# Novel mating-type-associated genes and gene fragments in the genomes of Mycosphaerellaceae and Teratosphaeriaceae fungi

Janneke Aylward[a,b], Minette Havenga[a,b], Brenda D. Wingfield[a], Michael J. Wingfield[a], Léanne L. Dreyer[c], Francois Roets[b], Emma T. Steenkamp[a]

[a] Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20 Hatfield, 0028, South Africa
[b] Department of Conservation Ecology and Entomology, Stellenbosch University, Private Bag X1 Matieland, 7602, South Africa
[c] Department of Botany and Zoology, Stellenbosch University, Private Bag X1 Matieland, 7602, South Africa

Email-addresses:
Janneke Aylward: janneke.aylward@fabi.up.ac.za (corresponding author)
Minette Havenga: havenga.minette@fabi.up.ac.za
Brenda D. Wingfield: brenda.wingfield@fabi.up.ac.za
Michael J. Wingfield: mike.wingfield@fabi.up.ac.za
Léanne L. Dreyer: ld@sun.ac.za
Francois Roets: fr@sun.ac.za
Emma T. Steenkamp: emma.steenkamp@fabi.up.ac.za

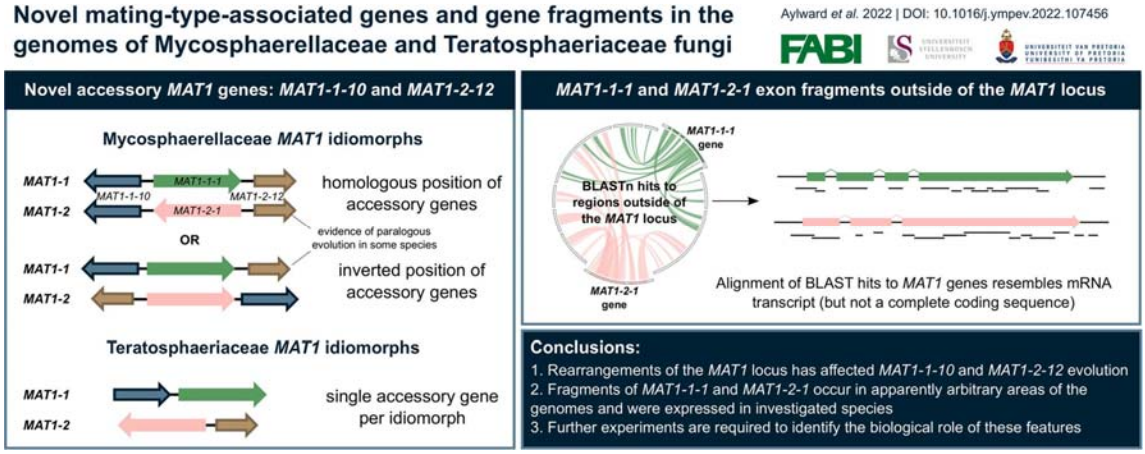
## Highlights

- Two unusual mating-type-associated features have evolved in Mycosphaerellales fungi.
- Some Mycosphaerellaceae have apparently paralogous genes in both *MAT1* idiomorphs.
- Almost 50% of species contained short *MAT* gene fragments outside of the *MAT1* locus.
- *MAT* fragments resembled mRNA transcripts and may play a regulatory role in mating.
- The paralogous *MAT* genes and fragments were expressed in representative species.

## Abstract

The mating-type (*MAT1*) locus encodes transcription factors essential for the onset of the sexual cycle in ascomycete fungi. This locus has been characterised in only a few heterothallic, plant pathogenic Mycosphaerellaceae and Teratosphaeriaceae. We used available genome sequences for Mycosphaerellales species to investigate the presence of two unique mating-type-associated features. The accessory *MAT1* genes, *MAT1-1-10* (*MATORF2*) and *MAT1-2-12* (*MATORF1*), typically occurred in both *MAT* idiomorphs of Mycosphaerellaceae species. In contrast, they were associated with only one idiomorph in Teratosphaeriaceae species. In *Pseudocercospora*, phylogenetic analyses showed that homologs present in different idiomorphs were paralogous and subject to different selective pressures, indicating that their evolution is linked to mating type. In almost half of the investigated Mycosphaerellales genomes, numerous short fragment sequences, almost identical to portions of the *MAT1-1-1* and *MAT1-2-1* genes, were present in multiple areas outside of the *MAT1* locus. Aligned to the *MAT1* genes, these sequences resembled an mRNA transcript. Fragment sequences were similar among species groups and occurred at the same genomic positions, implying that monophyletic groups have the same origins of these sequences. Although the functions of the *MAT* fragment sequences and accessory *MAT1* genes remain unknown, both were expressed in the representative Mycosphaerellaceae and Teratosphaeriaceae species that were investigated.

## Graphical abstract




**Keywords**: evolution; fragments; fungi; genes; mating; Mycosphaerellales

# 1. Introduction

The most abundant and best-known species in the Mycosphaerellaceae and Teratosphaeriaceae are plant pathogens. These fungi infect a wide range of economically important crops, including grains (Torriani et al., 2015), fruit (Marin et al., 2003), root vegetables (Weiland and Koch, 2004) and trees utilized in plantation forestry (Aylward et al., 2019; Bulman et al., 2016). The two families are accommodated in the Mycosphaerellales, an Order recently resurrected within Capnodiales *sensu lato* (Dothideomycetidae: Dothideomycetes; Abdollahzadeh et al., 2020; Crous et al., 2009). Although Mycosphaerellales includes fungi with ectophytic, saprophytic and pathogenic lifestyles (Abdollahzadeh et al., 2020), species in the Mycosphaerellaceae are predominantly plant pathogens (Videira et al., 2017). Their socioeconomic importance is reflected in databases containing whole genome sequence information for many species (Aylward et al., 2017). For example, in April 2020, Mycosphaerellaceae species represented more than 75% of the Mycosphaerellales having whole genome sequences in GenBank, with the remaining genomes (in descending order) being for species of Teratosphaeriaceae, Dissoconiaceae and Aeminiaceae.

Studies on plant pathogens commonly first consider their reproductive biology because this factor determines how diversity is distributed across populations (McDonald and Linde, 2002). For example, recombination associated with sexual reproduction generates genetic diversity that can, in turn, enable a population to overcome selection pressures (Ni et al., 2011). Sexual reproduction in Ascomycetes is controlled by the mating-type (*MAT1*) locus (Turgeon and Yoder, 2000; Wilken et al., 2017; Wilson et al., 2021). Heterothallic (obligately outcrossing) species have two non-allelic versions or idiomorphs (i.e., *MAT1-1* and *MAT1-2*) of the locus, at which sexually compatible individuals encode either the *MAT1-1-1* or *MAT1-2-1* gene, as well as other accessory genes (Coppin et al., 1997; Kronstad and Staben, 1997). Species that are homothallic (capable of selfing) harbour both the *MAT1-1-1* and *MAT1-2-1* genes in the same genome (Lin and Heitman, 2007; Wilson et al., 2015).

Among species in the Mycosphaerellales, the *MAT1* locus has been characterised in six genera. All of these characterised species are heterothallic and either reside in the Mycosphaerellaceae or Teratosphaeriaceae. In the case of the Teratosphaeriaceae, all are *Teratosphaeria* species (Aylward et al., 2020; Havenga et al., 2020), whereas those in the Mycosphaerellaceae include members of the genera *Cercospora* (Bolton et al., 2014; Groenewald et al., 2006), *Dothistroma* (Groenewald et al., 2007), *Passalora* (Stergiopoulos et al., 2007), *Pseudocercospora* (Arzanlou et al., 2010; Conde-Ferráez et al., 2007) and *Zymoseptoria* (Goodwin et al., 2003; Waalwijk et al., 2002).

Even though the two *MAT1* idiomorphs of heterothallic species are usually dissimilar and encode non-homologous genes (Turgeon and Yoder, 2000), regions of inverted sequence homology have been identified between the ends of the *Pseudocercospora* idiomorphs (Conde-Ferráez et al., 2007; see Fig. 2). These inverted repeats were later shown to encode two hypothetical open reading frames (ORFs) that were named *MATORF1* and *MATORF2* (Arzanlou et al., 2010). They were also identified in the *MAT1* idiomorphs of *Cercospora*, but were not inverted between *MAT1-1* and *MAT1-2* individuals in this species (Bolton et al.,

2014). In other words, *MATORF1* and *MATORF2* occur in both of the *MAT1* idiomorphs of the examined *Pseudocercospora* and *Cercospora* species. This is different from the situation in *Teratosphaeria* species where *MATORF2* is located in the *MAT1-1* idiomorph and *MATORF1* in the *MAT1-2* idiomorph (Aylward et al., 2020). These genes were consequently renamed *MAT1-1-10* and *MAT1-2-12*, respectively (Aylward et al., 2020; Havenga et al., 2020).

Some species of Mycosphaerellaceae are peculiar in terms of the genomic distribution of sequences that show homology to *MAT1*-encoded genes. In *Pseudocercospora*, Arzanlou et al. (2010) reported two genomic regions outside this locus containing 336-452 base pair (bp) sequences that are similar in sequence to portions of the *MAT1-1-1* and *MAT1-2-1* genes. Subsequently, Bolton et al. (2014) reported the occurrence of such sequences at multiple locations in the *Cercospora beticola* and *Pseudocercospora fijiensis* genomes. These *MAT* fragment sequences are apparently derived from exons of the *MAT1* genes, because near full-length *MAT1-1-1* and *MAT1-2-1* coding sequences can be reconstructed from the fragment sequences in each genome. Although the reconstructed gene sequences have high identity to those encoded at the *MAT1* locus, the origin and possible function of these additional *MAT* sequences remain unknown.

The *MAT1* locus and idiomorphs of Ascomycota have common structural features that tend to be conserved among related taxa (e.g. Wilken et al., 2018). The occurrence of additional *MAT* fragment sequences in the genomes of *Cercospora* and *Pseudocercospora*, together with the variability in copy-number and position of *MAT1-1-10* and *MAT1-2-12*, potentially present clues as to how this locus has evolved in the Mycosphaerellales. The aim of this study was to investigate the evolution of the *MAT1* locus, specifically in species of Mycosphaerellaceae and Teratosphaeriaceae. Consequently, we utilised genomes for species in the Mycosphaerellales that are publicly available on GenBank to reconstruct and compare the *MAT1* locus and its unique features in these two families. We thus investigated the physical placement and possible evolutionary history of the *MAT1-1-10* and *MAT1-2-12* genes in these fungi. We also considered the distribution and potential origin of the *MAT* fragment sequences outside of the *MAT1* locus and conducted a preliminary analysis of the expression of these features in representative species.

## 2. Materials and methods

### *2.1 Mycosphaerellales genomes and phylogeny*

Whole genome sequences for all Mycosphaerellales species available on GenBank up to April 2020 (Tables 1 and S1) were downloaded from the Genome repository of the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/genome/). Where more than one genome assembly per species was available, the annotated genome (if available) with the largest N50-value was chosen. Two *Cladosporium* (Cladosporiales) genomes were also included as outgroup.

Relationships among taxa were inferred using a phylogenomic approach in which multiple single-copy gene sequences were concatenated (Rokas et al., 2003; Waterhouse et al., 2017). For this purpose, BUSCO (Benchmarking Universal Single-Copy Orthologs) 3.0.2 (Simão et al., 2015) was used to identify homologs of the 290 genes included in the "Fungi *odb9*" dataset in all genomes. The sequences for "unduplicated" homologs present in all genomes (i.e. single-copy shared orthologs) were extracted with accessory scripts (extract_buscos_phylo.py and fetch_best_sequence.py) included in the BUSCO 3.0 distribution. The extracted sequences were then aligned with MAFFT 7.490 (Katoh and Standley, 2013), trimmed with trimAl 1.4.rev22 (Capella-Gutiérrez et al., 2009) and concatenated with FASconCAT-G 1.04 (Kück and Longo, 2014). ModelTest-NG and the Akaike information criterion (AIC; Darriba et al., 2020) were used to select the best substitution model for each gene partition in the alignment. A maximum likelihood (ML) phylogeny was computed with RAxML-NG 1.1 (Kozlov et al., 2019). To estimate the necessary number of bootstrap replicates, the RAxML-NG analysis included the "transfer bootstrap expectation" (TBE; Lemoine et al., 2018) metric, as well as "bootstopping", which is the bootstrap convergence test of Pattengale et al. (2010). Interactive Tree Of Life (iTOL) v5 (Letunic and Bork, 2019) was used to visualise the trees.

To investigate the probability of a homothallic or heterothallic Mycosphaerellales ancestor, the species tree was used to perform an ancestral character estimation with the Ape 5.4-1 (Paradis and Schliep, 2018) and Phytools 1.0-1 (Revell, 2012) packages in R 4.0.2 (R Core Team, 2020). Each taxon was labelled as heterothallic or homothallic, depending on whether one or both *MAT1* genes were detected in the genome (see Section 3.2). The thallic state at internal nodes was estimated with both a continuous-time Markov chain (Mk) model and a Bayesian stochastic character mapping approach. Because homothallic species were underrepresented in our dataset, an "all-rates-different" model was applied in both approaches, whereby the transition rate from heterothallism to homothallism, and *vice versa*, was permitted to differ, thus maximising the probability of obtaining the minority state.

### *2.2 Position, evolution and function of* MAT1-1-10 *and* MAT1-2-12

To minimise errors associated with the identification and characterisation of the *MAT1* locus, the *MAT1-1-10* and *MAT1-2-12* genes were investigated only in genera for which the *MAT1* locus had previously been described (Table S2). *Teratosphaeria nubilosa* was the only homothallic representative in this set. Therefore, two additional homothallic species, *Nothophaeocryptopus gaeumannii* and *Pallidocercospora crystallina*, were included for comparative purposes, although their *MAT1* loci have not been characterised previously. For each species, the region containing the *MAT1-1-1* and/or *MAT1-2-1* genes, together with 2000 bp of the upstream and downstream regions flanking them, were extracted from the respective genomes. If a genome of an isolate of the same species, but with the opposite mating type, was available on GenBank, its *MAT1* locus was also extracted. For species without gene predictions associated with their GenBank accessions, the *MAT1* locus was annotated with Augustus 3.3.3 (Stanke et al., 2006) using *Fusarium graminearum* gene models. The predicted genes were then identified and curated based on similarity to published *MAT1* sequences of these genera (Table S2).

The coding sequences of the *MAT1-1-10* and *MAT1-2-12* genes were compared among species by performing a codon-based alignment for each gene with MACSE v2.03 (Ranwez et al., 2018). These nucleotide alignments were then subjected to ML phylogenetic analyses on the NGPhylogeny.fr platform (Lemoine et al., 2019) using BMGE 1.12_1 (Criscuolo and Gribaldo, 2010) and PhyML with Smart Model Selection 1.8.1_1 (Lefort et al., 2017), applying 1000 bootstrap replicates. Protein conservation was evaluated using ConSeq (Berezin et al., 2004) on the ConSurf server (Ashkenazy et al., 2016). For each protein, HMMER 3.3 (http://hmmer.org/) was used to build a profile hidden Markov model (HMM) from the alignment and to search this HMM against the fungal sequences in the Identical Protein Groups resource on NCBI (https://www.ncbi.nlm.nih.gov/ipg/). The coding sequences were also compared to non-coding RNAs using cmscan 1.2.1 on the Rfam 14.1 database (Kalvari et al., 2017; https://rfam.xfam.org/).

EasyCodeml v1.21 (Gao et al., 2019; Yang, 2007) was used to test whether selective pressures differ between *MAT1-1-10* and *MAT1-2-12* gene copies in the different *Pseudocercospora MAT* idiomorphs. For each gene, the subtree consisting of the *Pseudocercospora* clade was taken from the phylogeny generated above, also retaining the *Teratosphaeria* genes as single-copy representatives. Site-specific selective constraints were modelled on the unrooted trees using the null model M2a_rel (Weadick and Chang, 2012) and Clade Model C (CmC; Bielawski and Yang, 2004). These models calculate the ratio ($\omega$) of nonsynonymous to synonymous substitution rates, considering three site classes and their proportions. The first two classes evolve, respectively, under purifying ($0 < \omega_0 < 1$) and neutral ($\omega_1=1$) selection. The third site class is free to evolve conservatively or under positive/diversifying selection ($\omega_2 > 0$). The null hypothesis of a constant selective pressure among clades was compared to three alternative hypotheses, each selecting either one or both of the *Pseudocercospora MAT1* gene clades as "foreground" branches that have a $\omega$ value different from the $\omega_2$ of the "background" branches. Hypotheses were compared with Likelihood Ratio Tests (LRTs; Weadick and Chang, 2012), applying a Bonferroni-corrected threshold value of $\alpha = 0.05/5 = 0.01$ (Haynes, 2013) to correct for all pairwise comparisons of hypotheses.

### *2.3 Characterisation of* MAT *fragments*

The presence of fragment sequences with high levels of similarity to *MAT1* genes and that occur outside of the *MAT1* locus in the Mycosphaerellales genomes was assessed with local BLASTn searches using the BLAST 2.9.0+ package (Altschul et al., 1990). Each genome was queried against a database of previously characterised Mycosphaerellaceae and Teratosphaeriaceae *MAT1* gene sequences (Fig. S1a), which included *MAT1-1-1*, *MAT1-2-1*, *MAT1-1-10* and *MAT1-2-12*, as well as 100 bp of the upstream and downstream sequences flanking each gene. The local BLAST database comprised sequences from *Cercospora beticola* (GenBank accessions KC960688 and KC960689), *Dothistroma pini* (DQ915449 and DQ915452), *Passalora fulva* (DQ659350 and DQ659351), *Pseudocercospora fijiensis* (DQ787015 and DQ787016), *Teratosphaeria gauchensis* (MN119558 and MN119559) and *Zymoseptoria tritici* (AF440399 and AF440398). The maximum Expect(e)-value was set to $1x10^{-6}$ and default values were retained for the remaining search parameters.

For each genome, the BLASTn result with the longest hit length (typically > 600 bp), as well results surrounding the area of the longest hit, was assumed to represent the *MAT1* locus. Shorter hits outside of the *MAT1* locus were recorded as "*MAT* fragments". Cases where such *MAT* fragments were not found were confirmed by extracting the *MAT1* locus from the query genome and adding it to the BLASTn database. Once all genomes had been investigated, the BLASTn search was repeated on the entire dataset using the updated database (see Fig. S1a).

The areas of the genome that contained *MAT* fragments were further investigated in available *MAT1-1* and *MAT1-2* representatives of each genus with a characterised *MAT1* locus, as well as the three homothallic species included previously (Fig. S1b). For this purpose, the identified *MAT* fragments (as well as 5000 bp up- and down-stream flanking regions) were extracted from each contig using CLC Genomics Workbench 20 (www.qiagenbioinformatics.com). Unless gene predictions were included in the GenBank accession, contigs containing these sequences were annotated with Augustus as described above. The genomic locations of the *MAT* fragments were compared between the species with reciprocal BLASTn searches of the fragment and flanking regions. Gene positions that overlapped with the fragment sequences were identified with BLASTx searches in Blast2GO 5.2.5 (Götz et al., 2008). Local alignments between the *MAT1* gene queries and *MAT* fragments were visualised with Circoletto 07.09.16 (Darzentas, 2010).

For all available species in the genera considered, the *MAT* fragment sequences from each genome were aligned with BLASTn to the *MAT1* gene(s) present in the respective genomes and, for heterothallic species, the opposite *MAT1* gene available in a different isolate. Where the alignment of *MAT* fragment sequences against their respective *MAT1* gene overlapped and nucleotides differed, the corresponding IUPAC degeneracy code was inserted to account for the polymorphism. An ML phylogeny of the full-length *MAT1* genes and their reconstructed counterparts was generated on the NGPhylogeny.fr platform as described above, but using MAFFT for alignment.

### *2.4 Expression of* MAT1 *genes and fragment sequences*

RNA-Seq data generated for previous *in vitro* studies of eight species were used to assess the expression of *MAT1* genes and *MAT* fragment sequences. Those for the Mycosphaerellaceae species *C. sojinae*, *C. zeina*, *D. septosporum*, *Pa. fulva*, *Ps. eumusae*, *Ps. fijiensis* and *Ps. musae* were available from the NCBI Sequence Read Archive (SRA; Table S3). Data for *T. destructans* (Teratosphaeriaceae) was generated for a study by Havenga *et al*. (in prep.) and was the only species for which data from isolates of opposite mating type were available. For *T. destructans*, *MAT1-1* (isolate CMW45661) and *MAT1-2* (isolate CMW44962) cultures were grown on nitrogen-deficient medium (Havenga et al., 2020) at 25 °C in the dark for 21 days. RNA was extracted and sequenced following the approach of Havenga et al. (2021).

All RNA-Seq libraries were quality-filtered with Trimmomatic v0.39 (Bolger et al., 2014) and mapped against both *MAT1* idiomorphs and the *MAT* fragment-containing regions of their corresponding species (Table S3) to assess expression. For the Mycosphaerellaceae species, if replicate libraries or libraries for the same strain grown in different *in vitro* conditions were

available, all were mapped against the genomic reads simultaneously, using STAR v2.7.5b (Dobin et al., 2013). Geneious R11.1.5 (https://www.geneious.com) was used to view the mapped reads and determine whether the genes and *MAT* fragments were expressed. Since the data from different species were not collected consistently and libraries were mapped together, differential expression was not considered, but genes and fragments were classified as expressed or not-expressed based on visual examination and a raw transcript count of at least 10. Visual examination entailed ensuring that the mapped RNA-Seq reads covered the entire gene or fragment and that SNPs were not consistently present in the mapped data.

A more in-depth analysis was conducted for *T. destructans*, because the data for strains of opposite mating type were generated for a comparative transcriptomics study. Using STAR, the reads of each replicate were mapped separately to five different templates. These were the whole genome sequences of (i) CMW45661 and (ii) CMW44962, their individual (iii) *MAT1-1* (GenBank accession MN531144) and (iv) *MAT1-2* (GenBank accession MN531146) idiomorphs and (v) the genomic regions in which *MAT* fragment sequences were identified. Expression was quantified in R version 3.6.0 (R Core Team, 2020) using the package Rsubread (Liao et al., 2019) and mapped alignments were viewed in Geneious.

## 3. Results

### *3.1 Mycosphaerellales genomes and phylogeny*

Genomes for 58 species in the Mycosphaerellales and two *Cladosporium* (Cladosporiales) genomes were obtained from NCBI and used in this study (Table 1). Based on the BUSCO "Fungi *odb9*" database, genome completeness ranged from 82.6% to 100%, with most genomes (50/60) being >95% complete (Fig. S2). At least 238 (>80%) of the 290 orthologs were identified as single copy in the genomes examined. The only exceptions were *Friedmanniomyces endolithicus*, *F. simplex* and *Hortaea werneckii* in which many (19-94%) of the 290 orthologs appeared to be duplicated. In comparison, at most 3/290 (1%) were duplicated in the remainder of the genomes.

Because the inclusion of paralogs may compromise the accuracy of species phylogenies (Sanderson and Shaffer, 2002), *F. endolithicus* and *H. werneckii* were excluded from the study. Although *F. simplex* was retained in subsequent analyses, the 57 genes duplicated in this species were discarded. Three genomes in which the *MAT1* locus could not be identified (Table 1), i.e. *Mycosphaerella populi*, *Ramularia coccinea* and *Piedraia hortae*, were also excluded. Accordingly, the Mycosphaerellales dataset included 143 single-copy orthologs that were shared among the remaining 53 taxa and two *Cladosporium* species. The best ML tree inferred from this dataset had a final LogLikelihood value of -1920259.91 and bootstrapping converged after 50 replicates. Family level relationships were resolved as expected (Abdollahzadeh et al., 2020) and the Mycosphaerellaceae and Teratosphaeriaceae separated into distinct and well-supported clades (Fig. 1). Ancestral state reconstruction indicated that the Mycosphaerellales ancestor was most likely (>90%) heterothallic, with several transitions to homothallism across the Order (Fig. S3).

**Table 1** Genomes of Mycosphaerellales and Cladosporiales species used in this study and the number of *MAT* fragment sequences identified in each.

| Species[a] | GenBank accession | Family | Mating type[b] | Number of *MAT* fragment sequences[c] | Number of contigs with *MAT* fragment sequences |
|---|---|---|---|---|---|
| **Cladosporiales (outgroup)** | | | | | |
| *Cladosporium cladosporioides* | GCA_002901145.1 | Cladosporiaceae | *MAT1-1* | 1 | 1 |
| *Cladosporium sphaerospermum* | GCA_000261425.2 | Cladosporiaceae | *MAT1-2* | 0 | 0 |
| **Mycosphaerellales** | | | | | |
| *Aeminium ludgeri* | GCA_004216415.1 | Aeminiaceae | *MAT1-1* | 7 | 4 |
| *Dissoconium aciculare* | GCA_010015565.1 | Dissoconiaceae | *MAT1-2* | 0 | 0 |
| *Ramichloridium luteum* | GCA_002368545.1 | Dissoconiaceae | *MAT1-2* | 0 | 0 |
| *Microcyclospora pomicola* | GCA_002786065.1 | Incertae sedis | *MAT1-1* | 0 | 0 |
| *Microcyclospora tardicrescens* | GCA_003012245.1 | Incertae sedis | *MAT1-2* | 0 | 0 |
| *Cercospora berteroae** | GCA_002933655.1 | Mycosphaerellaceae | *MAT1-2* | 20 | 7 |
| *Cercospora beticola** | GCA_002742065.1 | Mycosphaerellaceae | *MAT1-1* | 17 | 6 |
| *Cercospora canescens** | GCA_000347735.1 | Mycosphaerellaceae | *MAT1-2* | 6 | 5 |
| *Cercospora cf. flagellaris** | GCA_005356885.1 | Mycosphaerellaceae | *MAT1-2* | 16 | 7 |
| *Cercospora cf. sigesbeckiae** | GCA_002217505.1 | Mycosphaerellaceae | *MAT1-2* | 16 | 5 |
| *Cercospora kikuchii** | GCA_009193115.1 | Mycosphaerellaceae | *MAT1-1* | 16 | 7 |
| *Cercospora nicotianae** | GCA_002994015.1 | Mycosphaerellaceae | *MAT1-2* | 18 | 8 |
| *Cercospora sojina** | GCA_004299825.1 | Mycosphaerellaceae | *MAT1-1* | 13 | 5 |
| *Cercospora zeae-maydis** | GCA_010093985.1 | Mycosphaerellaceae | *MAT1-1* | 6 | 4 |
| *Cercospora zeina** | GCA_002844615.1 | Mycosphaerellaceae | *MAT1-1* | 8 | 5 |
| *Dothistroma pini** | GCA_002116355.1 | Mycosphaerellaceae | *MAT1-1* | 17 | 10 |
| *Dothistroma septosporum** | GCA_000340195.1 | Mycosphaerellaceae | *MAT1-2* | 17 | 5 |

| | | | | | |
|---|---|---|---|---|---|
| *Exutisphaerella laricina* | GCA_000504385.2 | Mycosphaerellaceae | *MAT1-1* | 2 | 2 |
| *Lecanosticta acicola* | GCA_002441625.1 | Mycosphaerellaceae | *MAT1-1* | 1 | 1 |
| *Microcyclosporella mali* | GCA_002785985.1 | Mycosphaerellaceae | *MAT1-1* | 0 | 0 |
| *Mycosphaerella arachidis* | GCA_001297265.1 | Mycosphaerellaceae | *MAT1-2* | 12 | 8 |
| *Mycosphaerella populi* | GCA_002153405.1 | Mycosphaerellaceae | ? | 0 | 0 |
| *Mycosphaerella sp.* | GCA_000504405.2 | Mycosphaerellaceae | *MAT1-2* | 0 | 0 |
| *Mycosphaerella sp. PB-2012b* | GCA_002116345.1 | Mycosphaerellaceae | *MAT1-2* | 0 | 0 |
| *Mycosphaerelloides madeirae* | GCA_002785995.1 | Mycosphaerellaceae | *MAT1-1/2* | 5 | 3 |
| *Nothophaeocryptopus gaeumannii** | GCA_002116385.1 | Mycosphaerellaceae | *MAT1-1/2* | 5 | 4 |
| *Pallidocercospora crystallina** | GCA_003666085.1 | Mycosphaerellaceae | *MAT1-1/2* | 8 | 6 |
| *Passalora fulva** | GCA_000301015.1 | Mycosphaerellaceae | *MAT1-2* | 24 | 11 |
| *Pseudocercospora eumusae** | GCA_001578235.1 | Mycosphaerellaceae | *MAT1-2* | 19 | 11 |
| *Pseudocercospora fijiensis** | GCF_000340215.1 | Mycosphaerellaceae | *MAT1-1* | 23 | 6 |
| *Pseudocercospora musae** | GCA_001578225.1 | Mycosphaerellaceae | *MAT1-1* | 17 | 11 |
| *Pseudocercospora pini-densiflorae** | GCA_000504365.2 | Mycosphaerellaceae | *MAT1-2* | 23 | 11 |
| *Ramularia coccinea* | GCA_004155245.1 | Mycosphaerellaceae | ? | 0 | 0 |
| *Ramularia collo-cygni* | GCF_900074925.1 | Mycosphaerellaceae | *MAT1-2* | 0 | 0 |
| *Ramularia endophylla* | GCA_002116395.1 | Mycosphaerellaceae | *MAT1-1* | 0 | 0 |
| *Sphaerulina musiva* | GCF_000320565.1 | Mycosphaerellaceae | *MAT1-2* | 0 | 0 |
| *Sphaerulina populicola* | GCA_000291705.1 | Mycosphaerellaceae | *MAT1-2* | 0 | 0 |
| *Zasmidium angulare* | GCA_002786045.1 | Mycosphaerellaceae | *MAT1-2* | 7 | 4 |
| *Zasmidium cellare* | GCA_010093935.1 | Mycosphaerellaceae | *MAT1-2* | 13 | 5 |
| *Zasmidium citri-griseum* | GCA_002786025.1 | Mycosphaerellaceae | *MAT1-2* | 6 | 4 |
| *Zymoseptoria ardabiliae* | GCA_000223765.2 | Mycosphaerellaceae | *MAT1-1* | 0 | 0 |
| *Zymoseptoria brevis* | GCA_000983655.1 | Mycosphaerellaceae | *MAT1-1* | 0 | 0 |
| *Zymoseptoria passerinii* | GCA_000223825.2 | Mycosphaerellaceae | *MAT1-1* | 0 | 0 |
| *Zymoseptoria pseudotritici* | GCA_000223685.2 | Mycosphaerellaceae | *MAT1-2* | 0 | 0 |
| *Zymoseptoria tritici* | GCA_900184105.1 | Mycosphaerellaceae | *MAT1-2* | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| *Acidomyces richmondensis* | GCA_003545705.1 | Teratosphaeriaceae | *MAT1-1* | 0 | 0 |
| *Baudoinia panamericana* | GCA_000338955.1 | Teratosphaeriaceae | *MAT1-2* | 0 | 0 |
| *Friedmanniomyces endolithicus* | GCA_005059855.1 | Teratosphaeriaceae | *MAT1-2* | 0 | 0 |
| *Friedmanniomyces simplex* | GCA_005059865.1 | Teratosphaeriaceae | *MAT1-2* | 0 | 0 |
| *Hortaea thailandica* | GCA_005059885.1 | Teratosphaeriaceae | *MAT1-2* | 0 | 0 |
| *Hortaea werneckii* | GCA_002127715.1 | Teratosphaeriaceae | *MAT1-1* | 0 | 0 |
| *Neohortaea acidophila* | GCA_010093505.1 | Teratosphaeriaceae | *MAT1-1* | 0 | 0 |
| *Piedraia hortae* | GCA_010093745.1 | Teratosphaeriaceae | ? | 0 | 0 |
| *Teratosphaeria destructans** | GCA_004009835.1 | Teratosphaeriaceae | *MAT1-2* | 3 | 2 |
| *Teratosphaeria gauchensis** | GCA_007113925.1 | Teratosphaeriaceae | *MAT1-2* | 17 | 7 |
| *Teratosphaeria nubilosa** | GCA_010093825.1 | Teratosphaeriaceae | *MAT1-1/2* | 7 | 5 |
| *Teratosphaeria pseudoeucalypti** | GCA_013403795.1 | Teratosphaeriaceae | *MAT1-2* | 4 | 3 |
| *Teratosphaeria zuluensis** | GCA_007113905.1 | Teratosphaeriaceae | *MAT1-1* | 16 | 6 |

[a] Asterisks indicate species used for further analysis of *MAT* fragments

[b] Question marks indicate species in which the *MAT1* locus could not be identified with the BLAST database (Supplementary File 1); "*MAT1-1/2*" denotes a homothallic species

[c] The number of distinct fragment sequences as determined with BLASTn (Supplementary File 1)

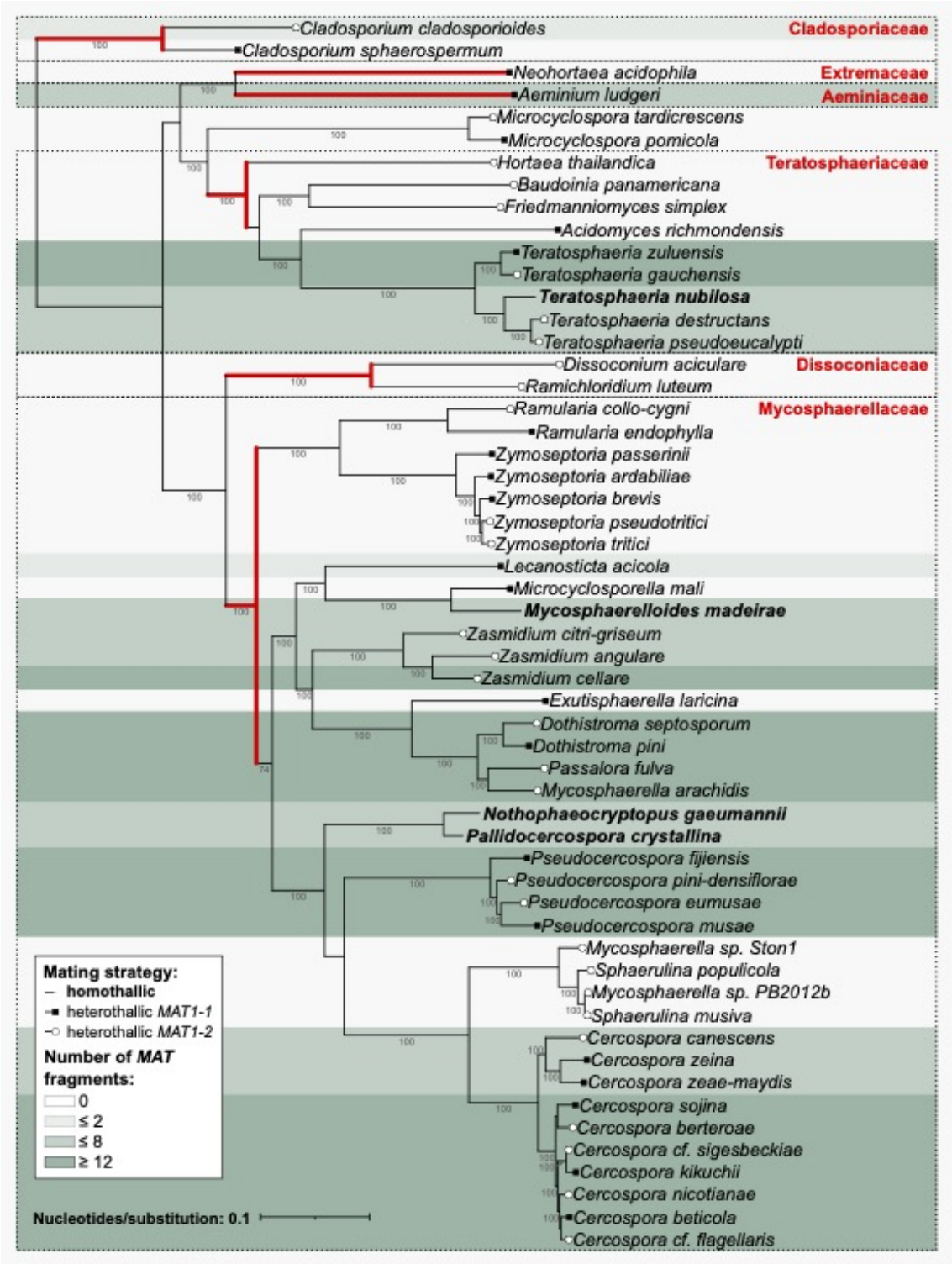



**Figure 1** Phylogenomic tree of Mycosphaerellales based on 143 shared, single-copy orthologs. Branches leading to a family have been highlighted and family names are indicated within dashed boxes. For heterothallic fungi, the shape of the branch tip reflects the mating type, while homothallic species are shown in bold. The colours on the phylogeny indicate presence and abundance of *MAT* fragments identified in the genomes (dark green = ≥ 12 fragment sequences on ≥5 different contigs;

lighter green = ≤ 8 fragment sequences on ≤ 6 contigs; lightest green, *Cladosporium cladosporioides*, *Lecanosticta acicola* and *Exutisphaerella laricina* = ≤ 2 sequences from a single *MAT1* gene only).

### 3.2 *Position, evolution and function of* MAT1-1-10 *and* MAT1-2-12

Using all of the genomes available from genera in which at least one species has a characterised *MAT1* locus, we investigated the placement of genes within and around this region. Overall, the arrangement and orientation of genes was similar within genera (Fig. 2). The *MAT1* idiomorph in the examined *Teratosphaeria* genomes contained either the *MAT1-1-10* or *MAT1-2-12* gene, depending on mating type (i.e., *MAT1-1* contained *MAT1-1-10* and *MAT1-2* contained *MAT1-2*-12). However, both *MAT1-1-10* and *MAT1-2-12* were present in the genomes of all Mycosphaerellaceae, regardless of mating type.

The *MAT1-2-12* gene in *Zymoseptoria* and both *MAT1-1-10* and *MAT1-2-12* in *Cercospora* formed part of the conserved regions flanking the *MAT1* locus of these fungi (Fig. 2). *MAT1-1-10* appeared to be completely absent from the *Zymoseptoria MAT1-2* idiomorph. In *Dothistroma*, *Passalora* and *Pseudocercospora*, the *MAT1-1-10* and *MAT1-2-12* genes formed part of the nonhomologous region between isolates of opposite mating type (i.e., they were part of the *MAT1* locus and not the region flanking it). However, in these three genera the entire *MAT1-2* idiomorph appeared to be in an inverted orientation relative to that of *Cercospora*. In other words, for *Dothistroma*, *Passalora* and *Pseudocercospora* the *MAT1-1-10* and *MAT1-2-12* genes formed the previously reported regions of inverted homology between the *MAT1-1* and *MAT1-2* idiomorphs (Conde-Ferráez et al., 2007).

Of the three homothallic species investigated, the genome of *Pal. crystallina* was the most contiguous and revealed that the *MAT1-1-1* and *MAT1-2-1* genes lie approximately 46 kilobase pairs (kbp) apart on the same contig (Fig. 2). In this genome, the *MAT1-2-1* gene was flanked by *MAT1-1-10* and *MAT1-2-12*, and the whole cluster was located adjacent to the *APC5* and *APN2* genes. Its *MAT1-1-1* counterpart lay downstream and no other genes typically associated with the *MAT1* locus could be detected around it. Gene synteny between *Pal. crystallina* and the other Mycosphaerellaceae species was also not sufficiently conserved to investigate the insertion of *MAT1-1-1* at this position. In the genome assembly of the second homothallic species, *N. gaeumannii*, the *MAT1* genes were identified on two short contigs, with no flanking sequence information available. However, a homolog of *MAT1-1-10* was predicted downstream of both the *MAT1-2-1* and the *MAT1-1-1* genes in this species.

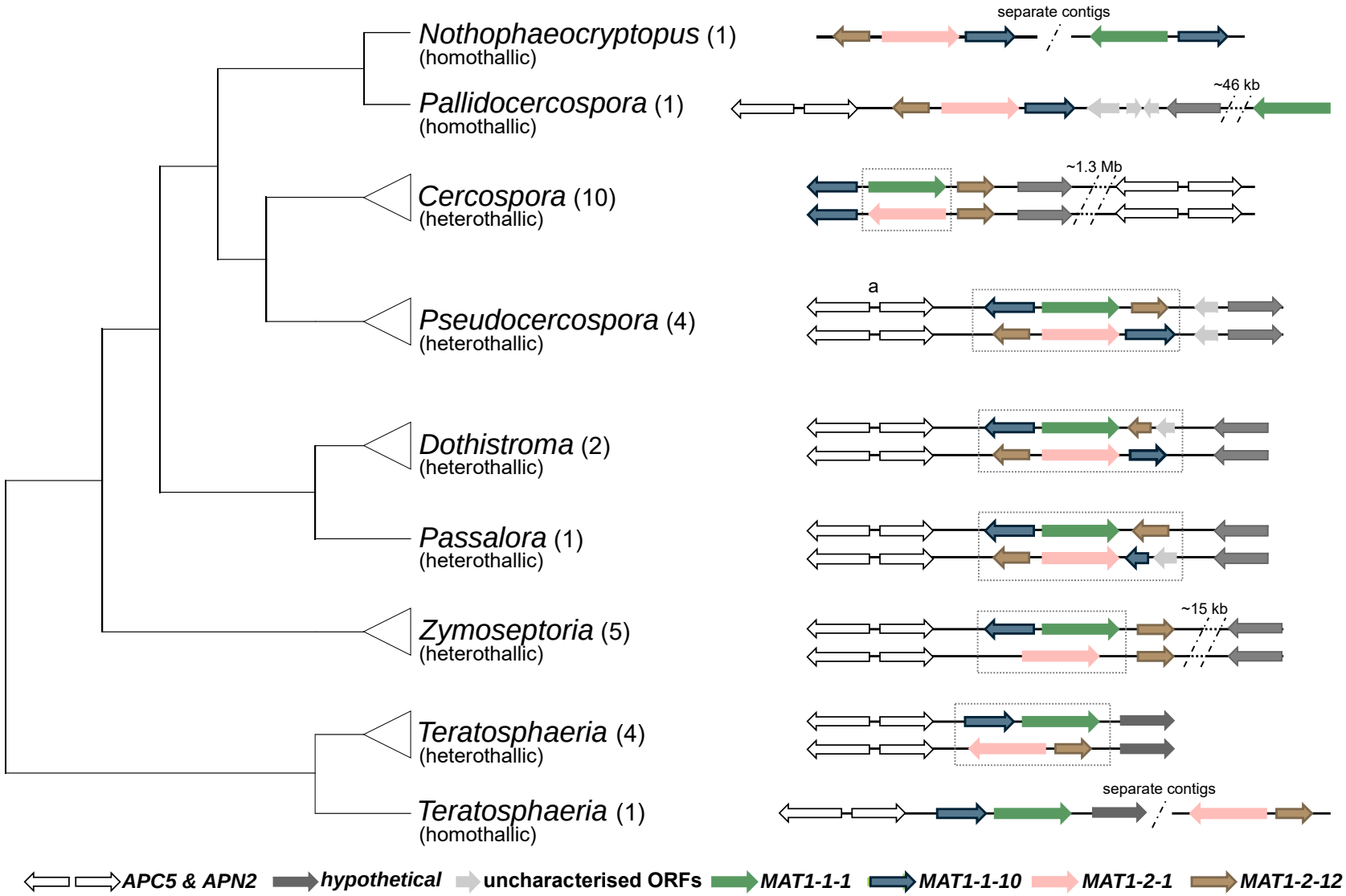

Nothophaeocryptopus (1)
(homothallic)
Pallidocercospora (1)
(homothallic)
Cercospora (10)
(heterothallic)
Pseudocercospora (4)
(heterothallic)
Dothistroma (2)
(heterothallic)
Passalora (1)
(heterothallic)
Zymoseptoria (5)
(heterothallic)
Teratosphaeria (4)
(heterothallic)
Teratosphaeria (1)
(homothallic)
separate contigs
~46 kb
~1.3 Mb
a
~15 kb
separate contigs
APC5 & APN2
hypothetical
uncharacterised ORFs
MAT1-1-1
MAT1-1-10
MAT1-2-1
MAT1-2-12

**Figure 2** Simplified representation of the arrangement of *MAT1* genes in Mycosphaerellaceae and Teratosphaeriaceae species. For each genus, arrows on black lines represent genes and gene directions in the *MAT1-1* (containing the green arrow) and *MAT1-2* idiomorphs (containing the pink arrow). The number in brackets after the genus name shows the number of species used to infer gene presence and order (see Table S2). The relationships among the genera were inferred from Figure 1. Boxes indicate the nonhomologous region between isolates of different mating type. Genes and their relative positions are not drawn to scale. The same hypothetical ORF is present in all genera, whereas uncharacterised ORFs differ. [a]The order of *APC5* and *APN2* is reversed in the *Pseudocercospora eumusae* genome sequence.

In the homothallic *T. nubilosa*, *MAT1* genes were located on two separate contigs, although the gene content and arrangement of these were similar to those of the respective idiomorphs in the heterothallic *Teratosphaeria* species. A region corresponding to the *MAT1-1* idiomorph was identified at the typical position of the *MAT1* locus (i.e., next to *APC5* and *APN2*), which lay at the edge of one contig. A region corresponding to the *MAT1-2* idiomorph occurred 110 kbp and 376 kbp from the respective edges of the other contig, indicating that, should these contigs form part of the same chromosome, the two idiomorphic regions would be distantly separated. The regions flanking the *T. nubilosa MAT1-2* idiomorphic region were also present in the other *Teratosphaeria* species, suggesting insertion of the *T. nubilosa MAT1-2* idiomorph at this position. Nevertheless, within these idiomorphic regions, *MAT1-2-12* was located next to *MAT1-2-1*, and *MAT1-1-10* was located next to *MAT1-1-1*.

ML phylogenetic analysis of the *MAT1-1-10* and *MAT1-2-12* genes grouped genera and closely related species together with high (>70%) statistical support in most cases (Fig. 3). However, in both phylogenies, the *Pseudocercospora* genes grouped according to mating type rather than species identity. A similar trend was also observed among certain *Cercospora* species (especially for *C. beticola* and its close relatives, as well as the closely related *C. zeina* and *C. zea-maydis*) where *MAT1-1* copies of the *MAT1-1-10* and *MAT1-2-12* genes were more closely related to each other than to their *MAT1-2* counterparts. Mating-type-dependent evolution of *MAT1-1-10* and *MAT1-2-12* was thus observed in *Pseudocercospora* and in some species of *Cercospora*.

Ratios ($\omega$) of nonsynonymous to synonymous substitution rates were used to test whether the mating-type-dependent evolution observed in *Pseudocercospora* was due to different selective pressures on the genes in the two *MAT1* idiomorphs. This was achieved by extracting the subtrees containing *Pseudocercospora* (together with *Teratosphaeria* as outgroup) in both the *MAT1-1-10* and *MAT1-2-12* phylogenies. Initially, the analysis was performed with only the three *Pseudocercospora* species for which genes were available for both mating types and was subsequently repeated to include *Ps. pini-densiflorae* (*MAT1-2*). For the *MAT1-2-12* gene, the most likely explanation ($P<0.001$) in both analyses was that significantly different selective forces were acting on the *MAT1-1* and *MAT1-2* copies (Hypothesis 3; Supplementary File 1). Approximately 14% of sites in the *MAT1-1* copies of *MAT1-2-12* showed evidence of diversifying selection ($\omega = 2.8$ and $\omega = 3.1$ for the respective subtrees), whereas the same proportion of sites displayed purifying selection ($\omega = 0.8$ for both subtrees) in the *MAT1-2*

copies, and even more so for the background (*Teratosphaeria*) branches ($\omega$ = 0.05 and $\omega$ = 0.06, respectively). For the *MAT1-1-10* gene, the two analyses indicated that all sites were under neutral or purifying selection, but did not allow statistically meaningful comparisons of selection between the *MAT1-1* and *MAT1-2* copies (Hypotheses 1 and 2; Supplementary File 1). For the one subtree, we obtained $\omega$ = 0 for 50% of sites using *MAT1-1* as the foreground branches, compared to $\omega$ = 0.1 for the background (*MAT1-2* and *Teratosphaeria*) branches (Hypothesis 1; Supplementary File 1). Analysis using the other subtree indicated evidence of purifying selection at 82% of sites with $\omega$ = 0.1 for the *MAT1-2* foreground branches versus $\omega$ = 0.2 for the background (*MAT1-1* and *Teratosphaeria*) branches (Hypothesis 2; Supplementary File 1).

None of our attempts to infer the possible biological roles of the identified *MAT1-1-10* and *MAT1-2-12* sequences revealed any functional identity, known conserved domains or similarity to non-coding RNAs. However, a comparison with the Gene Ontology knowledgebase (http://geneontology.org/) identified the *T. gauchensis* and *T. zuluensis* MAT1-2-12 proteins as "DNA binding" (GO: 0003677), which is consistent with a function as mating-type transcription factors (Debuchy et al., 2010). Despite the lack of known conserved domains, amino acid sequence alignments revealed conservation across specific residues in each protein (Fig. S4). ConSeq classified 9-14% of the amino acid residues in the two proteins as highly conserved and predicted that 8.3% of MAT1-1-10 and 6.3% of MAT1-2-12 residues have a functional or structural role (Supplementary File 2; Fig. S4). Three of these putative functional MAT1-2-12 residues were predicted to be under diversifying selection in the *Pseudocercospora MAT1-1* copies of *MAT1-2-12* (Fig. S4b). With BLASTp and hmmsearch, an additional five putative homologs of the MAT1-1-10 and 14 of the MAT1-2-12 proteins were identified from Mycosphaerellales species in GenBank, mostly from Teratosphaeriaceae and Mycosphaerellaceae (Fig. S5). All putative homologs were gene predictions derived from whole genome sequences and were characterised in GenBank as unknown or hypothetical proteins.

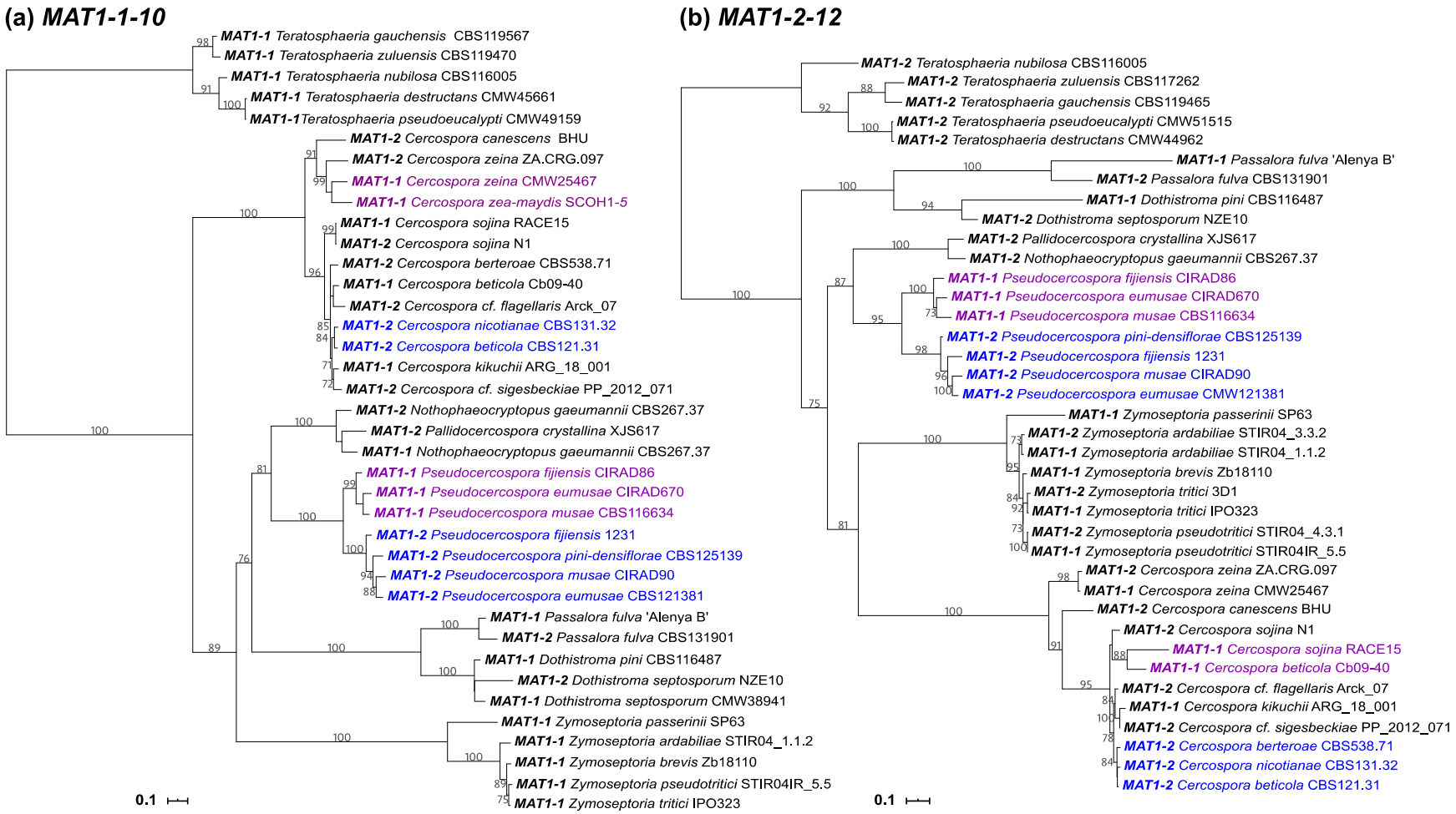

(a) MAT1-1-10
MAT1-1 Teratosphaeria gauchensis CBS119567
MAT1-1 Teratosphaeria zuluensis CBS119470
MAT1-1 Teratosphaeria nubilosa CBS116005
MAT1-1 Teratosphaeria destructans CMW45661
MAT1-1 Teratosphaeria pseudoeucalypti CMW49159
MAT1-2 Cercospora canescens BHU
MAT1-2 Cercospora zeina ZA.CRG.097
MAT1-1 Cercospora zeina CMW25467
MAT1-1 Cercospora zea-maydis SCOH1-5
MAT1-1 Cercospora sojina RACE15
MAT1-2 Cercospora sojina N1
MAT1-2 Cercospora berteroae CBS538.71
MAT1-1 Cercospora beticola Cb09-40
MAT1-2 Cercospora cf. flagellaris Arck_07
MAT1-2 Cercospora nicotianae CBS131.32
MAT1-2 Cercospora beticola CBS121.31
MAT1-1 Cercospora kikuchii ARG_18_001
MAT1-2 Cercospora cf. sigesbeckiae PP_2012_071
MAT1-2 Nothophaeocryptopus gaeumannii CBS267.37
MAT1-2 Pallidocercospora crystallina XJS617
MAT1-1 Nothophaeocryptopus gaeumannii CBS267.37
MAT1-1 Pseudocercospora fijiensis CIRAD86
MAT1-1 Pseudocercospora eumusae CIRAD670
MAT1-1 Pseudocercospora musae CBS116634
MAT1-2 Pseudocercospora fijiensis 1231
MAT1-2 Pseudocercospora pini-densiflorae CBS125139
MAT1-2 Pseudocercospora musae CIRAD90
MAT1-2 Pseudocercospora eumusae CBS121381
MAT1-1 Passalora fulva 'Alenya B'
MAT1-2 Passalora fulva CBS131901
MAT1-1 Dothistroma pini CBS116487
MAT1-2 Dothistroma septosporum NZE10
MAT1-1 Dothistroma septosporum CMW38941
MAT1-1 Zymoseptoria passerinii SP63
MAT1-1 Zymoseptoria ardabiliae STIR04_1.1.2
MAT1-1 Zymoseptoria brevis Zb18110
MAT1-1 Zymoseptoria pseudotritici STIR04IR_5.5
MAT1-1 Zymoseptoria tritici IPO323
0.1
(b) MAT1-2-12
MAT1-2 Teratosphaeria nubilosa CBS116005
MAT1-2 Teratosphaeria zuluensis CBS117262
MAT1-2 Teratosphaeria gauchensis CBS119465
MAT1-2 Teratosphaeria pseudoeucalypti CMW51515
MAT1-2 Teratosphaeria destructans CMW44962
MAT1-1 Passalora fulva 'Alenya B'
MAT1-2 Passalora fulva CBS131901
MAT1-1 Dothistroma pini CBS116487
MAT1-2 Dothistroma septosporum NZE10
MAT1-2 Pallidocercospora crystallina XJS617
MAT1-2 Nothophaeocryptopus gaeumannii CBS267.37
MAT1-1 Pseudocercospora fijiensis CIRAD86
MAT1-1 Pseudocercospora eumusae CIRAD670
MAT1-1 Pseudocercospora musae CBS116634
MAT1-2 Pseudocercospora pini-densiflorae CBS125139
MAT1-2 Pseudocercospora fijiensis 1231
MAT1-2 Pseudocercospora musae CIRAD90
MAT1-2 Pseudocercospora eumusae CMW121381
MAT1-1 Zymoseptoria passerinii SP63
MAT1-2 Zymoseptoria ardabiliae STIR04_3.3.2
MAT1-1 Zymoseptoria ardabiliae STIR04_1.1.2
MAT1-1 Zymoseptoria brevis Zb18110
MAT1-2 Zymoseptoria tritici 3D1
MAT1-1 Zymoseptoria tritici IPO323
MAT1-2 Zymoseptoria pseudotritici STIR04_4.3.1
MAT1-1 Zymoseptoria pseudotritici STIR04IR_5.5
MAT1-2 Cercospora zeina ZA.CRG.097
MAT1-1 Cercospora zeina CMW25467
MAT1-2 Cercospora canescens BHU
MAT1-2 Cercospora sojina N1
MAT1-1 Cercospora sojina RACE15
MAT1-1 Cercospora beticola Cb09-40
MAT1-2 Cercospora cf. flagellaris Arck_07
MAT1-1 Cercospora kikuchii ARG_18_001
MAT1-2 Cercospora cf. sigesbeckiae PP_2012_071
MAT1-2 Cercospora berteroae CBS538.71
MAT1-2 Cercospora nicotianae CBS131.32
MAT1-2 Cercospora beticola CBS121.31
0.1

**Figure 3** Maximum likelihood phylogenies of the (a) *MAT1-1-10* and (b) *MAT1-1-12* coding sequences. *MAT1-1* and *MAT1-2* indicates the idiomorph in which the copy of *MAT1-1-10* or *MAT1-2-12* was identified. Purple and blue denote *MAT1-1* and *MAT1-2* genes, respectively, that clustered together based on mating type rather than species identity. The trees were rooted on the branch leading to the *Teratosphaeria* genes.

### *3.3 Characterisation of* MAT *fragments*

Genomes with significant BLASTn hits to *MAT1* genes in genomic locations other than the full-length *MAT1* locus were considered to harbour *MAT* fragment sequences. No similarities to the *MAT1-1-10* and *MAT1-2-12* genes were identified outside of the *MAT1* locus, but fragment sequences of both the *MAT1-1-1* and *MAT1-2-1* genes occurred in many species. The sizes of these fragment sequences ranged from 28 to 445 bp (average 115.4 bp), with the majority (> 90%) being in the range of 60 to 210 bp (Supplementary File 3). Although short, most (75%) of these fragments had a nucleotide identity of > 90% (mean identity = 94.2%) to *MAT1-1-1* or *MAT1-2-1* genes and were, therefore, near identical to portions of these coding regions.

Overall, the *MAT* fragment sequences were identified in various members of the Mycosphaerellaceae and Teratosphaeriaceae, in the single representative of the Aeminiaceae (*Aeminium ludgeri*) and in the outgroup taxon, *Cladosporium cladosporioides* (Fig. 1; Table 1). In the Teratosphaeriaceae, fragment sequences were identified only in species of *Teratosphaeria*. In the Mycosphaerellaceae, fragments were absent from *Microcyclosporella mali* and the clades containing *Ramularia*, *Sphaerulina* and *Zymoseptoria* species.

When present, the abundance of *MAT* fragment sequences varied among genomes. Three species had fragments derived only from one of the *MAT1* genes, namely one *MAT1-1-1* fragment in each of *Cladosporium cladosporioides* and *Lecanosticta acicola*, and two *MAT1-2-1* fragments in *Exutisphaerella laricina*. In the remaining species, multiple fragments of both the *MAT1-1-1* and *MAT1-2-1* genes were identified in the same genome, even though most species were heterothallic. Based on their distribution, two broad categories of *MAT* fragment abundance were identified. Some species had ≥ 12 *MAT* fragments and these were typically found on > 5 contigs. Other species had ≤ 8 fragments, typically distributed across < 6 contigs (Fig. 1). Both the Mycosphaerellaceae and Teratosphaeriaceae had members in both categories, but closely related taxa mostly fell into the same category and, therefore, had similar abundances of *MAT* fragments. Similar abundances in closely related taxa, in combination with corroborating results from chromosome or near-chromosome level and less contiguous assemblies (Table S1), suggested that genome assembly quality did not significantly affect the identification of *MAT* gene fragments.

When aligned to the full-length *MAT1* genes, the tiling pattern of the *MAT* fragments resembled that of a spliced mRNA transcript (Fig. 4 and S5). Almost all of the *MAT* fragment sequences were homologous to exonic regions of the *MAT1* genes, with alignment gaps occupying positions around introns. Also, the fragment sequences often extended into the non-coding

regions directly upstream and downstream of the *MAT1* genes, which are reminiscent of 5' and 3' untranslated regions of mRNA transcripts. Accordingly, the process of rebuilding or reconstructing the respective genes from the identified *MAT* fragments was relatively straightforward.

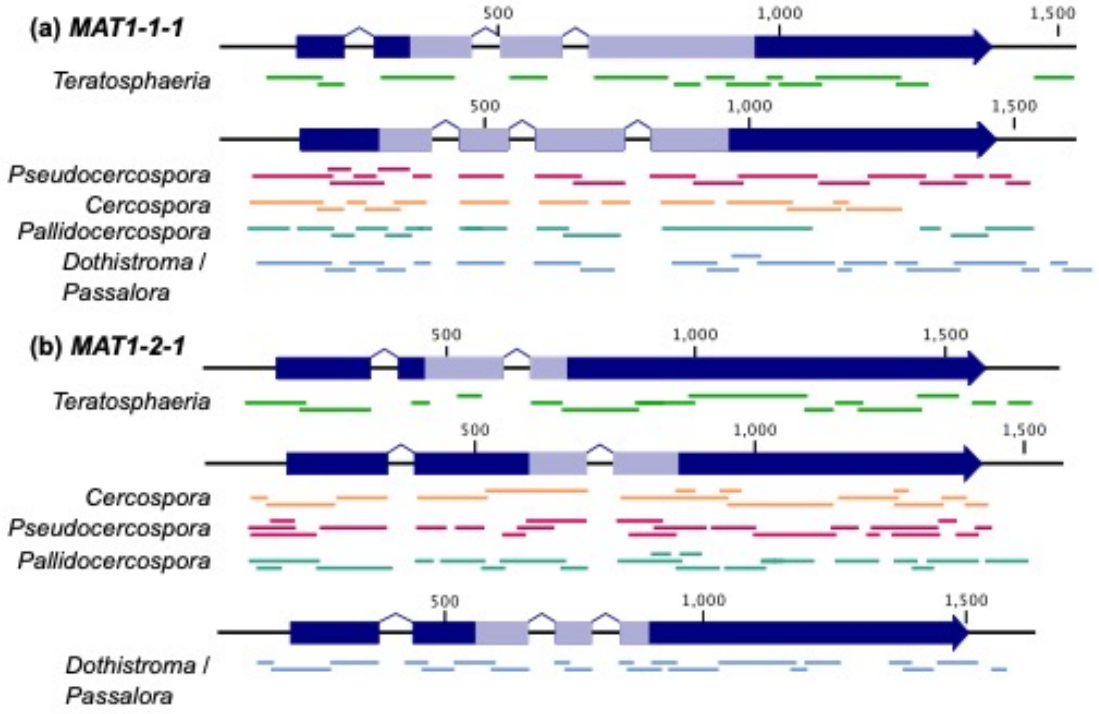



**Figure 4** *MAT1-1-1* (a) and *MAT1-2-1* (b) genes showing a summary of the BLASTn alignments of *MAT* fragment sequences identified in the genomes of Mycosphaerellales species with characterised *MAT1* loci and homothallic representatives. Dark blue arrows represent *MAT* coding sequences. Light regions on the arrows indicate the position of the conserved MAT alpha_HMG box domain (IPR006856) in *MAT1-1-1* and the HMG-box domain (IPR036910) in *MAT1-2-1*. Both *MAT1-1* and *MAT1-2* fragments originate from the same genome. Fragment sequences and alignments are summarised together for closely related species and Figure S5 provides a more detailed version.

For the genomes considered in this analysis, fragment sequences of both mating types were present in the same genome and, therefore, each genome yielded a reconstructed counterpart for *MAT1-1-1* and *MAT1-2-1*. These reconstructed sequences covered most (> 80 %, Supplementary File 4) of the exons of their corresponding *MAT1-1-1* and *MAT1-2-1* genes. Exceptions were the alignments of three *Teratosphaeria* species (*T. destructans*, *T. nubilosa* and *T. pseudoeucalypti*; Fig. S6), in which large portions of the *MAT1* genes were not covered by fragment sequences. Two other available *T. destructans* (GenBank accessions WBMN01 and WBMM01) and *T. pseudoeucalypti* genomes (JABBMY01 and JABASB01) confirmed that the low coverage was not a result of missing genome data (also see Fig. S7a). Even though

all the reconstructed sequences included a start codon, none formed a complete coding sequence, because all had nonsense mutations and most had at least one large gap relative to its corresponding *MAT1* gene.

The high similarity shared between *MAT* fragments and their corresponding *MAT1* genes was apparent in our phylogenetic analyses. The phylogeny of the reconstructed consensus sequences generally mirrored that of the respective *MAT1* genes (Fig. 5). Within closely related taxon groups, the reconstructed consensus sequences grouped together, but separate from the full-length *MAT1* genes of their "hosts". However, between closely related taxon groups, the reconstructed sequences mostly grouped with their host genes. In other words, the *MAT* fragment sequences were, firstly, more similar to fragment sequences of closely related species and, secondly, to their own *MAT1* genes, than to *MAT* fragment sequences in other species.

The *MAT* fragments occupied seemingly random locations in the genomes of the studied species. Fragments were found alone, as well as in close proximity to one another. When found together, the fragments were often a mixture of *MAT1-1-1* and *MAT1-2-1*-derived sequences (Fig. S7). In closely related species, BLASTn searches that included regions flanking the areas in which *MAT* fragments were found, indicated that homologous *MAT* fragments usually occupied the same genomic locations (e.g. Fig. S8). However, this positional conservation was only observed among closely related species and groups of species. In cases where BLAST identified certain *MAT* fragments in some species, but not in their close relatives, the homologous region was often present within the other genomes, but appeared to be degraded due to the presence of single nucleotide polymorphisms (SNPs) and indels (e.g. *Teratosphaeria*, see Fig. S9).

Although ORFs predicted next to the fragment-containing regions were most often similar among closely related species, ORF prediction within the fragment-containing regions was not consistent (Fig. S8). Some of the genes predicted in these regions contained significant similarity to *MAT1-1-1* and *MAT1-2-1* genes, however, the similarity of these ORFs to actual *MAT1* genes was limited to the small area containing the *MAT* fragment sequence. The majority of ORFs that overlapped *MAT* fragment regions did not have a hit in the NCBI non-redundant protein database or were classified as hypothetical proteins (Supplementary File 5). Although some of these predictions may be true ORFs, they do not represent true *MAT1* genes and the prediction of many of these ORFs appeared to be a result of the *MAT* fragment sequences "forcing" a prediction in the region (e.g. Fig. S8).

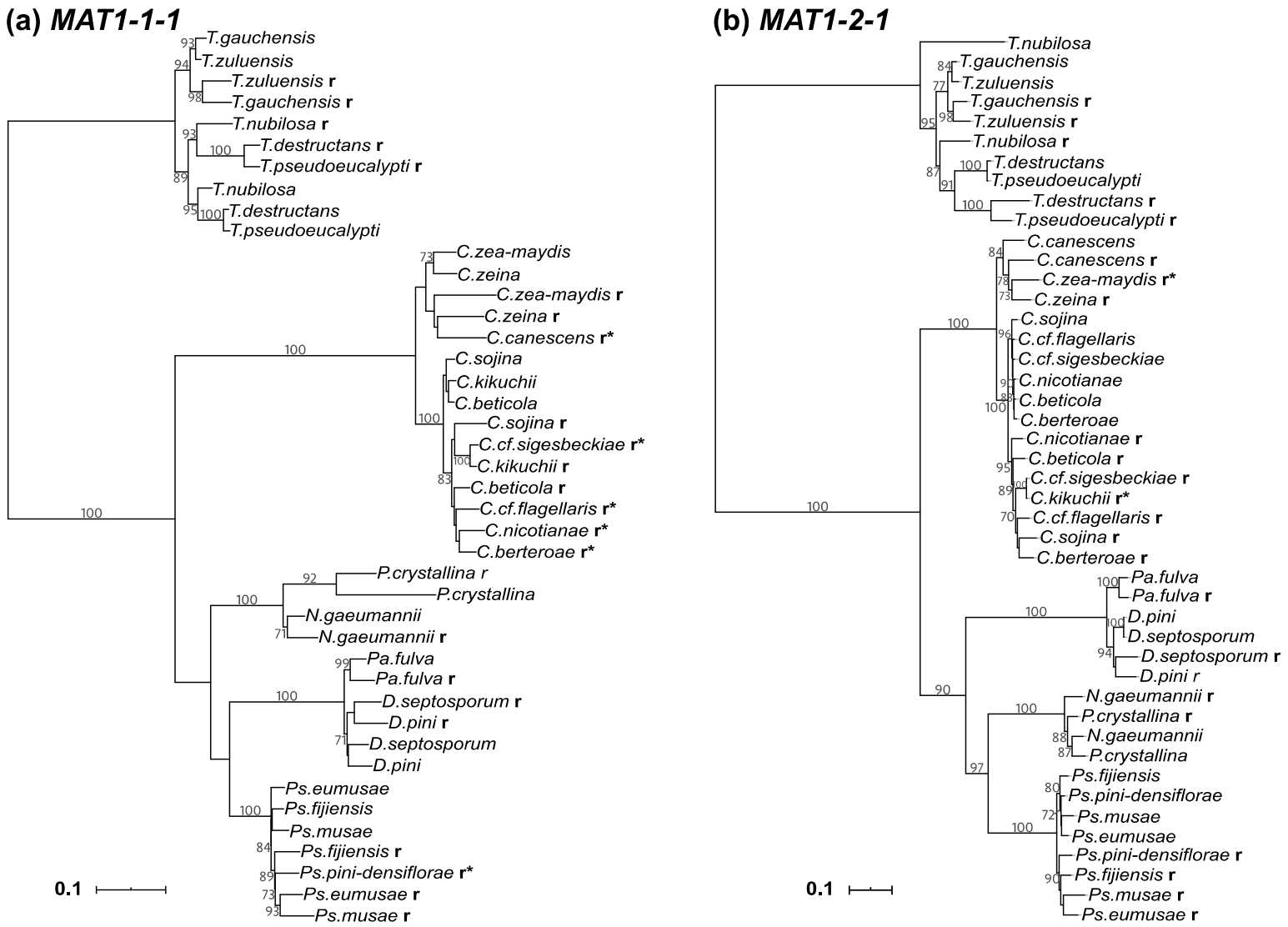

(a) MAT1-1-1
T.gauchensis
T.zuluensis
T.zuluensis r
T.gauchensis r
T.nubilosa r
T.destructans r
T.pseudoeucalypti r
T.nubilosa
T.destructans
T.pseudoeucalypti
C.zea-maydis
C.zeina
C.zea-maydis r
C.zeina r
C.canescens r*
C.sojina
C.kikuchii
C.beticola
C.sojina r
C.cf.sigesbeckiae r*
C.kikuchii r
C.beticola r
C.cf.flagellaris r*
C.nicotianae r*
C.berteroae r*
P.crystallina r
P.crystallina
N.gaeumannii
N.gaeumannii r
Pa.fulva
Pa.fulva r
D.septosporum r
D.pini r
D.septosporum
D.pini
Ps.eumusae
Ps.fijiensis
Ps.musae
Ps.fijiensis r
Ps.pini-densiflorae r*
Ps.eumusae r
Ps.musae r
0.1
(b) MAT1-2-1
T.nubilosa
T.gauchensis
T.zuluensis
T.gauchensis r
T.zuluensis r
T.nubilosa r
T.destructans
T.pseudoeucalypti
T.destructans r
T.pseudoeucalypti r
C.canescens
C.canescens r
C.zea-maydis r*
C.zeina r
C.sojina
C.cf.flagellaris
C.cf.sigesbeckiae
C.nicotianae
C.beticola
C.berteroae
C.nicotianae r
C.beticola r
C.cf.sigesbeckiae r
C.kikuchii r*
C.cf.flagellaris r
C.sojina r
C.berteroae r
Pa.fulva
Pa.fulva r
D.pini
D.septosporum
D.septosporum r
D.pini r
N.gaeumannii r
P.crystallina r
N.gaeumannii
P.crystallina
Ps.fijiensis
Ps.pini-densiflorae
Ps.musae
Ps.eumusae
Ps.pini-densiflorae r
Ps.fijiensis r
Ps.musae r
Ps.eumusae r
0.1

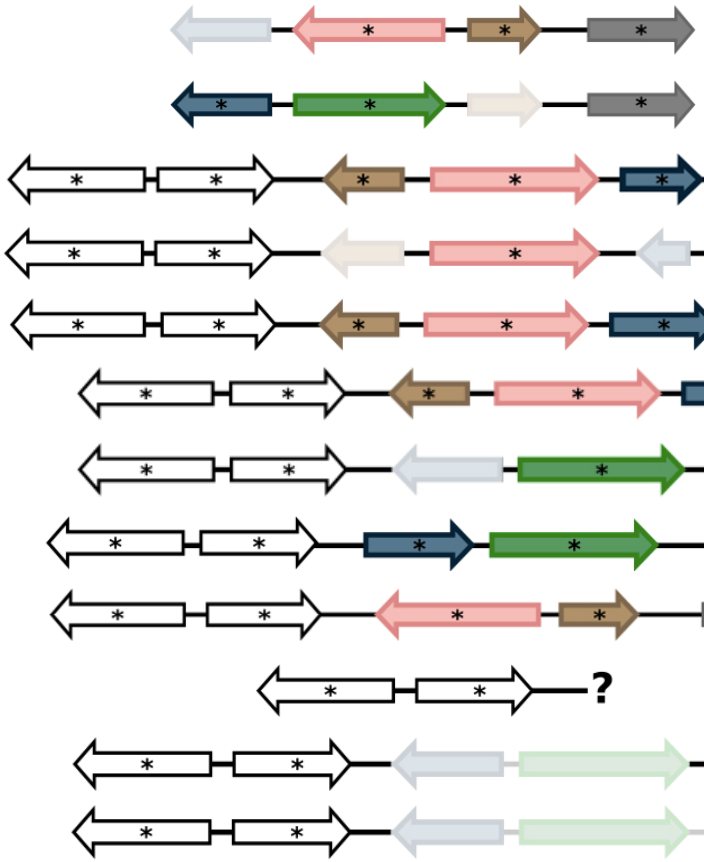
**Figure 5** Maximum likelihood phylogeny depicting the relationships among the sequences of the *MAT1-1-1* (a) and *MAT1-2-1* (b) genes and the sequences reconstructed (r) from the *MAT* fragments. For each species, reconstructed sequences in both trees were derived from the same genome. Asterisks indicate genes that were reconstructed based on the *MAT1* gene of a closely related species, since sequence information for the specific *MAT1* gene was not available. Trees were rooted using the *Teratosphaeria* sequences.

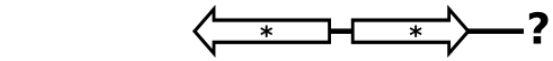
### *3.4 Expression of* MAT1 *genes and* MAT *fragment sequences*

Between 19 and 253 million quality-filtered RNA reads were available from the different species and the average read length ranged from 49 to 147 bp (Table S3). Mapping the RNA-Seq reads of each strain against its corresponding *MAT1* idiomorph sequence, indicated that the primary *MAT1* genes (*MAT1-1-1* or *MAT1-2-1*) were expressed in all *Cercospora*, *Dothistroma*, *Passalora*, *Pseudocercospora* and *Teratosphaeria* strains investigated (Table 2; Fig. 6a; Supplementary File 6). In all of these strains, except for *Pa. fulva*, expression of at least one accessory *MAT1* gene (*MAT1-1-10* or *MAT1-2-12*) was also detected. The RNA-Seq data for the *Zymoseptoria* species was sufficient to detect expression of genes flanking the *MAT1* locus, but the only *MAT1* gene for which expression could be detected was *MAT1-2-12* in *Z. brevis*.

The mapped RNA-Seq alignments of each strain against its own, as well as the opposite *MAT1* idiomorph sequence, suggested additional expression of the *MAT* fragment sequences (Table 2). This was apparent because the sequences of *MAT* fragments are not 100% identical to the *MAT1* genes (e.g. Supplementary Table 7-2), with SNPs that occur between the *MAT1* genes and RNA-Seq reads confirming that certain reads were a result of the expression of *MAT* fragments. This was subsequently verified by mapping the RNA-Seq data to the *MAT* fragment regions (see Supplementary File 7 for detailed analysis of *T. destructans* data). While the reads of each strain that mapped to the primary *MAT1* gene in the opposite idiomorph (i.e. the idiomorph absent from that strain) displayed high nucleotide identity to regions of the *MAT1* gene, SNPs and indels were consistently present and corresponded to the sequence of the *MAT* fragments (Fig. 6b and c). In *T. destructans*, for which transcriptome data from both *MAT1-1* and *MAT1-2* strains were available, we identified evidence for the expression of *MAT* fragments in both mating types (Table 2; Supplementary File 7). Additionally, expression appeared to be independent of mating type because *MAT* fragments derived from *MAT1-1-1* and *MAT1-2-1* were expressed in both isolates.

**Table 2** Expression of genes in and around the *MAT1* locus and of *MAT* fragment sequences, detected in RNA-Seq data

| Species | Isolate | *MAT*[a] | Genes expressed in and around the *MAT1* locus[b] | *MAT* fragments detected | |
|---|---|---|---|---|---|
| | | | | *MAT1-1-1* | *MAT1-2-1* |
| *Cercospora sojinae* | RACE1 | *MAT1-2** |  | 6/11 | 2/12 |
| *C. zeina* | CMW25467 | *MAT1-1* | | 2/11 | 5/12 |
| *Dothistroma septosporum* | NZE10 | *MAT1-2* | | 10/18 | 3/10 |
| *Passalora fulva* | CBS131901 | *MAT1-2* | | 8/19 | 5/24 |
| *Pseudocercospora eumusae* | CBS114824 | *MAT1-2* | | 13/18 | 8/19 |
| *Ps. fijiensis* | 14H1-11A | *MAT1-2** | | 8/12 | 7/30 |
| *Ps. musae* | CBS116634 | *MAT1-1* | | 10/18 | 11/22 |
| *Teratosphaeria destructans* | CMW45661 | *MAT1-1* | | 6/8 | 6/6 |
| *T. destructans* | CMW44962 | *MAT1-2* | | 8/8 | 6/6 |
| *Zymoseptoria ardabiliae* | STIR04_1.1.1 | ? | | 0/0 | 0/0 |
| *Z. brevis* | Zb18110 | *MAT1-1* | | 0/0 | 0/0 |
| *Z. passerini* | SP63 | *MAT1-1* | | 0/0 | 0/0 |

| | | | | | |
|---|---|---|---|---|---|
| *Z. pseudotritici* | STIR04_2.2.1 | ? |  | 0/0 | 0/0 |
| *Z. tritici* | IPO323 | *MAT1-1* |  | 0/0 | 0/0 |

[a] The mating type of isolates was known either from published *MAT1* idiomorph or genome sequences or inferred from the RNA-Seq mapping data (indicated with an asterisk). Expression of *MAT1* genes in *Z. ardabiliae* and *Z. pseudotritici* was insufficient to assign a putative mating type.

[b] Gene colours and placement follow that described in Fig. 2. Expressed genes are indicated with an asterisk, faded genes were not expressed. Some flanking genes were not evaluated because they were not directly adjacent to the *MAT1* locus (i.e. *APN2* and *APC5* in *Cercospora* and the hypothetical protein in *Zymoseptoria*) or not were not present in the *MAT1* idiomorph sequence used (i.e. *Pseudocercospora*).


**Figure 6** Alignment of the *Teratosphaeria destructans MAT1-1* and *MAT1-2* RNA-Seq reads against its *MAT1-1* and *MAT1-2* idiomorphs. (a) Mapping the reads of each isolate against its corresponding *MAT1* idiomorph confirmed the expression of both genes in each idiomorph. (b) Mapping these reads against the opposite idiomorph yielded a significant number of alignments in areas of the *MAT1-1-1* and *MAT1-2-1* genes that have homologous *MAT* fragments occurring elsewhere in the genome. (c) Analysis of SNPs (highlighted in blue) in these alignments showed that the RNA-Seq reads originate from the expression of *MAT* fragments (See Supplementary File 7 for further information).

## 4. Discussion

### *4.1 Two putative Mycosphaerellales-specific* MAT *genes*

This study showed that the genes *MAT1-1-10* and *MAT1-2-12*, initially identified as *MATORF1* and *MATORF2*, respectively (Arzanlou et al., 2010), consistently occur at the *MAT1* locus of the Teratosphaeriaceae and Mycosphaerellaceae. These genes were also identified in other species in which they had not previously been predicted, such as those of *Dothistroma* (Groenewald et al., 2007), and we showed that both have homologs in the genomes of Mycosphaerellaceae available on NCBI. In some instances, *MAT1-1-10* was either missing (*Zymoseptoria*) or truncated (*Passalora*) in the *MAT1-2* idiomorph of heterothallic species. Similarly, *MAT1-2-12* was truncated in the *MAT1-1* idiomorph of *Dothistroma*. Both genes were, therefore, present in the genomes of most heterothallic Mycosphaerellaceae, but only *MAT1-1-10* was consistently associated with the *MAT1-1* idiomorph and *MAT1-2-12* with the

*MAT1-2* idiomorph. This pattern is consistent with the *Teratosphaeria* species that encode only one of these genes per *MAT1* idiomorph (Aylward et al., 2020; Havenga et al., 2020), both of which we found were expressed in *T. destructans*. It is also consistent with the previously reported mating-type-dependent expression of these genes in *Ps. fijiensis* (Arzanlou et al., 2010). However, in the present study, mating-type-independent expression of these genes was also detected in *Dothistroma* and *Pseudocercospora*. These data may suggest that the *MAT1-2* copies of *MAT1-1-10* and the *MAT1-1* copies of *MAT1-2-12* are non-essential components of these respective mating types.

The *MAT1-1* and *MAT1-2* paralogues of *MAT1-1-10* and *MAT1-2-12* displayed mating-type-dependent evolution in *Pseudocercospora* and, to some extent, *Cercospora*. In the case of *Pseudocercospora*, Arzanlou et al. (2010) previously suggested that this pattern might be due to balancing selection that maintains trans-species polymorphism (Richman, 2000). This would also be consistent with our finding that different selective pressures act on *MAT1-2-12* in different idiomorphs. Purifying selection was detected in *MAT1-2* copies of *Pseudocercospora MAT1-2-12* genes, as opposed to the strongly diversifying selection in *MAT1-1* copies. In this case, the assumption would be that both copies are functional, likely with regards to the recognition of sexual compatibility, and that their polymorphism is advantageous (Richman, 2000). This would also imply that both are expressed, as found in this study. Alternatively, the divergence between *MAT1-1* and *MAT1-2* copies could be explained by mating-associated chromosomal regions being recombination "coldspots" with high rates of sequence evolution (Hsueh et al., 2006; Wik et al., 2008). In this case, the purifying selection on *MAT1-2* copies of *MAT1-2-12* supports the notion that it may be the only copy essential to the *MAT1-2* mating type. Clearly further work is needed to clarify the role of these Mycosphaerellales-specific *MAT1* genes in the biology of these fungi.

### *4.2 Short* MAT1 *coding sequences in some Mycosphaerellales genomes*

Our study confirmed that seemingly random fragment sequences of the *MAT1-1-1* and *MAT1-2-1* coding regions are present at multiple locations in the genomes of Mycosphaerellaceae and Teratosphaeriaceae species. The presence of *MAT* fragment sequences was not a ubiquitous or monophyletic phenomenon and their frequency varied among species. Bolton et al. (2014) and Arzanlou et al. (2010) previously identified short genomic sequences in *C. beticola* and *Ps. fijiensis*, respectively, that are homologous to portions of *MAT1* genes. The only other example of which we are aware and where parts of a gene occurs in different genomic locations is through a phenomenon known as *trans*-splicing. This is found in some plants, animals and protists where different exons encoded in separate regions of the genome are spliced together after transcription (Bonen, 1993). The *MAT* fragments, however, differ considerably from those associated with *trans*-splicing as they do not form a complete coding sequence and have full-length *MAT1-1-1* and *MAT1-2-1* gene representatives within the relevant genomes.

The fact that the *MAT* fragments could be assembled into sequences resembling mRNA transcripts may hint at how they came to be present in Mycosphaerellales genomes. Long interspersed nuclear elements (LINEs) are a class of retrotransposons known to cause *trans* effects by mobilizing other mRNA sequences that are not derived from LINEs (Esnault et al.,

2000). In this way, LINEs can reverse transcribe a mature mRNA molecule of a gene and cause it to be integrated into the genome, forming a processed pseudogene (Dewannieux and Heidmann, 2005). LINEs comprise a considerable portion of repeats (5-33%) in the genomes of *Pseudocercospora* species (Chang et al., 2016) and *Pa. fulva* (De Wit et al., 2012). However, they appear to be absent in *D. septosporum* (De Wit et al., 2012) and have not been investigated in most other Mycosphaerellales. The topic of LINE-mediated mobilization in Mycosphaerellales genomes would justify an independent study, but we hypothesise that a similar mechanism was responsible for integrating the *MAT* fragment sequences into these genomes.

If a process such as retrotransposon-mediated integration gave rise to the *MAT* fragment sequences, the question arises as to whether this process has taken place multiple times or in a common ancestor of Mycosphaerellales species. The presence of homologous *MAT* fragments at the same position in all genomes would support a common origin, but could not be established due to the mesosynteny in the genomes of even closely related Dothideomycetes (De Wit et al., 2012; Ohm et al., 2012). However, *MAT* fragment sequences at the same genomic positions within genera, together with their higher interspecific than intraspecific sequence similarity (Bolton et al., 2014), provide strong evidence for a common origin of these sequences for each monophyletic group in which they occur. The combined occurrence and close proximity of *MAT1-1-1* and *MAT1-2-1* fragment sequences in the genomes is unexpected for heterothallic species and further suggests that such an event would have taken place once, rather than multiple times. The availability of additional Mycosphaerellales genomes, especially for groups underrepresented in this study, may resolve this question in the future.

The expression of the *MAT* fragment sequences in several Mycosphaerellaceae species and *T. destructans*, as well as their widespread occurrence in Mycosphaerellales genomes suggests that these sequences have some function. One possibility is that the *MAT* fragments function as regulators of expression. For example, their expression yielded RNA molecules that, though not identical, are highly similar to the *MAT1-1-1* and *MAT1-2-1* genes and could potentially bind to their mRNA. Such binding capacity would be reminiscent of small RNAs (sRNAs) in eukaryotes that, when bound to multiprotein complexes, are able to identify and silence complementary mRNA sequences (Moazed, 2009). In this eukaryotic system, the sRNAs are derived from longer, double stranded RNA that is either transcribed within the cell or introduced from the outside (Almeida and Allshire, 2005). It is possible that RNA molecules expressed from the *MAT* fragments silence the expression of *MAT1-1-1* and *MAT1-2-1*, which are essential for onset of the sexual cycle (Ni et al., 2011). Although sexual reproduction is prominent in some species with *MAT* fragment sequences, such as *Ps. fijiensis* (Carlier et al., 1996), species without known sexual morphs, or with only rare occurrences of these morphs, are common in the Mycosphaerellales (Groenewald et al., 2006; Quaedvlieg et al., 2014). Therefore, the apparent absence of sexual reproduction in many Mycosphaerellales may represent a mystery that could be explained by such a suppressive mechanism. A well-designed RNA experiment that investigates potential interactions between the *MAT1* gene transcripts and *MAT* fragments could unravel the potential functional role of *MAT* fragments.

Whatever the function of *MAT* fragments, if indeed they have a function, it remains unclear how they retain the high level of similarity to corresponding regions of the resident *MAT1-1-1* and *MAT1-2-1* genes. Bolton et al. (2014) suggested that they are maintained through concerted evolution, a process that homogenises paralogous gene regions. If this is the case, a form of homologous recombination known as gene conversion (Chen et al., 2007) could be responsible for homogenizing the *MAT* fragment sequences with the *MAT1* genes. Gene conversion transfers DNA sequence information in one direction between similar sequences, e.g., from the *MAT1* genes to the *MAT* fragments, and a small number of events is necessary for effective homogenization (Liao, 1999). Although it has been reported during mitosis and meiosis, gene conversion during meiosis is more common and, congruent with the size of *MAT* fragments, tends to act on shorter sequences (Chen et al., 2007). In heterothallic species, *MAT1-1-1* and *MAT1-2-1* are only present together during the dikaryotic phase of the meiotic cycle, making it the only time at which both *MAT1-1* and *MAT1-2* fragments sequences could be homogenized. If *MAT* fragments are actively maintained in such a manner, it would provide further evidence of a biological role for these sequences.

### *4.3 Ancestral heterothallism*

As with other Ascomycota (Butler, 2007; Ni et al., 2011), the sexual mating strategies employed by the Mycosphaerellales are reflected by the architecture of their *MAT1* locus (see Fig. 2 and 7). In heterothallic Mycosphaerellales, the locus always had a primary *MAT1* gene (*MAT1-1-1* or *MAT1-2-1*) accompanied by at least one accessory gene (*MAT1-1-10* and/or *MAT1-2-12*). In some heterothallic Mycosphaerellaceae with *MAT1-1-10* and *MAT1-2-12* genes in both idiomorphs, the *MAT1-2* idiomorph had an inverted orientation compared to other species. This inversion is likely to have played a role in the evolution of the *MAT1-1-10* and *MAT1-2-12* genes, as it caused them to become part of the *MAT1* locus instead of the conserved flanking sequences. Also, the homothallic species examined harbour nonadjacent idiomorphic regions that correspond to the two mating types. Homothallic species with nonadjacent *MAT* idiomorphs are rare among Ascomycetes, with only few known examples of species with "unlinked" *MAT* loci, such as *Aspergillus nidulans*, *Cochliobolus cymbopogonis*, and *Neosartorya fischeri* (Lin and Heitman, 2007).

Heterothallism currently appears to be the dominant reproductive mode as well as the most likely ancestral mode of reproduction in the Mycosphaerellales, with only four homothallic species identified in this study. Beyond these genome sequences, additional examples of homothallic Mycosphaerellaceae and Teratosphaeriaceae species are known, e.g. *Paramycosphaerella marksii* and *Suberoteratosphaeria xenosuberosa* (Quaedvlieg et al., 2014), but they appear to be in the minority. Ancestral heterothallism (e.g. Fig. 7a) would be similar to the situation in *Cochliobolus* (Dothideomycetes: Pleosporales) where it is the most likely explanation for the prevalence of heterothallic species and the varied architecture of the *MAT1* locus of homothallic species (Lin and Heitman, 2007; Yun et al., 1999). Given a heterothallic ancestral state, *MAT* fragments could have been integrated into the genome of an early Mycosphaerellales species, mediated by LINEs as described above. To incorporate

fragments from both mating types, this would have needed to take place during the dikaryotic phase.

The fragment sequences identified in this study hinted towards a homothallic Mycosphaerellales ancestor. This is because fragment sequences of both *MAT1-1-1* and *MAT1-2-1* occurred in heterothallic and homothallic species and fragments of the 3' terminal of *MAT1-1-1* and 5' terminal of *MAT1-2-1* were found together (see Fig. S8; Bolton et al., 2014). This suggests either a close association between the two genes or a fused ancestral *MAT1* gene (Arzanlou et al., 2010; Bolton et al., 2014). Homothallism is, however, often a less parsimonious and, therefore, less popular explanation for the ancestral state of fungi (Li et al., 2020; Lin and Heitman, 2007). *Aspergillus* was previously regarded as a genus of predominantly homothallic species in which the genes flanking the *MAT1* locus and evidence for gene loss in heterothallic species implied ancestral homothallism (Butler, 2007; Lin and Heitman, 2007). However, genomic analyses have increased the number of known heterothallic aspergilli, producing models that favour ancestral heterothallism (Ojeda-López et al., 2018).

Based on the arrangement of genes in the *MAT1* locus and the presence of *MAT* fragments from both mating-type genes, both a heterothallic and homothallic ancestral state is plausible for Mycosphaerellales species (see Fig. 7). Conde-Ferráez et al. (2007) favoured a homothallic ancestor, suggesting that the inverted *MAT1-2* idiomorph result from an inversion event that separated the two idiomorphic regions in a homothallic ancestor. Such a means of transitioning from homothallism to heterothallism, however, does not explain the structure of the *MAT1* locus of *Cercospora* and *Teratosphaeria*. Extant homothallic Mycosphaerellales likely evolved from heterothallic ancestors because contemporary homothallic species have apparently unlinked *MAT* idiomorphic regions. Furthermore, the idiomorphic region that occurs downstream of the *APN2* and *APC5* genes differs between homothallic Teratosphaeriaceae and Mycosphaerellaceae, giving credence to the view that heterothallism is the ancestral state.

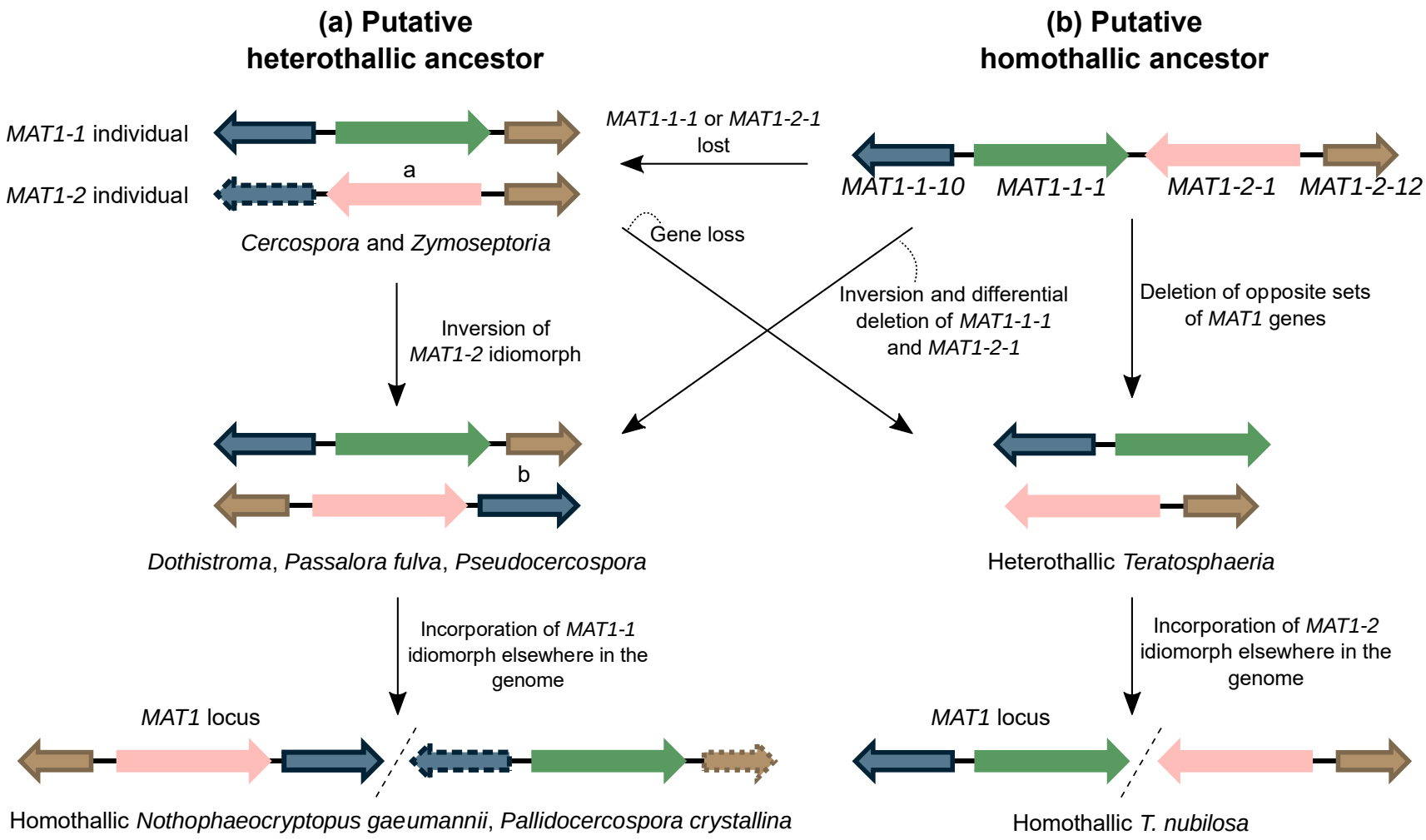

(a) Putative heterothallic ancestor
(b) Putative homothallic ancestor
MAT1-1 individual
MAT1-2 individual
a
Cercospora and Zymoseptoria
MAT1-1-1 or MAT1-2-1 lost
Gene loss
MAT1-1-10
MAT1-1-1
MAT1-2-1
MAT1-2-12
Inversion of MAT1-2 idiomorph
Inversion and differential deletion of MAT1-1-1 and MAT1-2-1
Deletion of opposite sets of MAT1 genes
b
Dothistroma, Passalora fulva, Pseudocercospora
Heterothallic Teratosphaeria
Incorporation of MAT1-1 idiomorph elsewhere in the genome
Incorporation of MAT1-2 idiomorph elsewhere in the genome
MAT1 locus
MAT1 locus
Homothallic Nothophaeocryptopus gaeumannii, Pallidocercospora crystallina
Homothallic T. nubilosa

**Figure 7** Proposed evolutionary pathways of the Mycosphaerellales *MAT1* locus. A putative heterothallic ancestor (a) would require two evolutionary pathways to produce the *MAT1* locus architectures observed today, whereas a putative homothallic ancestor containing both *MAT1-1* and *MAT1-2* genes at the *MAT1* locus (b) may have used two or three different pathways. In both cases, a transition from heterothallism back to homothallism likely produced the contemporary homothallic arrangements. Dotted arrows represent genes that may be absent in some species; the dashed lined between the idiomorphic regions of the homothallic species indicate that the regions are not adjacent to each other. [a]The direction of *MAT1-2-1* has changed in *Zymoseptoria*. [b]The direction of *MAT1-2-12* has changed in some species and *MAT1-2-12* or *MAT1-1-10* may be truncated (see Fig. 2).

## 5. Conclusions

Despite the economic importance of numerous species in the Teratosphaeriaceae and Mycosphaerellaceae, the *MAT1* locus has been studied in only a few taxa. Moreover, the species that have been investigated to date display novel features that have not previously been explored. Firstly, investigation of *MAT1-1-10* and *MAT1-2-12* revealed that these genes are likely consistently present within Teratosphaeriaceae and Mycosphaerellaceae species and that rearrangements of the *MAT1* locus has affected their evolution. Secondly, fragments of the *MAT1-1-1* and *MAT1-2-1* genes occur in apparently arbitrary areas of the genomes of many members in these families and are expressed in at least four genera. We hypothesise that these *MAT* fragment sequences were integrated into the genome in a similar way that processed pseudogenes are created and that they may be involved in regulating the activity of the *MAT1* genes. This study has characterised the *MAT1-1-10* and *MAT1-2-12* genes and the *MAT* fragments and further experiments are required to identify their biological role within Mycosphaerellales species.

## Acknowledgements

We are grateful to Prof. Dave Berger and Ms. Tanya Welgemoed for supplying the sequence of the *Cercospora zeina MAT1-2* idiomorph and to Dr. Andi Wilson for insightful comments on the manuscript. This work was supported by members of the Tree Protection Co-operative program (TPCP), the Department of Science and Innovation (DSI) - National Research Foundation (NRF) Centre of Excellence in Plant Health Biotechnology (CPHB), the SARChI chair in Fungal Genomics and the University of Pretoria.

## Data availability

The sequences of all *MAT1* idiomorphs used in this study, as well as the RNA-Seq dataset of *Teratosphaeria destructans*, are available through the National Centre for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/genbank/); see Supplementary Tables S2 and S3. Additions or changes to the gene annotations of sequences already present on NCBI are supplied here as Supplementary Data.

## Declaration of Competing Interest

The authors declare that they have no competing interests.

## References


Abdollahzadeh, J., Groenewald, J.Z., Coetzee, M.P.A., Wingfield, M.J., Crous, P.W., 2020. Evolution of lifestyles in Capnodiales. Stud. Mycol. 95, 381-414.

Almeida, R., Allshire, R.C., 2005. RNA silencing and genome regulation. Trends in Cell Biol. 15, 251-258.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Arzanlou, M., Crous, P.W., Zwiers, L.-H., 2010. Evolutionary dynamics of mating-type loci of *Mycosphaerella* spp. occurring on banana. Eukaryot. Cell 9, 164-172.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., Ben-Tal, N., 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 44, W344-W350.

Aylward, J., Havenga, M., Dreyer, L.L., Roets, F., Wingfield, B.D., Wingfield, M.J., 2020. Genomic characterization of mating type loci and mating type distribution in two apparently asexual plantation tree pathogens. Plant Pathol. 69, 28-37.

Aylward, J., Roets, F., Dreyer, L.L., Wingfield, M.J., 2019. Teratosphaeria stem canker of *Eucalyptus*: two pathogens, one devastating disease. Mol. Plant Pathol. 20, 8-19.

Aylward, J., Steenkamp, E.T., Dreyer, L.L., Roets, F., Wingfield, B.D., Wingfield, M.J., 2017. A plant pathology perspective of fungal genome sequencing. IMA Fungus 8, 1.

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., Casadio, R., Ben-Tal, N., 2004. ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinform. 20, 1322-1324.

Bielawski, J.P., Yang, Z., 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. 59, 121-132.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinform. 30, 2114-2120.

Bolton, M.D., de Jonge, R., Inderbitzin, P., Liu, Z., Birla, K., Van de Peer, Y., Subbarao, K.V., Thomma, B.P.H.J., Secor, G.A., 2014. The heterothallic sugarbeet pathogen *Cercospora beticola* contains exon fragments of both *MAT* genes that are homogenized by concerted evolution. Fungal Genet. Biol. 62, 43-54.

Bonen, L., 1993. Trans-splicing of pre-mRNA in plants, animals, and protists. FASEB J. 7, 40-46.

Bulman, L., Bradshaw, R., Fraser, S., Martín-García, J., Barnes, I., Musolin, D., La Porta, N., Woods, A., Diez, J., Koltay, A., 2016. A worldwide perspective on the management and control of Dothistroma needle blight. For. Pathol. 46, 472-488.

Butler, G., 2007. The Evolution of *MAT*: the Ascomycetes. In: Heitman, J., Kronstad, J.W., Taylor, J.W., Casselton, L.A. (Eds.), Sex in fungi: molecular determination and evolutionary implications. ASM Press, Washington, DC, pp. 1-18.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinform. 25, 1972-1973.

Carlier, J., Lebrun, M.-H., Zapater, M.-F., Dubois, C., Mourichon, X., 1996. Genetic structure of the global population of banana black leaf streak fungus, *Mycosphaerella fijiensis*. Mol. Ecol. 5, 499-510.

Chang, T.-C., Salvucci, A., Crous, P.W., Stergiopoulos, I., 2016. Comparative genomics of the Sigatoka disease complex on banana suggests a link between parallel evolutionary changes in *Pseudocercospora fijiensis* and *Pseudocercospora eumusae* and increased virulence on the banana host. PLoS Genet. 12, e1005904.

Chen, J.-M., Cooper, D.N., Chuzhanova, N., Férec, C., Patrinos, G.P., 2007. Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8, 762-775.

Conde-Ferráez, L., Waalwijk, C., Canto-Canché, B.B., Kema, G.H.J., Crous, P.W., James, A.C., Abeln, E.C.A., 2007. Isolation and characterization of the mating type locus of *Mycosphaerella fijiensis*, the causal agent of black leaf streak disease of banana. Mol. Plant Pathol. 8, 111-120.

Coppin, E., Debuchy, R., Arnaise, S., Picard, M., 1997. Mating types and sexual development in filamentous ascomycetes. Microbiolology and Mololecular Biolology Reviews 61, 411-428.

Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10, 210.

Crous, P.W., Schoch, C., Hyde, K., Wood, A., Gueidan, C., De Hoog, G., Groenewald, J., 2009. Phylogenetic lineages in the Capnodiales. Stud. Mycol. 64, 17-47.

Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., Flouri, T., 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. 37, 291-294.

Darzentas, N., 2010. Circoletto: visualizing sequence similarity with Circos. Bioinform. 26, 2620-2621.

De Wit, P.J.G.M., Van Der Burgt, A., Ökmen, B., Stergiopoulos, I., Abd-Elsalam, K.A., Aerts, A.L., Bahkali, A.H., Beenen, H.G., Chettri, P., Cox, M.P., 2012. The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. PLoS Genet. 8, e1003088.

Debuchy, R., Berteaux-Lecellier, V., Silar, P., 2010. Mating systems and sexual morphogenesis in ascomycetes. In: Borkovich, K.A., Ebbole, D.J. (Eds.), Cellular and Molecular Biology of Filamentous Fungi. ASM Press, Washington, DC, pp. 501-535.

Dewannieux, M., Heidmann, T., 2005. LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. Cytogenet. Genome Res. 110, 35-48.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinform. 29, 15-21.

Esnault, C., Maestre, J., Heidmann, T., 2000. Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. 24, 363-367.

Gao, F., Chen, C., Arab, D.A., Du, Z., He, Y., Ho, S.Y., 2019. EasyCodeML: A visual tool for analysis of selection using CodeML. Ecol. Evol. 9, 3891-3898.

Goodwin, S.B., Waalwijk, C., Kema, G.H.J., Cavaletto, J.R., Zhang, G., 2003. Cloning and analysis of the mating-type idiomorphs from the barley pathogen *Septoria passerinii*. Mol. Genet. Genom. 269, 1-12.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420-3435.

Groenewald, M., Barnes, I., Bradshaw, R.E., Brown, A.V., Dale, A., Groenewald, J.Z., Lewis, K.J., Wingfield, B.D., Wingfield, M.J., Crous, P.W., 2007. Characterization and distribution of mating type genes in the *Dothistroma* needle blight pathogens. Phytopathol. 97, 825-834.

Groenewald, M., Groenewald, J.Z., Harrington, T.C., Abeln, E.C.A., Crous, P.W., 2006. Mating type gene analysis in apparently asexual *Cercospora* species is suggestive of cryptic sex. Fungal Genet. Biol. 43, 813-825.

Havenga, M., Wingfield, B.D., Wingfield, M.J., Dreyer, L.L., Roets, F., Aylward, J., 2021. Genetic response to nitrogen starvation in the aggressive *Eucalyptus* foliar pathogen *Teratosphaeria destructans*. Current Genetics 67, 981-990.

Havenga, M., Wingfield, B.D., Wingfield, M.J., Roets, F., Dreyer, L.L., Tatham, C.T., Duong, T.A., Wilken, P.M., Chen, S., Aylward, J., 2020. Mating strategy and mating type distribution in six global populations of the *Eucalyptus* foliar pathogen *Teratosphaeria destructans*. Fungal Genet. Biol. 137, 103350.

Haynes, W., 2013. Bonferroni Correction. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (Eds.), Encyclopedia of Systems Biology. Springer New York, New York, NY, pp. 154-154.

Hsueh, Y.-P., Idnurm, A., Heitman, J., 2006. Recombination hotspots flank the *Cryptococcus* mating-type locus: implications for the evolution of a fungal sex chromosome. PLoS Genet. 2, e184.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., Petrov, A.I., 2017. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46, D335-D342.

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

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinform. 35, 4453–4455.

Kronstad, J.W., Staben, C., 1997. Mating type in filamentous fungi. Annu. Rev. Genet. 31, 245-276.

Kück, P., Longo, G.C., 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Frontiers in Zoology 11, 81.

Lefort, V., Longueville, J.-E., Gascuel, O., 2017. SMS: Smart model selection in PhyML. Mol. Biol. Evol. 34, 2422-2424.

Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., Gascuel, O., 2019. NGPhylogeny. fr: new generation phylogenetic services for non-specialists. Nucleic Acids Research 47, W260–W265.

Lemoine, F., Entfellner, J.-B.D., Wilkinson, E., Correia, D., Felipe, M.D., Oliveira, T.d., Gascuel, O., 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nat. 556, 452.

Letunic, I., Bork, P., 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47, W256-W259.

Li, J.Q., Wingfield, B.D., Wingfield, M.J., Barnes, I., Fourie, A., Crous, P.W., Chen, S.F., 2020. Mating genes in *Calonectria* and evidence for a heterothallic ancestral state. Persoonia 45, 163-176.

Liao, D., 1999. Concerted evolution: molecular mechanism and biological implications. Am. J. Hum. Genet. 64, 24.

Liao, Y., Smyth, G.K., Shi, W., 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 47, e47-e47.

Lin, X., Heitman, J., 2007. Mechanisms of homothallism in fungi and transitions between heterothallism and homothallism. In: Heitman, J., Kronstad, J.W., Taylor, J.W., Casselton, L.A. (Eds.), Sex in fungi: molecular determination and evolutionary implications. ASM Press, Washington, DC, pp. 35-57.

Marin, D.H., Romero, R.A., Guzmán, M., Sutton, T.B., 2003. Black Sigatoka: an increasing threat to banana cultivation. Plant Dis. 87, 208-222.

McDonald, B.A., Linde, C., 2002. Pathogen Population Genetics, Evolutionary Potential, and Durable Resistance. Annu. Rev. Phytopathol. 40, 349-379.

Moazed, D., 2009. Small RNAs in transcriptional gene silencing and genome defence. Nat. 457, 413-420.

Ni, M., Feretzaki, M., Sun, S., Wang, X., Heitman, J., 2011. Sex in fungi. Annu. Rev. Genet. 45, 405-430.

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., 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLoS Pathog. 8, e1003037.

Ojeda-López, M., Chen, W., Eagle, C., Gutiérrez, G., Jia, W., Swilaiman, S., Huang, Z., Park, H.-S., Yu, J.-H., Cánovas, D., 2018. Evolution of asexual and sexual reproduction in the aspergilli. Stud. Mycol. 91, 37-59.

Paradis, E., Schliep, K., 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinform. 35, 526-528.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., Stamatakis, A., 2010. How many bootstrap replicates are necessary? J. Comput. Biol. 17, 337-354.

Quaedvlieg, W., Binder, M., Groenewald, J.Z., Summerell, B.A., Carnegie, A.J., Burgess, T.I., Crous, P.W., 2014. Introducing the Consolidated Species Concept to resolve species in the *Teratosphaeriaceae*. Persoonia 33, 1-40.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Ranwez, V., Douzery, E.J., Cambon, C., Chantret, N., Delsuc, F., 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Mol. Biol. Evol. 35, 2582-2584.

Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3, 217-223.

Richman, A., 2000. Evolution of balanced genetic polymorphism. Mol. Ecol. 9, 1953-1963.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nat. 425, 798-804.

Sanderson, M.J., Shaffer, H.B., 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. 33, 49-72.

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. Bioinform. 31, 3210-3212.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B., 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34, W435-W439.

Stergiopoulos, I., Groenewald, M., Staats, M., Lindhout, P., Crous, P.W., De Wit, P.J.G.M., 2007. Mating-type genes and the genetic structure of a world-wide collection of the tomato pathogen *Cladosporium fulvum*. Fungal Genet. Biol. 44, 415-429.

Torriani, S.F.F., Melichar, J.P.E., Mills, C., Pain, N., Sierotzki, H., Courbot, M., 2015. *Zymoseptoria tritici*: a major threat to wheat production, integrated approaches to control. Fungal Genet. Biol. 79, 8-12.

Turgeon, B.G., Yoder, O., 2000. Proposed nomenclature for mating type genes of filamentous ascomycetes. Fungal Genet. Biol. 31, 1-5.

Videira, S., Groenewald, J., Nakashima, C., Braun, U., Barreto, R.W., de Wit, P.J., Crous, P., 2017. Mycosphaerellaceae–Chaos or clarity? Stud. Mycol. 87, 257-421.

Waalwijk, C., Mendes, O., Verstappen, E.C.P., de Waard, M.A., Kema, G.H.J., 2002. Isolation and characterization of the mating-type idiomorphs from the wheat septoria leaf blotch fungus *Mycosphaerella graminicola*. Fungal Genet. Biol. 35, 277-286.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35, 543-548.

Weadick, C.J., Chang, B.S., 2012. An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. Mol. Biol. Evol. 29, 1297-1300.

Weiland, J., Koch, G., 2004. Sugarbeet leaf spot disease (*Cercospora beticola* Sacc.). Mol. Plant Pathol. 5, 157-166.

Wik, L., Karlsson, M., Johannesson, H., 2008. The evolutionary trajectory of the mating-type (*mat*) genes in *Neurospora* relates to reproductive behavior of taxa. BMC Evol. Biol. 8, 109.

Wilken, P.M., Steenkamp, E.T., van der Nest, M.A., Wingfield, M.J., De Beer, Z.W., Wingfield, B.D., 2018. Unexpected placement of the *MAT1-1-2* gene in the *MAT1-2 i*diomorph of *Thielaviopsis*. Fungal Genet. Biol. 113, 32-41.

Wilken, P.M., Steenkamp, E.T., Wingfield, M.J., De Beer, Z.W., Wingfield, B.D., 2017. Which *MAT* gene? Pezizomycotina (Ascomycota) mating-type gene nomenclature reconsidered. Fungal Biology Reviews 31, 199-211.

Wilson, A.M., Wilken, P.M., van der Nest, M.A., Steenkamp, E.T., Wingfield, M.J., Wingfield, B.D., 2015. Homothallism: an umbrella term for describing diverse sexual behaviours. IMA Fungus 6, 207-214.

Wilson, A.M., Wilken, P.M., Wingfield, M.J., Wingfield, B.D., 2021. Genetic Networks That Govern Sexual Reproduction in the Pezizomycotina. Microbiol. Mol. Biol. Rev. 85, e00020-00021.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586-1591.

Yun, S.-H., Berbee, M.L., Yoder, O., Turgeon, B.G., 1999. Evolution of the fungal self-fertile reproductive life style from self-sterile ancestors. Proc. Natl. Acad. Sci. 96, 5592-5597.