New Putative Long Non-Coding RNAs (lncRNA) Revealed by Pan-Transcriptome of the Emerging Human Pathogenic Fungus *Talaromyces Marneffei*

David Aciole Barbosa
University of Mogi das Cruzes: Universidade de Mogi das Cruzes
Alexandre Santos Simeone
University of Mogi das Cruzes: Universidade de Mogi das Cruzes
Ana Carolina Humberto
University of Mogi das Cruzes: Universidade de Mogi das Cruzes
Yara Natercia Lima Faustino de Maria
University of Mogi das Cruzes: Universidade de Mogi das Cruzes
Regina Costa de Oliveira
University of Mogi das Cruzes: Universidade de Mogi das Cruzes
Daniela L. Jabes
University of Mogi das Cruzes: Universidade de Mogi das Cruzes
Luiz R. Nunes
Universidade Federal do ABC
Fabiano Bezerra Menegidio (fabiano.menegidio@biology.bio.br)
Universidade de Mogi das Cruzes https://orcid.org/0000-0002-4705-8352

Short Report

Keywords: Talaromyces marneffei, pan-transcriptome, transcription isomorphs, ncRNA, lncRNAs, RNA-seq

Posted Date: November 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1050608/v1

Abstract

Previous genomic/transcriptomic analyses of *Talaromyces marneffei* (TM) unravelled relevant pathogenicity-related elements, as well as chromosomal regions potentially involved with the production of non-coding RNAs (ncRNAs), which have been parsimoniously reported in fungi. This manuscript describes a comprehensive pan-transcriptome assembly for TM that identifies a series of previously undetected genetic elements in this emerging pathogenic fungus. Our results confirm that ~58.28% of the 9,480 genes currently annotated in the TM genome are, in fact, transcribed in vivo and that ~23.6% of them may display alternative isomorphs. Moreover, we identified 585 transcripts that do not match any gene currently mapped in the genome, represented by 90 coding transcripts and 140 ncRNAs, including 48 long non-coding RNAs (lncRNAs). Overall, we expect that the novel elements described herein may contribute to improve the currently available *Talaromyces* databases and foster studies aiming at characterizing lncRNA-mediated gene expression control in fungi.

1. Introduction

Talaromyces marneffei (TM), previously known as *Penicillium marneffei*, is a thermally dimorphic fungus viewed as an important emerging pathogen in endemic regions of Southeast Asia, Southern China and Eastern India [1–3]. Recent data showed that TM infection correlates with the highest mortality rates among HIV/AIDS patients, surpassing those observed in most HIV-associated complications [3–4]. TM infection has also been occasionally observed in individuals submitted to treatments involving immunosuppressive drugs (including cancer, systemic lupus erythematosus and bone marrow/solid organ transplants, among others) [5].

Omics analyses constitute some of the most effective tools currently available to properly understand the general biology and virulence mechanisms in pathogenic fungi. Accordingly, analysis of the ~28.9 Mb draft TM genome unravelled a series of relevant elements, which seem to be associated with development and virulence in this microorganism, including meiotic genes, proteins involved in pheromone response pathways and putative virulence factors [6]. Moreover, genome annotation identified chromosomal regions that potentially encode a series of non-coding RNAs (ncRNAs) in TM, while preliminary transcriptome analyses provided evidence for the existence of microRNA-like RNAs (miRNAs) in this microorganism [7–10]. Although miRNAs and other types of ncRNAs are important controllers of gene expression in plants and animals, their presence in fungi, as well as their role in controlling the outcome of relevant biological processes in such organisms has only started to be discovered [11–16].

2. Material And Methods

2.1. Raw Libraries collection

The TM mycelial-yeast pan-transcriptome described herein was built from RNA-seq libraries of *T. marneffei* PM1, available at the NCBI SRA database, under project ID PRJNA212740 [9, 6, 15].

2.2. Bioinformatics Analysis

Reads were submitted to a workflow previously described by [16], with minor modifications. Briefly, raw FASTQ sequencing data were processed in a Public Galaxy Server, available at usegalaxy.eu. Initially, the quality of raw sequences was assessed using FastQC [17] and MultiQC [18]. Fastp [19] was then used to remove low-quality reads (Q<30) and adapters. To remove sequences from the human host, reads were aligned with Bowtie2 [20] against a local database containing the human reference genome hg38, available at Galaxy Europe (usegalaxy.eu). To remove rRNA reads, the high-quality reads were aligned to sequences in the SILVA ribosomal RNA (rRNA) [21] and Rfam databases [22] using SortMeRNA [23]. NCBI UniVec database (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) was used to remove vector contaminants from libraries, with the aid of Bowtie2. Quality-filtered reads were then mapped to the latest version of the *T. marneffei* PM1 reference genome (GCA_000750115, ASM75011v1), available at the Ensembl Fungi database [24], using HISAT2 [25].

StringTie [25] was then used to assemble the mapped reads into transcripts, using the *de novo* transcriptome reconstruction method, allowing identification of all transcripts present in each sample (including currently annotated genes, as well as newly identified elements and isomorphs). StringTie merge [25] was next used to combine redundant transcription structures, providing a unified reference transcriptome, with unique identifiers. Cufflinks [26] was then used to estimate expression values (FPKM) for each element in the StringTie-generated transcriptome. Transcriptome completeness analysis was conducted with the aid of BUSCO using the Fungi odb10 reference database [27]. Finally, transcripts were classified into different Transcription Class Codes (TCCs), reflecting their respective nature/origin, with the aid of Cuffcompare [26], using a GFF3 reference annotation file for the *T. marneffei* PM1 genome (obtained from Ensembl Fungi). The final assembly derived from the StringTie/Cufflinks/Cuffcompare analyses was filtered by expression level and only elements displaying FPKM ≥ 1 and TCCs “=”, “j”, “i”, “u” and “x” were considered real transcripts and used in subsequent analyses, as suggested by [16].

2.2. Functional Annotation

Functional annotations were obtained with the aid of the Eukaryotic Non-Model Transcriptome Annotation Pipeline (EnTAP) [28]. Contigs were queried for similarity (blastx, e-value ≤ e-5) against the National Center for Biotechnology Information non-redundant protein database (NCBI nr); NCBI proteins reference database (RefSeq); Swiss-Prot curated database from UniProt Knowledgebase (UniProtKB) and the EggNOG proteins database. EnTAP ontology searches via EggNOG also helped to assign the biological function to the genes, identifying GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) terms. Additionally, InterProScan [29] mapped the transcripts to their respective protein families. Elements not annotated by EnTAP were evaluated with Infernal (cmscan) [30] and classified into the different families of ncRNAs defined in the Rfam database. After Infernal annotation, blastn (e-value ≤ e-5) searches compared possible novel transcripts with EST (expressed sequence tags) data available for *Talaromyces* and other *Trichocomaceae*, at NCBI.

2.3. lncRNA discovery and conservation

All real transcripts were evaluated for their respective coding potential (CP) with the aid of three tools: RNAsamba [31], Coding Potential Calculator (CPC2) [32] and Coding Potential Assessment Tool CPAT [33]. In this work, we considered putative long non-coding RNAs (lncRNA), those transcripts with length ≥200 bp, not annotated by EnTAP/Infernal and that were identified as non-coding by all CP tools. Finally, these putative lncRNAs were submitted to FEElnc [34], which evaluated their potential nature as lncRNAs, identified their respective partner genes and classified the new lncRNAs to their localization and the directionality, in relation to their partner RNAs. To discover conserved ncRNAs, we aligned putative ncRNA sequences against *Talaromyces* ncRNAs sequences (available at the Ensembl Fungi database) using blastn with a cut-off e-value ≤ e-5, coverage ≥ 90 and identity ≥ 50%.

2.4. Reproducibility of methods

To guarantee the reproducibility of the data presented in this manuscript, we provide all information about the versions, sources and references for all tools used to obtain the results presented herein in the Supplementary Table ST1-S0a. The workflow scheme is also provided in the Supplementary Table ST1-S0b.

3. Results And Discussion

After submitting approximately 100 million paired-end RNA-seq reads to the pipeline described above, we obtained a preliminary transcriptome assembly containing 9,275 elements, which are summarized in Supplementary Table ST1-S0c and shown in detail in Supplementary Table ST1-S1 (available at doi.org/10.6084/m9.figshare.14143757.v3). Transcriptome completeness identified 655/758 (86.4%) of the universal genes present in the Fungi odb10, supporting the high quality of this assembly (Fig. 1A; Table 1; Supplementary Table ST1-S2).

Transcripts were classified into different Transcription Class Codes (TCCs), reflecting their respective nature/origin (see Supplementary Table ST1-S0d) and filtered, to maintain only transcripts displaying FPKM ≥ 1 and belonging to TCCs “=”, “j”, “u”, “i” and “x”. This resulted in a reference TM pan-transcriptome containing 8,613 elements (Table 1; Supplementary Table ST1-S3). In total, 5,525 of these elements belong to TCC “=”, representing transcripts with exact match to a gene’s exon chain (Table 1; Supplementary Table ST1-S3). We also identified 2,503 putative mRNA isomorphs (derived from 2,236 genes), characterized as multi-exon transcripts, containing at least one exon-exon junction match (TCC “j”) (Table 1; Supplementary Table ST1-S3). This suggests that ~23.6% of the 9,480 CDSs currently mapped in the *T. marneffei* PM1 draft genome [6] may be alternatively processed through alternative splicing, differential polyadenylation, or alternative transcription start sites. Finally, we identified 585 transcripts that do not match any gene currently mapped in the TM genome (Table 1; Supplementary Table ST1-S3). These novel transcripts are represented by 82 elements from TCC “u” (transcripts mapping in intergenic regions), 2 elements from TCC “i” (intronic transcripts) and 501 elements from TCC “x” (transcripts displaying overlap with exonic regions, but mapped on the opposite strand). Since the most recent Ensembl genomic annotation for the *T. marneffei* PM1 genome identified 9,480 genes, the

585 novel elements identified herein represent an ~5.81% increase in the number of genes identified in this fungus.

The 8,613 transcripts in the reference pan-transcriptome were annotated with EnTAP against Swiss-Prot, EggNOG proteins, RefSeq and Nr proteins databases, resulting in matches for 8,219 transcripts (95.4%). Most transcripts (6,790; 78.8%) were functionally annotated with GO terms. Among these, 6,469 transcripts (75.1%) displayed matches to GO Biological Processes, while 5,189 (60.2%) displayed matches to GO Cellular Components and 6,351 (73.7%) to GO Molecular Functions. The ten most frequent functional groups within each GO category are shown in Fig. 1C. In addition, 2,353 (27.3%) isomorphs were annotated into at least one KEGG pathway term (Supplementary Table ST1-S4). InterProScan assigned protein family hits for 8,469 (98.3%) isomorphs (Supplementary Table ST1-S4). Next, transcripts not annotated by EnTAP were distributed into 10 RNA classes of the Rfam database, with the aid of Infernal, allowing their identification as non-coding RNAs (ncRNAs), based on their similarity to a series of such molecules, previously described in other organisms (Supplementary Table ST1-S4). Among these, are the small Cajal body RNAs (scaRNAs), involved in modifying small nucleolar RNAs (snoRNAs), such as snR80, widely observed in the Dikarya sub-kingdom [14]. We also identified snosnR61, described in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*; and the Afu transcripts, previously described in *Aspergillus fumigatus*, *Paracoccidioides brasiliensis* and other fungi [10–13, 15]. After Infernal annotation, we conducted blastn searches, comparing the sequences of all 585 novel transcripts with EST data available for *Talaromyces* and other *Trichocomaceae*, at NCBI. This analysis resulted in matches for 100 transcripts (Table 1; Supplementary Table ST1-S5), confirming the existence of a large proportion of such novel elements (~17.1%) by an independent approach, based on direct cDNA cloning (thus, reinforcing that they are not PCR artifacts produced during RNA-seq library construction.

All 8,613 transcripts were also evaluated for their coding potential (CP) with the aid of a custom pipeline, based on three CP prediction tools: CPAT, CPC2 and RNAsamba (see Material and Methods). These analyses led to the identification of 7,358 coding and 313 noncoding transcripts (Table 1; Supplementary Table ST1-S4). Additionally, there were 942 elements with undetermined coding potential, since their analyses by the CP-prediction tools provided conflicting results (Table 1; Supplementary Table ST1-S4). From the 7,358 coding elements, 90 were considered as derived from previously unmapped genes, since they do not correspond to any coding sequence (CDS) previously described in the TM genome (TCCs “u”, “x”) (Table 1; Supplementary Table ST1-S6). The 313 noncoding transcripts were further evaluated by FEElnc, which identified 102 lncRNAs among such elements, along with their corresponding partner mRNAs (Table 1; Supplementary Table ST1-S7-S8). Thus, the remaining 211 elements were considered novel putative non-coding RNAs (ncRNAs) (Table 1; Supplementary Table ST1-S9). Finally, the 313 putative ncRNAs were submitted to a blastn analysis against previously described ncRNA sequences from *Talaromyces* (available at the Ensembl Fungi database). These analyses resulted in matches for 125 putative ncRNAs reported herein (Table 1; Supplementary Table ST1-S10), indicating that our reference transcriptome identified 188 novel noncoding elements in the *Talaromyces* genus, including 48 novel lncRNAs (Table 1; Supplementary Table ST1-S11).

Declarations

Conflicts of interest/Competing interests

The authors report no conflicts of interest. The authors alone are responsible for the content and the writing of the paper.

Funding

This study was financed in part by scholarship grants from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil - CAPES (www.capes.gov.br/) (awarded to D.A.B., Y.N.L.F.M) and Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (www.gov.br/cnpq/pt-br) (awarded to A.S.S; A.C.H).

Availability of data and material

Supplementary Table S1 is available from the Figshare repository (doi.org/10.6084/m9.figshare.14143757.v3). Raw sequencing reads are available at NCBI Sequence Read Archive (SRA), under BioProject number PRJNA212740. Additional data derived from this study (including all intermediate data) is also available from the Open Science Framework (OSF) repository (DOI 10.17605/OSF.IO/XHN75).

References

1. Hu Y, Zhang J, Li X, Yang Y, Zhang Y, Ma J, Xi L (2013) *Penicillium marneffei* infection: an emerging disease in mainland china. Mycopathologia. https://doi.org/10.1007/s11046-012-9577-0
2. Stathakis A, Lim KP, Boan P, Lavender M, Wrobel J, Musk M, Heath CH (2015) *Penicillium marneffei* infection in a lung transplant recipient. Transpl Infect Dis. https://doi.org/10.1111/tid.12377
3. Chen Y, Huang A, Ao W, Wang Z, Yuan J, Song Q, Wei D, Ye H (2018) Proteomic analysis of serum proteins from HIV/AIDS patients with *Talaromyces marneffei* infection by TMT labeling-based quantitative proteomics. Clin Proteomics. https://doi.org/10.1186/s12014-018-9219-8
4. Jiang J, Meng S, Huang S et al., R (2019) Effects of *Talaromyces marneffei* infection on mortality of HIV/AIDS patients in southern China: a retrospective cohort study. Clin Microbiol Infect. https://doi.org/10.1016/j.cmi.2018.04.018
5. Chan JF, Lau SK, Yuen KY, Woo PC (2016) *Talaromyces* (*Penicillium*) *marneffei* infection in non-HIV-infected patients. Emerg. microbes & infect. https://doi.org/10.1038/emi.2016.18
6. Yang E, Chow WN, Wang G, Woo PC, Lau SK, Yuen KY, Lin X, Cai JJ (2014) Signature gene expression reveals novel clues to the molecular mechanisms of dimorphic transition in *Penicillium marneffei*. PLoS Genet. https://doi.org/10.1371/journal.pgen.1004662
7. Lin X, Ran Y, Gou L, He F, Zhang R, Wang P, Dai Y (2012) Comprehensive transcription analysis of human pathogenic fungus *Penicillium marneffei* in mycelial and yeast cells. Med Mycol.

https://doi.org/10.3109/13693786.2012.678398

8. Pasricha S, Payne M, Canovas D, Pase L, Ngaosuwankul N, Beard S, Oshlack A, Smyth GK, Chaiyaroj SC, Boyce KJ, Andrianopoulos A (2013) Cell-type-specific transcriptional profiles of the dimorphic pathogen *Penicillium marneffei* reflect distinct reproductive, morphological, and environmental demands. G3 (Bethesda). https://doi.org/10.1534/g3.113.006809
9. Yang E, Wang G, Woo PC, Lau SK, Chow WN, Chong KT, Tse H, Kao RY, Chan CM, Che X, Yuen KY, Cai JJ (2013) Unraveling the molecular basis of temperature-dependent genetic regulation in *Penicillium marneffei*. Eukaryot Cell. https://doi.org/10.1128/ec.00159-13
10. Lau SKP, Chow WN, Wong AYP, Yeung JMY, Bao J, et al. (2013) Identification of MicroRNA-Like RNAs in Mycelial and Yeast Phases of the Thermal Dimorphic Fungus *Penicillium marneffei*. PLoS Negl Trop Dis. https://doi.org/10.1371/journal.pntd.0002398
11. Jöchl C, Rederstorff M, Hertel J, Stadler PF, Hofacker IL, Schrettl M, Haas H, Hüttenhofer A (2008) Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis. Nucleic Acids Res. https://doi.org/10.1093/nar/gkn123
12. Canzler S, Stadler PF, Schor J (2018) The fungal snoRNAome. RNA. https://doi.org/10.1261/rna.062778.117
13. Logan MK, Burke MF, Hebert MD (2018) Altered dynamics of scaRNA2 and scaRNA9 in response to stress correlates with disrupted nuclear organization. Biol Open https://doi.org/10.1242/bio.037101
14. Parker S, Fraczek MG, Wu J, Shamsah S, Manousaki A, Dungrattanalert K, de Almeida RA, Invernizzi E, Burgis T, Omara W, Griffiths-Jones S, Delneri D, O'Keefe RT (2018) Large-scale profiling of noncoding RNA function in yeast. PLoS Genet. https://doi.org/10.1371/journal.pgen.1007253
15. Wang Q, Du M, Wang S, Liu L, Xiao L, Wang L, Li T, Zhuang H, Yang E (2018) MADS-Box Transcription Factor MadsA Regulates Dimorphic Transition, Conidiation, and Germination of *Talaromyces marneffei*. Front Microbiol. https://doi.org/10.3389/fmicb.2018.01781
16. Menegidio FB, Aciole Barbosa D, Alencar VC, Vilas Boas RO, Costa de Oliveira R, Jabes DL, Nunes LR (2021) Transcriptomic profiling identifies novel transcripts, isomorphs, and noncoding RNAs in *Paracoccidioides brasiliensis*. Med Mycol. https://doi.org/10.1093/mmy/myaa062
17. Andrews S (2010). FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 18 Jun 2021.
18. Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw354
19. Chen S, Zhou Y, Chen Y, Gu J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics. https://doi.org/10.1093/bioinformatics/bty560
20. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods. https://doi.org/10.1038/nmeth.1923
21. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. https://doi.org/10.1093/nar/gks1219

22. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, Petrov AI (2018) Non-Coding RNA Analysis Using the Rfam Database. Curr Protoc Bioinformatics. https://doi.org/10.1002/cpbi.51
23. Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts611
24. Howe KL, Contreras-Moreira B, De Silva N et al.(2020) Ensembl Genomes 2020-enabling non-vertebrate genomic research. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz890
25. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. https://doi.org/10.1038/nprot.2016.095
26. Trapnell C, Williams BA, Pertea G et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. https://doi.org/10.1038/nbt.1621
27. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2018) BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol. https://doi.org/10.1093/molbev/msx319
28. Hart AJ, Ginzburg S, Xu M, Fisher CR, Rahmatpour N, Mitton JB, Paul R, Wegrzyn JL (2020) EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. Mol Ecol Resour. https://doi.org/10.1111/1755-0998.13106
29. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res https://doi.org/10.1093/nar/gki442
30. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt509
31. Camargo AP, Sourkov V, Pereira GAG, Carazzolle MF (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. NAR Genom Bioinform. https://doi.org/10.1093/nargab/lqz024
32. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, Gao G (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx428
33. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. https://doi.org/10.1093/nar/gkt006
34. Wucher V, Legeai F, Hédan B et al. (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw1306

Tables

Table 1 is not available with this version.

Figures

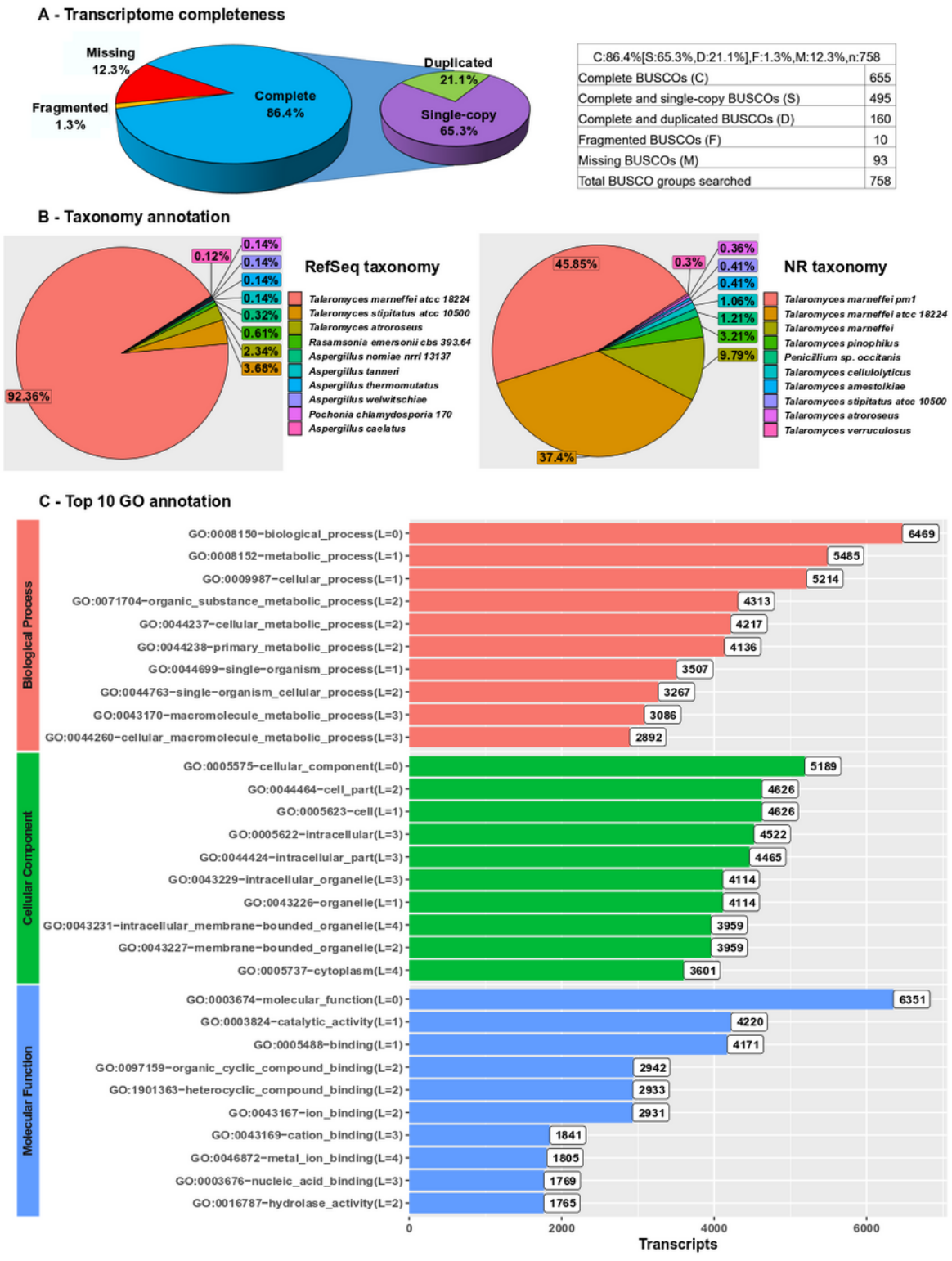

Figure 1

Homology search of Talaromyces marneffei PM1 transcripts. (A) Transcriptome completeness identifying 655/758 (86.4%) of the universal genes present in the reference Fungi odb10 lineage, supporting the high quality of the transcriptome assembly. (B) Species annotation distribution for the

best hits from NCBI Nr and NCBI RefSeq databases. (C) Gene ontology distribution for (i) biological process, (ii) cellular component and (iii) molecular function of assembled isoforms from the Talaromyces marneffei transcriptome.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TalaromycesmarneffeiPM1SupplementaryTableS1final.xlsx