scientific reports

Check for updates

OPEN

Exploring the hidden hot world of long non-coding RNAs in thermophilic fungus using a robust computational pipeline

Roger G. Silva[1], Paulo P. Amaral[2], Glória R. Franco[3] & Aristóteles Góes-Neto[1]✉

Long noncoding RNAs (lncRNAs) are versatile RNA molecules recently identified as key regulators of gene expression in response to environmental stress. Our primary focus in this study was to develop a robust computational pipeline for identifying structurally identical lncRNAs across replicates from publicly available bulk RNA-seq datasets. In order to demonstrate the effectiveness of the pipeline, we utilized the transcriptome of the thermophilic fungus *Thermothelomyces thermophilus* and assessed the expression pattern of lncRNAs in conjunction with Heat Shock Proteins (HSP), a well-known protein family critical for the cell's response to high temperatures. Our findings demonstrate that the identification of structurally identical transcripts among replicates in this thermophilic fungus ensures the reliability and accuracy of RNA studies, contributing to the validity of biological interpretations. Furthermore, the majority of lncRNAs exhibited a distinct expression pattern compared to HSPs. Our study contributes to advancing the understanding of the biological mechanisms comprising lncRNAs in thermophilic fungi.

Keywords Long non-coding RNA, Structurally identical transcripts, *Thermothelomyces thermophilus*, Transcriptome assembly

Fungi constitute a diverse and fascinating kingdom of organisms. We can cite the *Psilocybe*[1], which contains psychoactive compounds that can induce hallucination upon ingestion. There are also hundreds of parasitic fungi from the order Hypocreales, such as *Ophiocordyceps*, that infect insects and spiders, turning them into living zombies[2,3]. Additionally, there are bioluminescent fungi that can be used as a natural source of light[4]. Those intriguing complex eukaryotic organisms have also been found thriving in conditions where temperatures can induce effects on their genomes and influence various molecular processes. Thermophilic fungi are able to grow at high temperatures, typically between 40 and 60 ℃, unable to grow below 24 ℃, and can be found in extreme environments, such as compost piles and organic residues[5]. Their unique genomic features and different functions in environments and industrial facilities, due to their ability to produce thermotolerant enzymes[5], make them interesting and biotechnologically important organisms to be studied.

Thermophilic fungi have evolved to survive and thrive in high-temperature environments, which means that they might have developed specific genetic adaptations to help them cope with thermal stress since high temperatures can cause DNA denaturation, increasing rates of spontaneous mutations, and might affect gene expression[6]. One well-known genetic adaptation of organisms that have to tackle with heat stress is the highly conserved heat-shock proteins (HSP) family[7]. This family of stress-induced proteins protects cellular damage by stabilizing proteins in the cell, preventing proteins from being aggregated, assisting misfolded, damaged or newly synthesized proteins, and, ultimately, the removal of damaged proteins[7]. Furthermore, it has been well established that HSPs are essential for maintaining cellular homeostasis and protecting cells from various forms of stressors including heat stress[7].

Although fungal genome sizes vary greatly within their kingdom[8], likely due to numerous genome reduction and expansion events that may have occurred over millions of years, thermophilic genome reduction is noticeable when one compares their genome to the closest mesophilic counterparts[9]. Genome reduction in thermophilic

[1]Molecular and Computational Biology of Fungi Laboratory, Department of Microbiology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. [2]Institute of Education and Research, São Paulo, SP, Brazil. [3]Department of Biochemistry and Immunology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. ✉email: arigoesneto@icb.ufmg.br

fungi also includes reduction of intergenic and repetitive sequences, loss of protein-coding genes (PCG), and transposable elements[6]. Therefore, the distinctive features of fungal genomes, including those of thermophilic fungi, pose challenges for researchers, prompting us to question which other modifications thermophilic fungi have likely developed to survive in high-temperature environments.

Recently, important roles in cellular responses to environmental stress have been assigned to long noncoding RNAs (lncRNAs)[10–14]. These include regulating double-strand DNA breaks[10], differentially expressed lncRNAs under various stress stimuli[11], regulating gene expression in response to high water and CO_2 concentrations[12] as well as stress conditions in the nucleolus[13], and chromatin modification under low temperature[14]. All those functions support the stress environmental responses at the molecular and genomic levels within cells.

The singularity of RNA molecules encompasses not only its physical structure but also its abundance, functional diversity, and uniqueness[15]. RNA molecules can be found in a double-stranded fashioned form[15]; however, the great majority is uni-stranded, manufactured in the cell nucleus, and subsequently migrate into other cell compartments[15]. Beside the well-known messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) triad, acting in transcription and translation, there are a multitude of other RNA molecules within the cell that do not code for a protein[16]. Non-coding RNAs (ncRNAs) are an extensive group of RNA molecules classified into housekeeping and regulatory transcripts[17]. The former group contains the rRNA and the tRNA that are molecules involved in reading and linking amino acids, protein translation and transport whereas the latter can be further classified based on transcript length as small RNA (< 200 nt) and long non-coding RNA (≥ 200 nt), with other subsets within this group[17].

Evidenced by pervasively transcribed and abundant in plants[18], animals[19], fungi[20], bacteria[21] and even in viral genomes[22], lncRNAs are alternatively spliced transcripts longer than 200 nucleotides, which do not encode proteins, are functional in their primary, secondary, or tertiary structures, display low conservation and expression when compared to PCG, can encode small peptides, and are mostly transcribed by RNA polymerase II[15]. Once induced and folded, lncRNAs can interact with DNA, RNAs, and proteins, and regulate gene expression using different biological mechanisms[15]. Their characterization is still in its initial stage, and their evolutionary origins remain obscure, but some have been assessed by small or no open reading frame (ORF), others either 5′-capped, polyadenylated, or not, and undergo post-transcriptional modifications[15]. Additionally, they can act as thermal sensors and regulate gene expressions in response to thermal stress[23,24], and are classified according to their position in the genome, such as sense, antisense, intergenic, intronic, and bidirectional[15].

While numerous studies have explored the world of lncRNAs across various fungal species[25–30], elucidating their regulatory mechanisms[25], biological functions[26], and roles in fungal pathogenesis[27], as well as their interplay with pathogenic fungi and their hosts[28–30], our understanding of their roles in thermophilic fungi remains in its early stages. To date, only one study has identified in-silico lncRNAs in thermophilic fungi[31]. While this pioneering study represents a crucial first step towards elucidating the regulatory landscape of lncRNAs in thermophilic fungi, it emphasizes the need for robust approaches to studying these versatile non-coding molecules under temperature constraints.

Considering all the aforementioned challenges of the thermophilic fungal genome traits, an accurate transcript quantification on reference poorly-annotated models is itself a daunting task, given the numerous variables that can interfere, such as: (i) experimental design, (ii) biological replicates, (iii) RNA extraction, (iv) library preparation, (v) sequencing, (vi) data preprocessing, (vii) alignment, (viii) assembly, as well as other external factors. Yet, analysis of lncRNA adds another complexity level on top of all of those variables. Therefore, the primary focus of our study was to develop a robust and reliable computational pipeline for identifying structurally identical lncRNAs in thermophilic fungi across distinct replicates and investigate their relationship with HSPs, using publicly available bulk RNA-seq datasets of the thermophilic fungi *Thermothelomyces thermophilus* as a case study, which can and must be extended to any eukaryotic thermophilic organism.

Material and methods

Transcriptome data

According to Liu et al. (2022), transcriptome raw reads, RPKM (GSE184074_RPKM.xlsx), and raw counts (GSE184074_Readcounts.xlsx) spreadsheets containing reads counts were deposited in the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI), under accession number GSE184074. The transcriptome data were obtained from the SRA repository (https://www.ncbi.nlm.nih.gov/sra), using the parallel faster-dump tool (SRA-Toolkit). In total, 12 paired-end 150 bp poly-A cDNA RNA-seq libraries were downloaded for the thermophilic fungus *Thermothelomyces thermophilus*. The taxonomic analysis presented distribution of reads mapping to the *Thermothelomyces thermophilus ATCC 42464* species > 90%. The fungus was cultivated in four temperatures: 35, 40, 45, and 50 ℃ with three replicates for each experiment.

Computational pipeline

In order to identify lncRNA genes from transcriptome data, a computational pipeline was developed using Python 3.10.8 and the conda package management environment (version 4.12.0) with the bioconda channel added. A dedicated conda environment was created for installing all transcriptome tool packages and their dependencies used in this pipeline (Supplementary Material—Note 1). The *Thermothelomyces thermophilus* ATCC 42464 reference genome (ASM22609v1) along with its annotation file were downloaded from the NCBI website.

Quality, adapter removal, trimming, and filtering

After downloading the libraries, initial quality control of sequencing data was performed using FastQC[32] and Multiqc[33]. The BBDuk tool from Joint Genome Institute BBTools was then utilized for quality control filtering,

adapter removal, and decontamination[34]. BBTools is a bioinformatics toolkit used for processing and assessing DNA and RNA sequence data. It also has the capability of removing adapter sequences, filtering out low-quality reads, and removing contamination sequences from bacterial and eukaryotic ribosomal RNA. The set of adapters used in this cleaning stage came from the BBDuk repository.

All the library files displayed an overall Phred score above 30, according to FastQC application, indicating an error rate of one in every 1000 nucleobases sequenced. To clean the data, the BBduk tool was used with the parameters *QTRIM = 'rl', MINLEN = 50* and *TRIMQ = 20* which, respectively, trims the sequences on both ends, discards reads shorter than 50 bp, and removes low-quality reads below 20 Phred.

Decontamination

Ribosomal sequences were obtained from the RFam database[35] and used to generate a custom database of ribosomal RNA sequences, including 5S rRNA (*RF00001*), 5.8S rRNA (*RF00002*), tRNA (*RF00005*), Eukaryotic small subunit ribosomal RNA (*RF01960*), eukaryotic large subunit ribosomal RNA (*RF02543*), and bacterial large subunit ribosomal RNA (*RF02541*). The decontamination process using Kmers was also performed using the BBduk tool. This tool is able to remove or decontaminate ribosomal and bacterial contaminant sequences.

Strand detection

This step is essential and ensures that antisense lncRNA transcripts are not misclassified as PCG, as lncRNAs overlap on the opposite DNA strand with PCG features. Therefore, raw reads were processed before alignment using the Salmon pseudoalignment tool[36] with default parameters and *–gcBias –validateMappings –numGibbsSamples 200 –seqBias* flagged. The type of library can be found in the output file *salmon_quant.log*. A two-column text file was produced where it was saved each library name and its strandness respectively.

Sequencing reads alignment

Before aligning, it was necessary to create the organism genome index. This process involves three main steps: identifying splice sites (*hisat2_extract_splice_sites.py*), determining exon positions (*hisat2_extract_exons.py*), and constructing the genome index (*hisat2-build*) using the resulting splice sites and exon location files. The processed reads were subsequently aligned to the thermophilic fungal reference genome. The main following set of parameters were set in HISAT2[37] *–max-intronlen 100 –dta -a –secondary –no-mixed –no-discordant*. The parameters *–no-mixed* and *–no-discordant* were set to not align individual mates nor discordant alignments, respectively. The HISAT2 alignment tool generates an output file in SAM format used to store the alignment of sequences, which was subsequently sorted by coordinates and converted to BAM format and indexed for efficient data retrieval and manipulation as described in the Samtools manual. The intron length parameter was derived from the intron length observed in the JGI GeneModel data for closely related thermophilic fungi[38].

Transcript structure identification and assembly

The genome-guided transcriptome assembly StringTie2[39] tool was utilized for assembling the transcriptome libraries. In accordance with the protocol described in the StringTie2 paper, the tool was executed in three consecutive steps (transcriptome assembly, merging of transcriptomes into a non-redundant transcriptome, and transcript abundance estimation) to produce a meta-assembly transcriptome. For the initial step, StringTie2 was executed with the parameters *-j 5* that requires at least five spliced reads to be aligned across a junction, and *-c 10*, which sets a minimum coverage of 10 reads for a transcript to be predicted. All other parameters were set to their default value. The next step in the pipeline was the merge step. StringTie2 was executed with the option *-g 10*, whose value was selected due to thermophilic genome reduction characteristic. This parameter specifies a gap separation to merge neighbor transcripts, meaning that a gap between two transcripts less than 10 base pairs were merged together. In this step, *gffcompare* and *gffread*[40] were used for evaluating transcript assemblies and extracting FASTA files from the already merged assembled transcriptome. Finally, StringTie2 was performed for each library with the options *-e* and *-B* for estimating transcript abundances.

Transcriptome assessment

The merged meta-assembly FASTA file produced by the *gffread* was evaluated by the rnaQUAST tool[41] according to the reference genome and its annotation file.

Statistical identification of differentially expressed genes

The authors of StringTie2 provide a Python 3 script called *prepDE.py3* to generate a gene and transcript count matrix files to be used with the DESeq2 R package[42]. Before executing the script *prepDE.py3*, a text file was created listing the 12 SRA run IDs, a blank space, and their full paths to the merged GTF file for each SRA ID. The *prepDE.py3* script was executed with the -l parameter set to 146, which represents the average read length for each library. This value was verified using the *samtools view command* divided by the amount of the returned lines in the samtools command (Supplementary Material—Note 4).

The gene count matrix was then processed using DESeq2 according to DESeq2 vignette instructions and one treatment (temperature) as design formula. The fungal culture, which was cultivated at 35 ℃, was designated as the control, and it was compared to other experiments that were conducted with the fungus growing at 40, 45, and 50 ℃. Read counts below 10 were removed from the analysis before executing the DESeq2 analysis, and the results were filtered based on p-adj < 0.05 and – 1.5 < log2 fold change < 1.5. Genes that met the DESeq2 thresholds were considered differentially expressed (DEG). Principal component analysis (PCA) and hierarchical clustering using the DESeq2 package between groups were also performed.

Enrichment analysis

Functional enrichment analysis was conducted by STRING-DB[43]. The gene symbol list of all DEGs from the experiments at 35 ℃ (control) and 50 ℃ (treatment) were submitted to STRING analysis. STRING identified all genes as belonging to the *Thermothelomyces thermophilus ATCC 42464* species. The confidence score for the predicted protein–protein interaction was set to the high confidence level (0.700), and the maximum number of interactions to show in the first and second shells was set to none due to the large number of interactions in the network.

Heat shock proteins (HSP) enrichment

Gene symbols for annotated *Thermothelomyces thermophilus* HSP were retrieved from the Uniprot database[44] (Supplementary Material—Note 2) and used to select DEGs from the initial DESeq2 results. The shortlisted genes for the experiment 35 ℃ (control) and 50 ℃ (treatment) along with their log2 fold change, p-value and p-adj values were filtered by bash commands (Supplementary Material—Note 5).

Structurally identical transcripts

The *gffcompare* tool outputs a tracking file that lists structurally equivalent transcripts across all RNA-seq samples, allowing variations in the lengths of the first and last exons but requiring identical intron lengths due to alternative splicing events. The reporting of a transcript in the tracking file does not necessarily require its presence in all samples. The character '-' represents a transcript that was not included in the tracking file or was not expressed or detected in that particular sample. For the downstream analysis, only transcripts that were structurally identical across all replicates in each experiment (at 35, 40, 45, and 50 ℃) were selected and used for filtering lncRNAs and HSPs from the curated reads. Once the filtering step identified the reliable transcripts, they were linked back to their corresponding genes, and the analysis proceeded with DESeq2 at the gene level.

LncRNA identification

The *gffcompare*, when executed with *-r* option, classifies transcripts based on their position within the reference genome. By enabling this option, it outputs within the GTF file a field named "class_code" containing one single character. For downstream analysis of lncRNA sequences, the classes "u", "x", and "i" were selected, and these codes represent intergenic, antisense, and intragenic transcripts respectively. In the pipeline, *gffread* was used with the *-W* option to produce a more detailed FASTA file header for each sequence, including the *class_code* field in the heading content. This option facilitates sorting lncRNA sequences using only bash commands (Supplementary Material—Note 6) by just looking at the FASTA sequence header from the merged assembled transcripts file.

A local BLAST database from all Fungi PCG, downloaded from NCBI, was created, and the lncRNA gene catalog was compared to the local protein database using BLASTx[45] with *-max_target_seqs 10, -max_hsps 10* and *-evalue 1e-3*. Antisense lncRNA sequences were processed before executing BLASTx because those lncRNAs are localized on the opposite DNA strand and can overlap with protein-coding genes and, consequently, executing BLASTx would identify part of those sequences as belonging to opposite coding exons. Therefore, antisense lncRNA sequences were trimmed off their overlapping protein-coding portion before processing them into a BLASTx.

Additionally, intermediate lncRNA sequences were processed on a local installed Interproscan tool[46] with *-appl sfld, funfam, panther, prints, smart, pfam, pirsr, tigrfam, superfamily, cdd, antifam* options and *-goterms -pathways*. Any lncRNA sequence exhibiting similarity to protein families or domains present in any of the databases encompassed by the InterPro consortium were classified as protein-coding sequences and subsequently excluded from downstream analysis.

Finally, lncRNAs sequences that have not aligned nor exhibited similarity to any other functional protein were assessed by their coding potential in both strands with the CPC2[47] tool.

LncRNAs coding potential evaluation

Regarding the prediction of lncRNA coding potential, the pipeline made use of similar approaches previously employed[48–50], to assess intergenic and intragenic lncRNA coding potential, except for antisense lncRNAs whose overlapping PCG sequences were removed before performing the BlastX searching.

For all intermediated lncRNA transcripts, the pipeline executed the following steps:

- A identity-based coding prediction searching using BlastX;
- A functional domain and family protein signature screening using InterProScan on different databases;
- A Support Vector Machine Learning algorithm trained with four features (Fickett TESTCODE score, ORF length, ORF integrity and isoelectric point) CPC2 on both DNA strands.

Transcripts that exhibited any evidence of protein-coding potential by BlastX or carried any known protein domains or either displayed any coding potential on both strands were considered as potential coding transcripts and were excluded from the subsequent analyses. Notice that, for coding potential lncRNA evaluation, the pipeline does not use any ORF (Open Reading Frame) prediction length tool since the algorithm CPC2 already makes use of this feature for predicting transcript coding potential. Only non-coding transcripts from CPC2 were considered reliable for the lncRNA downstream analysis.

Weighted gene co-expression network analysis

In order to examine the co-expression patterns of protein coding genes and lncRNA genes between control conditions maintained at 35 ℃ and experimental treatments at 40, 45, and 50 ℃, read counts matrix was processed using the Weighted Gene Co-Expression Network Analysis (WGCNA) R package[51]. This algorithm was utilized to generate a weighted correlation matrix and subsequently identify sets of highly correlated genes (modules) that share similar expression patterns across samples.

Prior to that analysis, read counts below a threshold (< 10) were excluded and a variance-stabilizing transformation from the DESeq2 package was applied. A soft-threshold power value of 20 was set for a signed Topological Overlap Measure (TOM) to identify clusters of co-expressed genes with a scale-free topology. Genes with a high degree of similarity were assigned high TOM scores while those with lower levels of dissimilarity (1-TOM) were considered to be more distantly related. The *mergeCutHeight* was set to 0.25, which is the height where the dendrogram is cut and determines the number and size of the resulting gene modules. Spearman's correlation test, which is robust to outliers and nonlinear relationships, was applied to the PCG and lncRNA module relationships and p-value < 0.05 was considered statistically significant.

Finally, a module heatmap was generated and normalized expression data from three modules were further analyzed, each of which was found to be associated with HSPs and characterized by the expression of lncRNAs.

Results and discussion

We developed a computational pipeline that facilitates the identification of structurally similar long non-coding RNA (lncRNA) transcripts using samples from unprocessed transcriptome libraries in the thermophilic fungi *Thermothelomyces thermophilus* as a case study. This pipeline integrates various tools (Supplementary Material—Note 1) and algorithms to minimize errors and simplify the lncRNA analysis process.

The transcript assembly and quantification tool StringTie2 "merge" function is designed for assembling potentially different transcript fragments across samples, while Gffcompare tool is used to identify and track identical transcripts or fragments that are present in each sample. Therefore, this computational pipeline employs the tracking functionality present in the *gffcompare* tool to track identical transcripts across different samples. Since the main goal of this pipeline is tracking identical transcripts following the transcriptome assembly, all subsequent downstream steps were modified to use only identical transcripts across samples.

In order to evaluate the pipeline accuracy and determine the relationship between lncRNAs and HSPs, as well as their potential roles in thermal stress, it was essential to assess whether the fungus experienced thermal stress during its culturing. Thus, the publicly available transcriptome data were primarily used to identify differentially expressed HSPs between the control (35 ℃) and treatment (50 ℃) experimental groups. Subsequently, an enrichment analysis using the STRING database was performed and the results before and after reads curation were compared.

Thermal stress verification

High quality read count data (Supplementary Fig. 1) obtained from the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) database from the fungus *Thermothelomyces thermophilus* were processed (Supplementary Table 1) to detect differentially expressed HSP. For this step, 21 HSP computationally annotated genes were retrieved from the UniProt database, and their expression profiles were analyzed in *Thermothelomyces thermophilus*. The number of HSP differentially expressed from each pairwise experiment (35 × 40 ℃, 35 × 45 ℃ and 35 × 50 ℃) is listed in Supplementary Table 2, and it was consolidated in Supplementary Fig. 4.

DESeq2 analysis identified 1661 upregulated and 1594 downregulated genes between treatments (40, 45 and 50 ℃) and control (35 ℃) with adjusted p-value < 0.05 (Supplementary Fig. 2A-C). Moreover, the clustered heatmap plot (Supplementary Fig. 2D) showed that all replicates at 35 ℃ were clustered together and separated from the replicates at 50 ℃, the two more extreme temperatures. Principal Component Analysis (Supplementary Fig. 2B) was also performed, and Principal Component 1 (PC1) and 2 (PC2) together accounted for 80% of the variability in the data and showed replicates grouping together. This indicates that our preliminary exploratory data analyses are retrieving most of the important patterns and trends in the data.

Functional enrichment analysis of HSP before reads curation

In order to further explore the expression of the HSPs when comparing treatment (50 ℃) and control (35 ℃) under thermal stress, a functional enrichment analysis using STRING was performed on a set of 1179 DEGs. This analysis identified three out of four STRING clusters (Supplementary Table 3) consisting of 17, 15, and 11 members respectively belonging to protein refolding and chaperone binding clusters with False Discovery Rate (FDR) < = 0.05. These results suggest that HSPs were actively expressed in the treatment experiment, which is consistent with their known roles in responding to thermal stress.

Reads curation, alignment, assembly and assessment of the de novo transcriptome

Considering the initial exploratory analysis in the transcriptome data in which HSP gene expressions had revealed differences between control (35 ℃) and treatment (50 ℃), suggesting that the fungus was cultivated under thermal stress, the computational pipeline was employed to generate a de novo transcriptome assembly with high quality and accuracy. A series of ordered processing steps were performed on the raw reads, including filtering out low-quality reads, trimming adapter sequences, removing reads with a low Phred score and ribosomal decontamination, to prepare them for subsequent analysis. All the libraries retained more than 93% of the total reads (Supplementary Table 4) after cleaning and decontamination. Moreover, library strandness was also

identified, and all the transcriptome libraries were detected as likely IU, which means inward and unstranded libraries.

Supplementary Tables 1 and 4 provide information regarding the RNA-Seq libraries, including the library identifier, the temperature of fungal cultivation, the percentage of uniquely aligned reads, and other relevant details. Additionally, supplementary Fig. 3 shows a Multiqc graph comparing each RNA-Seq alignment. The HISAT2 aligner depicted uniquely paired read mappings ranging from 89.77 to 93.17%. The meta-assembled transcriptome, generated by StringTie2, was evaluated by rnaQUAST, which demonstrated 0.998% database coverage and 0 unaligned transcripts (Supplementary Table 5). Supplementary Table 6 and 7 compare DEGs as well as HSP before and after reads curation.

Our results corroborate with those reported by Liu, D. et al[52] and demonstrated the fungus was cultivated under thermal stress. Our traditional alignment methodology demonstrated its effectiveness in capturing these effects, especially when compared to the transcript-level expression quantification method initially presented by Liu, D. et al[52]. We observed minor fluctuations in the DEG counts before and after curation, resulting in an increase of 156 DEGs after the curation process. This implies that, while data curation contributed to enhanced data quality, it did not lead to significant alterations in the overall DEG landscape. In contrast, HSP genes exhibited a slight positive variation. After curation, a subtle increase in the number of HSP genes was noted in some pairwise comparisons, particularly in the 40×45 and 40×50 conditions, indicating that data curation helped in capturing additional HSP genes (Supplementary Fig. 4). Additionally, a high-quality assembly, from uniquely aligned reads, affirmed the extensive coverage of the database and the absence of unaligned transcripts, reinforcing the accuracy of our assembled transcriptome (Supplementary Table 5).

Identification of lncRNAs and PCG with similar structures

In order to predict lncRNAs in the fungi, the pipeline selected structurally identical transcripts present in all samples of each experiment. To accomplish this, the pipeline used the *gffcompare* track output file. This file contains structurally similar transcripts with variations in the length of the first and last exons but retaining identical intron length patterns. This variation may occur due to alternative splicing events, which lead to different transcripts isoforms with different exon lengths. It is noteworthy that StringTie2 merge function is intended to assemble transcript fragments that could vary significantly in each sample while *gffcompare* tracks only the identical transcripts/fragments in each sample. The difference between those approaches can negatively impact the identification of overlapping transcripts. For instance, the StringTie2 merge function detected 9154 overlapping protein-coding gene transcripts in the fungi genome, while *gffcompare* merging function identified 7741 transcripts overlapping those same genes. It is important to note that both assemblies were executed with the same set of parameters.

In order to demonstrate the potential impact of not selecting structurally identical transcripts on lncRNA identification, which are transcripts known to have lower expression levels than PCG, three *in-silico* experiments were conducted. Transcript data from each experiment was tracked and analyzed based on whether it was detected in one, two, or all three samples for each experiment (Fig. 1).

The output from the *gffcompare* merged GTF file depicted the assembled transfrag `TCONS_00000770`, class code "u" (intergenic transcript) with 3 exons. Notice that the biological experiments were performed using three replicates of the fungus cultivated at temperatures ranging from 35 to 50 ℃, totalling 12 samples. Each sample received an identifier starting with the letter *q* and an <*ID*> number. *gffcompare* labeled the first assembly file as q1, the second as q2, and successively until q12, according to its processing order and the number of processed samples. Looking at Supplementary Table 1, specifically at the *gffcompare* ID column's name, the

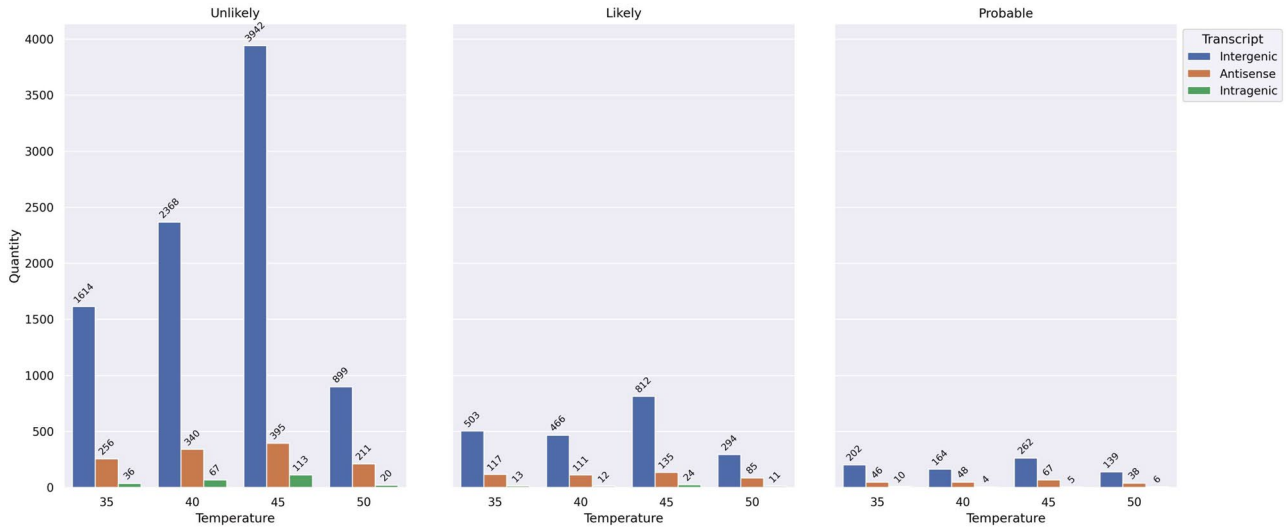

Figure 1. Number of lncRNA transcripts identified in one sample (Unlikely), two samples (Likely), or three samples (Probable) for experiments cultivated in different temperatures without assessing their coding potential.

identifier q1 belongs to the input file (GTF file) whose fungus was cultivated at 50 ℃; thus, this green intergenic transcript depicted in Fig. 2 was only identified in just one sample (out of three) for the same experiment and not expressed in any other samples even in different experiments. Transcripts appearing only in one sample for the same experiment were labeled as **Unlikely**.

In this another example (Fig. 2), transfrag `TCONS_00003679` (brown transfrags) with 2 exons, an antisense transcript (class code x), was expressed once in the 50 ℃ experiment (q2), three times in the 35 ℃ experiment (q3, q4, and q5), and once in the 40 ℃ experiment (q8). Setting the pipeline to recognize transcripts belonging to all samples for the same experiment, this transcript was labeled as **Probable** for the experiment at 35 ℃ only.

Figure 1 shows a comparison between lncRNA transcripts observed as **"Unlikely"** (one sample)**, "Likely"** (two samples) and **"Probable"** (three samples). The number of **"Unlikely"** transcripts, which are transcripts detected in only one sample, is much higher than those detected as **"Probable"** or in all samples for the same experiment. These values are raw values since those lncRNA transcripts have not been assessed yet for their coding potential by BlastX, InterproScan, or CPC2. Supplementary Table 8 summarizes the number of mRNAs and lncRNAs per sample and shows the total number of transcripts and their proportion per transcriptome. After the coding potential analysis, 399 transcripts (Supplementary Table 10) were considered putative lncRNA and utilized in downstream analyses.

The quantity of structurally identical lncRNAs revealed not only their presence in different temperature conditions but also sheds light on the variations in their expression levels. The classification scheme provides a clear distinction between lncRNAs that are consistently expressed in response to temperature changes and those that appear in a limited number of samples. The **"Probable"** lncRNAs, which are expressed in all three samples for each experiment, represent a set of transcripts that exhibit a robust and consistent presence, suggesting that these lncRNAs may play essential roles in the fungal response to thermal stress. On the other hand, the **"Unlikely"** lncRNAs, which are detected in only one sample, are a diverse group with the highest number of transcripts. These transcripts may represent sporadic or biological context-specific responses to temperature variations (Fig. 3 and supplementary Table 9). Hereafter, only **"Probable"** transcripts, lncRNAs and mRNAs, identified by the pipeline (Supplementary Table 9) were used for the downstream analysis.

In addition to the analysis of structurally identical lncRNAs, we extended our investigation to mRNA transcripts, specifically focusing on mRNAs that share structural identity across different temperature conditions, using the same pipeline. The results of this assessment are surprising, as these structurally identical mRNAs exhibited unique patterns of expression in response to temperature changes, following a distinct pattern when compared to their lncRNA counterparts (Supplementary Table 9). Moreover, functional enrichment analysis showed the enrichment of processes related to energy metabolism, which can be involved in repairing cell damage and also increase the production of secondary metabolites[53]. The quantity of "**Probable"** mRNAs consistently decreases with rising temperatures, starting with 6288 transcripts at 35 ℃, reaching 5382 transcripts at 45 ℃, which is its optimal growth temperature[54], having also the highest number of intergenic (262) and antisense (67) lncRNAs. Additionally, at the same temperature, differentially expressed lncRNAs showed the highest number, reaching 56 DEGs, being 51 lncRNAs up-regulated.

At the highest experimental temperature of 50 ℃, the number of **"Probable"** mRNA transcripts increased to 6422, contrasting to the reduced numbers of intergenic (139) and antisense (38) lncRNAs, which were the lowest among all temperatures. Notably, at 50 ℃, the fungal organisms expressed a higher number of HSP and witnessed a lower number of up-regulated lncRNAs, a characteristic response to thermal stress[55].

lncRNAs characterization

A comprehensive analysis of the features with all putative lncRNAs in the fungal genome was conducted to validate the effectiveness of the lncRNA detection method. It was compared their traits with those protein-coding transcripts (mRNA), including CG content, transcription length, expression level across different temperatures, number of exons and isoforms, and their localization within each fungal chromosome.

According to the GC content comparison between mRNAs and lncRNAs (Fig. 4A) across different temperatures, the analysis revealed that lncRNA GC% is higher than the overall genome GC% but lower than the mRNA GC%. Specifically, lncRNA GC% was around 55% while the organism overall genome GC% is 51.4491, and the mRNA GC% is around 60%. This comparison can provide insights into the stability of genomic regions the lncRNAs were localized in.

Moreover, the length of intragenic, intergenic, and antisense lncRNA transcript were compared to mRNA transcripts (Fig. 4B), showing that lncRNA transcripts were generally smaller than mRNAs and displayed the same median in all three lncRNA types. Furthermore, the lower extreme values of mRNA transcript lengths were much smaller than all lncRNA transcripts, even when compared to the upper extreme values of the same transcript class (mRNA), suggesting likely an inaccurate gene annotation.

Figure 4C displays the TPM expression levels of the fungal transcripts across different temperatures. The results suggest that the expression of these transcripts decreases gradually as the temperature increases from 35 to 50 ℃, indicating that the fungus was grown under thermal stress. Meanwhile, the lncRNA exon distribution followed the mRNA exon distribution, showing the majority of lncRNA transcripts are holding two exons and not exceeding four exons per transcripts (Fig. 4D).

Analysis of alternative splicing occurring in lncRNA and mRNA transcripts revealed that the majority of lncRNA loci had only one spliced isoform (Supplementary Fig. 5), which might indicate that alternative splicing is less prevalent in lncRNAs than in mRNAs. The intergenic transcripts `XLOC_000800`, `XLOC_005021`, and `XLOC_011814` exhibited the largest number of lncRNA isoforms, with three isoforms each (Supplementary Material—Github). Conversely, mRNA loci showed a higher degree of isoform diversity, with up to five isoform

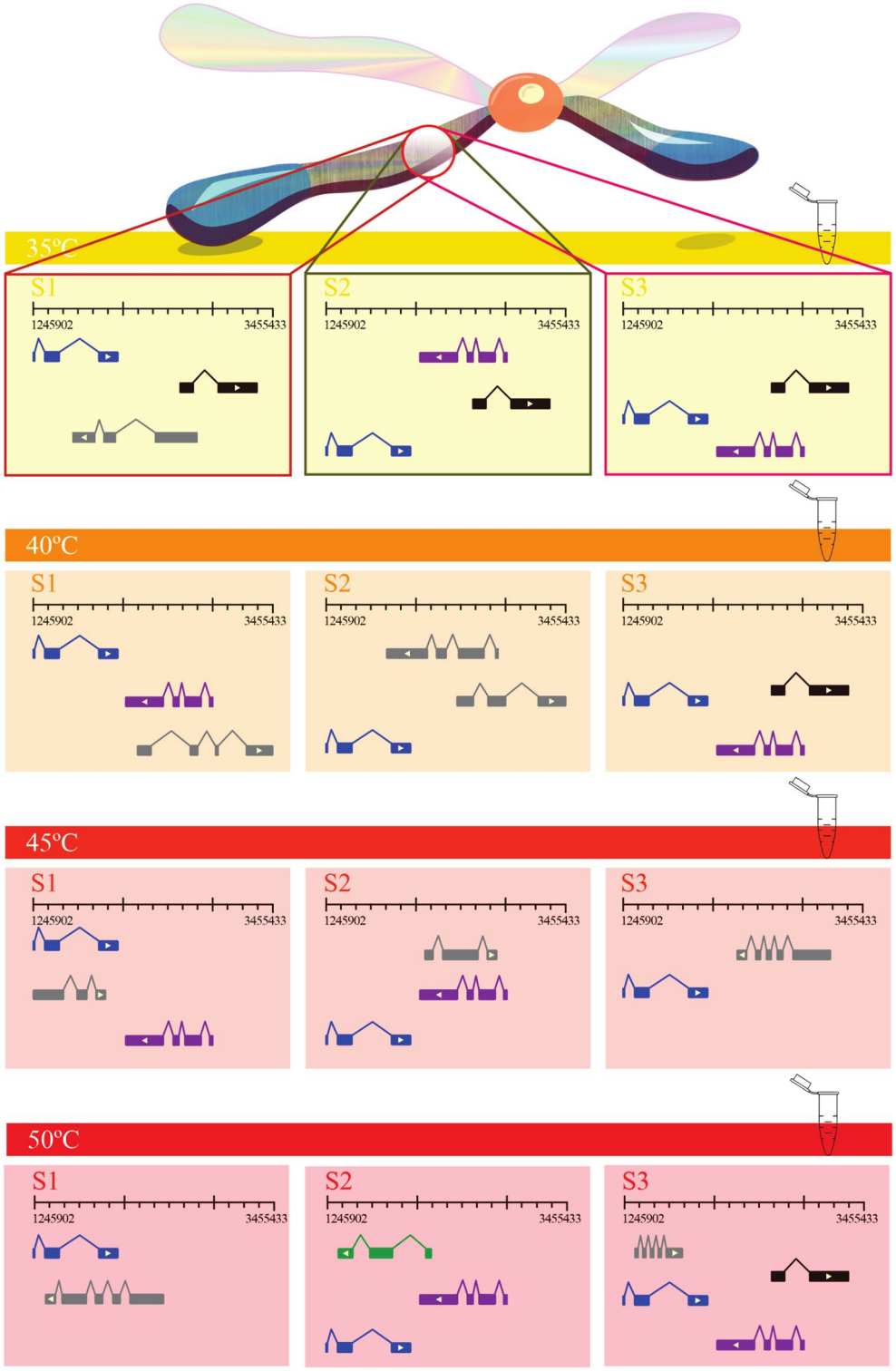

Figure 2. Comparison of transcripts among experiments at varied temperatures (35, 40, 45, and 50 ℃) revealing structural identical and non-identical transcripts. Colored transcripts demonstrated structurally identical transcripts found in each experiment.

variants detected, suggesting that alternative splicing could not be a common mechanism for generating lncRNA isoforms in this fungal species.

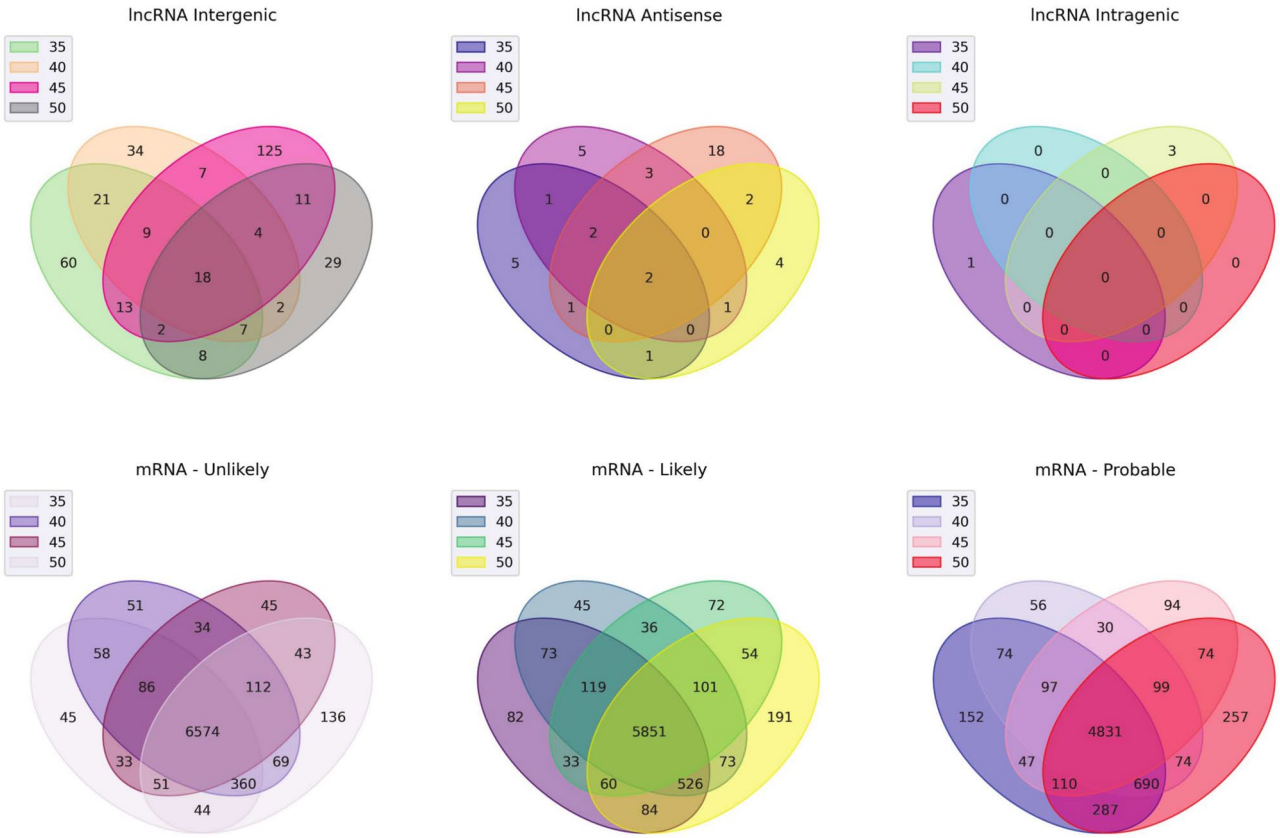

Figure 3. Venn diagrams showing the overlap of lncRNA and mRNA transcripts. The lncRNAs identified in intergenic, intronic, and antisense regions of experiments ranging from 35 to 50 ℃. Each diagram represents a specific type of transcript, with the number in the circle indicating the total number of transcripts identified. The overlap between the circles represents the number of transcripts that are present in multiple experiments.

This study found that the highest level of lncRNAs occurred in the 45 ℃ experiment for all types of lncRNAs with 189 intergenic, 28 antisense, and three intragenic transcripts detected (Supplementary Table 10). Venn diagrams were generated from lncRNAs expressed in all experiments to provide insights into the distribution and overlap of lncRNA transcripts across various temperature conditions (Fig. 3). This finding is consistent with previous studies[45] indicating that this fungus prefers to grow at higher temperatures (> 45 ℃). In contrast, mRNAs were the lowest level at 45 ℃, with 5382 detected transcripts (Supplementary Table 9), which is opposite to what was found for the lncRNAs. Nonetheless, at 50 ℃, mRNA transcripts were at the highest levels, with 6422 transcripts, while the quantity of lncRNA was the lowest, presenting only 91 transcripts (Fig. 3). These results raise the question of whether lncRNAs could be a strategic mechanism used by the fungus to modulate gene expression at milder temperatures. Besides, considering only energy consumption, why does the fungus rely on mRNAs to respond to thermal stress, given that mRNA undergoes translation, which is a more energy-consuming mechanism?

With respect to intergenic transcripts, there were 189 lncRNAs expressed at 45 ℃, which is 51 more transcripts expressed at 35 ℃. In comparison to the control (35 ℃), the hottest experiment (50 ℃) expressed 57 less intergenic lncRNAs, totalling only 81. Regarding the antisense transcripts, the temperature at which the greatest number of lncRNAs was observed was once again at 45 ℃, which was more than 2 times the expression of lncRNA in the control experiment. Furthermore, the experiment with the lowest number of antisense lncRNA transcripts was at 50 ℃, with only ten lncRNAs expressed. Interestingly, a similar pattern was observed for the intragenic lncRNAs, with three lncRNA expressed in the experiment at 45 ℃, one in the 35 ℃experiment, and none in the experiments at 40° and 50 ℃.

Finally, it is worth mentioning that 18 intergenic and two antisense lncRNAs, which were expressed in all four temperature experiments, can potentially serve as candidate housekeeping lncRNAs since they were expressed in all conditions, regardless of the temperature the fungus was cultivated.

Filtering out non-identical transcripts and differential expression analysis

This study performed a differential gene expression quantification analysis to determine whether the fungus was cultivated under thermal stress conditions when comparing the control and the experiment at the hottest temperature. Prior to describing the next differential gene expression quantification analyses using curated reads, it should be noted that the *prepDE.py3* script is not able to filter out **"Unlikely"** and **"Likely"** transcripts, and generates a gene count matrix only with structurally identical transcripts identified during the transcriptome

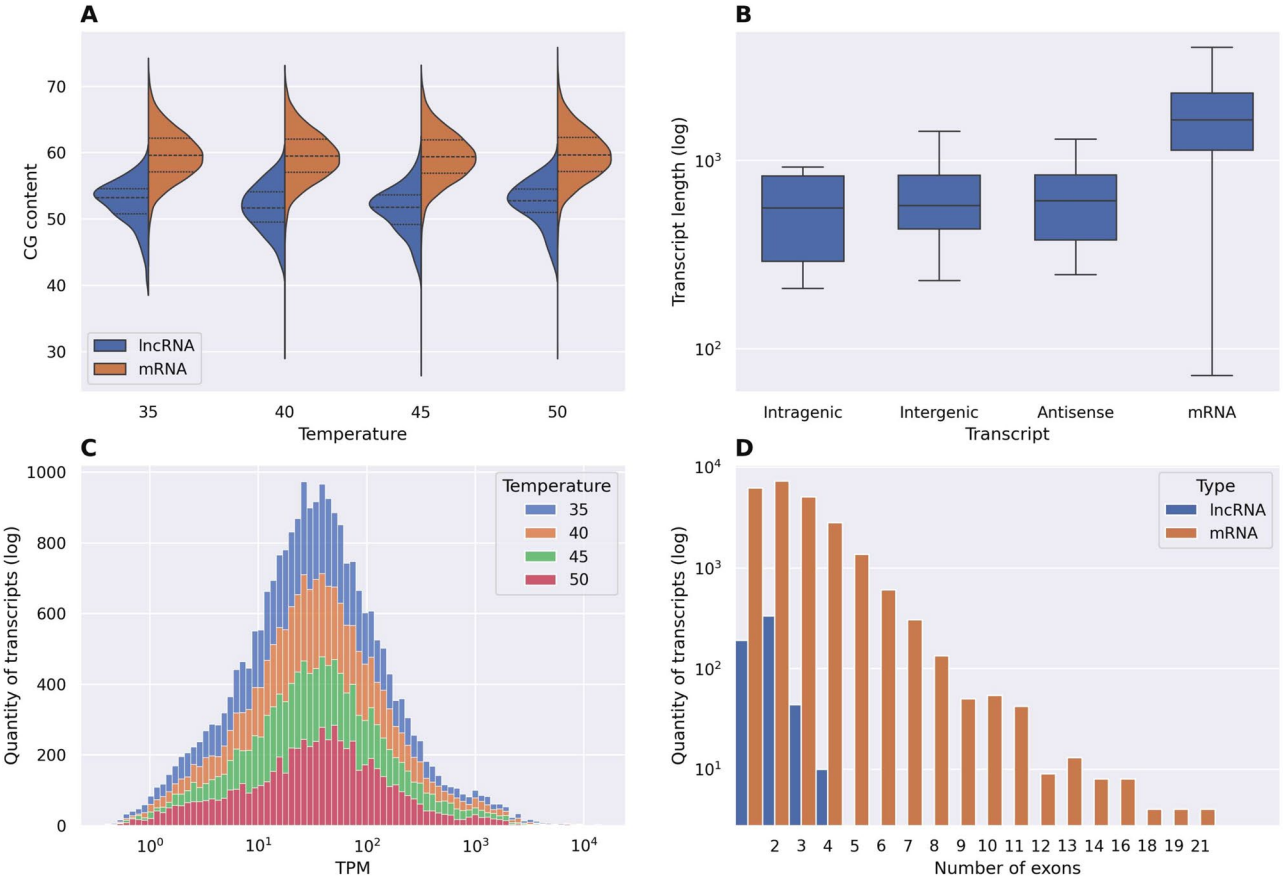

Figure 4. (**A**) mRNA and lncRNA CG-content comparison between temperatures. (**B**) Length comparison of different lncRNA classes and mRNA. (**C**) Quantity of transcripts across experiments. (**D**) Distribution of lncRNA exons.

assembly. Therefore, a Python script was used to filter out these transcripts as well as to retain gene isoforms with the highest read counts only (Supplementary Material—Github).

As previously discussed, the `TCONS_00003679` transcripts were expressed in three samples of the experiment cultivated at 35 ℃, but still present in two other samples labeled as q2 and q8, which respectively belong to the sample from 50 ℃ (q2) and 40 ℃ (q8). These two transcripts were not validated transcripts and, consequently, they should be excluded from the final DEG output results.

The resulting gene count matrix was then processed by DESeq2, and the results are displayed in Supplementary Fig. 6A-D. It should be noted that the heatmap grouped all three samples in an orderly manner according to each experiment temperature, particularly in the context of varying temperatures and emphasized the underlying patterns in gene expression data. Regarding the PC1 and PC2, they revealed less explainability (69%) compared to the prior PCA analysis (80%). This might be attributed to the decreased number of genes in the new gene count matrix and may have contributed to the reduction of variability explained by the PCA analyses.

Additionally, the DESeq2 results also demonstrated that the experimental investigation conducted at 45 ℃, when compared to the control group (35 ℃), yielded 56 differentially expressed lncRNAs, with 51 up-regulated and five down-regulated (Fig. 5). Surprisingly, differentially expressed lncRNAs at 50 ℃ exhibited an inverse profile contrasted to the previous experiment, with 10 down-regulated and nine up-regulated lncRNA genes. This inversion of lncRNA expression profile can be observed at the chromosome sets A and B (Fig. 6). Furthermore, even though the longest fungal chromosome is the chr1, it only harbored 10 differentially expressed lncRNAs at 45 ℃, while the chromosome 2, which is almost half its size, contained 19 differentially expressed lncRNAs, whereas seven lncRNAs were expressed in chromosomes 1 at 50 ℃. Interestingly, chromosome 1 harbors 10 HSP genes out of 21 HSP in the entire genome.

In general, our results suggest that temperature has a complex effect on lncRNA expression and, apparently, it is regulated by different chromosomes.

Functional enrichment analysis after reads curation

STRING enrichment analysis was conducted on the newly curated and validated set of 1046 differentially expressed genes from control (35 ℃) and treatment at 50 ℃. The STRING database recognized all PCG and reported the same three clusters related to protein refolding and chaperone binding clusters consisting of 17, 14, and 10 gene members with False Discovery Rate (FDR) < = 0.05 as in the previous analysis (Supplementary Table 11).

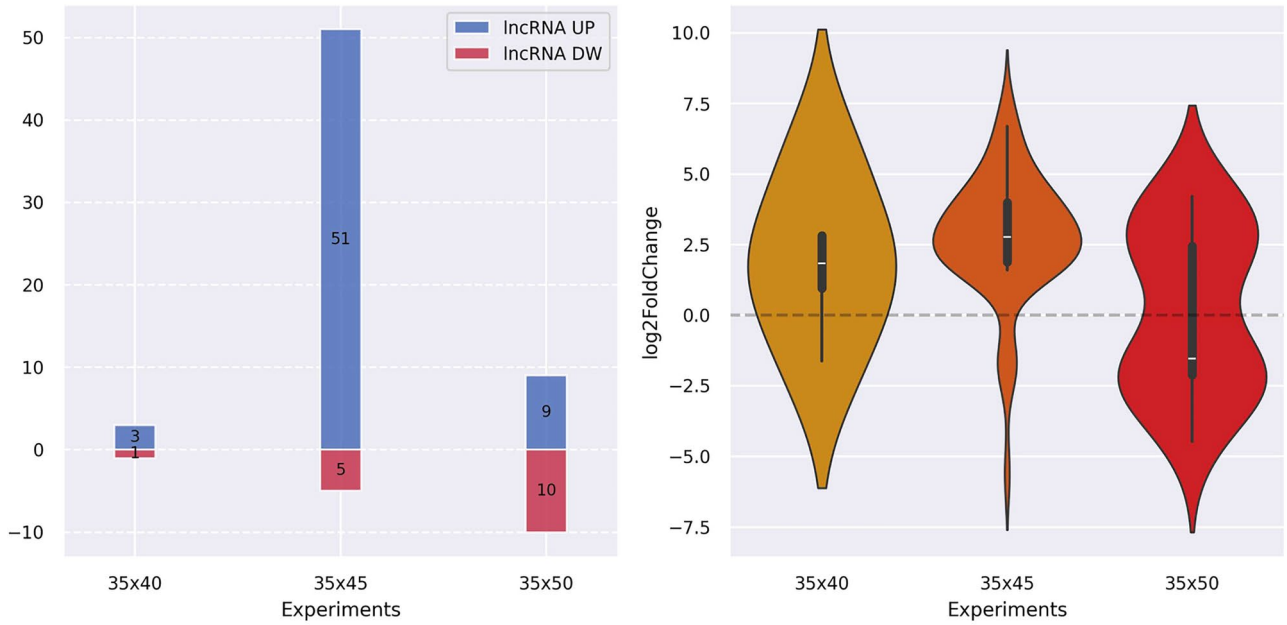

Figure 5. Distribution of differentially expressed lncRNAs and their log2 fold change between fungal experiments cultivated at 35 ℃ and exposed to thermal stress at 40, 45, and 50 ℃. The left panel shows stacked bar plots of the number of upregulated and downregulated lncRNAs for each experiment. The right panel shows violin plots of the LFC distribution for the differentially expressed lncRNAs.

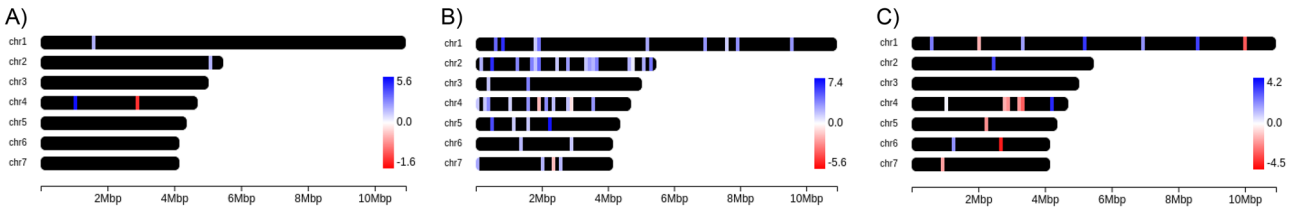

Figure 6. Distribution of differentially expressed lncRNAs within the 7 chromosomes of the fungus expressed in three different experiments (**A**) (35 × 40 ℃), (**B**) (35 × 45 ℃) and (**C**) (35 × 50 ℃). Highlighted genes represent lncRNA up-regulated (blue) or down-regulated (red)—(Anand & Lopez, 2022).

In addition, STRING also identified three novel GO Biological Process activity consisting of cellular carbohydrate metabolic process (36/149), carbohydrate catabolic process (38/162), and carbohydrate metabolic process (64/317) and FDR < 0.035. They are involved in energy production, structural maintenance, and overall metabolic adjustments necessary for fungal survival and adaptation.

Ultimately, our proposed pipeline yielded similar enriched pathways with slight reduction in the number of DEG before and after curation and validation. Moreover, it was able to identify novel and important enriched metabolic pathways that are essential for fungi.

LncRNA and temperature

Our study also evaluates lncRNA abundance and expression patterns across experiments and their contributions to the temperature response. Among the 500 most differentially expressed genes, 55 lncRNAs were distributed across the heatmap (Fig. 7). Six HSP genes, including two small heat shock proteins, one chaperonin Cpn60/GroEL, 2 ClpA/B family members, and one Hsp90 family member were up-regulated at high temperatures. Notably, the majority of lncRNA genes showed a down-regulation profile at high temperatures, with only a few displaying a similar expression pattern and grouped to the HSP genes.

Examining the cis-acting of lncRNAs with HSP, hierarchical clustering (Fig. 7) revealed an inversion of expression between these two transcripts. While HSP exhibited an up-regulated pattern at high temperatures, the expression of lncRNAs was down-regulated. To further explore the lncRNA and HSP transcript relationships in the fungus, Weighted Correlation Network Analysis showed that some lncRNAs share the same expression pattern with the cytochrome P450 family, an important stress-related gene family regulated in response to environmental stresses. Previous studies have shown that long noncoding RNAs (lncRNAs) and cytochrome P450 monooxygenases (P450s) are involved in the detoxification process[56]. Finally, lncRNAs exhibited a stronger correlation with CP450 proteins than with HSP (Fig. 9B).

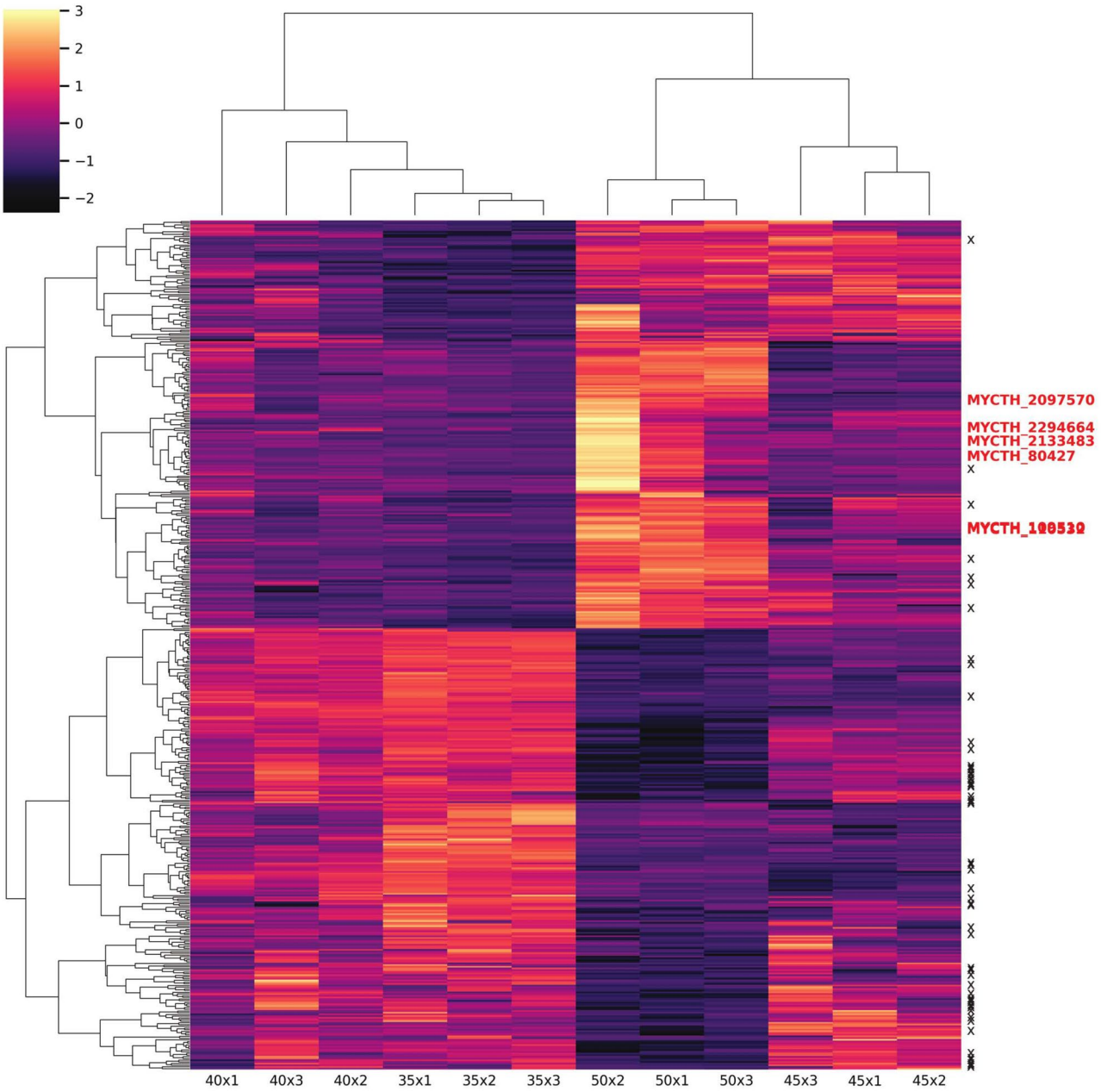

Figure 7. Cluster heatmap of the 500 top expressed genes across all samples. Z-score transformation was performed to each gene. The red labels represent HSP genes and the black asterisks (*) represent lncRNAs.

Weighted correlation network analysis

A WGCNA was constructed on a filtered and validated gene dataset consisting of 7261 genes after removing low count reads to study the potential roles of lncRNAs in thermal stress and their relationship with HSP. The analysis identified 10 (Fig. 8A) modules containing similar patterns of expression. Four modules (magenta, blue, brown and green colors) harbor around 80% of all the analyzed genes and the highest number of lncRNA respectively.

The remaining modules comprise a small number of clustered genes, including a limited number of lncRNAs. Ten HSP were clustered in the yellow module, with three HSP genes found in both blue and magenta modules, and one HSP each found in brown, green, and pink modules (Fig. 8B). A heatmap showing correlation between co-expression modules was plotted (Supplementary Fig. 7). Conversely, the yellow module, which contains the great majority of HSP genes, harbored 393 clustered genes and six lncRNA genes only. Interestingly, the modules that clustered the most quantities of PCG also contain a great amount of CP450 genes (six in magenta, eigh in blue, two in green and eight in brown). Oxidoreductase enzymes, such as the Cytochrome P450 monooxygenase, are involved in the fungal metabolism as well as response to stress in various organisms, including other fungi.

The oxidoreductase activity was investigated further due to its metabolic function in fungi. Computationally annotated gene sequences from the Oxidoreductase protein family (54 genes) were retrieved from the UniProt website, and used for a correlation analysis. Spearman's rank correlation coefficients were performed to investigate

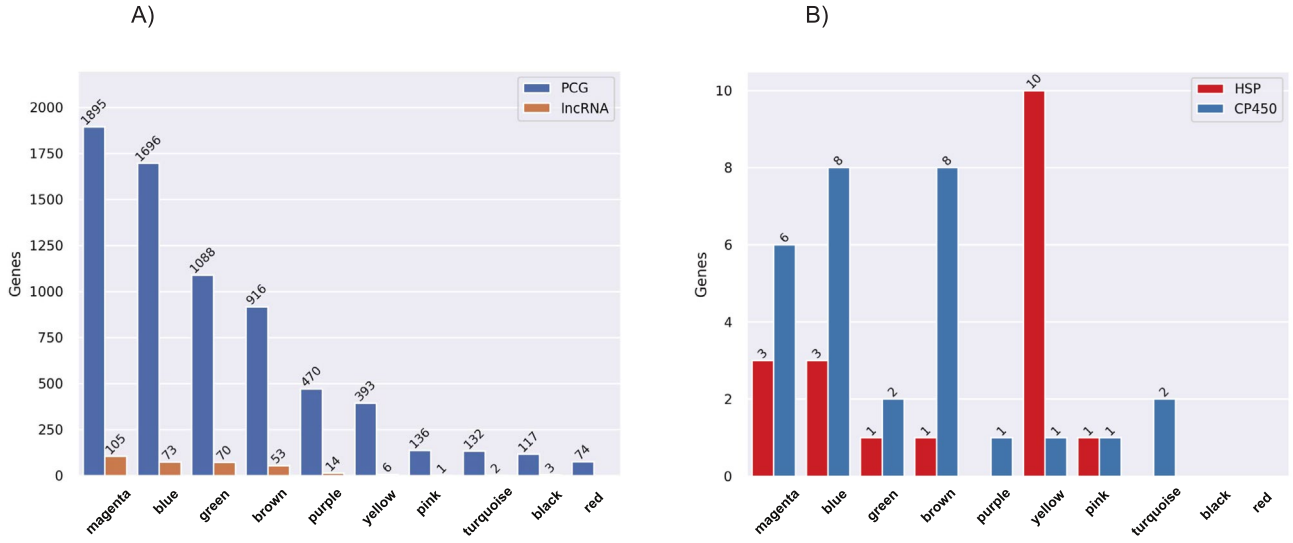

Figure 8. (**A**) The 10 modules identified by the WGCNA package. The blue bars represent mRNA and the orange bars represent lncRNAs. The bars were displayed in descending order according to the gene number in each cluster. (**B**) Bar plot showing the number of HSP (in red) and CP450 (in blue) genes in each module from the WGCNA analysis.

a possible regulatory involvement of lncRNAs with PCGs, HSPs, and CP450 in thermal stress response (Fig. 9A-B). The p-value was obtained to verify how likely the observed correlation is due to chance. The correlation between PCGs and lncRNAs (Fig. 9C) was strong and positive ($r = 0.9333$, $p = 0.00024$). Conversely, while the correlation between lncRNAs and HSPs was almost non-existent ($r = 0.26352$, $p = 0.6684$), the correlation between lncRNAs and CP450 was positive and stronger ($r = 0.69183$, $p = 0.0573$) than the correlation between lncRNA and HSP. This finding is an indication that lncRNAs may play a regulatory role in the expression of CP450 rather than HSPs under thermal stress conditions.

The modules Magenta, Blue, and Yellow were then selected for downstream analysis on STRING since those modules contain the major number of HSP (Fig. 8B). The Magenta Gene Ontology analysis on STRING showed the most significantly enriched GO terms were *Cellular process, Cellular metabolic process*, and *Cellular component biogenesis*, all of them having an FDR < 0.01. The Yellow Gene Ontology analysis demonstrated that those genes are involved in protein folding and refolding processes. Besides, the STRING analysis suggests that those genes may interact with chaperone proteins to aid in protein folding and refolding, corroborating with the KEGG pathway analysis, which indicates that those proteins could be involved in protein processing in the endoplasmic reticulum. Finally, the Blue Gene Ontology analysis suggests the majority of genes are involved in cellular metabolism, including nitrogen compound metabolic process and small molecule catabolic process, which are essential processes for fungi to survive and thrive in their environment.

Concluding remarks

Living organisms exhibit inherent variability. Therefore, replicated measurements are necessary[57] to enhance statistical power and assess the reproducibility of research findings affected by this inherent biological variability. Having multiple biological replicates increases the statistical robustness of the analysis and provides a more accurate estimation of variability within the samples, helping to distinguish experimental noise from technical artifacts. Consistent findings across replicates increase confidence in the reliability of the results, and this principle applies to RNA sequencing experiments as well[58]. Hence, after preparing the RNA experiments with biological replicates, accounting for biological variability, as well as mitigating technical variability introduced during RNA sample preparation, sequencing, and data analysis, it is important to note that there will still be a stochastic factor. Even under uniform conditions, individual cells may exhibit variability in gene expression levels. This stochasticity can lead to differences in gene expression profiles between biological replicates[59].

We have developed this pipeline to focus on reducing the variability in RNA-seq analysis by mitigating this stochastic effect. We achieved that by analyzing structurally identical transcripts as they represent the most commonly observed form of a gene in a particular condition through all replicates. This approach acts as a reference point for understanding the gene expression level and establishing its function. It provides a clearer picture of the organism's transcription activity when comparing treatment and control or biological replicates, increases accuracy of the analysis and strengthens confidence in the RNA-seq results. Therefore, structurally identical transcripts act as a control group to help distinguish true biological variation within the replicates from technical artifacts.

Furthermore, the pipeline does not ignore transcripts arising from alternative splicing events. Actually, if those transcripts were identified in the set of replicates, they would be classified as structurally identical and then they are reported by the pipeline. This allows researchers to identify which splicing events are significant and potentially influence gene function. Studying the expression and function of structurally identical transcripts

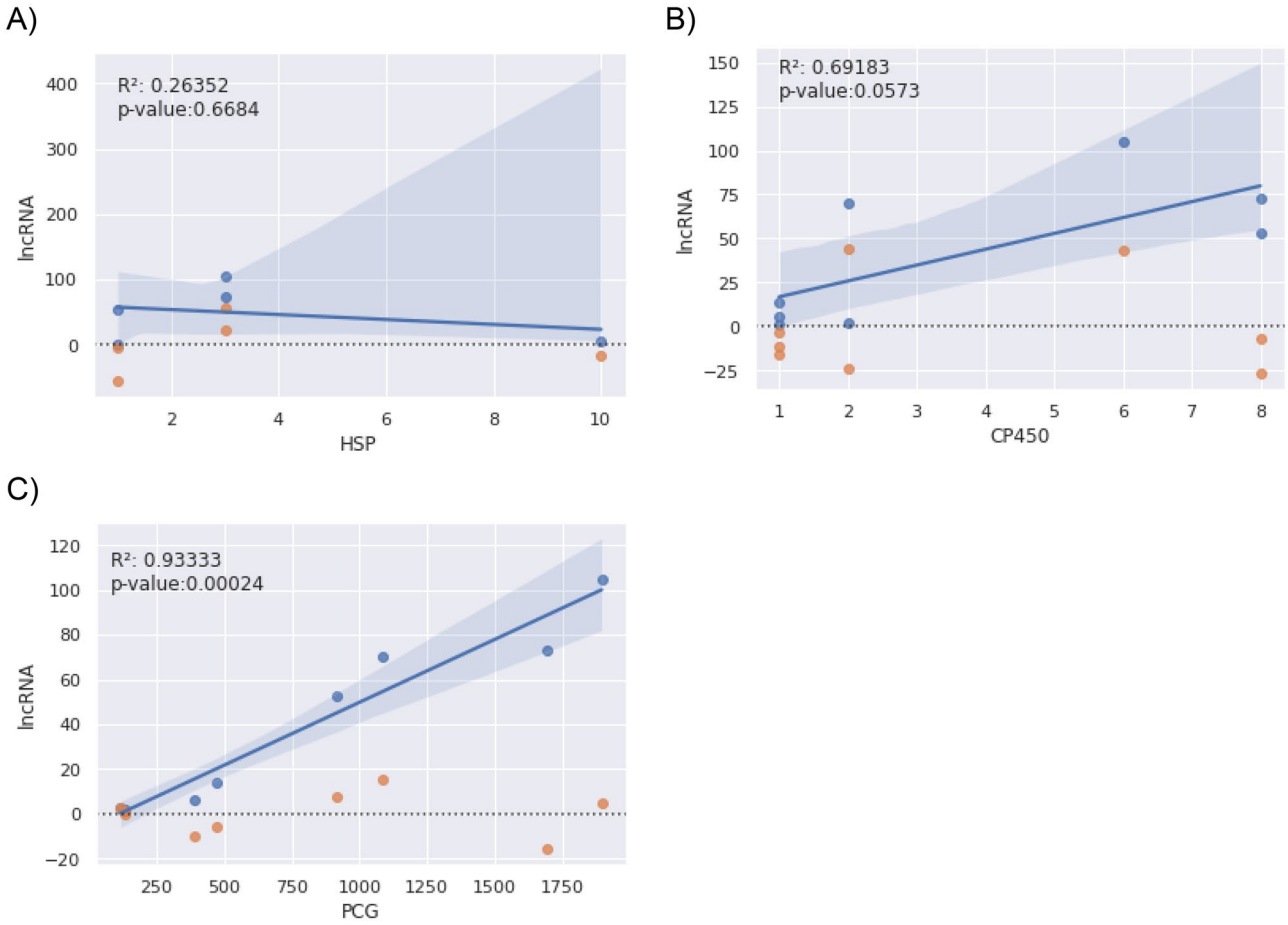

Figure 9. The graphs show Spearman's rank correlation and the relationship between lncRNAs, HSP (**A**) and CP450 (**B**), with the orange dots representing the residual errors. Data from the WGCNA analysis. (**C**) The graph represents Spearman's rank correlation. The blue line exhibited a linear relationship between PCG and lncRNA. The residual error is shown in orange.

arising from splicing events is crucial for understanding the roles of genes. This knowledge serves as the foundation for further exploration of alternative splicing mechanisms and their potential impact in the organism.

Adversely, it is absolutely possible that by comparing non-structurally identical transcripts, one might be comparing a canonical form of a gene with one its isoforms once these transcripts arise from the same gene but with variations in their structure due to alternative splicing events. This comparison could affect all downstream analysis. For example, if someone is comparing gene Transcripts Per Million (TPM) values from a triplicate experiment, it could be comparing two structurally identical transcripts with one non-identical transcripts (or isoform) and therefore violating statistical test assumptions. ANOVA test, for instance, requires homogeneity of variance or the variance among the groups should be approximately equal[60]. Non-structurally identical transcripts may violate this assumption if they exhibit different expression patterns or levels of variability between experimental conditions.

Moreover, DESeq2 assumes that the transcriptome data follows a negative binomial distribution and performs normalization to account for technical variations. However, the different distribution of a non-structurally identical transcript might not be fully normalized and potentially leading to an inflated fold change and false positive differently expressed (DE) result[58]. This analysis might struggle to distinguish the true difference from biological variability, potentially leading to miss DE genes or the analysis will recover only genes with the largest effect size.

To sum up, long non-coding RNA transcripts (lncRNAs) pose a level of complexity in RNA-seq analysis due to their lower abundance when compared to mRNA transcripts, their heterogeneity of expression across different cell types, tissues, and species, their low sequence conservation, their multitude of functions, and also their splice variants. For instance, the HOTAIRM1 lncRNA, which is localized in the HOX gene cluster, acts as a critical regulator of embryonic development and is known for its role in regulating axial patterning in vertebrates[61–64]. HOTAIRM1 has different splice variants each with different biological functions. Therefore, focusing on lncRNA transcripts with the same exon–intron structure allows for a more accurate assessment of true biological variability in gene expression levels across replicates. Furthermore, comparing identical transcripts minimizes technical and transcriptional noises, resulting in more reliable expression analysis of those transcripts. Finally, prioritizing structurally identical transcripts from RNA replicates with enough sequencing

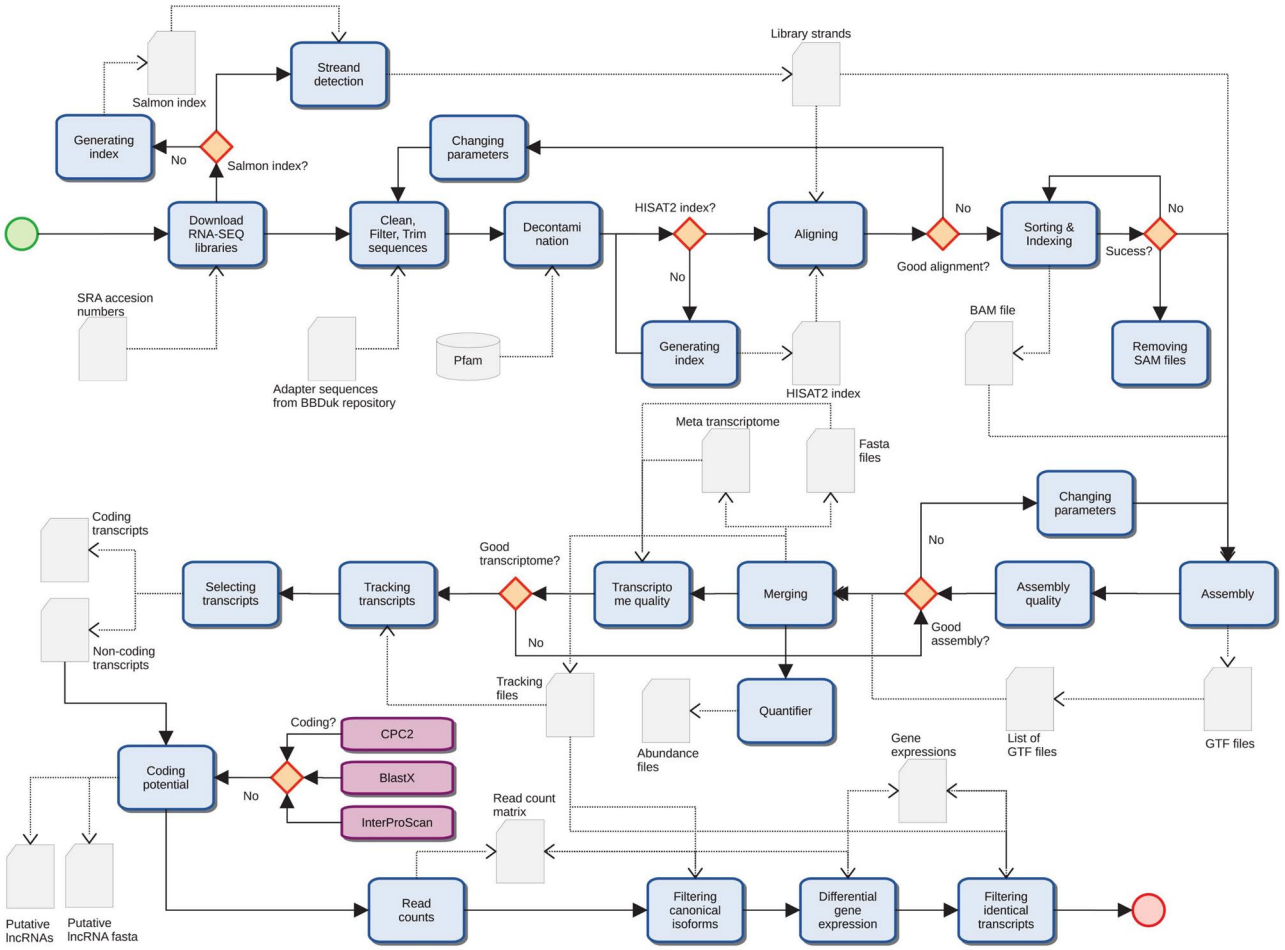

Figure 10. Detailed view of activities executed in the pipeline. Blue rectangles are activities and gray elements are input/output files.

depth enables accurate quantification of the overall expression of a gene in a specific condition or cell type, providing an additional and essential step in analyzing lncRNA transcripts.

Conclusions

In the rapidly evolving field of high-throughput genomics and transcriptomics sequencing[65], a challenge lies in distinguishing artifacts or biased elements from valuable biological transcripts within transcriptome data, particularly in the long non-coding RNA world where these molecules exhibit remarkable structural diversity with limited conservation across species, and play crucial roles in diverse biological processes[15]. In response to this challenge, we have developed a computational pipeline (Fig. 10) that employs a novel approach to track structurally identical lncRNA transcripts among different biological samples in fungi. This pipeline integrates multiple tools and algorithms to streamline the analysis of lncRNAs, identifies identical transcripts across different samples, ensuring a robust foundation for subsequent analyses and enables a more comprehensive understanding of their roles in fungal adaptation to extreme temperatures.

To sum up, our innovative pipeline offers a comprehensive framework for the study of lncRNAs in thermophilic fungi. Our findings provide valuable insights into the complex interplay between lncRNAs, temperature stress, and key genes involved in fungal thermal adaptation. Therefore, our study enhances the understanding of non-coding RNA biology in extreme environmental contexts and lays the groundwork for future investigations into the molecular mechanisms governing fungal responses to environmental challenges.

Data availability

The computational codes used in this study are available at GitHub (https://github.com/rogergsilva/structurally-identical-lncrnas/).

Received: 27 December 2023; Accepted: 18 July 2024
Published online: 27 August 2024

References

1. Clifton, J. M. *et al.* Psilocybin use patterns and perception of risk among a cohort of black individuals with opioid use disorder. *J. Psychedelic Stud.* **6**, 80–87 (2022).
2. Mendes-Pereira, T. *et al.* Disentangling the taxonomy, systematics, and life history of the spider-parasitic fungus Gibellula (Cordycipitaceae, Hypocreales). *J. Fungi* **9**, 457 (2023).
3. de Menezes, T. A. *et al.* Unraveling the secrets of a double-life fungus by genomics: *Ophiocordyceps australis* CCMB661 displays molecular machinery for both parasitic and endophytic lifestyles. *J. Fungi* **9**, 110 (2023).
4. Ke, H.-M. & Tsai, I. J. Understanding and using fungal bioluminescence—Recent progress and future perspectives. *Curr. Opin. Green Sustain. Chem.* **33**, 100570 (2022).
5. Maheshwari, R., Bharadwaj, G. & Bhat, M. K. Thermophilic fungi: Their physiology and enzymes. *Microbiol. Mol. Biol. Rev.* **64**, 461–488 (2000).
6. Patel, H. & Rawat, S. Thermophilic fungi: Diversity, physiology, genetics, and applications. In *New and Future Developments in Microbial Biotechnology and Bioengineering* (eds Patel, H. & Rawat, S.) 69–93 (Elsevier, 2021).
7. Tiwari, S., Thakur, R. & Shankar, J. Role of heat-shock proteins in cellular function and in the biology of fungi. *Biotechnol. Res. Int.* **2015**, 1–11 (2015).
8. Mohanta, T. K. & Bae, H. The diversity of fungal genome. *Biol. Proced. Online* https://doi.org/10.1186/s12575-015-0020-z (2015).
9. de Oliveira, T. B., Gostinčar, C., Gunde-Cimerman, N. & Rodrigues, A. Genome mining for peptidases in heat-tolerant and mesophilic fungi and putative adaptations for thermostability. *BMC Genom.* https://doi.org/10.1186/s12864-018-4549-5 (2018).
10. Thapar, R. Regulation of DNA double-strand break repair by non-coding RNAs. *Molecules* **23**, 2789 (2018).
11. Di, C. *et al.* Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. *Plant J.* **80**, 848–861 (2014).
12. Davati, N. & Ghorbani, A. Discovery of long non-coding RNAs in *Aspergillus flavus* response to water activity, CO2 concentration, and temperature changes. *Sci. Rep.* **13**, 1–13 (2023).
13. Pirogov, S. A., Gvozdev, V. A. & Klenov, M. S. Long noncoding RNAs and stress response in the nucleolus. *Cells* **8**, 668 (2019).
14. Tian, Y., Hou, Y. & Song, Y. LncRNAs elevate plant adaptation under low temperature by maintaining local chromatin landscape. *Plant Signal. Behav.* https://doi.org/10.1080/15592324.2021.2014677 (2022).
15. Mattick, J. S. *et al.* Long non-coding RNAs: Definitions, functions, challenges and recommendations. *Nat. Rev. Molecular Cell Biol.* **24**, 430–447 (2023).
16. Alberts, B. et al. From DNA to RNA. NCBI Bookshelf https://www.ncbi.nlm.nih.gov/books/NBK26887/ (2002).
17. Zhang, P., Wu, W., Chen, Q. & Chen, M. Non-Coding RNAs and their Integrated Networks. *J. Integr. Bioinform.* **16**, 20190027 (2019).
18. Samarfard, S. *et al.* Regulatory non-coding RNA: The core defense mechanism against plant pathogens. *J. Biotechnol.* **359**, 82–94 (2022).
19. Dou, J. *et al.* Genome-wide identification and functional prediction of long non-coding RNAs in Sprague-Dawley rats during heat stress. *BMC Genom.* https://doi.org/10.1186/s12864-021-07421-8 (2021).
20. Han, G. *et al.* Identification of long non-coding RNAs and the regulatory network responsive to *Arbuscular mycorrhizal* fungi colonization in maize roots. *Int. J. Mol. Sci.* **20**, 4491 (2019).
21. Harris, K. A. & Breaker, R. R. Large noncoding RNAs in bacteria. *Microbiol. Spectr.* https://doi.org/10.1128/microbiolspec.RWR-0005-2017 (2018).
22. Wang, Z., Zhao, Y. & Zhang, Y. Viral lncRNA: A regulatory molecule for controlling virus life cycle. *Non-coding RNA Res.* **2**, 38–44 (2017).
23. Hu, X. *et al.* Identification and characterization of heat-responsive lncRNAs in maize inbred line CM1. *BMC Genom.* https://doi.org/10.1186/s12864-022-08448-1 (2022).
24. Zhang, Y. *et al.* A long noncoding RNA HILinc1 enhances pear thermotolerance by stabilizing PbHILT1 transcripts through complementary base pairing. *Commun. Biol.* https://doi.org/10.1038/s42003-022-04010-7 (2022).
25. Zhang, Y. *et al.* Whole-transcriptome sequencing reveals that mRNA and ncRNA levels correlate with *Pleurotus cornucopiae* color formation. *Horticulturae* **10**, 60 (2024).
26. Li, R. *et al.* Pathogenicity-related long non-coding natural antisense transcripts in *Verticillium dahliae* during infections in cotton. *J. Phytopathol.* https://doi.org/10.1111/jph.13247 (2023).
27. Zang, F. *et al.* Responses of keratinocytes to *Trichophyton mentagrophyte* infection based on whole transcriptome analysis. *Mycoses* https://doi.org/10.1111/myc.13713 (2024).
28. Hovhannisyan, H. & Gabaldón, T. The long non-coding RNA landscape of Candida yeast pathogens. *Nat. Commun.* **12**, 1–13 (2021).
29. Riege, K. *et al.* Massive effect on LncRNAs in human monocytes during fungal and bacterial infections and in response to vitamins A and D. *Sci. Rep.* **7**, 40598 (2017).
30. Bruno, Mariolina *et al.* Comparative host transcriptome in response to pathogenic fungi identifies common and species-specific transcriptional antifungal host response pathways. *Computational Struct. Biotechnol. J.* **19**, 647–663 (2021).
31. Singh, A. *et al.* Global transcriptome characterization and assembly of the Thermophilic Ascomycete *Chaetomium thermophilum*. *Genes* **12**, 1549 (2021).
32. S., A. Babraham Bioinformatics. *FastQC A Quality Control tool for High Throughput Sequence Data* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
33. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
34. BBTools User Guide. *DOE Joint Genome Institute* https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/ (2016).
35. Kalvari, I. *et al.* Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucl. Acids Res.* **49**, D192–D200 (2020).
36. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
37. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
38. Thermophilic Fungi. https://mycocosm.jgi.doe.gov/Thermophilic_Fungi/Thermophilic_Fungi.info.html.
39. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* https://doi.org/10.1186/s13059-019-1910-1 (2019).
40. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).
41. Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V. & Prjibelski, A. D. rnaQUAST: A quality assessment tool for de novo transcriptome assemblies. *Bioinformatics* **32**, 2210–2212 (2016).
42. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* https://doi.org/10.1186/s13059-014-0550-8 (2014).

43. Szklarczyk, D. *et al.* The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucl. Acids Res.* **51**, D638–D646 (2022).
44. UniProt. https://www.uniprot.org/.
45. Johnson, M. *et al.* NCBI BLAST: A better web interface. *Nucl. Acids Res.* **36**, W5–W9 (2008).
46. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
47. Kang, Y.-J. *et al.* CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucl. Acids Res.* **45**, W12–W16 (2017).
48. Wang, L., Wang, J., Chen, H. & Hu, B. Genome-wide identification, characterization, and functional analysis of lncRNAs in *Hevea brasiliensis*. *Front. Plant Sci.* https://doi.org/10.3389/fpls.2022.1012576 (2022).
49. Hasan, S. *et al.* The long read transcriptome of rice (*Oryza sativa* ssp. *Japonica var*. Nipponbare) reveals novel transcripts. *Rice* https://doi.org/10.1186/s12284-022-00577-1 (2022).
50. Qian, J. *et al.* Long noncoding RNAs emerge from transposon-derived antisense sequences and may contribute to infection stage-specific transposon regulation in a fungal phytopathogen. *Mobile DNA* https://doi.org/10.1101/2023.06.13.544723 (2023).
51. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* https://doi.org/10.1186/1471-2105-9-559 (2008).
52. Liu, D. *et al.* Reconstruction and analysis of genome-scale metabolic model for thermophilic fungus *Myceliophthora thermophila*. *Biotechnol. Bioeng.* **119**, 1926–1937 (2022).
53. Barea, F. & Bonatto, D. Relationships among carbohydrate intermediate metabolites and DNA damage and repair in yeast from a systems biology perspective. *Mutat. Res. Fundam. Mol. Mech. Mutagen.* **642**, 43–56 (2008).
54. de Oliveira, T. B. & Rodrigues, A. Ecology of Thermophilic Fungi. *Springer International Publishing* https://link.springer.com/chapter/https://doi.org/10.1007/978-3-030-19030-9_3 (2019).
55. Tiwari, S., Thakur, R. & Shankar, J. Role of heat-shock proteins in cellular function and in the biology of fungi. *Biotechnol. Res. Int.* **2015**, 1–11 (2015).
56. Peng, T. *et al.* Functional investigation of lncRNAs and target cytochrome P450 genes related to spirotetramat resistance in Aphis gossypii Glover. *Pest Manag. Sci.* **78**, 1982–1991 (2022).
57. Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nature* https://doi.org/10.1038/nmeth.3091 (2014).
58. Schurch, N. J. *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use. *RNA (New York, N.Y.).* **22**, 839–51 (2016).
59. Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).
60. Kim, Y. J. & Cribbie, R. A. ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *Br. J. Math. Stat. Psychol.* **71**, 1–12 (2017).
61. Wang, X. Q. D. & Dostie, J. Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation. *Nucl. Acids Res.* **45**, 1091–1104 (2017).
62. Rea, J. *et al.* HOTAIRM1 regulates neuronal differentiation by modulating NEUROGENIN 2 and the downstream neurogenic cascade. *Cell Death Dis.* **11**, 1–15 (2020).
63. Zhang, X. *et al.* A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* **113**, 2526–2534 (2009).
64. Chen, Z.-H. *et al.* The lncRNA HOTAIRM1 regulates the degradation of PML-RARA oncoprotein and myeloid cell differentiation by enhancing the autophagy pathway. *Cell Death Differ.* **24**, 212–224 (2016).
65. D'Agostino, N., Li, W. & Wang, D. High-throughput transcriptomics. *Sci. Rep.* https://doi.org/10.1038/s41598-022-23985-1 (2022).

Acknowledgements
We would like to thank the Pro-Rectory of Research and the Graduation Program in Bioinformatics of UFMG.

Author contributions
R.S. developed the methods, performed computational analysis, analyzed the results, and designed and wrote the manuscript. G.F. and P.A and A.G-N helped in conceptualization, methodology, and data curation, and in its reviewing and editing. All authors contributed to the article and approved the submitted version.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-67975-x.

Correspondence and requests for materials should be addressed to A.G.-N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.