GENOME SEQUENCING AND COMPARATIVE ANALYSIS OF POTENTIALLY BIOTECHNOLOGICAL STRAINS FROM *TRICHODERMA*

Rafaela R. Rosolen[1,2], Maria Augusta C. Horta[3], Paulo H.C. de Azevedo[1,2], Carla C. da Silva[1], Gustavo H. Goldman[3], Danilo A. Sforca[1] & Anete P. de Souza[1,4*]

[1] *Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil.*
[2] *Graduate Program in Genetics and Molecular Biology, Institute of Biology, UNICAMP, Campinas, SP, Brazil.*
[3] *Faculty of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, SP, Brazil.*
[4] *Department of Plant Biology, Institute of Biology, UNICAMP, Campinas, SP, Brazil.*
[*] *Corresponding author's email address: anete@unicamp.br*

ABSTRACT

Trichoderma harzianum is widely used as a commercial biocontrol agent 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 the Th3844 and Th0179 strains. To assess the genetic diversity within the genus *Trichoderma*, the results of both strains were compared with strains of *Trichoderma atroviride* CBMAI-00020 (Ta0020) and *Trichoderma 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: Genomic analysis. Orthology analysis. Structural variants analysis. *Trichoderma*. Fungal biotechnology.

1 INTRODUCTION

Several *Trichoderma* species, including *Trichoderma atroviride*, *Trichoderma virens*, and *Trichoderma harzianum*, produce and secrete chemically diverse secondary metabolites (SMs) with bioactivity against various antagonistic microorganisms, including phytopathogens; therefore, they are used as biological control agents[1]. More 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[2-4]. Previous studies have suggested the great potential of *T. harzianum* IOC-3844 (Th3844) and *T. harzianum* CBMAI-0179 (Th0179) strains as hydrolytic enzyme producers when compared to *T. atroviride* CBMAI-0020 (Ta0020) and *Trichoderma reesei* CBMAI-0711 (Tr0711)[3, 5]. Although Th3844 and Th0179 demonstrated a high biotechnology potential, genomic information regarding both strains remains unclear. In this study, Pacific Biosciences (PacBio)[6] technology was used to obtain highly contiguous de novo assemblies of Th3844 and Th0179 and identify the genetic variation between them. To expand knowledge on the genetic diversity within the genus *Trichoderma*, the results obtained for *T. harzianum* strains were compared to those from Ta0020 and Tr0711.

After performing whole-genome annotation, we investigated the contents of carbohydrate-active enzymes (CAZymes)[7], as well as, based on previous transcriptomes[3, 5], their gene expression levels under cellulose and glucose growth conditions. We also inspected the 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 variation that is inherited and polymorphic in species[8], between them and *T. reesei* QM6a, the reference genome[9]. 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 could be applied to deeply investigate the evolution and basic biology of the evaluated strains, as well as the biotechnology potential that they could offer to the scientific community and industry. For instance, from an evolutionary perception, through the genomes, we could highlight the genetic differences between the evaluated strains; which could explain the phenotype differences between them. From a biotechnology perspective, the genomic information provided by this study is a valuable resource for the bioprospecting of new enzymes.

2 MATERIAL & METHODS

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 culture conditions and DNA extraction and sequencing protocols used for the evaluated strains were described in a previously

published article[10]. Briefly, the genomes were assembled de novo using Canu software (v.2.1). Gene prediction was performed using AUGUSTUS (v.3.3.3) 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 (v.2.31.11). The predicted genes were functionally annotated by searching for homologous sequences in the UniProt, eggNOG-mapper v.2, and Protein Annotation with Z score (PANNZER2) databases. For the annotation of CAZymes, we used CDSs as homology search queries against the database of the dbCAN2 server. Coverages were estimated with QualiMap (v.2.2.2c) using minimap2 v. 2.17 + galaxy4, which were both implemented on the Galaxy platform. SMGCs in the Th3844, Th0179, Ta0020, and Tr0711 genomes were predicted using antiSMASH fungal version v.6.1.0. The complete methodology regarding genome assembly, gene prediction, and functional annotation is available in a previously published article[10].For the orthology analysis, we used the software OrthoFinder v2.5.2. The resulting tree from the OrthoFinder analysis was visualized and edited using Interactive Tree of Life (iTOL) v6. Next, SVs were identified by aligning the PacBio HiFi reads from Th3844, Th0179, Ta0020, and Tr0711 with the *T. reesei* QM6a reference genome[9] using the software Map with BWA-MEM v.0.7.17.2. Variants were called using Sniffles (v.1.0.12 + galaxy0) and annotated using SnpEff (v.4.3 + T. galaxy1). The gene expression profile of the CAZymes identified in the assembled genomes was investigated and the methodology used is described in a previously published article[10]. The references of the programs described all over the methodology of this study are cited in a previous article. Due to the reduced space of this abstract, we opted not to cite it here.

3 RESULTS & DISCUSSION

In the present study, we introduced the whole-genome sequences of Th3844, Th0179, Ta0020, and Tr0711. Overall, the genomes of the evaluated *Trichoderma* spp. varied in the 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. In other words, the ecological behavior of the mycoparasites *T. atroviride* and *T. harzianum*, compared to the plant wall degrader *T. reesei*, is reflected by the sizes of the respective genomes. Therefore, due to the great genetic variability of the evaluated strains, we investigated the presence of SMGCs and CAZymes in all their genome, as explored next.

Many *Trichoderma* species are the most prominent producers of SMs with antimicrobial activity against phytopathogenic fungi[1]. 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[11]. 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.

Regarding the CAZyme content, the results found here for the Th3844, Th0179, Ta0020, and Tr0711 genomes (**Figure 1**) follow the same profile as that of a previous study[12], 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 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 their lignocellulose-degrading enzyme potential but cannot be directly related to their 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[13]. For this reason, CAZy analysis is associated with functional approaches, such as enzymatic activity assays, which provide valuable insight into the actual behaviour of the concerned species on specific lignocellulose substrates.

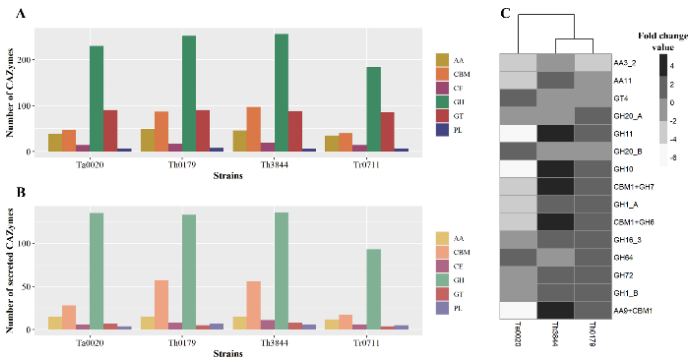
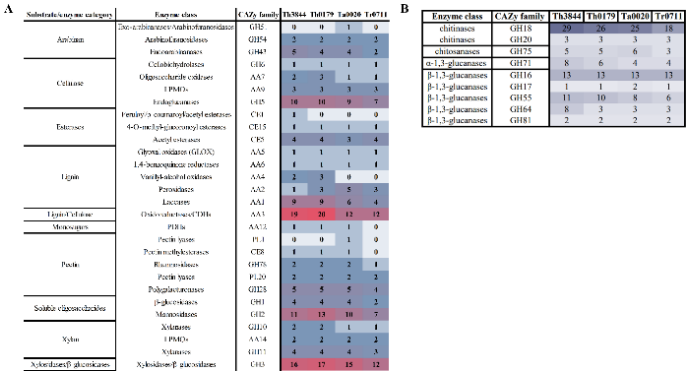


Figure 1 Distribution of CAZymes in *Trichoderma* spp. and evaluation of their expression in Th3844, Th0179, and Ta0020 cells by RNA-Seq.

Considering the differentially expressed genes (DEGs) (**Figure 1**), 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. Moreover, the CAZyme families related to biomass degradation as well as mycoparasitism were investigated in greater depth (**Figure 2**). 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 (**Figure 2**). 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.



Substrate/enzyme category	Enzyme class	CAZy family	Th3844	Th0179	Ta0020	Tr0711
Arbutan	Dextranobiases/Arabinofuranosidases	GH2	9	9	7	9
	Arabinofuranosidases	GH54	2	2	2	2
	Fucosylxylases	GH95	5	4	4	2
Cellulose	Cellobiohydrolases	GH6	1	1	1	1
	Oligosaccharide cellulases	AA7	2	3	1	1
	LPMOs	AA9	3	3	3	3
	Endoglucanases	GH5	10	10	9	7
Esterases	Feruloyl/p-coumaroylacetyl-esterases	CE1	1	0	0	0
	4-O-me(3)-β-glucuronoyl esterases	CE15	1	1	1	1
	Acetylesterases	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 of CDHs	AA3	19	20	12	12
Mannoscapes	PDHs	AA8	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
Xyloglucans/β glucosidases	Xyloglucanases/β glucosidases	GH5	16	17	15	12

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

Figure 2 Comparison of the biomass-degrading and mycoparasitic enzymatic repertoires predicted for the *Trichoderma* isolate genomes.

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 compared with the other strains. In relation to the other evaluated strains, Th3844 and Th0179 exhibited more orthologs in common with Ta0020 than with Tr0711. 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. In relation to the SVs analysis, a total of 12,407 (Th3844), 12,763 (Th0179), 11,650 (Ta0020), and 7,103 (Tr0711) SVs were identified for each strain. Although the *T. harzianum* strains are phylogenetically close[14], comparison of the SVs identified from the mapping of both genomes against *T. reesei* QM6a reveals genetic variability across the strains[9]. These SVs included different phenomena that affect gene sequences, such as break ends, deletions, multiple nucleotides and InDels, duplications, insertions, and inversions. For all evaluated strains, the most frequently presented SV categories were multiple nucleotides and an InDel, followed by deletions and insertions.

4 CONCLUSION

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.

REFERENCES

[1]ALMEIDA, D. A., M. A. C. HORTA, J. A. FERREIRA FILHO, N. F. MURAD & A. P. DE SOUZA (2021) The synergistic actions of hydrolytic genes reveal the mechanism of *Trichoderma harzianum* for cellulose degradation. *Journal of Biotechnology,* 334**,** 1-10.

[2]ARDUI, S., A. AMEUR, J. R. VERMEESCH & M. S. HESTAND (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research,* 46**,** 2159-2168.

[3]ARNTZEN, M. Ø., O. BENGTSSON, A. VÁRNAI, F. DELOGU, G. MATHIESEN & V. G. EIJSINK (2020) Quantitative comparison of the biomass-degrading enzyme repertoires of five filamentous fungi. *Scientific Reports,* 10**,** 20267.

[4]CANTAREL, B. L., P. M. COUTINHO, C. RANCUREL, T. BERNARD, V. LOMBARD & B. HENRISSAT (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research,* 37**,** D233-D238.

[5]CHEN, M., Q. LIU, S.-S. GAO, A. E. YOUNG, S. E. JACOBSEN & Y. TANG (2019) Genome mining and biosynthesis of a polyketide from a biofertilizer fungus that can facilitate reductive iron assimilation in plant. *Proceedings of the National Academy of Sciences,* 116**,** 5499-5504.

[6]DELABONA, P. D. S., C. A. CODIMA, J. RAMONI, M. P. ZUBIETA, B. M. DE ARAÚJO, C. S. FARINAS, J. G. D. C. PRADELLA & B. SEIBOTH (2020) The impact of putative methyltransferase overexpression on the *Trichoderma harzianum* cellulolytic system for biomass conversion. *Bioresource Technology,* 313**,** 123616.

[7]FANELLI, F., V. C. LIUZZI, A. F. LOGRIECO & C. ALTOMARE (2018) Genomic characterization of *Trichoderma atrobrunneum* (*T. harzianum* species complex) ITEM 908: insight into the genetic endowment of a multi-target biocontrol strain. *BMC genomics,* 19**,** 1-18.

[8]HORTA, M. A. C., J. A. F. FILHO, N. F. MURAD, E. DE OLIVEIRA SANTOS, C. A. DOS SANTOS, J. S. MENDES, M. M. BRANDÃO, S. F. AZZONI & A. P. DE SOUZA (2018) Network of proteins, enzymes and genes linked to biomass degradation shared by *Trichoderma* species. *Scientific Reports,* 8**,** 1341.

[9]LI, J.-X., F. ZHANG, D.-D. JIANG, J. LI, F.-L. WANG, Z. ZHANG, W. WANG & X.-Q. ZHAO (2020) Diversity of Cellulase-Producing Filamentous Fungi From Tibet and Transcriptomic Analysis of a Superior Cellulase Producer *Trichoderma harzianum* LZ117. *Frontiers in Microbiology,* 11.

[10]MARTINEZ, D., R. M. BERKA, B. HENRISSAT, M. SALOHEIMO, M. ARVAS, S. E. BAKER, J. CHAPMAN, O. CHERTKOV, P. M. COUTINHO & D. CULLEN (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. Hypocrea jecorina). *Nature biotechnology,* 26**,** 553-560.

[11]MILLS, R. E., K. WALTER, C. STEWART, R. E. HANDSAKER, K. CHEN, C. ALKAN, A. ABYZOV, S. C. YOON, K. YE, R. K. CHEETHAM, A. CHINWALLA, D. F. CONRAD, Y. FU, F. GRUBERT, I. HAJIRASOULIHA, F. HORMOZDIARI, L. M. IAKOUCHEVA, Z. IQBAL, S. KANG, J. M. KIDD, M. K. KONKEL, J. KORN, E. KHURANA, D. KURAL, H. Y. K. LAM, J. LENG, R. LI, Y. LI, C.-Y. LIN, R. LUO, X. J. MU, J. NEMESH, H. E. PECKHAM, T. RAUSCH, A. SCALLY, X. SHI, M. P. STROMBERG, A. M. STÜTZ, A. E. URBAN, J. A. WALKER, J. WU, Y. ZHANG, Z. D. ZHANG, M. A. BATZER, L. DING, G. T. MARTH, G. MCVEAN, J. SEBAT, M. SNYDER, J. WANG, K. YE, E. E. EICHLER, M. B. GERSTEIN, M. E. HURLES, C. LEE, S. A. MCCARROLL, J. O. KORBEL & P. GENOMES (2011) Mapping copy number variation by population-scale genome sequencing. *Nature,* 470**,** 59-65.

[12]ROSOLEN, R. R., A. H. AONO, D. A. ALMEIDA, J. A. FERREIRA FILHO, M. A. C. HORTA & A. P. DE SOUZA (2022) Network Analysis Reveals Different Cellulose Degradation Strategies Across *Trichoderma harzianum* Strains Associated With XYR1 and CRE1. *Frontiers in Genetics,* 13.

[13]ROSOLEN, R. R., M. A. C. HORTA, P. H. C. DE AZEVEDO, C. C. DA SILVA, D. A. SFORCA, G. H. GOLDMAN & A. P. DE SOUZA (2023) Whole-genome sequencing and comparative genomic analysis of potential biotechnological strains of *Trichoderma harzianum*, *Trichoderma atroviride*, and *Trichoderma reesei*. *Molecular Genetics and Genomics,* 298**,** 735-754.

[14]YAO, X., H. GUO, K. ZHANG, M. ZHAO, J. RUAN & J. CHEN (2023) *Trichoderma* and its role in biological control of plant fungal and nematode disease. *Frontiers in Microbiology,* 14.

ACKNOWLEDGEMENTS

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.