UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática

Roger Gomes da Silva

LONG NON-CODING RNA IN THERMOPHILIC FUNGI

Belo Horizonte
2024

Roger Gomes da Silva

LONG NON-CODING RNA IN THERMOPHILIC FUNGI

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Bioinformatics in the Biological Sciences Institute of Federal University of Minas Gerais.

Supervisor: Dr. Aristóteles Góes Neto
Co-supervisor: Dr. Glória Regina Franco

Belo Horizonte

Setembro de 2024

043 Silva, Roger Gomes da.
Long non-coding RNA in thermophilic fungi [manuscrito] / Roger Gomes da Silva. – 2024.
112 f. : il. ; 29,5 cm.

Orientador: Dr. Aristóteles Góes Neto. Co-orientadora: Dr. Glória Regina Franco.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Fungos. 3. RNA Longo não Codificante. 4. Expressão Gênica. I. Góes Neto, Aristóteles. II. Franco, Glória Regina. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS

ATA

INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Roger Gomes da Silva

Às quatorze horas do dia **30 de setembro de 2024**, reuniu-se, npor vídeoconferência através do aplicativo Zoom, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Long non-coding RNA in thermophilic fungi"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Aristóteles Góes Neto**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	Instituição	Indicação
Dr. Aristóteles Góes Neto	Universidade Federal de Minas Gerais	Aprovado
Dra. Glória Regina Franco	Universidade Federal de Minas Gerais	Aprovado
Dra. Cristiane Paula Gomes Calixto	Universidade de São Paulo	Aprovado
Dr. Bruno Silva Andrade	Universidade Estadual do Sudoeste da Bahia	Aprovado
Dra. Sara Cuadros Orellana	Universidad Católica del Maule, Chile	Aprovado
Dr. Rodrigo Juliani Siqueira Dalmolin	Universidade Federal do Rio Grande do Norte	Aprovado
Dr. Alessandro de Mello Varani	Universidade Estadual Paulista Júlio de Mesquita Filho	Aprovado

Pelas indicações, o candidato foi considerado: Aprovado

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 30 de setembro de 2024.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Professor do Magistério Superior**, em 01/10/2024, às 13:35, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiane Paula Gomes Calixto**, **Usuário Externo**, em 01/10/2024, às 13:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Silva Andrade**, **Usuário Externo**, em 01/10/2024, às 18:51, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gloria Regina Franco**, **Professora do Magistério Superior**, em 02/10/2024, às 16:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rodrigo Juliani Siqueira Dalmolin**, **Usuário Externo**, em 02/10/2024, às 23:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alessandro de Mello Varani**, **Usuário Externo**, em 08/10/2024, às 13:57, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3591914** e o código CRC **E258270E**.

Referência: Processo nº 23072.255795/2024-29 SEI nº 3591914

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Roger Gomes da Silva

"Long non-coding RNA in thermophilic fungi"

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Aristóteles Góes Neto - Orientador
Universidade Federal de Minas Gerais

Profa. Glória Regina Franco - Coorientadora
Universidade Federal de Minas Gerais

Profa. Cristiane Paula Gomes Calixto
Universidade de São Paulo

Prof. Bruno Silva Andrade
Universidade Estadual do Sudoeste da Bahia

Profa. Sara Cuadros Orellana
Universidad Católica del Maule, Chile

Prof. Rodrigo Juliani Siqueira Dalmolin
Universidade Federal do Rio Grande do Norte

Prof. Alessandro de Mello Varani
Universidade Estadual Paulista Júlio de Mesquita Filho

Belo Horizonte, 30 de setembro de 2024.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Professor do Magistério Superior**, em 01/10/2024, às 13:35, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiane Paula Gomes Calixto**, **Usuário Externo**, em 01/10/2024, às 13:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Silva Andrade**, **Usuário Externo**, em 01/10/2024, às 18:52, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gloria Regina Franco**, **Professora do Magistério Superior**, em 02/10/2024, às 16:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rodrigo Juliani Siqueira Dalmolin**, **Usuário Externo**, em 02/10/2024, às 23:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alessandro de Mello Varani**, **Usuário Externo**, em 08/10/2024, às 13:56, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3591974** e o código CRC **F7E5E238**.

Referência: Processo nº 23072.255795/2024-29 SEI nº 3591974

Acknowledgements

Before I express my gratitude to those who have occupied an important role in guiding me through the academic journey, I find myself compelled to delve into the profound beauty of life and the relentless rule of science in trying to organize the chaos that surrounds us all.

In an attempt to simplify the complexity of the organism I am about to discuss, though acknowledging its extraordinary intricacy which I have limited comprehension of, I recognize it as a framework within which we exist and our consciousness interacts with the external world. We, *Homo sapiens*, encompass an entire universe within itself. It is estimated that the human body consists of approximately 30 to 40 trillion cells, categorized into around 200 general cell types, such as muscles, nerves, and blood cells. Beyond these general cells, deeper within this human walking body universe, lie hundreds of distinct cell types, including immune system cells, hepatocytes, epithelial cells, neurons, and many more. Within these cells, but not all, a vast array of molecules coexists. The renowned DNA, with its structural variants A-DNA, B-DNA, Z-DNA, its close cousin RNA, which has been mentioned to have preceded the DNA molecule in evolutionary terms with a family of molecules including tRNA, rRNA, and mRNA, and other diverse molecule family such as proteins, lipids, metabolites, ions, and signaling molecules are integral components of this intricate system. The list, extensive and ever-expanding, continues to be shaped by researchers as they fit new pieces into this colossal jigsaw puzzle of life. I can’t provide an exhaustive list of molecules due to my limited knowledge and to spare you from potential boredom.

Why do I share these insights? Firstly, because I am passionate about genetics and how we can employ computational tools to piece together this giant puzzle. Secondly, and most importantly, contemplating the marvel of the human body and the meticulous regulation of its myriad molecules in performing their designated functions lead me to a profound belief: the notion that this intricate human body machine is not a masterpiece of chance. Yes, I do believe in God, and I am immensely grateful to Him for orchestrating the circumstances that allowed me to study a small fraction of His creation. The significance of a single amino acid changing in a specific location within the human genome, exemplified by conditions such as sickle cell anemia, emphasizes the delicate balance in His creation.

In writing these words, I have the privilege of studying His creation, an opportunity to explore a minuscule yet significant part of His perfection, constructing ideas and hypotheses that culminate in endeavors to comprehend a small parcel of the earliest knowledge embedded in one of His creations.

Next, I am grateful to my research advisors, Dr. Aristóteles Góes Neto and Dr. Glória Regina Franco, for their support, patience, and for generously sharing their comments, knowledge, and time. I deeply appreciate your mentorship and advice over the years.

Furthermore, I would like to express my eternal gratitude to my parents, Creusa and Francisco. Though you are not here to hold me in your arms and celebrate this moment together, my heart embraces you both. I also want to thank my dear sister Geisy and brother-in-law Rogério, along with my nieces Amanda and Mariana. Also, life has blessed me with another family, and I extend my thanks to Waldir, Sônia, Adriana, Júlio, João, Maria, Sabrina, Matheus, Pedro, Thales, and Rafael.

My special gratitude goes to Patrícia, Isabela and Artur from the deepest place of my heart. Your presence, support, and love in my life throughout these years have been immeasurable. I am forever indebted to each of you for all the crucial moments in your lives that I may have missed due to my studies, for all the sacrifices you've made, allowing me to embrace the opportunities that have led me to where I am today.

By extension, I want to express my appreciation to everyone in the Bioinformatics Program, including the Institute of Biological Sciences, Federal University of Minas Gerais and the Research Development Foundation - FUNDEP.

“Dreams without goals are just dreams. On the road to achieving your dreams, you must apply discipline but more importantly, consistency because without commitment you'll never start, but without consistency, you'll never finish.” - Denzel Washington

Resumo

Organismos que vivem em ambientes extremos, como aqueles encontrados em altas temperaturas, sempre despertaram nossa curiosidade na tentativa de compreender quais mudanças biológicas permitiram esses organismos prosperarem em condições ambientais extremas. A maioria dos organismos termofílicos pertence aos domínios Bacteria e Archaea. No entanto, há um pequeno e interessante grupo no domínio Eukarya, como os fungos termofílicos, capazes de se desenvolver em temperaturas entre 45°C e 60°C e que não conseguem crescer em temperaturas ordinárias (15°C–25°C).

Recentemente, os RNAs longos não codificantes (lncRNAs) emergiram como reguladores da expressão gênica, particularmente em resposta ao estresse ambiental. No entanto, a sua função biológica em organismos que habitam ambientes extremos, como os fungos termofílicos, permanece pouco compreendida. Neste trabalho, buscamos investigar a função e o significado evolutivo das lncRNAs em fungos termofílicos e mesofílicos, com foco em suas potenciais contribuições para sua adaptação térmica.

Foi desenvolvido um pipeline computacional para aprimorar a análise de dados de RNA-seq, identificando lncRNAs estruturalmente idênticos entre replicatas biológicas. Esse método reduz a variabilidade, ao mesmo tempo, melhorando a confiabilidade da expressão gênica desses transcritos, proporcionando análises mais precisas da atividade transcricional e minimizando a expressão gênica estocástica e ruído em decorrência de artefatos técnicos. Em fungos termofílicos, como *Thermothelomyces thermophilus*, essa abordagem permitiu observar padrões de expressão distintos entre lncRNAs e proteínas de choque térmico (HSPs), principais reguladores das respostas ao estresse celular. E ainda, transcritos estruturalmente idênticos demonstraram ser uma referência confiável para análises subsequentes, garantindo comparações estatísticas robustas, considerando ainda, os eventos de splicing alternativo.

Em um estudo comparativo entre o fungo termofílico *T. thermophilus* e o mesofílico *Chaetomium globosum*, identificamos diferenças significativas nos repertórios de lncRNAs desses fungos. O fungo termofílico exibiu uma quantidade 70% maior de lncRNAs intergênicos, quando comparados com o mesofílico, sugerindo um potencial papel na adaptação desses organismos. Curiosamente, uma das características dos organismos que prosperam em altas temperaturas é a redução do genoma, isso também se aplica aos fungos termofílicos. Porém, o

tamanho médio desses transcritos intergênicos permaneceu, praticamente, idêntico quando se compara o fungo termofílico e mesofílico, indicando uma possível conservação em sua estrutura. Além disso, foi identificado um aumento de, aproximadamente, três vezes no número de isoformas de lncRNAs no fungo termofílicos, particularmente em regiões intergênicas, o que sugere pode sugerir um mecanismo de adaptação ao estresse térmico.

Nossa análise também revelou uma discrepância nos padrões de expressão gênica entre fungos termofílicos e mesofílicos, sendo que os fungos termofílicos apresentaram mais transcritos com baixa abundância (TPM), sugerindo uma regulação gênica precisa de genes, provavelmente uma característica relacionada à temperatura. Essa observação destaca a importância regulatória dos lncRNAs, que, apesar de sua pouca abundância, desempenham papeis cruciais na regulação da expressão gênica sob condições extremas. Além disso, a presença de motifs conservados nessas sequências de lncRNAs termofílicos e mesofílicos, sugerindo um papel regulatório entre esses organismos, sendo que esses motifs provavelmente contribuem para a formação da estrutura secundária das moléculas de RNA e interações das mesmas com proteínas.

Também foi analisada a expressão das RNA polimerases (RNAP) I, II e III, que são críticas para a regulação transcricional de RNAs não codificantes. A expressão da RNAP III foi significativamente elevada em fungos termofílicos em todas as temperaturas que esses organismos foram cultivados, ressaltando a importância desses transcritos no perfil transcricional do fungo termofílico. Curiosamente, os fungos termofílicos apresentaram transcritos mais longos de RNAP I e III, sugerindo uma outra possível adaptação evolutiva relacionada com o aumento da eficiência e a especificidade transcricional sob condições de alta temperatura. Por fim, este trabalho fornece insights sobre os mecanismos moleculares e regulatórios que sustentam a adaptação térmica em fungos, com foco particular no papel dos lncRNAs. Além disso, revela a importância estrutural, funcional e evolutiva das lncRNAs e das RNA polimerases nos fungos termofílicos no processo de adaptação desses organismos aos ambientes extremos.

Palavras-chaves: bioinformática, fungos termofílicos e mesofílicos, análise de transcriptomas, RNA longos não codificantes, genômica funcional.

Abstract

Organisms living in harsh environments, such as those found in high temperatures, have always made our sense of curiosity tweak at trying to understand which internal biological changes make them thrive in extreme environmental conditions. Most thermophilic organisms are found in Bacteria and Archaea domains. Nonetheless, there is a small and interesting group in Eukarya, such as the thermophilic fungi, which are able to develop at temperatures between 45°C and 60°C and cannot grow at ordinary temperatures (15°C–25°C).

Long non-coding RNAs (lncRNAs) have recently emerged as critical regulators of gene expression, particularly in response to environmental stresses. However, their role in organisms inhabiting extreme environments, such as thermophilic fungi, remains poorly understood. This thesis investigates the function and evolutionary significance of lncRNAs in thermophilic and mesophilic fungi, with a focus on their potential contributions to thermal adaptation.

We developed a computational pipeline designed to enhance RNA-seq analysis by identifying structurally identical lncRNAs across replicates. This method reduces variability, improving the reliability of gene expression studies by providing more accurate assessments of transcriptional activity and minimizing the influence of stochastic gene expression and technical noise. In thermophilic fungi, such as *Thermothelomyces thermophilus*, this approach allowed us to observe distinct expression patterns between lncRNAs and Heat Shock Proteins (HSPs), key regulators of cellular stress responses. Importantly, structurally identical transcripts were shown to act as reliable reference points for downstream analysis, ensuring robust statistical comparisons while accounting for alternative splicing events.

In a comparative study between the thermophilic *T. thermophilus* and the mesophilic *Chaetomium globosum*, we uncovered significant differences in the lncRNA repertoires of these fungi. Thermophilic fungi exhibited a 70% higher quantity of intergenic lncRNAs, suggesting their potential role in temperature adaptation through fine-tuning of gene expression. Interestingly, despite genome reduction in thermophilic fungi, the median length of intergenic lncRNAs remained consistent between species, indicating conservation in lncRNA structure. Furthermore, a threefold increase in thermophilic lncRNA isoforms, particularly in intergenic regions, points to the potential role of alternative splicing as a mechanism for adapting to thermal stress.

Our analysis also revealed a discrepancy in gene expression patterns between thermophilic and mesophilic fungi, with thermophilic species showing a higher abundance of transcripts at lower TPM values, suggesting fine regulation of key genes involved in temperature adaptation. This observation highlights the regulatory importance of lncRNAs, which, despite their lower abundance, play crucial roles in the regulation of gene expression under extreme conditions. Additionally, the presence of conserved sequence motifs in both thermophilic and mesophilic lncRNAs further suggests a shared regulatory role, with these motifs likely contributing to RNA secondary structure formation and RNA-protein interactions.

We also analyzed the expression of RNA polymerases (RNAP) I, II, and III, which are critical for understanding the transcriptional regulation of non-coding RNAs. The expression of RNAP III was significantly elevated in thermophilic fungi across all tested temperatures, underscoring the importance of non-coding loci in the thermophilic transcriptional landscape. Interestingly, thermophilic fungi exhibited longer RNAP I and III transcripts, suggesting evolutionary adaptations that enhance transcriptional efficiency and specificity under high-temperature conditions.

Altogether, this thesis provides novel insights into the molecular mechanisms underlying thermal adaptation in fungi with a particular focus on the role of lncRNAs. Moreover, it reveals the structural, functional, and evolutionary significance of lncRNAs and RNA polymerases in thermophilic fungi, contributing to our understanding of how non-coding transcripts function in organismal adaptation to extreme environments.

Keywords: bioinformatics, mesophilic and thermophilic fungus, transcriptome analysis, long non-coding RNAs, functional genomics.

List of figures

Chapter 1

Chapter 3

Figure 7: Venn diagrams showing the variation of mRNAs according to their occurrence in each sample in experiments ranging from 35°C to 50°C. The overlap between the circles represents the number of transcripts that are present in multiple experiments.

Figure 8: (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.

Figure 9: The graph shows the isoform comparison between lncRNA and mRNA. The axis x represents the number of isoforms and the logarithm of transcripts on axis y.

Figure 10: After reads curation and all transcripts validated. A) Scatterplot of log fold changes vs. mean normalized counts generated by DESeq2 package for *Thermothelomyces thermophilus* cultivated under different temperatures, adjusted p-value < 0.05, with 7262 nonzero total read count, LFC > 0 (up): 1350, and LFC < 0 (down): 1324. B) Heatmap clustering of replicate samples from the 35, 40, 45 and 50°C experiments generated by the DESeq2 package. C) Principal component analysis showing characteristics of samples per experiment. Each dot indicates a sample cultivated at a certain temperature and samples grouped by color.

Figure 11: Distribution of differentially expressed lncRNAs and their log2 fold change between fungal experiments cultivated at 35°C and exposed to thermal stress at 40°C, 45°C, and 50°C. 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 12: Distribution of differentially expressed lncRNAs within the 7 chromosomes of the fungus expressed in three different experiments A(35 x 40°C), B(35 x 45°C) and C(35 x 50°C). Highlighted genes represent lncRNA up-regulated (blue) or down-regulated (red) - (Anand & Lopez, 2022).

Figure 13: Cluster heatmap of the 500 top expressed genes across all samples. Z-score transformation was performed for each gene. The red symbols represent HSP genes and the black stars (*) represent lncRNAs.

Figure 14: 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.

Figure 15: Bar plot showing the number of HSP (in red) and CP450 (in blue) genes in each module eigengene from the WGCNA analysis.

Figure 16: Cluster dendrogram from 7261 genes showing different colors below the dendrogram which indicates different co-expression modules. The heatmap shows a correlation between each experiment and module eigengenes from WGCNA analysis.

Figure 17: 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.

Figure 18: 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.

Chapter 4

Figure 1: FastQC results of the 12 RNA-seq libraries displaying the relationship between base position (x-axis) and Phred score (y-axis). The graph was generated using Multiqc.

Figure 2: Comparison of transcript abundance in mRNA and Intergenic, Antisense, and Intragenic lncRNAs across thermophilic and mesophilic fungi.

Figure 3: The bar graph shows a comparison between alternative splicing isoforms in thermophilic and mesophilic transcriptomes for mRNA and lncRNAs.

Figure 4: Distribution of CG content with outliers for mRNA and lncRNAs transcripts across thermophilic and mesophilic fungus.

Figure 5: The graph compares the length distribution of different RNA categories in thermophilic and mesophilic fungus.

Figure 6: The bar graph illustrates the transcript abundance measured in transcripts per million (TPM) values for different RNA categories in the mesophilic and thermophilic fungi.

Figure 7: TPM values for each mesophilic fungus scaffold and thermophilic fungus chromosomes. The graph A) shows TPM values for all mRNA transcripts, while the graph B) presents TPM values for all types of lncRNAs.

Figure 8: Boxplot comparing the distance in base pairs from a protein coding genes Transcription Stop Site (TTS) to an Intergenic lncRNA Transcription Start Site (TSS) and from an Intergenic lncRNA Transcription Termination Site (TTS) to a protein coding genes Transcription Start Site (TSS) across the mesophilic and thermophilic fungi.

Figure 9: The venn diagram showing the number of shared orthologous protein coding genes between the genomes of the fungi.

Figure 10: Ideogram representation of syntenic regions between the genomes of the fungi.

Figure 11: Motif Analysis of Mesophilic and Thermophilic lncRNA Sequences. A) Motifs found in Mesophilic lncRNA sequences, comprising 29 base pairs (bp), identified in 127 sites with an e-value of 1.7e-054. B) Motifs found in Thermophilic lncRNA sequences, also spanning 29 bp, discovered in 168 sites with an e-value of 4.9e-072.

List of tables

Chapter 1

Chapter 3

Chapter 4

List of Abbreviations

lncRNA - Long non-coding RNA
TSS - Transcription Start Site
TTS - Transcription Termination Site
HSP - Heat Shock Protein
WGCNA - Weighted Correlation Network Analysis
LUCA - Last Universal Common Ancestor
CG - Cytosine-Guanine
AT - Adenine-Thymine
RNA - Ribonucleic acid
DNA - Deoxyribonucleic Acid
tRNA - Transfer RNA
rRNA - ribosomal RNA
mRNA - Messenger RNA
HGT - Horizontal Gene Transfer
ncRNA - non-coding RNAs
ORF - Open Reading Frame
PCG - Protein-Coding Genes
cDNA - Complementary DNA
DEG - Differentially Expressed Genes
SRA - Sequence Read Archive

Contents

Chapter 1 - Review of thermophily

The origin of life on Earth is still a mystery! Many hypotheses have arisen, but science remains undecided. It has not been possible yet to ascertain with confidence how organic compounds combined to form living matter. However, looking back in time, considering the humongous evidence in history, chemical, and/or biological evolutionary processes have been present since the beginning of life on “Pale Blue Dot”. Despite the incredible variations of organisms on Earth, at the fundamental level, biological life is built upon those six essential ingredients: carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur. Those chemical elements serve as building blocks for assembling a more complex molecular system, which is known as a cell[1].

Prokaryotes, which were likely the first forms of cellular life on Earth, are unicellular organisms that lack a nucleus and other organelles[2]. They have developed strategies that allow them to survive in extreme environmental conditions such as hot springs, toxic wastes, salt lakes, and deep-sea hydrothermal vents. Looking at the most extreme conditions on Earth, bacteria have been documented thriving in permafrost soil and in sea ice[4] at temperatures of around -20°C. On the other extreme, the heat-tolerant prokaryote isolated from deep-sea black smokers grows at temperatures up to 113°C and does not survive at temperatures below 90°C[4]. Moreover, their thermal stability might help them survive atmospheric pressure much higher than the Earth’s atmosphere[2]. With such resistance to temperature and severe environmental conditions where only few organisms can live in, these organisms usually appear on the lower branches of any phylogenetic tree of life, after the Last Universal Common Ancestor (LUCA) has arisen. It has been hypothesized that such temperature resistance could have been inherited from the first living organisms on Earth[5].

Table 1 shows the temperature range into which these thermoresistant organisms can be classified into. This classification is sometimes controversial because organisms can start growing in one range and span to another, turning their identification accordingly to their optimal growth temperature, a challenging task.

Table 1: The table classifies organisms based on their growth temperature.

Temp/Scale	Cryophilic	Psychrophilic	Mesophilic	Thermophilic	Extremophilic
Range	-30° to -2°C	-1° to +10°C	+11° to +45° C	+46 to +75°C	> 75°C

Those organisms as many other living organisms have to perform metabolism and reproduction in order to keep them alive. Despite many other variables involved in these biological processes that have to be precisely well regulated, they also have to deal with the consequences of temperature on their cellular components. This fact focuses our attention on their RNA, DNA, proteins, and, specifically membrane molecules since biological material degrades rapidly at high temperatures[2].

1.1. Organism strategies

But, which adaptations have those organisms developed to couple with temperature? Evidence suggests that Bacteria and Archaea depend upon different strategies to maintain their molecular machineries functioning; therefore, we can categorize the modifications in two main groups. The first set of modifications accomplishes modifications on cytoplasm, membranes, cell surface complexes and even small molecules present in the cytoplasm. The second group, which is our main goal, embraces changes occurring in DNA, RNA, and proteins[2].

1.2. Overview of temperature adaptations in Archaea and Bacteria domains

The elevated Cytosine-Guanine (CG) nucleic acid content is supposingly one evolutionary modification that happened to heat-lovers organisms because of their higher thermal stability when compared to the Adenine-Thymine (AT) base pair. The thermal stability comes from the stronger stacking interaction between the nucleotides with the presence of triple hydrogen bonds between the bases[6]. In Bacteria, for instance, high GC content has been correlated with multiple factors, including tolerance to higher temperature[7]. Dinucleotide composition also has been cited as a possible factor of thermal stability, demonstrating correlation between dinucleotide composition and optimal growth temperature from transfer RNA (tRNA) and DNA of thermophilic bacteria and archaea[8]. In Archaea, a simple combination of

purine and pyrimidine dinucleotide composition is linearly correlated with the organisms' optimal growth temperature[9].

RNA thermal adaptation mechanisms, namely RNA thermometers, have been extensively studied in bacteria. Those RNA temperature sensors change their conformation when temperature is increased, blocking RNA binding sites in regulatory regions of mRNAs, turning on protein synthesis response to heat stress[10]. It is plausible to think that ribosomal RNA (rRNA), tRNA, and protein coding messenger RNA (mRNA) molecules would also be susceptible to the same temperature constraints as the organism's DNA. Hence, those non-coding RNA molecules show a significant correlation with temperature and their GC content, mostly in base-paired and extended loop regions as well as in specific regions of the double-stranded RNA[11]. Nevertheless, increasing GC content in tRNA molecules could lead to their misfolding and dysfunction, which ultimately can account for ineffective protein synthesis. Thus, the increase in GC content *per se* would not explain their thermal molecular stability.

On the other hand, if thermophilic organism's DNA had adopted the same strategy for its RNA molecules, considering an increase of its GC content, the DNA would have had a significant increase of amino acids encoded by Alanine, Arginine, Glycine and Proline, all GC-rich codons, while the amino acids with GC-poor codons such as Lysine, Isoleucine, Tyrosine, and Phenylalanine would be less frequent in thermophilic proteins, but such pattern has not been observed. Nonetheless, DNA dinucleotide composition has been hypothesized as a possible (but not a clearly diagnostic) adaptation of high temperature organisms. Alterations in dinucleotides would be preferable than changes in mononucleotides because dinucleotide composition alters double strand supercoiling[8], and, thus, alters the DNA rigidity. Other authors describe that synonymous codon usage in prokaryotes is different from mesophiles to thermophiles species, suggesting that thermophilic genomes are more frequently spotted the trinucleotides AGG, ATA and AGA rather than CGT and CGA[9,12]. Nonetheless, there has not been found any noticeable correlation between the GC content in prokaryotic genome composition and their optimal environmental growth temperature, and, hence, the variation appears to be the result of mutational biases[13,14].

Thermophiles have other mechanisms to keep their molecular integrity other than increasing their GC content. Thermostable proteins have been the most studied topic, perhaps because of their uncountable industrial applications. Investigations on protein structure amongst

thermophilic and mesophilic homologues have demonstrated some adaptation to higher temperatures. One adaptation is the disulfide bond. Disulfide bridge is a covalent bond between sulfur atoms and two cysteine amino acids. It was demonstrated that those bridges help protein stabilization in some thermophilic organisms and are present in higher numbers than in mesophiles[15]. Archaeal organisms also make use of disulfide bonds as protein stabilization but only in some hyperthermophilic species[15]. Another mechanism that has been previously studied is the amino acid **IVYWREL** occurrences, which vary with temperature range, and it is supposedly determinant of thermophilic adaptations[16]. Factors such as electrostatic interactions[17], hydrogen bonds[18], deletions in exposed loop regions[19], distinctive amino acid compositions[20,21] have been also suggested to be responsible for thermostability. Therefore, it is expected that protein structure of thermophiles has more stability and rigidity at high temperature than mesophilic proteins. Genome-wide studies also reported that thermophilic proteins usually tend to be shorter than their mesophilic homologs[22], and; therefore, contributing to the genome size reduction in thermophilic organisms[23,24].

Even though there have been quite interesting and even informative studies about thermotolerant mechanisms at molecular level, there is still a lack of knowledge on how those molecular changes affect the organism as a whole. Whatever the temperature does to the thermophilic organisms, DNA will probably suffer denaturation at high temperatures, and, consequently, it would lead to mutations[25] and double strand breaks[26,27]. Interestingly, Meyer (2021) and Speth et al., (2022) have shown that bacterial thermophilic DNA displays lower mutation rates when compared to mesophilic bacterial genome, leading us to think in three possible hypotheses: (i) bacterial thermophilic DNA repair system might be more efficient on repairing genome mutations than in mesophilic organisms, as suggested by studies on *Pyrococcus abyssi*[27], *Sulfolobus islandicus*[28], and *Persephonella*[29]; (ii) the microorganisms might have an internal mechanism to protect its DNA from temperature damage; or (iii) it seems that mutations tend to be more deleterious in thermophilic than in mesophilic organisms, and, consequently, the mutation rate might be reduced in thermophilic organisms. The latter hypothesis is supported by observations on the compactness of the thermophilic genome further discussed.

Genome reduction is another trait of thermophiles[23]. Their genomes are usually smaller than mesophiles and; therefore, a more compact genome may help the DNA repair mechanism to deal with mutations and indels effectively. This genome reduction would have influenced the

intergenic region contraction as well as collaborated to the protein length reduction hypothesis. Gene losses are also another genome reduction contributor. Mechanisms such as glycerol metabolism, urease complex, rhamnose pathway, genes from fructose complex (transport and utilization) and citric acid cycle have all been lost over evolutionary history of those organisms[30-32]. This reduction could be a consequence of the high cost of their functional maintenance, adaptation to a high environmental temperature, or even an elimination of genes that have low thermal stability. Curiously, genome reduction is a feature of endosymbiotic and pathogen-host dependent lifestyles[33].

Genes assembled into operons have been an adaptation to environmental changes observed in bacteria and archaea thriving in high temperatures. For instance, *T. maritima* (hyperthermophilic bacterium) has high density of operons in its genome and, once formed, these operons tend to have more stability than operons in mesophilic organisms[34]. Genes regulated by operons could be energetically more economic and respond quickly to stressors. Also, because of thermophilic genome reduction and operon adaptation, it is proposed that there might be a group of regulators (global regulators) that can be used to regulate different genes or gene clusters at once[32].

Heat shock proteins (HSP) are a highly conserved group of proteins that are activated in response to temperature stress and protect cells against thermal stress[32,34]. Gene upregulations belonging to the HSP group have been extensively observed at the transcriptional level in the three domains[15,35-37]. Moreover, proteomic analysis has revealed that those proteins can interact with other proteins forming complexes and play a protective role on the protein they interact with[15]. Chaperones have been observed in thermophilic organisms in response to temperature stress[38]. They can be triggered by HSF and these proteins give assistance to protein folding and refolding due to temperature stress. Wang et al. (2012), after detailed *T. maritima* proteome examination, concluded that chaperone proteins and a specific elongation factor, at above growth optimal temperature, were up-regulated and heavily abundant[39].

Bacteria and Archaea share a ubiquity adaptation mechanism. Horizontal Gene Transfer (HGT) permits DNA content being exchanged among organisms of different species. It is considered an important mechanism for bacterial genome evolution[40]. Indeed, it is proposed that thermophiles would not exist without HGT[40]. HGT is a widespread phenomenon in all domains of life, and, in spite of its importance and significance, it is generally assumed that closely related

organisms are engaged in genetic exchange more frequently than distantly related ones[40], meaning both thermophilic partners could have shared thermal traits.

1.3. Thermophilic in Eukaryota domain

Among prokaryotic organisms living at extreme temperatures, there are eukaryotes thriving at temperatures both below and above those typical for most life forms[41,42]. Thermophilic organisms are present in all domains, but only a small number of fungal species have been described as living in desert soils and decomposing plant residues at harsh temperatures[43]. Indubitably, they are not as extreme as bacteria and archaea thriving above 120°C[42], but they can reach growth temperatures of 40°C up to 60°C[44].

The thermophilic fungi is a small and interesting group in Ascomycota booming at temperatures that compromise cell membrane stability and other biological structures[45]. Phylogenetically closely related to mesophilic fungi (Figure 1), which usually grow best in moderate temperatures (20–30°C), these heat-lovers fungal species are on the spot of scientific and commercial interests due to their potential sources of thermostable enzymes[46]. They feed on where organic matter decomposes and cannot grow below 40°C, clearly showing they are biologically adapted for living at that temperature range[44].

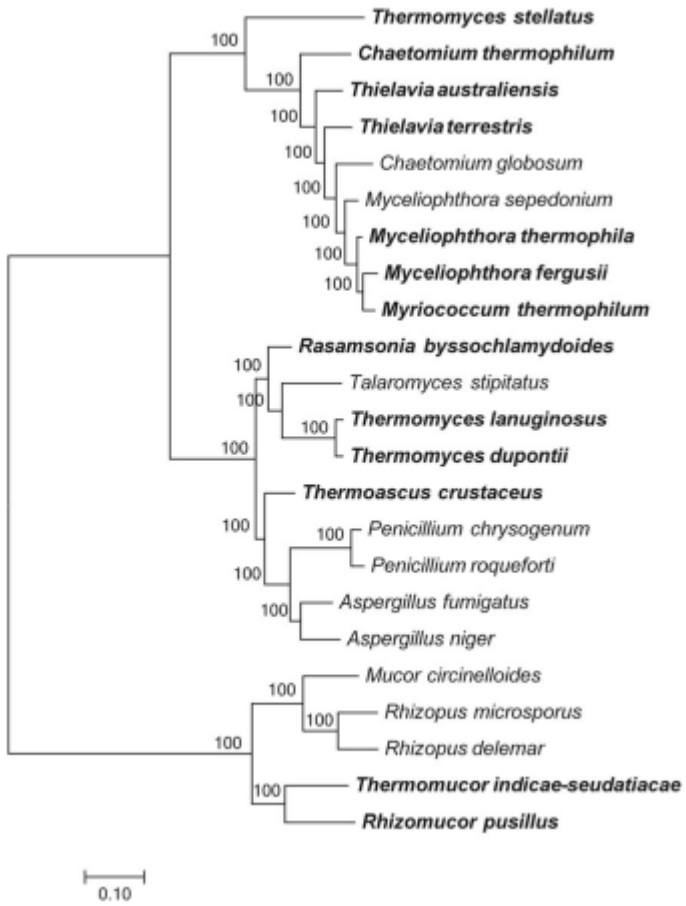


Figure 1: Phylogenetic tree implemented in PhyML 3.3 using Likelihood-Ratio Test of phylogenetically related mesophilic and thermophilic species (in bold), adapted from de Oliveira et. al.[63], based on fungal proteomes.

Genome size seems to be one of those adaptive processes that have happened to those organisms. Despite fungal genome sizes vary within its kingdom from 7.2 Mb of *M. restricta*[47] to 2,054Mb of *G. confusum*[48], likely due to many reduction and expansion genome events that happened over millions of years, it is noticeable a genome reduction in thermophilic genomes when they are compared to their closest mesophilic species[44,41]. Although the thermophilic fungi genomes were driven to a more compact form, their habitat is terrestrial and, consequently, they must deal with many variables under temperature constraints (nutrient concentrations, competing species, gasses, chemicals, water, etc.), demonstrating their ability to adapt to several factors and not only temperature stress[45].

With regarding to their genome GC content, comparative study between the thermophiles *M. thermophila* and *T. terrestris*[49] and the closely related mesophilic *C. globosum* has demonstrated that those thermophilic fungi have higher GC ratio in their coding sequence

regions and even more, when comparing the third nucleotide position (GC3) to the mesophilic orthologous sequences. Even though it is an interesting finding, there is no correlation with thermophilic prokaryotes[49]. Yet, the research group states that their analyses “were unable to find differences that can convincingly be interpreted as the molecular bases that underpin fungal thermophily”, neither in nucleotide nor in protein coding sequences, from those usually observed in thermophilic prokaryotes, not even in heat shock proteins, oxidative stress, membrane biosynthesis, chromatin structure and modification, nor fungal cell wall metabolism[49].

One likely thermophilic fungi adaptation occurs on their membrane composition. Those organisms produce lipids with higher saturated fatty acids and lower levels of unsaturated fatty acids since unsaturated fatty acids tend to be liquid at room temperature[41]. This change in membrane composition might give a glimpse of why thermophilic fungi are not able to grow at temperatures below 20°C. Nevertheless, this phenomenon does not occur in any thermophilic fungi; some thermal fungi are able to grow at mesophilic temperatures, showing that this membrane adaptation is not the main reason for these fungi to grow at a minimum temperature[41].

Moreover, genes interrupted by short introns, reduction of intergenic regions, loss of protein-coding genes, repetitive sequences, transposable elements, diversity of gene structure are characteristics of all the fungal genomes[49]. Those notable features have demonstrated a challenge for fungal studies and even more for thermophilic fungi investigations. Thus, what would be the likely modifications those thermophilic fungi have developed to survive at high temperatures? Recently, it has been assigned to long noncoding RNAs (lncRNAs) important functions in the cellular response to stressful environmental conditions such as regulation of double-strand DNA breaks[50], differentially expressed lncRNAs under various stress stimuli[51], regulating gene expression in response to stress conditions in the nucleolus[52] and other functions[53-55]. All those functions have been assigned to support stress environmental responses at the molecular and genomic levels within the cells.

1.4. What are Long non-coding RNAs?

For decades it was previously thought that the non-protein coding fraction of the genome was non-functional and accordingly labeled ‘junk’ DNA. This was due to the observations that only about 1.2% of the human genome were translated into proteins. In 2012, the ENCODE Project Consortium sought to uncover the purpose of this ‘junk’ and; therefore, characterize all functional elements. Strikingly, they classified different RNA types that covered 62% of the entire human genome. Most of the RNA belong outside (intronic and intergenic elements) of the protein-coding regions (ENCODE, 2012).

The singularity of RNA molecules encompasses not only its physical structure, but also its abundance, functional diversity and uniqueness. Although it is possible to find RNA molecules in a double-stranded fashioned form, the great majority is uni-stranded, manufactured in the cell nucleus and migrated into other cell parts[56]. 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 can be splitted in two major groups: coding and non-coding RNAs (Figure 2). Coding RNAs, usually referred as mRNAs, encodes a protein that functions as enzymes, signal transductions, antibodies, oxygen transporters, etc[56]. Aside from the coding RNA, non-coding RNAs (ncRNAs) are an extensive group classified into housekeeping and regulatory transcripts. The former contains the rRNA and the tRNA that are involved in reading and linking amino acids, protein translation and transport whereas the latter can be further classified based on transcript length as small (< 200 nt) and long non-coding RNA (> 200 nt) with other subsets within each group[57].

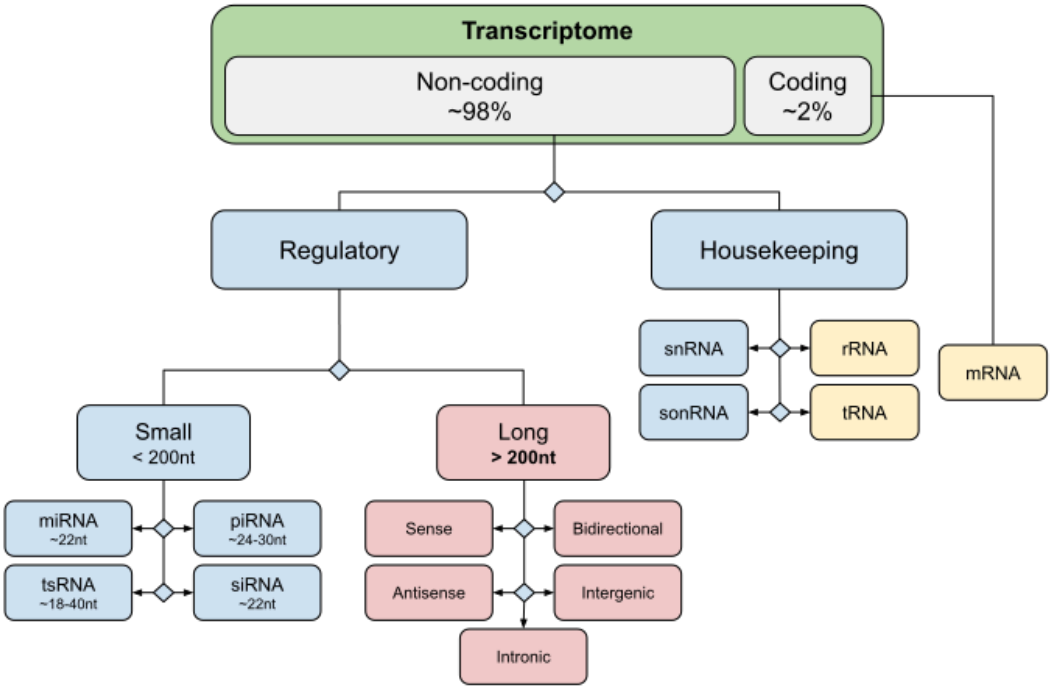


Figure 2: Overall taxonomy of non-coding RNA grouped into subclasses - RNA molecules can be grouped into two main groups: coding and non-coding. The coding RNAs usually referred to mRNA are molecules that encode proteins. Apart from those ones, non-coding RNAs are split into two main groups: regulatory and housekeeping RNAs, which can be split further into other subgroups[57].

Evidenced by the pervasively transcribed and abundant in plants, animals, fungi, prokaryotes and even in viral genomes, some long non-coding RNAs (lncRNA) are alternatively spliced transcripts longer than 200 nucleotides, do not encode proteins, are functional in their primary, secondary, or tertiary structures, display low conservation and expression when compared to protein-coding genes, and mostly transcribed by RNA polymerase II. Once induced and folded, lncRNAs interact with DNA, RNAs, and proteins, and can regulate gene expression using different biological mechanisms. Its characterization is at the initial stage and their evolutionary origins remain obscure, but some have been assessed by little or none open reading frame (ORF), some are 5-capping, poly or not adenylated, and undergo post-transcriptional modifications. Furthermore, they can also be classified according to their position in the genome (Figure 3), such as sense, antisense, intergenic, intronic, and bidirectional[58].

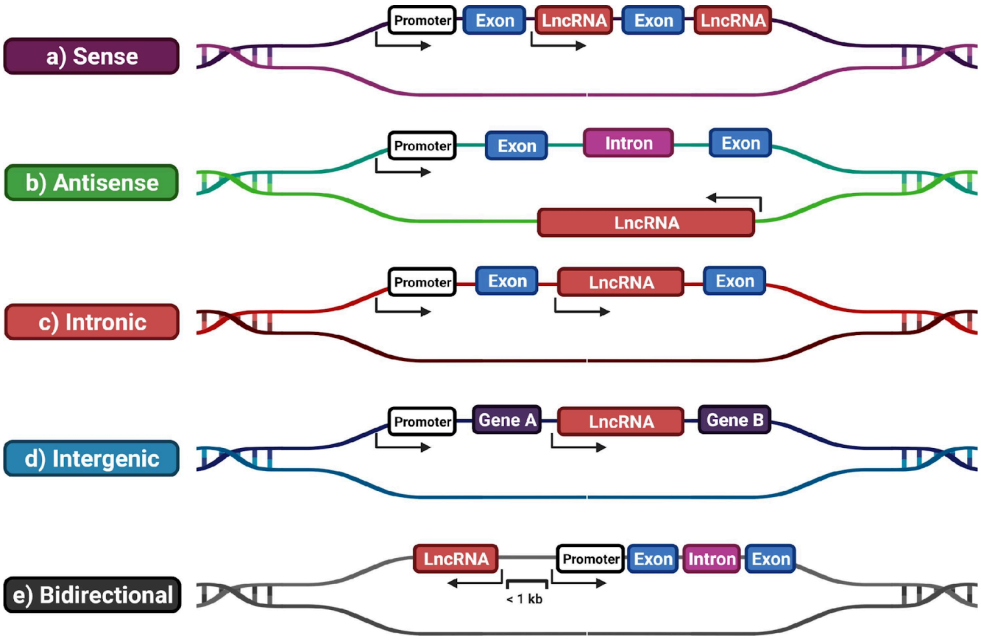


Figure 3: LncRNAs and their subclasses and genomic structure - Transcripts classified as lncRNAs have diverse primary sequence structure and orientation within the genome. They come in the form of: a) sense lncRNA transcripts overlap with the same strand of genes that are translated into protein; b) antisense lncRNA transcripts overlap with an opposite strand of the gene that is translated into protein; c) intronic lncRNAs are generated from between exon sequences of protein-coding genes; d) intergenic lncRNAs are generated between two genes that are translated into protein; and e) bidirectional lncRNAs share the same promoter with protein-coding genes, but are expressed from the opposite strand of it, and the starting point of the transcription of these lncRNAs cannot be more than 1000 base pairs from the opposite strand of genes that are able to translate into proteins.

Despite having defined characteristics, lncRNAs are poorly annotated. Moreover, the subclasses (or biotypes as they are labeled in Ensembl) are not uniformly annotated within genomic databases and misannotation compounds further increase the problem. When compared to a protein-coding gene investigation where a wealth of curated annotations and databases are publicly available, any lncRNA project leaves researchers directionless and achieving their accomplishments are even more challenging due to the absence of those resources.

1.5. LncRNA as likely stress regulators

Temperature is one of the most influential traits and influences almost everything within an organism, spanning from biological functions to cellular integrity to biomolecule structures. What makes thermophilic fungi organisms so fascinating is that as the temperature increases, there is a direct impact on disrupting hydrogen bonds and hydrophobic interactions and,

consequently, resulting in the denaturation of proteins and nucleic acids. Recently, lncRNAs have been recognized as major gene expression regulators and involved with distinct organism heat-stress responses[52]; however, their mechanisms of action remain largely unknown.

Thermophilic fungi have developed adaptations that grant them the ability to grow under severe high temperatures conditions[44]. Thermophily in fungi could be related to protein thermostability, protein compactness and hydrophobicity, sequence-based mutations with increased number of charged residues, rapidly synthesis of Heat Shock Protein (HSP), amino acid and nucleotide composition, genome size reduction, thermotolerance genes, rapidly turnover of essential metabolites and others[41].

Nowadays, only one study has explored the roles of lncRNAs in thermophilic fungi and their functions under thermal stress conditions[62]. Nevertheless, there are studies of lncRNA response to heating stress in plants[59,60,64], and scarce studies in mammals[60] and mesophilic fungi[61]. For instance, Earth's temperature has increased lately and trees worldwide have been exposed to heat-stress conditions. Researchers using a computational approach demonstrated the existence of a putative regulatory interaction between miRNA-lncRNA–mRNA in the tree's response to heat stress, showing that miRNAs may negatively regulate lncRNAs activity on mRNA involved in the heat shock protein synthesis[59].

Moreover, an integrative data approach of *Arabidopsis thaliana* lncRNAs under different stress conditions using expression patterns, epigenetic signatures, and sequence and structural features, demonstrated that lncRNAs might present differentially expression patterns under stress temperatures when compared with protein coding gene expressions[64]. The same research group also found that those differentially expressed lncRNAs have enriched sequence and structural motifs that would help RNA-binding proteins to bind to and cause lncRNA structural rearrangement in response to stress conditions[64].

Considering the heat stress affects not only the organism's genome and lncRNAs play important roles in gene expression at multiple levels under temperature stress, Hu et al. (2022) predicted target genes from differentially expressed lncRNAs between control and heat stress treatments in maize. They have found that those genes are enriched into important biological processes and pathways, such as: photosynthesis, metabolism, translation, stress response,

hormone signal transduction, and spliceosome, showing that their lncRNA have important functions in heat response.

Although investigations have accumulated evidence of lncRNA functions in plants under thermal stress, there are not many thermal stress studies in Ascomycota fungi, and even less for lncRNA roles in thermophilic fungal species. *C. thermophilum*, a thermophilic filamentous ascomycete, which is able to grow at 50–52 °C, had its transcriptome characterized[62] by deep RNA sequencing. This study identified 4567 non-coding new genes, but lncRNA downstream functional analysis targeting those molecules was not performed. Yet, the understanding of the fungal response to temperature stress and the involvement of lncRNAs in this context are still limited.

Chapter 2 - Project hypothesis and aims

Detailed molecular phylogenetic study (Figure 1) has identified close relatives of thermophilic and mesophilic fungi species. This result suggests that the thermophilic *Myceliophthora thermophila* and the mesophilic *Chaetomium globosum* are good models to investigate the consequences of temperature on fungal genomes[63].

Nowadays, only a few studies have reported the role of lncRNAs in fungi and almost nothing about their functions specifically in thermophilic fungi. Thermophilic fungi thriving at high temperatures might induce distinct lncRNA responses that contribute to different cellular regulation processes against stressful heat conditions. Hence, researching lncRNA functions related to environmental temperature stress will improve our understanding of how high-temperature impacts genomes, their regulation roles, and contributions to thermophilic genomes.

My project hypothesis is that lncRNAs are key thermal regulators of thermophilic cellular functions whose roles have not been elucidated in those organisms. I hypothesize that mesophilic and thermophilic fungi exhibit distinct expression profiles of lncRNAs that are correlated with their adaptation to temperature, suggesting a regulatory role for these molecules in temperature tolerance. In my thesis, I will focus on polyA+ lncRNA identification, annotation, and functional prediction in thermophilic fungi. In order to test my hypothesis and elucidate my

research questions, I will investigate the role and biological significance of thermal-associated lncRNAs throughout:

Aim 1 - The development of a computational pipeline for identifying lncRNAs across replicates from publicly available bulk RNA-seq datasets;

Aim 2 - Identification and computational characterization of polyA+ lncRNAs in phylogenetically related thermophilic and mesophilic fungi;

Aim 3 - Conducting comparative genomics studies to understand the patterns of lncRNA sequence, structural prediction and motifs across the fungal lineages to gain insights into their evolutionary history and functional conservation.

Chapter 3 - Exploring the hidden hot world of long non-coding RNAs in thermophilic fungus using a robust computational pipeline

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

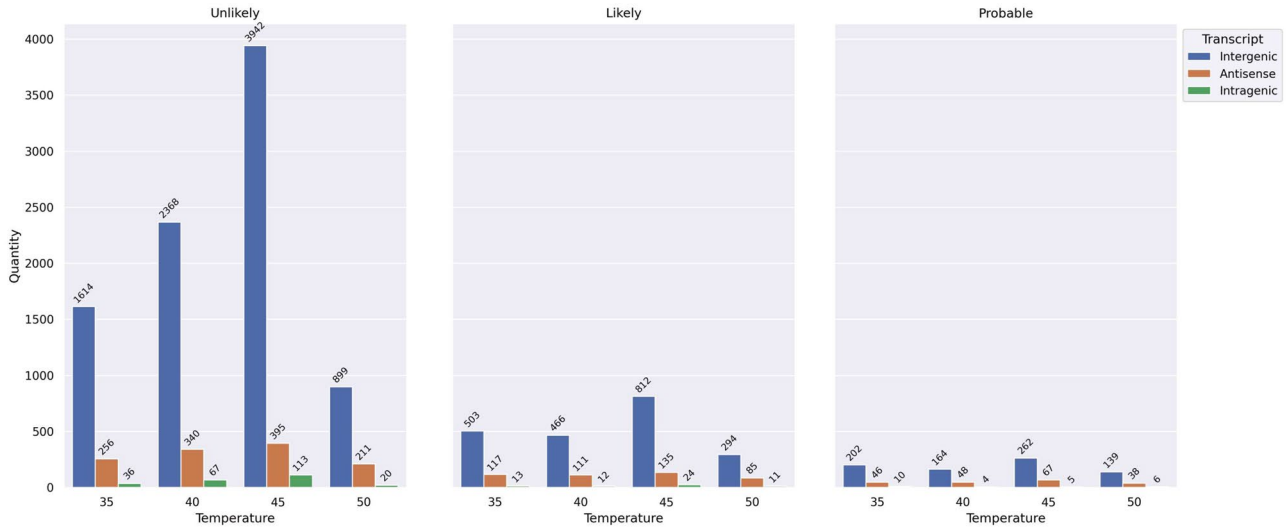


Figure 1. Number of lncRNA transcripts identified in one sample (Unlikely), two samples (Likely), or three samples (Probable) for experiments cultivated in different temperatures without assessing their coding potential.

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

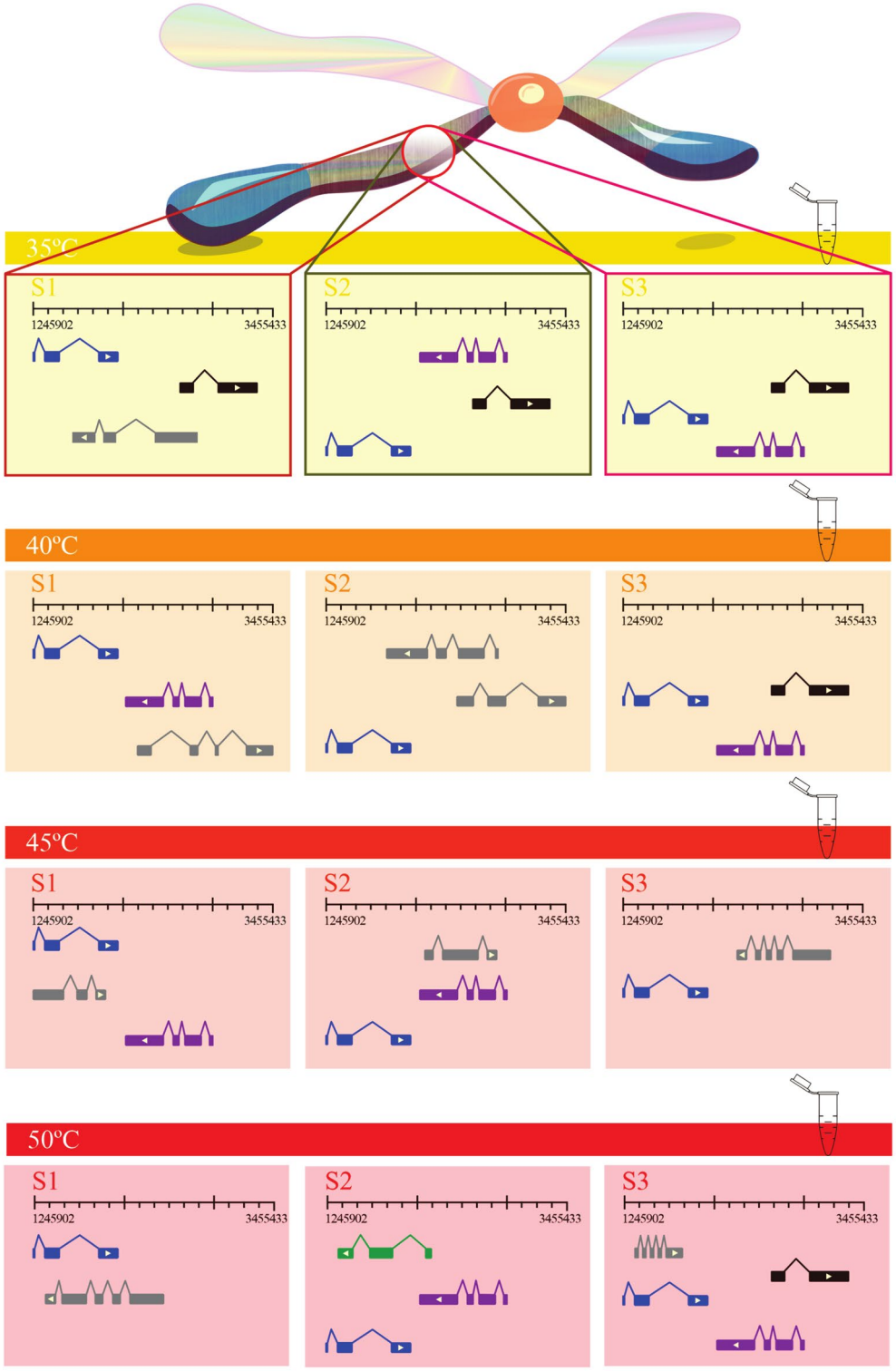


Figure 2. Comparison of transcripts among experiments at varied temperatures (35, 40, 45, and 50 ℃) revealing structural identical and non-identical transcripts. Colored transcripts demonstrated structurally identical transcripts found in each experiment.

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

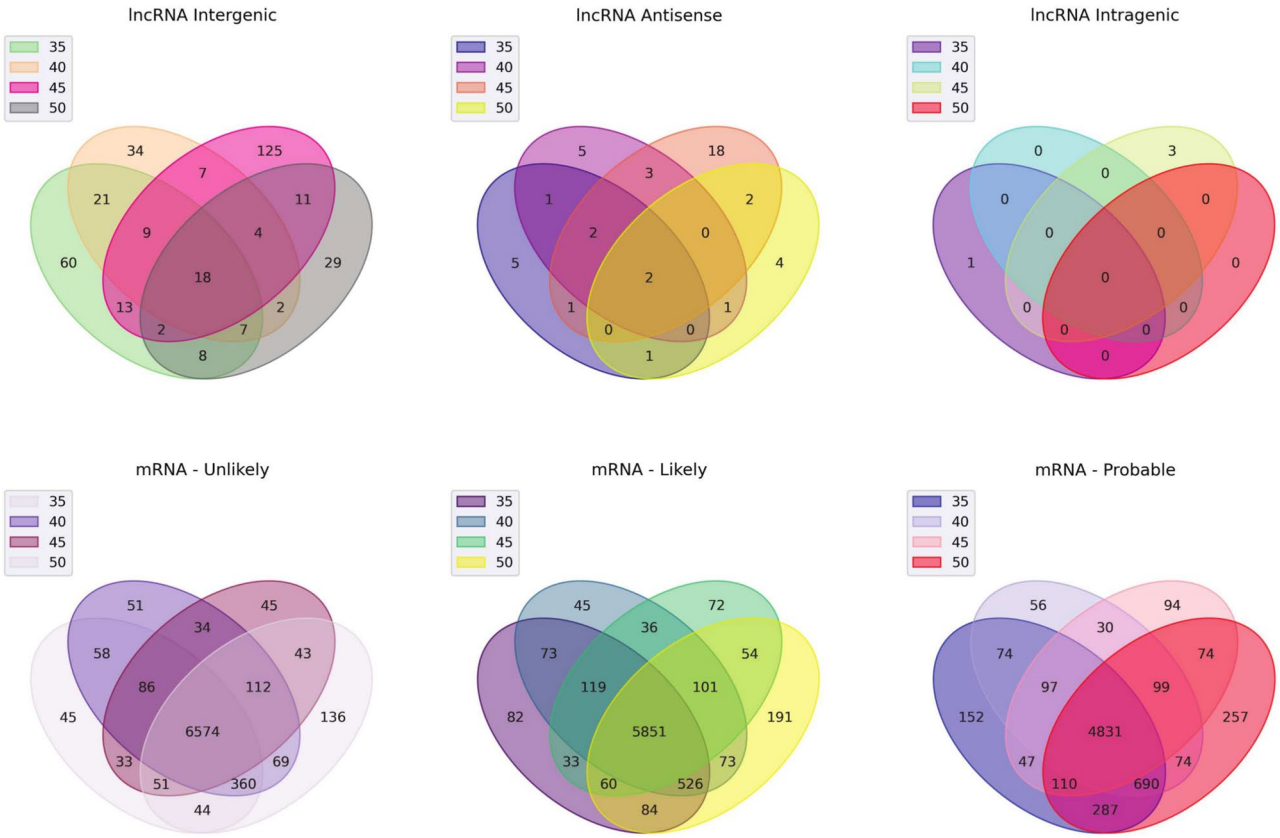


Figure 3. Venn diagrams showing the overlap of lncRNA and mRNA transcripts. The lncRNAs identified in intergenic, intronic, and antisense regions of experiments ranging from 35 to 50 ℃. Each diagram represents a specific type of transcript, with the number in the circle indicating the total number of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

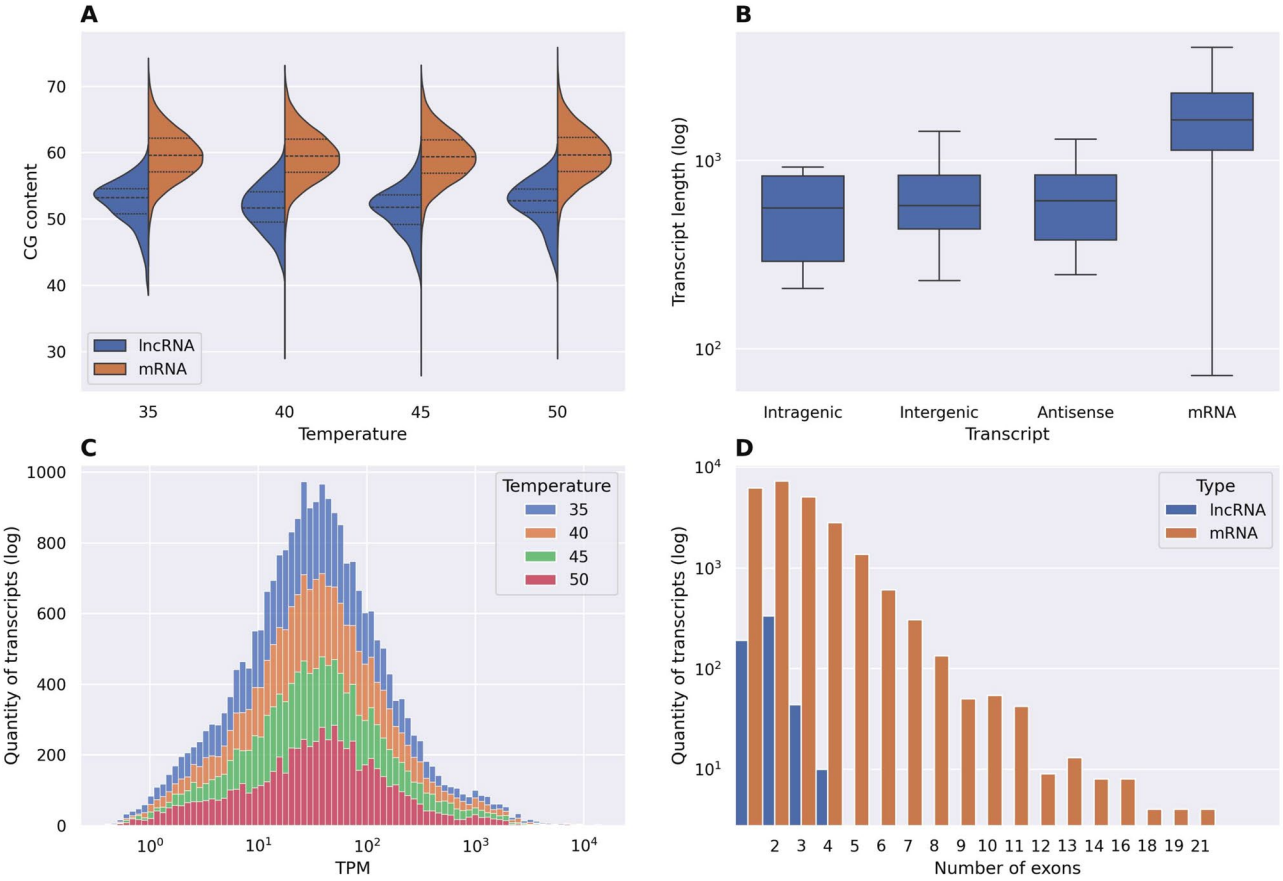


Figure 4. (**A**) mRNA and lncRNA CG-content comparison between temperatures. (**B**) Length comparison of different lncRNA classes and mRNA. (**C**) Quantity of transcripts across experiments. (**D**) Distribution of lncRNA exons.

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).

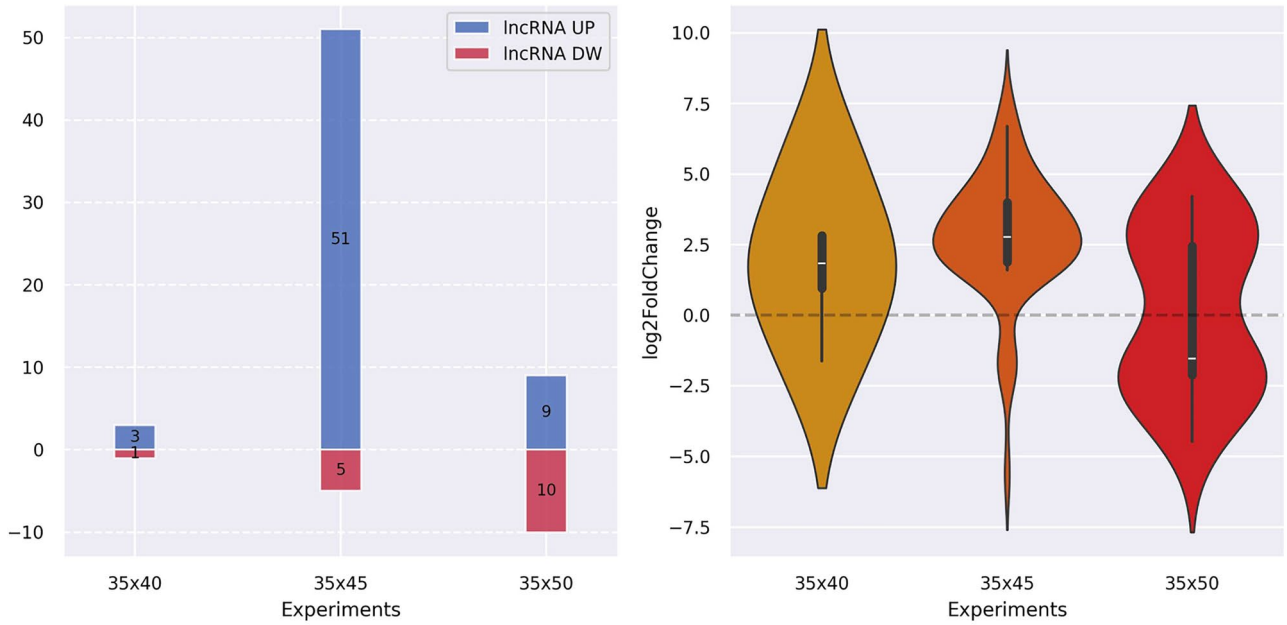


Figure 5. Distribution of differentially expressed lncRNAs and their log2 fold change between fungal experiments cultivated at 35 ℃ and exposed to thermal stress at 40, 45, and 50 ℃. The left panel shows stacked bar plots of the number of upregulated and downregulated lncRNAs for each experiment. The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

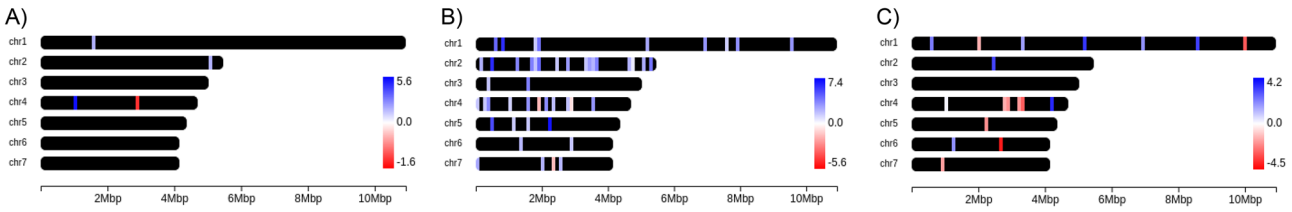


Figure 6. Distribution of differentially expressed lncRNAs within the 7 chromosomes of the fungus expressed in three different experiments (**A**) (35 × 40 ℃), (**B**) (35 × 45 ℃) and (**C**) (35 × 50 ℃). Highlighted genes represent lncRNA up-regulated (blue) or down-regulated (red)—(Anand & Lopez, 2022).

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).

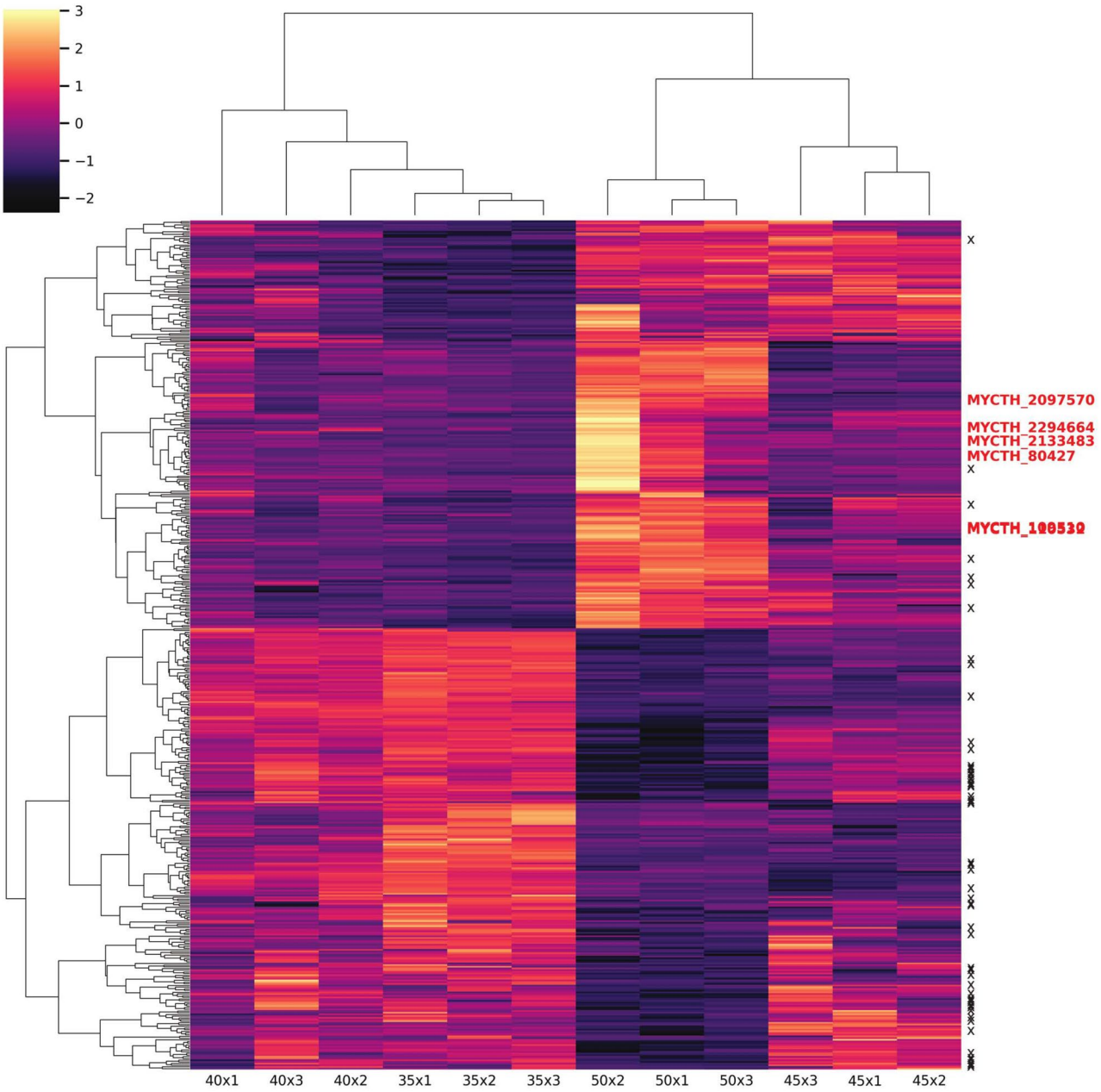


Figure 7. Cluster heatmap of the 500 top expressed genes across all samples. Z-score transformation was performed to each gene. The red labels represent HSP genes and the black asterisks (*) represent lncRNAs.

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

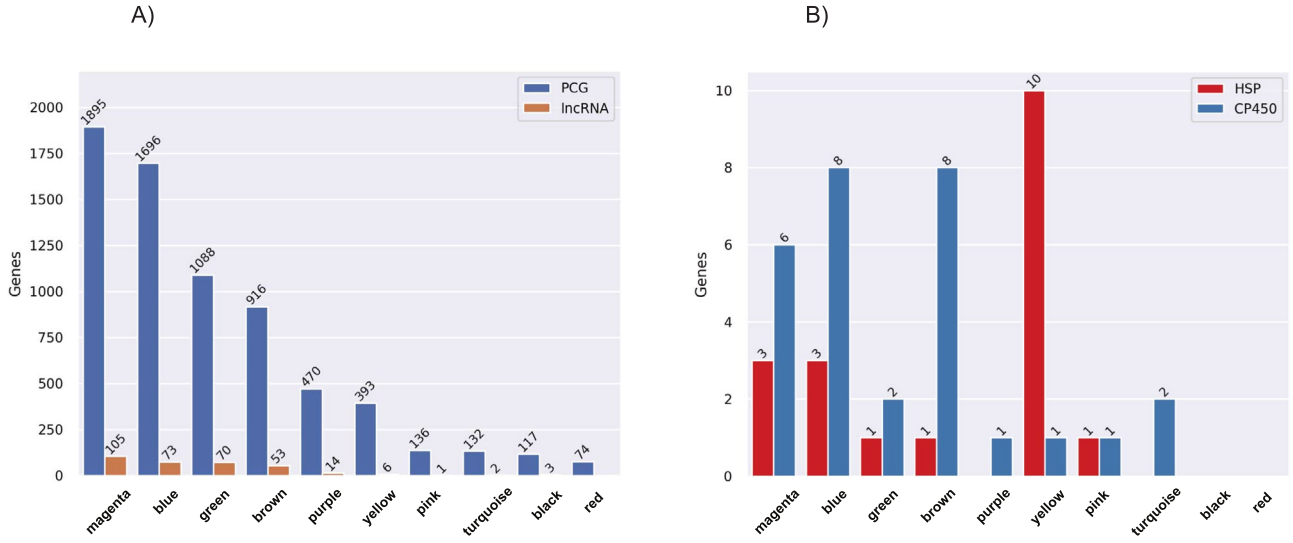


Figure 8. (**A**) The 10 modules identified by the WGCNA package. The blue bars represent mRNA and the orange bars represent lncRNAs. The bars were displayed in descending order according to the gene number in each cluster. (**B**) Bar plot showing the number of HSP (in red) and CP450 (in blue) genes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

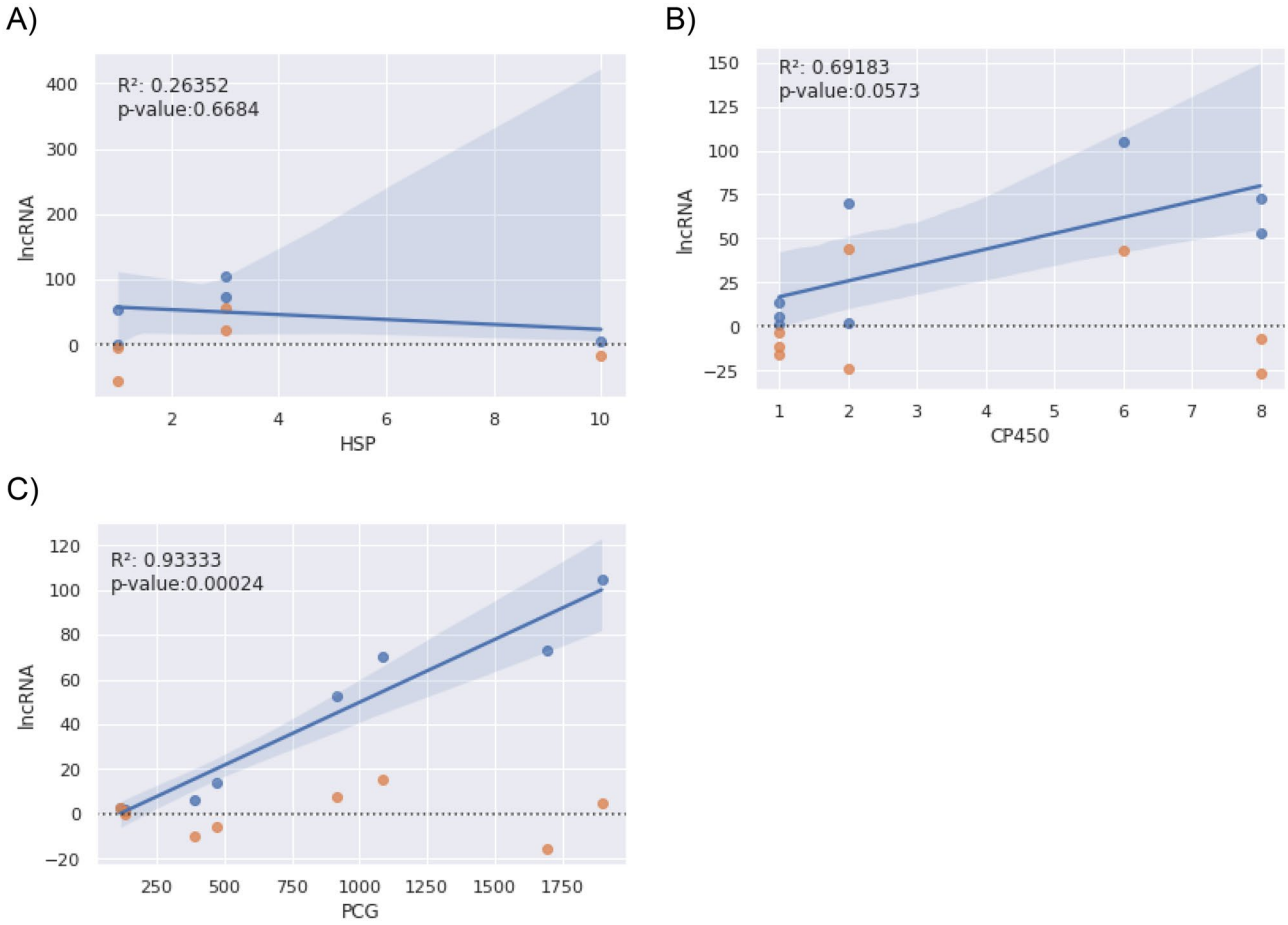


Figure 9. The graphs show Spearman's rank correlation and the relationship between lncRNAs, HSP (**A**) and CP450 (**B**), with the orange dots representing the residual errors. Data from the WGCNA analysis. (**C**) The graph represents Spearman's rank correlation. The blue line exhibited a linear relationship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

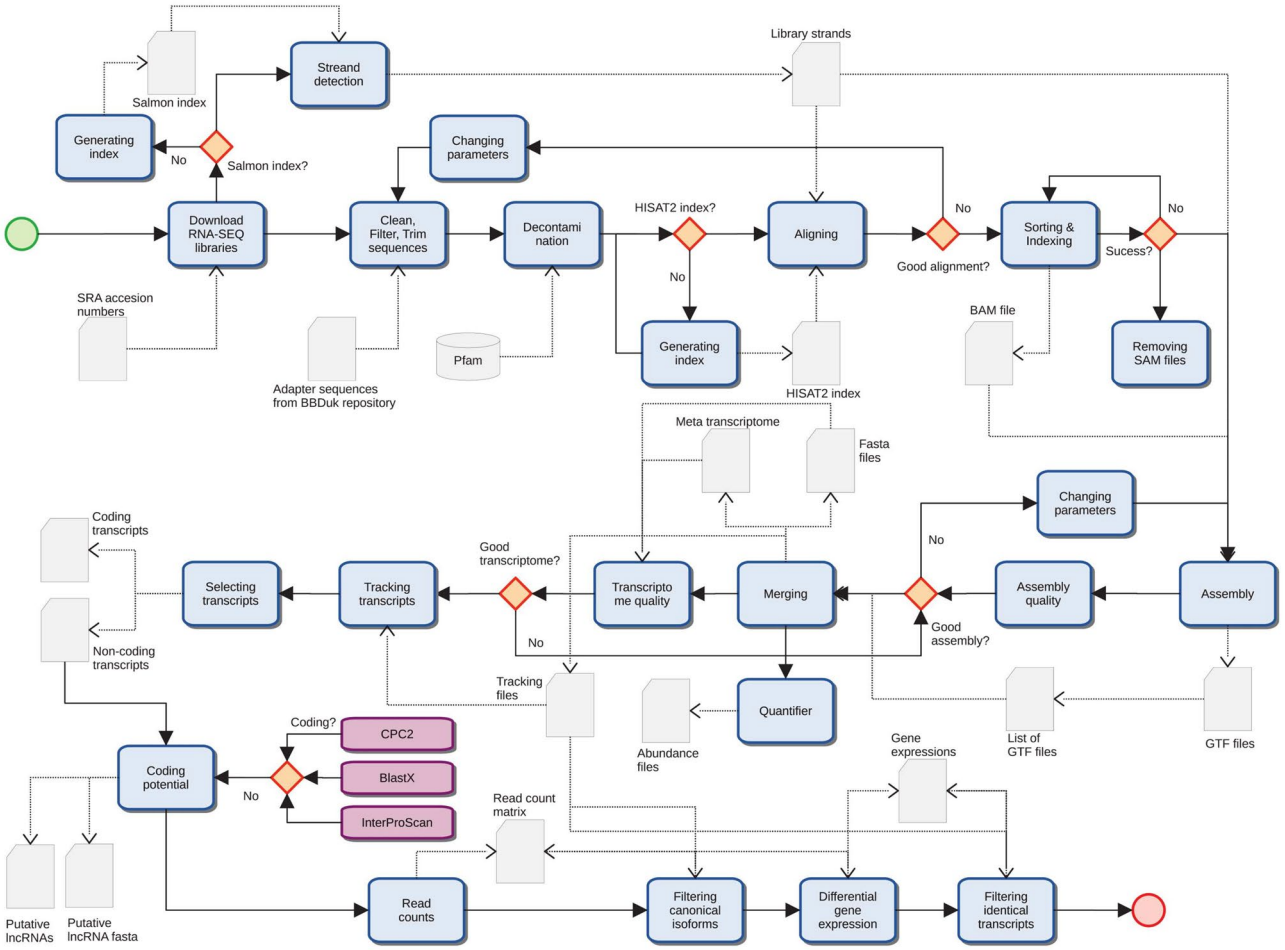


Figure 10. Detailed view of activities executed in the pipeline. Blue rectangles are activities and gray elements are input/output files.

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.

Chapter 4 - Comparing lncRNA expression in fungi that like it warm and those that like it hot!

Roger Silva[1], Glória R. Franco[2], Aristóteles Góes-Neto[3]

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. Department of Biochemistry and Immunology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil.

Keywords

Long non-coding RNA, thermophilic and mesophilic fungi, transcriptome assembly

Abstract

Fungi demonstrate the ability to thrive across a diverse range of temperatures and environments. Thermophilic fungi, a group adapted to living under thermal stress, have developed unique molecular mechanisms to cope with extreme conditions. Recently, long non-coding RNAs (lncRNAs), a class of RNA molecules longer than 200 nucleotides and not encoding proteins, have emerged as important regulators of gene expression, playing a role in the response of organisms to heat stress. In this paper, we explore the role of lncRNAs as potential genomic adaptations of thermophilic fungi by comparing their repertoire with that of mesophilic counterparts, which inhabit moderate temperatures. Additionally, we have analyzed the expression of RNA polymerases within these fungi. For achieving our aims, we have used different available lncRNA tools for characterizing these RNA molecules. Collectively, our results indicate that the genomes of these fungi contain putative syntenic lncRNAs which are enriched in shared sequence motifs. Additionally, we have identified likely thermal adaptation in the RNA polymerases of the thermophilic species.

INTRODUCTION

Fungi diverged from the animal kingdom approximately 1 to 1.5 billion years ago[1] and this evolutionary relationship is well-supported by substantial scientific evidence[2-3]. This divergence marked the split of the Opisthokonta supergroup into two major lineages: the Holomycota (also called Holofungi or Nucletmycea), which includes Fungi and their unicellular relatives and the Holozoa, which encompasses Metazoa (animals) and their unicellular relatives like choanoflagellates. The fungi kingdom is incredibly diverse, encompassing a wide range of organisms from unicellular yeasts to multicellular molds, as well as those that can thrive in extreme environments, including both cold and hot temperatures.

Cryophilic fungi are an ecological group of fungi that thrive in cold environments (<0°C), such as the Antarctic continent[4]. They have adapted to living in temperatures that would be inhospitable to most other organisms, by producing antifreeze proteins and unsaturated fatty acids with high catalytic activities that help protect their cells from freezing. On the hot side of Earth, thermophilic fungi are predominantly found in soil and other habitats characterized by the decomposition of various plant materials and grow at temperatures extending up to 62°C[5], demonstrating their remarkable adaptability to high-temperature environments. These fungi produce thermotolerant enzymes that can break down complex organic matter even at high temperatures.

In between these two extreme temperatures, mesophilic fungi prefer temperatures that are moderate, such as those found in soil and decaying organic matter. They have developed an arsenal of enzymes that allow them to break down tough organic materials, such as cellulose and lignin, which are essential for maintaining healthy ecosystems.

Furthermore, fungi are not just important for their ecological and industrial roles. They are also fascinating creatures. Thermophilic fungi is a small and interesting group in Ascomycota phylum booming at temperatures that challenge cell membrane stability and other biological structures[5]. Phylogenetically closely related to mesophilic fungi[6], which usually grow best in moderate temperatures (20–45°C), these heat-lovers fungal species are on the spot of scientific and commercial interests due to their potential sources of thermostable enzymes[6]. They feed on where organic matter decomposes and cannot grow below 30°C, clearly showing they are biologically adapted for living at that temperature range.

Temperature influences almost everything within an organism, spanning from biological functions to cellular integrity. Genome size seems to be one of those adaptive processes that have happened to these organisms to cope with elevated temperatures. Despite fungi genome sizes vary greatly among different species, from 2.9 mega base pairs of the Encephalitozoon intestinalis[7] to 893.2 Mbp from Gymnosporangium confusum[8], there is noticeable genome reduction in thermophilic genomes when they are compared to the closest mesophilic species[9]. Moreover, genes interrupted by short introns, reduction of intergenic regions, loss of protein-coding genes, repetitive sequences, transposable elements, diversity of gene structure are all characteristics of the fungal genomes[9]. Those notable features have demonstrated a challenge for fungal studies and even more for thermophilic fungi investigations.

Recently, long non-coding RNAs have emerged as significant gene expression regulators and are involved in distinct organism heat-stress responses[10-12]. Those single-stranded RNAs are longer than 200 nucleotides and do not encode for proteins. They can be classified according to their position in the genome as sense, antisense, intergenic, intronic, and bidirectional[13]. Although those molecules once were thought to be “junk” RNA, it is now known as important regulatory mechanisms in gene expression, chromatin structure and involvement with other cellular processes. Their functions are often dependent on their specific sequence, structure and its ability to interact with other cellular molecules, specifically in thermophilic organisms that make them important thermal markers. However, their mechanisms of action remain largely unknown[13].

Considering the thermophilic genome reduction event which would have influenced the intergenic region contraction as well as collaborated to the protein length reduction hypothesis, gene losses are also another genome reduction contributor that might include lncRNA genes. This reduction could be a consequence of the high cost of their functional maintenance, adaptation to a high environmental temperature, or even an elimination of genes that have low thermal stability. The present study compares lncRNAs between thermophilic and mesophilic fungi and their implication in genome reduction.

MATERIAL AND METHODS

Transcriptome data collection and assembly

In this study, it was analyzed 22 single-end 75bp poly-A cDNA RNA-seq libraries from the mesophilic fungus *Chaetomium globosum* (SRA Study SRP198385) and 12 paired-end 150bp poly-A cDNA RNA-seq from the *Thermothelomyces thermophilus* (SRA Study SRP336977), a thermophilic fungus, obtained from the NCBI Sequence Read Archive (SRA).

Mesophilic fungi libraries were inspected by FastQC[14] and MultiQC[15]. Illumina adapter sequences were removed, including low quality reads from both ends below Phred 30 using BBDuk[16]. A ribosomal decontamination with sequences downloaded from the RFam database[17] was applied also using BBDuk. Reads shorter than 50 bp were removed. Salmon pseudo-alignment tool[18] with default parameters and --gcBias --validateMappings --numGibbsSamples 200 --seqBias flagged was used to identify library strandness. The remaining reads were then mapped *de novo* using the genome based splice-aware aligner HISAT2[19] with --no-mixed --no-discordant parameters. StringTie2[20] was applied to assemble the mapped reads and generate a meta-transcriptome, according to the StringTie2 protocol. The *gffcompare* and *gffread*[21] tools were used for evaluating transcript assemblies, extracting FASTA files from the already merged assembled transcriptome and track structurally identical transcripts between the RNA-seq libraries. Ultimately, the meta-assembly FASTA file was evaluated by the rnaQUAST tool[22] according to the reference genome and its annotation file (Table 1). Regarding the thermophilic libraries, we applied the methods described by Silva, R. G., et. al[23].

Identification of lncRNAs

The *gffcompare* tool classifies transcripts based on their position within the reference genome. Transcript sequences whose length were longer than 200 bp were classified as "u", “x”, and “i” by *gffcompare* were selected. These codes stand for intergenic, antisense, and intragenic transcripts respectively. A Fungi protein-coding genes BLAST database was built locally and lncRNA genes were compared to this local protein database using BLASTx. Antisense lncRNA sequences were preprocessed before BLASTx alignment, according to methods described in Silva, R. G., et. al[23].

Additionally, InterproScan[24] was used to detect Pfam protein domains in intermediate lncRNA sequences. Any lncRNA sequence exhibiting similarity to protein domains or family were excluded. Ultimately, transcripts classified as non-coding by CPC2[25] tool in both strands without being aligned by BLASTx nor detected belonging to any protein domain or family by Interproscan were maintained.

Orthology and synteny analysis

A phylogenetic study[6] analyzed the proteomes of thermophilic and mesophilic fungi, revealing their evolutionary relationship and relatedness in the phylogenetic tree. To investigate the presence of conserved lncRNAs in both genomes, the OrthoMCL tool[26] was used for ortholog gene prediction. The analysis was performed using the Synima pipeline[27], which encompassed the execution of all-versus-all protein alignments through BLASTp[23] searches. Additionally, the DAGchainer tool[28] was employed to identify conservation blocks within the compared genetic data of the thermophilic and mesophilic fungi.

Identification of conserved lncRNA sequences between fungus

To discover conserved lncRNA within the two fungus species, a BLAST[23] database was built with all lncRNA DNA sequences for each fungus. Then, a reciprocal best hit sequence alignment was conducted using the following parameters: e-value ≤ $1x10^{-3}$, identity and query coverage ≥ 50%, and the difference in length between each pair of sequences ≥ 50% for thermophilic-mesophilic and mesophilic-thermophilic fungi sequences. The best hit for each alignment was retained.

Secondary structure conservation

The multifaceted functions of lncRNAs are frequently linked to their capacity to recognize and bind to different molecular targets. In this context, their secondary and tertiary structures play crucial roles. To compare lncRNA secondary structures, BEAGLE[29] was applied to study the secondary structure similarities between the putative orthologous lncRNAs. This tool calculates a z-score value for each pairwise comparison and, accordingly to their authors, z-score ≥ 3 and p-value ≤ 0.01 was considered a significant match.

Protein interaction analysis

LncRNA-protein interactions play crucial roles in diverse cellular processes, including gene expression regulation[30]. The specific binding between lncRNAs and proteins is often mediated through motifs or sequence elements present in the lncRNA molecule. Therefore, the catRAPID omics v2.1[31] computation-based tool was used for predicting the interaction of lncRNA-protein, by computing scores of the lncRNA-protein interaction propensities using physicochemical properties, such as secondary structure, hydrogen bonding and intermolecular force.

LncRNA localization

To investigate the subcellular localization of lncRNAs, DeepLncLoc computational tool was used. DeepLncLoc[32] is a deep learning framework designed for predicting the subcellular localization of lncRNA sequences. Only the housekeeping lncRNAs were analyzed by the DeepLncLoc web server.

Motif discovery analysis

Recent studies have observed that lncRNA conservation might be restricted to small regions within the overall sequence[33]. Therefore, to evaluate the degree of motif similarities between the lncRNAs from both fungi, lncRNA sequences were assessed using the MEME online suite[34]. Motifs were considered significant with e-value < 0.05. Additionally, the motifs were searched in the Fungi JASPER database[35] to identify likely transcription factor binding sites. It was also used the LncLOOM[36] algorithm, which is designed to identify deeply conserved short sequence motifs in ordered way within orthologous sequences. The parameters flagged in this tool were --startw 30, which sets lncLOOM for scanning motifs of length 30 nucleotides long and decreases until 6 nucleotides. The other parameter was -r 100 which defines the number of random interactions for statistical purposes.

Analysis of RNA polymerase expressions

RNA Polymerase is an enzyme responsible for synthesizing RNA from a DNA template during the process of transcription. Therefore, to investigate differences in RNA Polymerase gene expressions between the two fungi, the largest subunits of RNA polymerases (RNAP) I, II and III

from mesophilic and thermophilic fungi were retrieved computationally from the OrthoDB database (https://www.orthodb.org/) based on their annotations. The subunits RPA1 for RNAP I, RPB1 for RNAP II and RPC1 for RNAP III were selected because those subunits contain the catalytic site or the site where the polymerization of RNA occurs. Subsequently, all retrieved genes were searched and analyzed on the UniProt website, having demonstrated a minimum of 50% identity in the Uniprot Reference Clusters. Transcripts per million (TPM) expression for each gene symbol was then queried in the StringTie assembled transcript tracking file. Deseq2[37] analysis method was applied to transcript count matrix files from libraries of thermophilic fungi. Kruskal-Wallis H and post-hoc pairwise Dunn's test analyses, both non-parametric tests, were performed on TPM values of thermophilic RNA polymerase genes.

RESULTS

Characterizing features of mesophilic and thermophilic lncRNAs

Considering the phylogenetic relatedness of the mesophilic and thermophilic selected fungal species, a comparative analysis on both transcriptomes was performed to determine the abundance and diversity of their transcripts and to characterize their lncRNAs. This analysis uncovered intriguing findings. Figure 2 presents the total number of transcripts, for both mRNAs and lncRNAs, and it demonstrates that the transcript repertoire of the thermophilic fungus is significantly larger than the mesophilic fungus, except for the mesophilic intragenic lncRNAs, which showed a difference of 2 transcripts more than the thermophilic. The results indicated that there were three types of lncRNAs in both transcriptomes, 350 intergenic, 45 antisense and 4 intragenic for the thermophilic fungus, whilst for the mesophilic lncRNAs, there were 206 intergenic, 21 antisense and 6 intragenic.

Currently, the mesophilic genome assembly is at the scaffold level and likely missing gene annotations. Nevertheless, a more precise quantification will be achievable when both genomes are fully assembled at the chromosome level. Moreover, both fungi should be cultivated at temperatures higher than their optimal growth temperatures, which was the case only in the thermophilic experiment. Despite these considerations, the thermophilic transcriptome exhibited the expression of 75% of all annotated protein-coding genes (9,292), whereas the mesophilic transcriptome expressed only 41% of its annotated protein-coding genes (11,048). This divergence in expression raises intriguing questions regarding the underlying reasons for such

differences in transcription pattern between the two fungal species. Therefore, further investigations are needed to elucidate the molecular mechanisms and environmental factors that contribute to these observed characteristics.

When it comes to analyzing alternative splicing transcripts, limited depth in RNA library data could have higher chance of missing lowly expressed or rare isoforms, but, on the other hand, by prioritizing structural identical isoforms, the analysis increases the likelihood of capturing isoforms that contribute the most to the overall gene expression pattern. Therefore, Figure 3 demonstrates the differences in isoform abundance between the thermophilic and mesophilic fungus. The thermophilic mRNA exhibited approximately three times more isoforms (1233) compared to the mesophilic mRNA (407). This pattern was also observed in intergenic lncRNAs, with 61 isoforms identified for thermophilic intergenic lncRNAs and only 22 for mesophilic intergenic lncRNAs. Interestingly, no antisense isoforms were detected in the mesophilic fungus, while 6 were identified in the thermophilic fungus. Additionally, this analysis did not identify any intragenic isoforms in these fungi. These findings might provide evidence on the differential isoform pattern between thermophilic and mesophilic fungi, suggesting that thermophilic fungi may rely on isoforms to cope with elevated temperatures.

The GC content of each fungus was calculated and plotted for mRNA and the analyzed lncRNAs (Figure 4). It is evident that the CG content of mRNAs is notably higher than that of the lncRNAs, reaching approximately 60% for both fungi. It is worth noting that one possible adaptation for thermophilic fungi to thrive at higher temperatures is an increase in the dinucleotide C-G. However, comparing their mRNA, they exhibit similar CG content distributions. Additionally, looking at the lncRNA molecules, they display comparable interquartile distributions, with minor variation in the median for each distribution. Intragenic lncRNAs showed a slightly larger interquartile range, likely due to the number of samples in the distribution.

Protein-coding gene losses contribute to genome reduction in thermophilic fungi. Interestingly, when comparing transcript lengths (Figure 5) between the thermophilic and mesophilic fungi, it is noticeable that the median length of transcripts in the thermophilic fungus is located close to the third quartile of mRNA lengths in the mesophilic fungus. The thermophilic median is approximately 1700 bp, while the mesophilic median is around 1100 bp, indicating that the mRNA transcripts from the thermophilic fungus are longer than those from the mesophilic counterpart. While the median is centered for both antisense transcripts in the fungi, the

intergenic transcript distribution is slightly right skewed, suggesting the presence of shorter transcripts in both samples for this category. Overall, the distributions of antisense and intergenic lncRNAs are similar for both fungi, with the mesophilic transcripts exhibiting a broader dispersion compared to thermophilic transcripts.

Regarding the expression levels between the mesophilic and thermophilic fungi, Figure 6 compares mRNA and lncRNA transcripts, revealing higher expression of thermophilic mRNA compared to the mesophilic fungus. Despite the use of the transcripts per million (TPM) normalization method to plot the expression values, TPM calculations rely on the total number of mapped reads. Consequently, variations in library depth can impact TPM values, leading to inflated values for certain transcripts[42]. Additionally, only the comparison of antisense lncRNA TPM values showed comparable results, with 929 for the mesophilic fungus and 1280 for the thermophilic fungus when comparing the same antisense transcripts. In contrast, other transcript types, such as mRNA, intergenic, and intragenic lncRNAs, displayed substantially different values between the two species. For instance, intergenic transcripts exhibited higher abundance in the mesophilic fungus (63,678 TPM) compared to the thermophilic fungus (10,961 TPM). Similarly, intragenic transcripts followed a similar pattern, with greater expression in the mesophilic fungus (550 TPM) than in the thermophilic fungus (95 TPM), representing an approximately five-fold difference in expression levels between the two species. To analyze in detail the TPM values for mRNA and lncRNAs, Figure 7 depicts expression levels of mRNAs and lncRNAs across mesophilic scaffolds and thermophilic chromosomes in each fungus. The discrepancy in TPM values, showing a higher abundance of thermophilic transcripts at lower values, when compared to the mesophilic, could be an indicator of specific gene expression levels or potential variation in transcriptional activity between the organisms.

Finally, the distances in base pairs between the Transcription Termination Site (TTS) of protein coding genes and the closest downstream Transcription Start Site (TSS) of intergenic lncRNAs were compared in mesophilic and thermophilic fungi. Additionally, the distance from the TTS of intergenic lncRNAs to the closest downstream Transcription Start Site (TSS) of protein coding genes was also computed (Figure 8). The graph reveals that the distance from the TTS of protein coding genes to intergenic lncRNAs is smaller in the mesophilic fungus compared to the thermophilic fungus. A similar trend is observed when comparing the distance between the TTS of an intergenic lncRNA to protein coding genes and the transcription termination site of lncRNAs, suggesting that there is a relatively shorter distance between intergenic lncRNAs and

protein coding genes in the mesophilic fungus compared to the thermophilic fungus. The maintenance of intergenic lncRNA elements in the thermophilic genome, even after the genome reduction event, may indicate their functional significance in important regulatory processes of the thermophilic fungus.

LncRNA sequence conservation

Although the evolutionary conservation of lncRNA nucleotide sequences across different species can vary from poorly conserved to highly conserved compared to protein-coding gene sequences[43], the conservation of lncRNA sequences within these fungi was analyzed. Out of a total of 350 lncRNAs for the thermophilic fungus and 233 for the mesophilic fungus, this analysis identified a limited number of conserved lncRNAs and one lncRNA isoform that aligned to the same lncRNA in both databases using BLAST tool. This outcome suggests that the majority of lncRNAs show limited conservation between the two species, an indication of rapid sequence evolution of the lncRNAs, possibly due to an acquiring species-specific functions, regulatory roles or even thermal adaptation.

Secondary structure conservation

The lack of sequence conservation in the previous analysis of lncRNAs suggests that their functional elements may reside in other features, such as their secondary structure. In this regard, pairwise alignments of mesophilic lncRNAs with all other thermophilic lncRNAs were analyzed. Applying the threshold recommended by the authors of the BEAGLE algorithm (z-score > 3), only 5 lncRNAs were considered structurally similar across the mesophilic and thermophilic fungi. The BEAGLE algorithm identified 416 alignments with a z-score ≥ 3 and p-value ≤ 0.01. Nevertheless, considering the importance of structural identity when aligning RNA secondary structures to determine functional similarity, a structural identity threshold of ≥ 50% was also applied to the previous selection, resulting in five alignments. Remarkably, one mesophilic lncRNA aligned to three different thermophilic lncRNAs. Based on these results, it appears that this approach did not provide significant evolutionary inferences between the two sets of fungi lncRNAs.

Finding orthologous lncRNAs

Since the previous approaches failed to identify orthologous lncRNAs based on sequence similarity or secondary sequence conservation, further investigation into orthology and synteny was deemed valuable and subsequently conducted. The orthologous study revealed the presence of 6,498 shared genes (Figure 9) between the mesophilic and thermophilic fungi, indicating a significant overlap in protein coding gene content and emphasizing the substantial conservation observed among these organisms. In addition, the syntenic study (Figure 10) aimed to determine the presence of lncRNAs within genomic regions between orthologous genes. The analysis identified 74 likely loci for intergenic lncRNAs in mesophilic adjacent orthologous genes and 172 likely loci for intergenic lncRNAs in the thermophilic fungi. Moreover, a subset of 8 housekeeping lncRNAs was identified between the adjacent orthologous genes, representing potential candidates for conserved regulatory roles essential to both fungal species.

Subcellular localization of lncRNAs

The housekeeping lncRNAs were utilized in the DeepLncLoc tool, which predicts lncRNA localization in cellular compartments within a cell, including Cytoplasm, Nucleus, Exosome, Ribosome, and Cytosol. Using this tool, the localization patterns between mesophilic and thermophilic lncRNAs were compared. Nonetheless, the results displayed a different subcellular localization for eight pairs of lncRNA and did not provide any insights into the evolutionary relationship of housekeeping lncRNAs. Only two pairs exhibited localization in the same cellular region within the cell, which were identified in the cytoplasm (TCONS_00000833, TCONS_00003919) and nucleus (TCONS_00018980, TCONS_00007770). Therefore, this tool was not so effective in identifying a consistent subcellular localization pattern for the housekeeping lncRNAs and a complementary approach is needed.

Protein interaction with lncRNAs

One established function of lncRNAs is to act as protein scaffolds, bringing protein into close proximity forming functional complexes of RNA-binding proteins. The catRAPID tool estimates the binding propensity of protein-RNA pairs by combining secondary structure, hydrogen bonding and van der Waals contributions. It was computed the protein-RNA potential interaction between all protein coding-genes and all sets of mesophilic and thermophilic lncRNAs. The analysis did not exhibit any significant predicted interaction between mesophilic and thermophilic lncRNAs and any proteins, not even using the shortened list of lncRNAs found previously.

Motifs within the lncRNAs

Despite the limited sequence conservation observed within both sets of lncRNAs, as well as the lack of substantial evolutionary inferences based on their secondary structures, subcellular localizations, and protein interactions, this study aimed to investigate the potential enrichment of shared sequence motifs. To explore this aspect, two motif discovery tools were employed to analyze the sets of lncRNAs. Conserved sequence motifs were analyzed by the MEME suite, which detected one motif containing 29 nt long in 127 sequences with an e-value of 1.7e-054 for the mesophilic lncRNA sequences. For the thermophilic lncRNA sequences, a motif of 29 nt in length was detected in 168 sites with an extremely low e-value of 4.9e-072. However, after having analyzed these finding motifs in the Fungi JASPER database through the MEME suite, e-value (calculated by multiplying the p-value by the total number of target motifs in all the target databases) and q-value (False Discovery Rate) exhibited non-significant values. Additionally, LncLOOM, an algorithm designed to identify orderly biologically relevant short conserved motifs within lncRNAs was also executed. This algorithm requires a set of orthologous lncRNAs to be compared to and the authors demonstrated that functional elements require motif conservation, and the order of these motifs are conserved across long evolutionary distances. Therefore, the shortened list of lncRNAs previously found was processed by lncLOOM, and this approach was able to identify motifs with each pair of lncRNAs. Table 2 describes the mesophilic and thermophilic lncRNAs, the number of motifs found in order, and their respective lengths. Based on the findings of the lncLOOM paper, which examined Cyrano lncRNAs across 17 species and identified 9 conserved motifs in each sequence, a threshold of 9 motifs was applied in this analysis. Consequently, the first, second, fourth, and fifth lncRNAs were filtered out, resulting in the identification of 4 orthologous lncRNAs that contained 9 or more motifs. Concluding, despite the limited sequence conservation, lncLOOM algorithm and orthologous analysis highlighted the motif conservation and motif order in functional elements across evolutionary lncRNAs and was able to identify four biologically relevant transcripts within the mesophilic and thermophilic lncRNAs.

RNA polymerase expression analysis

RNA polymerases are essential for gene expression in all living organisms. RNA polymerase I is responsible for transcribing ribosomal RNA (rRNA) genes, which are essential for ribosome

biogenesis and protein synthesis. RNA polymerase II is essential for transcribing protein-coding genes in both mesophilic and thermophilic fungi and is highly conserved across eukaryotes, including fungi. Along with these two RNA polymerases, RNA polymerase III transcribes genes encoding small non-coding RNAs, such as tRNAs and 5S ribosomal RNA and some small nuclear RNAs (snRNAs). These three classes of RNAs are the main types of RNA polymerases typically found in fungi. Furthermore, they consist of multiple subunits, both small and large, that come together to form a functional enzyme. At Sordariales (Ascomycota ⇢ Sordariomycetes) level, the OrthoDB classified the fungi RNAPs orthologous genes from ten different species, each presenting single copies and none multi-copy in those ten species.

Our analysis showed that DESEq2 did not detect any differences in the expression of RNAP genes among the thermophilic RNA libraries. The mesophilic RNA libraries were not analyzed by DESEq2 because they were not prepared under the same cultivation temperature conditions as the thermophilic libraries. Additionally, structurally identical RNAP I and RNAP III were not found in any mesophilic transcriptomes, only RNAP II (identified by the gene CHGG_01704). Therefore, mesophilic RNAPs were not analyzed due to the lack of structurally identical transcripts between different classes of RNAP. Moreover, no structurally identical samples were found for RNAP I and RNAP III in the experiment conducted at 45°C in thermophilic fungi. Consequently, RNA II expressions for experiments carried out at 45°C in all samples were excluded from subsequent analysis.

We were interested in two questions. Firstly, if there is any difference in gene expression of RNAP I, II, or III among thermophilic samples at different temperatures; and secondly, if there are any differences in RNA polymerase expressions among them in order to address those questions, we applied Kruskal-Wallis H-test and Dunn's test pairwise comparisons. The Dunn’s test is commonly used as a post-hoc test for comparing multiple group means. Table 4 and 5 present the results of both tests. It shows that only RNAP III samples rejected the null hypothesis for all tested temperatures. Additionally, RNAP I and RNAP II expressions cultivated at 50°C, and RNAP II and RNAP III expressions cultivated at 40°C did not meet the threshold of 0.05 and, consequently, the null hypotheses were not rejected.

DISCUSSION

The mesophilic fungus *Chaetomium globosum* and its evolutionary counterpart, *Thermothelomyces thermophilus*, a thermophilic species, present an opportunity for studying lncRNAs due to the potential to observe adaptive evolutionary mechanisms at the molecular level. Since these fungal species share a common ancestor[9], their adaptation to distinct environmental temperatures likely involved regulatory changes at the genomic level, including alterations in protein-coding genes or changes in gene expression patterns mediated by non-coding RNAs[6]. Utilizing our recently developed computational pipeline for identifying structurally identical lncRNAs, which adds another level to lncRNA identification and characterization, we have applied this approach to publicly available bulk RNA-seq datasets from mesophilic and thermophilic libraries. This investigation aimed to elucidate the conservation and divergence of lncRNAs between these closely related fungi, shedding light on the evolutionary dynamics of non-coding elements in response to thermal environmental stress.

In our analysis, we explored the lncRNA repertoires within these two fungal species and compared them in each fungus. Primarily, we focused on elucidating the role of lncRNAs in temperature adaptation, particularly regarding their abundance, diversity, and potential regulatory functions. Significant intergenic lncRNA quantity was observed in thermophilic fungi, ~70% higher than the mesophilic fungi, suggesting their potential involvement in mediating responses to temperature stress. Interestingly, the length of intergenic lncRNA transcripts apparently exhibited the same median in both fungi, a genetic location supposed to be shrinked in thermophilic fungi. The detected higher number of thermophilic isoforms (~3x), particularly in intergenic regions, suggests a potential role for using isoform splicing in fine-tuning gene expression under elevated temperatures, likely due the genome reduction in these organisms. Moreover, Figure 8 demonstrates the distance in base pairs from Transcription Start Site (TSS) and intergenic lncRNA Transcription Termination Site (TTS) is longer in thermophilic fungus than in mesophilic fungus, which is another interesting finding since genome reduction is typical of thermophilic fungus. This result highlights the importance of intergenic regions in the regulatory landscape of thermophilic fungi.

The expression analysis between these two species displayed an interesting finding. Figure 7 showed the discrepancy in TPM values in thermophilic organisms, specially the higher abundance of transcripts at lower TPM values, implying that certain genes may be tightly

regulated or highly expressed even at relatively low levels. This pattern may reflect the importance of certain genes or pathways in the adaptation of thermophilic fungi to high-temperature environments and highlights that lncRNAs, which presents low abundance when compared to protein coding genes and are involved in fine-tuning gene expression by regulating the transcription, splicing, or translation of protein-coding genes, could be important players in regulating thermophilic gene expression. Moreover, these differences may also involve transcription factors, epigenetic modifications, or other regulatory elements in response to environmental cues or evolutionary pressures. Hence, understanding the regulatory mechanisms underlying these expression patterns can contribute to our knowledge of fungal adaptation and evolution.

Examining the motifs found within each fungal species, our approach identified biologically relevant motifs within both mesophilic and thermophilic lncRNAs, suggesting potential roles in regulatory processes despite sequence divergence. Despite the sequence differences, both motifs share some common features, such as the presence of repetitive sequences and alternating nucleotide patterns. These conserved features may indicate functional elements or structural motifs that are conserved across lncRNAs in both thermophilic and mesophilic fungi. Repetitive sequences and alternating nucleotide patterns have been associated with RNA secondary structure formation, RNA-protein interactions, and regulatory functions in gene expression[38-41]. Thus, the differences in motif sequences between the thermophilic and mesophilic fungi may reflect adaptations to their respective environmental conditions and regulatory requirements. Nonetheless, further analysis is needed to determine the exact roles and significance of these conserved features in lncRNA biology.

Furthermore, we have analyzed the expression of RNA polymerase I, II, and III, which are crucial for understanding the transcriptional regulatory mechanisms underlying temperature adaptation, evolutionary adaptation, and non-coding transcripts. RNA polymerase III shows significant expression across all tested temperatures, aligning with our other findings and showing the importance of non-coding loci in the thermophilic fungi. Additionally, Table 3 shows that RNAP I and RNAP III length from mesophilic fungi are smaller than the thermophilic counterparts. This likely demonstrates another thermal adaptation from this fungi which could impact in facilitating transcriptional activity at high temperatures as well as in its efficiency and specificity to transcribe their genes. This evolutionary change could also impact regulatory elements within the thermophilic genome, such as promoters and enhancers may need to be

adapted to accommodate these differences in size, linking again to the importance of intergenic regions.

CONCLUSION

This comparative analysis of lncRNAs in mesophilic and thermophilic fungi offers a wealth of information on the regulatory mechanisms underlying temperature adaptation. By elucidating the functional roles, expression patterns, and genomic organization of lncRNAs, this study significantly contributes to our understanding of the molecular basis of fungal adaptation to environmental stressors. Besides, it further paves the way for future studies aiming to unravel the intricate mechanisms of lncRNAs and fungal gene regulation and their evolutionary dynamics.

While lncRNAs may indeed be expressed at lower levels compared to protein-coding genes, their lower abundance does not necessarily diminish their biological importance. Instead, lncRNAs probably play critical regulatory roles in diverse cellular processes despite their relatively modest expression levels. In spite of the fact that available lncRNA tools still present challenges in elucidating their evolutionary history, even within these closely related organisms, fungi play critical roles in ecosystems as decomposers, symbionts, and pathogens and represent one of the earliest branching lineages among eukaryotes in the evolutionary timeline, and the study lncRNA in fungi is still in its infancy.

FIGURES

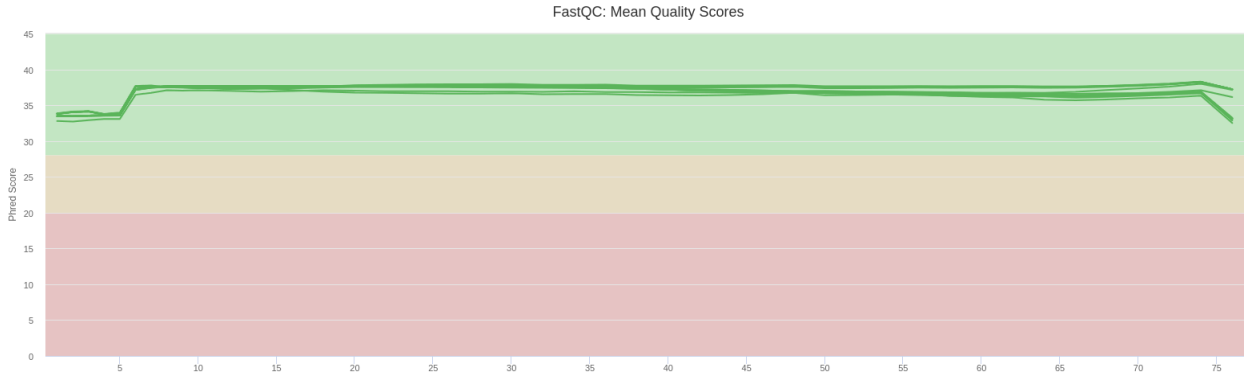


Figure 1: FastQC results of the 22 RNA-seq libraries displaying the relationship between base position (x-axis) and Phred score (y-axis). The graph was generated using Multiqc from FastQC analyses.

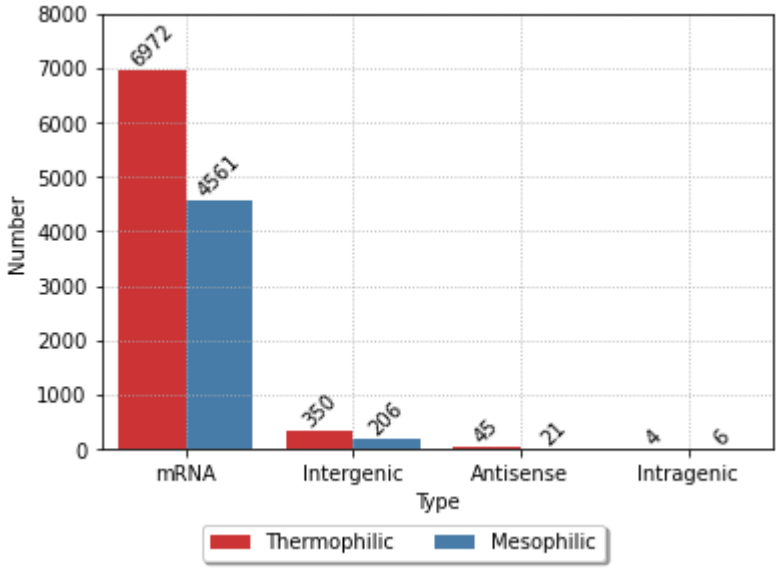


Figure 2: Comparison of transcript quantity in mRNA and Intergenic, Antisense, and Intragenic lncRNAs across thermophilic and mesophilic fungi.

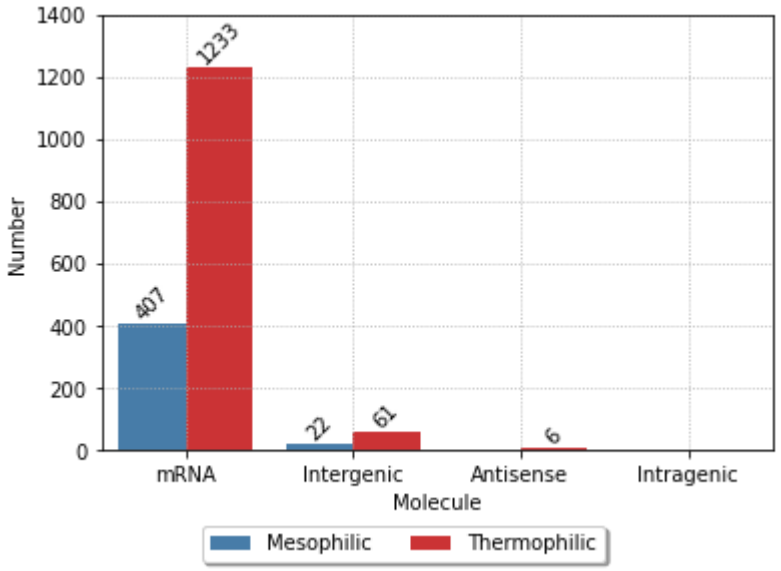


Figure 3: The bar graph shows a comparison between alternative splicing isoforms in thermophilic and mesophilic transcriptomes for mRNA and lncRNAs.

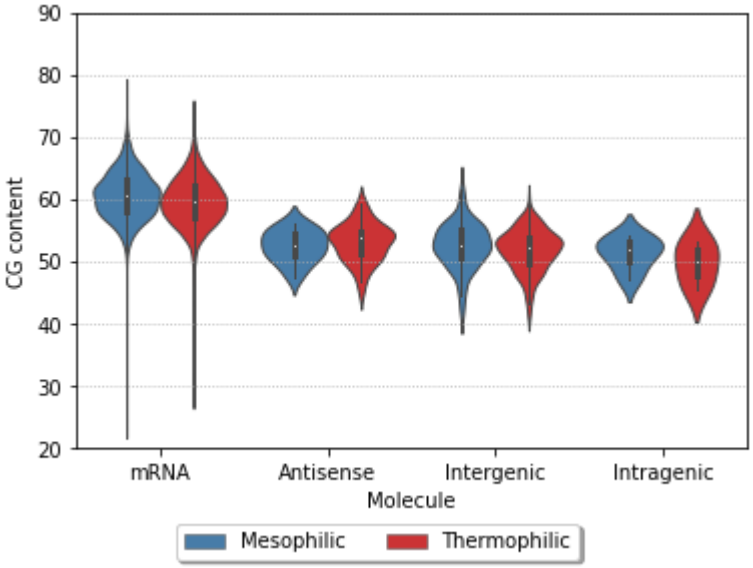


Figure 4: Distribution of CG content with outliers for mRNA and lncRNAs transcripts across thermophilic and mesophilic fungus.

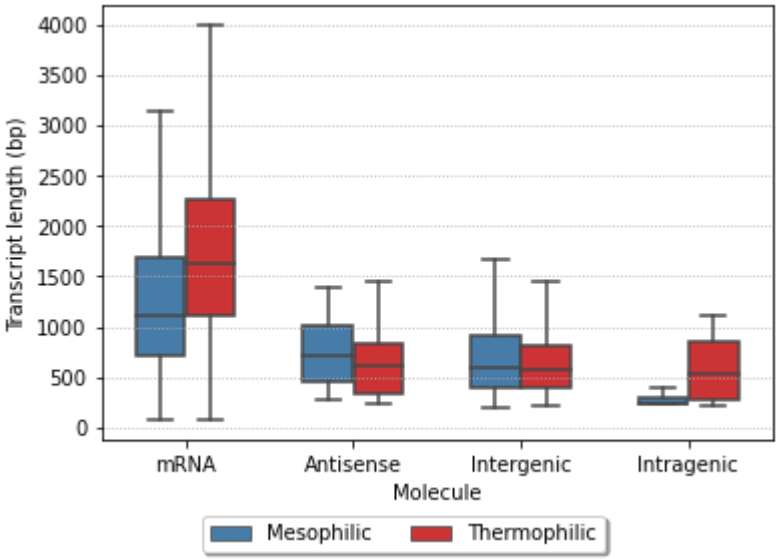


Figure 5: The graph compares the length distribution of different RNA categories in thermophilic and mesophilic fungus.

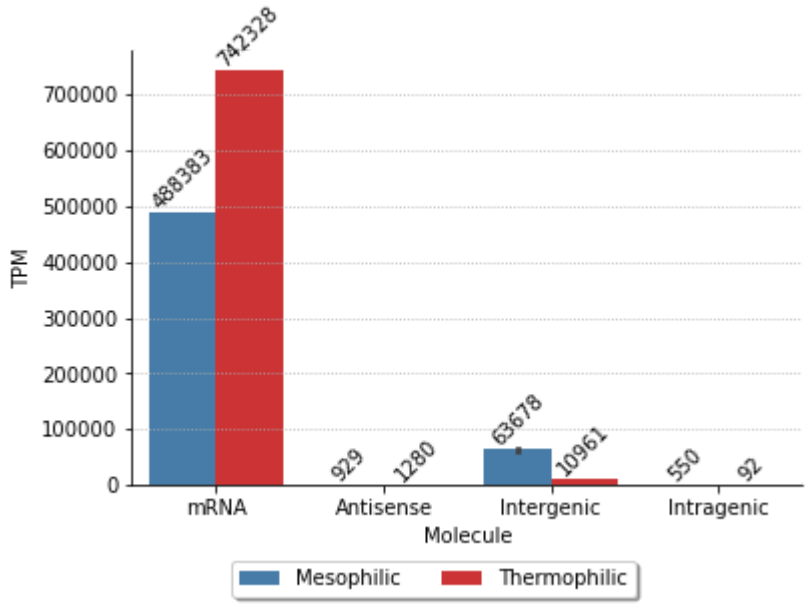


Figure 6: The bar graph illustrates the transcript abundance measured in transcripts per million (TPM) values for different RNA categories in the mesophilic and thermophilic fungi

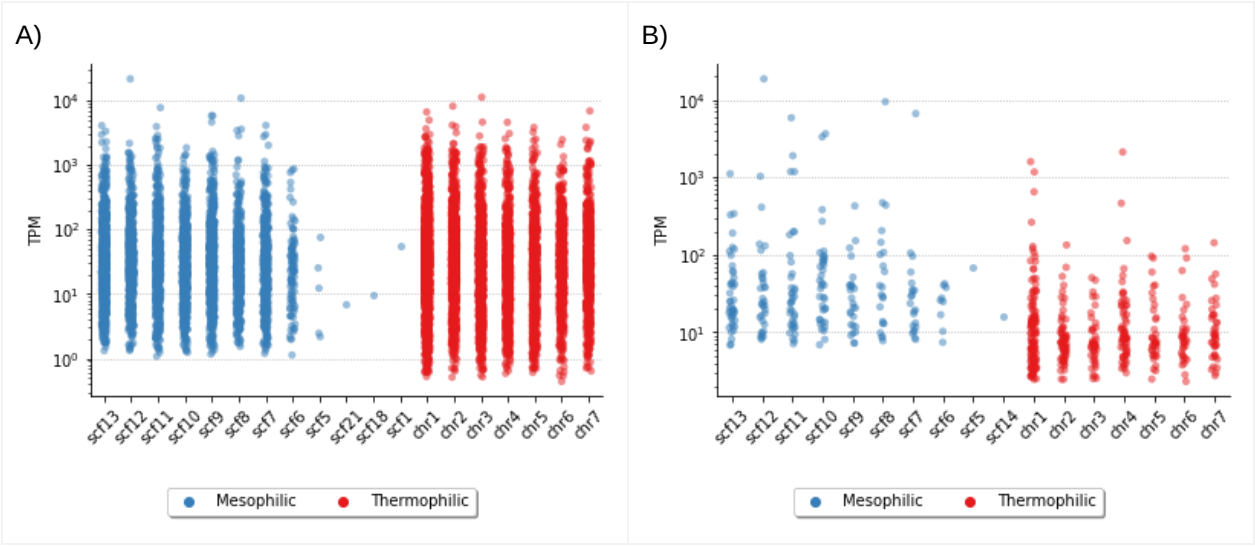


Figure 7: TPM values for each mesophilic fungus scaffold and thermophilic fungus chromosomes. The graph A) shows TPM values for all mRNA transcripts, while the graph B) presents TPM values for all types of lncRNAs.

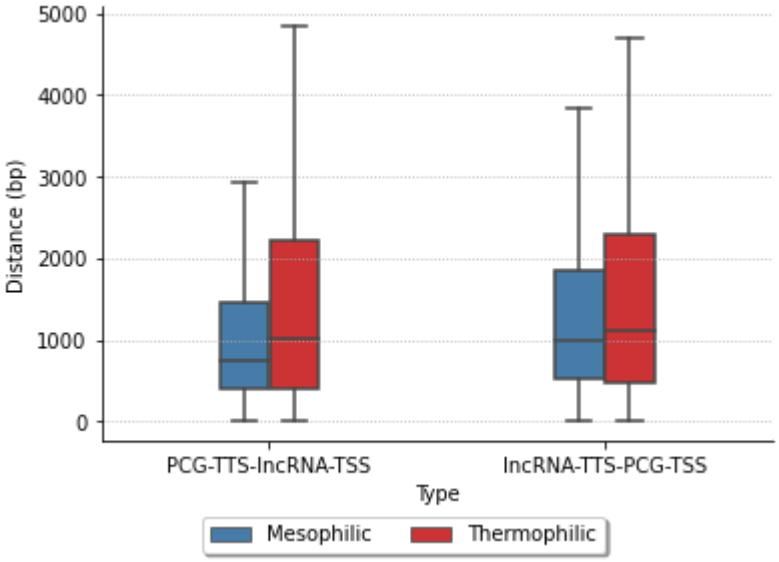


Figure 8: Boxplot comparing the distance in base pairs from a protein coding genes Transcription Termination Site (TTS) to an Intergenic lncRNA Transcription Start Site (TSS) and from an Intergenic lncRNA Transcription Termination Site (TTS) to a protein coding genes Transcription Start Site (TSS) across the mesophilic and thermophilic fungi.

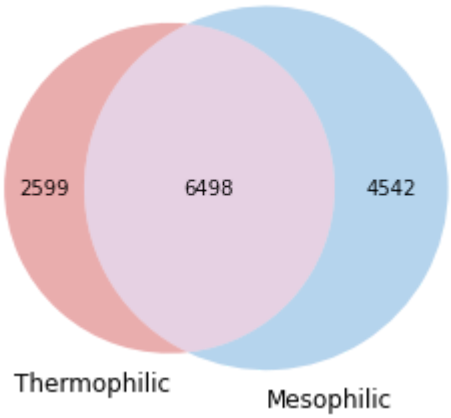


Figure 9: The venn diagram showing the number of shared orthologous protein coding genes between the genomes of the fungi.

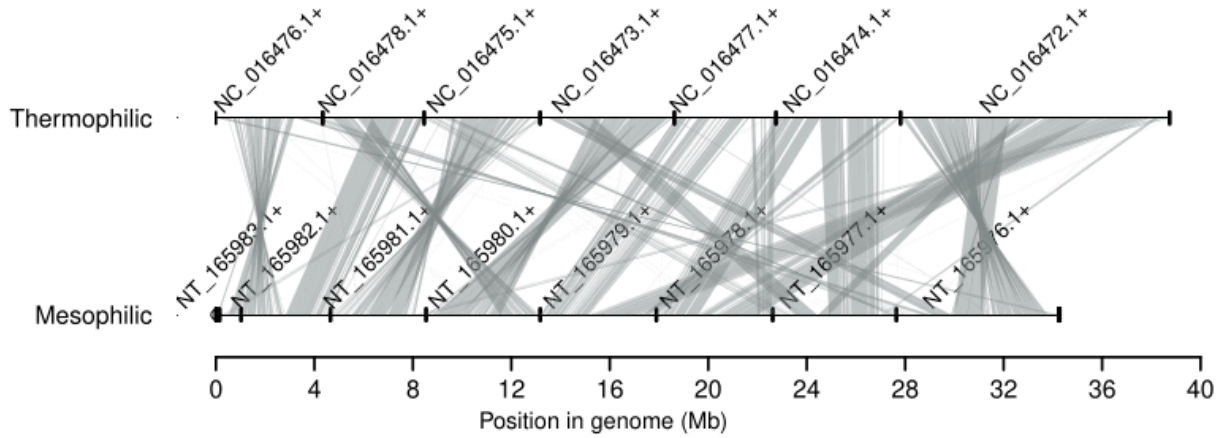
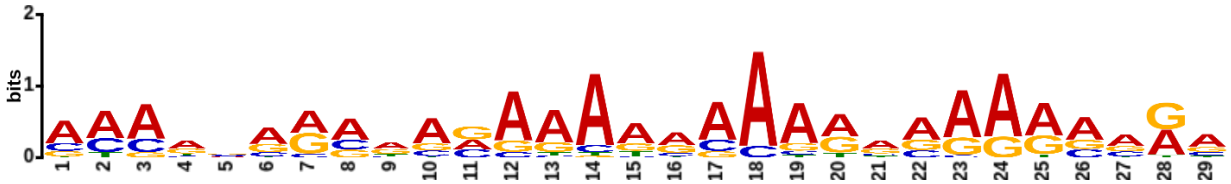


Figure 10: Ideogram representation of syntenic regions between the genomes of the fungi.

A)



B)

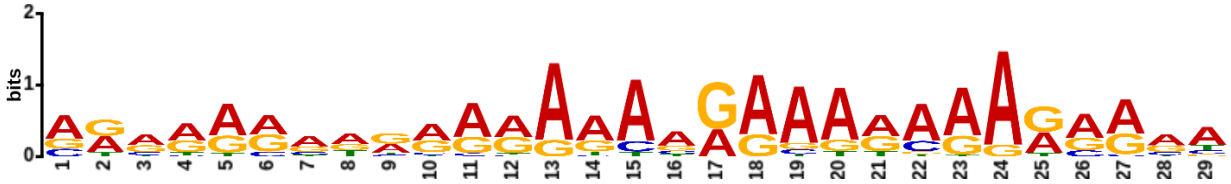


Figure 11: Motif Analysis of Mesophilic and Thermophilic lncRNA Sequences. A) Motifs found in Mesophilic lncRNA sequences, comprising 29 base pairs (bp), identified in 127 sites with an e-value of 1.7e-054. B) Motifs found in Thermophilic lncRNA sequences, also spanning 29 bp, discovered in 168 sites with an e-value of 4.9e-072.

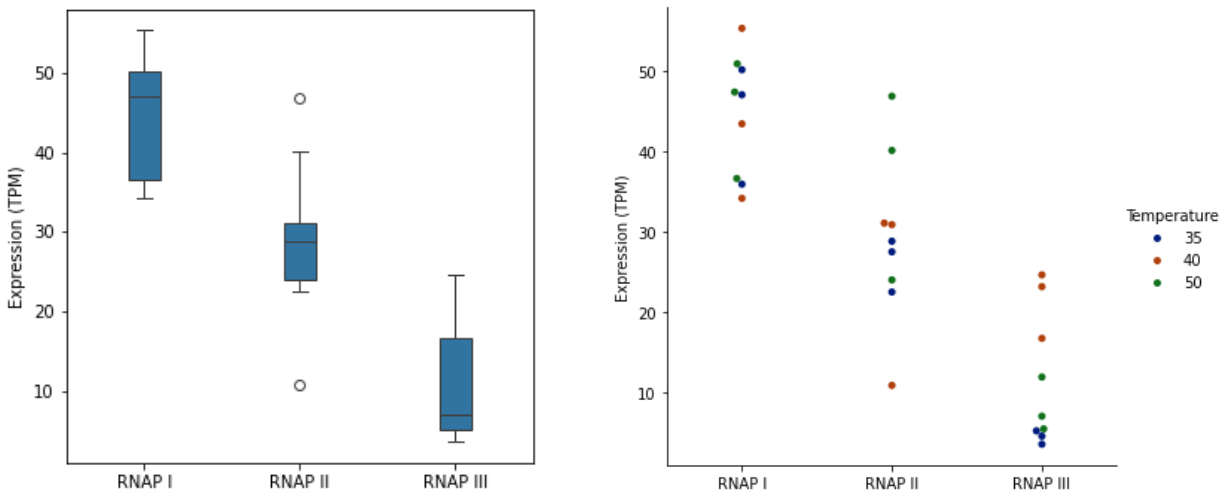


Figure 12: Expression of RNA Polymerase I, II, and III in TPM values across experiments conducted on the thermophilic fungus cultivated at temperatures of 35°C, 40°C, and 50°C.

TABLES

Table 1: This table exhibits alignment results reported by rnaQuast for the merged assembled transcriptome. It shows values for aligned transcripts greater than 500 and 1000 base pairs, total of aligned transcripts, database coverage and uniquely and multiply aligned transcripts.

Species	Genes	Transcripts							
		> 500 bp	> 1000 bp	Total	Aligned	Database coverage	Uniquely aligned	Multiply aligned	Unaligned
C. globosum	11,232	23430	17874	27229	27229	0.859	27183	46	0

Table 2: All putative orthologous lncRNAs between adjacent orthologous genes with their respectives motifs and sizes.

Mesophilic lncRNA	Thermophilic lncRNA	Motifs	Motif's lengths
TCONS_00000452	TCONS_00004091	5	8,9,7,8,7
TCONS_00000833	TCONS_00003919	6	6,8,8,6,7,7
TCONS_00002358	TCONS_00008942	9	10,6,6,9,8,7,8,10,6
TCONS_00003032	TCONS_00009020	7	7,7,7,6,8,6,7
TCONS_00017694	TCONS_00003064	8	7,6,6,9,7,8,7,8
TCONS_00018837	TCONS_00001345	14	6,8,6,6,10,8,8,7,9,8,8,6,8,7
TCONS_00023834	TCONS_00014608	13	8,6,7,8,6,7,6,6,7,7,6,8,6
TCONS_00023932	TCONS_00015086	9	6,7,6,8,6,6,7,6,8

Table 3: Orthologos RNA polymerases from mesophilic and thermophilic fungi

Mesophilic Gene symbol	Thermophilic Gene symbol	Type of RNA	Evolutionary rate	Length (Mesophilic / Thermophilic)
CHGG_10294	MYCTH_2312447	RNAP I	0.87	1475 / 1721
CHGG_01704	MYCTH_2294525	RNAP II	0.77	1756 / 1754
CHGG_08628	MYCTH_2307483	RNAP III	0.86	1290 / 1555

Table 4: Kruskal-Wallis H-test table summarizing the p-value results from RNA polymerase I,II and III expressions for thermophilic fungi cultivated at different temperatures.

RNAP	35X40	35X50	40x50
RNAP I	0.82725	0.51269	0.82725
RNAP II	0.51269	0.27523	0.27523
RNAP III	0.04953	0.04953	0.04953

Table 5: Multiple comparison of means performed by Kruskal-Wallis H-test, comparing thermophilic fungi RNAP I, II and III.

Group 1	Group 2	35	40	50
RNAP I	RNAP II	0.04953	0.04953	0.27523
RNAP I	RNAP III	0.04953	0.04953	0.04953
RNAP II	RNAP III	0.04953	0.51269	0.04953

Supplementary information

Packages and their version used in this chapter

BBDuk tool - 39.01
DESeq2 - 1.40.2
HISAT2 - 2.2.1
StringTie2 - 2.2.1
Gffcompare - v0.11.2
Gffread - v0.12.7
Samtools - 1.17
Salmon - v1.9.0
rnaQUAST - v.2.2.2
FastQC - v0.11.9
Multiqc - v1.14
mstrg_prep.pl - https://gist.github.com/gpertea/b83f1b32435e166afa92a2d388527f4b
Pandas - 2.0.3
Numpy - 1.25.2
Matplotlib - 3.7.2
Seaborn - 0.12.1
Biopython - 1.80
Venn - 0.1.3
Scipy - 1.11.2
Statistics - 3.4

Python - 3.10.8

REFERENCES

1. Taylor, J. W. & Berbee, M. L. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia***98**, 838–849 (2006).

2. Brown, M. W. *et al.* Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biology and Evolution***10**, 427–433 (2018).

3. Wainright, P. O., Hinkle, G., Sogin, M. L. & Stickel, S. K. Monophyletic Origins of the Metazoa: an Evolutionary Link with Fungi. *Science***260**, 340–342 (1993).

4. Ordóñez-Enireb, E. *et al.* Antarctic fungi with antibiotic potential isolated from Fort William Point, Antarctica. *Scientific Reports***12**, (2022).

5. Maheshwari, R., Bharadwaj, G. & Bhat, M. K. Thermophilic Fungi: Their Physiology and Enzymes. *Microbiology and Molecular Biology Reviews***64**, 461–488 (2000).

6. 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 Genomics***19**, (2018).

7. Keeling, P. J., Fast, N. M. & Corradi, N. Microsporidian Genome Structure and Function. *Wiley Online Library* https://onlinelibrary.wiley.com/doi/10.1002/9781118395264.ch7.

8. Tavares, S. *et al.* Genome size analyses of Pucciniales reveal the largest fungal genomes. *Frontiers in Plant Science***5**, (2014).

9. Berka, R. M. *et al.* Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris. *Nature Biotechnology***29**, 922–927 (2011).

10. Hu, X. *et al.* Identification and characterization of heat-responsive lncRNAs in maize inbred line CM1. *BMC Genomics***23**, (2022).

11. Zhao, J., He, Q., Chen, G., Wang, L. & Jin, B. Regulation of Non-coding RNAs in Heat Stress Responses of Plants. *Frontiers in Plant Science***7**, (2016).

12. Zhao, Z. *et al.* Long Non-Coding RNAs: New Players in Plants. *International Journal of Molecular Sciences***23**, 9301 (2022).

13. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology***24**, 430–447 (2023).

14. Babraham Bioinformatics. *FastQC A Quality Control tool for High Throughput Sequence Data* http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

15. 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).

16. BBTools User Guide. *DOE Joint Genome Institute* https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/ (2016).

17. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research***49**, D192–D200 (2020).

18. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods***14**, 417–419 (2017).

19. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology***37**, 907–915 (2019).

20. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols***11**, 1650–1667 (2016).

21. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research***9**, 304 (2020).

22. 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).

23. Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Research***36**,

W5–W9 (2008).

24. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research***47**, D351–D360 (2018).

25. Kang, Y.-J. *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research***45**, W12–W16 (2017).

26. Fischer, S. *et al.* Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. *Current Protocols in Bioinformatics***35**, (2011).

27. Farrer, R. A. Synima: a Synteny imaging tool for annotated genome assemblies. *BMC Bioinformatics***18**, (2017).

28. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics***20**, 3643–3646 (2004).

29. Mattei, E., Pietrosanto, M., Ferrè, F. & Helmer-Citterich, M. Web-Beagle: a web server for the alignment of RNA secondary structures: Figure 1. *Nucleic Acids Research***43**, W493–W497 (2015).

30. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology***22**, 96–118 (2020).

31. Armaos, A., Colantoni, A., Proietti, G., Rupert, J. & Tartaglia, G. G. catRAPIDomics v2.0: going deeper and wider in the prediction of protein–RNA interactions. *Nucleic Acids Research***49**, W72–W79 (2021).

32. Zeng, M. *et al.* DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding. *Briefings in Bioinformatics***23**, (2021).

33. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nature Reviews Genetics***17**, 601–614 (2016).

34. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Research***43**, W39–W49 (2015).

35. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research***50**, D165–D173 (2021).

36. Ross, C. J. *et al.* Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. *Genome Biology***22**, 1–31 (2021).

37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology***15**, 1–21 (2014).

38. Trigiante, G., Blanes Ruiz, N. & Cerase, A. Emerging Roles of Repetitive and Repeat-Containing RNA in Nuclear and Chromatin Organization and Gene Expression. *Frontiers in cell and developmental biology***9**, 735527 (2021).

39. Repeat-associated RNA structure and aberrant splicing. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms***1862**,.

40. Swinnen, B., Robberecht, W. & Van Den Bosch, L. RNA toxicity in non-coding repeat expansion disorders. *The EMBO Journal***39**, (2019).

41. de Groot, N. S. *et al.* RNA structure drives interaction with proteins. *Nature Communications***10**, 1–13 (2019).

42. Oziolor, E., Arat, S. & Martin, M. Annotation depth confounds direct comparison of gene expression across species. *BMC Bioinformatics***22**, 1–15 (2021).

43. Ross, C. J. & Ulitsky, I. Discovering functional motifs in long noncoding RNAs. *Wiley Interdisciplinary Reviews: RNA***13**,.

Conclusion and future perspectives

Undoubtedly, lncRNA research is a fascinating field due to the multitude of functions this singular molecule has been associated with. Even though they do not code for proteins, less conserved across species, and were once dismissed as “junk DNA”, these molecules have emerged as crucial players in a wide range of biological processes, including chromatin modification, transcriptional regulation, RNA processing, post-transcriptional regulation and epigenetic regulation. This versatility helps explain why some scientists believe RNA may have preceded DNA in early life forms.

In this PhD research project, Chapter 3 aimed to develop a novel computational pipeline for robust identification of structurally identical lncRNAs across replicates in public bulk RNA-seq datasets. RNA-seq replicates are multiple measurements taken from the same biological sample, and they help assess the variability of biological systems. This variability can arise from subtle changes during sample preparation, biological fluctuations in gene expression, technical variations introduced during library preparation steps, or sequencing errors. Replicates are essential for obtaining reliable results and capturing both biological and technical variabilities, thereby increasing statistical power and reproducibility. Today, lncRNA identification and annotation are challenging tasks due to their low expression levels when compared their expressions to protein coding genes, and their myriad of isoforms they can express. Furthermore, the inner nature of lncRNAs make them difficult to distinguish from artifacts in RNA-seq data. With the use of this pipeline, we can increase confidence in lncRNA identification and annotation using bulk RNA-seq data by filtering out noise naturally found in RNA-seq analysis and focusing on consistent structural patterns in transcripts.

In this chapter, we also employed the new computational pipelined and explored the lncRNA diversity in the thermophilic fungi *Thermothelomyces thermophilus*, specifically their potential regulatory interactions between HSP, a family of proteins produced by cells in response to exposure to thermal stressful conditions to help refolding other proteins or target proteins for degradation when their conditions are beyond repair. Altogether, these proteins help the cell maintain its function under tough conditions. Additionally to their interactions with HSP, we characterized the landscape of these transcripts in the fungal organism under different temperature conditions, contributing to our understanding of the

diversity and dynamics of lncRNAs in response to environmental changes. Moreover, lncRNAs were classified into groups based on their expression patterns, a classification scheme which provided insights into the variability of lncRNA expression in this fungi. Along with that, correlation analysis identified correlations between lncRNAs and the cytochrome P450 stress-related protein family, a large and diverse superfamily of enzymes found in all kingdoms of life. These enzymes are linked to stress response pathways particularly heat stress and oxidative stress.

Additionally, it was demonstrated that the pipeline can also be used for the identification of structurally identical mRNAs and lncRNAs. The pipeline levares of consistent identification of the same transcripts across multiple replicates, ensuring the reliability and high-confidence of the downstream analysis and also reducing the impact of technical variability of sequencing artifacts.

The chapter 4 focuses on two closely related fungi, the *Chaetomium globosum*, a mesophilic fungi, and its thermophilic cousin *Thermothelomyces thermophilus*. These two fungi offer an unique opportunity to observe their adaptive evolutionary mechanisms at the molecular level on lncRNA repertories. The novel computational pipeline developed in the previous chapter was employed, adding a level of sophistication to the lncRNA identification and characterization. Our study characterized the lncRNA arsenal within the two fungi, trying to elucidate the role of lncRNAs in the fungi temperature adaptation once an it has been stated that genome reduction is a possible adaptation of the genomes of thermophilic fungi compared to related mesophiles. Genome reduction could have several potential implications for the organism's lncRNAs such as those lncRNA molecules that are found in intergenic regions, introns of protein-coding genes and also antisense which are transcribed from the opposite strand of protein-coding genes. With higher gene density in thermophilic fungi, the presence or absence of certain lncRNAs may be correlated with thermal adaptation strategies.

Comparative studies showed the thermophilic fungus has a significantly larger transcriptome compared to the mesophilic fungus, except for mesophilic intragenic long non-coding RNAs. Additionally, we have found intergenic, antisense, and intragenic lncRNAs in both fungi, but thermophilic fungus has a higher number of each type compared to the mesophilic fungus. Moreover, the comparison of alternative splicing

transcripts between mesophilic and thermophilic revealed approximately three times more isoforms in both mRNA and intergenic lncRNAs compared in the thermophilic fungus and absence of antisense isoforms in the mesophilic fungus, suggesting a potential adaptive mechanism in response to elevated temperatures.

GC content is considered a potential adaptive mechanism in thermophilic fungi and higher GC content is often associated with increased thermal stability of nucleotides. However, we found similar mRNA GC content distribution between the two fungi. Furthermore, the interquartile distributions of GC content in lncRNAs are comparable between the two fungi, with minor variations in the median for each distribution, showing that GC content may not be a major determinant of thermal adaptation. Notably, the median length of transcripts in the thermophilic fungus is longer than those found in the mesophilic fungus. Interestly, both thermophilic and mesophilic fungi exhibit conservation in the antisense and intergenic lncRNAs distribution of transcript lengths, suggesting that these transcript lengths in both fungi there may be under selective pressure pottentally to maintain specific regulatory functions across different environmental conditions.

By analyzing the Transcription Termination Site of protein-coding genes and the Transcription Start Site of intergenic long non-coding RNAs, our analysis indicated that the distance between these two transcripts is shorter in the mesophilic fungus compared to the thermophilic fungus. Similarly, the distance from the Transcription Termination Site of lncRNAs and the Transcription Start Site of protein-coding genes is also shorter in the mesophilic fungus, corroborating with the maintenance idea of intergenic lncRNA elements, even after genome reduction events in the thermophilic fungus, may indicate their functional significance in regulatory processes. Additionally, synteny studies between mesophilic and thermophilic revealed a significant overlap of protein-coding genes (6498 from 9292), being considered a high degree of conservation between the fungi, therefore it is expected that lncRNAs may play regulatory roles with conserved intergenic regions in thermophilic, potentially influencing its physiological adaptation to the thermal stress.

In addition, we analyzed the role of highly conserved RNA Polymerases (I, II and III) in both fungi, being each polymerase responsible for transcribing specific types of genes, such as ribosomal RNA genes, protein-coding genes, and genes encoding small

non-coding RNAs. Our results show that there are differences in RNA polymerase III (RNAP III) expression among thermophilic samples at different temperatures, contributing to our hypothesis of intergenic regulatory regions in thermophilic fungi may take part in its thermotolerance function in thermophilic fungi. Also, differences in the gene length of RNA polymerases between mesophilic and thermophilic fungi suggest evolutionary adaptations related to transcriptional activity and efficiency at high temperatures, potentially affecting regulatory elements within the genome, such as promoters and enhancers by facilitating transcriptional activity under different environmental conditions.

In summary, this PhD research project investigated the intricate world of lncRNAs in mesophilic and thermophilic fungi and how they contribute to fungal adaptation in response to temperature. Our study developed a new computational pipeline to identify structurally identical transcripts, enabling a precise identification and characterization of lncRNA transcripts amidst the genomic complexity and technical variability inherent in RNA-seq datasets. Moreover, the comparative analyses between mesophilic and thermophilic fungi unveiled significant differences in transcriptome sizes, isoform diversity, and expression patterns, shedding light on evolutionary strategies to cope with environmental stressors, unraveling the lncRNA functional significance in complex biological systems.

Finally, during this research project, an intriguing question emerged. Although lncRNAs are recognized for their significant roles in gene regulation, their conservation across species remains an area urging for new explorations. Therefore, our future work will focus on studying antisense lncRNAs overlapped by protein-coding genes. This new avenue of research will illuminate the evolutionary significance and functional consequences of lncRNA and protein-coding gene preservation and interactions.

REFERENCES

1. Krijt, S. *et al.* Chemical habitability: Supply and retention of life's essential elements during planet formation. *arXiv.org* https://arxiv.org/abs/2203.10056 (2022).

2. Cooper, G. M. *The cell: A molecular approach*. (Sinauer Associates, Incorporated, 2018).

3. Schwartz, W. M. Shilo (editor), strategies of microbial life in extreme environments (life science research report 13. Dahlem Konferenzen). 513 S., 42 abb., 24 tab. Weinheim-New York 1979. Verlag Chemie. DM 72,00. *Zeitschrift für allgemeine Mikrobiologie***21**, 270–270 (1981).

4. Speth, D. R. *et al.* Microbial communities of Auka hydrothermal sediments shed light on vent biogeography and the evolutionary history of thermophily. *The ISME Journal***16**, 1750–1764 (2022).

5. Canganella, F. & Wiegel, J. Extremophiles: From abyssal to terrestrial ecosystems and possibly beyond. *Naturwissenschaften***98**, 253–279 (2011).

6. Yakovchuk, P. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research***34**, 564–574 (2006).

7. Musto, H. *et al.* Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochemical and Biophysical Research Communications***347**, 1–3 (2006).

8. Nakashima, H. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *Journal of Biochemistry***133**, 507–513 (2003).

9. Nakashima, H. & Kuroda, Y. Differences in dinucleotide frequencies of thermophilic genes encoding water soluble and membrane proteins. *Journal of Zhejiang University SCIENCE B***12**, 419–427 (2011).

10. Sharma, A. *et al.* RNA thermometers and other regulatory elements: Diversity and importance in bacterial pathogenesis. *WIREs RNA***13**, (2022).

11. Meyer, M. M. Revisiting the relationships between genomic G + C content, RNA

secondary structures, and optimal growth temperature - PubMed. *Journal of molecular evolution***89**, (2021).

12. Hickey, D. A. & Singer, G. A. Genomic and proteomic adaptations to growth at high temperature. *Genome Biology***5**, 1–7 (2004).

13. Hurst, L. D. & Merchant, A. R. High guanine-cytosine content is not an adaptation to high temperature: A comparative analysis amongst prokaryotes. *Proceedings of the Royal Society B: Biological Sciences***268**, (2001).

14. Galtier, N. & Lobry, J. R. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution***44**, 632–636 (1997).

15. Meng, B. *et al.* Proteomic analysis on the temperature-dependent complexes in Thermoanaerobacter tengcongensis. *PROTEOMICS***9**, 3189–3200 (2009).

16. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. *PLOS Computational Biology***3**, (2007).

17. Perutz, M. F. & Raidt, H. Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature***255**, 256–259 (1975).

18. Vogt, G., Woell, S. & Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *Journal of Molecular Biology***269**, 631–643 (1997).

19. Thompson, M. J. & Eisenberg, D. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability - PubMed. *Journal of molecular biology***290**, (1999).

20. Kreil, D. P. & Ouzounis, C. A. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Research***29**, (2001).

21. Fukuchi, S. & Nishikawa, K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *Journal of Molecular Biology***309**, 835–843 (2001).

22. Kumar, S. & Nussinov, R. How do thermophilic proteins deal with heat? - PubMed. *Cellular and molecular life sciences : CMLS***58**, (2001).

23. Silva, R. G., Amaral, P. P., Franco, G. R. & Góes-Neto, A. Exploring the hidden hot world of long non-coding RNAs in thermophilic fungus using a robust computational pipeline. *Scientific Reports* **14**, 1–17 (2024).

24. Saha, D., Panda, A., Podder, S. & Ghosh, T. C. Overlapping genes: A new strategy of thermophilic stress tolerance in prokaryotes. *Extremophiles***19**, 345–353 (2014).

25. Chu, X.-L. *et al.* Temperature responses of mutation rate and mutational spectrum in an Escherichia coli strain and the correlation with metabolic rate. *BMC Evolutionary Biology***18**, 1–8 (2018).

26. Kantidze, O. L., Velichko, A. K., Luzhin, A. V. & Razin, S. V. Heat stress-induced DNA damage. *Acta Naturae***8**, (2016).

27. Cohen, G. N. *et al.* An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi - PubMed. *Molecular microbiology***47**, (2003).

28. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon Sulfolobus islandicus. *Molecular Biology and Evolution***22**, 2354–2361 (2005).

29. Mino, S. *et al.* Biogeography of Persephonella in deep-sea hydrothermal vents of the Western Pacific. *Frontiers in Microbiology***4**, (2013).

30. Blaby, I. K. *et al.* Experimental evolution of a facultative thermophile from a mesophilic ancestor. *Applied and Environmental Microbiology***78**, 144–155 (2012).

31. Dutta, A. & Chaudhuri, K. Analysis of tRNA composition and folding in psychrophilic, mesophilic and thermophilic genomes: Indications for thermal adaptation. *FEMS Microbiology Letters***305**, 100–108 (2010).

32. Wang, Q., Cen, Z. & Zhao, J. The survival mechanisms of thermophiles at high temperatures: An angle of omics. *Physiology***30**, 97–106 (2015).

33. Nicks, T. & Rahn-Lee, L. Inside out: Archaeal ectosymbionts suggest a second model of reduced-genome evolution. *Frontiers in Microbiology***8**, (2017).

34. Yoon, S. H. *et al.* Parallel evolution of transcriptome architecture during genome reorganization. *Genome Research***21**, 1892–1904 (2011).

35. Tiwari, S., Thakur, R. & Shankar, J. Role of heat-shock proteins in cellular function and in the biology of fungi. *Biotechnology Research International***2015**, 1–11 (2015).

36. Stephanou, A. & Latchman, D. S. Transcriptional modulation of heat-shock protein gene expression. *Biochemistry Research International***2011**, 1–8 (2011).

37. Roy, S. K. & Nakamoto, H. Cloning, Characterization, and Transcriptional Analysis of a Gene Encoding an α-Crystallin-Related, Small Heat Shock Protein from the Thermophilic Cyanobacterium Synechococcus vulcanus. *Journal of Bacteriology***180**, 3997–4001 (1998).

38. Chen, K. *et al.* Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proceedings of the National Academy of Sciences***114**, 11548–11553 (2017).

39. Wang, Z. *et al.* The Temperature Dependent Proteomic Analysis of Thermotoga maritima. *PLoS ONE***7**, e46463 (2012).

40. Caro-Quintero, A. & Konstantinidis, K. T. Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria. *The ISME Journal***9**, 958–967 (2014).

41. Maheshwari, R., Bharadwaj, G. & Bhat, M. K. Thermophilic fungi: Their physiology and enzymes. *Microbiology and Molecular Biology Reviews***64**, (2000).

42. Merino, N. *et al.* Living at the Extremes: Extremophiles and the limits of life in a planetary context. *Frontiers in Microbiology***10**, (2019).

43. Thanh, V. N. *et al.* Surveying of acid-tolerant thermophilic lignocellulolytic fungi in Vietnam reveals surprisingly high genetic diversity. *Scientific Reports***9**, (2019).

44. Salar, R. K. *Thermophilic fungi: Basic concepts and biotechnological applications*. (2018).

45. Niehaus, F., Bertoldo, C., Kähler, M. & Antranikian, G. Extremophiles as a source of novel enzymes for industrial application. *Applied Microbiology and Biotechnology***51**, 711–729 (1999).

46. Segal-Kischinevzky, C. *et al.* Yeasts inhabiting extreme environments and their biotechnological applications. *Microorganisms***10**, 794 (2022).

47. Wu, G. *et al.* Genus-Wide comparative genomics of malassezia delineates its phylogeny, physiology, and niche adaptation on human skin. *PLOS Genetics***11**, e1005614 (2015).

48. Tavares, S. *et al.* Genome size analyses of Pucciniales reveal the largest fungal genomes. *Frontiers in Plant Science***5**, (2014).

49. Berka, R. M. *et al.* Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris. *Nature Biotechnology***29**, 922–927 (2011).

50. Thapar, R. Regulation of DNA double-strand break repair by non-coding rnas. *Molecules***23**, 2789 (2018).

51. Di, C. *et al.* Characterization of stress-responsive lncRNAs inArabidopsis thalianaby integrating expression, epigenetic and structural features. *The Plant Journal***80**, 848–861 (2014).

52. Pirogov, S. A., Gvozdev, V. A. & Klenov, M. S. Long Noncoding RNAs and Stress Response in the Nucleolus. *Cells***8**, (2019).

53. Hu, X. *et al.* Identification and characterization of heat-responsive lncRNAs in maize inbred line CM1. *BMC Genomics***23**, (2022).

54. Song, X. *et al.* Comparative analysis of long noncoding RNAs in angiosperms and characterization of long noncoding RNAs in response to heat stress in Chinese cabbage. *Horticulture Research***8**, (2021).

55. Morris, K. V. *Long non-coding rnas in human disease*. (Springer, 2016).

56. Bruce, A. *et al.Molecular Biology of the Cell: Seventh international student edition with registration card*. (W.W. Norton & Company, 2022).

57. Amin, N., McGrath, A. & Chen, Y.-P. P. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence***1**, 246–256 (2019).

58. Mattick, J. S. *et al.* Long non-coding RNAs: Definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology* 1–17 (2023) doi:10.1038/s41580-022-00566-8.

59. Xu, J. *et al.* Third-generation sequencing found LncRNA associated with heat shock protein response to heat stress in Populus qiongdaoensis seedlings. *BMC Genomics***21**, (2020).

60. Dou, J. *et al.* Genome-wide identification and functional prediction of long non-coding RNAs in Sprague-Dawley rats during heat stress. *BMC Genomics***22**, (2021).

61. Wang, Z., Jiang, Y., Wu, H., Xie, X. & Huang, B. Genome-Wide Identification and Functional Prediction of Long Non-coding RNAs Involved in the Heat Stress Response in Metarhizium robertsii. *Frontiers in Microbiology***10**, (2019).

62. Singh, A. *et al.* Global Transcriptome Characterization and Assembly of the Thermophilic Ascomycete Chaetomium thermophilum. *Genes***12**, (2021).

63. 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 genomics***19**, 152 (2018).

64. Di, C. *et al.* Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. *The Plant Journal***80**, 848–861.

Paper 1 - Rebuttal letter

To the Assigned Editor of Scientific Reports.

Dear Editor,

We thank the Editor for having considered our work for publication in the Scientific Reports journal, and we would like to thank the Reviewers for their valuable comments. Herein, we provide a specific rebuttal for each point raised by the two Reviewers and the general comments of the In-house Editor.

Thank you for the update and for the opportunity to revise my manuscript. I appreciate the feedback provided by the reviewers and will start working on the revisions promptly. Please let me know if there are any specific guidelines or deadlines I should be aware of during this process.

In-house Editor's comment:
In-house editorial comment: To aid our readers, and to maximise the accessibility of your. manuscript, the title should have a clear, precise scientific meaning and should not contain a colon. Where possible, the title should be read as one concise sentence. Please could you re-write the title ensuring that it is informative and appropriate.

Edited as suggested.

Reviewer #1's comments
Major remarks:

- *While the authors are able to obtain some biologically interpretable results (most of which are not really new knowledge or could be expected from what is already known regarding lncRNAs biology), the focal point is the pipeline itself. The pipeline, however, raises a couple of concerns. Firstly, I don't think the authors have sufficiently convinced the reader that considering structurally-identical or similar transcripts is beneficial in a broader perspective – what is lacking is employing the pipeline in the other settings, beyond a single fungi species. Moreover, this approach seems to lead to a significant reduction in the number of lncRNAs being considered – how does this impact the biological side? There are also some flaws in the pipeline itself as well as in the manuscript (see below). Also, the authors haven't compared their pipeline with similar solutions. Finally, the pipeline's code itself is published (GitHub), but it is not really usable, as it lacks proper documentation and the code is adapted to the published study only – it is unclear how one could adapt it to own studies. Taking into consideration the*

overall scientific quality - the quality and impact of research - I believe the manuscript is not fit for publishing with Scientific Reports.

We thank the Reviewer #1 for the point raised regarding not having adequately demonstrated the broader benefits of considering structurally identical or similar transcripts, specifically, not showing evidence of the utility of our pipeline with different settings beyond a single fungal species.

Living organisms exhibit inherent variability. Therefore, replicated measurements are necessary[1] 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[2]. 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[3].

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. On the other hand, the appearance of a large number of transcripts with unexpected structures in the transcriptome might indicate technical issues during RNA-seq library preparation or sequencing itself[4]. Therefore, structurally identical transcripts act as a control group to help distinguish true biological variation within the replicates from technical artifacts.

Additionally, 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 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 isoform since these transcripts arise from the same gene but with variations in their structure due to alternative splicing. 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[5]. Non-structurally identical transcripts may violate this assumption if they exhibit different expression patterns or levels of variability between experimental conditions.

Furthermore, DESeq2 assumes that the transcriptome data follow 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[2]. 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.

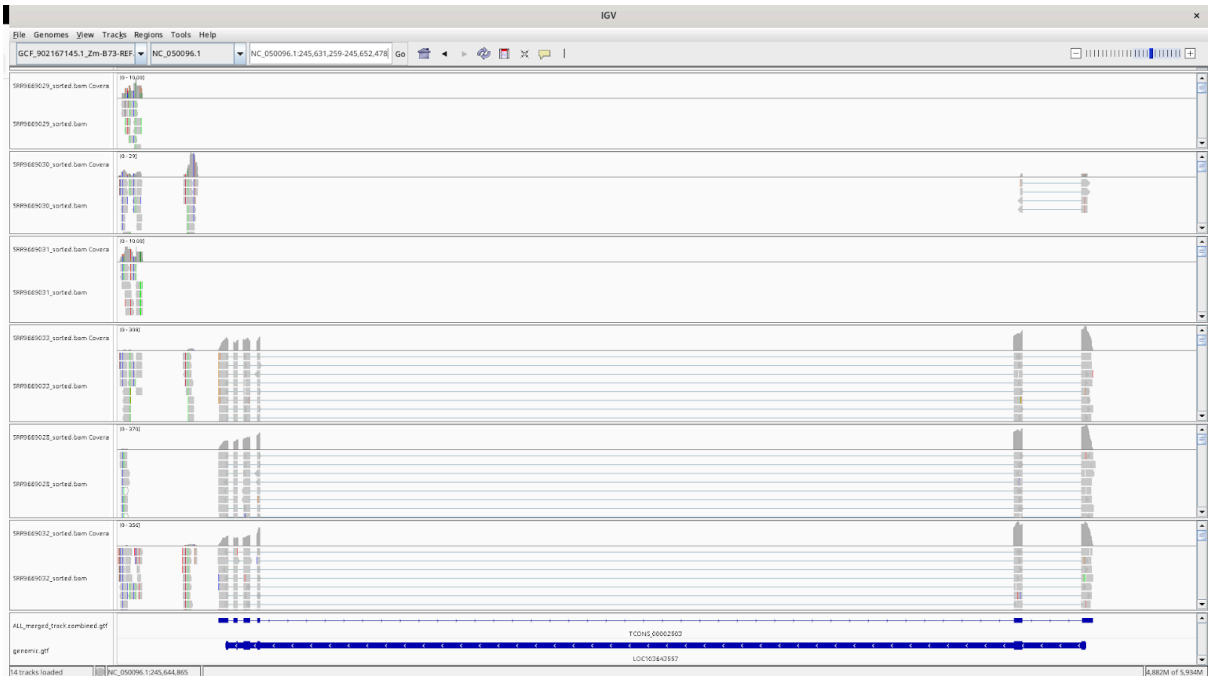
Long non-coding RNA transcripts (lncRNAs) pose a level of complexity in RNA-seq analysis due to their lower abundance when compared to mRNA transcripts: their heterogeneity of expression across different cell types, tissues, and species, their low sequence conservation, their multitude of functions, 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. HOTAIRM1 has different splice variants[6-9]. The HOTAIRM1-1 isoform (unspliced transcript) activates UTX/MLL and silences PRC2 complexes, and is required for conformational changes of the HOX cluster after retinoic acid induction[6]. On the other hand, the isoform HOTAIRM1-1 (3 exons) recruits PRC2 to regulate NEUROG2, with its function mainly in neuronal differentiation[7]. Myeloid differentiation requires HOTAIRM1-2 (2 exons) to regulate HOXA1 and HOXA4 expression[7,8]. The HOTAIRM1 (three variants, 1, 3, 5) acts as sponges of miRNA-20a/106b and miR-125b and activates ULK1, E2F1, DRAM2 genes, thereby regulating autophagy[9]. 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.

Additionally, prioritizing structurally identical transcripts 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 and mRNA transcripts.

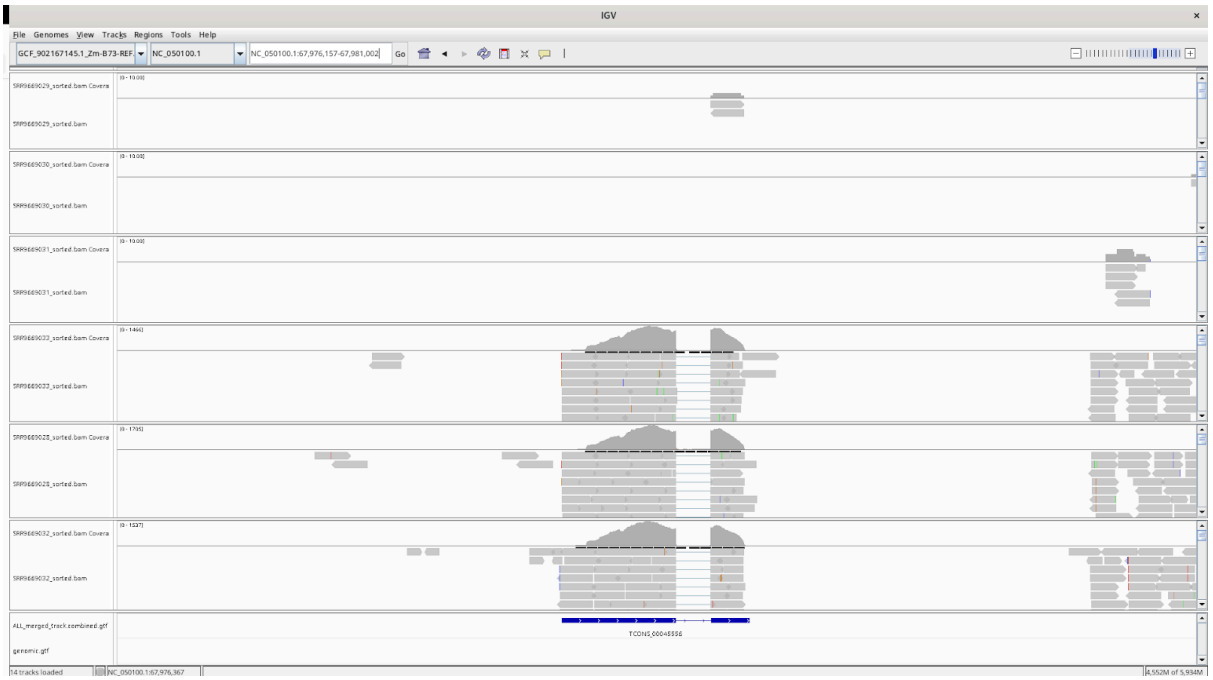
The *"significant reduction in the number of lncRNAs"* is the result of selecting only structurally identical transcripts. When planning and conducting an experiment with three biological replicates, the aim is to capture observed effects that represent genuine phenomena in organisms of the same type, treated or grown under identical conditions[10]. It is expected that the same set of genes will be expressed in each replication, thereby reducing the stochastic factor and increasing the statistical power in RNA-seq experiments. Excluding non-structurally identical transcripts improves the quality of RNA-seq data by reducing noise and ambiguity in transcript quantification, enhancing the accuracy and reliability of gene expression measurements, ultimately leading to more robust and interpretable results. Prioritizing canonical isoforms in the analysis increases the likelihood of capturing isoforms that contribute the most to the overall gene expression pattern and are most commonly observed in each replicate. This reduction not only refines the accuracy of variability estimation within the samples but also aids in distinguishing genuine biological signals from experimental noise and technical artifacts. Consequently, it enhances the reliability and interpretability of the results obtained.

As requested, we have applied our pipeline to a species beyond the fungi kingdom. Specifically, we have downloaded and processed BioProject PRJNA553580, which focuses on maize root inoculated with an AM fungus. Although this BioProject does not involve a fungus organism as requested, we chose it because the authors validated four differentially expressed lncRNAs using Real-Time PCR (see Paper supplementary Figure 1). We believe that fungal genomes have their own specificities, and while our pipeline was parameterized for them, the authors have provided lncRNA fasta files and utilized strand-specific RNA-sequencing libraries.

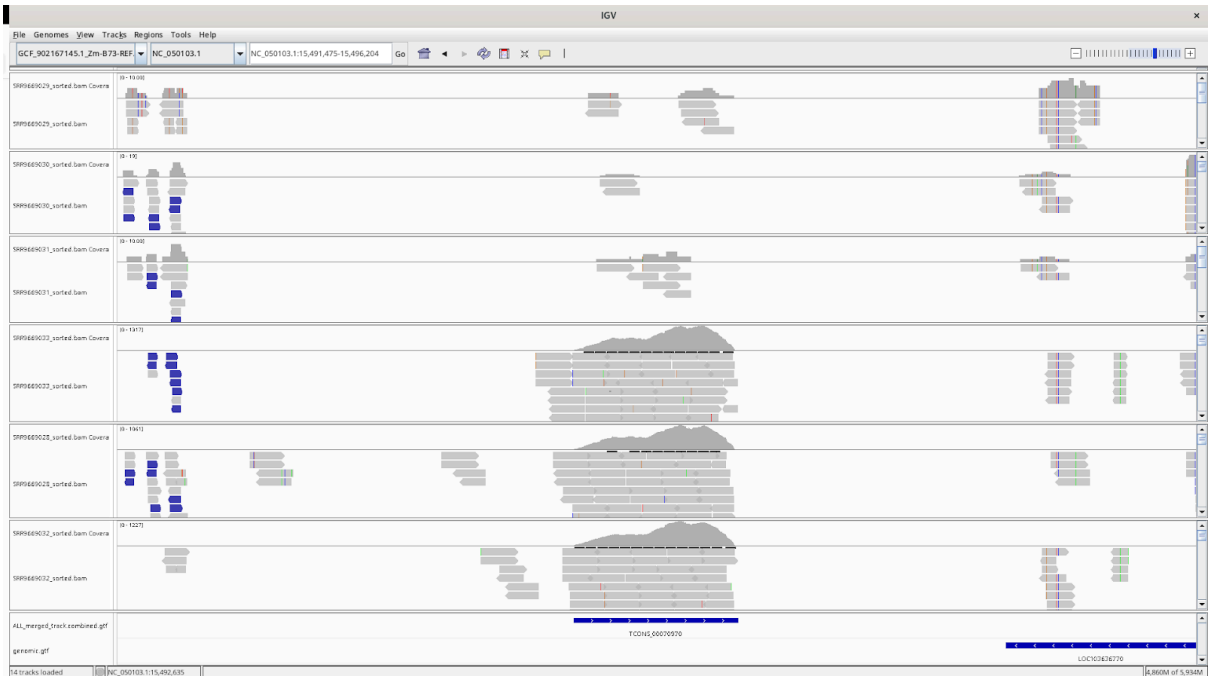
Paper Transcript ID: TCONS_00025975



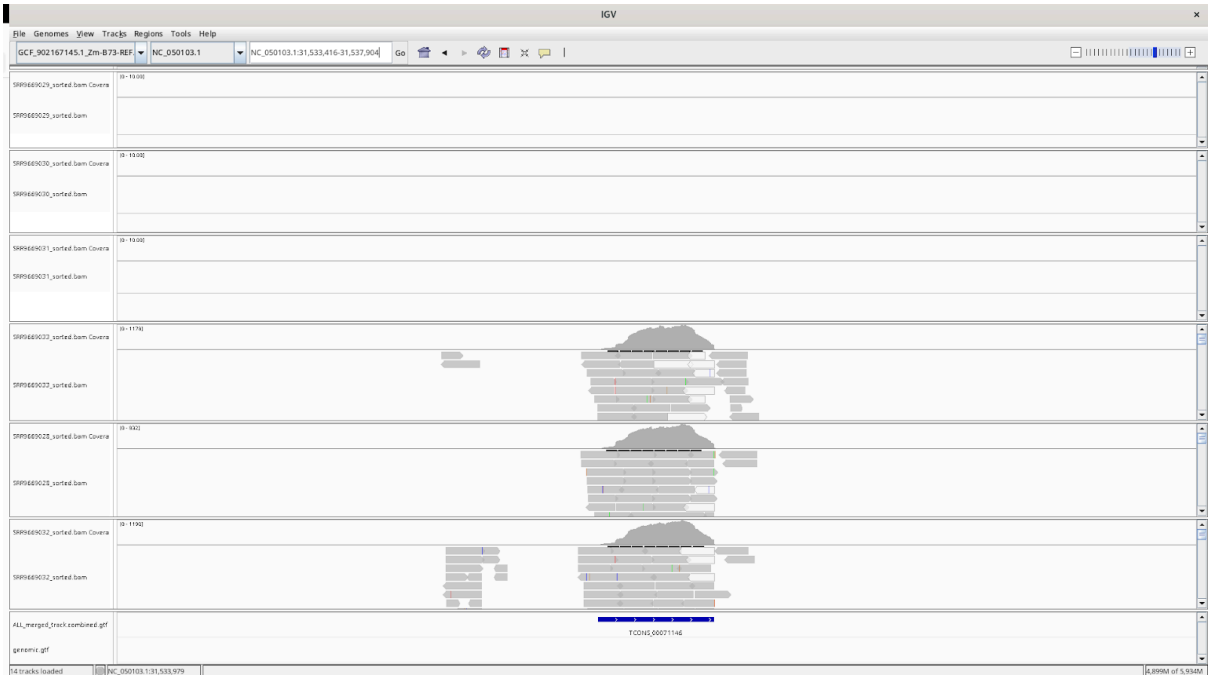
Paper Transcript ID: TCONS_00125081



Paper Transcript ID: TCONS_00165851


IGV
File Genomes View Tracks Regions Tools Help
GCF_902167145.1_Zm-B73-REF
NC_050103.1
NC_050103.1:15,491,475-15,496,204
Go
SRR6669029_sorted.bam Covera
SRR6669029_sorted.bam
SRR6669030_sorted.bam Covera
SRR6669030_sorted.bam
SRR6669031_sorted.bam Covera
SRR6669031_sorted.bam
SRR6669032_sorted.bam Covera
SRR6669032_sorted.bam
SRR6669028_sorted.bam Covera
SRR6669028_sorted.bam
SRR6669032_sorted.bam Covera
SRR6669032_sorted.bam
ALL_merged_track.combined.gtf
TCONS_00070970
genemir.gtf
LOC103626770
14 tracks loaded
NC_050103.1:15,492,635
4,860M of 5,934M

Paper Transcript ID: TCONS_00175390



As you can verify, our pipeline has catched all four Real-Time PCR validated lncRNAs. LncRNA fasta sequences from our transcriptome assembly can be found below along with the lncRNA ids from the paper.

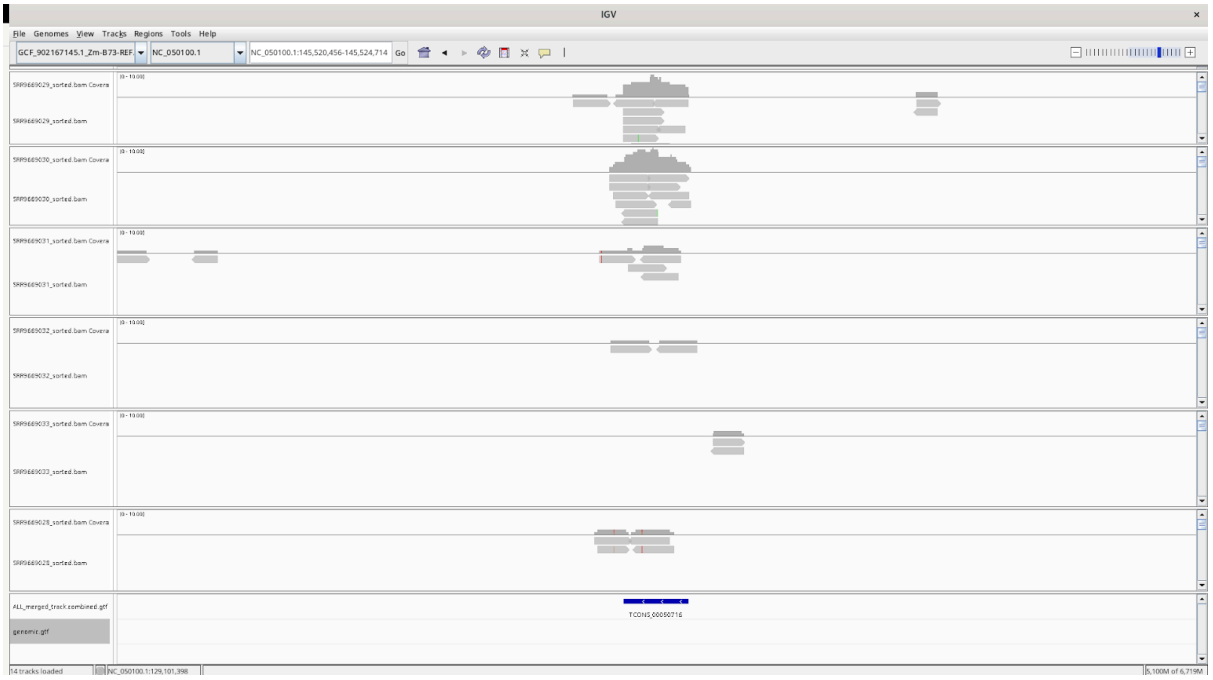
>TCONS_00025975 832bp (from paper)
>TCONS_00002503 tss_id=TSS3973;num_samples=3
GTCACTATCTTGACCCACAATCCTTTATTTATTGCAATTCCAACAGTAGTATGAAGTATTTCCATGTTTT
CTTTGATCTCGTGCAAAGACCGGAATAAAGACAATTACTTTACTGTAAAAAGGATAAGGAGACCAAGATC
TCTATATTGCAGTAACCAGTAGTAGTCTAGACGTTGGATTCCATGGCAATGGATAATCCTGCATTCTGCT
AGTACTAATAAAGACCATTTCCGACTGGTGCAGCCGGGTTTACCAGTCCTAGCCTGGCGCAGTCTGCCCC
AAACGTGCAGTTGCCGCCGATGCAAGGCATGGAGGTGAGGTTGAACATGTTTCTCAAGCCCGGCTGGCAG
CGGCACTGGTAACTGGAGCCGGCGCCCTTCTCGCAGTCACCTCCCTGGCCGCAATTGACGAGGGCGCAGG
GGTCTGTGAGAGGAATGCTCCGCATCGGCACGGTAGGGTTGAAGCAAGCACTGCTGAAGGTGCAGTTGGG
CACGATGCAGGGCATGAAGGGTAGGACCGTTAGGTTGAGCAACTTCGGCTGGGACCAGCCCGGGTCGCAG
GTGCACTTGTAGGAagtcgtgacctcggggatgatcCCCGGCACCTCACTGCAGGTGCCCTTGCCGCACT
CCGCCGTGTCGCAGATAGGGCTACTGTcggttgctgctgctgctgttcctGCTCTACCTCCTGCGAAGCC
AAGCAGCAGcaacatcaggagcagcatcacaTCTGGGCTAGCCATGGCCACCATCACCGTCTCGGCAAGC
TCCCGCTGCGATGTGCAGGTCAGTGCGCGTACGTTCGTGCGTTCTTCTAGCGTGGCTCTCTGACGGACGC
ACACGATCGAGCGACCCCACGACGACGACGATGCAGATGCAGTGCCGGCACCGGCCGGG

>TCONS_00125081 499bp **(from paper)**
>TCONS_00045556 tss_id=TSS35074;num_samples=3
TCCGAATCAGTCATTCTATAATAAATATTCAACCATTCAGTGGCGGATACAATATCGTGCGTGCCGAGGT
TGCAGCCGCGGTCGCCCCCAATAACACTTATACAATATCGCGCCCCAATAACACATGTATATATAATAAA
AGCTCACAAAGCTTAACTAGAACAGGTTACACCGTTACAACCTCTTTCTGACCACGACGTTCAGCAAATT
TGCGGATGTATGTAGTATCAACTAAACCAGCTGTTACCATTGTATAAATATTATTTTATACCACTATTAC
ACGGATGGGAGGCCGAGTAAGAGGATGGGATCGACGACATCACGGCGAGCTCAGCTAGCAGTTGGTGAAG
CACCAGCAACGAGGTGGGGAGTCAAAGCAAGCGCCGCCGACGTTGTTGCTGTCTTCCCCTTTGCATGTAT
CTAGGCAGGGGGTGTAGGAGGTGTCGCACAAGCCATGGTAGGAGCCGCTCAGGTGCCTGCAGGTGCATAT
GTCTGCCCCAACGGAGGCCATCTCGGTGGTGACGATGGCCGCGAGTAGAAGGAAGATAGCAGCGGTGGCA
GAGAGAGTCCTTGGTGACGGCGCCATTGGAGGATGCTTAATAGCCTTTTATTTACTGCAAGGTGGTATGT
TAATTGGAGACGAATGTAatggcctagtgaggagattgctGAATGAGTGAGTGATTCATCGA

>TCONS_00165851 755bp **(from paper)**
>TCONS_00070970 tss_id=TSS53453;num_samples=3
GTTGGAATaagggcaaaggtataaatgtatgtATCATAATTTATATGTACACCGTATCAGACTGAACGTT
AACTCGAAATATAGACCACGTAAACGTGACAAAAGCAATCCAAAGAAAACGCAGCTACTCTCGTCACACG
ACGTACAGGCATGCATGTGAGCATGTTTGATGTGTCTGATTCTGCATGCACGAGAACACGTATAGTTACC
TACTAGTAAAACGTTCCAGCAGCCACGTCATCACGTCCCAGTATTATTCATTGATCTAATAACACTAATA
TATATAGCTCACTAGCATCGCAAGCACACCGATCGAAGAAGAGACGCACGGACGGCTCCGATGAGTCGGT
GGTTGCATCGATTCGCGACGTTGCCATGCTGTGCTGGTGCTATATACGATACTACGGGGAATCGTCGGAC
GCCGGggcagcgtcgtcgtcgtcggacgCGCAGCAGTTCGGCCCCGGCCTGCAGCATCGGTCGCCGGGGT
CCTCGCCATCAGTCACCACCGCGTGTGGCAGAGCCGCCGGGAGGACGCTGGCTCTCAGCACAGCTCGCGG
CACGGCCGCAGCGCCGGCGCTCATCTCGGCGCCGGAGATGAGTATGACGCACACGAGAGCTACGACTAGC
AGTTTGGAGACCTGCATGCTGCCGCTTGTTATTTAATTTTCCGGTTACTTCTCACTCAGGATGGCCTGCC
TTTCTTTGCATGGACGAGTGCTGTGCTGAT

>TCONS_00175390 477bp **(from paper)**
>TCONS_00071146 tss_id=TSS53749;num_samples=3
TCACTCGTCAACACAAGACATCTAATTGTTTATAATTCAATATATTGAGTAGGACACCACTCAATACATA
TTCAACACGGACACTTAAAAGTATGAAATCAACAATGAAAGCAAGATCAGACATCACACTCGTATTATTC
AATTGATGCTTGCCTTTCATATATCAAGGCATAAAAAGGTGTAAAGGTAGCCTTTGCTAGGCAACCTAGA
GAATATATGGAATTTCTTCAACGGTTTCCACATTAGTGTTTGGGTCCACGCTGAAGTAGATGACAACACG
TATAGTTCCGGTTGCGAGTTGGTCAGCAGGCCTCGGGGTCTCGTTCAAATCCACATCTAGTGACTCAAAG
AATGTTACATCATCCCgttgttcttctatttttcaatTGCATCATTAATCTCCCAGCCAATCAACTCGG
GCCAACTGGTCTTGTCCGGCATAGACATGATCACGTACTATCCTTAGCTTTTCTCACTTTTGAAACTGT

Additionally, we have attached a screenshot of the IGV tool showing one lncRNA identified as intergenic by our pipeline, but without being structurally identical among two replicates only. If someone does not apply our pipeline, these two transcripts could be identified as belonging to all samples mistakenly without using our pipeline.



Finally, reviewer #2 suggested two papers to compare our pipeline with. The paper entitled "Identification and function annotation of long intervening noncoding RNAs"

developed a pipeline specifically for the identification of novel long intergenic non-coding RNAs (lincRNAs) and the prediction of their functions. Also, the paper "Plant-LncPipe: a computational pipeline providing significant improvement in plant lncRNA identification" retrained several models, including CPAT, PLEK, and LncFinder, using plant datasets.

Our pipeline is based on transcriptome data obtained from RNA-seq libraries, enabling it to compare real transcripts among biological replicates. This approach provides a more accurate representation of gene expression patterns compared to methods that do not use identical transcripts. Moreover, our pipeline uses publicly available and well-documented RNA analysis tools. This integration leverages established tools and algorithms, ensuring reliability and reproducibility in the analysis process. It also facilitates transparency and ease of use for researchers familiar with these tools.

We have addressed the premises of our pipeline and incorporated part of the comment into the Discussion section in the concluding remarks of the manuscript. Additionally, we have discussed some shortcomings of not using structurally identical transcripts during RNA-seq analysis. Therefore, we believe this pipeline should be utilized before any other downstream computational pipelines. The two aforementioned pipelines would yield better outcomes if complemented with our pipeline.

Minor remarks:

- *"Genome size appears to be another adaptive process that has occurred in these organisms". – genome size is not a process.*

  We thank the Reviewer for the suggestion. We amended the sentence to “Genome size appears to be another evolutionary adaptation in these organisms.”

- *"Recently, important roles in cellular responses to environmental stress have been assigned to long noncoding RNAs (lncRNAs)11-14. These include regulating double-strand DNA breaks11, differentially expressed lncRNAs under various stress stimuli12, regulating gene expression in response to stress conditions in the nucleolus13, and chromatin modification under low temperature14. All those functions support the stress environmental responses at the molecular and genomic levels within cells." - these are selected examples from plants and animals. How about fungi?"*

  We thank the Reviewer #1 for the point raised. We have changed the text and added another reference to cite the lncRNAs in fungi (lines 70-75).

- *“others are 5-capping, poly- or not adenylated”: language*

Edited as suggested.

- *"…converted to BAM format and indexed for efficient data retrieval and manipulation" – how exactly was it done (software, versions, parameters)*

  We have uploaded the script onto the github website with the commands used for converting the SAM file to the BAM file.

- *"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." – why such strict criteria were applied? How this relates to other studies?*

  When analyzing spliced reads, there might be a balance between sensitivity (capturing real splicing events) and specificity (avoiding false positives). Choosing lower thresholds (2-3 reads) but with a more relaxed False Discovery Rate (FDR), acknowledging the possibility of including some false positives. Or using a stricter FDR threshold between 5 and 10 reads aligned across a junction gives a reliable detection with higher confidence in identified junctions[12]. Moreover, some protocols exhibit reduced sensitivity when examining junctions covered by fewer than five reads[12].

- *"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 is actually "gap between transcripts to merge together (default: 250)"; why such a low value was applied, compared with default 250?*

  According to the StringTie manual, the -g parameter is "Minimum locus gap separation value. Reads that are mapped closer than this distance are merged together in the same processing bundle. Default: 50 (bp)". Additionally, we have chosen a lower gap value because of the reduction of genome size which is considered a hallmark of thermophilic organisms, including this fungi, as cited in the manuscript.

- *"The merged meta-assembly FASTA file produced by the gffread was evaluated by the rnaQUAST tool33 according to the reference genome and its annotation file." –since there is a reference annotations file, was it used to guide the transcriptome assembly with StringTie?*

  Yes. StringTie uses the reference annotation file as a guide to assembly and tag the genes with names from that file[17].

- *How exactly was rnaQUAST implemented?*

  The implementation of rnaQUAST can be found in the already uploaded script. As reference, the command line used in rnaQUAST was: rnaQUAST.py --transcripts {param['fasta']} --reference {param['geno']} --gtf {param['gff']} -o {param['output']}

- *"Read counts below 10 were removed from the analysis before executing the DESeq2 analysis" – this is not a standard procedure, please justify*

  DESeq2 used with default values already applies an independent filtering in the analysis of differentially expressed genes. This function discards low count reads depending on the distribution of small p-values over the filter statistic. Generally, this method is less stringent than applying a hard cutoff on read counts directly once the independent filtering takes into account the overall data distribution and avoids arbitrarily discarding potentially informative genes with slightly lower counts. Applying a low count filter before DESeq2 analysis can remove genes with very low expression levels that are unlikely to be biologically relevant[15]. This can further reduce noise in the data and potentially improve the signal-to-noise ratio. Furthermore, focusing on genes with a higher probability of having reliable expression estimates potentially leads to more robust results.

- *logFC2 |1.5|. – this is unclear.*

  Edited as suggested.

- *"Principal component analysis (PCA) and hierarchical clustering using the DESEq2 package between groups were also tested."*

  Edited as suggested.

- *specie. – a typo*

  Edited as suggested.

- *Instead of STRING, it would be much better to annotate the transcriptome (e.g. with Trinotate) and try other approaches.*

  We appreciate the Reviewer #1 suggestion to explorer Trinotate for transcriptome annotation. Our primary interest was in understanding potential interactions between lncRNAs in conjunction with Heat Shock Proteins (HSP). STRING provides a vast

amount of interaction data from various sources, including protein-protein interactions, protein-chemical interactions, genetic interactions, and co-expression data with a user-friendly interface with well literature-supported information. We are always open to exploring alternative approaches, and Trinotate could be a valuable tool for future projects.

- *"For the downstream analysis, only transcripts that were structurally identical across all replicates in each experiment (35°C, 40°C, 45°C and 50°C) were selected and used for filtering lncRNAs and HSP from the curated reads." - since DE analysis was at the gene level, how was information about structurally identical transcripts (not genes) was integrated within this analysis?*

  Thank the Reviewer #1 for this insightful comment. We initially filtered structurally identical transcripts across all replicates within each temperature condition (35°C, 40°C, 45°C, and 50°C). Once the filtering step identified reliable transcripts, we linked them back to their corresponding genes. With the transcripts linked back to genes, we could then proceed with DE analysis using DESeq2 at the gene level. We have made changes to the manuscript in order to provide clarity about this procedure.

- *"For downstream analysis of lncRNA sequences, the classes "u", "x", and "i" were selected, and these codes represent intergenic, antisense, and intragenic transcripts respectively." → the class code "i" denotes intronic transcripts, not intragenic ones. Besides, it is unclear why only these class codes were accepted. Why exact matches ("=") or alternative splicing isoforms ("j") were absolutely excluded from considerations? Were the reference genes 100% protein coding?*

  We would like to extend our appreciation to Reviewer #1 for bringing the intragenic/intronic typo to our attention. It has been corrected in the manuscript. Also, we bring to your attention that the class code "=" or exact matches was depicted on Figure 3 and summarized on 454-455, 477-478, 527-529, 705-708 lines and it was also quantified on Supplementary Table 8. Regarding the alternative splicing isoforms ("j") transcripts, those transcripts were not analyzed because they might be alternative splicing isoforms of protein-coding genes rather than lncRNAs[16].

- *"The gffcompare, when executed with -r option, classifies transcripts based on their position within the reference genome." - what transcriptome annotations were used as a reference?*

  We used the reference genome/annotation files in the analysis. We added this information in the section material and methods, subsection Computational pipeline, lines 144-146.

- *"Antisense lncRNA sequences were processed before executing BLASTx because those lncRNA 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." – why not simply filter blastx results taking into consideration the relative orientation of query and subject (+/+ vs +/-)?*

  While filtering BLASTx results based on orientation (+/+ vs. +/-) seems a viable option, trimming the overlapping protein-coding portion from antisense lncRNAs before BLASTx might be more efficient and accurate because BLASTx searches for protein-like sequences in a nucleotide database. Even with orientation filtering, antisense lncRNAs might still generate partial alignments with protein-coding sequences on the opposite strand due to overlapping regions. Moreover, trimming the protein-coding gene overlapping portion reduces the search space for the unique portions of antisense lncRNA in BLASTx as well as the number of irrelevant comparisons against overlapping protein-coding regions.

- *Blastx is not enough, because it does not take into consideration the actual ORFs, so one would expect huge numbers of false positives, especially that relatively loose threshold for E-value (1e-3) was applied. Instead, one should look for ORFs and encoded proteins and then use blastp agains protein databases, e.g. SwissProt.*

  We thank Reviewer #1 for raising this point. The Blastx tool is used to compare a protein query sequence against a protein sequence database. It converts a nucleotide sequence into a protein by translating it in all six possible reading frames. Open reading frame (ORF) analysis is done by the CPC2 tool, which utilizes four features for classifying non-coding transcripts. These features are: the Fickett TESTCODE score, open reading frame (ORF) length, ORF integrity, and isoelectric point (pI). Therefore, ORF characteristics were taken into account when putative lncRNA sequences were filtered out. Weak or short ORFs with low conservation are unlikely to be protein[21]. Apart from that, the reduction of false positive lncRNAs can be achieved by applying a combination of in-silico approaches, as it was done in the pipeline. Additionally, the Blastx tool has been used extensively for classified transcripts as lncRNAs[18-20].

Reviewer #2's comments

Major comments:

1. *The figures need to be improved. Such as ImageGP (https://doi.org/10.1002/imt2.5) can generate high quality figures and with reproducible scripts.*

We have plotted all images in higher resolution as requested.

2. *The structure of the results appears not in concise. The authors should have presented their findings in around 3 mainly sections will be better.*

   We have edited our paper to mainly 3 sections as requested.

3. *MATERIAL AND METHODS-Transcriptome data. The authors used data from the SRP336977 (SRR15886248-SRR15886259) project, but this is the poly-A cDNA RNA-seq libraries. It is well-established in the field that lncRNAs are generally not polyadenylated, and as such, poly(A) RNA sequencing may not accurately represent the lncRNA landscape. The standard approach for lncRNA discovery and analysis should involve the use of rRNA-depleted, strand-specific libraries, which provide a more complete and unbiased view of the transcriptome, including non-polyadenylated species (https://doi.org/10.1186/s12864-016-2365-3; Doi:10.1016/j.ygeno.2019.11.009; Doi:10.3389/FPLS.2023.1118011). In the absence of strand-specific libraries, lncRNAs on antisense cannot be obtained. It is obviously inappropriate to develop an analysis process with unsuitable data, and the authenticity of lncRNA obtained under these conditions was not verified experimentally in this paper. Therefore, the reviewers believe that the design of the paper is flawed.*

   The reviewer #2 raised a valid point. Poly(A) RNA-seq libraries primarily capture polyadenylated transcripts, which are mainly mRNAs. However, many other lncRNAs are polyadenylated and exhibit mRNA-like characteristics[23]. Furthermore, many lncRNAs are localized in the cytoplasm and are associated with ribosomes through 5′ untranslated regions[24-26]. These rRNA-depletion methods could remove these transcripts or despite the researcher efforts to deplete rRNA, some rRNA molecules may remain in the sequencing library, which could affect the sensitivity of lncRNA detection. Moreover, it has been reported that lncRNAs can be transcribed by polymerase I, II, and III, resulting in lncRNA transcripts with specific traits. Thus, creating a computational pipeline to address a specific class of lncRNA based on the RNA kit extraction the researcher had used would impose significant limitations on the lncRNA landscape.

   Additionally, our computational pipeline is highly parameterizable and well documented. These features provide users flexibility to change the script to their specific transcriptome analysis needs, aiming to explore different experimental conditions, study diverse biological systems, or focus on specific transcript types. The pipeline facilitates researchers to adapt the pipeline according to their objectives with ease and confidence, still retaining the ability of tracking structurally identical transcripts.

We have implemented a decontamination step in the pipeline. The decontamination process using Kmers as probes was used to identify and remove contaminant sequences from the transcriptome data. For that, we have used the BBDuk tool from the BBMap package and have set the k-mers parameter to 31-mers.

4. *Lack of Rigorous Validation: Even if poly(A) data were appropriate for lncRNA analysis, the manuscript does not provide sufficient validation of the identified lncRNAs. Validation steps should include, but are not limited to, experimental verification of the lncRNA candidates, such as Northern blotting or RT-qPCR, and functional assays to provide evidence of biological significance.*

   We have answered this question in the reviewer #1 major comment.

5. *The researchers developed a robust computational process to address this problem, aiming to identify lncrnas with the same structure across different replication samples to ensure the reliability and accuracy of RNA studies in thermophilic fungi. That with the existing lncRNA process (doi:10.1093/bib/bbw046, https://doi.org/10.1093/hr/uhae041) compared to do what improvements, had better have benchmarking analysis*

   We appreciate the suggestions from Reviewer #2 to compare our pipeline with others.
   The paper entitled "Identification and function annotation of long intervening noncoding RNAs" developed a pipeline specifically for the identification of novel long intergenic non-coding RNAs (lincRNAs) and the prediction of their functions. While the paper "Plant-LncPipe: a computational pipeline providing significant improvement in plant lncRNA identification" retrained several models, including CPAT, PLEK, and LncFinder, using plant datasets.

   Our pipeline is based on transcriptome data obtained from RNA-seq libraries, enabling it to compare real transcripts among biological replicates. This approach provides a more accurate representation of gene expression patterns compared to methods that do not use identical transcripts. Moreover, our pipeline uses publicly available and well-documented RNA analysis tools. This integration leverages established tools and algorithms, ensuring reliability and reproducibility in the analysis process. It also facilitates transparency and ease of use for researchers familiar with these tools.

   In response to Reviewer #1's major remark, we have addressed the premises of our pipeline and incorporated part of the comment into the Discussion section in the concluding remarks of the manuscript. Additionally, we have discussed some shortcomings of not using structurally identical transcripts during RNA-seq analysis. Therefore, we believe this pipeline should be utilized before any other downstream

computational pipelines. The two aforementioned pipelines would yield better outcomes if complemented with our pipeline.

6. *The software, tested data and results are required to be uploaded on GitHub for peers to use, and conda and/or docker installation modes are recommended for software with complex dependencies. We will take software Star, Fork, and downloads of GitHub as one of the audience indicators. Software installation and User tutorial are required in Readme.md or Wiki in GitHub.*

   We have uploaded the software, tested data and the results to Github as well as documented each step/method.

7. *A video of software download, installation, operation, and result display is required with a computer or server without any related software installed, to make sure that any new user can perform the whole process according to the tutorial.*

   We have written in detail how to use the pipeline.

Minor comments:

1. *In the Introduction, a large number of Thermophilic biological problems are introduced in the Introduction, which is too much, and it is recommended to simplify.*

   Edited as suggested.

2. *In the Discussion, there is a lack of critical examination of how the findings align or contrast with the field of lncRNA research.*

   We thank Reviewer #2 for this suggestion. We have made amendments to the manuscript and added a subsection in the discussion section.

3. *The stringtie2 step needs to filter transcripts of single exons, otherwise there will be a large number of false positives.*

   We have decided not to filter single-exon transcripts because a portion of bona fide lncRNAs are indeed single-exonic, and filtering them out could lead to the exclusion of genuine non-coding RNAs from our analysis. Recently, the RNA Atlas[22] has applied a very stringent sequencing and pipeline in the human transcriptome from 300 human tissues and cell lines and shown that out of 5471 lncRNAs, 89% were single-exon genes rather than contaminating DNA. Therefore, we have chosen not to exclude these transcripts, as such exclusion might cause the analysis to miss potentially interesting evolutionary insights or conserved functional elements.

4. *About "LncRNA identification and coding potential: the classes "u", "x", and "i" were selected". Authors should show the number of all transcripts classified, and the proportion of these.*

   We thank Reviewer #2 for this suggestion. We have made amendments to the manuscript and added this relevant information as Supplementary Table 8, referring to it in the text (lines 463-464).

5. *The transcript is marked with a "j" class code, indicating that at least one intron of the transcript is the same as the intron of the annotated gene, while others may be different, thus suggesting that such a transcript may be a novel isoform of the annotated gene. In addition, the classification of "i,o,u,x" conforms to the characteristics of lncRNA and can be used for lncRNA recognition. Thus, the five classes of transcripts "i,j,o,u,x" indicate possible new transcripts. The author needs to explain why only i u x is used.*

   We have opted to analyze only the transcript class code "x", "i", "u" because it has been suggested that transcripts of the 'c', 'k', 'j', 'm', 'n' or 'o' class may represent new

isoforms of known genes rather than lncRNAs[16]. However, the pipeline is easily parameterized to access other structurally identical transcripts, including any transcript classified by StringTie tool.

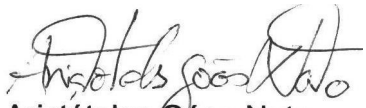
6. *"as small RNA (< 200 nt) and long non-coding RNA (> 200 nt), with other subsets within this group"(In the introduction). One is... 200, the other should be ≥200.*

   Edited as suggested.

7. *The manuscript could improve by providing a more comprehensive review of existing studies on lncRNAs in thermophilic organisms, particularly how these findings compare or contrast with previous research, to better situate its contributions within the existing body of knowledge.*

   We thank reviewer #2 for this valuable suggestion. Although it would be interesting to provide a more comprehensive review of existing studies on lncRNAs in thermophilic organisms, we believe that thermophilic bacteria and fungi respond to temperature stress differently. For example, while thermophilic fungi can thrive at temperatures up to 60°C, there are some bacteria surviving at temperatures usually above 80°C. Therefore, a review about thermophilic organisms would be out of the scope of this paper. However, we have added (lines 103-110) information about the only study on thermophilic fungi in the paper to better situate the readers and the studies that have been done on thermophilic fungi.

Sincerely (on behalf of all the authors),



Prof. Dr. Aristóteles Góes-Neto
Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, Brazil, CEP 31270-901
E-mails: arigoesneto@icb.ufmg.br / arigoesneto@gmail.com
CV: http://lattes.cnpq.br/6134133834289438
ORCID: 0000-0002-7692-6243

1. Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nature* https://www.nature.com/articles/nmeth.3091 (2014).

2. 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).

3. Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics***6**, 451–464 (2005).

4. Varabyou, A., Salzberg, S. L. & Pertea, M. Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments. *Genome research***31**, 301–308 (2021).

5. Kim, Y. J. & Cribbie, R. A. ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology***71**, 1–12.

6. Wang, X. Q. D. & Dostie, J. Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation. *Nucleic acids research***45**, 1091–1104 (2017).

7. Rea, J. *et al.* HOTAIRM1 regulates neuronal differentiation by modulating NEUROGENIN 2 and the downstream neurogenic cascade. *Cell Death & Disease***11**, 1–15 (2020).

8. Zhang, X. *et al.* A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood***113**, 2526–34 (2009).

9. 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 &*

*Differentiation***24**, 212–224 (2016).

10. Bell, G. Replicates and repeats. *BMC Biology***14**, 1–2 (2016).

11. Han, G. *et al.* Identification of Long Non-Coding RNAs and the Regulatory Network Responsive to Arbuscular Mycorrhizal Fungi Colonization in Maize Roots. *International journal of molecular sciences***20**, 4491 (2019).

12. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods***10**, 1185–1191 (2013).

13. Patel, H. & Rawat, S. Thermophilic fungi: Diversity, physiology, genetics, and applications. in *New and Future Developments in Microbial Biotechnology and Bioengineering* 69–93 (Elsevier, 2021).

14. Berka, R. M. *et al.* Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris. *Nature Biotechnology***29**, 922–927 (2011).

15. Ying Sha, Phan, J. H. & Wang, M. D. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, 2015).

16. Sheng, L., Ye, L., Zhang, D., Cawthorn, W. P. & Xu, B. New Insights Into the Long Non-coding RNA SRA: Physiological Functions and Mechanisms of Action. *Frontiers in medicine***5**, 244 (2018).

17. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols***11**, 1650–1667 (2016).

18. Nabi, A., Dilekoglu, B., Adebali, O. & Tastan, O. Discovering misannotated lncRNAs using deep learning training dynamics. *Bioinformatics***39**, (2022).

19. Shukla, B. *et al.* lncRNADetector: a bioinformatics pipeline for long non-coding RNA identification and MAPslnc: a repository of medicinal and aromatic plant lncRNAs. *RNA Biology***18**, 2290–2295 (2021).

20. Jannesar, M. *et al.* A genome-wide identification, characterization and functional analysis of salt-related long non-coding RNAs in non-model plant Pistacia vera L. using transcriptome high throughput sequencing. *Scientific Reports***10**, 1–23 (2020).

21. Kang, Y.-J. *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research***45**, W12–W16 (2017).

22. Lorenzi, L. *et al.* The RNA Atlas expands the catalog of human non-coding RNAs. *Nature Biotechnology***39**, 1453–1465 (2021).

23. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology***24**, 430–447 (2023).

24. van Heesch, S. *et al.* Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biology***15**, 1–12 (2014).

25. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology***22**, 96–118 (2020).

26. Carlevaro-Fita, J., Rahim, A., Guigó, R., Vardy, L. A. & Johnson, R. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA (New York, N.Y.)***22**, 867–82 (2016).