Comparative genome analysis reveals adaptation to the ectophytic lifestyle of sooty blotch and flyspeck fungi

Chao Xu[1, 2], Rong Zhang[1], Guangyu Sun[1*], and Mark L. Gleason[3*]

[1]State Key Laboratory of Crop Stress Biology in Arid Areas and College of Plant Protection, Northwest A&F University, Yangling, Shaanxi, China

[2]Department of Plant Pathology, Henan Agricultural University, Zhengzhou, Henan, China

[3]Department of Plant Pathology and Microbiology, Iowa State University, Ames, Iowa, USA

*Co-corresponding authors: Guangyu Sun, State Key Laboratory of Crop Stress Biology in Arid Areas and College of Plant Protection, Northwest A&F University, Yangling, Shaanxi, China, 86-029-87092075, sgy@nwsuaf.edu.cn; Mark L. Gleason, Department of Plant Pathology and Microbiology, Iowa State University, Ames, Iowa, USA, +1 5152940579, mgleason@iastate.edu

Data deposition

De novo genome assemblies of five plant pathogenic fungi have been deposited at NCBI under the following accessions: *Mycosphaerella madeirae* CBS 112895, NHNV00000000; *Zasmidium citri* CBS 116366, NHNW00000000; *Microcyclosporella mali* UMD1a, NHNU00000000; *Zasmidium angulare* GA2-2.7B1a, NHNX00000000; and *Microcyclospora pomicola* SP1-49Fa, NHNY00000000.

Abstract

Sooty blotch and flyspeck (SBFS) fungi are a distinctive group of plant pathogens which, although phylogenetically diverse, occupy an exclusively surface-dwelling niche. They cause economic losses by superficially blemishing the fruit of several tree crops, principally apple, in moist temperate regions worldwide. In this study, we performed genome-wide comparative analyses separately within three pairs of species of ascomycete pathogens; each pair contained an SBFS species as well as a closely related but plant-penetrating parasite (PPP) species. Our results showed that all three of the SBFS pathogens had significantly smaller genome sizes, gene numbers and repeat ratios than their counterpart PPPs. The pathogenicity-related genes encoding MFS transporters, secreted proteins (mainly effectors and peptidases), plant cell wall degrading enzymes and secondary metabolism enzymes were also drastically reduced in the SBFS fungi compared with their PPP relatives. We hypothesize that the above differences in genome composition are due largely to different levels of acquisition, loss, expansion and contraction of gene families and emergence of orphan genes. Furthermore, results suggested that horizontal gene transfer may have played a role, although limited, in the divergent evolutionary paths of SBFS pathogens and PPPs; repeat-induced point mutation could have inhibited the propagation of transposable elements and expansion of gene families in the SBFS group, given that this mechanism is stronger in the SBFS fungi than in their PPP relatives. These results substantially broaden understanding of evolutionary mechanisms of adaptation of fungi to the epicuticular niche of plants.

Key words: Comparative genomics, genome assembly, gene family evolution, host adaptation, plant pathogen

Introduction

Living plants are home to a large variety of fungal species with diverse lifestyles, including mutualism, commensalism and parasitism (Redman et al. 2001). Among them, the plant-pathogenic fungi attract special concern because their economic costs to agricultural producers can be extremely high. On apples and some other fruit trees, a specialized group of plant pathogens, collectively called the sooty blotch and flyspeck (SBFS) complex, colonize the epicuticular wax layer and form darkly pigmented mycelial mats and fruiting bodies (Belding et al. 2000; Williamson and Sutton 2000; Xu et al. 2016). In spite of causing no cell damage due to the absence of cell wall penetration, these pathogens can still inflict substantial economic losses by blemishing fruit, which leads to downgrading of their fresh-market value (Díaz Arias et al. 2010).

Over 80 species are included in the SBFS group, most of which come from several genera of the order Capnodiales (Dothideomycetes, Ascomycota) and occur worldwide on the fruit, leaves, twigs and stems of numerous cash crops and native plants (Gleason et al. 2011). Unlike classical plant pathogens that invade living host cells and then absorb nutrients, SBFS fungi just attach to plant surfaces and subsist primarily on tissue leachates (Williamson and Sutton 2000). Their unique ability to partially penetrate but not breach the cuticle describes a niche called "ectophytic parasitism" that differs from both exclusively surface-dwelling and plant-penetrating microbes (Xu et al. 2016). Although SBFS fungi have been studied for nearly 200 years (Williamson and Sutton 2000), their evolutionary origins and physiological adaptations to environments remain largely unexplored. Encouragingly, a recent study used ancestral state reconstruction analysis to show that the major SBFS lineages appear to have evolved from ancestral fungi that were classic cell-penetrating plant parasites (Ismail et al. 2016). This assertion was largely supported by our subsequent work indicating that *Peltaster fructicola*, a widely prevalent SBFS species, had undergone reductive evolution (i.e., genome contraction) accompanied by the loss of many pathogenicity-related genes (Xu et al. 2016). However, it remains to be determined whether this evolutionary pattern is generalizable across other SBFS taxa.

To close this gap in understanding of SBFS evolution, we selected three pairs of species in the order Capnodiales (*Mycosphaerella madeirae* and *Microcyclosporella mali*; *Zasmidium citri* and *Zasmidium angulare*; *Teratosphaeria nubilosa* and *Microcyclospora pomicola*), each containing a species known to be a plant-penetrating parasite (PPP) as well as an SBFS species that was as closely related as possible to the paired PPP (Frank et al. 2010; Ismail et al. 2016; Li et al. 2012). Both *Mycosphaerella madeirae* (Mycma) and *T. nubilosa* (Ternu) cause necrotic leaf spots on *Eucalyptus* foliage (Aguin et al. 2013; Hunter et al. 2009), and *Zasmidium citri* (Zasci) causes leaf and fruit lesions on most citrus and related hosts (Mondal and Timmer 2006). In contrast, *Microcyclosporella mali* (Micma), *Zasmidium angulare* (Zasan) and *Microcyclospora pomicola* (Micpo) are SBFS species that colonize the surfaces of apples (Frank et al. 2010; Li et al. 2012). The genomic data of Ternu is available in the JGI fungal genome portal MycoCosm (Nordberg et al. 2014), whereas draft assemblies of all the others were novel to this paper.

After an organism is completely sequenced, its genome size normally becomes the first character to be of concern, especially for the SBFS fungi in which there is a high probability of occurrence of relatively small genomes (e.g., *P. fructicola*) (Xu et al. 2016). Although smaller in genome size than most other eukaryotes, fungi vary approximately from 10 to 900 Mb and average 37.7 Mb overall (Tavares et al. 2014). Such variations are often viewed as adaptive, since changes (expansion or contraction) in genome size of fungi could have an impact on their parasitism or pathogenicity (D'hondt et al. 2011; Kelkar et al 2012). For example, the smallest fungal genomes, which have lost most mobile elements and retain only the genes involved in basic metabolisms, usually come from free-living saprophytes in the family

Saccharomycetaceae (Dujon 1996), whereas the largest genomes (Pucciniales) with abundant transposable elements and protein-coding genes are obligately parasitic and pathogenic (Tavares et al. 2014). In other words, the genome size of an organism could depend largely on its particular developmental and ecological need (Petrov 2001). However, it is still not clear how the genomes of fungi in the SBFS complex fit into their unique ectophytic niche and which evolutionary forces help to alter their genome sizes.

In this study, we described genome contents of the five newly sequenced species and performed genome-wide comparative analyses within three separate monophyletic clades formed by them as well as Ternu. Through examining the orphaned genes and gene families that had undergone expansion/contraction and gain/loss events, we found that, compared with their PPP counterparts, significantly more orphan genes and gene families involved in pathogen-host interactions were lost and contracted during the evolutionary transition to SBFS species with an ectophytic lifestyle. In addition, recent horizontal gene transfer (HGT) and repeat-induced point mutation (RIP), both of which have been demonstrated to play a central role in the evolution of several ascomycetous genomes (Hane et al. 2015; Nguyen et al. 2015), were identified in all the above species. From the perspective of genomic evolution, our results provide clues for unveiling adaptive mechanisms in the SBFS complex and other surface-dwelling fungi.

Results

Genome sequencing and general features

The overall assemblies (ranging from 23.5 Mb to 45.0 Mb) of Mycma, Micma, Zasci, Zasan and Micpo were sequenced at a minimum of 54-fold coverage and a maximum of 157-fold coverage using Illumina HiSeq 2500 technology. Theoretical sizes of their genomes were estimated to be 24.3–46.9 Mb according to the k-mer analysis, indicating that most of the genomes are included in the above assemblies. More detailed parameters of their assemblies and genes, such as statistics of the scaffolds, GC contents, repeat ratios and protein-coding genes, are displayed in table 1. As for Ternu, 7.88% of its published 28.4 Mb assembly with 10,998 predicted protein-coding genes were identified as repetitive sequences in this study. The three SBFS species (Micma, Zasan and Micpo) were inferior to their respective PPP relatives (Mycma, Zasci and Ternu) in genome size (28.0 Mb vs. 33.7 Mb; 37.9 Mb vs. 45.0 Mb; 23.5 Mb vs. 28.4 Mb), number of genes (11,572 vs. 14,017; 14,946 vs. 17,275; 9,169 vs. 10,998), and repeat ratio (1.57% vs. 6.09%; 3.60% vs. 8.35%; 1.95% vs. 7.88%).

In functional annotation, 82–89% of the proteomes of the above six fungi showed sequence similarities (e-value <1e-05) to entries deposited in the NR database of NCBI; 8–11% were mapped in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database; 44–55% were sorted using the EuKaryotic Orthologous Groups (KOG) classification; and 57–68% were assigned to the Gene Ontology (GO) terms (table 1 and supplementary figure S1). The distributions of proteins into different KOG categories are similar between the two species in each pair, except that all three SBFS pathogens are relatively deficient (gap of >30 genes per species) in four specific function classes: amino acid transport and metabolism (E); carbohydrate transport and metabolism (G); secondary metabolite biosynthesis, transport and catabolism (Q); and general function prediction only (R) (supplementary figure S2).

Homolog clustering and phylogeny

Identification of homolog groups was performed using 237,587 protein-coding genes from 20 fungal species. In total, 201,135 proteins built 20,388 protein/gene families (each containing at least two members), whereas the remaining 36,452 unclassified proteins were regarded as originating from orphan genes (lacking homology to other sequences in the dataset). The amino acid sequences of 3,718 single-copy orthologous groups were used to construct a genome-based maximum-likelihood (ML) tree in which

Aureobasidium pullulans, the only non-Capnodiales species, was treated as the outgroup (Figure 1A). The topology of the phylogenetic tree was strongly supported by 100% bootstrap values on all its branches. According to this phylogram, Mycma and Micma adjoined each other and their divergence time was estimated to be ~15.6 million years ago (Mya); Zasci and Zasan split ~21.8 Mya and clustered together with a saprophytic fungus, *Zasmidium cellare*; Ternu, Micpo and another saprophyte, *Baudoinia compniacensis*, collectively formed a monophyletic group that descended from a common ancestor ~73.3 Mya. The six investigated species were segregated into three separate clades, of which Mycma and Micma were the two most closely related species, followed by Zasci and Zasan, and finally, Ternu and Micpo.

To further explore the relationships between the two species of each pair, interspecific intersections of homologous protein groups were calculated (Figure 1B). We found that 9,130 homologous protein groups were shared by Mycma and Micma, which accounted for 87% of the gene families of the former (10,488) and 94% of the latter (9,699), respectively. A total of 11,343 common protein groups occupied 88% of the gene families of Zasci (12,934) and 92% of Zasan (12,372). Furthermore, the intersection set containing 6,505 homologous protein groups was responsible for 80% of the gene families of Ternu (8,148) and 86% of Micpo (7,599). These results broadly met our expectation that the closer the relatedness between two species of one pair, the higher the proportion of gene families they will share. In view of the above evidence, we also speculated that the genome structures of the three pairs of closely related species would have different levels of synteny. By anchoring proteins of one genome to their homologs in the genome of the other species in the pair, 612 collinear blocks with five or more contiguous genes were detected between Mycma and Micma; 744 collinear blocks were detected between Zasci and Zasan; and 145 collinear blocks were detected between Ternu and Micpo (supplementary figure S3). Despite extensive collinearity for the first two pairs, distinct genomic rearrangements probably arose due to the accumulation of a series of mutational events (deletions, duplications, insertions, inversions or translocations) over evolutionary time (Fierro and Martin 1999).

Deficiency of genes associated with host-penetrating parasitism in SBFS pathogens

Transporters are a diverse group of transmembrane proteins which transfer ions, small molecules and macromolecules from source to sink, resulting in cellular uptake and extrusion of compounds (Saier et al. 2016). Each of our investigated genomes encoded a large number of transporters belonging to at least 49 superfamilies of the Transporter Classification (TC) system (supplementary table S1). Nevertheless, the SBFS species had considerably fewer such proteins than their respective PPP relatives (591 in Micma vs. 724 in Mycma; 873 in Zasan vs. 982 in Zasci; 432 in Micpo vs. 523 in Ternu). Most (52–78%) of these quantitative differences were caused by the major facilitator superfamily (MFS) (296 in Micma vs. 396 in Mycma; 511 in Zasan vs. 596 in Zasci; 195 in Micpo vs. 242 in Ternu) (supplementary table S1 and Figure 2A). It has been reported that MFS as well as ATP-binding cassette (ABC) transporters were associated with host-penetrating parasitism by conferring protection against plant defense compounds and fungitoxic antibiotics (Hayashi et al. 2002). This linkage was also confirmed by our finding that an overwhelming majority (>87%) of the predicted MFS transporters have similarities to the experimentally verified proteins catalogued in the pathogen-host interaction (PHI) database. However, we did not observe any notable reduction in the ABC transporters (including those with PHI homologs) of the SBFS pathogens relative to their paired plant-penetrating pathogenic species (supplementary table S1).

PPP fungi secrete numerous proteins (especially peptidases and effectors) during colonization. *In silico* analysis identified 633 secreted proteins in Mycma, 1,142 in Zasci and 484 in Ternu, which surpassed their respective SBFS relatives Micma (488), Zasan (887) and Micpo (339) (Figure 2B). Similarly, the amount of secreted peptidases in each PPP was slightly larger than that in its SBFS relative. However, in

only one (S10) of the 24 predicted MEROPS subfamilies did all three PPPs surpass their SBFS counterparts (13 in Mycma vs. 8 in Micma; 17 in Zasci vs. 9 in Zasan; 11 in Ternu vs. 9 in Micpo) (supplementary table S2). Serine carboxypeptidases of the S10 subfamily are functionally critical to PPPs because they are expected to work in inhospitable extracellular environments and digest defense-related proteins from host plants (Adhikari et al. 2013). In addition, acting as the virulence/avirulence factors that facilitate infection or trigger plant immune responses, candidate secreted effector proteins (CSEPs) were screened out of the secretomes using 200 amino acids as the upper limit for protein size. Their counts were estimated to be 142 in Mycma vs. 77 in Micma, 269 in Zasci vs. 189 in Zasan, and 188 in Ternu vs. 60 in Micpo, which is evidence for a shortage of CSEPs for the SBFS fungi relative to the PPP fungi. Of the predicted CSEPs, an average of 92.1% had no PFAM domains, which exceeds the 46.3% for secretomes and 39.8% for proteomes, reflecting the fact that the functions of effectors are frequently unknown. Furthermore, the average percentage (3.8%) of cysteine residues in the CSEPs was higher than in the other proteins (2.1% for secretomes and 1.4% for proteomes), according with the fact that effectors are commonly cysteine-rich (supplementary figure S4).

Celluloses, hemicelluloses, pectins and cutins are needed to build epidermal cell walls, the protective barriers of plants against diverse pathogen attacks. Accordingly, plant cell wall degrading enzymes (PCWDEs) can be divided into five categories based on substrate specificity, i.e., the enzymes that break down both celluloses and hemicelluloses, both pectins and hemicelluloses, and hemicelluloses, pectins, or cutins alone. Our research showed that the three SBFS species lagged their PPP relatives not only in the overall count of PCWDEs (75 in Micma vs. 98 in Mycma; 140 in Zasan vs. 160 in Zasci; 34 in Micpo vs. 75 in Ternu) but also in the number of each individual PCWDE class except cutinases (Figure 2C). In addition, for all six species, approx. 36 to 49% of their PCWDEs were not included in the above secretomes, suggesting that these proteins may just participate in intracellular metabolic processes or are more likely to be secreted by unconventional mechanisms (Nickel and Seedorf 2008).

Secondary metabolites (SMs) are crucial small molecules (including mycotoxins, antibiotics and pharmaceuticals) that typically facilitate the adaptation of various fungi to distinct habitats. Their biosynthesis is controlled mainly by five kinds of core enzymes that catalyze the first committed steps of corresponding metabolic pathways: polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), hybrid PKS-NRPSs, terpene cyclases (TCs) and dimethylallyl tryptophan synthases (DMATSs). We found 47 core SM biosynthetic enzymes in Mycma (29 PKSs, 13 NRPSs, 2 hybrid PKS-NRPSs, 1 DMATS and 2 TCs) versus 29 in Micma (15 PKSs, 12 NRPSs, 0 hybrid PKS-NRPS, 0 DMATS and 2 TCs), 49 in Zasci (25 PKSs, 19 NRPSs, 1 hybrid PKS-NRPS, 1 DMATS and 3 TCs) versus 34 in Zasan (17 PKSs, 15 NRPSs, 0 hybrid PKS-NRPS, 0 DMATS and 2 TCs), and 23 in Ternu (9 PKSs, 10 NRPSs, 1 hybrid PKS-NRPS, 0 DMATS and 3 TCs) versus 19 in Micpo (7 PKSs, 10 NRPSs, 1 hybrid PKS-NRPS, 0 DMATS and 1 TC) (Figure 2D), which demonstrates the numerical disadvantage of the three SBFS species versus their respective PPP relatives in the number of core SM genes. The high proportion (>83%) of core genes encoding PKSs and NRPSs in their genomes accords with the fact that fungi, in contrast to plants, produce mostly polyketides and non-ribosomal peptides (Collemare et al. 2014). Moreover, 100% of the PKSs, NRPSs and hybrid PKS-NRPSs have homologs in the PHI database, although the DMATSs and TCs could not be matched to any PHI-base proteins. These results support the view that fungal SMs facilitate invasion by PPPs.

Evolution of gene families

During evolution, the genetic composition of an organism is generally shaped by five types of events: acquisition, loss, expansion and contraction of gene families, and emergence of orphan genes (Hahn et al.

2007; Tautz and Domazet-Loso 2011). Since their most recent common ancestor (MRCA) bifurcated at node 5, Mycma has acquired 52 gene families, lost 561 families, expanded 378 families, and contracted 19 gene families with 2,612 orphan genes emerging; in contrast, Micma has gained 8, lost 1,306, expanded 98, contracted 48 gene families and produced 1,444 orphan genes (Figure 3A). From their MRCA node 2, Zasci (a PPP) experienced 105 acquisition events compared to 416 for the SBFS fungus Zasan, losses were 1,061 and 1,934 respectively, expansions totaled 590 and 287 respectively, contraction events were 41 and 181, respectively, and Zasci totaled 2,783 orphan genes compared to 1,664 for Zasan. Furthermore, as Ternu and Micpo evolved from their MRCA node 15, a total of 69 gene families were acquired, 1,758 were lost, 237 were expanded, 38 were contracted, and 2,394 orphan genes were detected in Ternu, whereas 38 gene families were acquired, 2,276 were lost, 156 were expanded, 48 were contracted, and 1,188 orphan genes were identified for Micpo.

In this study, we focused on four categories of pathogenicity-related genes encoding MFS & ABC transporters, secreted proteins, PCWDEs and core SM enzymes, and regarded them and all the other proteins having PHI-base homologs as the factors that can affect outcomes of PHIs. When using this specific definition to annotate the orphan genes and the gene families undergoing acquisition, loss, expansion and contraction, we found that many of them may play roles in the interactions between pathogens and host plants. For Mycma, 17 acquired, 115 lost, 156 expanded and 11 contracted gene families and 317 orphan genes were flagged as the factors involved in PHIs; for Micma, 1 acquired, 447 lost, 41 expanded and 28 contracted gene families and 107 orphan genes were flagged; for Zasci, 18 acquired, 328 lost, 206 expanded and 30 contracted gene families and 398 orphan genes were flagged; for Zasan, 128 acquired, 584 lost, 135 expanded and 109 contracted gene families and 215 orphan genes were flagged; for Ternu, 23 acquired, 596 lost, 120 expanded and 28 contracted gene families and 253 orphan genes were flagged; for Micpo, 5 acquired, 811 lost, 69 expanded and 28 contracted gene families and 101 orphan genes were flagged (Figure 3B). Clearly, the loss and contraction of the gene families with PHI flags could blunt aggressivity of pathogens, whereas the other three evolutionary events could contribute to enhancing this capability.

Horizontal gene transfer

Horizontal gene transfer (HGT) allows gene exchange among genetically unrelated organisms, which in fungi has been correlated with the gain of virulence and adaptive traits (Mehrabi et al. 2011). To explore whether this lateral transmission contributed to the different niche adaptations between SBFS pathogens and their PPP relatives, we identified HGT candidates in their full sets of species-specific genes (i.e., the genes from newly acquired families plus orphan genes). In total, three HGTs were detected for Mycma, one for Micma, three for Zasci, three for Zasan, three for Ternu, and one for Micpo, of which over 70% (ten) were derived from bacterial sources (supplementary table S3). Most of these HGT genes can code only extremely small proteins ($\leqslant$150 aa), much shorter than their BLAST hits in the nr database. Probably because of this, 79% (11) of the HGTs had no functional annotation (PFAM domain), of which one was predicted to be a CSEP in Ternu, and none of them possessed homologs in the PHI-base. The only three HGTs with PFAM domains were annotated as an L-PSP endoribonuclease in Zasci, a UbiA prenyltransferase and a glutathione S-transferase (probably degraded) in Zasan.

Repeat-induced point mutations

Compared with their PPP relatives, the SBFS pathogens harbored significantly fewer repetitive sequences, especially interspersed repeats including transposable elements (TEs) and unclassified duplications (supplementary table S4). As repeat-induced point mutation (RIP) is a fungal-specific genome defense mechanism against transposons by rendering them mutational and eventually inactive (Hane and Oliver

2008), we performed dinucleotide frequency analyses of the identified TE families to test whether a RIP-like mechanism existed in the investigated species and had contributed to the relatively low percentages of repetitive elements in the SBFS fungi. In Micma, the CpA and TpG dinucleotide RIP-targets were largely depleted, and in the RIP dinucleotide products only TpA showed a corresponding increase (Figure 4A). This suggests that CpA to TpA (or TpG to TpA in the complementary strand) is the dominant form of CpN toTpN dinucleotide mutation in Micma, as observed in *Neurospora crassa* (Galagan et al. 2003). However, the absence of a marked decrease or increase of any dinucleotide frequencies in Mycma indicates absence of the RIP process. A similar pattern of change in dinucleotide abundance of Micma is shared by both Zasci and Zasan, which proves that they also have the CpA dinucleotide mutational bias (Figure 4B). As for Ternu and Micpo, besides CpA and TpG, several more dinucleotides such as CpC and CpG declined considerably in frequency. Nevertheless, it is clear that both of these species have undergone RIP with CpA as the predominant mutation site, because TpA was generated preferentially over all the other dinucleotides (Figure 4C).

The above results are corroborated by analyses of two different RIP indices. The TpA/ApT index measures the RIP products, TpA with correction for false positives due to A:T rich regions (a higher value implies a stronger RIP response); the other RIP index, (CpA+TpG)/(ApC+GpT), is similar in principle to TpA/ApT but estimates the depletion of the RIP targets CpA and TpG (a lower value is indicative of a stronger RIP) (Hane and Oliver 2008). The TpA/ApT index was calculated to be 0.80 in Mycma ($\geqslant$0.61 of the corresponding non-repetitive control sequences, indicating RIP mutations), 1.93 in Micma ($\geqslant$0.59), 2.12 in Zasci ($\geqslant$0.88), 2.07 in Zasan ($\geqslant$0.79), 1.60 in Ternu ($\geqslant$0.68), and 1.71 in Micpo ($\geqslant$0.63); the (CpA+TpG)/(ApC+GpT) index was calculated to be 1.30 in Mycma ($\leqslant$1.26 of the corresponding non-repetitive control sequences, indicating RIP mutations), 0.11 in Micma ($\leqslant$1.29), 0.06 in Zasci ($\leqslant$1.10), 0.06 in Zasan ($\leqslant$1.16), 0.41 in Ternu ($\leqslant$1.28), and 0.40 in Micpo ($\leqslant$1.24). Except Mycma, RIP indices for TpA/ApT are two times more than the control levels whereas the (CpA+TpG)/(ApC+GpT) indices are far below the control levels, indicating that RIP is lacking for Mycma but exists in all the other five fungi. Moreover, a ratio of the TpA/ApT index to that of the corresponding non-repetitive control was calculated for Zasci (2.42), Zasan (2.63), Ternu (2.34) and Micpo (2.72), respectively (Figure 4). This might mean that the two SBFS pathogens, Zasan and Micpo, possess slightly stronger RIP activity than their PPP relatives.

The only enzyme currently known to be essential for RIP is a DNA methyltransferase (DMT), encoded by the *rid* (RIP-defective) gene first identified in *N. crassa* (Freitag et al. 2002). The genomes of all six investigated species contain rid homologs, which share approx. 38% (Mycma), 38% (Micma), 41% (Zasci), 35% (Zasan), 43% (Ternu) and 33% (Micpo) identities with the *N. crassa* rid (accession no. AF500227). All of the ten characteristic DMT motifs shown by Freitag et al. (2002) were also found in the putative rid protein sequences (Figure 5). Apparently, presence of a *rid* gene greatly enhances the possibility for an effective RIP machinery; in Mycma, however, the presence of a *rid* homolog alone does not confer RIP activity.

Discussion

In this study, we conducted comparative genome analyses of three pairs of related fungal species (all of them were first sequenced here except Ternu), each of which includes an SBFS pathogen and a closely related PPP. The genetic relationships between these paired fungi were shown to be much closer than that between our previous research objects, i.e., the SBFS pathogen *P. fructicola* and its PPP relative *Zymoseptoria tritici*, whose MRCA event dates back to over 100 Mya (Xu et al. 2016). Given that the major SBFS lineages evolved from plant-parasitic ancestors (Ismail et al. 2016), it is illuminating to

understand how their genomic divergence reflected divergent strategies developed to deal with host plants. Compared with their related PPPs, without exception, these three SBFS pathogens have relatively smaller genome sizes, due primarily to their lower repeat contents and smaller numbers of genes; and all three are largely deficient in the protein-coding genes for synthesis, transport and metabolism of several biomolecules including amino acids, carbohydrates and secondary metabolites, which are partly associated with the host-penetrating parasitism. These deficiencies were also observed in the genome of *P. fructicola* (Xu et al. 2016), suggesting environment-adaptive convergent evolution. That is to say, SBFS fungi that generally dwell in the harsh conditions (e.g., oligotrophy and dehydration) associated with plant epicuticular wax layers tend to possess genomes of austerity (Kelley et al. 2014), adapted for minimizing energy expenditures.

In many plant-pathogen interactions, fungal-infected plants are either kept alive to ensure a prolonged supply of various nutritional substances for the pathogens (biotrophy) or destroyed and subsequently fed on by the pathogens (necrotrophy). Some other pathogens (hemibiotrophs) start with a biotrophic phase but switch to necrotrophic behaviors at later infection stages. Collectively, these categories comprise nearly all known niches of plant pathogenic fungi, and are referred to here as PPPs. Despite some resemblance to biotrophs, however, SBFS pathogens (recently termed ectophytic fungi) decidedly do not belong to any of the previously described PPP niches because they can penetrate only part of the cuticle and never invade living plant cells (Xu et al. 2016). It is therefore unsurprising that in our survey all of the SBFS pathogens were predicted to have dramatically fewer pathogenicity-related genes (encoding MFS transporters, secreted proteins, PCWDEs and core SM enzymes) than their PPP relatives. Moreover, as plants develop diverse defense mechanisms to resist different PPPs (e.g., hypersensitive cell death against biotrophic fungi), a major challenge for pathogens is how to avoid being recognized by their hosts and then triggering corresponding immune responses (Glazebrook 2005). SBFS fungi may evade such counterattacks from host plants because they produce very limited pathogenic factors that could serve as elicitors, and even these have no direct contact with host cells. That is probably why these ectophytic organisms manage to survive in the absence of the more abundant nutrients available to PPPs.

The fact that SBFS fungi possess significantly fewer protein-coding genes than their PPP relatives appears to be, to a large extent, responsible for their relative shrinkage of genome content. Our results suggest that this difference in gene quantity was caused by a series of evolutionary events including the acquisition, loss, expansion and contraction of gene families and the emergence of orphan genes that occurred after their reproductive isolation. Subtracting losses from acquisitions (-509 for Mycma, -1,298 for Micma, -956 for Zasci, -1,518 for Zasan, -1,689 for Ternu, and -2,238 for Micpo), we found that each investigated SBFS species eventually experienced a larger decrease in the number of gene families relative to their PPP near-relatives . Second, each species has undergone many more expansions than contractions; but when comparing an SBFS pathogen with its PPP relative, the former has always undergone fewer expansions and more contractions. This means that although all species have gained genes, the SBFS pathogens gained fewer. Last, the PPPs have many more orphan genes that possibly allow organisms to adapt to constantly changing ecological conditions (Tautz and Domazet-Loso 2011). In essence, gene family acquisition and expansion and emergence of orphan genes are all ascribed to the creation of new genes, which can result from several factors, including 1) gene fusion and *de novo* origination (Long et al. 2003), 2) accelerated nucleotide divergence of a new duplicate or member of a previously existing gene family (Copley et al. 2003), and 3) horizontal gene transfer (HGT) (Nguyen et al. 2015). In contrast, both loss and contraction of gene families result from the loss of pre-existing genes, which is mainly due to 1) deletion or pseudogenization and 2) the rapid evolution of protein sequences, such that the genes are no

longer identified as belonging to the same family (Albalat and Canestro 2016). To sum up, the SBFS pathogens experienced more gene loss but less gene creation over the course of evolution, which ultimately led to their smaller proteomes. The above genomic mechanisms have been theorized to account for the adaptations to parasitic growth in fungi (Meerupati et al. 2013). First, the formation of novel genes, which could have specific roles during host infection, is likely indicative of a gain in pathogenicity. Second, weakened parasitism could result from gene loss. This is evident for the SBFS pathogens and PPPs considered here, by seeking out the genes that relate to PHIs from their orphan genes and gene families undergoing acquisition, loss, expansion, and contraction. We found that compared with their PPP relatives, the SBFS pathogens acquired fewer gene families (except Zasan), lost more gene families, expanded fewer gene families, contracted more gene families (except Micpo) and acquired fewer orphan genes, which were predicted to be involved in PHIs. That is, the SBFS pathogens have lost more PHI-related genes but supplemented fewer such functional elements, which eventually led to their relative lack of pathogenic factors.

Although HGT was previously thought to be more common as an important evolutionary force in prokaryotes, it has recently been described in an increasing number of eukaryotes (Nguyen et al. 2015). Many horizontally transferred genes in fungi were found to be related to the metabolism of sugars, nitrogen and amino acids as well as the protein secretion and secondary metabolism, which could help adapt to changing environments (Richards et al. 2011). In this study, we focused on the HGT events that occurred after the separation of SBFS pathogens and their PPP relatives, and therefore just searched the newly acquired genes. In effect for each species, only few HGT candidates were identified, and they had no significant difference in quantity. Of all the 14 HGT genes, one (in Ternu) was identified to code a CSEP potentially involved in the PHI and one (in Zasci) was annotated as an L-PSP endoribonuclease that was reported to be related with increased survival in the parasite *Leishmania infantum* (Pires et al. 2014) and pathogenicity in other fungi (Huser et al. 2009). Based on above mentioned, we hypothesize that HGT could make a limited contribution to the infectivity of the PPPs we studied, but not enhance the functional repertoire of related SBFS fungi.

The SBFS pathogens contain fewer transposable elements (TEs) than their PPP relatives, which might be due to the differences in repeat-induced point mutation (RIP). The main outcome of RIP is to induce transition mutations (C↔T or G↔A) in repeated sequences, which in fact are not random with particular CpN dinucleotides (e.g., CpA and CpG) being preferentially altered over others (Hane and Oliver 2008). RIP was absent in the PPP Mycma because its genome did not show any form of dominant transitions from C:G to T:A, despite the retention of a homolog of the *N. crassa rid* gene, the only gene known to be required for RIP (Freitag et al. 2002). Though all the other five fungi exhibit a clear dinucleotide bias towards CpA-to-TpA changes, both of the remaining two PPPs (Zasci and Ternu) presented relatively weaker RIP activity. An efficient RIP mechanism was predicted to inhibit creating new genes through duplication, and thereby massive expansion of gene families may become rather difficult (Meerupati et al. 2013). This should help to explain why the SBFS pathogens had fewer expanded gene families than their PPP relatives. RIP is often considered to operate only in successive sexual cycles until the sequence identity between pairs of repeated sequences could no longer be recognized (Meerupati et al. 2013). Of the above six fungi, however, only the PPPs have been reported to reproduce sexually (Crous et al. 2004; Hunter et al. 2009; Mondal and Timmer 2006), whereas all the SBFS pathogens were observed to propagate only in an asexual manner (Frank et al. 2010; Li et al. 2012). There are several possible explanations for this conflict: first, RIP may have taken place in a sexual ancestor that subsequently became asexual; second, the "RIP-like" mutations that occur in asexual fungi may be actually caused by other mechanisms producing

similar results; and third, some species, previously thought to be asexual, undergo sexual recombination that may be cryptic or alternatively do so only under very special conditions (Braumann et al. 2008). It is premature to conclude which of these mechanisms may have operated in the SBFS fungi, because all three SBFS species we studied have been identified only recently and further research on them is needed. Moreover, a single mating type (MAT) idiomorph (either *MAT1-1* or *MAT1-2*) has been detected in each of the SBFS pathogens (unpublished data), suggesting that they are potentially heterothallic. More studies are required to understand the importance of the RIP mechanism in the evolution of SBFS fungal genome as well as to determine the full range of their reproductive modes.

Materials and Methods

Fungal species and DNA preparation

Five fungal species were selected for *de novo* genome sequencing: *Mycosphaerella madeirae* (CBS 112895) and *Z. citri* (CBS 116366) accessed from the CBS Fungal Biodiversity Centre, the Netherlands; and *Microcyclosporella mali* (UMD1a), *Z. angulare* (GA2-2.7B1a), and *Microcyclospora pomicola* (SP1-49Fa) archived in the Gleason Laboratory of the Department of Plant Pathology and Microbiology, Iowa State University, USA. Prior to any formal experiments, single spore cultures were prepared for each species to ensure homozygosity, and then maintained on potato dextrose agar (PDA) plates at 22 °C. Highly purified genomic DNA was extracted from the fungal mycelia using the Wizard® Genomic DNA Purification Kit (Promega) according to the manufacturer's instructions. Quality and quantity of the DNA were evaluated using standard 1% agarose gel electrophoresis, as well as spectrophotometrically with NanoDrop 2000 (Thermo Fisher Scientific, USA).

Genome sequencing and assembly

Paired-end 150 bp sequencing of five genomic DNA libraries (500 bp inserts) was performed on the Illumina HiSeq2500 platform (rapid mode) at the DNA Facility of Iowa State University, resulting in raw overlapping forward and reverse reads. The reads containing primers/adapters or ambiguous bases (non-ATCG) and low-quality reads with over 30% poor quality bases (score below 20) were removed using two Perl scripts (IlluQC.pl and AmbiguityFiltering.pl) included in the NGS QC Tookit v2.3.3 package (Patel and Jain 2012). The software Musket was used for error correction in the remaining reads (Liu et al. 2013), and the tools FastUniq (for paired reads) and fastx_collapser (for unpaired reads) (http://hannonlab.cshl.edu/fastx_toolkit/) were used to remove duplicates introduced mainly by PCR amplification (Xu et al. 2012). These filtered clean reads were assembled using both the ABySS assembler version 1.9.0 (Simpson et al. 2009) and the SSPACE-Standard version 3.0 (Boetzer et al. 2011). Theoretical genome sizes could be estimated by calculating the 21-kmer multiplicity with JELLYFISH version 2.0 (Marcais and Kingsford 2011). Collinear blocks of pairs of relative genomes were detected by MCScanX to identify putative homologous chromosomal regions and display the chromosomal rearrangement (Wang et al. 2012).

Repeats and repeat-induced point mutation (RIP) analysis

Gaps within the draft assemblies were closed by GapFiller using the distance information of paired-read raw data (Boetzer and Pirovano 2012). Repetitive elements in the genomes were then identified using RepeatMasker with both the latest Repbase fungal library (http://www.girinst.org/) and *ab initio* libraries generated by RepeatModeler (Tarailo-Graovac and Chen 2009). Of the recognized transposable elements (TEs) including transposons and retrotransposons, only those over 400 bp in length were used for the RIP analysis. RIP indices were determined with the software RIPCAL by reference against the non-repetitive control families (Hane and Oliver 2008). Identification of DNA methyltransferase-like (DMT) sequences

were performed using the BLASTp program with the *Neurospora crassa* RID (AF500227) protein as the query against six genome databases (one for each species) (Braumann et al. 2008).

Gene prediction and annotation

Modified assemblies containing only the scaffolds larger than 200 bp were used as inputs for gene model prediction and other subsequent analyses. Gene calling was performed using the MAKER2 pipeline (Holt and Yandell 2011), which combines three different *ab initio* predictors: GeneMark-ES (Ter-Hovhannisyan et al. 2008), Augustus (Stanke et al. 2006), and SNAP (Korf 2004). The program GeneMark-ES utilized unsupervised training while Augustus and SNAP were trained using two iterative runs of MAKER2. In addition, proteomes of all sequenced Capnodiales fungi downloaded from the JGI website were mapped to our genomes using Exonerate to help improve the gene structural annotations.

Predicted gene models were functionally annotated using an all-in-one bioinformatics software suite Blast2GO, based on the Gene Ontology vocabulary (Gotz et al. 2008). This high-throughput functional annotation pipeline integrates various annotation strategies and tools, including BLASTp alignment against GenBank with non-redundant set (Altschul et al. 1990), GO (Gene Ontology) terms mapping (Ashberner 2000), InterProScan (Jones et al. 2014), Enzyme Commission (EC) numbers assigning (Bairoch 2000), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2004). Predicted protein sequences were also mapped to the KOG classification system (Tatusov et al. 2003). Transfer RNAs (tRNAs) were predicted with tRNAscan-SE (Lowe and Eddy 1997).

Annotation of specific gene categories

Secretomes were identified using a custom pipeline (Xu et al. 2016), combining SignalP v. 4.1 (Petersen et al. 2011), TMHMM v. 2.0 (Krogh et al. 2001), TargetP 1.1 Server (Emanuelsson et al. 2000), GPIsom (Fankhauser and Maser 2005) and WoLF PSORT (Horton et al. 2007). Sequences of the secreted proteins were subjected to a MEROPS Batch BLAST analysis (e-value <1e-04) (Release 10.0; http://merops.sanger.ac.uk/) for predicting peptidases (Rawlings et al. 2016), and false positives were eliminated by parsing hits (whether or not the annotations involved peptidases) obtained following a BLASTp search (e-value <1e-04) on the NCBI nr protein database. Candidate secreted effector proteins (CSEPs) were defined here as the secreted proteins smaller than 200 aa in length, and a CSEP was labeled as "cysteine-rich" when the percentage of cysteine residues in that protein was at least twice as high as the average cysteine content of the proteome of that species. Membrane transporters were identified using the transporter prediction server TransportTP (e-value <1e-05) (Li et al. 2009). Prediction of genes coding for nonribosomal peptide synthase (NRPS), polyketide synthase (PKS), hybrid NRPS-PKS, dimethylallyl tryptophan synthase (DMATS), and terpene cyclase (TC) was performed using the programs SMURF (Khaldi et al. 2010) and antiSMASH 2.0 (Blin et al. 2013) in combination (union). Identification of carbohydrate-active enzyme (CAZyme) genes was achieved using the web server and database dbCAN (Yin et al. 2012). Four types of plant cell wall-degrading enzymes (PCWDEs), including cellulase, hemicellulase, pectinase and cutinase, were predicted using our previously described protocol (Xu et al. 2016). Potential pathogenicity and virulence-related genes were identified by BLASTP similarity searches against the pathogen-host interaction database (PHI-base) version 4.2 using a cutoff of <1e-05 (Winnenburg et al. 2006).

Homology and phylogenomic analysis

Besides the five species sequenced in this study, an additional 15 fungi, including 14 Capnodiales species and one Dothideales species (genomic data available mainly from the JGI website), were selected to perform the identification of gene families and phylogenomic analysis (table 2). Results from BLASTp (cutoff <1e-5) were used as input for clustering the sequences of both orthologs and paralogs with

OrthoMCL version 2.0.9 (Li et al. 2003). For the phylogenomic construction, only families containing exactly one gene copy for each of the 20 genomes were used, because families with paralogs may hinder correct phylogenetic inference. Multiple alignments of the protein sequences belonging to the same families were parallelly performed in batch mode using MAFFT (Katoh and Standley 2013). The sequences were concatenated and poorly aligned regions of each alignment were removed using Gblocks (Talavera and Castresana 2007). The trimmed alignment was subsequently used for phylogenetic reconstruction using maximum likelihood method implemented in RAxML with 1,000 bootstrap replications under the PROGAMMAILG model estimated by ProtTest v3.4 (Darriba et al. 2011; Stamatakis 2014). The RAxML tree was visualized in FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/). The divergence times were roughly estimated by MEGA v6.0 using the previously published splitting time (109 Myr) between *P. fructicola* and *Z. tritici* as a calibration point (Tamura et al. 2013; Xu et al. 2016). An ultrametric tree based on the RAxML tree was generated using Mesquite 3.04 (Maddison and Maddison 2015). Gene family expansions and contractions were estimated with the CAFE software version 3.0 using the ultrametric tree and the OrthoMCL gene families as input (De Bie et al. 2006). The ancestral set as well as gains and losses of gene families were inferred by reconciliation of the binary presence or absence profiles of fungal members constituting an orthologous group with the species phylogeny using the DOLLOP program (Felsenstein 2005), assuming irreversibility of the loss of an orthologous group.

Identification of species-specific HGT candidates

The species/lineage-specific genes (subproteomes) were searched using blastp (v2.2.31+) against the GenBank nr database (latest version). The blastp outputs were then subjected to HGT-Finer (*R* threshold ranging from 0.2 to 0.9, *Q* value <0.01) to identify HGT candidates (Nguyen et al. 2015). In addition, to eliminate potential DNA contamination, we examined the flanking sequences (or genes) of a candidate HGT; if these adjacent sequences are homologous to sequences of relative species of the HGT receptor, it is probably a true HGT event.

Literature Cited

Adhikari BN, et al. 2013. Comparative genomics reveals insight into virulence strategies of plant pathogenic oomycetes. PLoS One 8: e75072.

Aguin O, Sainz MJ, Ares A, Otero L, Mansilla JP. 2013. Incidence, severity and causal fungal species of *Mycosphaerella* and *Teratosphaeria* diseases in *Eucalyptus* stands in Galicia (NW Spain). Forest Ecol Manag. 302: 379-389.

Albalat R, Canestro C. 2016. Evolution by gene loss. Nat Rev Genet. 17: 379-391.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215: 403-410.

Ashberner M. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 5: 25-29.

Bairoch A. 2000. The ENZYME database in 2000. Nucleic Acids Res. 28: 304-305.

Belding RD, Sutton TB, Blankenship SM, Young E. 2000. Relationship between apple fruit epicuticular wax and growth of *Peltaster fructicola* and *Leptodontidium elatius*, two fungi that cause sooty blotch

disease. Plant Disease 84: 767-772.

Blin K, et al. 2013. antiSMASH 2.0-a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 41: W204-W212.

Boetzer M, Henkel CV, Jansen HJ, 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.

Braumann I, van den Berg M, Kempken F. 2008. Repeat induced point mutation in two asexual fungi, *Aspergillus niger* and *Penicillium chrysogenum*. Curr Genet. 53: 287-297.

Collemare J, et al. 2014. Secondary metabolism and biotrophic lifestyle in the tomato pathogen *Cladosporium fulvum*. PLoS One 9: e85877.

Copley RR, Goodstadt L, Ponting C. 2003. Eukaryotic domain evolution inferred from genome comparisons. Curr Opin Genet Dev. 13: 623-628.

Crous PW, Groenewald JZ, Mansilla JP, Hunter GC, Wingfield MJ. 2004. Phylogenetic reassessment of *Mycosphaerella* spp. and their anamorphs occurring on *Eucalyptus*. Stud Mycol. 50: 195-214.

Díaz Arias MMD, et al. 2010. Diversity and biogeography of sooty blotch and flyspeck fungi on apple in the eastern and midwestern United States. Phytopathology 100: 345-355.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27: 1164-1165.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22: 1269-1271.

D'hondt L, Höfte M, Van Bockstaele E, Leus L. 2011. Applications of flow cytometry in plant pathology for genome size determination, detection and physiological status. Mol Plant Pathol. 12: 815-828.

Dujon B. 1996. The yeast genome project: what did we learn? Trends Genet. 12: 263-270.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 300: 1005-1016.

Fankhauser N, Maser P. 2005. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics 21: 1846-1852.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html

Fierro F, Martin JF. 1999. Molecular mechanisms of chromosomal rearrangement in fungi. Crit Rev Microbiol. 25: 1-17.

Frank J, et al. 2010. *Microcyclospora* and *Microcyclosporella*: novel genera accommodating epiphytic fungi causing sooty blotch on apple. Persoonia 24: 93-105.

Freitag M, Williams RL, Kothe GO, Selker EU. 2002. A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. Proc Natl Acad Sci U S A. 99: 8802-8807.

Galagan JE, et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422: 859-868.

Glazebrook J. 2005. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. Annu Rev Phytopathol. 43: 205-227.

Gleason ML, et al. 2011. A new view of sooty blotch and flyspeck. Plant Disease 95: 368-383.

Gotz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36: 3420-3435.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. PloS Genet. 3: 2135-2146.

Hane JK, Oliver RP. 2008. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics 9: 478.

Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP. 2015. Repeat-induced point mutation: a fungal-specific, endogenous mutagenesis process. In: van den Berg MA, Maruthachalam K, editors. Genetic Transformation Systems in Fungi. Cham: Springer International Publishing. p. 55-68.

Hayashi K, Schoonbeek H-j, De Waard MA. 2002. Bcmfs1, a novel major facilitator superfamily transporter from *Botrytis cinerea*, provides tolerance towards the natural toxic compounds camptothecin and cercosporin and towards fungicides. Appl Environ Microb. 68: 4996-5004.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.

Horton P, et al. 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 35: W585-587.

Hunter GC, Crous PW, Carnegie AJ, Wingfield MJ. 2009. *Teratosphaeria nubilosa*, a serious leaf disease pathogen of *Eucalyptus* spp. in native and introduced areas. Mol Plant Pathol. 10: 1-14.

Huser A, Takahara H, Schmalenbach W, O'Connell R. 2009. Discovery of pathogenicity genes in the

crucifer anthracnose fungus *Colletotrichum higginsianum*, using random insertional mutagenesis. Mol Plant Microbe In. 22: 143-156.

Ismail SI, et al. 2016. Ancestral state reconstruction infers phytopathogenic origins of sooty blotch and flyspeck fungi on apple. Mycologia 108: 292-302.

Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30: 1236-1240.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32: D277-D280.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30: 772-780.

Kelkar YD, Ochman H. 2012. Causes and consequences of genome expansion in fungi. Genome Biol Evol. 4: 13-23.

Kelley JL, et al. 2014. Compact genome of the *Antarctic midge* is likely an adaptation to an extreme environment. Nat Commun. 5: 4611.

Khaldi N, et al. 2010. SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 47: 736-741.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5: 59.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305: 567-580.

Li HQ, Benedito VA, Udvardi MK, Zhao PX. 2009. TransportTP: a two-phase classification approach for membrane transporter prediction and characterization. BMC Bioinformatics 10: 418.

Li HY, et al. 2012. *Dissoconiaceae* associated with sooty blotch and flyspeck on fruits in China and the United States. Persoonia 28: 113-125.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178-2189.

Liu YC, Schroder J, Schmidt B. 2013. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics 29: 308-315.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4: 865-875.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955-964.

Maddison WP, Maddison DR. 2015. Mesquite: a modular system for evolutionary analysis. Version 3.04. http://mesquiteproject.org.

Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764-770.

Meerupati T, et al. 2013. Genomic mechanisms accounting for the adaptation to parasitism in nematode-trapping fungi. PloS Genet. 9: e1003909.

Mehrabi R, et al. 2011. Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. Fems Microbiol Rev. 35: 542-554.

Mondal SN, Timmer LW. 2006. Relationship of the severity of citrus greasy spot, caused by *Mycosphaerella citri*, to ascospore dose, epiphytic growth, leaf age, and fungicide timing. Plant Disease 90: 220-224.

Nguyen M, Ekstrom A, Li XQ, Yin YB. 2015. HGT-Finder: a new tool for horizontal gene transfer finding and application to *Aspergillus* genomes. Toxins 7: 4035-4053.

Nickel W, Seedorf M. 2008. Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells. Annu Rev Cell Dev Bi. 24: 287-308.

Nordberg H, et al. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42: D26-D31.

Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7: e30619.

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

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. Trends Genet. 17: 23-28.

Pires SD, et al. 2014. Identification of virulence factors in *Leishmania infantum* strains by a proteomic approach. J Proteome Res. 13: 1860-1872.

Rawlings ND, Barrett AJ, Finn R. 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 44: D343-D350.

Redman RS, Dunigan DD, Rodriguez RJ. 2001. Fungal symbiosis from mutualism to parasitism: who controls the outcome, host or invader? New Phytol. 151: 705-716.

Richards TA, Leonard G, Soanes DM, Talbot NJ. 2011. Gene transfer into the fungi. Fungal Biol Rev. 25: 98-110.

Saier MH, et al. 2016. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res. 44: D372-D379.

Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19: 1117-1123.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312-1313.

Stanke M, et al. 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. 34: W435-W439.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56: 564-577.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30: 2725-2729.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 25: 4.10.11-14.10.14.

Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. Nat Rev Genet. 12: 692-702.

Tavares S, et al. 2014. Genome size analyses of Pucciniales reveal the largest fungal genomes. Front Plant Sci. 5: 422.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 18: 1979-1990.

Wang YP, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40: e49.

Williamson SM, Sutton TB. 2000. Sooty blotch and flyspeck of apple: Etiology, biology, and control. Plant Disease 84: 714-724.

Winnenburg R, et al. 2006. PHI-base: a new database for pathogen host interactions. Nucleic Acids Res. 34: D459-D464.

Xu C, et al. 2016. *Peltaster fructicola* genome reveals evolution from an invasive phytopathogen to an ectophytic parasite. Sci Rep. 6: 22926

Xu HB, et al. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS One 7: e52249.

Yin YB, et al. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40: W445-W451.

Figure legends

Fig. 1. Phylogenetic relationship and analysis of gene families. (A) Circular phylogenetic tree of 20 fungal species constructed using the maximum likelihood (ML) method. Clades of the three monophyletic groups are respectively highlighted by different colours, including Mycma and Micma (red); Zasci and Zasan (blue); Ternu and Micpo (yellow). The black arrows represent the divergence time of each two lineages with Myr used to abbreviate million years. All the internal nodes are labeled with arabic numerals in hollow circles. (B) Venn diagrams of the predicted gene families in the fungi of each monophyletic group versus those of all the other fungal species. The predicted gene families of all fungi except the corresponding monophyletic group were combined and named "Complex".

Fig. 2. The numbers of plant pathogenicity-related protein-coding genes in the six investigated fungi. (A) Membrane transporters. ABC transporters, MFS transporters and all the other transporters were indicated using three different shades of colours, respectively. (B) Secreted proteins (SPs). Peptidases, candidate secreted effector proteins (CSEPs) and all the other SPs were indicated using three different shades of colours, respectively. (C) Plant cell wall degrading enzymes. Cutinases and degrading enzymes for only pectins, for only hemicelluloses, for both pectins and hemicelluloses, and for both celluloses and hemicelluloses were indicated using five different shades of colours, respectively. (D) Secondary metabolite biosynthesis core enzymes. PKSs, NRPSs, PKS-NRPSs, DMATSs and TCs were indicated using five different shades of colours, respectively. PKS, polyketide synthase; NRPS, nonribosomal peptide synthase; TC, terpene cyclase; DMATS, dimethyl allyl tryptophan synthase.

Fig. 3. Evolution of gene families and orphan genes in the six investigated fungi. (A) Four gene family events (including acquisition, loss, expansion and contraction) occurring during the evolution of three pairs of fungal species from their respective most recent common ancestors. This unrooted tree is a simplified version of the circular phylogenetic tree in Fig. 1A. (B) The numbers of gene families (having experienced acquisition, loss, expansion and contraction) and orphan genes putatively involved in pathogen-host interactions (PHIs).

Fig. 4. Fold changes (Log_{10} scale) in dinucleotide abundances (left) and RIP indices (right) for repeat families of six investigated fungi compared to their non-repetitive control sequences. (A) Mycma and Micma. (B) Zasci and Zasan. (C) Ternu and Micpo.

Fig. 5. Sequence comparison of the *Neurospora crassa* RID (AF500227) protein and its homologous sequences identified in the genomes of six investigated fungi. Only the catalytic DMT domains are shown and the ten DMT motifs are outlined and numbered with roman numerals. Asterisks indicate positions occupied by identical amino acids, while dots represent positions showing conservative amino acid exchanges.

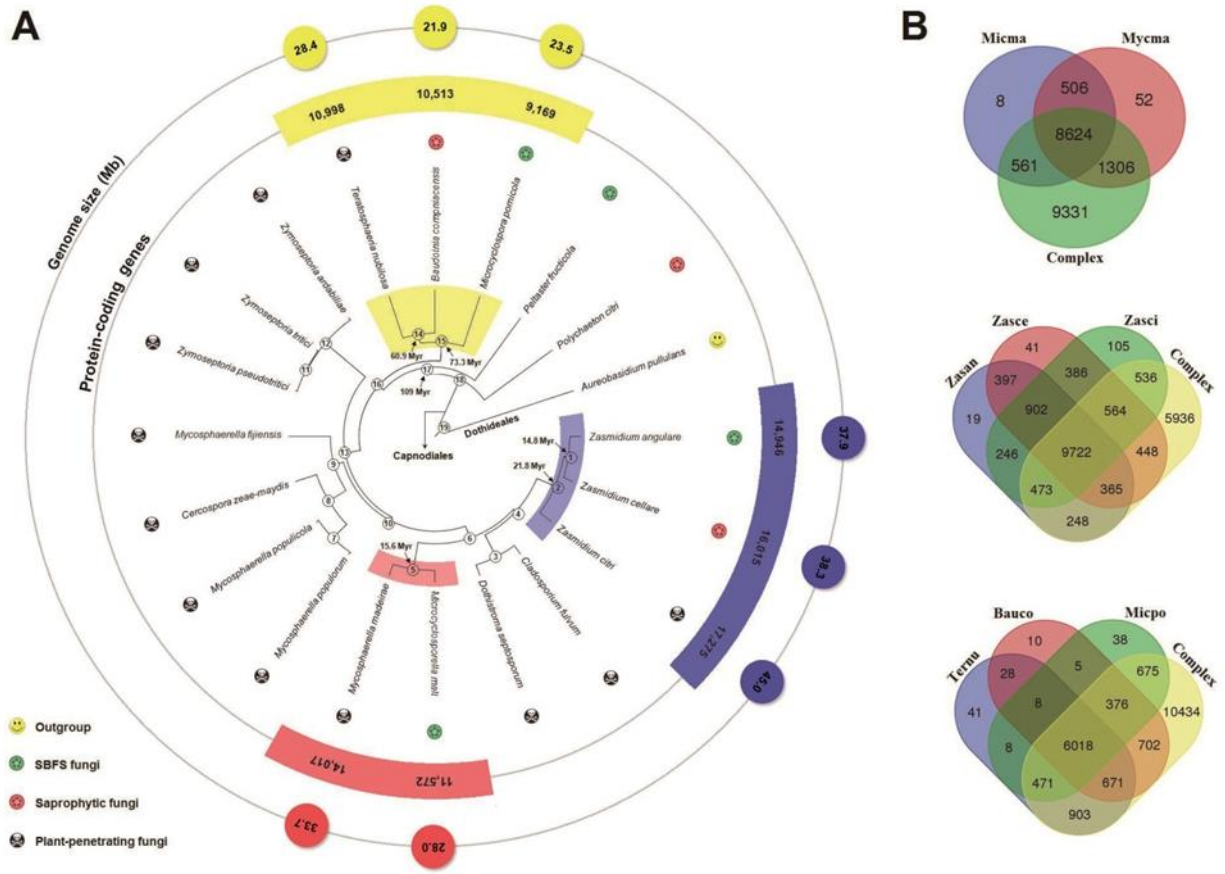

Fig. 1

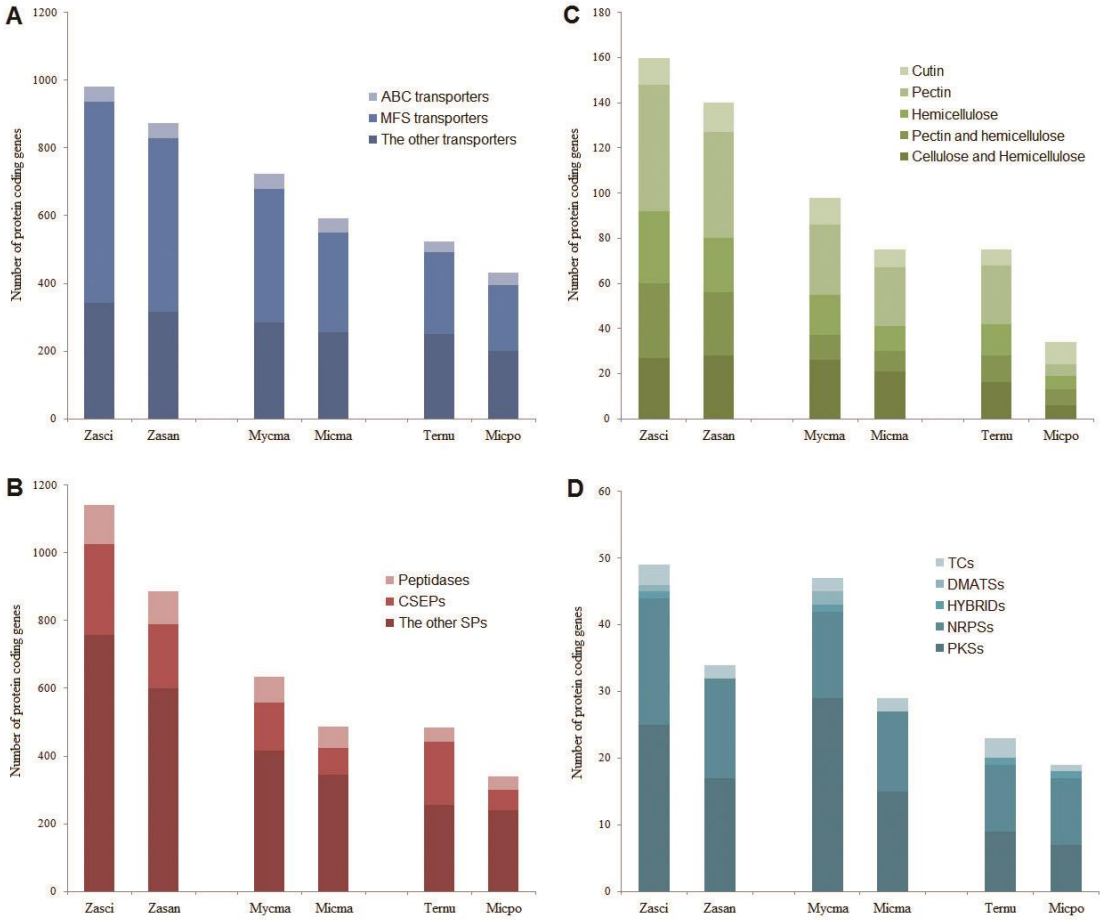

Fig. 2

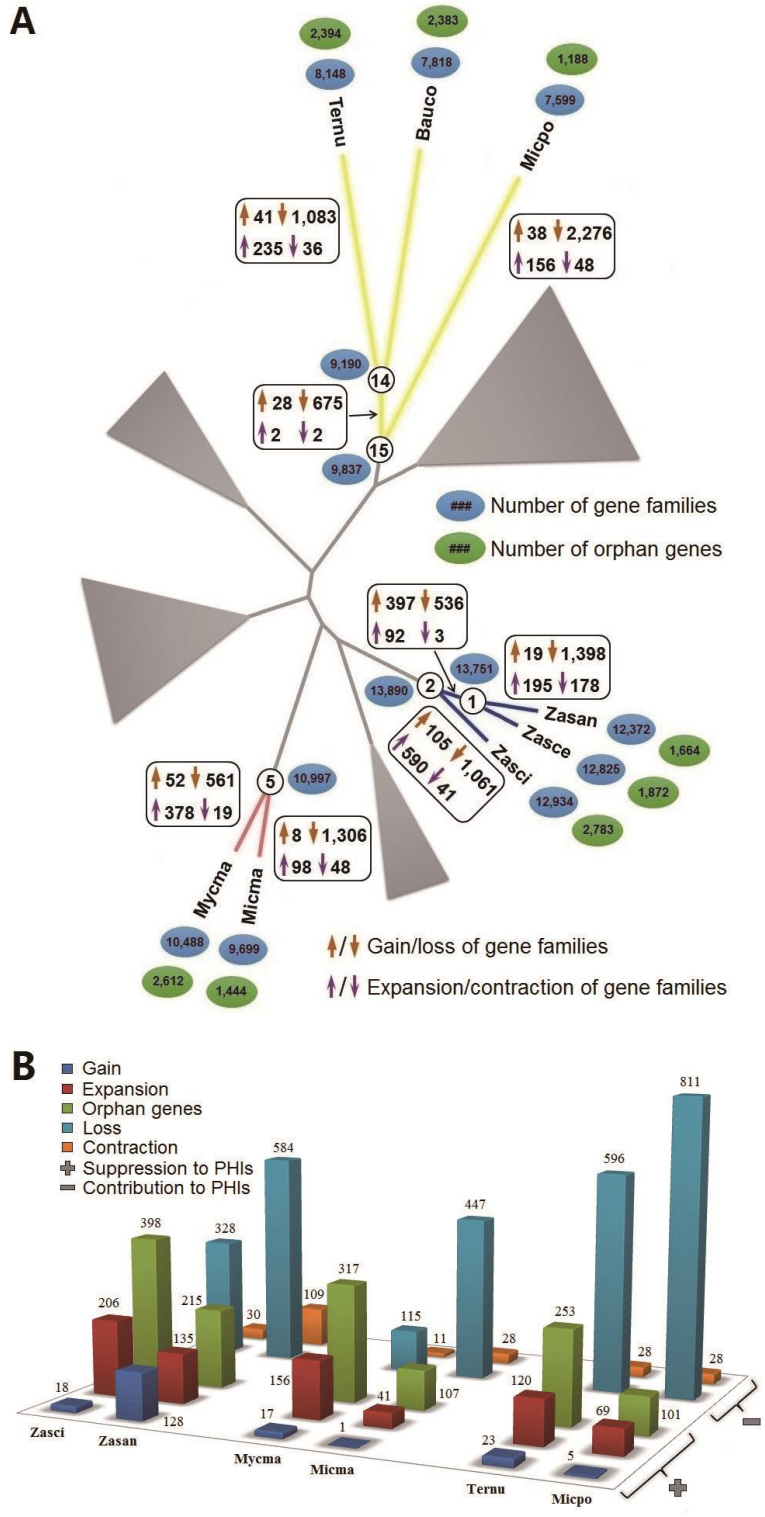

Fig. 3

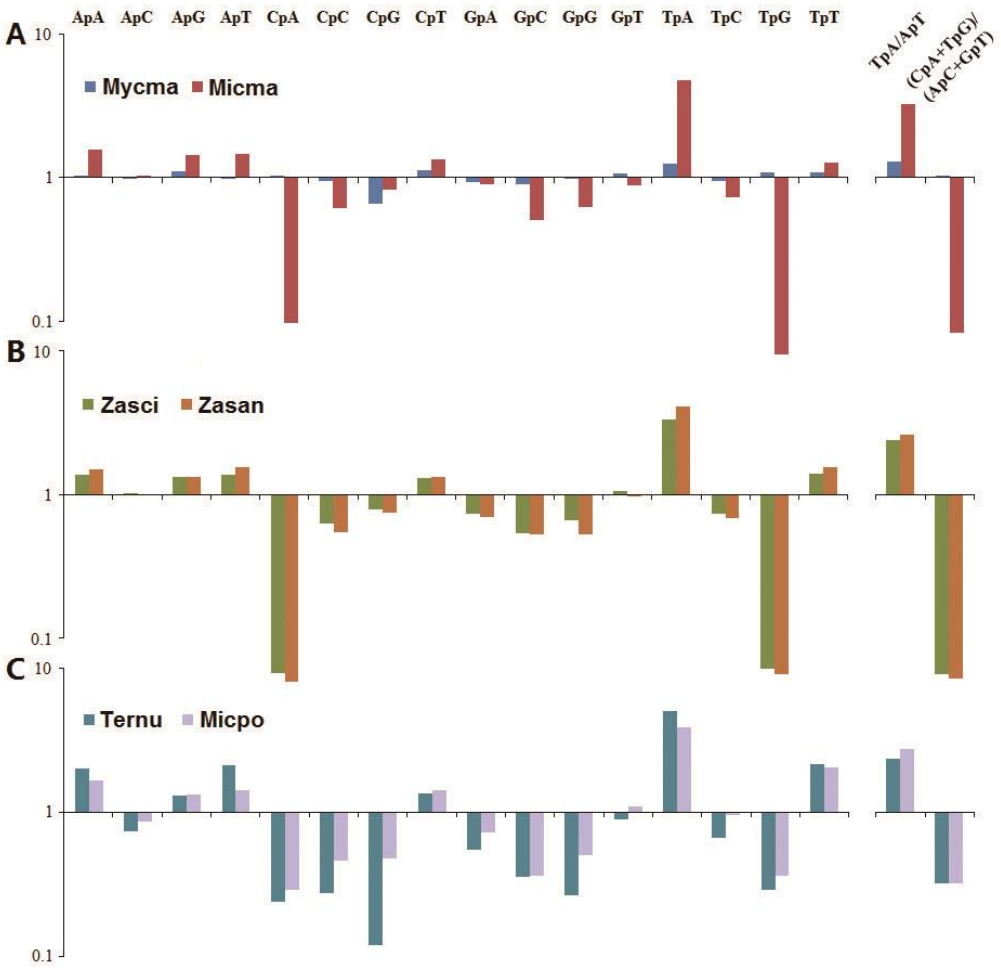

Fig. 4

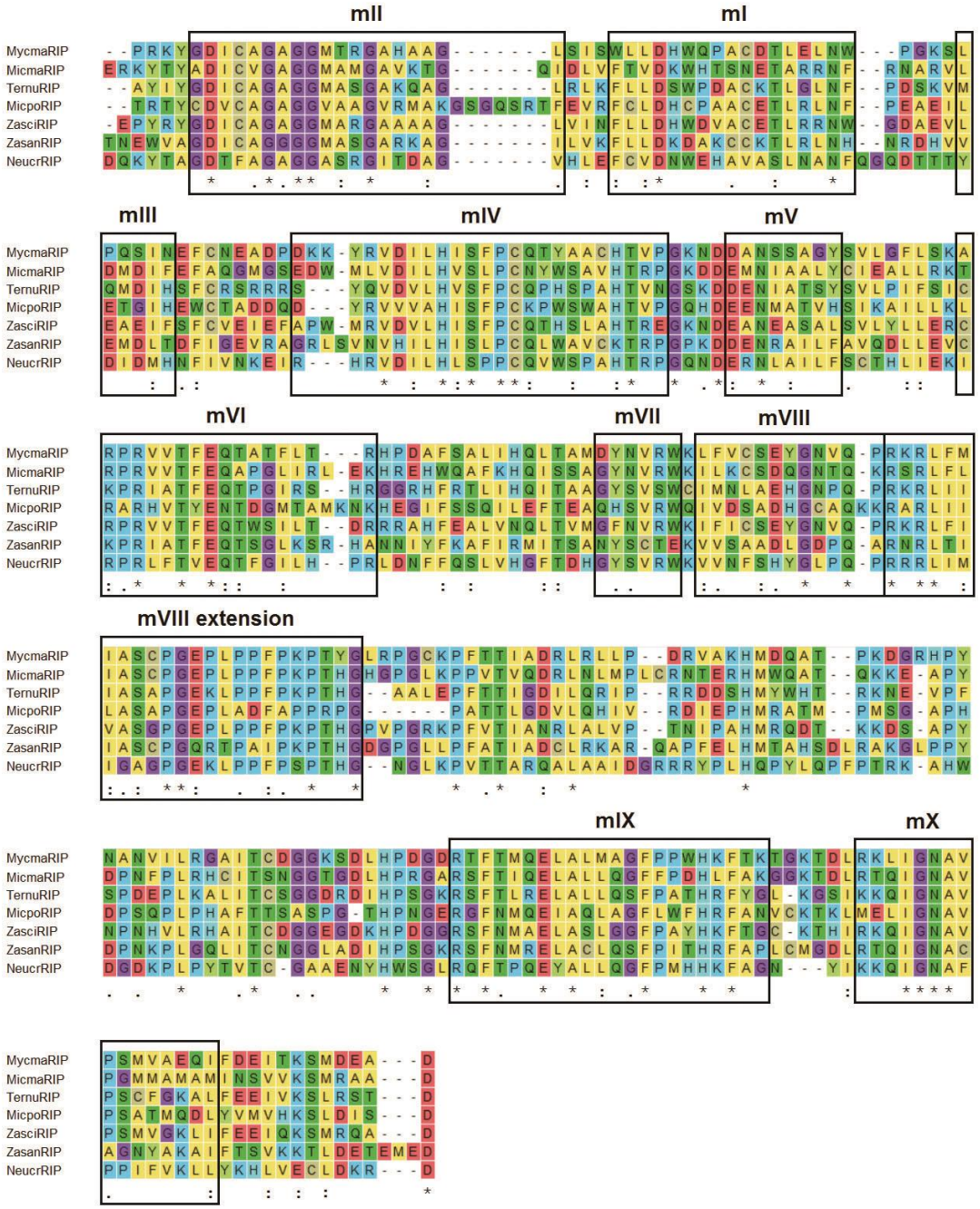

Fig. 5

Table Legends

Table 1. Genome properties of six fungi compared in this study

Parameter	Mycma	Micma	Zasci	Zasan	Ternu[a]	Micpo
Assembly size (Mb)	33.7	28.0	45.0	37.9	28.4	23.5
Genome estimated size (Mb)	36.7	31.9	46.9	41.8	-	24.3
Estimated coverage (fold)	143	65	157	54	-	124
Number of gaps	258	99	125	163	-	75
Gap length (bp)	5,405	2,510	2,964	8,233	-	3,754
GC content (%)	52.72	52.73	51.82	52.45	-	52.93
Number of scaffolds (>200 bp)	867	122	277	359	-	109
Scaffold N50 (kb)	335	1,515	631	1,171	-	557
Max. scaffold size (kb)	1,798	2,115	1,847	2,523	-	2,538
Min. scaffold size (bp)	208	206	215	166	-	202
Mean scaffold size (kb)	39	229	162	106	-	216
Number of protein-coding genes	14,017	11,572	17,275	14,946	10,998	9,169
Mean exon length (bp)	549	623	569	594	-	679
Repeat ratio (%)	6.09	1.57	8.35	3.6	7.88	1.95
Number of tRNAs	116	79	167	110	-	74
NR (percentage in total genes)	11,800 (84%)	10,082 (87%)	14,511 (84%)	13,074 (87%)	9,038 (82%)	8,200 (89%)
KEGG (percentage in total genes)	1,178 (8%)	1,149 (10%)	1,391 (8%)	1,305 (9%)	1,065 (10%)	994 (11%)
GO (percentage in total genes)	8,279 (59%)	7,336 (63%)	9,836 (57%)	9,198 (62%)	6,822 (62%)	6,194 (68%)
KOG (percentage in total genes)	6,467 (46%)	5,805 (50%)	7,635 (44%)	7,163 (48%)	5,352 (49%)	5,063 (55%)

[a]The genome of this fungus was not sequenced here but previously.

Table 2. Species used in the phylogenetic analysis and identification of gene families

Species[a]	Taxonomy	Lifestyle/trophic mode	Proteome	Source of genomic data
Microcyclospora pomicola	Capnodiales	Ectophyte (SBFS)	9,169	This study
Microcyclosporella mali	Capnodiales	Ectophyte (SBFS)	11,572	This study
Mycosphaerella madeirae	Capnodiales	Plant-penetrating parasite	14,017	This study
Zasmidium angulare	Capnodiales	Ectophyte (SBFS)	14,946	This study
Zasmidium citri (*Mycosphaerella citri*)	Capnodiales	Plant-penetrating parasite	17,275	This study
Aureobasidium pullulans	Dothideales	Epiphyte or endophyte	11,866	JGI MycoCosm
Baudoinia compniacensis	Capnodiales	Saprophyte	10,513	JGI MycoCosm
Cercospora zeae-maydis	Capnodiales	Plant-penetrating parasite	12,020	JGI MycoCosm
Cladosporium fulvum	Capnodiales	Plant-penetrating parasite	14,127	JGI MycoCosm
Dothistroma septosporum (*Mycosphaerella pini*)	Capnodiales	Plant-penetrating parasite	12,580	JGI MycoCosm
Mycosphaerella fijiensis	Capnodiales	Plant-penetrating parasite	13,107	JGI MycoCosm
Mycosphaerella populicola (*Septoria populicola*)	Capnodiales	Plant-penetrating parasite	9,739	JGI MycoCosm
Mycosphaerella populorum (*Septoria musiva*)	Capnodiales	Plant-penetrating parasite	10,223	JGI MycoCosm
Peltaster fructicola	Capnodiales	Ectophyte (SBFS)	8,334	NCBI
Polychaeton citri (*Capnodium citri*)	Capnodiales	Saprophyte	10,582	JGI MycoCosm
Teratosphaeria nubilosa	Capnodiales	Plant-penetrating parasite	10,998	JGI MycoCosm
Zasmidium cellare	Capnodiales	Saprophyte	16,015	JGI MycoCosm
Zymoseptoria pseudotritici	Capnodiales	Plant-penetrating parasite	11,044	JGI MycoCosm
Zymoseptoria ardabiliae	Capnodiales	Plant-penetrating parasite	10,788	JGI MycoCosm
Zymoseptoria tritici (*Mycosphaerella graminicola*)	Capnodiales	Plant-penetrating parasite	10,933	JGI MycoCosm

[a]The anamorph, teleomorph or synonym of a certain species is provided in parenthesis.