidases implicated in regulation of peptide signalling) and CAZymes (families CBM18 and GH18 involved in chitin degradation).

A large number of core and accessory genes were predicted to encode proteins associated with the activity of TEs. For the core gene set, searches against the reference sequences in GyDB database, recovered 1009 putative genes involved in TE activity or their movement (Supplementary Table 13). Likewise, the accessory gene set included 36–80 sequences with similarity to GyDB database genes (e.g., genes encoding reverse transcriptases, retroelement integrases and *Gag*-like proteins involved in the replication and integration of certain TEs) (Wilhelm and Wilhelm, 2001; Novikova, 2009; Llorens et al., 2011). However, the frequency of these genes differed significantly between the core and accessory gene sets (Chi-squared test, $P < 0.05$) for isolates CMW 4068, CMW 24685 and CMW 13890. Also, Fisher's exact test indicated that the accessory gene set was significantly ($P < 0.05$) enriched in functions related to DNA integration processes that are known to play a role in the insertion of transposable elements in protein coding genes (Plissonneau et al., 2018).

3.4. Genome conservation and synteny

Alignments of the Illumina assemblies against the 10 largest contigs of the hybrid assembly for isolate CMW 4068 showed interrupted stretches of high similarity. Similarity between the Illumina assemblies and the hybrid assembly ranged from 96.2 % for contig 7 of CMW 13890 to 99.4 % for contig 1 of CMW 17620 (Supplementary Table 14). On most of the contigs, large stretches of similarity (i.e., LCBs identified with Mauve) were interrupted by areas that were unalignable. This was particularly evident on contigs 1, 4 and 7 (Fig. 3 and Supplementary Fig. S6). Also, a fraction of the individual Illumina reads for CMW 13890 (1.9 %), CMW 17620 (1.1 %), CMW 17274 (2.9 %) and CMW 24685 (1.5 %) did not map to the hybrid assembly for isolate CMW 4068 (Supplementary Table S15). The proportion of these “unmapped” fractions was generally somewhat lower than those observed using pairwise BLAST comparisons of the Illumina assemblies (Fig. 2). This may be because the repetitive nature of the individual genomes influences how many reads map to the reference genome, although we also cannot exclude the possible influence of differences in genome completeness and genome size on these data.

Overall, a high level of synteny (Fig. 4A) was observed among the five *C. albifundus* genomes based on the presence and order of orthologous genes using SynChro (Drillon et al., 2014). Despite this high level of gene order conservation, numerous regions lacking detectable long-range synteny were also observed relative to the 10 largest contigs of the CMW 4068 hybrid assembly (Figs. 3 and 4 and Supplementary Fig. 2). SynChro detected synteny breaks (>11.1 kb on average) on contigs 1, 4, 6, 7, 8 and 9 (Fig. 3 and Supplementary Fig. S6). Additionally, SynChro detected inversions between the hybrid assembly and Illumina assemblies: three inversions for CMW 17274, 7 inversions each for CMW 17620 and CMW 13890, and 8 inversions for CMW 24685.

The regions lacking detectable long-range synteny often occurred in SNP-dense regions and/or regions with large variation in read depth (Fig. 3; Supplementary Fig. S6). For example, the three small synteny breaks (from positions 0.53–0.55 Mb, 0.63–0.67 Mb, and 1.92–1.96 Mb in the hybrid assembly) and one large synteny break (positions 1.11–1.36 Mb in the hybrid assembly) on contig 7 were all localized in areas characterized by a higher number of SNPs in one or more of the compared genomes. The fact that these breaks typically occurred in regions with highly variable read depth (e.g., the read depth for the large syntenic break on contig 7 ranged from 20 to 170 reads/5000 bases among the genomes compared) suggests that the breaks were primarily caused by the presence of repetitive sequences in these areas.

3.5. Genomic distribution of host-associated and accessory genes in the CMW 4068 hybrid assembly

Relative to the total number of genes encoded per contig, the abundance of accessory genes (i.e., those that were missing in one or more of the four Illumina assemblies) varied substantially across the 10 largest contigs of the hybrid assembly (Supplementary

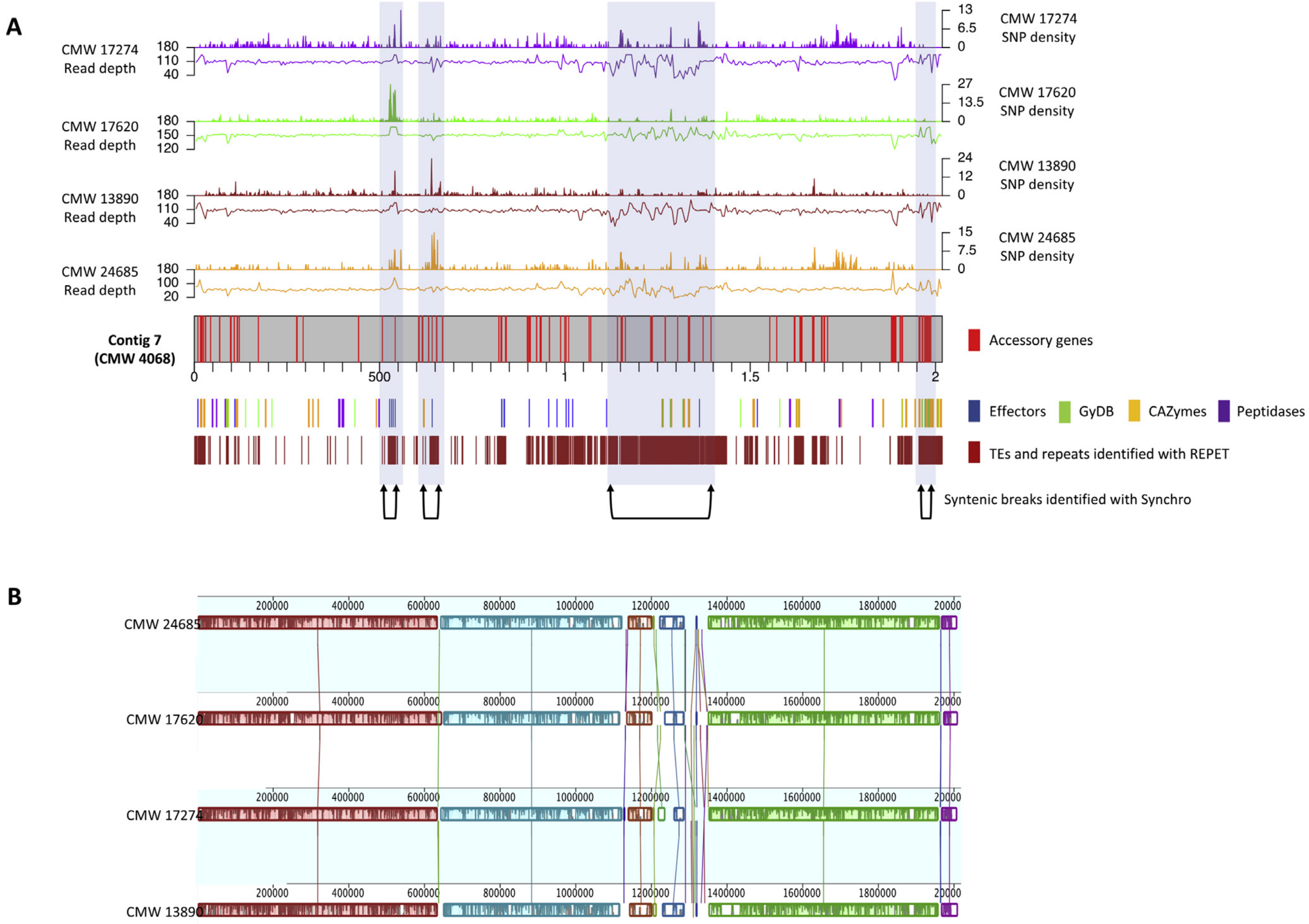


Fig. 3. Visualization of single nucleotide polymorphism (SNP) density, read depth, distribution of genes associated with host interactions and movement of TEs, as well as TEs and repeat regions along contig 7 of the *Ceratocystis albifundus* reference genome for isolate CMW 4068. (A) The peaks at the top represent SNP distribution that was defined by non-overlapping 5000 bp intervals across the sequence and measuring the SNP content as the number of SNPs per interval using the R/Bioconductor package KaryoplotEr (Gel and Serra, 2017). The peaks at the bottom represent read depth calculated using a sliding window of 5000 bp. The positions of transposable elements and repeat regions are indicated with dark brown vertical lines, putative CAZymes identified using dbCAN database (Yin et al., 2012) are indicated with yellow lines, putative peptidases identified using the MEROPS (Rawlings et al., 2016) are indicated with purple lines, putative effectors identified using EffectorP v1.0 (Sperschneider et al., 2016) are indicated with blue horizontal lines and genes with significant similarity to those previously shown to be involved in the movement and activity of TEs that are present in the Gypsy DataBase (Llorens et al., 2011) are indicated with green horizontal lines. Synteny breaks between the reference and query genomes were determined using SynChro (Drillon et al., 2014). (B) MAUVE visualisation of synteny between the five *Ceratocystis albifundus* genomes. Pairwise alignments of genomes were generated using the MAUVE plugin implemented in Geneious v. 7 (Kearse et al., 2012). Locally Collinear Blocks are marked with the same colour and connected by straight lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Fig. S6). For example, accessory genes were much more abundant on contigs 6 (15.1 %) and 7 (19.2 %) than contigs 2 (8.9 %), 3 (5.4 %) and 8 (8.9 %) (Fig. 3 and Supplementary Fig. S6). Although the accessory genes were generally distributed across contigs, they often appeared to be localized in regions lacking detectable long-range synteny and that SynChro identified as syntenic breaks (e.g., positions 0.78–0.89 Mb on contig 4, positions 0.13–0.50 Mb and 1.94–1.99 Mb on contig 6 and on contig 7 positions 1.92–1.96 Mb). Accordingly, Chi-squared tests of independence rejected the null expectation that the frequency of accessory genes located in regions lacking long-range synteny is the same as in the rest of the contig (P-value < 0.05).

The genes encoding products potentially involved in host interactions and in the activity and movement of TEs appeared to be randomly positioned on the 10 largest contigs of the hybrid assembly. These included CAZymes, peptidases and putative effectors, as well as GyBD-identified genes involved in movement and activity of TEs (Fig. 3, Supplementary Fig. S6 and Supplementary Tables 11–13). However, the non-syntenic regions appeared to be enriched for putative effectors (contig 7), CAZyme (contigs 6 and 8) and TE-associated (contigs 1, 4 and 6) genes (Fig. 3 and Supplementary Fig. S6).

3.6. *TE identification, annotation and distribution in the CMW 4068 hybrid assembly*

Based on their predicted transposition mechanisms, the transposable elements and repeat sequences for the hybrid assembly were classified as Class I, Class II or “NoClass” for those that could not be assigned into either of the two classes by REPET (Fig. 4B; Supplementary Table 17). Most of the annotated TEs represented Class I transposons (commonly referred to as retrotransposons), which utilize a copy-and-paste mechanism for transposition via an RNA intermediate (Wicker et al., 2007; Amselem et al., 2015). Those TEs annotated as being Class II transposons (commonly referred to as DNA transposons) likely use a cut-and-paste mechanism involving transposases and DNA intermediates (Wicker et al., 2007; Amselem et al., 2015). The annotated Class I and Class II TEs were

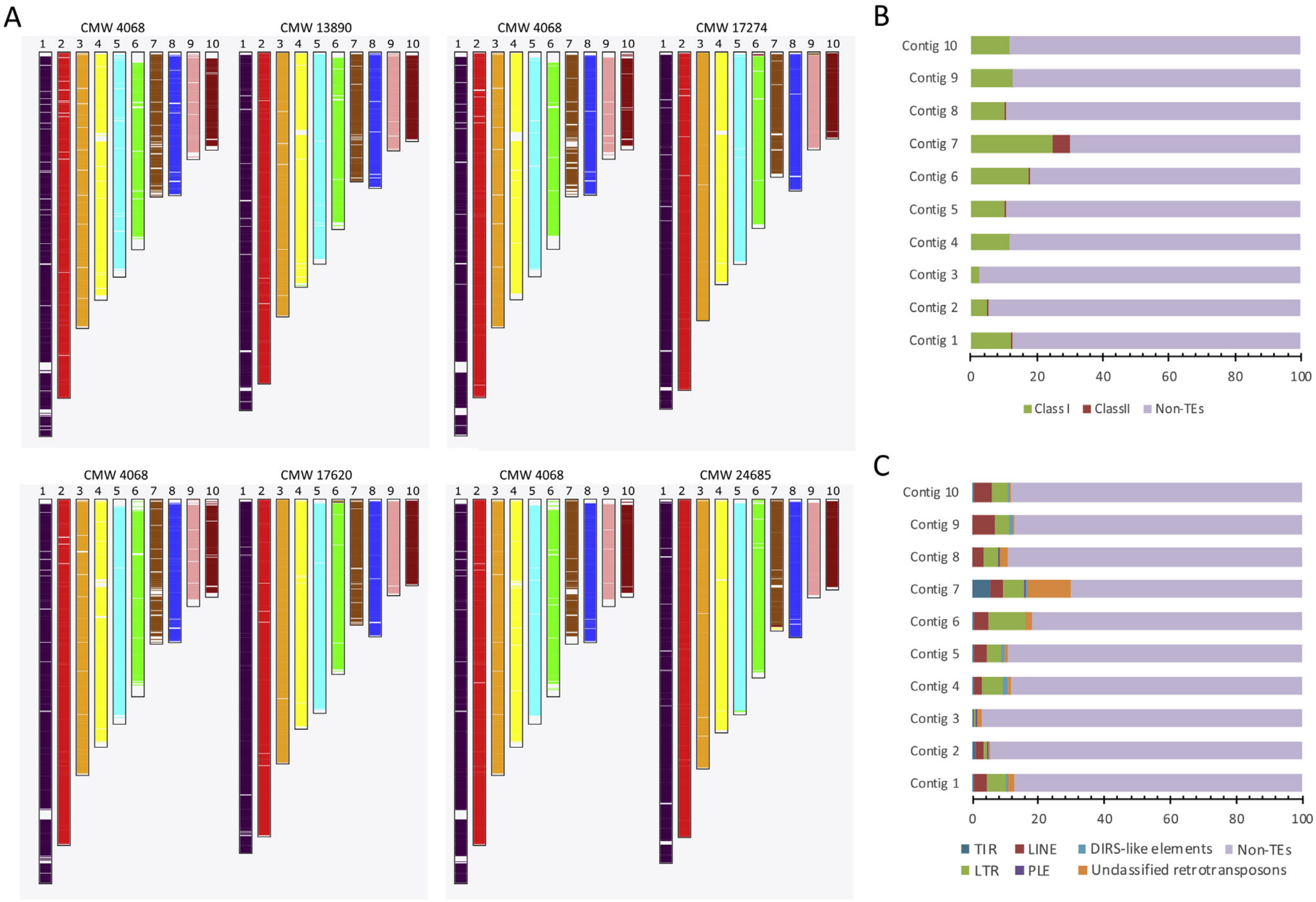


Fig. 4. (**A**) Genome organization of the five *Ceratocystis albifundus* genomes. Conserved synteny blocks were defined between pairwise combinations of the five genomes (i.e., the CMW 4068 hybrid assembly against each of the Illumina assemblies) using Synchro (options: 0 3; 0 for all pairwise, and 2 for delta of RBH genes) (Drillon et al., 2014). This program defined orthology relationships between genes from different isolates on the basis of bidirectional hits in a BLASTp comparison (reciprocal best hits). Different colours are used to differentiate gene contents in different ancestral *C. albifundus* contigs. The colour white indicated the absence of orthologs in the other isolate. Classes (**B**) and orders (**C**) of TEs identified in the hybrid assembly of *Ceratocystis albifundus* CMW 4068 based on the classification scheme of Wicker et al. (2007). The various elements were denoted as follows: TIR = Terminal Inverted Repeats; MITE = Miniature Inverted Repeat; LINE = Long Interspersed Nuclear Elements; LTR = Long Terminal Repeat; LARD = Large retrotransposon derivatives; TRIM = Terminal repeat transposons in miniature; SINE = Short Interspersed Elements; and Unclassified = non-autonomous retrotransposons. Those in Class I included LTR, LINES, SINES, LARD and TRIM), Class II included TIR and MITE, while the unclassified TEs formed part of NoClass. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

further grouped into orders (Wicker et al., 2007), based on their insertion mechanism, structure and encoded proteins (Fig. 4C; Supplementary Table 17). The annotated TEs were represented by 5 orders, namely LTR [Long Terminal Repeat], LINE [Long Interspersed Nuclear Elements], TIR [Terminal Inverted Repeats], DIRS-like elements [Dictyostelium intermediate repeat sequence] and PLE [Penelope-like elements].

A large portion (16.3 % for the 16 contigs > 2 Mb) of the hybrid assembly is represented by TEs. The occurrence and distribution of the various TEs differed substantially among contigs (Figs. 3 and 4B, C, Supplementary Table 17). For instance, contig 7 contained a much higher proportion of TEs (30 %) than contig 2 of which only 5.5 % were dedicated to TEs. The more TE-dense regions also appeared to occur in areas lacking detectable long-range synteny (Fig. 3A and Supplementary Fig. 2). On contig 7, for example, all four of the syntenic breaks co-localize with TE-dense regions (Fig. 3A). In terms of TE distribution, we observed some clustering (e.g., in syntenic breaks and at the ends of the contigs), however, many TEs also seemed to be spread out across contigs (Fig. 3A and Supplementary Fig. 2).

The ends of almost all of the contigs in the CMW 4068 hybrid assembly were rich in TEs and repeats. Chi-squared tests showed that the frequencies of these elements within the terminal 20 000 bases of each contig were significantly different from those outside these regions ($P < 0.05$). Furthermore, within the terminal 20 000 bases, five of the large contigs (i.e., 4, 5, 6, 12 and 14) contained 20–38 copies of the telomeric repeat 5′ TTAGGG 3' motif (Fulnečková et al., 2013; van Wyk et al., 2018), but only at one of their ends. It is therefore unlikely that any of our contigs represent chromosome-sized scaffolds, indicating that the Pacbio long-reads were not adequate for sequencing across the repetitive regions and assembling to chromosome level.

4. Discussion

The results of this study demonstrated that the genome of *C. albifundus* is comprised of core and accessory subgenomic compartments. This is similar to what has been observed in other fungi (Croll and McDonald 2012; Gladieux et al., 2014; Ma et al., 2013; Dong et al., 2015; Ohm et al., 2012) and may correspond to the eu- and heterochromatic DNAs of eukaryotes (Vanrobays et al., 2018). In our study, this was evidenced by the stretches of individual genomes that were highly variable and lacking synteny, and that were interspersed with conserved regions of high sequence similarity.

Comparisons of our five Illumina assemblies also revealed large proportions of genes common to all isolates of *C. albifundus* (*viz.* forming part of the core subgenomic compartment), as well as genes (4.6–7.6 % of those predicted per isolate) that were present in only some isolates (*viz.* forming part of the accessory subgenomic compartment). Such individual or lineage-specific patterns of gene presence/absence have also been reported for the heterochromatic regions of other eukaryotes (Fortna et al., 2004; Dopman and Hartl, 2007). Our study thus represents an important step towards characterizing the pangenome (i.e., the combined core and accessory genomes) of *C. albifundus*, as these subgenomic compartments have important and distinctly different roles in the biology and evolution of a pathogen (Plissonneau et al., 2018).

The core and the accessory compartments of *C. albifundus* encoded different gene repertoires, consistent with previous reports (Plissonneau et al., 2016, 2018). Core genes were predicted to mostly encode basal or housekeeping functions (see Table 2), while accessory genes encoded putative products required for access to plant-derived nutrients (Lee and Sheppard, 2016; Ohm et al., 2012; Zhao et al., 2013) and host-pathogen interactions (Desjardins and Hohn, 1997; Mukherjee et al., 2012). The accessory genes also encoded putative products for signal transduction in sensing and responding to environmental conditions, thus enabling growth and survival in a specific biological niche (Braunsdorf et al., 2016). This knowledge provides a foundation for future functional studies that aim to clarify the roles of these proteins, and the possibility of manipulating them to improve disease management.

Structural analyses of the *C. albifundus* genomes suggested that the accessory subgenomic regions are distributed throughout the genome rather than located on specific chromosomes (Ma et al., 2013; Goodwin et al., 2011; Leclair et al., 1996; Tzeng et al., 1992; Hatta et al., 2002). The percentage and distribution of accessory genes in this genome is comparable to what has been reported in *Zymoseptori tritici* (Plissonneau et al., 2016, 2018). The accessory subgenomic compartment of *C. albifundus* may contain 4.9–8.3 % of all the genes predicted in an isolate (in *Z. tritici* 1.8–8.5 % of genes lack homologues in one or more isolates) (Plissonneau et al., 2016, 2018). Like in *Z. tritici*, many *C. albifundus* accessory genes also co-occurred in areas lacking long-range synteny (Plissonneau et al., 2016, 2018). However, *Ceratocystis* differs from fungi such as *Fusarium*, where isolate or lineage-specific genomic regions are localized to specific chromosomes (Ma et al., 2010, 2013).

The accessory subgenomic compartments in fungi are often enriched for genes involved in virulence and pathogenicity (Ma et al., 2013; Faino et al., 2016; Plissonneau et al., 2016, 2018). This may also be true for *C. albifundus*, since many non-syntenic regions contained genes encoding putative CAZymes and effectors (Ohm et al., 2012; Dong et al., 2015; Fouché et al., 2018; Laurie et al., 2012; Faino et al., 2016). Effector genes commonly reside in rapidly evolving accessory compartments in genomes of filamentous plant pathogens (Laurie et al., 2012; Dong et al., 2015; Faino et al., 2016; Fouché et al., 2018). As shown before, putative effector genes were identified in the accessory compartment of *C. albifundus*. Also, consistent with what has been observed for the accessory genes in other fungal pathogens (Raffaele and Kamoun, 2012), some of the host-associated functions (i.e., encoding effectors, CAZymes and peptidases) encoded on the accessory set of *C. albifundus* evolve under diversifying selection. These data therefore suggest that some of the accessory genes in *C. albifundus* encode products that modulate host immune responses, promote infection and effective colonization of plant hosts and that allow adaptation of the fungus to changing environments (Stergiopoulos and de Wit, 2009).

Our study suggests that *C. albifundus* has a two-speed-genome (Croll and McDonald, 2012; Dong et al., 2015). This is because the regions enriched with accessory genes in the *C. albifundus* genome appeared to be co-localised with synteny breaks, and these breaks bear all of the hallmarks of fast-evolving regions. They contained unique sequences and genes, displayed low sequence similarity and high SNP density, as well as vary greatly in Illumina sequencing read depth. This is similar to what has been observed in *Verticillium dahliae* (Faino et al., 2016). Also, disruptions in long-range synteny have been linked to the divergence of *Saccharomyces cerevisiae* and its wild relative *S. paradoxus* (Yue et al., 2017). In *C. albifundus*, the presence of such a fast-evolving subgenomic compartment may help to explain the host range and high levels of genetic diversity reported in previous studies (Roux et al., 2001, 2007; Barnes et al., 2005).

In fungal pathogens, the accelerated evolutionary rates of accessory subgenomic compartments are often linked to the presence of TEs (Croll and McDonald, 2012; Raffaele and Kamoun, 2012; Vanheule et al., 2016; Faino et al., 2016). Consistent with this view, the accessory genes of *C. albifundus* were overrepresented (38–65 %) for genes that encode proteins involved in the movement or activity of TEs. Also, the genomic locations of synteny breaks and TE density were clearly coordinated. The TE content of *C. albifundus* (16.3 % for the hybrid reference genome) was higher than what was reported for many other fungi. These include other necrotrophic pathogens such as *Botrytis cinerea* (0.7–2.2 %), *Stagonospora nodorum* (2.4 %), *Alternaria brassicicola* (5.6 %), as well as hemibiotrophs like *Dothistroma septosporum* (0.7 %), *Cochliobolus sativus* (5.4 %) and *Mycosphaerella populorum* (3.6 %) (Grandaubert et al., 2014; Ohm et al., 2012; Amselem et al., 2011, 2015). In fact, the TE content in the *C. albifundus* genome was similar to fungi in which large-scale TE invasions have been reported. These include hemibiotrophs such as *Leptosphaeria maculans* (25 %) and *Mycosphaerella graminicola* (11.7 %), as well as ectomycorrhizal fungi such as *Laccaria bicolor* (24 %) and *Tuber melanosporum* (58 %) (Ohm et al., 2012; Labbe et al., 2012; Raffaele and Kamoun, 2012; Grandaubert et al., 2014).

The findings presented in this study suggest that TEs likely played a significant role in the evolution of *C. albifundus*. The genome of this fungus is relatively rich in these elements and some of the genetic diversity observed in *C. albifundus* might have been the product of varied TE activities over time. Through their activity, TEs could have shaped the genomic landscape of *C. albifundus* by causing chromosomal rearrangements, deletions and duplications (Daboussi, 1997; Daboussi and Capy, 2003). In addition to gene and genomic plasticity, TE activity might also have caused phenotypic diversity through epigenetic mechanisms and/or other changes in gene regulation (Daboussi and Capy, 2003). TEs and the dynamic accessory subgenomic compartment in which they occur are thus likely to have been important sources of diversity and adaptive phenotypes in *C. albifundus* (Croll and McDonald, 2012; Gladieux et al., 2014; Ma et al., 2013; Dong et al., 2015).

Data availability

The Whole Genome Shotgun projects have been deposited at DDBJ/EMBL/GenBank under accession numbers MAOA00000000, MANZ00000000, MANY00000000, MANX00000000 and MANW00000000. The hybrid assembly for isolate CMW 4068 was used to update the existing Illumina sequence for this isolate, and is available as version MAOA02000000.

Acknowledgements

We thank the University of Pretoria, the South African Department of Science and Technology (DST), National Research Foundation (NRF), Centre of Excellence in Tree Health Biotechnology, the

DST-NRF SARChI Chair in Fungal Genomics, and the Genomics Research Institute (University of Pretoria) for financing the sequencing of the genomes. This study was based on research supported by multiple grants from the NRF of South Africa, and these include Grant specific unique reference number (UID) 83924. The Grant holders acknowledge that opinions, findings and conclusions or recommendations expressed in publications generated by NRF supported research are that of the authors, and that the NRF accepts no liability whatsoever in this regard.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.funbio.2019.02.002.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Amselem, J., Cuomo, C.A., Van Kan, J.A., Viaud, M., Benito, E.P., Couloux, A., Coutinho, P.M., De Vries, R.P., Dyer, P.S., Fillinger, S., Fournier, E., 2011. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS Genet. 7, e1002230.

Amselem, J., Lebrun, M.-H., Quesneville, H., 2015. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics 16, 141.

Aylward, J., Wingfield, B.D., Dreyer, L.L., Roets, F., Wingfield, M.J., Steenkamp, E.T., 2017. Contrasting carbon metabolism in saprotrophic and pathogenic microascalean fungi from Protea trees. Fungal Ecol. 30, 88–100.

Bao, Z., Eddy, S.R., 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 12, 1269–1276.

Barnes, I., Gaur, A., Burgess, T., Roux, J., Wingfield, B.D., Wingfield, M.J., 2001. Microsatellite markers reflect intra-specific relationships between isolates of the vascular wilt pathogen *Ceratocystis fimbriata*. Mol. Plant Pathol. 2, 319–325.

Barnes, I., Nakabonge, G., Roux, J., Wingfield, B.D., Wingfield, M.J., 2005. Comparison of populations of the wilt pathogen *Ceratocystis albifundus* in South Africa and Uganda. Plant Pathol. 54, 189–195.

Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M., Phillippy, A.M., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630.

Biely, P., 2012. Microbial carbohydrate esterases deacetylating plant polysaccharides. Biotechnol. Adv. 30, 1575–1588.

Biémont, C., 2010. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 186, 1085–1093.

Bland, N.D., Pinney, J.W., Thomas, J.E., Turner, A.J., Isaac, R.E., 2008. Bioinformatic analysis of the neprilysin (M13) family of peptidases reveals complex evolutionary and functional relationships. BMC Evol. Biol. 23, 1.

Boetzer, M., Pirovano, W., 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinf. 15, 211.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., Pirovano, W., 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579.

Boetzer, M., Pirovano, W., 2012. Toward almost closed genomes with GapFiller. Genome Biol. 13, R56.

Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., Volff, J.-N., 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res. 16, 203–215.

Braunsdorf, C., Mailänder-Sánchez, D., Schaller, M., 2016. Fungal sensing of host environment. Cell Microbiol. 18, 1188–2000.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B., 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 37, D233–D238.

Casacuberta, J.M., Santiago, N., 2003. Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311, 1–11.

Chiapello, H., Mallet, L., Guérin, C., Aguileta, G., Amselem, J., Kroj, T., Ortega-Abboud, E., Lebrun, M.H., Henrissat, B., Gendrault, A., Rodolphe, F., 2015. Deciphering genome content and evolutionary relationships of isolates from the fungus *Magnaporthe oryzae* attacking different host plants. Genome Biol. Evol. 7, 2896–2912.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 18, 3674–3676.

Croll, D., McDonald, B.A., 2012. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog. 8, e1002608.

Daboussi, M.-J., 1997. Fungal transposable elements and genome evolution. Genetica 100, 253–260.

Daboussi, M.-J., Capy, P., 2003. Transposable elements in filamentous fungi. Annu. Rev. Microbiol. 57, 275–299.

Daboussi, M.J., 1996. Fungal transposable elements: generators of diversity and genetic tools. J. Genet. 75, 325–339.

Darling, A.C., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14, 1394–1403.

De Beer, Z.W., Duong, T., Barnes, I., Wingfield, B.D., Wingfield, M.J., 2014. Redefining *Ceratocystis* and allied genera. Stud. Mycol. 79, 187–219.

Desjardins, A.E., Hohn, T.M., 1997. Mycotoxins in plant pathogenesis. Mol. Plant Microbe Interact. 10, 147–152.

Dong, S., Raffaele, S., Kamoun, S., 2015. The two-speed genomes of filamentous pathogens: waltz with plants. Curr. Opin. Genet. Dev. 35, 57–65.

Dopman, E., Hartl, D., 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S.A. 104, 19920–19925.

Drillon, G., Carbone, A., Fischer, G., 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PLoS One 9, e92621.

Edgar, R.C., Myers, E.W., 2003. PILER: identification and classification of genomic repeats. Bioinformatics 21, 152–158.

Ellinghaus, D., Kurtz, S., Willhoeft, U., 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinf. 9, 18.

Faino, L., Seidl, M.F., Shi-Kunne, X., Pauper, M., van den Berg, G.C., Wittenberg, A.H., Thomma, B.P., 2016. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res. 26, 1091–1100.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37.

Flutre, T., Duprat, E., Feuillet, C., Quesneville, H., 2011. Considering transposable element diversification in *de novo* annotation approaches. PLoS One 6, e16526.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., Karimpour-Fard, A., 2004. Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. 2, e207.

Fouché, S., Plissonneau, C., Croll, D., 2018. The birth and death of effectors in rapidly evolving filamentous pathogen genomes. Curr. Opin. Microbiol. 46, 34–42.

Fudal, I., Ross, S., Brun, H., Besnard, A.-L., Ermel, M., Kuhn, M.-L., Balesdent, M.-H., Rouxel, T., 2009. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Mol. Plant Microbe Interact. 22, 932–941.

Fulnečková, J., Ševčíková, T., Fajkus, J., Lukešova, A., Lukeš, M., Vlček, Č., Lang, B.F., Kim, E., Eliaš, M., Sýkorova, E., 2013. A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. Genome Biol. Evol. 5, 468–483.

Gel, B., Serra, E., 2017. karyoploteR: an R/Bioconductor package to plot customizable linear genomes displaying arbitrary data. BioRxiv 122838.

Ghannoum, M., 2000. Potential role of phospholipases in virulence and fungal pathogenesis. Clin. Microbiol. Rev. 13, 122–143.

Gish, W., States, D.J., 1993. Identification of protein coding regions by database similarity search. Nat. Genet. 3, 266–272.

Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguileta, G., Vienne, D.M., Rodríguez de la Vega, R.C., Branco, S., Giraud, T., 2014. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. Mol. Ecol. 23, 753–773.

Goodwin, S.B., M'Barek, S.B., Dhillon, B., Wittenberg, A.H., Crane, C.F., Hane, J.K., Foster, A.J., Van der Lee, T.A., Grimwood, J., Aerts, A., Antoniw, J., 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7, e1002070.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91.

Grandaubert, J., Lowe, R.G., Soyer, J.L., Schoch, C.L., Van de Wouw, A.P., Fudal, I., Robbertse, B., Lapalu, N., Links, M.G., Ollivier, B., Linglin, J., 2014. Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal pathogens. BMC Genomics 15, 1.

Hall, T., 2011. BioEdit: an important software for molecular biology. GERF Bull. Biosci. 2, 60–61.

Hardison, R.C., 2003. Comparative genomics. PLoS Biol. 1, 156–160.

Harris, R.S., 2007. Improved Pairwise Alignment of Genomic DNA. The Pennsylvania State University.

Hartl, L., Zach, S., Seidl-Seiboth, V., 2012. Fungal chitinases: diversity, mechanistic properties and biotechnological potential. Appl. Microbiol. Biotechnol. 93, 533–543.

Hatta, R., Ito, K., Hosaki, Y., Tanaka, T., Tanaka, A., Yamamoto, M., Akimitsu, K., Tsuge, T., 2002. A conditionally dispensable chromosome controls host-specific pathogenicity in the fungal plant pathogen *Alternaria alternata*. Genetics 161, 59–70.

Heath, R.N., Wingfield, M.J., Wingfield, B.D., Meke, G., Mbaga, A., Roux, J., 2009. *Ceratocystis* species on *Acacia mearnsii* and *Eucalyptus* spp. in eastern and southern Africa including six new species. Fungal Divers. 34, 41–67.

Kanehisa, M., Sato, Y., Morishima, K., 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726–731.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649.

Kidwell, M.G., Lisch, D.R., 2000. Transposable elements and host genome evolution. Trends Ecol. Evol. 15, 95–99.

Labbe, J., Murat, C., Morin, E., Tuskan, G.A., Le Tacon, F., Martin, F., 2012. Characterization of transposable elements in the ectomycorrhizal fungus *Laccaria bicolor*. PLoS One 7, e40197.

Laurie, J.D., Ali, S., Linning, R., Mannhaupt, G., Wong, P., Güldener, U., Münsterkötter, M., Moore, R., Kahmann, R., Bakkeren, G., Schirawski, J., 2012. Genome comparison of barley and maize smut fungi reveals targeted loss of RNA silencing components and species-specific presence of transposable elements. Plant Cell 24, 1733–1745.

Leclair, S., Ansan-Melayah, D., Rouxel, T., Balesdent, M., 1996. Meiotic behavior of the minichromosome in the phytopathogenic ascomycete *Leptosphaeria maculans*. Curr. Genet. 30, 541–548.

Lee, D.H., Roux, J., Wingfield, B.D., Wingfield, M.J., 2015. Variation in growth rates and aggressiveness of naturally occurring self-fertile and self-sterile isolates of the wilt pathogen *Ceratocystis albifundus*. Plant Pathol. 64, 1103–1109.

Lee, D.H., Roux, J., Wingfield, B.D., Barnes, I., Mostert, L., Wingfield, M.J., 2016. The genetic landscape of *Ceratocystis albifundus* populations in South Africa reveals a recent fungal introduction event. Fungal Biol. 120, 690–700.

Lee, M.J., Sheppard, D.C., 2016. Recent advances in the understanding of the *Aspergillus fumigatus* cell wall. J. Microbiol. 54, 232–242.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M., Henrissat, B., 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol. Biofuels 6, 1.

Li, F., Shi, H.-Q., Ying, S.-H., Feng, M.-G., 2015. Distinct contributions of one Fe- and two Cu/Zn-cofactored superoxide dismutases to antioxidation, UV tolerance and virulence of *Beauveria bassiana*. Fungal Genet. Biol. 81, 160–171.

Llorens, C., Futami, R., Covelli, L., Dominguez-Escriba, L., Viu, J.M., Tamarit, D., Aguilar-Rodriguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., Maumus, F., 2011. The Gypsy database (GyDB) of mobile genetic elements: Release 2.0. Nucleic Acids Res. 39 (Suppl. 1), D70–D74.

Ma, L.J., Van Der Does, H.C., Borkovich, K.A., Coleman, J.J., Daboussi, M.J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P.M., 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464, 367–373.

Ma, L.J., Geiser, D.M., Proctor, R.H., Rooney, A.P., O Donnell, K., Trail, F., Gardiner, D.M., Manners, J.M., Kazan, K., 2013. *Fusarium* pathogenomics. Annu. Rev. Microbiol. 67, 399–416.

Manning, V.A., Pandelova, I., Dhillon, B., Wilhelm, L.J., Goodwin, S.B., Berlin, A.M., Figueroa, M., Freitag, M., Hane, J.K., Henrissat, B., Holman, W.H., 2013. Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. G3 Genes Genom. Genet. 3, 41–63.

Metsalu, T., Vilo, J., 2015. Clustvis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 43, W566–W570.

Mukherjee, P.K., Horwitz, B.A., Kenerley, C.M., 2012. Secondary metabolism in *Trichoderma* – a genomic perspective. Microbiology 158, 35–45.

Novikova, O., 2009. Chromodomains and LTR retrotransposons in plants. Commun. Integr. Biol. 2, 158–162.

Ohm, R.A., Feau, N., Henrissat, B., Schoch, C.L., Horwitz, B.A., Barry, K.W., Condon, B.J., Copeland, A.C., Dhillon, B., Glaser, F., Hesse, C.N., 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen *Dothideomycetes* fungi. PLoS Pathog. 8, e1003037.

Paris, S., Wysong, D., Debeaupuis, J.-P., Shibuya, K., Philippe, B., Diamond, R.D., Latge, J.-P., 2003. Catalases of *Aspergillus fumigatus*. Infect. Immun. 71, 3551–3562.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786.

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartmann, F.E., Croll, D., 2017. Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. Front. Plant Sci. 8, 119.

Plissonneau, C., Hartmann, F.E., Croll, D., 2018. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol. 16, 5.

Plissonneau, C., Stürchler, A., Croll, D., 2016. The evolution of orphan regions in genomes of a fungal pathogen of wheat. mBio 7 e01231–16.

Quesneville, H., Nouaud, D., Anxolabéhère, D., 2003. Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes. J. Mol. Evol. S50-S59.

Raffaele, S., Kamoun, S., 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat. Rev. Microbiol. 10, 417–430.

Rawlings, N.D., Barrett, A.J., Finn, R.D., 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 44, D343–D350.

Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U.S.A. 106, 19126–19131.

Roux, J., Harrington, T.C., Steimel, J.P., Wingfield, M.J., 2001. Genetic variation in the wattle wilt pathogen *Ceratocystis albifundus*. Mycoscience 42, 327–332.

Roux, J., Heath, R.N., Labuschagne, L., Nkuekam, G.K., Wingfield, M.J., 2007. Occurrence of the wattle wilt pathogen, *Ceratocystis albifundus* on native South African trees. Forerst Pathol. 37, 292–302.

Roux, J., Dunlop, R., Wingfield, M.J., 1999. Susceptibility of elite *Acacia mearnsii* families to Ceratocystis wilt in South Africa. J. Forest Res. 4, 187–190.

Roux, J., Meke, G., Kanyi, B., Mwangi, L., Mbaga, A., Hunter, G.C., Nakabonge, G., Heath, R.N., Wingfield, M.J., 2005. Diseases of plantation forestry trees in eastern and Southern Africa. South Afr. J. Sci. 101, 409–413.

Shi, X., Faino, L., van den Berg, G., Thomma, B., Seidl, M., 2018. Evolution within the fungal genus *Verticillium* is characterized by chromosomal rearrangement and gene loss. Environ. Microbiol. 20, 1362–1373.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.

Sperschneider, J., Gardiner, D.M., Dodds, P.N., Tini, F., Covarelli, L., Singh, K.B., Manners, J.M., Taylor, J.M., 2016. EffectorP: predicting fungal effector proteins from secretomes using machine learning. New Phytol. 210, 743–761.

Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B., 2004. AUGUSTUS: a web server for gene finding in Eukaryotes. Nucleic Acids Res. 32, W309–W312.

Steenkamp, E.T., Wingfield, M.J., McTaggart, A.R., Wingfield, B.D., 2018. Fungal species and their boundaries matter – definitions, mechanisms and practical implications. Fungal Biol. Rev. 32, 104–116.

Stergiopoulos, I., de Wit, P.J., 2009. Fungal effector proteins. Annu. Rev. Phytopathol. 8, 233–263.

Supek, F., Bosnjak, M., Skunca, N., Smuc, T., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6, e21800.

Tanabe, S., Ishii-Minami, N., Saitoh, K.-I., Otake, Y., Kaku, H., Shibuya, N., NishizawaY, Minami, E., 2011. The role of catalase-peroxidase secreted by *Magnaporthe oryzae* during early infection of rice cells. Mol. Plant Microbe Interact. 24, 163–171.

Tzeng, T.H., Lyngholm, L.K., Ford, C.F., Bronson, C.R., 1992. A restriction fragment length polymorphism map and electrophoretic karyotype of the fungal maize pathogen *Cochliobolus heterostrophus*. Genetics 130, 81–96.

van der Nest, M.A., Beirn, L.A., Crouch, J.A., Demers, J.E., De Beer, Z.W., De Vos, L., Gordon, T.R., Moncalvo, J.-M., Naidoo, K., Sanchez-Ramirez, S., Roodt, D., 2014a. Draft genomes of *Amanita jacksonii, Ceratocystis albifundus, Fusarium circinatum, Huntiella omanensis, Leptographium procerum, Rutstroemia sydowiana*, and *Sclerotinia echinophila*. IMA Fungus 5, 473–486.

van der Nest, M.A., Bihon, W., De Vos, L., Naidoo, K., Roodt, D., Rubagotti, E., Slippers, B., Steenkamp, E.T., Wilken, P.M., Wilson, A., Wingfield, M.J., Wingfield, B.D., 2014b. Draft genome sequences of *Diplodia sapinea, Ceratocystis manginecans*, and *Ceratocystis moniliformis*. IMA Fungus 5, 135–140.

van der Nest, M.A., Steenkamp, E.T., McTaggart, A.R., Trollip, C., Godlonton, T., Sauerman, E., Roodt, D., Naidoo, K., Coetzee, M.P.A., Wilken, P.M., 2015. Saprophytic and pathogenic fungi in the Ceratocystidaceae differ in their ability to metabolize plant-derived sucrose. BMC Evol. Biol. 15, 1.

van Wyk, S., Wingfield, B.D., De Vos, L., Santana, Q., Van der Merwe, N., Steenkamp, E.T., 2018. Multiple independent origins for a subtelomeric locus associated with growth rate in *Fusarium circinatum*. IMA Fungus 9, 27–36.

Vanheule, A., Audenaert, K., Warris, S., van de Geest, H., Schijlen, E., Höfte, M., De Saeger, S., Haesaert, G., Waalwijk, C., van der Lee, T., 2016. Living apart together: crosstalk between the core and supernumerary genomes in a fungal plant pathogen. BMC Genomics 17, 670.

Vanrobays, E., Thomas, M., Tatout, C., 2018. Heterochromatin positioning and nuclear architecture. Annu. Rev. Plant Biol. 46, 157–190.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982.

Wilhelm, M., Wilhelm, F.X., 2001. Reverse transcription of retroviruses and LTR retrotransposons. CMLS Cell. Mol. Life Sci. 58, 1246–1262.

Wilken, P.M., Steenkamp, E.T., Wingfield, M.J., De Beer, Z.W., Wingfield, B.D., 2013. *Ceratocystis fimbriata*: draft nuclear genome sequence for the plant pathogen, *Ceratocystis fimbriata*. IMA Fungus 4, 357–358.

Wingfield, B.D., Barnes, I., de Beer, Z.W., De Vos, L., Duong, T.A., Kanzi, A.M., Naidoo, K., Nguyen, H.D., Santana, Q.C., Sayari, M., Seifert, K.A., 2015. IMA Genome-F 5: draft genome sequences of *Ceratocystis eucalypticola, Chrysoporthe cubensis, C. deuterocubensis, Davidsoniella virescens, Fusarium temperatum, Graphilbum fragrans, Penicillium nordicum,* and *Thielaviopsis musarum*. IMA Fungus 6, 493–506.

Wingfield, B.D., Ambler, J.M., Coetzee, M., De Beer, Z.W., Duong, T.A., Joubert, F., Hammerbacher, A., McTaggart, A.R., Naidoo, K., Nguyen, H.D., Ponomareva, E., 2016a. Draft genome sequences of *Armillaria fuscipes, Ceratocystiopsis minuta, Ceratocystis adiposa, Endoconidiophora laricicola, E. polonica* and *Penicillium freii.* IMA Fungus 7, 217–227.

Wingfield, B.D., Duong, T.A., Hammerbacher, A., van der Nest, M.A., Wilson, A., Chang, R., De Beer, W., Steenkamp, E.T., Wilken, M.P., Naidoo, K., Wingfield, M.J., 2016b. Draft genome sequences for *Ceratocystis fagacearum, C. harringtonii, Grosmannia penicillata*, and *Huntiella bhutanensis*. IMA Fungus 7, 317–323.

Winnenburg, R., Urban, M., Beacham, A., Baldwin, T.K., Holland, S., Lindeberg, M., Hansen, H., Rawlings, C., Hammond-Kosack, K.E., Köhler, J., 2008. PHI-base update: additions to the pathogen-host interaction database. Nucleic Acids Res. 36 (Suppl. 1), D572–D576.

Wittenberg, A.H.J., Van der Lee, T.A.J., Schouten, H.J., 2009. Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. PLoS One 4, 1–37.

Yang, Z., Nielson, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19, 908–917.

Yang, Z., Nielson, R., Goldman, N., Pedersen, A., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., Xu, Y., 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40, W445–W451.

Yue, J.X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., Bergström, A., Coupland, P., Warringer, J., Lagomarsino, M.C., Fischer, G., 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat. Genet. 49, 913.

Zerbino, D.R., 2010. Using the Velvet *de novo* assembler for short-read sequencing technologies. Curr. Protoc. Bioinf. 11, 5.1–11.5.12.

Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

Zerillo, M.M., Adhikari, B.N., Hamilton, J.P., Buell, C.R., Levesque, C.A., Tisserat, N., 2013. Carbohydrate-active enzymes in *Pythium* and their role in plant cell wall and storage polysaccharide degradation. PLoS One 8, e72572.

Zhao, Z., Liu, H., Wang, C., Xu, J.R., 2013. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics 14, 1.