Whole-Genome Sequencing and Comparative Genomic Analysis of Potential Biotechnological Strains of *Trichoderma harzianum*, *Trichoderma atroviride*, and *Trichoderma reesei*

Rafaela Rossi Rosolen[a,b] Maria Augusta Crivelente Horta[a,c], Paulo Henrique Campiteli de Azevedo[a,b], Carla Cristina da Silva[a], Danilo Augusto Sforca[a], Gustavo Henrique Goldman[c], and Anete Pereira de Souza[a,d*]

[a]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, Campinas, SP, Brazil

[b]Graduate Program in Genetics and Molecular Biology, Institute of Biology, UNICAMP, Campinas, SP, Brazil

[c]Faculty of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, SP, Brazil

[d]Department of Plant Biology, Institute of Biology, UNICAMP, Cidade Universitária Zeferino Vaz, Rua Monteiro Lobato, Campinas, SP, Brazil

***Corresponding author**

E-mail: anete@unicamp.br

ORCID

RRR: 0000-0002-8964-4034; **MACH:** 0000-0003-2159-5296; **PHCA:** 0000-0001-9564-9336; **DAS:** 0000-0003-2304-8712; **GHG:** 0000-0002-2986-350X; and **APS:** 0000-0003-3831-9829.

Acknowledgments

We are grateful to CBMAI Campinas and SP for conceiving the fungal isolates used in the current study; the Center of Molecular Biology and Genetic Engineering (CBMEG) at the University of Campinas and SP for the use of the center and laboratory space; and the São Paulo Research Foundation (FAPESP), the Coordination of Improvement of Higher Education Personnel (CAPES, Computational Biology Program), and the Brazilian National Council for Technological and Scientific Development (CNPq) for supporting the project and researchers.

Abstract

Trichoderma atroviride and *Trichoderma harzianum* are widely used as commercial biocontrol agents against plant diseases. Recently, *T. harzianum* IOC-3844 (Th3844) and *T. harzianum* CBMAI-0179 (Th0179) demonstrated great potential in the enzymatic conversion of lignocellulose into fermentable sugars. Herein, we performed whole-genome sequencing and assembly of Th3844 and Th0179 strains. To assess the genetic diversity within the genus *Trichoderma*, the results of both strains were compared with strains of *T. atroviride* CBMAI-00020 (Ta0020) and *T. reesei* CBMAI-0711 (Tr0711). The sequencing coverage value of all genomes evaluated in this study was higher than that of previously reported genomes for the same species of *Trichoderma*. The resulting assembly revealed total lengths of 40 Mb (Th3844), 39 Mb (Th0179), 36 Mb (Ta0020), and 32 Mb (Tr0711). A genome-wide phylogenetic analysis provided details on the relationships of the newly sequenced species with other *Trichoderma* species. Structural variants revealed genomic rearrangements among Th3844, Th0179, Ta0020, and Tr0711 relative to the *T. reesei* QM6a reference genome and showed the functional effects of such variants. In conclusion, the findings presented herein allow the visualization of genetic diversity in the evaluated strains and offer opportunities to explore such fungal genomes in future biotechnological and industrial applications.

Keywords: *Trichoderma*, genomes, comparative analysis, orthology analysis, structural variant analysis

1 Introduction

Fungi of the genus *Trichoderma* are characterized by their considerable nutritional versatility (Sharma et al. 2019), which allows them to be employed in a wide range of biotechnological applications (Kidwai and Nehra 2017). For example, *Trichoderma reesei* is the primary fungal source of the industrial cellulases and hemicellulases present in enzymatic cocktails (Bischof et al. 2016). In addition to enzymatic activities, the biocontrol capacity against plant pathogenic fungi has been widely explored in *Trichoderma harzianum* and *Trichoderma atroviride* (Medeiros et al. 2017; Saravanakumar et al. 2017). Recently, *T. harzianum* strains were explored for their enzymatic potential and were demonstrated to be useful for improving lignocellulosic conversion into sugars during second-generation ethanol (2G ethanol) production (Almeida et al. 2021; Delabona et al. 2020; Motta et al. 2021; Zhang et al. 2020).

The phenotypic and ecological heterogeneity across fungi of the same genera is reflected, in part, by the diversity observed within their genomes (Priest et al. 2020). In this way, the diverse and important roles of fungi, as well as the technological advances in next-generation sequencing, have motivated broad efforts to sequence several fungal genomes (Hagestad et al. 2021; Kumar et al. 2021; Nagel et al. 2021; Varga et al. 2019; Wu et al. 2018). Since the genome of *T. reesei* QM6a was first presented (Martinez et al. 2008), *Trichoderma* sequencing studies have increased with the goal of better understanding the biological and ecological roles of *Trichoderma* to improve their applications (Druzhinina et al. 2018; Horta et al. 2014; Kubicek et al. 2011, 2019; Li et al. 2017; Schmoll et al. 2016). Such a growing number of sequenced species may help reveal the molecular basis for the specific features of diverse *Trichoderma* strains.

Previously, the transcriptional profiles of two *T. harzianum* strains, *T. harzianum* IOC-3844 (Th3844) and *T. harzianum* CBMAI-0179 (Th0179), were analyzed under cellulose degradation conditions and compared with those from *T. atroviride* CBMAI-0020 (Ta0020) and *T. reesei* CBMAI-0711 (Tr0711) (Almeida et al. 2021; Horta et al. 2018). Such studies have suggested the great potential of both *T. harzianum* strains as hydrolytic enzyme producers, and this was similar to Tr0711, while Ta0020 showed a low cellulolytic ability. Furthermore, differences in the transcriptional regulation of hydrolytic enzyme expression were observed between Th3844 and Th0179 (Rosolen et al. 2022), highlighting the genetic differences between these strains. Although previous studies investigated the Th3844 genomic regions, which are related to biomass degradation, through bacterial artificial chromosome (BAC) library construction (Crucello et al. 2015; Filho et al. 2017), genomic information regarding the hydrolytic strains of *T. harzianum*, Th3844 and Th0179, remains unclear.

In this study, Pacific Biosciences (PacBio) (Ardui et al. 2018) technology was used to obtain highly contiguous de novo assemblies and to describe the genetic variation present among Th3844 and Th0179. To expand knowledge on the genetic diversity within the genus *Trichoderma*, the results obtained for *T. harzianum* strains were compared to those from *T. atroviride* and *T. reesei*. The chosen species are appropriate for the study's goal because while *T. atroviride* is a biocontrol species that is distantly related to the lignocellulolytic species *T. reesei* (Druzhinina et al. 2006), representing a well-defined phylogenetic species (Dodd et al. 2003), *T. harzianum sensu lato* is also commonly used in biocontrol but constitutes a complex of several cryptic species (Chaverri et al. 2015; Druzhinina et al. 2010).

After performing whole-genome annotation, we investigated the contents of carbohydrate-active enzymes (CAZymes) (Cantarel et al. 2009) and secondary metabolite biosynthetic gene clusters (SMGCs) that were distributed among the studied genomes. To thoroughly investigate the genetic variability across the four evaluated strains, we explored the structural variants (SVs), which represent a major form of genetic and phenotypic variations that are inherited and polymorphic in species (Mills et al. 2011), between them and *T. reesei* QM6a, the reference genome (Martinez et al. 2008). In addition, by performing a comparative genomic analysis across the genus *Trichoderma* and more evolutionarily distant genera, the orthologs and the orthogroups across them were identified, and the rooted gene tree based on the single-copy orthologs was inferred.

The genomic resources we provide herein significantly extend our knowledge regarding the evolution and basic biology of the evaluated strains, and this may increase their biotechnological employment. The results from this study might also increase the availability of genomic data, which can be used for comparative studies aiming to correlate the phenotypic differences in the genetic diversity of *Trichoderma* species; therefore, the study may help improve the search for enzymes with enhanced properties and provide aid toward improving the production of chemicals and enzymes in such fungi.

2 Materials and methods

2.1 Fungal strains and culture conditions

The species originated from the Brazilian Collection of Environment and Industry Microorganisms (CBMAI), which is located in the Pluridisciplinary Center for Chemical, Biological, and Agricultural Research (CPQBA) at the

University of Campinas (UNICAMP), Brazil. The identity of *Trichoderma* isolates was authenticated by CBMAI based on phylogenetic studies of their internal transcribed spacer (ITS) region, translational elongation factor 1 (*tef1*), and RNA polymerase II (*rpb2*) marker gene. The phylogenetic tree based on the ITS region was previously published (Rosolen et al. 2022); moreover, the methodology used to model the phylogenetic trees based on the *tef1* and *rpb2* marker genes is described in Supplementary Material 1. Briefly, Th3844, Th0179, Ta0020, and Tr0711 strains were cultivated on potato dextrose agar (PDA) solid medium (ampicillin 100 µg/ml and chloramphenicol 34 µg/ml) for 3 days at 28 °C. Conidia were harvested, and an initial spore solution was used to inoculate 500 mL of potato dextrose broth (PDB) medium. The cultivation process was performed in biological triplicates for 72 h at 28 °C and 200 rpm for all evaluated strains. Then, mycelial samples were harvested using Miracloth (Millipore), frozen using liquid nitrogen, and stored at -80 °C. Frozen material was used for DNA extraction.

2.2 DNA extraction and sequencing

The ground fungal tissue was suspended in 800 µl of lysis buffer (1 M Tris-HCl [pH 8.0], 0.5 M EDTA [pH 8.0], 5 M NaCl, N-lauroylsarcosine, β-mercaptoethanol, and H_2O), and phenol:chloroform:isoamyl alcohol (25:24:1) (Sigma, US) was then added. After centrifugation at 4 °C and 15,871 x g for 10 min, the aqueous layer was collected, and genomic DNA was precipitated through the addition of isopropanol. DNA was harvested by centrifugation at 4 °C and 15,871 x g for 10 min, and the pellet was washed with 70% ethanol and centrifuged at 4 °C and 15,871 x g for 5 min. After a second wash with 95% ethanol and centrifugation at 4 °C and 15,871 x g for 5 min, the pellet was dried at room temperature and dissolved in TE buffer. Before the quality control steps, the DNA was subjected to RNase treatment, which involved the addition of 2 µl of RNase to the TE buffer (final concentration of 50 µg/mL) and incubation of the samples for 1 h at room temperature before storage. The methodology used for DNA extraction used was based on previously established protocols (Oliveira et al. 2015; Peterson et al. 2000) and adapted according to our fungal samples.

The quantity of the extracted gDNA was determined by measuring the absorbance at 260 nm using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific) and Qubit Fluorometer (Thermo Fisher Scientific). The quality of extracted gDNA was assessed through 0.8% agarose gel electrophoresis. HiFi sequencing libraries were prepared according to the PacBio protocol (Wenger et al. 2019), and sequencing was performed at the Arizona

Genomics Institute (AGI; Tucson, USA) using a SMRT DNA sequencing system from PacBio (PacBio RSII platform).

2.3 Genome assembly

The data were transferred to a local server, and the genomes were assembled de novo using Canu software (v.2.1) (-pacbio – hifi, and a genome estimate equal to 40 Mb for all evaluated strains), which was developed for long-read sequencing (Koren et al. 2017). Genome integrity was assessed using the Quality Assessment Tool (QUAST) (Gurevich et al. 2013) (v.5.0.2) and Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao et al. 2015) (v.4.1.4) tools. The Nucmer alignment tool from the MUMmer (v.4.0.0beta2) toolbox (Kurtz et al. 2004; Marcais et al. 2018) was used to perform the whole-genome alignments between the evaluated strains.

2.4 Gene prediction and functional annotation

Gene prediction was performed using AUGUSTUS (v.3.3.3) (Stanke et al. 2006) through gene models, which were built from *T. harzianum* T6776, *T. atroviride* IMI206040, and *T. reesei* QM6a (TrainAugustus (v.3.3.3)), together with MAKER (Cantarel et al. 2008) (v.2.31.11). Such programs are implemented on the Galaxy platform. The predicted genes were functionally annotated by searching for homologous sequences in the UniProt (The UniProt Consortium 2021), eggNOG-mapper v.2 (Cantalapiedra et al. 2021), and Protein Annotation with Z score (PANNZER2) (Toronen et al. 2018) databases. Transmembrane proteins were predicted using TMHMM v.2.0 (Krogh et al. 2001). For the annotation of CAZymes, we used CDSs as homology search queries against the database of the dbCAN2 server (Zhang et al. 2018), which integrates the (I) DIAMOND (E-value < 1e-102) (Buchfink et al. 2015), (II) HMMER (E-value < 1e-15, coverage > 0.35) (Finn et al. 2011), and Hotpep (frequency > 2.6, hits > 6) (Busk et al. 2017) tools. We considered all CDSs as true hits if they were predicted by at least two tools. Coverages were estimated with QualiMap (Okonechnikov et al. 2016) (v.2.2.2c) using minimap2 (Li 2018) v. 2.17 + galaxy4, which were both implemented on the Galaxy platform (Afgan et al. 2018). SMGCs in the Th3844, Th0179, Ta0020, and Tr0711 genomes were predicted using antiSMASH fungal version v.6.1.0 (Blin et al. 2021).

2.5 Ortholog identification and clustering

The proteomes of Th3844, Th0179, Ta0020, and Tr0711 were compared with the *Trichoderma* spp. proteomes that are available on NCBI databases. *Fusarium* spp., *Aspergillus* spp., and *Neurospora* spp. were used as outgroup. For this analysis, we used the software OrthoFinder (Emms and Kelly 2015, 2019) v2.5.2, which clustered the protein sequences of fungi into orthologous groups and allowed the phylogenetic relationships between them to be identified. The consensus species tree was inferred using the STAG algorithm (Emms and Kelly 2018) and rooted using the STRIDE algorithm (Emms and Kelly 2017), which are implemented on the OrthoFinder program. The resulting tree from the OrthoFinder analysis was visualized and edited using Interactive Tree of Life (iTOL) v6 (Letunic and Bork 2007).

2.6 Long-read structural variant analysis

SVs were identified by aligning the PacBio HiFi reads from Th3844, Th0179, Ta0020, and Tr0711 with the *T. reesei* QM6a reference genome (Martinez et al. 2008) using the software Map with BWA-MEM (Li and Durbin 2010) v.0.7.17.2 with (-x pacbio, sorted by chromosomal coordinates). The duplicate reads in the BAM file were identified and marked using the tool MarkDuplicates (Broad Institute) v.2.18.2.2. Variants were called using Sniffles (Sedlazeck et al. 2018) v.1.0.12+galaxy0, allowing for a minimum support of 10 (--min_support), maximum number of splits of 7 (--max_num_splits), maximum distance of 1000 (--max_distance), minimum length of 30 (--min_length), minimum mapping quality of 20 (--minmapping_qual), and CCS reads option (--ccs_reads). SVs were annotated using SnpEff (v.4.3+T. galaxy1) (Cingolani et al. 2012), which allowed the effects of variants in genome sequences to be categorized. Such tools are implemented on the Galaxy platform (Afgan et al. 2018).

2.7 Expression profile of CAZymes

The gene expression profile of the CAZymes identified in the assembled genomes was investigated using data from previous transcriptomes (Almeida et al. 2021; Horta et al. 2018). The gene expression of CAZymes from Th3844, Th0179, Ta0020, and Tr0711 was modeled using transcripts per million (TPM) data, which were validated through qPCR and described by Almeida et al. (2021) and Horta et al. (2018). This gene expression dataset was obtained from three biological replicates under cellulose and glucose growth conditions.

To search for the corresponding homologs, we mapped all transcripts from (I) Th3844 and Th0179, (II) Ta0020, and (III) Tr0711 against the nucleotide sequences from (I) *T. harzianum* T6776 (Baroncelli et al. 2015), (II)

T. atroviride IMI 206040 (Kubicek et al. 2011), and *T. reesei* v2.0 (Martinez et al. 2008), respectively. We then selected the transcripts corresponding to the CAZymes identified in the assembled genomes, and their expression patterns under cellulose and glucose growth conditions were evaluated. The corresponding orthologs of CAZyme-encoding genes among the genomes of Th3844, Th0179, Ta0020, and Tr0711 were obtained using the results from the OrthoFinder analysis, and the results were visualized using the R package pheatmap (Kolde 2019).

3 Results

3.1 Molecular phylogeny of the evaluated **Trichoderma** *spp. isolates*

We previously investigated the phylogenetic relationships among Th3844, Th0179, Ta0020, and Tr0711 based on the ITS sequences (Rosolen et al. 2022). Here, we modeled phylogenetic trees based on the *rpb2* and *tef1* sequences of 16 and 26 *Trichoderma* spp., respectively, including those of our study strains (Supplementary Material 1: Supplementary Fig. 1). According to the results, high genetic proximity was observed between Th3844 and Th0179. In contrast, both strains were phylogenetically distant from Ta0020 and Tr0711, which were grouped with other *T. atroviride* and *T. reesei* strains, respectively. *Neurospora* spp. and *Fusarium* spp. were used as outgroups of the *rpb2* phylogenetic tree. No outgroups were used in the *tef1* phylogenetic tree because the orthologs in most distant genera, including *Neurospora* spp. and *Fusarium* spp., presented significantly different transcript sequences, which did not allow alignment between their nucleotide sequences with those of *Trichoderma* spp.

3.2 Strain cultivations and evaluation of extracted DNA

First, we cultivated Th3844, Th0179, Ta0020, and Tr0711. After 3 days of culture at 28 °C, Tr0711 presented the highest production of spores, followed by Th0179, Th3844, and Ta0020 (Supplementary Material 1: Supplementary Fig. 2). Additionally, after 7 days of culture at 28 °C, the four strains showed a different growth pattern, i.e., Th3844 (Fig. 1a) presented a production of spores similar to Th0179 (Fig. 1b) and Tr0711 (Fig. 1c), and Ta0020 (Fig. 1d) showed slow sporulation. The DNA from the evaluated *Trichoderma* isolates was then extracted, and its integrity and quality were assessed (Supplementary Material 1: Supplementary Table 1 and Supplementary Material 1: Supplementary Fig. 3).

3.3 Assembled genomic features and general comparison across **Trichoderma** *spp.*

In the present study, we introduced the whole-genome sequences of Th3844, Th0179, Ta0020, and Tr0711 (Table 1). Overall, the genomes of the evaluated *Trichoderma* spp. varied in a number of contigs (14–26), sizes (32–40 Mb) and gene contents (8,796–11,322 genes). In comparison with the other strains, Th0711 contains the smallest gene repertoire, while Th3844 contains the highest gene repertoire. To assess the completeness and integrity of the assembled genomes, BUSCO analysis was performed. For all evaluated strains, over 90% of genes were complete. Although the genome of Th3844 presented a considerable number of missing genes (9.7%), its assembled genome exhibited a lower degree of fragmentation compared to that of other genomes.

Considering the genome sequencing coverage, genome size, GC content, length metrics (N50 and L50 values), assembly level, and number of genes, the genomes of Th3844, Th0179, Ta0020, and Tr0711 were compared to other fungal genome references (Baroncelli et al. 2015; Chung et al. 2021; Kubicek et al. 2011; Li et al. 2017; Martinez et al. 2008) (Table 2). All *T. harzianum* genomes were similar in size and GC content. The same profile was observed for the *T. atroviride* and *T. reesei* genomes. Large differences were found for the genome sequencing coverage, in which Th3844, Th0179, Ta0020, and Tr0711 presented higher values than those reported for other strains in the literature (Baroncelli et al. 2015; Kubicek et al. 2011; Li et al. 2017). In regard to quality, except for Tr0711, the genomes assembled in this study showed a lower degree of fragmentation compared to those previously available (Baroncelli et al. 2015; Chung et al. 2021; Kubicek et al. 2011; Li et al. 2017; Martinez et al. 2008).

We also performed alignment analyses to evaluate the similarities and variations between the genomes of the studied strains (I) and *T. reesei* QM6a, which is a model organism for the lignocellulose deconstruction system and has a genome that is assembled at the chromosomal level (Li et al. 2017) (Fig. 2), and (II) and their respective reference genomes (Supplementary Material 1: Supplementary Fig. 4). The profile of alignment across the genomes illustrated the degree of divergence across the studied strains with *T. reesei* QM6a and with the closest related strain of each evaluated fungus. For each evaluated strain, we observed alignment in different regions of the *T. reesei* QM6a reference genome. Using *T. reesei* QM6a as a reference genome, we found a total of (I) 12% aligned bases for Th3844, (II) 13% aligned bases for Th0179, (III) 8% aligned bases for Ta0020, and (IV) 95% aligned bases for Tr0711. The total percentage of aligned bases for the Th3844 and Th0179 genomes with the *T. harzianum* T6776

genome was approximately 83.5% and 89%, respectively. In addition, for the Ta0020 genome, the total percentage of aligned bases to the *T. atroviride* IMI206040 reference genome was approximately 88%.

3.4 Functional annotation, secondary metabolite gene cluster diversity, and CAZyme distribution and quantification in evaluated **Trichoderma** *spp. genomes*

After performing the genome assembly and gene prediction steps, functional annotation was accomplished by a homology search. The functional category distribution regarding the clusters of orthologous groups of proteins (COGs) (Tatusov et al. 2000) is shown in Fig. 3 and Supplementary Material 1: Supplementary Table 2. Disregarding the (S) function unknown category, the top 5 functional categories were (G) carbohydrate metabolism and transport, (O) posttranslational modification, protein turnover, chaperone functions, (Q) secondary structure, (E) amino acid transport and metabolism, and (U) intracellular trafficking, secretion, and vesicular transport. Overall, the evaluated strains seem to have similar COG assignment profiles. The complete functional annotation of the four genomes is available in Supplementary Material 2: Supplementary Table 3.

AntiSMASH predicted the presence of different types of biosynthetic gene clusters (BGCs) that encode potential secondary metabolites (SMs). The BGCs of Th3844, Th0179, Ta0020, and Tr0711 are listed in Supplementary Material 3: Supplementary Table 4. In total, (I) 48, (II) 63, (III) 42, and (IV) 30 BGCs were predicted for (I) Th3844, (II) Th0179, (III) Ta0020, and (IV) Tr0711, respectively. Among these, (I) 17, (II) 21, (III) 10, and (IV) 9 clusters showed similarity to BGCs with a known function. For Th3844, five gene clusters containing 100% of the genes from the known cluster were identified (Table 3), whereas for the other strains, six (Th0179), four (Ta0020), and three (Tr0711) gene clusters were recognized.

The BGCs identified could be categorized into several classes, including the following: fungal RiPP with POP or UstH peptidase types and a modification (fungal-RiPP); nonribosomal peptide synthetases (NRPSs); NRPS-like fragment (NRPS-like); type I polyketide synthase (T1PKS); terpene; hybrid NRPS, T1PKS; NRPS, T1PKS, beta-lactone; NRPS, T1PKS, NRPS-like; NRPS-like; and terpene, NRPS-like. Overall, among the tested strains, Th0179 showed the highest number of BGCs, followed by Th3844, Ta0020, and Tr0711 (Fig. 4). Additionally, the T1PKS class was the most represented among all evaluated strains. Interestingly, the NRPS class was more highly represented in Th0179 compared with the other strains.

Within filamentous fungi, the genus *Trichoderma* is a model system for the production of CAZymes, which include glycoside hydrolases (GHs), carbohydrate esterases (CEs), glycosyltransferases (GTs), polysaccharide lyases (PLs), auxiliary activities (AAs), and carbohydrate-binding modules (CBMs) (Cantarel et al. 2009). Therefore, we performed a classification at the genomic level of the total CAZyme content among the new potential biotechnological strains evaluated in this study (Supplementary Material 4: Supplementary Table 5). To detect similarities and differences between the strains, their CAZyme profiles were compared (Fig. 5a). Overall, the results indicated a great variability in the CAZyme content related to the biomass degradation process and mycoparasitism among the Th3844, Th0179, Ta0020, and Tr0711 genomes, as explored next. Among the main CAZyme classes detected in all strains, GHs were overrepresented, and Th3844 (256) and Th0179 (252) had the highest number, followed by Ta0020 (230) and Tr0711 (184). This CAZyme class was followed by GTs as follows: (I) 88 (Th3844), (II) 90 (Th0179), (III) 90 (Ta0020), and (IV) 86 (Tr0711). However, the analysis of the CAZymes with predicted signal peptides revealed that only a few GTs were secreted, as follows: (I) 8 (Th3844), (II) 9 (Th0179), (III) 7 (Ta0020), and (IV) 4 (Tr0711) (Fig. 5b). The distribution of the different CAZyme families among strains was investigated and is available in Supplementary Material 1: Supplementary Fig. 5.

We also investigated the expression profile of the CAZymes identified in the assembled genomes under cellulose and glucose growth conditions. Considering the differentially expressed genes (DEGs) under cellulose growth conditions, we found 35, 78, 31, and 14 CAZyme genes for Th0179, Th3844, Ta0020, and Tr0711, respectively. The classification of the CAZyme genes of Th0179, Th3844, Ta0020, and Tr0711 under cellulose fermentative conditions along with their gene ID number, protein product number, corresponding homolog in the reference genome, fold change values, e-values, and potentially secreted protein are described in Supplementary Material 1: Supplementary Tables 6-9, respectively. The DEGs under cellulose growth conditions that presented orthologs across Th3844, Th0179, and Ta0020 are represented in Fig. 5c, and the identification of the corresponding ortholog for each strain is shown in Supplementary Material 1: Supplementary Table 10. Curiously, Tr0711 showed only one CAZyme-encoding gene from the AA3_2 subfamily with orthologs in the other evaluated strains (Th3844_011000, Th0179_011191, Ta0020_000129, and Tr0711_006613). Overall, Th3844 presented more CAZymes with upregulated expression under cellulose degradation conditions than Th0179 and Ta0200, whereas a significant number of CAZymes from the last strain exhibited downregulated expression under such growth conditions (Fig. 5c). We would also like to highlight that important hemicellulases, such as GH10 and GH11,

showed a notable difference in their expression profile between Ta0020 and Th3844. The same finding was observed for AA9 with the CBM1 domain (Fig. 5c). In contrast, CAZymes from the GT4 and GH20 families that exhibited upregulated expression under cellulose degradation conditions were identified for Ta0020 (Fig. 5c).

Moreover, the CAZyme families related to biomass degradation as well as mycoparasitism were investigated in greater depth (Fig. 6). In relation to lignocellulose depolymerization, CAZymes from the GH5, AA1, AA3, GH2, and GH3 families were well represented among all the strains, and the highest numbers were found in Th3844 and Th0179 (Fig. 6a). Regarding mycoparasitism activity, although CAZymes from the GH18 family, which are related to chitin degradation, were present in all evaluated strains, Tr0711 exhibited the smallest number of such enzymes in its genome. Additionally, although other CAZyme classes, which are related to mycoparasitic interactions, were present in the genomes of Th3844, Th0179, Ta0020, and Tr0711, the amount of each CAZyme family differed among the strains (Fig. 6b).

3.5 Orthology analysis, phylogenetic profiling, and structural variant analyses

The predicted proteomes of Th3844, Th0179, Ta0020, and Tr0711 were compared with those of 19 other *Trichoderma* spp. and with those from more phylogenetically distant fungi, including *Fusarium* spp., *Aspergillus* spp., and *Neurospora* spp. (outgroups) (Supplementary Material 5: Supplementary Table 11). We identified a total of 18,349 orthogroups, which encompassed 313,444 genes (95%) in a total of 329,927 genes, i.e., the number of unassigned genes was equal to 16,483 (5%). Moreover, we detected 3,378 orthogroups among all present species, and 1,864 of these orthogroups consisted entirely of single-copy genes. Furthermore, 408 species-specific orthogroups, i.e., orthogroups that consist entirely of genes from one species, encompassing 1,114 genes were detected. Fifty percent of all genes were in orthogroups with 29 or more genes (G50 was 29) and were contained in the largest 4,609 orthogroups (O50 was 4609). The analysis of the orthologous relationships across the evaluated strains revealed that both *T. harzianum* strains shared the highest number of orthologous genes among them than with the other strains. In relation to the other evaluated strains, Th3844 and Th0179 exhibited more orthologs in common with Ta0020 than with Tr0711 (Table 4). To explore the evolutionary history of Th3844, Th0179, Ta0020, and Tr0711, a rooted species tree was inferred using the 1,864 single-copy orthologous genes conserved in the 29 analyzed fungi (Fig. 7 and Supplementary Material 1: Supplementary Fig. 6). Such phylogenetic analysis indicated that Tr0711 was most closely related to *Trichoderma parareesei* CBS 125925, while Ta0020 was most closely

related to *T. atroviride* IMI206040. On the other hand, Th0179 and Th3844 were phylogenetically close to *Trichoderma lentiforme* CFAM-422 and to other *T. harzianum* strains (TR274 and CB226.95), respectively.

The SVs of the evaluated *Trichoderma* isolates were identified by mapping the long reads of the fungi against the reference genome of *T. reesei* QM6a (Martinez et al. 2008). A total of 12,407 (Th3844), 12,763 (Th0179), 11,650 (Ta0020), and 7,103 (Tr0711) SVs were identified for each strain, showing substitution rates of ~1/2,674 nucleotides (Th3844), ~1/2,585 nucleotides (Th0179), ~1/2,832 nucleotides (Ta0020), and ~1/4,655 nucleotides (Tr0711). These SVs included different phenomena that affect gene sequences, such as break ends, deletions, multiple nucleotides and InDels, duplications, insertions, and inversions (Supplementary Material 6: Supplementary Table 12 and Fig. 8a). Compared with the other evaluated strains, Tr0711 presented a low number of SVs, while Th0179 displayed the highest number. For all evaluated strains, the most presented SV categories were multiple nucleotides and an InDel, followed by deletions and insertions (Fig. 8a).

To thoroughly investigate the functional effects of the identified SVs, we performed an annotation of the structural rearrangements, which were placed into different classes based on their predicted effects on protein function. Details of these effects can be found in Supplementary Material 7: Supplementary Tables 13-16, and the most prevalent effects are represented in Fig. 8b. For all evaluated strains, the majority of variants presented a modifier impact, which was higher at downstream and upstream genomic locations. Such an effect was more accentuated for both *T. harzianum* strains. In addition, SVs present in transcripts, genes, and intergenic regions were well represented for Ta0020.

With the aim of elucidating the SVs identified against the *T. reesei* QM6a reference genome in the evaluated genomes, we selected the CAZymes identified as DEGs under cellulose and glucose growth conditions and with orthologs in Th3844, Th0179, and Ta0020 for detailed analyses. After searching for corresponding homologs in the *T. reesei* QM6a reference genome, we investigated SVs in the following genes: GH72 (XM_006965422.1), GH1 (XM_006963374.1), GH20 (XM_006963001.1), GH16 (XM_006966546.1), GH1 (XM_006964776.1), and GH20 (XM_006969114.1). Considering such genes, we identified SVs in (I) 4 (Th3844), (II) 5 (Th0179), and (III) 4 (Ta0020) genes, and the majority had modifier impact.

4 Discussion

Diversity within the genus *Trichoderma* is evident from the wide range of phenotypes exhibited by the fungi as well as the various ecological roles and industrial purposes the fungi serve (Nakkeeran et al. 2021). Because of their various applications, different *Trichoderma* species have become model organisms for a broad spectrum of physiological phenomena, such as plant cell wall degradation and enzyme production (Fang et al. 2021), biocontrol (Zin and Badaluddin 2020), and response to light (Schmoll 2018). Within the genus *Trichoderma*, *T. harzianum* has been used as a commercial biocontrol agent against plant diseases (Fraceto et al. 2018). In addition to their mycoparasitic activities, hydrolytic enzymes from *T. harzianum* strains have demonstrated great potential in the conversion of lignocellulosic biomass into fermentable sugars (Almeida et al. 2021; Delabona et al. 2020; Motta et al. 2021; Zhang et al. 2020). Recently, different types of enzymatic profiles across *Trichoderma* species were reported, in which Th3844 and Th0179 presented a higher hydrolytic potential during growth on cellulose than Ta0020 (Almeida et al. 2021); furthermore, differences between Th3844 and Th0179 concerning the transcriptional regulation coordinated by XYR1 and CRE1 during cellulose degradation were reported (Rosolen et al. 2022). Because such diversity in enzyme response might be related to transcriptomic and genomic differences, we aimed to provide foundations for further studies that could investigate such variations at a genomic level.

Herein, we presented high-quality genome assemblies for two *T. harzianum* strains with hydrolytic potential (Th3844 and Th0179) and, for comparative purposes, for a mycoparasitic species (Ta0020) and saprotrophic species (Tr0711). Thus, *T. reesei* and *T. atroviride* strains were used to assess the genetic differences in the genus *Trichoderma*. With the aim of obtaining high-quality genomes, we employed Canu software for the genome assemblies, and this software has been applied with success in the assembly of fungal genomes (Courtine et al. 2020; Gan et al. 2020; Montoliu-Nerin et al. 2020). Except for Tr0711, the resulting genomic assemblies displayed the highest coverage scores and the lowest fragmentation values compared to those of *Trichoderma virens* Gv29-8 (Kubicek et al. 2011) as well as to *T. harzianum* T6776 (Baroncelli et al. 2015) and *T. atroviride* IMI206040 (Kubicek et al. 2011), which were used as the reference genome in preceding studies from our research group (Almeida et al. 2021; Horta et al. 2018). Although the genomes were not assembled at the chromosome level, the quality of the Th3844, Th0179, Ta0020, and Tr0711 genome assemblies based on the BUSCO value was over 90%. Only the Th3844 genome exhibited 9.7% missing genes. However, even chromosome-level genome assembly does not necessarily achieve a complete BUSCO score (Chung et al. 2021). This may be because the assembly is not 100% accurate, but at the same time, the BUSCO value may not be a perfect indicator to assess genomic qualities.

Despite its limitations, without a definitive alternative, BUSCO is still an essential genomic quality assessment tool that includes up-to-date data from many species.

After evaluating the quality of the four assembled genomes, we performed gene prediction and functional annotation for the datasets. The ecological behavior of the mycoparasites *T. atroviride* and *T. virens*, compared to the plant wall degrader *T. reesei*, is reflected by the sizes of the respective genomes; *T. atroviride* (36 Mb) and *T. virens* (39 Mb) were somewhat larger than the weakly mycoparasitic *T. reesei* (33 Mb) (Kubicek et al. 2011; Martinez et al. 2008). Herein, compared to Th3844 (40 Mb), Th0179 (39 Mb), and Ta0020 (36 Mb), the genome of Tr0711 (32 Mb) was smaller, which might be conceivably as the gene function is lost to mycoparasitism during the evolution of *T. reesei* (Kubicek et al. 2011). In relation to the number of genes, our results showed that Tr0711 presented a smaller gene content than Th3844, Th0179, and Ta0020. Such findings corroborated a previous study (Xie et al. 2014), in which 9,143 and 11,865 genes were predicted for *T. reesei* and *T. atroviride*, respectively. In relation to the *T. reesei* QM6a reference genome, the genomes of Th3844, Th0179, Ta0020, and Tr0711 displayed significant structural reorganization, which was more greatly accentuated by an increased phylogenetic distance. Interestingly, this structural reorganization was also observed within strains of the same species, highlighting their genetic diversity.

4.1 Comparative and functional genomics

To obtain insights regarding the functional profile of Th3844, Th0179, Ta0020, and Tr0711, COG analyses of proteins from their genomes were performed. “Carbohydrate metabolism and transport” was a notable COG term for all evaluated strains, suggesting that the genomic arsenal of these fungi is connected to their ability to use carbon sources that are available in the environment. Such characteristics are well known for the saprophytic fungal *T. reesei* (Arntzen et al. 2020), and recent studies have observed the same characteristics for *T. harzianum* (Almeida et al. 2021; Delabona et al. 2020). “Posttranslational modification, protein turnover, chaperone functions” was the second most notable COG term that was present in the four evaluated genomes. Posttranslational modifications (PTMs), which are used by eukaryotic cells to diversify their protein functions and dynamically coordinate their signaling networks, encompass several specific chemical changes that occur on proteins following their synthesis (Ramazi and Zahiri 2021). The production of intact and functional proteins is a prerequisite for large-scale protein production, and extensive host-specific PTMs often affect the catalytic properties and stability of recombinant

enzymes. The high extracellular secretion capability and eukaryotic PTM machinery make *Trichoderma* spp. particularly interesting hosts (Wei et al. 2021). In this context, PTMs are a major factor in the cellulolytic performance of fungal cellulases (Amore et al. 2017; Beckham et al. 2012; Dana et al. 2014), and the impact of plant PTMs on the enzyme performance and stability of the major cellobiohydrolase Cel7A from *T. reesei* has already been determined (van Eerde et al. 2020). In addition, PTMs, especially phosphorylation, of the proteins involved in plant biomass degradation, including CRE1, play an essential role in signal transduction to achieve carbon catabolite repression (CCR) (Han et al. 2020; Horta et al. 2019). Thus, describing this class of *Trichoderma* genomes is essential to understand the impact of alternative PTMs on the catalytic performance and stability of recombinant enzymes. We also would like to highlight that "secondary structure" and "amino acid transport and metabolism", which are related to PTMs, were overrepresented COG terms.

SMs from microorganisms may play an antifungal role against agriculturally important phytopathogenic fungi (Khan et al. 2020a). Many *Trichoderma* species are the most prominent producers of SMs with antimicrobial activity against phytopathogenic fungi (Khan et al. 2020b; Morais et al. 2022). Therefore, we performed genome-wide identification and analysis of SMGCs in the Th3844, Th0179, Ta0020, and Tr0711 genomes. Considering the diversity of bioactive molecules isolated from the genus and given the vast biosynthetic potential that emerged from the antiSMASH analysis conducted in our study, the four evaluated strains, particularly Th0179, have high potential to produce bioactive molecules that warrant their use as biocontrol agents against plant pathogens. Furthermore, in both *T. harzianum* strains investigated in this study, namely, Th3844 and Th0179, genome mining identified the biosynthesis of tricholignan A, which is a natural product that helps plants assimilate iron from the soil (Chen et al. 2019). Similar results were reported for the biofertilizer strain *T. harzianum* t-22 (Chen et al. 2019). Detailed information about these SMs, when grouped together, enhances the understanding of their efficient utilization and allows further exploration of new bioactive compounds for the management of plant pathogenic fungi.

The comparative genomics of *Trichoderma* spp. suggested that mycoparasitic strains, such as *T. vires* and *T. atroviride*, presented a set of genes, including CAZymes and genes encoding SMs, that were more highly expressed and related to mycoparasitism than the saprotrophic species *T. reesei* (Kubicek et al. 2011). Although such fungi are widely used in industry as a source of cellulases and hemicellulases, they have a smaller arsenal of genes that encode the CAZymes related to biomass deconstruction than other lignocellulolytic fungi (Martinez et al. 2008). Regarding the CAZyme content, the results found here for the Th3844, Th0179, Ta0020, and Tr0711

genomes follow the same profile as that of a previous study (Fanelli et al. 2018), in which the CAZyme genetic endowment of some strains from *T. harzianum*, including B97 and T6776, was significantly higher than that of *T. atroviride* IMI206040, *T. reesei* QM6a, and *T. virens* Gv-29-8. However, by normalizing the CAZyme counts by the total gene counts for each strain, we found similar values among the evaluated fungi, as follows: (I) 3.8% (Th3844), (II) 3.6% (Th0179), (III) 3.7% (Ta0020), and 3.7% (Tr0711). Such results indicate that the probable differences regarding the CAZyme distribution might be related to the specific CAZyme families of a strain, as explored below, and not necessarily related to the total CAZyme content across the compared genomes. Furthermore, the presence of putative CAZyme-encoding genes in the genomes of Th0179 and Th3844 provides insight into its lignocellulose-degrading enzyme potential but cannot be directly related to its real degradation ability. In fact, since fungal species rely on different strategies, it has been observed that the number of genes related to the degradation of a given polysaccharide is not necessarily correlated to the extent of its degradation (Arntzen et al. 2020; Kjaerbolling et al. 2020). For this reason, CAZy analysis is associated with functional approaches, such as enzymatic activity assays, which provide valuable insight into the actual behavior of the concerned species on specific lignocellulose substrates.

In relation to the CAZyme families that are directly associated with the deconstruction of plant biomass, the genomes of Th3844, Th0179, Ta0020, and Tr0711 contained genes encoding GH5, which includes cellulases that are most highly abundant in fungi (Li and Walton 2017); GH3, which includes β-glucosidases that are frequently secreted into the medium (Guo et al. 2016); and AA3, which is a member of the enzyme arsenal that is auxiliary to GHs (Levasseur et al. 2013). Lytic polysaccharide monooxygenases (LPMOs), which are classified into the CAZy auxiliary activity families AA9-AA11 and AA13-AA16, are copper-dependent enzymes that also perform important roles in lignocellulose degradation (Couturier et al. 2018; Monclaro and Filho 2017). Herein, each of the Th3844, Th0179, Ta0020, and Tr0711 genomes exhibited three AA9 and two AA14 enzymes. Compared to other fungi, the genomes of *Trichoderma* species harbor a high number of chitinolytic genes (Kubicek et al. 2011, 2019), reflecting the importance of these enzymes in the mycoparasitic characteristics of fungi. From the *Trichoderma* genomes that have been analyzed in detail thus far, the fungal chitinases that belong to the family GH 18 are significantly expanded in *T. virens*, *T. atroviride*, *T. harzianum*, *Trichoderma asperellum*, *Trichoderma gamsii*, and *Trichoderma atrobrunneum* (Fanelli et al. 2018; Kubicek et al. 2011). Similarly, the number of chitosanases (GH75) is enhanced, and there are at least five corresponding genes; in contrast, most other fungi have only one or two corresponding

genes (Kappel et al. 2020; Kubicek et al. 2011). Furthermore, β-1,3-glucanases that belong to GH families 16, 17, 55, 64, and 81 are expanded in *Trichoderma* mycoparasites compared to other fungi (Fanelli et al. 2018; Kubicek et al. 2011). Here, the CAZyme families that are related to mycoparasitic activity were present in the four genomes studied, and the most were in Th3844 and Th0179.

The transcriptome analysis identified the DEGs and the differentially expressed CAZyme genes in all strains. Although *T. reesei* is the primary source of hydrolytic enzymes that form the enzymatic cocktails, Tr0711 exhibited the smallest number of identified DEGs under cellulose growth conditions, whereas Th3844 presented the highest number, followed by Th0179 and Ta0020. Each strain presented a different set of genes with different expression levels, which can be attributed to between-strain differences in the regulatory mechanisms of hydrolysis.

4.2 Insights into the evolutionary history

Comparative genomics could be a powerful tool for studying fungal evolution and promoting insights into their genetic diversity. In this context, identifying orthologous relationships among sequences is an essential step for obtaining a more thorough understanding of the genetic correlation of particular fungi. Thus, by applying orthology analyses, we could identify orthogroups and orthologs between the evaluated strains, as well as among some other *Trichoderma* spp. and filamentous fungi that are more genetically distant. In the genus *Trichoderma*, several lifestyles have been documented, including saprotrophy, which is a lifestyle that is also observed in other filamentous fungi, such as *Neurospora* spp., *Aspergillus* spp., and *Fusarium* spp. (Arntzen et al. 2020; Correa et al. 2020; Najjarzadeh et al. 2021). Thus, the proteomes of some species and strains of such genera were also included in the orthology analysis. Through our results, we may infer that some genus-specific genes are necessary for specific lifestyles and are shared by fungi that have the same lifestyle but are in quite different evolutionary orders. The phylogenetic tree modeled from the orthologous analysis revealed a low bootstrap value for the clade formed by aligning the sequence proteomes from *T. harzianum*. Furthermore, the presence of other *Trichoderma* species, such as *T. lentiforme*, *T. guizhouense*, and *T. simmonsii*, was observed. These observations might be explained by the complex speciation process within the *T. harzianum* species group (Druzhinina et al. 2010); therefore, the phylogenetic position is uncertain for fungi from these species. However, the molecular identification of Th3844 and Th0179 based on the ITS and *tef1* sequences has already been reported (Rosolen et al. 2022), which confirmed that

both strains were phylogenetically close to other *T. harzianum* strains. Tr0711 was grouped with *T. pararessei* CBS 125925, and both were phylogenetically in proximity to other *T. reesei* strains.

While short reads perform well in the identification of single nucleotide variants (SNVs) and small insertions and deletions (InDels), they are not well suited for detecting changes in larger sequences (Mahmoud et al. 2019). SVs, which include insertions, deletions, duplications, inversions, or translocations that affect ≥ 50 bp (Mills et al. 2011), are more amenable to long-read sequencing (Mitsuhashi and Matsumoto 2020; Sakamoto et al. 2020). In fungi, SV analysis was employed successfully using reads from third-generation sequencing (Badet et al. 2021; Basile et al. 2021). In this study, we detected SVs by aligning our four genomes against *T. reesei* QM6a (Martinez et al. 2008); although the genomes were assembled at a scaffolding level, unlike the *T. reesei* QM6a reference genome, which was assembled at the chromosome level, we opted to proceed with this dataset because its annotation file was available (Li et al. 2017). As expected, due to phylogenetic proximity to the reference genome (Rosolen et al. 2022), Tr0711 exhibited a lower number of SVs than the other strains. However, although *T. atroviride* is phylogenetically distant from *T. reesei* (Rosolen et al. 2022), Ta0020 exhibited fewer SVs than both *T. harzianum* strains, and this result might be explained by the uncertain phylogenetic position of fungi in these species (Druzhinina et al. 2010). Although the *T. harzianum* strains are phylogenetically close (Rosolen et al. 2022), comparison of the SVs identified from the mapping of both genomes against *T. reesei* QM6a reveals genetic variability across the strains (Martinez et al. 2008). Such results are consistent with the findings of previous studies in that genetic variations between fungal strains of the same species are not uncommon (Andersen et al. 2011; de Vries et al. 2017; Thanh et al. 2019).

Considering their basic and economic importance, the high-quality genomes found herein might be helpful for better understanding the diversity within the genus *Trichoderma*, as well as improving the biotechnological applications of such fungi. Furthermore, the comparative study of multiple related genomes might be helpful for understanding the evolution of genes that are related to economically important enzymes and for clarifying the evolutionary relationships related to protein function.

Statements and declarations

Funding

Financial support for this work was provided by the São Paulo Research Foundation (FAPESP - Process number 2015/09202-0 and 2018/19660-4) and the Coordination for the Improvement of Higher Education Personnel (CAPES, Computational Biology Program - Process number 88882.160095/2013-01). RRR received a PhD fellowship from CAPES (88887.482201/2020-00) and FAPESP (2020/13420-1), PHCA received a PhD fellowship from CAPES (88887.612254/2021-00), MACH received a postdoctoral fellowship from FAPESP (2020/10536-9), and APS received a research fellowship from the Brazilian National Council for Technological and Scientific Development (CNPq- Process number 312777/2018-3).

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Ethics approval

Not applicable.

Informed consent

Not applicable.

Data availability

All data generated or analyzed in this study are included in this published article (and its supplementary information files). The raw datasets and the assembled genomes were deposited in the NCBI Sequence Read Archive and can be accessed under BioProject number PRJNA781962 and BioSample number as follows: SAMN23309297 (Th3844), SAMN23309298 (Th0179), SAMN23309299 (Ta0020), and SAMN23309300 (Tr0711).

Authors' contribution

RRR: Writing – original draft, methodology, formal analysis, and visualization. **MACH:** Conceptualization, methodology, and writing – review & editing. **PHCA:** Methodology and resources. **CCS:** Methodology and formal

analysis. **DAS:** Methodology and resources. **GHG:** Writing - review & editing. **APS:** Conceptualization, supervision, review & editing, and funding acquisition.

References

Afgan E, Baker D, Batut B et al (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46:W537-W544. https://doi.org/10.1093/nar/gky379

Almeida DA, Horta MAC, Ferreira Filho JA, Murad NF, de Souza AP (2021) The synergistic actions of hydrolytic genes reveal the mechanism of *Trichoderma harzianum* for cellulose degradation. J Biotechnol 334:1-10. https://doi.org/10.1016/j.jbiotec.2021.05.001

Amore A, Knott BC, Supekar NT et al (2017) Distinct roles of N- and O-glycans in cellulase activity and stability. Proc Natl Acad Sci U S A 114:13667-13672. https://doi.org/10.1073/pnas.1714249114

Andersen MR, Salazar MP, Schaap PJ et al (2011) Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. Genome Res 21:885-897. https://doi.org/10.1101/gr.112169.110

Ardui S, Ameur A, Vermeesch JR, Hestand MS (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 46:2159-2168. https://doi.org/10.1093/nar/gky066

Arntzen MO, Bengtsson O, Varnai A, Delogu F, Mathiesen G, Eijsink VGH (2020) Quantitative comparison of the biomass-degrading enzyme repertoires of five filamentous fungi. Sci Rep 10:20267. https://doi.org/10.1038/s41598-020-75217-z

Badet T, Fouche S, Hartmann FE, Zala M, Croll D (2021) Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen. Nat Commun 12:3551. https://doi.org/10.1038/s41467-021-23862-x

Baroncelli R, Piaggeschi G, Fiorini L, Bertolini E, Zapparata A, Pe ME, Sarrocco S, Vannacci G (2015) Draft whole-genome sequence of the biocontrol agent Trichoderma harzianum T6776. Genome Announc 3:e00647-15. https://doi.org/10.1128/genomeA.00647-15

Basile A, De Pascale F, Bianca F, Rossi A, Frizzarin M, De Bernardini N, Bosaro M, Baldisseri A, Antoniali P, Lopreiato R, Treu L, Campanaro S (2021) Large-scale sequencing and comparative analysis of oenological *Saccharomyces cerevisiae* strains supported by nanopore refinement of key genomes. Food Microbiol 97:103753. https://doi.org/10.1016/j.fm.2021.103753

Beckham GT, Dai Z, Matthews JF, Momany M, Payne CM, Adney WS, Baker SE, Himmel ME (2012) Harnessing glycosylation to improve cellulase activity. Curr Opin Biotechnol 23:338-345. https://doi.org/10.1016/j.copbio.2011.11.030

Bischof RH, Ramoni J, Seiboth B (2016) Cellulases and beyond: the first 70 years of the enzyme producer Trichoderma reesei. Microb Cell Fact 15:106. https://doi.org/10.1186/s12934-016-0507-6

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 49:W29-W35. https://doi.org/10.1093/nar/gkab335

Broad Institute Picard tools. http://broadinstitute.github.io/picard/

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59-60. https://doi.org/10.1038/nmeth.3176

Busk PK, Pilgaard B, Lezyk MJ, Meyer AS, Lange L (2017) Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. BMC Bioinformatics 18:214. https://doi.org/10.1186/s12859-017-1625-9

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. bioRxiv 2021.06.03.446934. https://doi.org/10.1101/2021.06.03.446934

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37:D233-D238. https://doi.org/10.1093/nar/gkn663

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18:188-196. https://doi.org/10.1101/gr.6743907

Chaverri P, Branco-Rocha F, Jaklitsch W, Gazis R, Degenkolb T, Samuels GJ (2015) Systematics of the *Trichoderma harzianum* species complex and the re-identification of commercial biocontrol strains. Mycologia 107:558-590. https://doi.org/10.3852/14-147

Chen M, Liu Q, Gao SS, Young AE, Jacobsen SE, Tang Y (2019) Genome mining and biosynthesis of a polyketide from a biofertilizer fungus that can facilitate reductive iron assimilation in plant. Proc Natl Acad Sci U S A 116:5499-5504. https://doi.org/10.1073/pnas.1819998116

Chung D, Kwon YM, Yang Y (2021) Telomere-to-telomere genome assembly of asparaginase-producing *Trichoderma simmonsii*. BMC Genom. 22:830. https://doi.org/10.1186/s12864-021-08162-4

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92. https://doi.org/10.4161/fly.19695

Correa CL, Midorikawa GEO, Filho EXF, Noronha EF, Alves GSC, Togawa RC, Silva-Junior OB, Costa M, Grynberg P, Miller RNG (2020) Transcriptome profiling-based analysis of carbohydrate-active enzymes in aspergillus terreus involved in plant biomass degradation. Front Bioeng Biotechnol 8:564527. https://doi.org/10.3389/fbioe.2020.564527

Courtine D, Provaznik J, Reboul J, Blanc G, Benes V, Ewbank JJ (2020) Long-read only assembly of *Drechmeria coniospora* genomes reveals widespread chromosome plasticity and illustrates the limitations of current nanopore methods. Gigascience 9:giaa099. https://doi.org/10.1093/gigascience/giaa099

Couturier M, Ladeveze S, Sulzenbacher G et al (2018) Lytic xylan oxidases from wood-decay fungi unlock biomass degradation. Nat Chem Biol 14:306-310. https://doi.org/10.1038/nchembio.2558

Crucello A, Sforca DA, Horta MA, dos Santos CA, Viana AJ, Beloti LL, de Toledo MA, Vincentz M, Kuroshu RM, de Souza AP (2015) Analysis of genomic regions of *Trichoderma harzianum* IOC-3844 related to biomass degradation. PLoS One 10:e0122122. https://doi.org/10.1371/journal.pone.0122122

Dana CM, Dotson-Fagerstrom A, Roche CM, Kal SM, Chokhawala HA, Blanch HW, Clark DS (2014) The importance of pyroglutamate in cellulase Cel7A. Biotechnol Bioeng 111:842-847. https://doi.org/10.1002/bit.25178

de Vries RP, Riley R, Wiebenga A et al (2017) Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus*. Genome Biol 18:28. https://doi.org/10.1186/s13059-017-1151-0

Delabona PDS, Codima CA, Ramoni J, Zubieta MP, de Araujo BM, Farinas CS, Pradella J, Seiboth B (2020) The impact of putative methyltransferase overexpression on the *Trichoderma harzianum* cellulolytic system for biomass conversion. Bioresour Technol 313:123616. https://doi.org/10.1016/j.biortech.2020.123616

Dodd SL, Lieckfeldt E, Samuels GJ (2003) *Hypocrea atroviridis* sp. nov., the teleomorph of *Trichoderma atroviride*. Mycologia 95:27-40. https://doi.org/10.1080/15572536.2004.11833129

Druzhinina IS, Chenthamara K, Zhang J et al (2018) Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus *Trichoderma* from its plant-associated hosts. PLoS Genet 14:e1007322. https://doi.org/10.1371/journal.pgen.1007322

Druzhinina IS, Kopchinskiy AG, Kubicek CP (2006) The first 100 *Trichoderma* species characterized by molecular data. Mycoscience 47:55-64. https://doi.org/10.1007/s10267-006-0279-7

Druzhinina IS, Kubicek CP, Komon-Zelazowska M, Mulaw TB, Bissett J (2010) The *Trichoderma harzianum* demon: complex speciation history resulting in coexistence of hypothetical biological species, recent agamospecies and numerous relict lineages. BMC Evol Biol 10:94. https://doi.org/10.1186/1471-2148-10-94

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. https://doi.org/10.1186/s13059-015-0721-2

Emms DM, Kelly S (2017) STRIDE: species tree root inference from gene duplication events. Mol Biol Evol 34:3267-3278. https://doi.org/10.1093/molbev/msx259

Emms DM, Kelly S (2018) STAG: species tree inference from all genes. bioRxiv 267914. https://doi.org/10.1101/267914

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. https://doi.org/10.1186/s13059-019-1832-y

Fanelli F, Liuzzi VC, Logrieco AF, Altomare C (2018) Genomic characterization of *Trichoderma atrobrunneum* (*T. harzianum* species complex) ITEM 908: insight into the genetic endowment of a multi-target biocontrol strain. BMC Genom. 19:662. https://doi.org/10.1186/s12864-018-5049-3

Fang H, Li C, Zhao J, Zhao C (2021) Biotechnological advances and trends in engineering Trichoderma reesei towards cellulase hyperproducer. Biotechnol Bioprocess Eng 26:517-528. https://doi.org/10.1007/s12257-020-0243-y

Filho JAF, Horta MAC, Beloti LL, Dos Santos CA, de Souza AP (2017) Carbohydrate-active enzymes in *Trichoderma harzianum*: a bioinformatic analysis bioprospecting for key enzymes for the biofuels industry. BMC Genom. 18:779. https://doi.org/10.1186/s12864-017-4181-9

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29-W37. https://doi.org/10.1093/nar/gkr367

Fraceto LF, Maruyama CR, Guilger M, Mishra S, Keswani C, Singh HB, de Lima R (2018) Trichoderma harzianum-based novel formulations: potential applications for management of Next-Gen agricultural challenges. J. Chem. Technol. Biotechnol. 93:2056-2063. https://doi.org/10.1002/jctb.5613

Gan X, Cao D, Zhang Z, Cheng S, Wei L, Li S, Liu B (2020) Draft genome assembly of floccularia luteovirens, an edible and symbiotic mushroom on qinghai-tibet plateau. G3 (Bethesda) 10:1167-1173. https://doi.org/10.1534/g3.120.401037

Guo B, Sato N, Biely P, Amano Y, Nozaki K (2016) Comparison of catalytic properties of multiple beta-glucosidases of Trichoderma reesei. Appl Microbiol Biotechnol 100:4959-4968. https://doi.org/10.1007/s00253-016-7342-x

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Hagestad OC, Hou L, Andersen JH et al (2021) Genomic characterization of three marine fungi, including *Emericellopsis atlantica* sp. nov. with signatures of a generalist lifestyle and marine biomass degradation. IMA Fungus 12:21. https://doi.org/10.1186/s43008-021-00072-0

Han L, Tan Y, Ma W, Niu K, Hou S, Guo W, Liu Y, Fang X (2020) Precision engineering of the transcription factor Cre1 in Hypocrea jecorina (Trichoderma reesei) for efficient cellulase production in the presence of glucose. Front Bioeng Biotechnol 8:852. https://doi.org/10.3389/fbioe.2020.00852

Horta MA, Vicentini R, Delabona Pda S, Laborda P, Crucello A, Freitas S, Kuroshu RM, Polikarpov I, Pradella JG, Souza AP (2014) Transcriptome profile of *Trichoderma harzianum* IOC-3844 induced by sugarcane bagasse. PLoS One 9:e88689. https://doi.org/10.1371/journal.pone.0088689

Horta MAC, Filho JAF, Murad NF, Santos EDO, Santos CAD, Mendes JS, Brandao MM, Azzoni SF, de Souza AP (2018) Network of proteins, enzymes and genes linked to biomass degradation shared by *Trichoderma* species. Sci Rep 8:1341. https://doi.org/10.1038/s41598-018-19671-w

Horta MAC, Thieme N, Gao Y et al (2019) Broad Substrate-Specific Phosphorylation Events Are Associated With the Initial Stage of Plant Cell Wall Recognition in Neurospora crassa. Front Microbiol 10:2317. https://doi.org/10.3389/fmicb.2019.02317

Kappel L, Munsterkotter M, Sipos G, Rodriguez CE, Gruber S (2020) Chitin and chitosan remodeling defines vegetative development and *Trichoderma* biocontrol. PLoS Pathog 16:e1008320. https://doi.org/10.1371/journal.ppat.1008320

Khan RAA, Najeeb S, Hussain S, Xie B, Li Y (2020a) Bioactive secondary metabolites from Trichoderma spp. against phytopathogenic fungi. Microorganisms 8:817. https://doi.org/10.3390/microorganisms8060817

Khan RAA, Najeeb S, Mao Z, Ling J, Yang Y, Li Y, Xie B (2020b) Bioactive secondary metabolites from trichoderma spp. Against phytopathogenic bacteria and root-knot nematode. Microorganisms 8:401. https://doi.org/10.3390/microorganisms8030401

Kidwai MK, Nehra M (2017) Biotechnological applications of *Trichoderma* species for environmental and food security. In: Gahlawat SK, Salar RK, Siwach P, Duhan JS, Kumar S, Kaur P (eds) Plant biotechnology: recent advancements and developments. Springer, Singapore, pp 125–156

Kjaerbolling I, Vesth T, Frisvad JC et al (2020) A comparative genomics study of 23 *Aspergillus* species from section Flavi. Nat Commun 11:1106. https://doi.org/10.1038/s41467-019-14051-y

Kolde R (2019) Pheatmap: pretty heatmaps. R package version 1.2

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722-736. https://doi.org/10.1101/gr.215087.116

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567-580. https://doi.org/10.1006/jmbi.2000.4315

Kubicek CP, Herrera-Estrella A, Seidl-Seiboth V et al (2011) Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. Genome Biol 12:R40. https://doi.org/10.1186/gb-2011-12-4-r40

Kubicek CP, Steindorff AS, Chenthamara K et al (2019) Evolution and comparative genomics of the most common *Trichoderma* species. BMC Genom. 20:485. https://doi.org/10.1186/s12864-019-5680-7

Kumar AG, Manisha D, Sujitha K, Peter DM, Kirubagaran R, Dharani G (2021) Genome sequence analysis of deep sea *Aspergillus sydowii* BOBA1 and effect of high pressure on biodegradation of spent engine oil. Sci Rep 11:9347. https://doi.org/10.1038/s41598-021-88525-9

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128. https://doi.org/10.1093/bioinformatics/btl529

Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels 6:41. https://doi.org/10.1186/1754-6834-6-41

Li B, Walton JD (2017) Functional diversity for biomass deconstruction in family 5 subfamily 5 (GH5_5) of fungal endo-beta1,4-glucanases. Appl Microbiol Biotechnol 101:4093-4101. https://doi.org/10.1007/s00253-017-8168-x

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-595. https://doi.org/10.1093/bioinformatics/btp698

Li WC, Huang CH, Chen CL, Chuang YC, Tung SY, Wang TF (2017) Trichoderma reesei complete genome sequence, repeat-induced point mutation, and partitioning of CAZyme gene clusters. Biotechnol Biofuels 10:170. https://doi.org/10.1186/s13068-017-0825-x

Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. Genome Biol 20:246. https://doi.org/10.1186/s13059-019-1828-7

Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018) MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Martinez D, Berka RM, Henrissat B et al (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat Biotechnol 26:553-560. https://doi.org/10.1038/nbt1403

Medeiros HA, Araujo Filho JV, Freitas LG, Castillo P, Rubio MB, Hermosa R, Monte E (2017) Tomato progeny inherit resistance to the nematode *Meloidogyne javanica* linked to plant growth induced by the biocontrol fungus *Trichoderma atroviride*. Sci Rep 7:40216. https://doi.org/10.1038/srep40216

Mills RE, Walter K, Stewart C et al (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59-65. https://doi.org/10.1038/nature09708

Mitsuhashi S, Matsumoto N (2020) Long-read sequencing for rare human genetic diseases. J Hum Genet 65:11-19. https://doi.org/10.1038/s10038-019-0671-8

Monclaro AV, Filho EXF (2017) Fungal lytic polysaccharide monooxygenases from family AA9: recent developments and application in lignocelullose breakdown. Int J Biol Macromol 102:771-778. https://doi.org/10.1016/j.ijbiomac.2017.04.077

Montoliu-Nerin M, Sanchez-Garcia M, Bergin C, Grabherr M, Ellis B, Kutschera VE, Kierczak M, Johannesson H, Rosling A (2020) Building de novo reference genome assemblies of complex eukaryotic microorganisms from single nuclei. Sci Rep 10:1303. https://doi.org/10.1038/s41598-020-58025-3

Morais EM, Silva AAR, Sousa FWA, Azevedo IMB, Silva HF, Santos AMG, Junior JEAB, Carvalho CP, Eberlin MN, Porcari AM, Araujo F (2022) Endophytic *Trichoderma* strains isolated from forest species of the Cerrado-Caatinga ecotone are potential biocontrol agents against crop pathogenic fungi. PLoS One 17:e0265824. https://doi.org/10.1371/journal.pone.0265824

Motta MLL, Filho JAF, de Melo RR, Zanphorlin LM, Santos CAD, de Souza AP (2021) A novel fungal metal-dependent alpha-L-arabinofuranosidase of family 54 glycoside hydrolase shows expanded substrate specificity. Sci Rep 11:10961. https://doi.org/10.1038/s41598-021-90490-2

Nagel JH, Wingfield MJ, Slippers B (2021) Increased abundance of secreted hydrolytic enzymes and secondary metabolite gene clusters define the genomes of latent plant pathogens in the *Botryosphaeriaceae*. BMC Genom. 22:589. https://doi.org/10.1186/s12864-021-07902-w

Najjarzadeh N, Matsakas L, Rova U, Christakopoulos P (2021) How carbon source and degree of oligosaccharide polymerization affect production of cellulase-degrading enzymes by Fusarium oxysporum f. sp. lycopersici. Front Microbiol 12:652655. https://doi.org/10.3389/fmicb.2021.652655

Nakkeeran S, Marimuthu T, Renukadevi P, Brindhadevi S, Jogaiah S (2021) 24 - Exploring the biogeographical diversity of *Trichoderma* for plant health. In: Jogaiah S (ed) Biocontrol agents and secondary metabolites. Woodhead Publishing, Sawston, pp 537–571

Okonechnikov K, Conesa A, Garcia-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32:292-294. https://doi.org/10.1093/bioinformatics/btv566

Oliveira RR, Viana AJ, Reategui AC, Vincentz MG (2015) Short communication an efficient method for simultaneous extraction of high-quality RNA and DNA from various plant tissues. Genet Mol Res 14:18828-18838. https://doi.org/10.4238/2015.December.28.32

Peterson DG, Tomkins JP, Frisch DA, Wing RA, Paterson AH (2000) Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. J. Agric. Genom. 5:1-100

Priest SJ, Yadav V, Heitman J (2020) Advances in understanding the evolution of fungal genome architecture. F1000Res 9:F1000 Faculty Rev–1776. https://doi.org/10.12688/f1000research.25424.1

Ramazi S, Zahiri J (2021) Posttranslational modifications in proteins: resources, tools and prediction methods. Database (Oxford) 2021:baab012. https://doi.org/10.1093/database/baab012

Rosolen RR, Aono AH, Almeida DA, Ferreira Filho JA, Horta MAC, De Souza AP (2022) Network analysis reveals different cellulose degradation strategies across Trichoderma harzianum strains associated with XYR1 and CRE1. Front Genet 13:807243. https://doi.org/10.3389/fgene.2022.807243

Sakamoto Y, Sereewattanawoot S, Suzuki A (2020) A new era of long-read sequencing for cancer genomics. J Hum Genet 65:3-10. https://doi.org/10.1038/s10038-019-0658-5

Saravanakumar K, Li Y, Yu C, Wang QQ, Wang M, Sun J, Gao JX, Chen J (2017) Effect of Trichoderma harzianum on maize rhizosphere microbiome and biocontrol of Fusarium Stalk rot. Sci Rep 7:1771. https://doi.org/10.1038/s41598-017-01680-w

Schmoll M (2018) Regulation of plant cell wall degradation by light in *Trichoderma*. Fungal Biol Biotechnol 5:10. https://doi.org/10.1186/s40694-018-0052-7

Schmoll M, Dattenbock C, Carreras-Villasenor N et al (2016) The genomes of three uneven siblings: footprints of the lifestyles of three trichoderma species. Microbiol Mol Biol Rev 80:205-327. https://doi.org/10.1128/MMBR.00040-15

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC (2018) Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15:461-468. https://doi.org/10.1038/s41592-018-0001-7

Sharma S, Kour D, Rana KL et al (2019) Trichoderma: biodiversity, ecological significances, and industrial applications. In: Yadav AN, Mishra S, Singh S, Gupta A (eds) Recent advancement in white biotechnology through fungi: volume 1: diversity and enzymes perspectives. Springer International Publishing, Cham, pp 85–120

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212. https://doi.org/10.1093/bioinformatics/btv351

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. https://doi.org/10.1093/nar/gkl200

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33-36. https://doi.org/10.1093/nar/28.1.33

Thanh VN, Thuy NT, Huong HTT, Hien DD, Hang DTM, Anh DTK, Huttner S, Larsbrink J, Olsson L (2019) Surveying of acid-tolerant thermophilic lignocellulolytic fungi in Vietnam reveals surprisingly high genetic diversity. Sci Rep 9:3674. https://doi.org/10.1038/s41598-019-40213-5

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480-D489. https://doi.org/10.1093/nar/gkaa1100

Toronen P, Medlar A, Holm L (2018) PANNZER2: a rapid functional annotation web server. Nucleic Acids Res 46:W84-W88. https://doi.org/10.1093/nar/gky350

van Eerde A, Varnai A, Jameson JK, Paruch L, Moen A, Anonsen JH, Chylenski P, Steen HS, Heldal I, Bock R, Eijsink VGH, Liu-Clarke J (2020) In-depth characterization of *Trichoderma reesei* cellobiohydrolase TrCel7A produced in *Nicotiana benthamiana* reveals limitations of cellulase production in plants by host-specific post-translational modifications. Plant Biotechnol J 18:631-643. https://doi.org/10.1111/pbi.13227

Varga T, Krizsan K, Foldi C et al (2019) Megaphylogeny resolves global patterns of mushroom evolution. Nat Ecol Evol 3:668-678. https://doi.org/10.1038/s41559-019-0834-1

Wei H, Wu M, Fan A, Su H (2021) Recombinant protein production in the filamentous fungus *Trichoderma*. Chin J Chem Eng 30:74-81. https://doi.org/10.1016/j.cjche.2020.11.006

Wenger AM, Peluso P, Rowell WJ et al (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155-1162. https://doi.org/10.1038/s41587-019-0217-9

Wu L, McCluskey K, Desmeth P et al (2018) The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. Gigascience 7:giy026. https://doi.org/10.1093/gigascience/giy026

Xie BB, Qin QL, Shi M et al (2014) Comparative genomics provide insights into evolution of trichoderma nutrition style. Genome Biol Evol 6:379-390. https://doi.org/10.1093/gbe/evu018

Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46:W95-W101. https://doi.org/10.1093/nar/gky418

Zhang Y, Yang J, Luo L, Wang E, Wang R, Liu L, Liu J, Yuan H (2020) Low-cost cellulase-hemicellulase mixture secreted by Trichoderma harzianum EM0925 with complete saccharification efficacy of lignocellulose. Int J Mol Sci 21:371. https://doi.org/10.3390/ijms21020371

Zin NA, Badaluddin NA (2020) Biological functions of *Trichoderma* spp. for agriculture applications. Ann Agric Sci 65:168-178. https://doi.org/10.1016/j.aoas.2020.09.003

Tables

Table 1. Genome assembly and annotation statistics.

	Th3844	**Th0179**	**Ta0020**	**Tr0711**
Number of reads	418,031	504,913	458,142	282,252
Genome size (bp)	40,219,724	39,170,259	36,411,897	32,448,670
GC content (%)	47.5	49.4	49.5	53.5
N50 reads	3,607,994	2,983,622	3,146,023	1,694,659
L50 contigs	5	6	5	7
Number of contigs	15	18	14	26
Complete BUSCOs (%)	90.1	98.7	99	99.1
Complete and single-copy BUSCOs (%)	89.6	98.1	98.8	98.9
Complete and duplicated BUSCOs (%)	0.5	0.6	0.2	0.2
Fragmented BUSCOs (%)	0.2	0.2	0.1	0.1
Missing (%)	9.7	1.1	0.9	0.8
Number of predicted genes	10,786	11,322	10,082	8,796
Number of annotated genes	10,611	11,065	9,547	8,495

Table 2. Comparison of the genome features of *Trichoderma* spp. genomes.

Species	Strain	Coverage	NCBI accession	Genome size (Mb)	GC content (%)	N50 reads	L50 contigs	Assembly level	Genes	Reference
T. harzianum	IOC-3844	164 x	SAMN23309297	40	47.5	3,607,994	5	Contigs (15)	10,786	This study
T. harzianum	CBMAI-0179	219 x	SAMN23309298	39	49.4	2,983,622	6	Contigs (18)	11,322	This study
T. harzianum	T6776	85 x	SAMN02851310	39	48.5	68,846	178	Scaffolds (1,572)	11,501	(Baroncelli et al. 2015)
T. harzianum	CBS 226.95	120 x	SAMN00761861	41	47.6	2,414,909	7	Scaffolds (532)	14,269	-
T. virens	Gv29–8	8 x	SAMN02744059	39	49.2	1,836,662	8	Scaffolds (93)	12,405	(Kubicek et al. 2011)
T. simmonsii	GH-Sj1	-	SAMN15516371	40	48.13	6,451,197	3	Scaffolds (7)	13,296	(Chung et al. 2021)
T. atroviride	CBMAI-0020	229 x	SAMN23309299	36	49.5	3,146,023	5	Contigs (14)	10,082	This study
T. atroviride	IMI206040	8 x	SAMN02744066	36	49.7	2,007,903	6	Contigs (29)	11,809	(Kubicek et al. 2011)
T. reesei	CBMAI-0711	143 x	SAMN23309300	32	53.5	1,694,659	7	Contigs (26)	8,796	This study
T. reesei	QM6a	80 x	SAMN05250858	35	51.0	18,236	3	Chromosomes (7)	10,877	(Li et al. 2017)
T. reesei	QM6a	9 x	SAMN02746107	33	52.8	1,219,543	9	Scaffolds (77)	9,109	(Martinez et al. 2008)

Table 3. Potential biosynthetic gene clusters in Th3844, Th0179, Ta0020, and Tr0711 showing 100% similarity with the known clusters predicted by antiSMASH.

Contig	Region	Type	From	To	Most similar known cluster
			Th3844		
tig00000165	Region 7.1	T1PKS	432,98	473,452	naphthopyrone
tig00000165	Region 7.2	NRPS,T1PKS	723,511	822,263	alternariol
tig00000165	Region 7.6	NRPS	2,686,328	2,731,748	dimethylcoprogen
tig00000267	Region 14.1	T1PKS	276,364	334,365	tricholignan A
tig00000268	Region 15.2	T1PKS	1,451,269	1,497,959	naphthopyrone
			Th0179		
tig00000001	Region 1.4	NRPS	454,959	510,636	AbT1
tig00000076	Region 4.2	T1PKS	2,694,282	2,742,501	clavaric acid
tig00000184	Region 7.1	T1PKS	245,46	285,906	naphthopyrone
tig00000184	Region 7.2	NRPS,T1PKS	543,777	635,971	alternariol
tig00000184	Region 7.6	NRPS	2,223,632	2,269,064	dimethylcoprogen
tig00000216	Region 9.1	T1PKS	158,954	216,896	tricholignan A
			Ta0020		
tig00000043	Region 3.4	NRPS	2,250,033	2,299,455	enniatin
tig00000127	Region 7.2	terpene	2,081,716	2,104,094	clavaric acid
tig00000139	Region 8.1	NRPS	90,107	131,179	dimethylcoprogen
tig00000163	Region 10.1	T1PKS	54,348	96,76	naphthopyrone
			Tr0711		
tig00000040	Region 2.1	T1PKS	134,208	178,258	clavaric acid
tig00000131	Region 6.1	NRPS	1,576,642	1,622,080	dimethylcoprogen
tig00000197	Region 9.1	T1PKS	313,591	354,94	alternariol

Table 4. Distribution of Th3844, Th0179, Ta0020, and Tr0711 orthologs.

	Th3844	**Th0179**	**Ta0020**	**Tr0711**
Th3844	-	9,729	7,927	7,330
Th0179	9,729	-	8,755	8,141
Ta0020	7,927	8,755	-	7,784
Tr0711	7,330	8,141	7,784	-

Figure captions

Fig. 1 *Trichoderma* isolates evaluated in this study. (a) Th3844, (b) Th0179, (c) Tr0711, and (d) Ta0020 cultivated on potato dextrose broth (PDA) solid medium at 28 °C. The four strains showed a different growth pattern, i.e., Th3844 (a) presented a production of spores similar to Th0179 (b) and Tr0711 (c), and Ta0020 (d) showed slow sporulation. Ta0020: *T. atroviride* CBMAI-0020; Th0179: *T. harzianum* CBMAI-0179; Th3844: *T. harzianum* IOC-3844; Tr0711: *T. reesei* CBMAI-0711

Fig. 2 Comparisons between the genomes of the analyzed *Trichoderma* isolates and the *T. reesei* QM6a reference genome. Dot plots of the assemblies of (a) Th3844, (b) Th0179, (c) Ta0020, and (d) Tr0711 that were generated by Canu (y-axis) against those from *T. reesei* QM6a (x-axis) that are available in the NCBI database. Ta0020: *T. atroviride* CBMAI-0020; Th0179: *T. harzianum* CBMAI-0179; Th3844: *T. harzianum* IOC-3844; Tr0711: *T. reesei* CBMAI-0711

Fig. 3 COG functional category distribution of the *Trichoderma* isolates considered. The plot shows the number of genes in the genomes of (a) Th3844, (b) Th0179, (c) Ta0020, and (d) Tr0711, in which a COG classification was obtained. The size of the boxes represents the abundance of the genes at the level of individual COG families. Only the COG functional categories with more than a hundred counts were represented. Ta0020: *T. atroviride* CBMAI-0020; Th0179: *T. harzianum* CBMAI-0179; Th3844: *T. harzianum* IOC-3844; Tr0711: *T. reesei* CBMAI-0711

Fig. 4 Quantitative comparison of the enzyme classes involved in SMs synthesis. Different classes of enzymes related to the production of SMs were predicted from the Th3844, Th0179, Ta0020, and Tr0711 genomes. Among all the strains, Th0179 exhibited the highest number of BGCs, followed by Th3844, Ta0020, and Tr0711. Although the T1PKS class was the most represented among all evaluated strains, the NRPS class was more highly represented in Th0179 than in the other strains. Fungal-RiPP: fungal RiPP with POP or UstH peptidase types and a modification; NRPS: nonribosomal peptide synthetase; T1PKS: type I polyketide synthase; Ta0020: *T. atroviride* CBMAI-0020; Th0179: *T. harzianum* CBMAI-0179; Th3844: *T. harzianum* IOC-3844; Tr0711: *T. reesei* CBMAI-0711

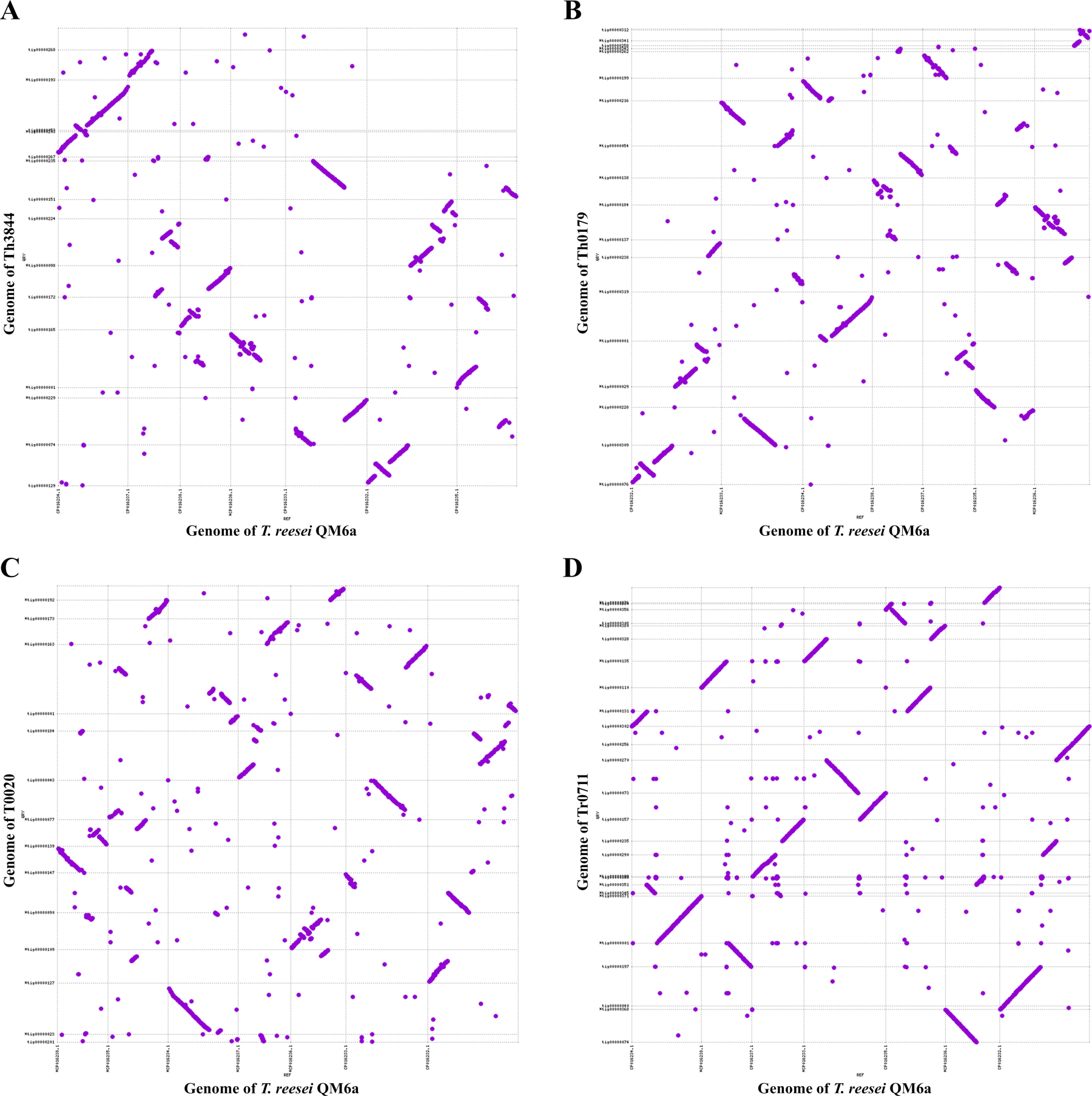
Fig. 5 Distribution of CAZymes in *Trichoderma* spp. and evaluation of their expression in Th3844, Th0179, and Ta0020 cells by RNA-Seq. (a) The predicted CAZymes from the assembled genomes were classified according to the CAZy database. (b) The secreted CAZymes were grouped according to their CAZyme class. (c) Heatmap plotted based on the CAZymes of Th3844, Th0179, and Ta0020 showing upregulated and downregulated expression

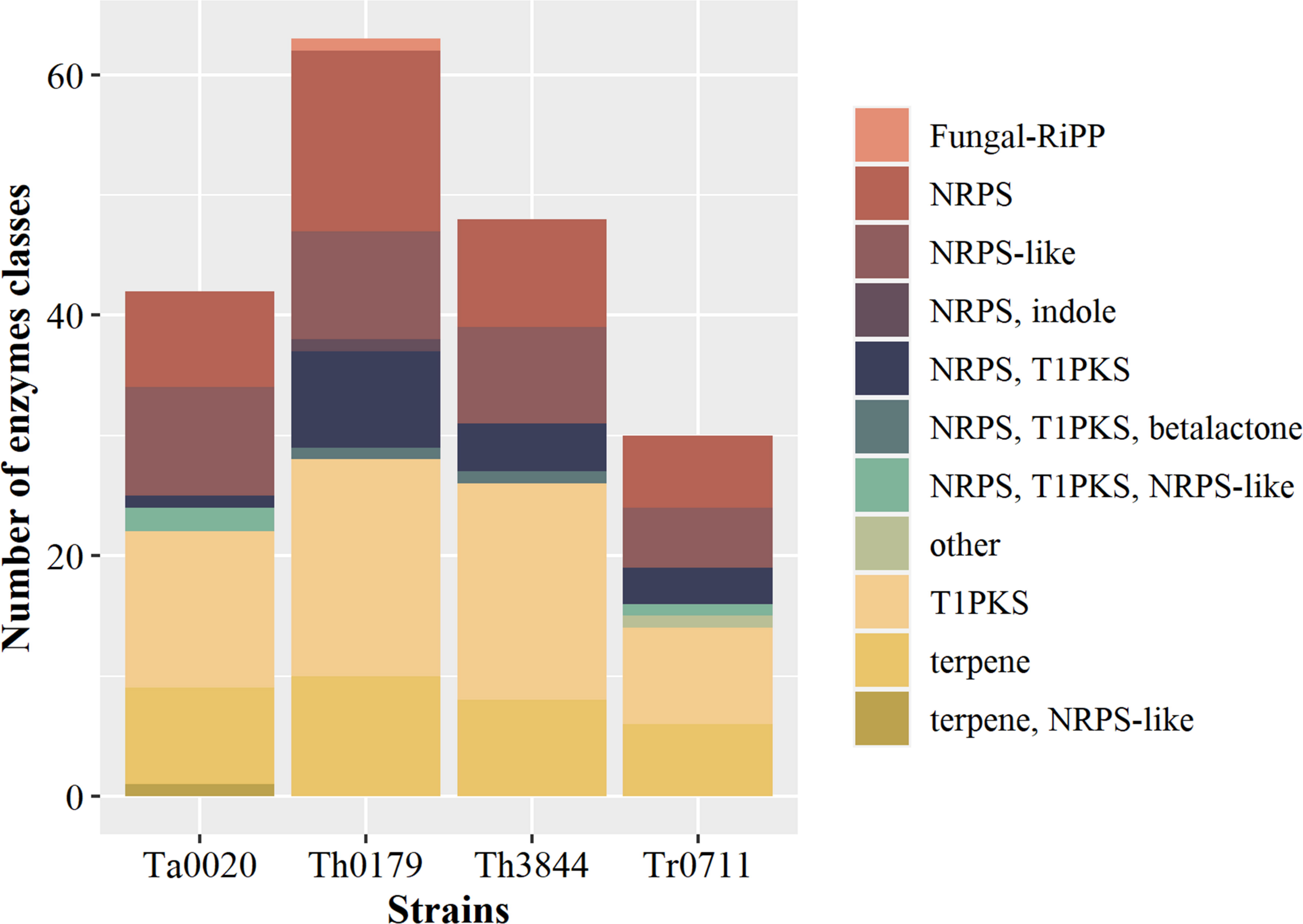
under cellulose and glucose growth conditions, as indicated by increase or decrease in the gray color in the subtitle. The expression of the CAZyme classes of enzymes with orthologs identified between Th3844, Th0179, and Ta0020 was quantified. CAZymes: carbohydrate-active enzymes; Ta0020: *T. atroviride* CBMAI-0020; Th0179: *T. harzianum* CBMAI-0179; Th3844: *T. harzianum* IOC-3844; Tr0711: *T. reesei* CBMAI-0711. AA: auxiliary activity; CBM: carbohydrate-binding module; EC: carbohydrate esterase; GH: glycosyl hydrolase; GT: glycosyl transferase; PL: polysaccharide lyase

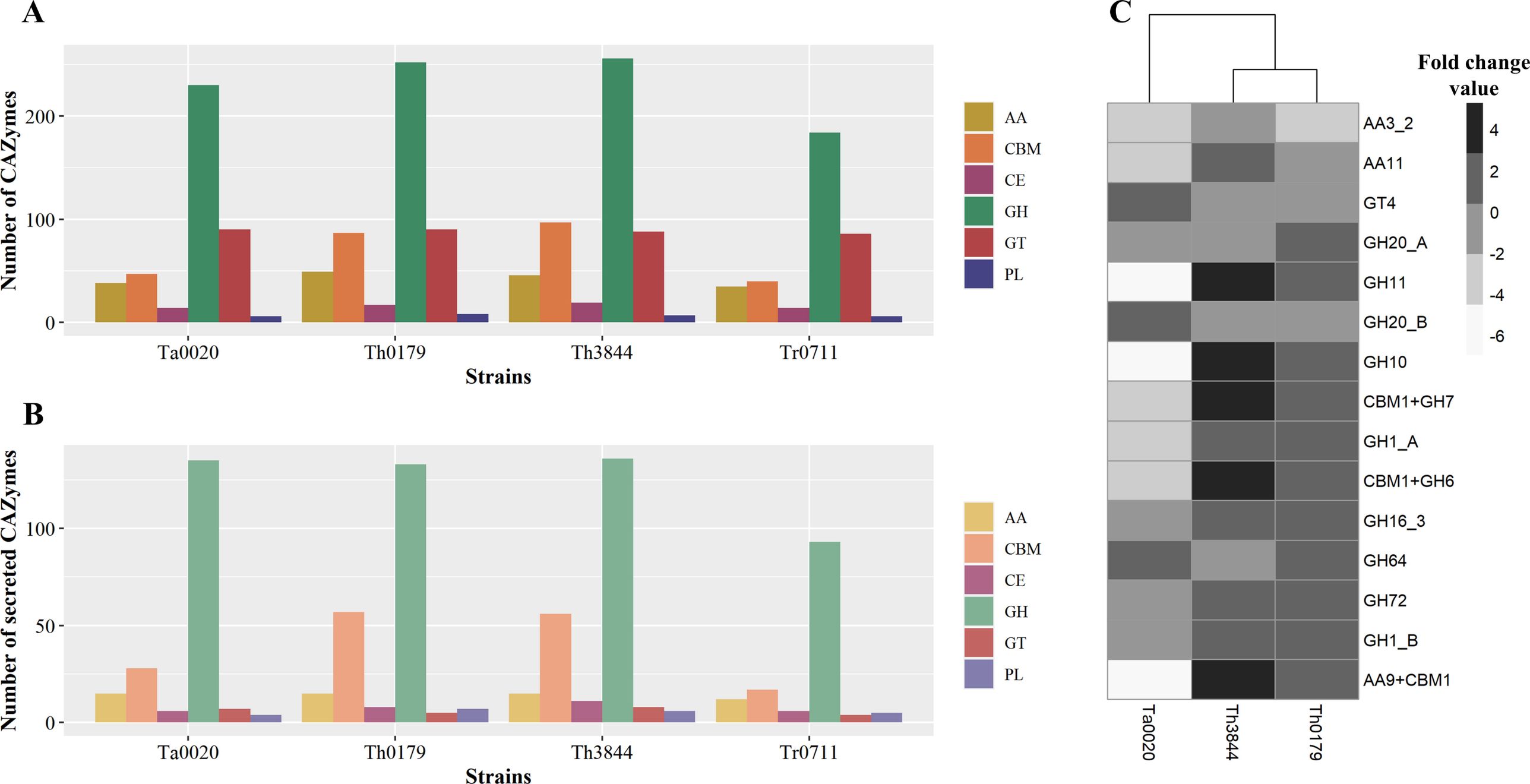
Fig. 6 Quantitative comparison of the biomass-degrading and mycoparasitic enzymatic repertoires predicted for the *Trichoderma* isolate genomes. (a) Heatmap of the number of enzymes in each CAZY family from the Th3844, Th0179, Ta0020, and Tr0711 genomes. This map includes only the enzymes/proteins related to biomass degradation. (b) Heatmap of the number of enzymes in each CAZy family from the Th3844, Th0179, Ta0020, and Tr0711 genomes. This map includes only the enzymes related to mycoparasitic activity. Ta0020: *T. atroviride* CBMAI-0020; Th0179: *T. harzianum* CBMAI-0179; Th3844: *T. harzianum* IOC-3844; Tr0711: *T. reesei* CBMAI-0711; LPMOs: lytic polysaccharide monooxygenases; CDHs: cellobiose dehydrogenases; PHDs: pyranose dehydrogenases

Fig. 7 Phylogenetic relationships of *Trichoderma* spp. as inferred by an orthology analysis. The phylogenetic tree modeled by OrthoFinder software was based on the concatenation of 2,229 single-copy orthogroups. In addition to the proteomes of Th3844, Th0179, Ta0020, and Tr0711, this methodology shows the inferred relationships among 19 *Trichoderma* spp., for which the proteomes are available in the NCBI database. *Fusarium* spp., *Aspergillus* spp., and *Neurospora* spp. were used as the outgroup. Bootstrap values are shown at the nodes

Fig. 8 Structural variant heterogeneity across the genomes of the evaluated *Trichoderma* spp. (a) Long-read alignment-based structural variant (SV) analyses among the evaluated *Trichoderma* isolates and *T. reesei* QM6a showed breakends (BNDs), deletions (DELs), multiple nucleotides and InDels (MIXEDs), duplications (DUPs), insertions (INSs), and inversions (INVs) among the genomes. (B) Functional effects of the identified SVs. Ta0020: *T. atroviride* CBMAI-0020; Th0179: *T. harzianum* CBMAI-0179; Th3844: *T. harzianum* IOC-3844; Tr0711: *T. reesei* CBMAI-0711


A
Genome of Th3844
Genome of *T. reesei* QM6a
B
Genome of Th0179
Genome of *T. reesei* QM6a
C
Genome of T0020
Genome of *T. reesei* QM6a
D
Genome of Tr0711
Genome of *T. reesei* QM6a


Number of enzymes classes
60
40
20
0
Ta0020
Th0179
Th3844
Tr0711
Strains
Fungal-RiPP
NRPS
NRPS-like
NRPS, indole
NRPS, T1PKS
NRPS, T1PKS, betalactone
NRPS, T1PKS, NRPS-like
other
T1PKS
terpene
terpene, NRPS-like

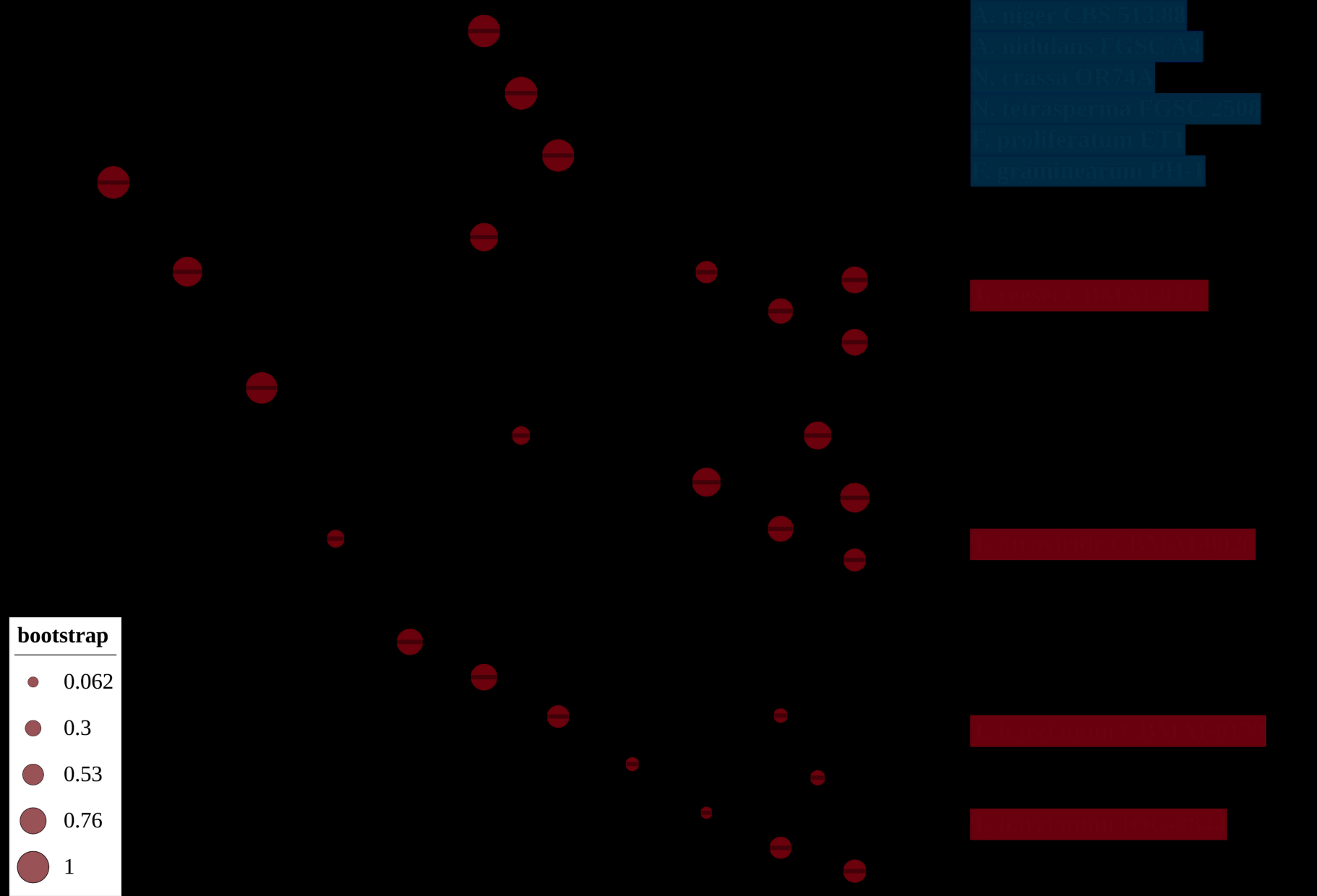

A
Number of CAZymes
200
100
0
Ta0020
Th0179
Th3844
Tr0711
Strains
AA
CBM
CE
GH
GT
PL
B
Number of secreted CAZymes
100
50
0
Ta0020
Th0179
Th3844
Tr0711
Strains
AA
CBM
CE
GH
GT
PL
C
Fold change value
4
2
0
-2
-4
-6
AA3_2
AA11
GT4
GH20_A
GH11
GH20_B
GH10
CBM1+GH7
GH1_A
CBM1+GH6
GH16_3
GH64
GH72
GH1_B
AA9+CBM1
Ta0020
Th3844
Th0179

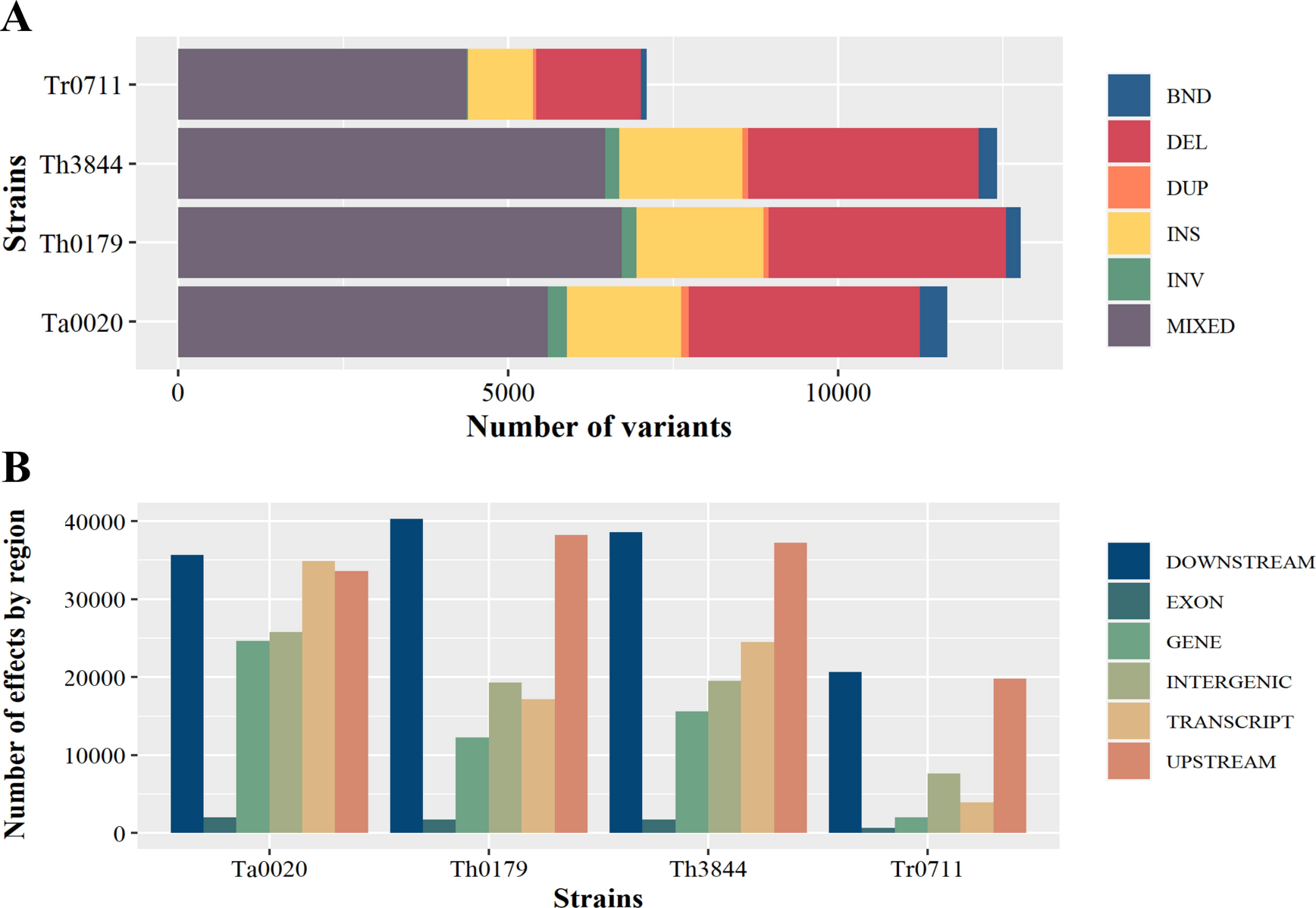
A

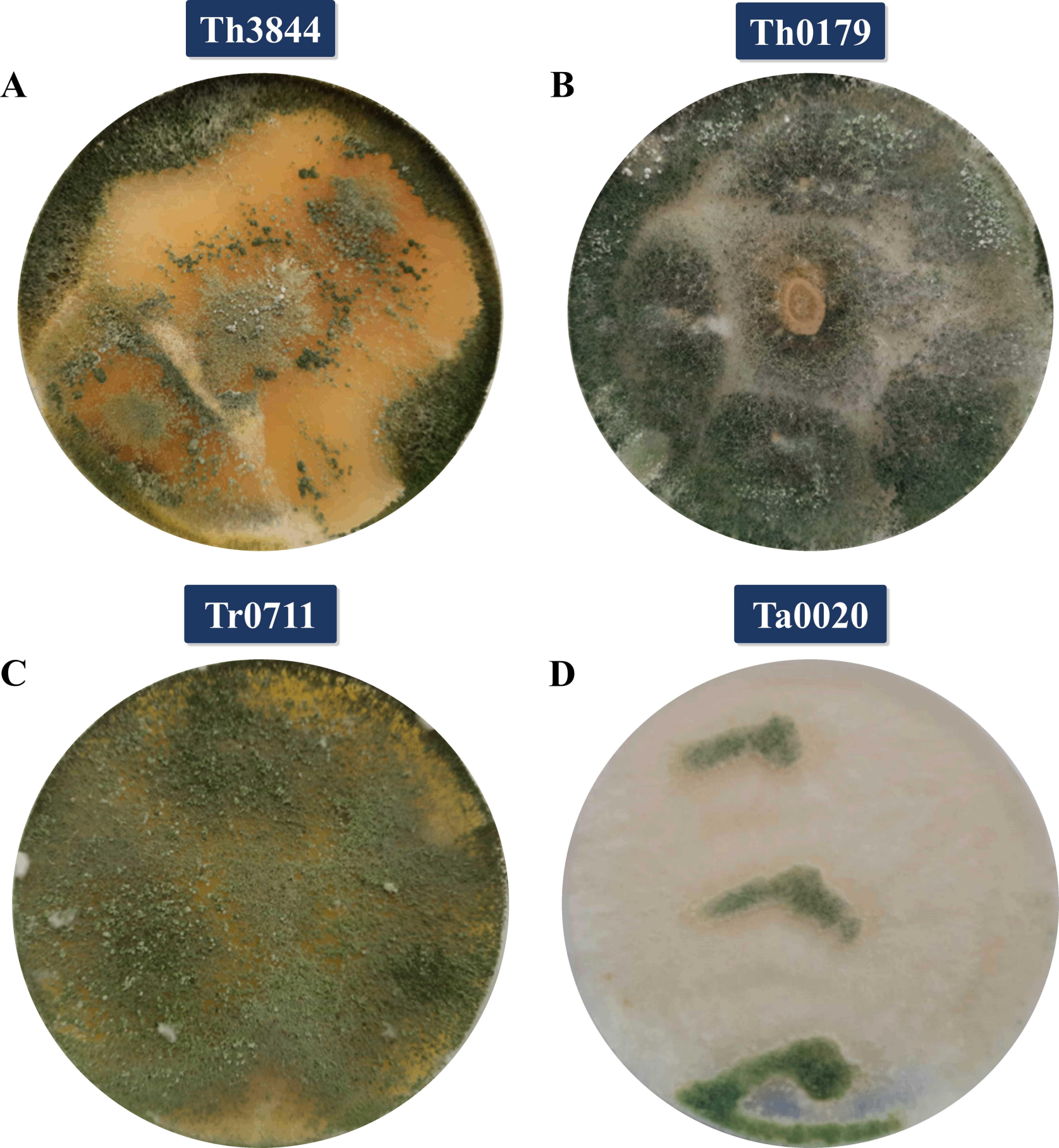
Substrate/enzyme category	Enzyme class	CAZy family	Th3844	Th0179	Ta0020	Tr0711
Arabinan	Exo-arabinanases/Arabinofuranosidases	GH51	0	0	1	0
	Arabinofuranosidases	GH54	2	2	2	2
	Endoarabinanases	GH43	5	4	4	2
Cellulose	Cellobiohydrolases	GH6	1	1	1	1
	Oligosaccharide oxidases	AA7	2	3	1	1
	LPMOs	AA9	3	3	3	3
	Endoglucanases	GH5	10	10	9	7
Esterases	Feruloyl/p-coumaroyl/acetyl esterases	CE1	1	0	0	0
	4-O-methyl-glucoronoyl esterases	CE15	1	1	1	1
	Acetyl esterases	CE5	4	4	3	4
Lignin	Glyoxal oxidases (GLOX)	AA5	1	1	1	1
	1,4-benzoquinone reductases	AA6	1	1	1	1
	Vanillyl-alcohol oxidases	AA4	2	3	0	0
	Peroxidases	AA2	1	3	5	3
	Laccases	AA1	9	9	6	4
Lignin/Cellulose	Oxidoreductases/CDHs	AA3	19	20	12	12
Monosugars	PDHs	AA12	1	1	1	0
Pectin	Pectin lyases	PL1	0	0	1	0
	Pectin methylesterases	CE8	1	1	1	0
	Rhamnosidases	GH78	2	2	2	1
	Pectin lyases	PL20	2	2	2	2
	Polygalacturonases	GH28	5	5	5	4
Soluble oligosaccharides	β-glucosidases	GH1	4	4	4	2
	Mannosidases	GH2	11	13	10	7
Xylan	Xylanases	GH10	2	2	1	1
	LPMOs	AA14	2	2	2	2
	Xylanases	GH11	4	4	4	3
Xylosidases/β-glucosidases	Xylosidases/β-glucosidases	GH3	16	17	15	12

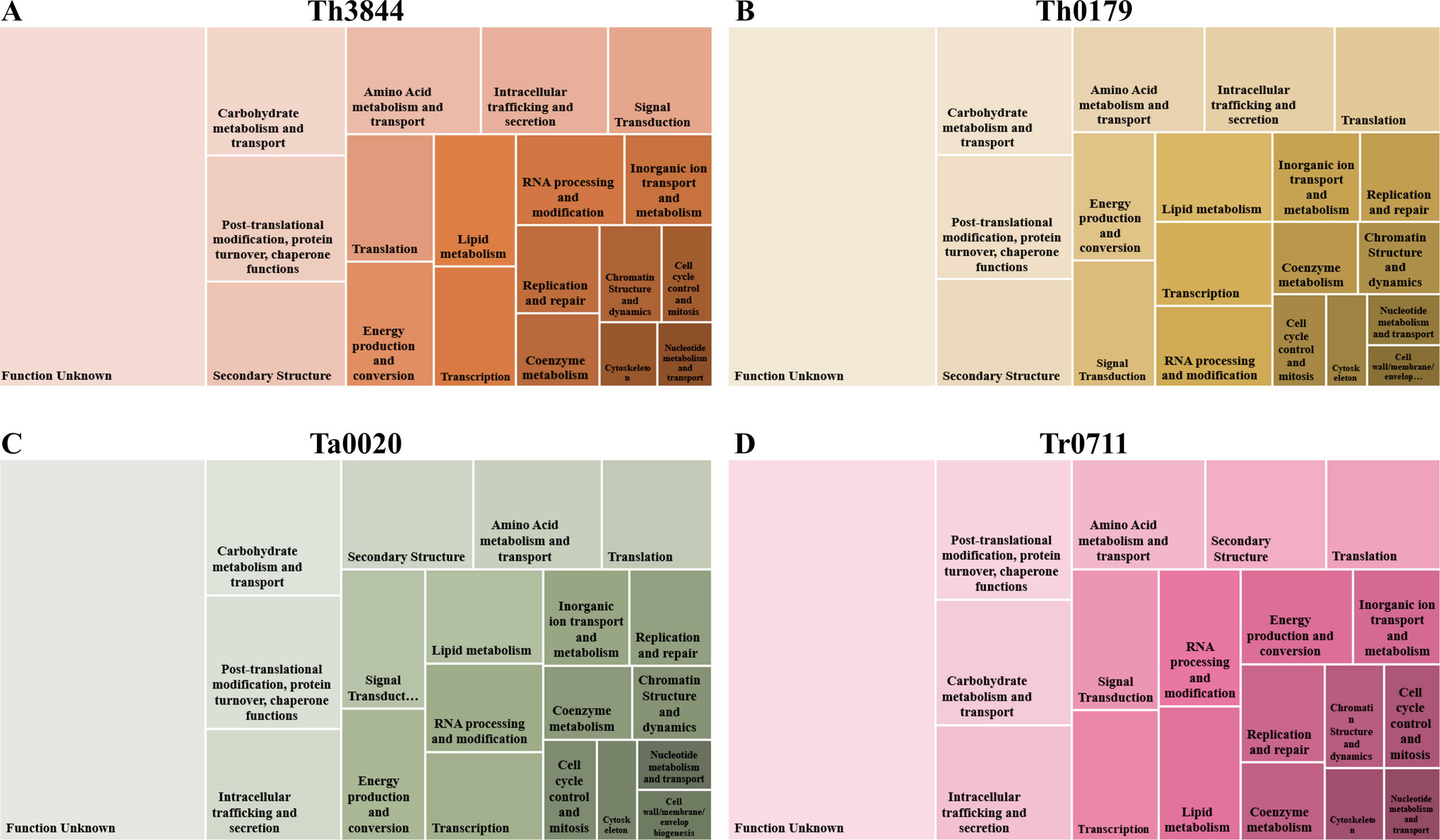
B

Enzyme class	CAZy family	Th3844	Th0179	Ta0020	Tr0711
chitinases	GH18	29	26	25	18
chitinases	GH20	3	3	3	3
chitosanases	GH75	5	5	6	3
α-1,3-glucanases	GH71	8	6	4	4
β-1,3-glucanases	GH16	13	13	13	13
β-1,3-glucanases	GH17	1	1	2	1
β-1,3-glucanases	GH55	11	10	8	6
β-1,3-glucanases	GH64	8	3	3	3
β-1,3-glucanases	GH81	2	2	2	2


A. niger CBS 513.88
A. nidulans FGSC A4
N. crassa OR74A
N. tetraspermia FGSC 2508
F. proliferatum ET1
F. graminearum PH-1
bootstrap
0.062
0.3
0.53
0.76
1


A
Strains
Tr0711
Th3844
Th0179
Ta0020
0
5000
10000
Number of variants
BND
DEL
DUP
INS
INV
MIXED
B
Number of effects by region
40000
30000
20000
10000
0
Ta0020
Th0179
Th3844
Tr0711
Strains
DOWNSTREAM
EXON
GENE
INTERGENIC
TRANSCRIPT
UPSTREAM


Th3844
A
Th0179
B
Tr0711
C
Ta0020
D


A
Th3844
Function Unknown
Carbohydrate metabolism and transport
Post-translational modification, protein turnover, chaperone functions
Secondary Structure
Amino Acid metabolism and transport
Intracellular trafficking and secretion
Signal Transduction
Translation
Lipid metabolism
Energy production and conversion
Transcription
RNA processing and modification
Inorganic ion transport and metabolism
Replication and repair
Chromatin Structure and dynamics
Cell cycle control and mitosis
Coenzyme metabolism
Cytoskeleto n
Nucleotide metabolism and transport
B
Th0179
Function Unknown
Carbohydrate metabolism and transport
Post-translational modification, protein turnover, chaperone functions
Secondary Structure
Amino Acid metabolism and transport
Intracellular trafficking and secretion
Translation
Energy production and conversion
Signal Transduction
Lipid metabolism
Transcription
RNA processing and modification
Inorganic ion transport and metabolism
Replication and repair
Chromatin Structure and dynamics
Coenzyme metabolism
Cell cycle control and mitosis
Cytosk eleton
Nucleotide metabolism and transport
Cell wall/membrane/ envelop…
C
Ta0020
Function Unknown
Carbohydrate metabolism and transport
Post-translational modification, protein turnover, chaperone functions
Intracellular trafficking and secretion
Secondary Structure
Amino Acid metabolism and transport
Translation
Signal Transduct…
Energy production and conversion
Lipid metabolism
RNA processing and modification
Transcription
Inorganic ion transport and metabolism
Replication and repair
Coenzyme metabolism
Chromatin Structure and dynamics
Cell cycle control and mitosis
Cytosk eleton
Nucleotide metabolism and transport
Cell wall/membrane/ envelop biogenesis
D
Tr0711
Function Unknown
Post-translational modification, protein turnover, chaperone functions
Carbohydrate metabolism and transport
Intracellular trafficking and secretion
Amino Acid metabolism and transport
Secondary Structure
Translation
Signal Transduction
Transcription
RNA processing and modification
Lipid metabolism
Energy production and conversion
Inorganic ion transport and metabolism
Replication and repair
Coenzyme metabolism
Chromati n Structure and dynamics
Cell cycle control and mitosis
Cytoskeleto n
Nucleotide metabolism and transport