communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-08024-9

Phylogenomics and adaptive evolution of the *Colletotrichum gloeosporioides* species complex

Ziying Ma[1,2,6], Fang Liu[1,2,6], Clement K. M. Tsui[3,4,5] & Lei Cai[1,2]

The *Colletotrichum gloeosporioides* species complex (CGSC) is one of the most devastating fungal phytopathogens, and is composed of three main clades: Kahawae, Musae, and Theobromicola. Despite the diversity of CGSC, there is limited understanding on their evolutionary mechanisms. By analysing 49 newly assembled genomes, we found that the expansion of transposable elements, especially long terminal repeat retrotransposons, facilitates the expansion of genome size and genetic variation. In-depth analyses suggested that an intra-chromosomal inversion may have been the driving force behind the divergence of Kahawae clade from its ancestor. Within the Kahawae clade, the narrow-hosted quarantine species *C. kahawae* has undergone extensive chromosomal rearrangements mediated by repetitive sequences, generating highly dynamic lineage-specific genomic regions compared to the closely related broad-hosted species *C. cigarro*. The findings of this study highlight the role of chromosomal rearrangements in promoting genetic diversification and host adaptation, and provide new perspectives for understanding the evolution of phytopathogenic fungi.

Colletotrichum species are severe phytopathogenic fungi that can infect hundreds of plant species[1], and they cause destructive anthracnose diseases on agricultural and horticultural crops, including red rot, crown and stem rot, seeding blights, and brown blotch[2,3], as well as post-harvest diseases of diverse fruit[4–8]. In addition, several representatives are opportunistic pathogens to humans and animals[2,9].

Colletotrichum has been divided into 16 complexes and 15 singletons based on morphological features and rRNA phylogenetic analysis, and the *C. gloeosporides* species complex (CGSC) is the most species-rich complex with a broad host range[9,10]. Fifty-six species have been accepted in CGSC and they can be categorized to three major phylogenetic clades, i.e., Musae, Kahawae, and Theobromicola[9–11]. Most species in CGSC are polyphagous. For example, *C. gloeosporioides*, *C. fructicola*, and *C. siamense* infect hundreds of different crop species worldwide. In contrast, several members exhibit higher levels of specificity towards particular host plants, including *C. kahawae* and *C. musae*[11], which only infect *Coffea arabica* and *Musa* spp., respectively. Despite the variation in host range and diversity, the genomic architecture and molecular basis of host specificity, as well as the evolutionary trajectories between specialists and generalists recognized in CGSC, remain poorly characterized. Recent comparative genomic studies suggested that lineage-specific (LS) losses of pathogenicity genes and the contraction of gene families may account for the narrow host range in *Colletotrichum*[12–14]. However, most species included in these investigations were distantly related, and belonged to different *Colletotrichum* species complexes or even genera[12,15]. Thus, this apparent genomic variation may not reflect the ongoing evolutionary pathway among closely related species in CGSC.

C. kahawae is a highly destructive pathogen in CGSC that triggers Coffee Berry Disease in Africa[16–18] where the infection causes up to 80% harvest loss if no control measures are adopted[19]. *C. kahawae* also pose a serious threat to other major coffee-producing countries such as Brazil, Colombia, Guatemala, Honduras, India, Indonesia, Perú, and Vietnam[20], thus has been listed as a quarantine pathogen in Australia and China[21,22]. Evolutionary analyzes based on six nuclear loci indicated that the rapid and recent speciation of *C. kahawae* was driven by a host jump event to coffee berries from a generalist group of seemingly harmless relatives[23]. *C. kahawae* shares a most common ancestor with *C. cigarro* (corresponding to "UG1")[17,23]. Significant enrichment of non-synonymous mutations in *C. kahawae* and related species, including *C. cigarro*, was detected based on RAD-seq analyzes, and four distinctive candidate genes were associated with pathogenicity for coffee berries[24]. In addition, flow cytometry analysis revealed that the genome size of *C. kahawae* has expanded compared to that

[1]State Key Laboratory of Microbial Diversity and Innovative Utilization, Institute of Microbiology, Chinese Academy of Sciences, Beijing, PR China. [2]University of Chinese Academy of Sciences, Beijing, PR China. [3]Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada. [4]Infectious Disease Research Laboratory, National Centre for Infectious Diseases, Singapore, Singapore. [5]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [6]These authors contributed equally: Ziying Ma, Fang Liu. e-mail: cail@im.ac.cn

of *C. cigarro*[19,25]. However, how the host jump event affected genome evolution and speciation in *C. kahawae* is largely unclear. Insight into the evolutionary processes associated with speciation is key to an understanding of the emergence of novel pathogenicity properties in CGSC and has directly translatable relevance to the epidemiology and potentially the disease control.

Advancements in genome sequencing and assembly technologies have revealed that the *Colletotrichum* genomes contain varying numbers of repetitive-rich and rapidly evolving minichromosomes[26–29], which may arise through mechanisms such as segmental duplications, chromosome rearrangements and core chromosome recombination[29,30], and they are thought to contribute to the pathogen's adaptation to its environment or host plants. Previous studies have shown that large-scale chromosome rearrangements and the formation of minichromosomes may play a key role in enhancing genetic diversity and genomic plasticity in the sexually reproducing species *C. fructicola*[30,31]. However, the genome-wide evolutionary processes among asexual and host-specific CGSCs, such as *C. camelliae*, *C. kahawae*, and *C. musae*[2,11,23,32], remain largely unknown. In contrast to sexual reproduction, strictly asexual organisms are thought to be evolutionarily less flexible, relying mainly on random mutations to adapt to environmental changes, and are often considered as evolutionary dead ends[33,34]. Nevertheless, recent studies have indicated that chromosomal rearrangements may lead to reproductive isolation and act as a general mechanism for adaptation in organisms that propagate asexually, including *C. destructivum*, *C. higginsianum*[29,35], *Verticillium dahliae,* and *Fusarium oxysporum*[36–41].

Determining patterns of variation at the genomic and chromosomal level is crucial for revealing the evolutionary genetics of fungal plant pathogens. Prior to the present work, whole-genome sequences were available only for ten species from the CGSC group[9] that cover a small fraction of the species diversity in the complex. Also, most available genomes were assembled using short-read data only, which has presented a challenge to decipher the evolutionary mechanisms at the chromosome level. In this study, we sequenced de novo 49 genomes of 37 species with different host ranges in CGSC. Fourteen species among the three major clades were assembled with the aid of long reads to the near-chromosome level, thus completing a genome compendium of this species complex. We employed comparative genomic approaches to assess whether large-scale chromosomal rearrangements and gene copy number variation are associated with host adaptation and pathogenicity in CGSC. We aim to identify genomic signatures that differentiates narrow host range (specialist) from broad host range species (generalist). Moreover, we estimated the divergence time among the representatives in CGSC and studied the genetic mechanisms that drive rapid divergence, speciation, and adaptive evolution in closely related species with contrasting host ranges, particularly between the characteristic specialist pathogen *C. kahawae* and its broad host species relative *C. cigarro*.

Results

Genome panorama of CGSC

To understand the evolution of genome complexity and architecture on pathogenicity and adaptation of CGSC species (CGSCs) that cause anthracnose diseases (Fig. 1), whole genome sequences of 49 strains from 37 species of different host ranges were assembled and analyzed (Supplementary Data 1, 2). The BUSCO completeness was > 98.8% in each assembly. The average GC content of CGSCs was 51.09% with a range from 41.32% to 53.58%. The genome sizes of CGSCs ranged from 49.83 to 84.23 Mbp, encompassing 11,719 to 15,272 protein-coding genes (Fig. 2a and Supplementary Data 1), and were typically larger than those of other *Colletotrichum* representatives (average of 60.94 Mbp vs. 53.54 Mbp) as reported in ref. 9. We found that, except for *C. xanthorrhoeae*, all CGSCs

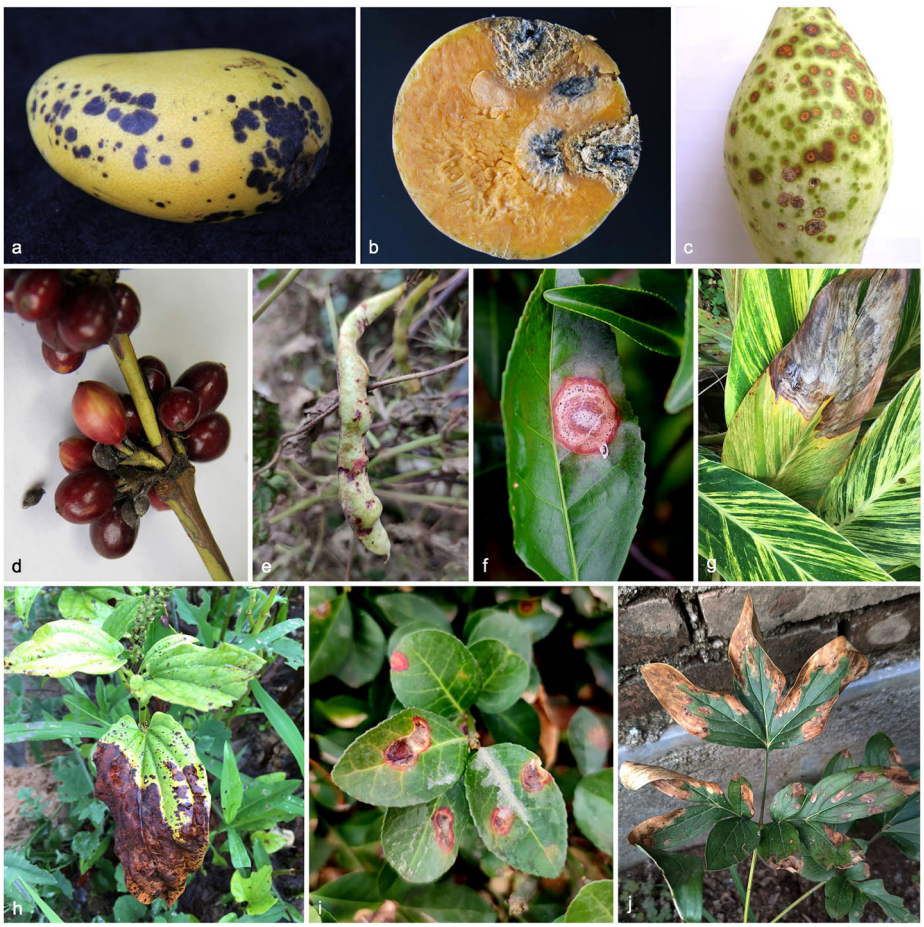


Fig. 1 | Symptoms of anthracnose caused by *Colletotrichum*. **a** *Mangifera indica*, **b** *Cucurbita moschata*, **c** *Lagenaria siceraria*, **d** *Coffea arabica*, **e** *Phaseolus vulgaris*, **f** *Camellia sinensis*, **g** *Alpinia pusilla*, **h** *Saururus chinensis*, **i** *Ilex* sp., **j** *Paeonia lactiflora*. All photos are originally from this study.

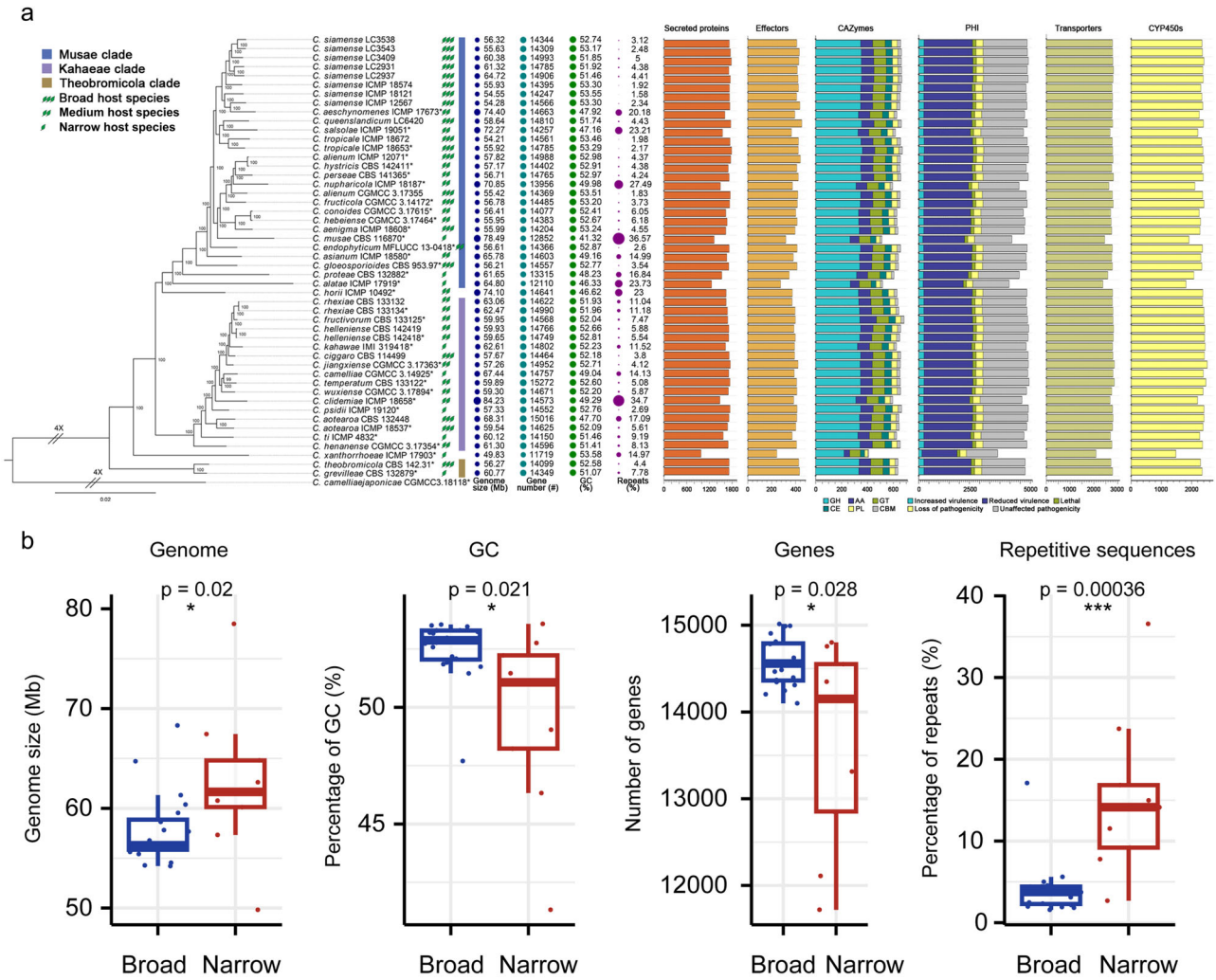


Fig. 2 | Phylogenomic relationship and genome features of the *C. gloeosporioides* species complex (CGSC). a The phylogenomic tree generated from 7303 single-copy homologous genes agrees with previous multi-locus trees[9,11] in the delineation of the major clades: Musae, Kahawae, and Theobromicola. Ex-type strains are indicated with asterisk at the end of the taxa labels. Ten species that infect more than a dozen host plant families display broad host range, nine species that infect a single host plant genus or species appear to be narrow host species[9–11], and the other 15 species that infect host plants in an intermediate range were considered as medium host species. Key genome features are illustrated using bubble plots, and their sizes are scaled in each panel and are not comparable across panels. The bar plots represent the number of genes predicted to encode secreted proteins, effectors, CAZymes, PHI, transporters, and CYP450s. Different classes of CAZymes and PHI are shown in different colors. CAZyme classes: auxiliary activities (AA), carbohydrate esterases (CE), glycoside hydrolases (GH), glycosyl transferases (GT), polysaccharide lyases (PL), and carbohydrate-binding molecules (CBM). **b** Comparison of genome size, GC content, gene number, and repetitive sequences content between twenty broad and nine narrow host range species. The statistical significance was calculated using the Wilcoxon test (*$P < .05$; ***$P < .001$).

with a repetitive DNA content exceeding 7.5% had genome sizes larger than 60 Mbp, and in most genomes of CGSCs larger than 60 Mbp, transposable elements (TEs) accounted for more than 50% of the repetitive sequences. These findings suggest that the expansion of TEs, particularly long terminal repeat retrotransposons, contributed to genome expansion (Fig. 3). A similar pattern was also observed among the different strains within the species (Fig. 2a).

The pan-genome of CGSC across the 49 studied genomes comprised 705,986 homologous genes that were grouped into 21,172 protein families (Supplementary Fig. 16). The core genome, namely protein families with at least one member in all compared strains, comprised 8327 protein families (422,037 genes), which constituted 39.33% of the pan-genome. The homologous protein families shared among the species branching from the corresponding node varied from 8319 to 13,557. In contrast, the number of species-specific genes, namely genes without homologs in other genomes, was relatively low, ranging from eight (*C. cigarro*) to 410 (*C. aeschynomenes*). The total number of species-specific genes accounted for only 0.55% of the pan-genome. Compared with *Aspergillus* and *Aspergillus* section *Nigri*[42], the proportion of species-specific genes in CGSC was relatively lower, suggesting low genetic diversity within CGSC.

A total of 3,698 secondary metabolite gene clusters (SMGCs) were predicted across 49 genomes of CGSC, averaging 75 SMGCs per genome, a number comparable to the prolific *Aspergillus* section *Flavi*[43]. Of these, only 718 SMGCs were associated with 49 known compound families, averaging 15 SMGCs per genome. Within these families, the precursor to melanin, 1,3,6,8-tetrahydroxynaphthalene, and siderophore dimethylcoprogen were conserved across all 49 CGSC genomes. In contrast, 47 SMGCs were variable in the genomes of CGSCs, e.g., those similar to clusters for fujikurin and leucinostatin were found only in *C. musae* and *C. xanthorrhoeae*, respectively (Supplementary Fig. 17b). Ribosomally synthesized and post-translationally modified peptides are an unexplored class of bioactive natural products found only in the Kahawae clade (Supplementary Fig. 17b and Supplementary Data 1), within which *C. kahawae* encodes the largest number of terpenes in the CGSC. These results suggest that *Colletotrichum* has considerable potential to produce diverse novel bioactive secondary metabolites.

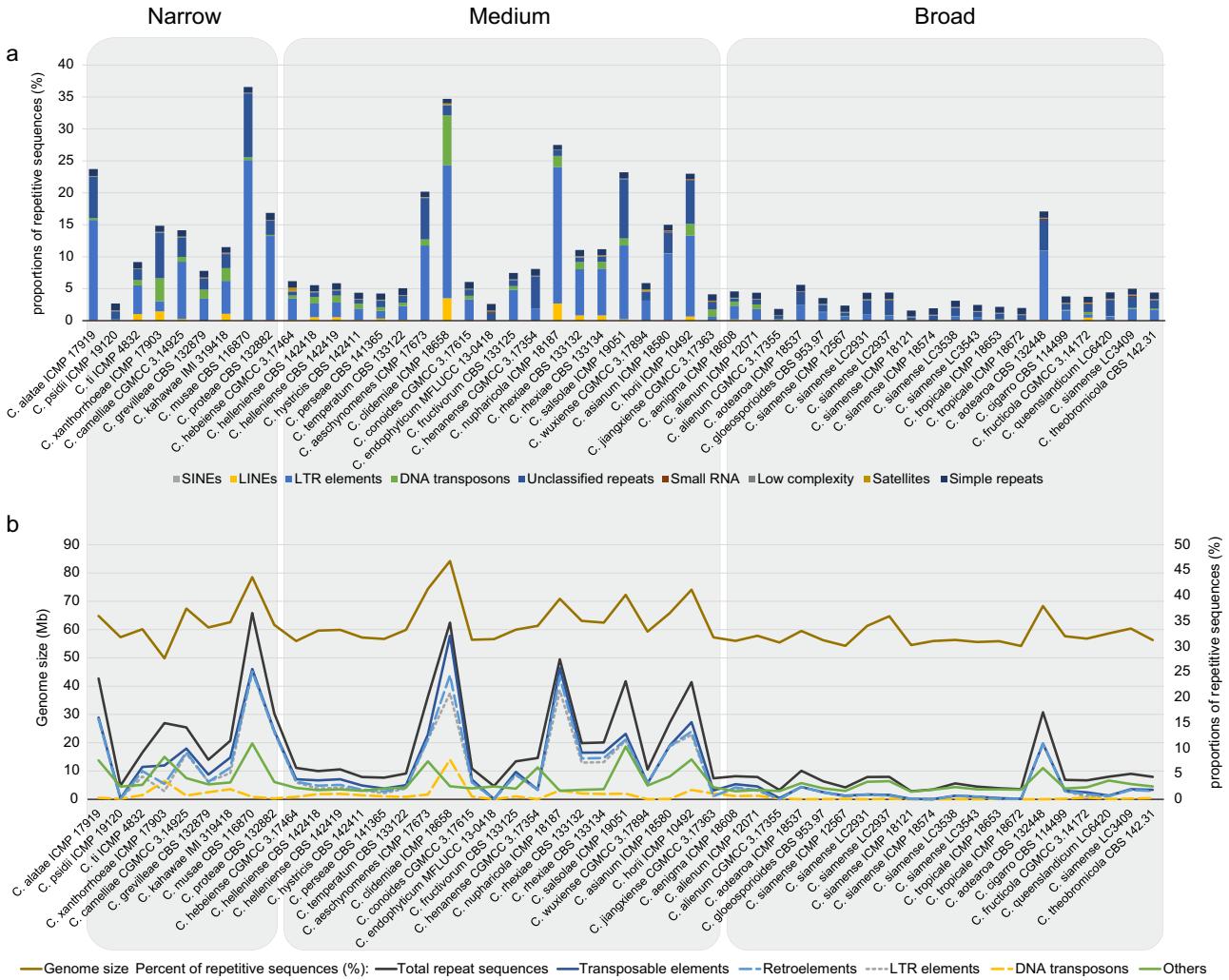


Fig. 3 | Distribution and composition of repetitive DNA sequences in each genome. The light-gray blocks refer to species of narrow hosts (Narrow), medium hosts (Medium), and broad hosts (Broad). **a** The bar plot illustrates the proportions (%) of each class of repetitive sequences. SINEs short interspersed nuclear elements, LINEs long interspersed nuclear elements, LTR long terminal repeats. **b** The line chart illustrates the relationship between genome size and proportions of main categories of repetitive sequences. Others indicate the repetitive sequences except transposable elements.

Combined with long-read sequencing data, we successfully assembled 12 high-quality genomes with fewer than 30 contigs, along with two additional assemblies featuring 31 and 38 contigs. The detection of telomeric repeats at both ends of specific contigs signified assembly at the chromosome level (Supplementary Figs. 1–14). Notably, our investigation revealed the prevalence of minichromosomes in CGSCs, which are typically less than 2 Mb in size, and differ significantly from the core chromosomes in genomic characteristics such as gene density, transposon content, and GC content[44], and that they vary in both number and sequence composition (Supplementary Data 3). For instance, 10 genomes exhibited the assembly of one to twelve minichromosomes; exceptions were noted for *C. musae*, *C. jiangxiense*, *C. kahawae*, and *C. theobromicola* (Supplementary Figs. 1–14). Subsequent analysis revealed that minichromosomes had an almost threefold lower gene density compared to core chromosomes (0.83 genes/10 kbp and 2.36 genes/10 kbp, respectively). Correspondingly, pathogenicity genes, which are necessary for disease development but not essential for the pathogen to complete its lifecycle in vitro[45], including cytochrome P450 enzymes (CYP450s), transporters, secreted proteins, effectors, genes associated with pathogen-host interaction (PHI), and secondary metabolite gene clusters (SMGCs)[14,46] are predominantly located on the core chromosomes. In contrast, the proportion of pathogenicity genes was significantly lower or even absent on the minichromosomes ($P < 0.01$, Wilcoxon test) (Supplementary Fig. 19a). In contrast, minichromosomes were notably enriched in repetitive sequences (29.67%) compared to core chromosomes (12.06%), in particularly with a significant abundance of LTRs (Supplementary Fig. 19b). However, the percentage of SINEs was significantly lower on minichromosomes than on core chromosomes (Supplementary Fig. 19b). Unlike other phytopathogenic fungi such as *Fusarium oxysporum* and *Magnaporthe oryzae*, pathogenicity genes in CGSCs were predominantly located on the core chromosomes, whereas the proportion of pathogenicity genes on the minichromosomes was significantly lower or even absent ($P < 0.01$, Wilcoxon test) (Supplementary Fig. 19a).

Species with narrow host range generally have high content of repetitive sequences and low number of pathogenicity genes

Following previous studies[9–11], we define here ten species that infect more than a dozen host plant families as having a broad host range, nine species that infect one to three species in a single host plant genus as having a narrow host range, and other species as having a medium host range (Fig. 2a and Supplementary Data 2). In CGSC, species of narrow host ranges possessed significantly larger genome size, higher repetitive sequence ratio, lower GC percentage, and fewer protein coding genes than those with broad host ranges ($P < 0.05$, Wilcoxon test) (Fig. 2b and Supplementary Fig. 15). Specifically, all species with high repetitive sequence content (> 7.5%) and larger genomes (> 60 Mbp) had narrow or medium host ranges. For example, host-specific *C. kahawae* exhibited a repetitive DNA content of 11.52% and

a genome that was larger (62.61 Mbp) than the sister species *C. cigarro* (57.67 Mbp), which is a broad host species with only 3.8% repetitive DNA. These results suggest that CGSCs with narrow host ranges may promote gene optimization and accelerate their evolutionary journey through repeated sequence expansion, ensuring their adaptive survival in narrow ecological niches.

Additional analysis of the gene content associated with pathogenicity among CGSCs with varying host plant ranges revealed that the narrow host species exhibited significantly fewer genes encoding secreted proteins, effectors, carbohydrate-active enzymes (CAZymes), PHI factors, transporters, and SMGCs compared to the broad host species ($P < 0.05$, Wilcoxon-test), but comparable number of CYP450 genes (Supplementary Fig. 15). However, there were exceptions, for instance, the total numbers of SMGCs were identical in the narrow host-range species *C. kahawae* and the closely related species with a broad host range, *C. cigarro*. Nonetheless, *C. kahawae* displayed an expansion in non-ribosomal peptide synthetase (NRPS) and NRPS-like gene clusters, while the polyketide synthase (PKS) loci were contracted (Fig. 2 and Supplementary Data 1). This suggests that the NRPS gene clusters may play an important role in the differences in pathogenicity between these two sister species, despite their short divergence time.

Also, the number of pathogenicity genes was significantly lower in species colonizing a single monocotyledonous plant genus (*C. xanthorrhoeae*, *C. alatae*, and *C. musae*) than in those infecting dicotyledons or co-infecting monocotyledonous and dicotyledonous plants ($P < 0.01$, Wilcoxon test, Fig. 2, Supplementary Figs. 17, 18, and Supplementary Data 4). In addition, the numbers of genes for CAZymes involved in hemicellulose and pectin hydrolyzation were contracted in these three monocot pathogens, especially in *C. xanthorrhoeae*.

Unbalanced gene family expansion/contraction

To estimate the divergence time and evolutionary history of CGSCs, we generated a phylogenomic tree using fifteen genomes assembled with both short- and long-read data of this study, together with published genomes of the well-known *C. graminicola* and *C. higginsianum* species complexes, and five other Sordariomycetes fungi. The crown age of CGSC was estimated at 8.54 Mya (95% HPD, 7.51–9.56 Mya). The estimated divergence time between the two major clades, namely Musae and Kahawae, was ~5.9 Mya (95% HPD, 5.19–6.62 Mya). Subsequently, the two clades underwent rapid divergence in the recent past, leading to the radiation of a number of species (Fig. 4a).

In CGSC, gene family contraction was more common than expansion in the "early" speciation event, i.e., at deep nodes, than during "later" speciation in the Musae clade during the evolution of *Colletotrichum* (Fig. 4a, b). Moreover, unbalanced gene family expansion/contraction and gene gain/loss occurred within CGSC: contractions were prevalent in the respective most recent common ancestor (MRCA) of the Kahawae and Theobromicola clades, but gene expansion and gene gain patterns dominated in most species in both clades. In other words, it seems that the ancestors of the Kahawae and Theobromicola clades underwent marked gene contractions, but their descendants experienced high levels of gene gain during the evolution of interactions with different hosts. In contrast, gene family expansion and contraction were not observed in the MRCA of the Musae clade, but gene family contraction and gene loss were evident in all descendant lineages of the Musae clade.

Gene family contraction was often accompanied by gene loss in most species of CGSC, but the pattern was opposite in *C. kahawae* and *C. cigarro*. Although *C. kahawae* underwent gene family contraction rather than expansion, it acquired the largest number of genes within CGSC (Fig. 4b). GO enrichment analysis indicated that the genes in expanded gene families of *C. kahawae* were implicated mainly in oxidoreductase activity, iron ion binding, heme binding, ATP binding, and oxidoreductase activity ($P < 0.05$) (Fig. 4d). In addition, *C. kahawae* experienced the most duplication events in CGSC, and GO terms of 363 genes in the 263 duplication events related mainly to nucleic acid binding, zinc ion binding, DNA integration, protein binding, DNA integration, and catalytic activity including ATPase and hydrolase activities ($P < 0.05$) (Fig. 4c, f). Three hundred and fifty-one of the 363 genes were in expanded gene families. Compared with 21 other species, 226 species-specific genes with the function of nucleic acid binding and zinc ion binding ($P < 0.05$) were discovered in *C. kahawae* (Fig. 4c, e).

Chromosomal rearrangement may be a key mechanism in the evolution of CGSCs

To unravel the mechanisms steering the divergence of the major clades in CGSC and their progeny, we conducted collinear analyzes by comparing the 12 high-quality genome assemblies from Musae and Kahawae clades with that of *C. grevilleae*, a basal species of these clades. A large intra-chromosomal inversion (~1.27 Mbp) was detected in the members of the Kahawae clade, namely *C. aotearoa*, *C. camelliae*, *C. cigarro*, *C. kahawae*, and *C. jiangxiense*, in the same collinear region of *C. grevilleae* contig05 (Fig. 5a, shown as red triangle). In contrast, the Musae clade showed very good collinearity on this contig. Although an intra-chromosomal inversion (~0.68 Mbp) was also observed in *C. horii*, situated at the base of the Musae clade, the position and length of this rearrangement differ from the inversion detected in the Kahawae clade (Supplementary Fig. 20). Therefore, it is highly likely that the large chromosomal inversion (Fig. 5a) was a possible driving force behind the ancestral divergence of the Kahawae clade in CGSC.

Given above results, we speculate that chromosomal rearrangement may be a key mechanism contributing to the formation of recently diverged species within the Kahawae clade. A further in-depth comparative genomic analysis between *C. kahawae* and *C. cigarro*, diverged ~0.58 Mya, confirmed our conjecture and revealed extensive inter- and intra-chromosomal rearrangements (Fig. 6), including nine large intra-inversions (> 100 kbp), 31 small intra-inversions (10–100 kbp), and 15 inter-chromosomal translocations. Within *C. kahawae*, 741 genes were located within the intra-inversions, encompassing genes implicated in oxidoreductase activity, carbohydrate metabolism, and hydrolase activity ($P < 0.05$) (Supplementary Fig. 21a). Only three genes (0.40%) associated with phospholipase, cystathionine beta-synthase, and oxidoreductase activities were under positive selection ($Ka/Ks > 1$, Supplementary Fig. 21c).

After that, we wanted to learn more about what genomic traits are associated with chromosomal inversions. The results of the statistical analyses showed that the proportion of repeat elements and pathogenicity genes within the intra-inversions in *C. kahawae* showed no significant difference from the rest of the genome, except for the genes encoding secreted proteins (average percentage 15.51% vs. 10.43%, $P < 0.05$, Wilcoxon test, Supplementary Fig. 22). Furthermore, there were 28 breakpoints at the ends of inter-chromosomal translocations, and 80 breakpoints were identified at intra-chromosomal inversions (> 10 kbp) in *C. kahawae* compared to *C. cigarro* (Fig. 6c). The average number of coding genes per 10 kbp in the vicinity of the breakpoints (10 kbp upstream and downstream) was lower than in the rest of the genome (1.91 vs. 2.25, $P < 0.05$, Wilcoxon test) (Supplementary Fig. 23), while repetitive sequences were significantly enriched in these regions (31.51% vs. 12.54%, $P < 0.01$, Wilcoxon test). A total of 423 genes were located around the breakpoints, with most involved in membrane, ATPase-coupled transmembrane transporter activity, hydrolase activity, and mycotoxin biosynthesis ($P < 0.05$) (Supplementary Fig. 21b). Moreover, species-specific genes tended to be located at the flanks of the intra-inversions rather than at the inter-chromosomal translocations (Fig. 6d and Supplementary Fig. 24). Thus, the evolutionary event of chromosomal inversions is closely linked to changes in repetitive sequences, pathogenicity-related genes, and species-specific genes.

Characterization of the rapidly evolving LS regions of *C. kahawae*

Chromosomal synteny analysis comparing *C. kahawae* and *C. cigarro* revealed notably diverse lineage-specific (LS) regions in the former, spanning 8.75 Mbp and constituting 14.98% of the entire genome. These LS regions encompassed five large (> 100 kbp) segments, covering most of the chromosomal breakpoints and the ends of chromosomes or contigs. Furthermore, a substantial proportion of species-specific genes (50.26%) and

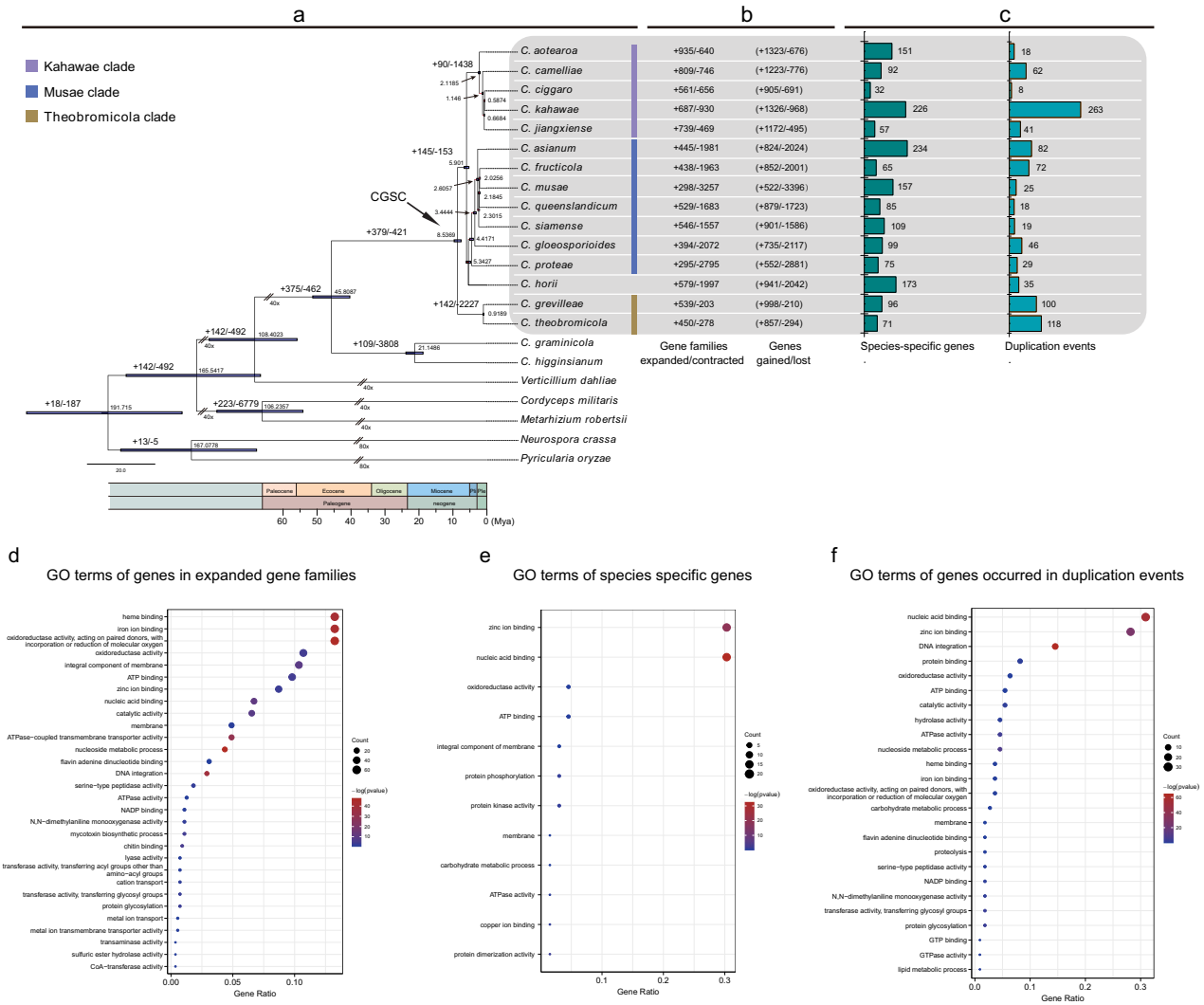


Fig. 4 | Maximum clade credibility chronogram and comparison of gene families of CGSCs. a The chronogram generated from the BEAST analysis based on the concatenated alignment of 872 single-copy orthologous genes from 17 *Colletotrichum* species and five other Sordariomycetes fungi. The mean divergence time is given in millions of years at each node and the blue shadow shows the 95% HPD. The number of gene families that expanded or contracted in each evolutionary branch after speciation is shown on the corresponding branch, with "+" indicading expansion and "−" indicading contraction. The red branches indicate that the numbers of gene family expansion and contraction are both zero here. Mya: million years ago; Pli: Pliocene; Ple: Pleistocene. **b** Left column: the number of expanded (+) and contracted (−) gene families of each lineage. Right column (numbers in brackets): gene gain (+) and loss (−) in corresponding to gene family expansion and contraction on the left side. **c** The number of species-specific genes and duplication events in each species (Source data is based on the result of Orthofinder). Gene ontology (GO) enrichment of *C. kahawae* in expanded gene families (**d**), species-specific genes (**e**), genes in duplication events (**f**). The *x*-axis is the gene ratio of the analyzed genes. The 30 GO terms of genes with the lowest *P* value in expanded gene families (**d**), the total GO terms of species-specific genes (**e**), and genes in duplication events (**f**) are shown on the *y*-axis, respectively. GO functional enrichment analysis was performed by the R package ClusterProfile.

duplicated genes (78.79%) in *C. kahawae* were situated within the LS regions (Fig. 6c).

Compared to the 14 genomes assembled to the near-chromosome level, the length of LS regions in *C. kahawae* exhibited positive correlation with the genetic distances between these species and *C. kahawae* (Supplementary Fig. 25a, b). Several LS regions, totaling 4.33 Mbp, were conserved. The 384 genes located in the conserved LS regions were associated with GO terms primarily related to nucleic acid and zinc ion binding, DNA integration, ATPase activity, and serine-type carboxypeptidase activity ($P < 0.05$), including genes encoded zinc finger proteins, DDE superfamily endonucleases, and integrases (Supplementary Fig. 25c, d and Supplementary Data 5). Gene density in LS regions was significantly lower compared to the bulk of the *C. kahawae* genome, and these regions were significantly enriched in repetitive sequences and pathogenicity genes that were predicted to be CYP450s and SMGCs (Supplementary Fig. 26). However, genes related to secreted proteins, CAZymes, transporters, and PHI were significantly less abundant in the LS regions than in the rest of the genome (Supplementary Fig. 26).

Discussion

Ecological adaptation and corresponding evolutionary mechanism of CGSCs

Our phylogenomic and molecular clock analyses suggested that CGSCs underwent rapid diversification since the late Miocene epoch (8.54 Mya) (Fig. 4a). The high-intensity monsoon and intense tectonic activity in the Hengduan Mountains and Himalayas during this period expanded habitat availability, facilitating the radiation of plants[47] and, consequently, the diversification of plant-associated microbes[48]. However, unlike the evolutionary trajectory observed in the entomopathogen *Metarhizium* spp., which underwent a transition from specialist to generalist through intermediate host range species[49], specialists and generalists within the CGSC had multiple origins with respect to the genome tree (Fig. 2a), reflecting a

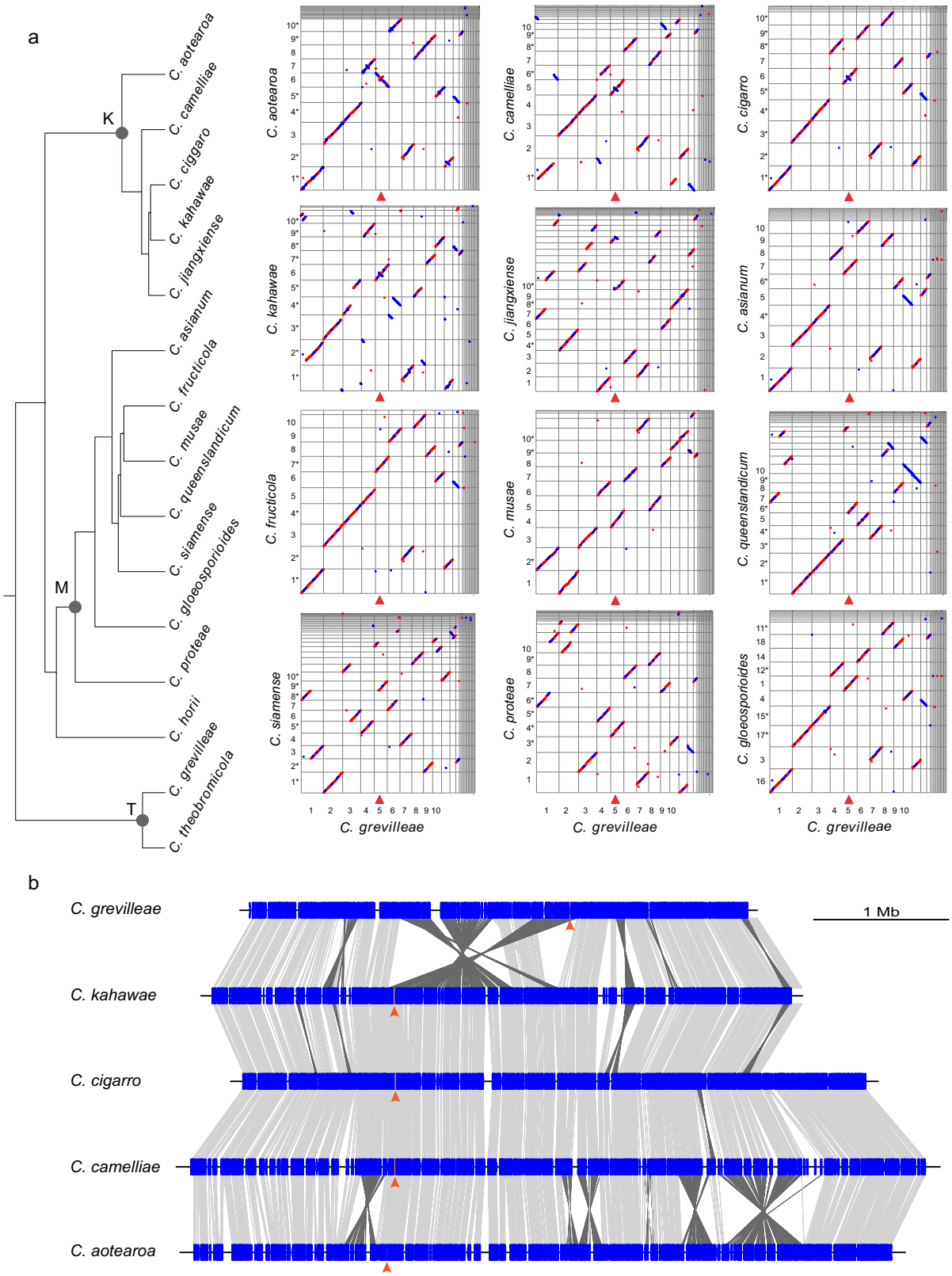
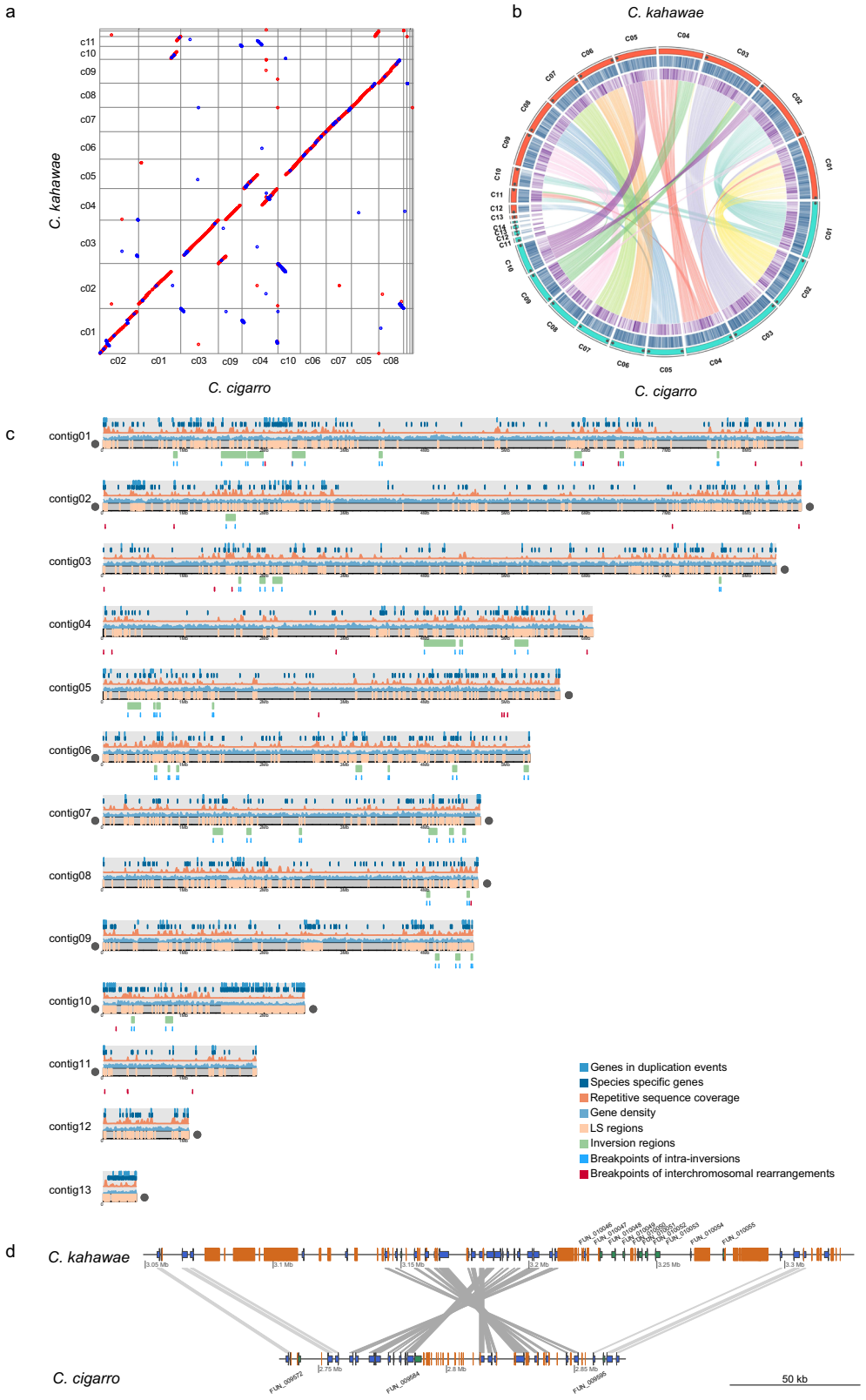


Fig. 5 | Syntenic relationships among species in the Kahawae clade (K), Musae clade (M), and Theobromicola clade (T) of CGSC. a Dot plot represents the genome syntenic regions between *C. grevilleae* (*x*-axis) and species in Kahawae/Musae clade (*y*-axis). The red and blue dots represent sequences aligning the forward and reverse hits. The red triangles indicate contigs with or without genomic rearrangement at the collinear region of *C. grevilleae* contig05. The number from 1 to 10 beside and below the dot plot indicates the number of the 10 longest contigs in each assembly, except that of *C. gloeosporioides*, in which the contig number is marked following the original number of the downloaded genome from GenBank (accession number: GCA_003243855.1). The stars after the number indicated the reverse-complemented contigs. **b** Syntenic relationships among five representative genomes in CGSC. The blue lines represent the genes of *C. grevilleae* in contig05 and the corresponding regions in *C. kahawae* contig06, *C. cigarro* contig06, *C. camelliae* contig05, and *C. aotearoa* contig06. The light and dark gray ribbons indicate the collinear blocks between the two adjacent contigs. Orange arrows indicate the locations of genes annotated as centromere protein H.



more complex evolutionary process and host adaptation. In particular, genes associated with secreted proteins, effectors, CAZymes, and PHI factors were significantly expanded in CGSC generalists compared to specialists (Supplementary Fig. 15). This expansion likely contributes to the ability of generalists to infect multiple hosts, a pattern also seen in other fungal pathogens[49].

CAZymes play a pivotal role in degrading plant cell walls during host penetration and infection[50]. The variability in the number of these enzymes was evident between *C. higginsianum* and *C. graminicola*, which infect distinct host plants[51]. Our findings indicated a reduction in genes associated with pectate lyase in monocot pathogens (Supplementary Fig. 17a, Supplementary Data 4), attributed to the lower pectin content in

Fig. 6 | Genome architecture of *C. kahawae* and its syntenic relationship with its sister species *C. cigarro*. **a** Syntenic relationships between *C. kahawae* and *C. cigarro*. A dot plot represents the genome syntenic regions between 14 contigs of *C. cigarro* (*x*-axis) and 13 contigs of *C. kahawae* (*y*-axis). The red and blue dots represent sequences aligning the forward and reverse hits. The last four unlabeled contigs of *C. cigarro* are contig11 to contig14, and the last two unlabeled contigs of *C. kahawae* are contig12 and contig13. **b** Circos diagram illustrating collinear blocks between chromosomes of *C. kahawae* and *C. cigarro* using proteins sequences. The outermost circle represents 13 contigs of *C. kahawae* (upside) and 14 contigs of *C. cigarro* (downside), with the dark gray dots indicating the positions of telomere. The second (dark blue) and third (purple) circle represents the heatmap of gene and repetitive sequence density, respectively. **c** Distribution of genes and LS regions of *C. kahawae* compared with its relative *C. cigarro*. The gray bars represent the assembled contigs of *C. kahawae*, and the flanking gray dots indicate that telomeric sequences are detected on that side of contigs. The light orange regions within gray bars represent the LS regions of *C. kahawae* compared with *C. cigarro*. Above the gray bars, from top to bottom, the light blue and dark blue lines represent the genes occurred in duplication events and species-specific genes, respectively, and each line represents a gene; the two ribbon plots represent gene density and repetitive sequence coverage (with a window of 10 kbp). Below the genome, the green bands indicate intra-arrangement regions in *C. kahawae*. The light blue lines and red lines indicate the breakpoints of intra-inversions and interchromosomal-rearrangements, respectively. **d** Example of species-specific gene of *C. kahawae* at the flank of intra-inversions between *C. kahawae* contig06 (3,054,643 to 3,321,682 bp) and *C. cigarro* contig06 (2,737,281 to 2,867,416 bp). Gray ribbons indicate regions of homology between the two genes. Blue and green arrows indicate the common genes and species-specific genes, respectively. Orange blocks indicated repetitive sequences.

the primary cell walls of monocots compared to dicots[15,52–54]. Notably, the absence of the PL11 polysaccharide lyase responsible for the initial cleavage of rhamnogalacturonan, a major component of plant cell wall pectin[55,56], was observed in the three monocot pathogens (*C. xanthorrhoeae*, *C. alatae*, and *C. musae*). Conversely, Baroncelli et al.[12] reported an over-representation of polysaccharide lyases in dicotyledonous pathogens. Additionally, the numbers of genes encoding auxiliary activities and carbohydrate-binding modules, including the CBM24 gene, were reduced in the three monocot pathogens. Hence, the arsenal needed for degradation between dicot and monocot pathogens differs due to the differences in plant cell wall chemistry.

Gene duplication, a prevalent genomic phenomenon, followed by acquiring new gene function often serves as a key driver of evolutionary innovation and ecological adaptation[57,58]. The identification of gene duplication events in *C. kahawae*, mostly located in LS regions, provides intriguing insights into the adaptive strategy and genetic divergence of this fungal species[59]. Its duplicated genes are associated with essential functions, such as nucleic acid and zinc ion binding, which is a hallmark of zinc-finger transcription factors, as well as catalytic and hydrolase activities, highlighting their relevance in shaping the adaptive landscape. The presence of glycosyl hydrolase among the duplicated genes further underscores their involvement in plant biomass degradation[60,61], a crucial aspect of the ecological niche. Therefore, the interplay between gene duplication and ecological adaptation adds complexity to our understanding of how fungi like *C. kahawae* navigate their ecological niches and underscores the multifaceted nature of genomic evolution in response to environmental changes.

The interplay between gene gain and gene loss also holds profound implications for the evolutionary dynamics of species, particularly in the context of ecological adaptation[62]. Our analysis indicates that gene loss emerges as a dominant force in the evolution of *Colletotrichum*, aligning with previous observations in this genus[63] and several other groups[64,65]. This pervasive gene loss suggests a mechanism promoting adaptive phenotypic diversity, where shedding unnecessary genetic baggage could be advantageous for navigating different ecological niches[66,67]. Intriguingly, the Kahawae clade, and *C. kahawae* in particular, deviates from this trend and shows a prevalence of gene gain. Notably, the expansion of CYP450 gene families, known for their versatile roles in cellular processes[68,69], signifies a potential key to adaptive evolution following host jump event. The contrasting reduction in pathogenicity-related genes in *C. kahawae*, including those involved in transport, PHI responses, CAZymes, and secretion, highlights a nuanced evolutionary strategy. It suggests that, in the transition from generalist to specialist lifestyles, gene loss may operate as a selective force, eliminating redundancies and streamlining the genetic repertoire to align with the specialized ecological niche. This intricate balance between gene gain and loss emerges as a dynamic driver in the evolution of *Colletotrichum* species, which has also been found in other pathogen[49], shaping their adaptive trajectories in response to varying environmental and host-specific challenges.

Genomic rearrangements and enrichment of repetitive DNA are potential drivers of CGSC divergence and speciation

Chromosomal synteny analysis is a useful approach for deciphering genome evolution, particularly when applied to closely related species[70–72]. A few studies have been performed to address the chromosomal rearrangements among species within the CGSCs. For instance, interspecific and intraspecific chromosomal rearrangements were observed within species of Musae clade in CGSC[30]. Moreover, TE-mediated chromosome rearrangements have also been reported between *C. destructivum* and *C. higginsianum*[29] and at the intraspecific level in *C. higginsianum*[35]. However, one of the key innovations of this study is the discovery of chromosomal rearrangement between the big clades, which has not been reported in previous studies[29,31,35]. In the present study, mesosynteny on the same collinear chromosome was observed between species belonging to the Kahawae clade and the basal Theobromicola clade (*C. grevilleae*), where the chromosome content was preserved but the order and orientation of the genes were shuffled. While the Musae clade showed very good collinearity on this contig. Although an inversion is also occurred on that chromosome in *C. horii*, it is different that of the Kahawae clade in the position and length (Supplementary Fig. 20). Given that large inversions have been reported to play a remarkable role in genetic divergences and evolution[73,74], we hypothesized that the two intra-inversions observed in *C. horii* and the Kahawae clade were resulted from different evolutionary events of chromosomal rearrangements in the MRCA of GCSCs. We have not observed similar chromosomal rearrangement events as in *C. horii* in other species. Therefore, the ancestral species of *C. horii* may have undergone chromosomal rearrangement events that were detrimental to environmental adaptation and did not lead to further speciation events. This may also explain the inconsistent phylogenetic position of *C. horii* inferred in previous studies based on multi-locus phylogenetic analyzes, which placed it either basal to the Musae clade, the Kahawae clade, or both[9–11], and in our current analysis based on phylogenomic analysis, which placed it basal to the Musae clade. In addition, centromere protein H, one of the foundation kinetochore proteins, was found within this large inversion region that occurred in the Kahawae clade (Fig. 5b), suggesting a possible correlation between genomic rearrangement and mitotic crossover[75,76]. This observed pattern may be similar to that of *V. dahliae*, where genomic rearrangements are associated with centromeres during the evolution of the species[77].

Within the Kahawae clade, the host-specific species *C. kahawae* exhibits extensive chromosomal rearrangements when compared to its closely related species *C. cigarro*, which has a broad host range. The rearranged regions in *C. kahawae* are characterized by an enrichment of genes associated with secreted proteins, along with the presence of genes related to oxidoreductase activity, carbohydrate metabolism, and mycotoxin biosynthesis. Notably, a gene encoding cystathionine beta-synthase in *C. kahawae* was under positive selection. This gene, known for its role in the pathogenicity of *Aspergillus niger* on pear fruit[78], located within the rearrangement regions may play key roles in virulence and host interaction in CGSC. Furthermore, a key species-specific gene of *C. kahawae*, located at the flank of intra-inversions, belongs to the dynamin-like GTPase

superfamily. This superfamily is involved in various cellular processes and is crucial for the development of appressoria in *Magnaporthe oryzae* and for virulence in the plant symbiotic fungus *M. robertsii*[79–81]. Hence, the extensive chromosomal rearrangements in *C. kahawae* are likely drivers of the evolutionary adaptation of virulence genes, facilitating the ongoing arms race with host plant immune systems.

The repetitive sequences were significantly enriched in the flanks of chromosomal rearrangements (breakpoints) and the LS regions of *C. kahawae* (Supplementary Figs. 23, 26), generating a genomic environment that favors the generation of chromosomal rearrangements and may contribute to sequence diversity and genome plasticity in *C. kahawae*[39,82,83]. The estimated divergence time of *C. kahawae* and *C. cigarro* from their common ancestor was ~0.58 Mya (Fig. 4a), and correlated with the emergence of *Coffea arabica* (0.543–1.08 Mya), the host plant of *C. kahawae*[84]. Therefore, we hypothesized that the ancestral population of *C. kahawae* potentially underwent a host jump from the most common ancestor with *C. cigarro*[23], possibly accompanied by chromosomal rearrangements following the speciation events of *C. arabica*.

The generation of minichromosomes enriched in rearrangements and repetitive sequences could serve as a mechanism to increase genetic variation, compensating for the absence of sexual reproduction[85]. This is because the high variability of repetitive sequences makes them more prone to mutations, which could facilitate the emergence of highly pathogenic genes on minichromosomes[27,86]. Among the CGSCs genomes assembled to near-chromosomal levels, especially in species reported to primarily have asexual stages (e.g., *C. aotearoa*, *C. camelliae*, *C. grevilleae*, *C. horii*, *C. proteae*, and *C. queenslandicum*), 1–12 minichromosomes were identified (Supplementary Figs. 1–14). In contrast, Wang et al.[30] reported that CGSCs have an average of 14 complete chromosomes with telomeres at both ends, which means that each strain has an average of four minichromosomes. However, they did not mention the number of minichromosomes in each strain. The number of minichromosomes varies between different species and even between different strains of the same species, indicating the genetic instability and diversity of genome structure of CGSCs. The varying number of minichromosomes observed between different species, and even between different strains of the same species, indicates the genetical instability and diversity of genome architecture in CGSCs. Although telomere-to-telomere minichromosomes of *C. kahawae* were not assembled successfully, this species has been confirmed as a true clonal pathogen[23,32]. Historical demographic reconstructions and DNA polymorphism patterns consistently indicate a sudden and severe decline of the effective population size of *C. kahawae*[23], signifying an intensified genetic drift, in line with the expansion of TEs in this species (Fig. 3). Repetitive sequences may facilitate the generation of genetic diversity by promoting mitotic crossover which in turn generates new genetic combinations[39]. Similar genome architectures have also been reported in *F. oxysporum* and *V. dahliae*, contributing to gene flow and acting as a driving force for adaptive evolution[39,74,87–89].

This study has certain limitations. First, PacBio reads were not available for all the samples. Since repetitive sequences are common in fungi, this could present technical challenges for genome assembly from Illumina reads and sequence alignment. Second, gene prediction and annotation are dependent on the quality of genome assemblies and number of overlapping genes. The number of predicted genes in fragmented genomes could be lower than the complete genomes due to the potential larger number of fragmented/smaller-sized contigs. For instance, the genome of *C. xanthorrhoeae*, which was assembled with Illumina data, resulted in a lower gene count of 11,719, reflecting the highly fragmented nature of the assembly. As a result, the number of species-specific genes, as exemplified by *C. xanthorrhoeae*, *C. aeschynomenes*, and *C. clidemiae*, could be lower due to more fragmented genome assemblies. In future ongoing research, a more robust genome assembly approach, such as combining PacBio or Hi-C sequencing data, would be essential for a detailed and accurate gene analysis.

Methods

Genome sequencing and assembly

The genomes of 49 strains that represented 37 available species in CGSC were sequenced (Supplementary Data 1). Single spore of these strains was isolated[90] and cultivated in potato dextrose broth at 25°C with shaking at 150 rpm for 3–5 d. Biomass was collected by filtration through sterile gauze, and samples were freeze dried and stored at −80 °C. Total genomic DNA was extracted using a modified cetyltrimethylammonium bromide protocol[91] and the quality and concentration of the DNA were measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., MA, USA) and Qubit 3.0 Fluorometer (Thermo Fisher). The DNA libraries were sequenced on the Illumina NovaSeq 6000 platform to produce 150 bp paired-end reads (Annoroad Gene Technology Co., Ltd, Yiwu, China). Raw reads were assessed by FastQC v.0.11.8[92], and the quality-trimmed reads were assembled de novo using SPAdes v.3.12.0[93] with a "careful" mode. Contigs < 200 bp were removed from subsequent analyzes[94].

Among the 37 species, 14 species with different host ranges and belonging to the three major clades of CGSC, namely *C. aotearoa*, *C. asianum*, *C. camelliae*, *C. cigarro*, *C. fructicola*, *C. grevilleae*, *C. horii*, *C. jiangxiense*, *C. kahawae*, *C. musae*, *C. proteae*, *C. queenslandicum*, *C. siamense*, and *C. theobromicola* (Supplementary Data 1) were also sequenced using PacBio RS II platforms (Nextomics Biosciences Co., Ltd., Wuhan, China). The PacBio raw reads were assembled using Canu v.1.4.1 with default parameters[95], and the assemblies from PacBio were polished with the Illumina reads using Pilon v.1.23.2[96]. Contigs belonging to the mitochondrial genomes were removed by performing the BLASTn search (e-value ≤ 1e − 5) against the NCBI nr/nt databases (downloaded on 23/08/2019). Telomeric sequences were identified using a high stringency BLASTn with repeats of TTAGGG or CCCTAA. Repeats ≥ 60 bp length that were located at the end of contigs were considered as telomeric sequences[97]. The quality of genome assembly was assessed using QUAST v.5.0.2[98], and the completeness of genomes was estimated using BUSCOs v.2.0.1[99] to search for a set of 3725 Sordariomycete_odb9 universal single-copy orthologous genes.

Repeat analysis and gene annotation

Repetitive DNA elements were identified de novo using RepeatModeler v.2.0[100], which incorporates RECON v.1.0.8[101], RepeatScout v.1.0.8[102], TRF v.4.0.9[103], and RepeatMasker v4.1.0 (http://www.repeatmasker.org) combined database Dfam_3.1[104].

Gene prediction was performed using the Funannotate pipeline (v.1.7.0, https://github.com/nextgenusfs/funannotate). Repetitive sequences in assemblies were soft-masked using the command funannotate mask with default parameters. The soft-masked assemblies were predicted by running the command funannotate predict with --busco_db Sordariomycetes and --busco_seed_species *Verticillium longporum* and default parameters. The predicted protein-coding gene models were annotated using eggNOG-mapper-v2[105]. Pfam domains were identified using Pfam database v.32.0[106]. Gene ontology (GO) terms were transferred from the Pfam entries with Pfam2GO mapping (version data: 2020/11/28)[107], and the GO enrichment analysis was performed using ClusterProfiler package in R[108].

To identify putative CAZymes, the protein genes were searched using run_dbcan v.2.0.6[109], which integrates three tools: HMMER[110] search against dbCAN-HMM database v.8 (e-value < 1e−15, coverage > 0.35), DIAMOND[111] search against CAZy database v.0731 2018 (e-value < 1e−102)[112] and Hotpep[113] search against the PPR (peptide pattern recognition) library (frequency > 2.6, hits > 6)[114]. We considered genes identified by at least two tools as CAZyme candidates.

SMGCs were identified using the antiSMASH v.5.0 web server fungal version[115] with additional BLAST searches against the MIBiG database[116] to predict the association between the detected clusters and the production of known or novel bioactive compounds.

Prediction of secreted proteins was performed using SignalP v.5.5[117], TMHMM v.2.0[118], and TargetP v.2.0[119] with default parameters to assess the presence of signal peptides, transmembrane domains, and subcellular

localization for protein sequences. Proteins with a signal peptide but no transmembrane domain and that did not target the mitochondrion were considered as secreted proteins. EffectorP v.1.0[120] was used to predict fungal effectors from the predicted secreted proteins. The CYP450s, transporters, and proteins involved in PHI were annotated via BLASTp (e-value ≤ 1e−10) against the Fungal Cytochrome P450 Database[68], Transporters Classification Database (downloaded on 18/04/2020)[121], and PHI Database v.4.8[122].

Phylogenetic analysis and divergence time estimation

Orthologous genes in all the genomes were inferred using Orthofinder v.2.3.3[123] with default parameters[124]. The genes clustered in one orthologous group were considered as a gene family. Each single-copy gene family was aligned using MAFFT v.7.407[125], trimmed using Gblocks v.0.91b (−t = p)[126], and concatenated. The final concatenated alignment was used to construct a phylogenetic tree using RAxML v.8.2.12[127] with 1000 replicates under the JTT model advised by ProtTest v.3.4.2[42,128,129].

The genome assemblies of 14 *Colletotrichum* species using a combination of PacBio and paired-end Illumina sequence data in this study, along with the genomes of *C. gloeosporioides* (accession number: GCA_003243855.1), *C. graminicola* (GCF_000149035.1), and *C. higginsianum* (GCF_001672515.1) retrieved from GenBank were used to construct the time-scaled phylogeny of CGSC. *Cordyceps militaris* (accession number: GCF_000225605.1), *Metarhizium robertsii* (GCF_000187425.2), *Neurospora crassa* (GCF_000182925.2), *Pyricularia oryzae* (GCF_000002495.2), and *V. dahliae* (GCF_000150675.1) were used as outgroups and for time calibration. Three ancestral nodes were used for time calibration: the common ancestral node between *Cordyceps-Metarhizium* divergence at 146–206 Mya[130], the common ancestral node between *C. fructicola* and *C. graminicola* divergence at 22.09–47.68 Mya, and the common ancestral node between *C. graminicola* and *C. higginsianum* divergence at 12.3 Mya (http://www.timetree.org/). Following the pipeline of Hacquard et al.[130], single-copy gene families with a phylogenetic signal > 0.7[131] were used to construct a phylogenetic tree using RAxML v.8.2.12 with JTT substitution model and the most appropriate model for each maker. MCMC analyzes were then performed with a chain length of 10,000,000 sampling on every 5000 generations in BEAST v.2.6.3[132]. The results from two independent MCMC runs were combined with LogCombiner. Twenty percent of the first total trees identified with TreeAnnotator were discarded and the node heights were estimated using mean heights to obtain the maximum clade credibility tree[124]. The final tree was visualized in FigTree v.1.4.0 (http://tree.bio.ed.ac.uk/software/Figtree).

Gene family gains and losses

Gene family expansion and contraction of the 17 *Colletotrichum* species and five other Sordariomycetes fungi used for aforesaid divergence estimation were assessed using CAFÉ v.4.2.1[133] with a *P* value of 0.01 and an automatic estimation of the birth and death (λ) value. The result of CAFÉ was calculated refer to the existing scripts (https://github.com/hahnlab/cafe_tutorial/tree/main).

Genomic similarity and synteny

Genome synteny of the 14 high-quality genomes assembled using both PacBio RSII and Illumina data was analyzed using both MUMmer v.3.23[134] and MCScanX[135]. The pairwise genome alignments were generated by MUMmer using NUCmer with default parameters (except for –mum). The dot plot of genome synteny was visualized by mummerplot with the filtered alignment file (minimum alignment length > 1000 bp and alignment identify > 90%). In addition, collinear blocks between whole-genome protein sequences were assessed using BLASTp (e-value ≤ 1e−10) and MCScanX with default parameters. The collinear blocks between two genomes or within species were visualized using R package circlize[136], and the gene collinearity between contigs was visualized using R package genoPlotR[137].

The coverage breadth of the whole-genome alignments generated by MUMmer on 1 kb nonoverlapping windows was assessed using BEDTools v.2.26.0 to determine the LS regions. The regions were considered LS with an alignment breadth less than 0.2[39]. The gene density and repeat coverage of genomes were calculated using BEDTools with a window of 10 kbp. The result of the genome architecture was visualized using R package karyoploteR[138].

To search the segmental genomic duplication of *C. kahawae*, the whole-genome nucleotide alignments of *C. kahawae* were generated by MUMmer as above method. The regions longer than 5 kbp and with identity more than 90% were considered as large segmental genomic duplications.

Identification of positive selection genes

To test for adaptive molecular evolution and identify amino acid sites under diversifying selection of the specialist pathogen *C. kahawae* driven by host jumping, the ratio of nonsynonymous to synonymous substitutions (*Ka/Ks*) of orthologous genes between *C. kahawae* and the close relative *C. cigarro* was calculated with ParaAT v.2.0[139] and KaKs_calculator v.2.0[140] using MAFFT v.7.407 for multiple-sequence alignment with default parameters. Genes with *P* value > 0.05, *Ks* value > 2.0, and *Ka/Ks* ratio > 45 or NA were discarded because of the risk of substitution saturation and few nonsynonymous sites or substitutions[141,142].

Statistics and reproducibility

The sample sizes shown in figures were determined based on statistical requirements and sample availability. Each group had at least three replicates. All data values used in the figures were contained in Supplementary Data. For statistical analysis, Wilcoxon test was used to compare the findings obtained from two different types of genomes, and has been described in figure legends. The significance levels are indicated in the figures by exact *p* values or asterisk ($*p < 0.05$; $**p < 0.01$; $***p < 0.001$; $p > 0.05$ or ns not significant).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All new genomes published in the study are available in the Genome Warehouse (GWH, https://ngdc.cncb.ac.cn/gwh/#) under the BioProject PRJCA033952 and in the National Microbiology Data Center (NMDC, https://nmdc.cn/) under the BioProject NMDC10018299, and the GWH accession number of assembled genomes are listed in Supplementary Data 1.

Received: 24 September 2024; Accepted: 31 March 2025;
Published online: 10 April 2025

References

1. Dean, R. et al. The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.* **13**, 414–430 (2012).
2. Cannon, P. F., Damm, U., Johnston, P. R. & Weir, B. S. *Colletotrichum* - current status and future directions. *Stud. Mycol.* **73**, 181–213 (2012).
3. Jayawardena, R. S. et al. Notes on currently accepted species of *Colletotrichum*. *Mycosphere* **7**, 1192–1260 (2016).
4. McGovern, R. J., Seijo, T. E., Hendricks, K. & Roberts, P. D. New report of *Colletotrichum gloeosporioides* causing postbloom fruit drop on citrus in Bermuda. *Can. J. Plant Pathol.* **34**, 187–194 (2012).
5. Lima, N. B. et al. First report of mango anthracnose caused by *Colletotrichum karstii* in Brazil. *Plant Dis.* **97**, 1248–1248 (2013).
6. Munda, A. First report of *Colletotrichum fioriniae* and *C. godetiae* causing apple bitter rot in Slovenia. *Plant Dis.* **98**, 1282–1282 (2014).
7. Dias, L. R. C., Brito, R. A. S., Melo, T. A. & Serra, I. M. R. S. First report of papaya fruit anthracnose caused by *Colletotrichum okinawense* in Brazil. *Plant Dis.* **104**, 573–573 (2020).

8. Khruengsai, S. et al. Antifungal properties of volatile organic compounds produced by *Daldinia eschscholtzii* MFLUCC 19-0493 isolated from *Barleria prionitis* leaves against *Colletotrichum acutatum* and its post-harvest infections on strawberry fruits. *PeerJ* **9**, e11242 (2021).
9. Liu, F. et al. Updating species diversity of *Colletotrichum*, with a phylogenomic overview. *Stud. Mycol.* **101**, 1–56 (2022).
10. Talhinhas, P. & Baroncelli, R. *Colletotrichum* species and complexes: geographic distribution, host range and conservation status. *Fungal Divers* **110**, 109–198 (2021).
11. Weir, B. S., Johnston, P. R. & Damm, U. The *Colletotrichum gloeosporioides* species complex. *Stud. Mycol.* **73**, 115–180 (2012).
12. Baroncelli, R. et al. Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. *BMC Genomics* **17**, 555 (2016).
13. Buiate, E. A. S. et al. A comparative genomic analysis of putative pathogenicity genes in the host-specific sibling species *Colletotrichum graminicola* and *Colletotrichum sublineola*. *BMC Genomics* **18**, 67 (2017).
14. Lelwala, R. V., et al. Comparative genome analysis indicates high evolutionary potential of pathogenicity genes in *Colletotrichum tanaceti*. *PLoS ONE* **14**, e0212248 (2019).
15. Baroncelli, R. et al. Genome evolution and transcriptome plasticity is associated with adaptation to monocot and dicot plants in *Colletotrichum* fungi. *Gigascience* **13**, giae036 (2024).
16. Waller, J. M., Bridge, P. D., Black, R. & Hakiza, G. Characterization of the coffee berry disease pathogen *Colletotrichum Kahawae sp. nov*. *Mycol. Res.* **97**, 989–994 (1993).
17. Batista, D. et al. Legitimacy and implications of reducing *Colletotrichum kahawae* to subspecies in plant pathology. *Front. Plant Sci.* **7**, 2051 (2017).
18. Jayawardena, R. S. et al. One stop shop IV: taxonomic update with molecular phylogeny for important phytopathogenic genera: 76–100 (2020). *Fungal Divers* **103**, 87–218 (2020).
19. Pires, A. S. et al. Cytogenomic characterization of *Colletotrichum kahawae*, the causal agent of coffee berry disease, reveals diversity in minichromosome profiles and genome size expansion. *Plant Pathol.* **65**, 968–977 (2016).
20. Ferrucho, R. L., Marín-Ramírez, G. A., Ochoa-Corona, F. & Angel, C. C. A. PCR-based detection for the quarantine fungus *Colletotrichum* kahawae, a biosecurity threat to the coffee (*Coffea arabica*) industry worldwide. *Plant Dis.* **108**, 2615–2624 (2024).
21. Cooke T., Persley D. & House S. *Diseases of Fruit Crops in Australia* (CSIRO Publishing, 2009).
22. Zhao, P. et al. Fungi of quarantine concern for China I: dothideomycetes. *Persoonia* **47**, 45–105 (2021).
23. Silva, D. N. et al. Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae*. *Mol. Ecol.* **21**, 2655–2670 (2012).
24. Vieira, A. et al. Aggressiveness profiling of the coffee pathogen *Colletotrichum kahawae*. *Plant Pathol.* **68**, 358–368 (2019).
25. Cabral, A. et al. Pathological, morphological, cytogenomic, biochemical and molecular data support the distinction between *Colletotrichum cigarro comb. et stat. nov*. and *Colletotrichum kahawae*. *Plants* **9**, 502 (2020).
26. Dallery, J. F. et al. Gapless genome assembly of *Colletotrichum higginsianum* reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. *BMC Genomics* **18**, 667 (2017).
27. Bhadauria, V. et al. Genetic map-guided genome assembly reveals a virulence-governing minichromosome in the lentil anthracnose pathogen *Colletotrichum lentis*. *N. Phytol.* **221**, 431–445 (2019).
28. Becerra, S., Baroncelli, R., Boufleur, T. R., Sukno, S. A. & Thon, M. R. Chromosome-level analysis of the *Colletotrichum graminicola* genome reveals the unique characteristics of core and minichromosomes. *Front. Microbiol.* **14**, 1129319 (2023).
29. Lapalu, N. et al. Complete genome of the Medicago anthracnose fungus, *Colletotrichum destructivum*, reveals a mini- chromosome-like region within a core chromosome. *Microb. Genomics* **10**, 001283 (2024).
30. Wang, H. M. et al. The evolution of mini-chromosomes in the fungal genus *Colletotrichum*. *mbio* **14**, e0062923 (2023).
31. Gan, P. et al. Telomeres and a repeat-rich chromosome encode effector gene clusters in plant pathogenic *Colletotrichum* fungi. *Environ. Microbiol.* **23**, 6004–6018 (2021).
32. Vieira, A., Silva, D. N., Varzea, V., Paulo, O. S. & Batista, D. Novel insights on colonization routes and evolutionary potential of *Colletotrichum kahawae*, a severe pathogen of *Coffea arabica*. *Mol. Plant Pathol.* **19**, 2488–2501 (2018).
33. Burt, A. Perspective: sex, recombination, and the efficacy of selection–was Weismann right?. *Evolution* **54**, 337–351 (2000).
34. McDonald, B. A. & Linde, C. Pathogen population genetics, evolutionary potential, and durable resistance. *Annu. Rev. Phytopathol.* **40**, 349–379 (2002).
35. Tsushima, A. et al. Genomic plasticity mediated by transposable elements in the plant pathogenic fungus *Colletotrichum higginsianum*. *Genome Biol. Evol.* **11**, 1487–1500 (2019).
36. Yun, S. H., Arie, T., Kaneko, I., Yoder, O. C. & Turgeon, B. G. Molecular organization of mating type loci in heterothallic, homothallic, and asexual *Gibberella/Fusarium* species. *Fungal Genet. Biol.* **31**, 7–20 (2000).
37. Fradin, E. F. & Thomma, B. Physiology and molecular aspects of *Verticillium* wilt diseases caused by *V. dahliae* and *V. albo-atrum*. *Mol. Plant Pathol.* **7**, 71–86 (2006).
38. Hane, J. K. et al. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biol.* **12**, R45 (2011).
39. de Jonge, R. et al. Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res.* **23**, 1271–1282 (2013).
40. Stukenbrock, E. H. Evolution, selection and isolation: a genomic view of speciation in fungal plant pathogens. *N. Phytol.* **199**, 895–907 (2013).
41. Plissonneau, C. et al. Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. *Front. Plant Sci.* **8**, 119 (2017).
42. Vesth, T. C. et al. Investigation of inter- and intraspecies variation through genome sequencing of *Aspergillus* section *Nigri*. *Nat. Genet.* **50**, 1688–1695 (2018).
43. Kjaerbolling, I. et al. A comparative genomics study of 23 *Aspergillus* species from section *Flavi*. *Nat. Commun.* **11**, 1106 (2020).
44. Plaumann, P. L. & Koch, C. The many questions about mini chromosomes in *Colletotrichum* spp. *Plants* **9**, 641 (2020).
45. Idnurm, A. & Howlett, B. J. Pathogenicity genes of phytopathogenic fungi. *Mol. Plant Pathol.* **2**, 241–255 (2001).
46. Rao, S. & Nandineni, M. R. Genome sequencing and comparative genomics reveal a repertoire of putative pathogenicity genes in chilli anthracnose fungus *Colletotrichum truncatum*. *PLoS ONE* **12**, e0183567 (2017).
47. Ding, W. N., Ree, R. H., Spicer, R. A. & Xing, Y. W. Ancient orogenic and monsoon-driven assembly of the world's richest temperate alpine flora. *Science* **369**, 578–581 (2020).
48. Sato, H., Tanabe, A. S. & Toju, H. Host shifts enhance diversification of ectomycorrhizal fungi: diversification rate analysis of the ectomycorrhizal fungal genera Strobilomyces and Afroboletus with an 80-gene phylogeny. *N. Phytol.* **214**, 443–454 (2017).
49. Hu, X. et al. Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. *Proc. Natl. Acad. Sci. USA* **111**, 16796–16801 (2014).

50. Zhao, Z. T., Liu, H. Q., Wang, C. F. & Xu, J. R. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC Genomics* **14**, 274 (2013).
51. O'Connell, R. J. et al. Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nat. Genet.* **44**, 1060–1065 (2012).
52. Kubicek, C. P., Starr, T. L. & Glass, N. L. Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. *Annu. Rev. Phytopathol.* **52**, 427–451 (2014).
53. Yokoyama, R. A genomic perspective on the evolutionary diversity of the plant cell wall. *Plants* **9**, 1195 (2020).
54. Lorrai, R. & Ferrari, S. Host cell wall damage during pathogen infection: mechanisms of perception and role in plant-pathogen interactions. *Plants* **10**, 399 (2021).
55. Caffall, K. H. & Mohnen, D. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydr. Res.* **344**, 1879–1900 (2009).
56. Ochiai, A., Itoh, T., Mikami, B., Hashimoto, W. & Murata, K. Structural determinants responsible for substrate recognition and mode of action in family 11 polysaccharide lyases. *J. Biol. Chem.* **284**, 10181–10189 (2009).
57. Paterson, A. H. et al. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
58. Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B Biol. Sci.* **279**, 5048–5057 (2012).
59. Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. & Ravikesavan, R. Gene duplication as a major force in evolution. *J. Genet.* **92**, 155–161 (2013).
60. Espagne, E., Balhadere, P., Penin, M. L., Barreau, C. & Turcq, B. HET-E and HET-D belong to a new subfamily of WD40 proteins involved in vegetative incompatibility specificity in the fungus Podospora anserina. *Genetics* **161**, 71–81 (2002).
61. Prieto, A. et al. Fungal glycosyl hydrolases for sustainable plant biomass valorization: Talaromyces amestolkiae as a model fungus. *Int. Microbiol.* **24**, 545–558 (2021).
62. Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12**, 1048–1059 (2002).
63. Gan, P. et al. Genus-wide comparative genome analyses of *Colletotrichum* species reveal specific gene family losses and gains during adaptation to specific infection lifestyles. *Genome Biol. Evol.* **8**, 1467–1481 (2016).
64. Sharma, R., Mishra, B., Runge, F. & Thines, M. Gene loss rather than gene gain is associated with a host jump from monocots to dicots in the Smut Fungus *Melanopsichium pennsylvanicum*. *Genome Biol. Evol.* **6**, 2034–2049 (2014).
65. Mourkas, E. et al. Agricultural intensification and the evolution of host specialism in the enteric pathogen *Campylobacter jejuni*. *Proc. Natl. Acad. Sci. USA* **117**, 11018–11028 (2020).
66. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**, e00036–12 (2012).
67. Albalat, R. & Canestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
68. Cresnar, B. & Petric, S. Cytochrome P450 enzymes in the fungal kingdom. *Biochim. Biophys. Acta* **1814**, 29–35 (2011).
69. Shin, J., Kim, J. E., Lee, Y. W. & Son, H. Fungal cytochrome P450s and the P450 complement (CYPome) of *Fusarium graminearum*. *Toxins* **10**, 112 (2018).
70. Lyons, E. et al. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781 (2008).
71. Tang, H. B. et al. Perspective - Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
72. Cheng, F., Wu, J., Fang, L. & Wang, X. W. Syntenic gene analysis between Brassica rapa and other Brassicaceae species. *Front. Plant Sci.* **3**, 198 (2012).
73. Hou, J., Friedrich, A., de Montigny, J. & Schacherer, J. Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in *Saccharomyces cerevisiae*. *Curr. Biol.* **24**, 1153–1159 (2014).
74. Shi-Kunne, X., Faino, L., van den Berg, G. C. M., Thomma, B. & Seidl, M. F. Evolution within the fungal genus *Verticillium* is characterized by chromosomal rearrangement and gene loss. *Environ. Microbiol.* **20**, 1362–1373 (2018).
75. Amor, D. J., Kalitsis, P., Sumer, H. & Choo, K. H. A. Building the centromere: from foundation proteins to 3D organization. *Trends Cell Biol.* **14**, 359–368 (2004).
76. Chovanec, P. & Yin, Y. A mapping platform for mitotic crossover by single-cell multi-omics. *Methods Enzymol.* **661**, 183–204 (2021).
77. Seidl, M. F. et al. Repetitive elements contribute to the diversity and evolution of centromeres in the fungal genus Verticillium. *mBio* **11**, e01714–e01720 (2020).
78. Guo, S. X. et al. Functional characterization of a cystathionine β-synthase gene in sulfur metabolism and pathogenicity of *Aspergillus niger* in pear fruit. *J. Agric. Food Chem.* **67**, 4435–4443 (2019).
79. Zhong, K. L. et al. MoDnm1 dynamin mediating peroxisomal and mitochondrial fission in complex with MoFis1 and MoMdv1 is important for development of functional appressorium in *Magnaporthe oryzae*. *PLoS Pathog.* **12**, e1005823 (2016).
80. Ramachandran, R. & Schmid, S. L. The dynamin superfamily. *Curr. Biol.* **28**, R411–R416 (2018).
81. Xie, X. Y. et al. DNM1, a dynamin-related protein that contributes to endocytosis and peroxisome fission, is required for the vegetative growth, sporulation, and virulence of *Metarhizium robertsii*. *Appl. Environ. Microbiol.* **86**, e01217–e01220 (2020).
82. Möller, M., Habig, M., Freitag, M. & Stukenbrock, E. H. Extraordinary genome instability and widespread chromosome rearrangements during vegetative growth. *Genetics* **210**, 517–529 (2018).
83. Covo, S. Genomic instability in fungal plant pathogens. *Genes* **11**, 421 (2020).
84. Bawin, Y. et al. Phylogenomic analysis clarifies the evolutionary origin of *Coffea arabica*. *J. Syst. Evol.* **59**, 953–963 (2021).
85. Temporini, E. D. & VanEtten, H. D. Distribution of the pea pathogenicity (PEP) genes in the fungus *Nectria haematococca* mating population VI. *Curr. Genet.* **41**, 107–114 (2002).
86. Plaumann, P. L., Schmidpeter, J., Dahl, M., Taher, L. & Koch, C. A dispensable chromosome is required for virulence in the hemibiotrophic plant pathogen *Colletotrichum higginsianum*. *Front. Microbiol.* **9**, 1005 (2018).
87. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.* **8**, e1002608 (2012).
88. Faino, L. et al. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res.* **26**, 1091–1100 (2016).
89. Fokkens L., Shahi S., Connolly L. R., Stam R. & Rep M. The multi-speed genome of *Fusarium oxysporum* reveals association of histone modifications with sequence divergence and footprints of past horizontal chromosome transfer events. *bioRxiv*, https://doi.org/10.1101/465070 (2018).
90. Zhang, K., Su, Y.-Y. & Cai, L. An optimized protocol of single spore isolation for fungi. *Cryptogam. Mycologie* **34**, 349–356 (2013).
91. Allen, G. C., Flores-Vergara, M. A., Krasnyanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325 (2006).
92. Andrews S. FASTQC. A quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

93. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
94. Mousavi-Derazmahalleh, M. et al. Prediction of pathogenicity genes involved in adaptation to a lupin host in the fungal pathogens *Botrytis cinerea* and *Sclerotinia sclerotiorum* via comparative genomics. *BMC Genomics* **20**, 385 (2019).
95. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
96. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
97. Li, F. et al. Emergence of the Ug99 lineage of the wheat stem rust pathogen through somatic hybridisation. *Nat. Commun.* **10**, 5068 (2019).
98. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
99. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.***1962**, 227–245 (2019).
100. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
101. Bao, Z. R. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
102. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
103. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
104. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 22 (2021).
105. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
106. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
107. Mitchell, A. et al. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, D213–D221 (2015).
108. Yu, G. C., Wang, L. G., Han, Y. Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).
109. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).
110. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
111. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
112. Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
113. Busk, P. K., Pilgaard, B., Lezyk, M. J., Meyer, A. S. & Lange, L. Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinforma.* **18**, 214 (2017).
114. Busk, P. K. & Lange, L. Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs. *Appl. Environ. Microbiol.* **79**, 3380–3391 (2013).
115. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
116. Kautsar, S. A. et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
117. Armenteros, J. J. A. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
118. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
119. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
120. Sperschneider, J. et al. EffectorP: predicting fungal effector proteins from secretomes using machine learning. *N. Phytol.* **210**, 743–761 (2016).
121. Saier, M. H., Reddy, V. S., Tamang, D. G. & Vastermark, A. The transporter classification database. *Nucleic Acids Res.* **42**, D251–D258 (2014).
122. Urban, M. et al. PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. *Nucleic Acids Res.* **45**, D604–D610 (2017).
123. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
124. Sierra-Patev, S. et al. A global phylogenomic analysis of the shiitake genus Lentinula. *Proc. Natl. Acad. Sci. USA* **120**, e2214076120–e2214076120 (2023).
125. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
126. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
127. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
128. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
129. Miyauchi, S. et al. Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. *Nat. Commun.* **11**, 5125 (2020).
130. Hacquard, S. et al. Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nat. Commun.* **7**, 11362 (2016).
131. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).
132. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comp. Biol.* **10**, e1003537 (2014).
133. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
134. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
135. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
136. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. *Circlize* implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
137. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).

138. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
139. Zhang, Z. et al. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **419**, 779–781 (2012).
140. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteom. Bioinforma.* **8**, 77–80 (2010).
141. Chen, G. H. et al. Genome-wide analysis of proline-rich extension-like receptor protein kinase (PERK) in *Brassica rapa* and its association with the pollen development. *BMC Genomics* **21**, 401 (2020).
142. Wu, Z. H. et al. Analysis of six chloroplast genomes provides insight into the evolution of *Chrysosplenium* (Saxifragaceae). *BMC Genomics* **21**, 621 (2020).

Acknowledgements

We appreciate the help from Xingxing Shen (Zhejiang University), Weiwei Zhang, Xiaoling Zhang, Kuan Li, Yingfeng Luo, Qi Wu, Junmin Liang, Shilin Han, Yuanjie Li and Da Li (CAS, China) for various technical assistance. We are very grateful to Pedro W. Crous (Westerdijk Fungal Biodiversity Institute, The Netherlands) and Weir Bevan (Landcare Research, New Zealand) for providing several strains. Clement K.M. Tsui is grateful to CAS PIFI for the award of visiting scientist fellowship. This work was supported by National Natural Science Foundation of China (32322001, 32330002), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0810000), the National Science and Technology Fundamental Resources Investigation Program of China (2021FY100900), the Youth Innovation Promotion Association of Chinese Academy of Sciences (2021085), and Beijing Nova Program (20230484305).

Author contributions

F.L., L.C., and Z.M. conceived the project. F.L. collected the samples. Z.M. and F.L. performed the data analyzes. Z.M. performed the laboratory experiments and bioinformatic analyzes. F.L. and Z.M. wrote the manuscript with the help from L.C. and C.K.M.T. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08024-9.

Correspondence and requests for materials should be addressed to Lei Cai.

Peer review information *Communications Biology* thanks Nicolas Lapalu and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Matteo Dell'Acqua and Michele Repetto.