Global Transcriptome Characterization and Assembly of Thermophilic Ascomycete *Chaetomium thermophilum*

Amit Singh[1*] and Géza Schermann[1*], Sven Reislöhner[1], Nikola Kellner[1], Ed Hurt[1,2], Michael Brunner[1,2]

[1] Heidelberg University Biochemistry Center, Im Neuenheimer Feld 328, D-69120 Heidelberg, Germany
[2] Corresponding authors: ed.hurt@bzh.uni-heidelberg.de, michael.brunner@bzh.uni-heidelberg.de
[*] Contributed equally

A correct genome annotation is fundamental for research in the field of molecular and structural biology. The annotation of the reference genome *Chaetomium thermophilum* has been reported previously, but it is limited to open reading frames (ORFs) of genes and contains only a few noncoding transcripts. In this study, we identified and annotated by deep RNA sequencing full-length transcripts of *C.thermophilum.* We annotated 7044 coding genes and a large number of noncoding genes (n=4567). Astonishingly, 23% of the coding genes are alternatively spliced. We identified 679 novel coding genes and corrected the structural organization of more than 50% of the previously annotated genes. Furthermore, we substantially extended the Gene Ontology (GO) and Enzyme Commission (EC) lists, which provide comprehensive search tools for potential industrial applications and basic research. The identified novel transcripts and improved annotation will help understanding the gene regulatory landscape in *C.thermophilum*. The analysis pipeline developed here can be used to build transcriptome assemblies and identify coding and noncoding RNAs of other species. The R packages for gene and GO annotation database can be found under https://www.bzh.uni-heidelberg.de/brunner/Chaetomium_thermophilum.

Introduction

The *Chaetomium thermophilum* is a thermophilic filamentous ascomycete, with the ability to grow at 50-52ºC from degrading plant material (1). *C.thermophilum* produces, owned by it's lingocellulolytic life style, different thermostable enzymes such as cellulase, xylanase, laccase, chitinases and proteases (2–6). The thermostability of these enzymes makes *C.thermophilum* a model organism of choice in various pharmaceutical and food processing industries. In the past decade *C.thermophilum* has attracted the attention for different applications such as starch degradation, hydrolysis of cellulose for bioethanol production as well as other applications requiring enzymatic activities at higher temperatures (6–14). Additionally, many *C.thermophilum* proteins and complex protein assemblies have been reported in highest resolutions in various crystallisation and cryo-electron microscopy studies, which improved our understanding of the structural organization and function of protein complexes. These include the Crm1 export factor, the splicing factor Cwc27, mRNA export factor Mex67-Mtr2, the FACT complex, the eukaryotic RAC chaperone, the nuclear pore Nsp1-channel complex, and the 90S pre-ribosomal complex (15–22). The genome of *C.thermophilum* was first reported by (23) and afterwards substantially improved in its annotation by (24). It has a size of 28.3 Mb and assembled into 20 scaffolds containing 7165 protein coding and 387 noncoding transcripts. Despite of recent advances in sequencing technologies (25–28) the *C.thermophilum* genome has not been substantially upgraded. The reported *C.thermophilum* genome lacks proper annotation of untranslated regions (UTRs), and the majority of intron-exon structures are computationally predicted rather than experimentally determined. Given the substantial increase in the number of genomic studies on *C.thermophilum* a comprehensive genome annotation will be helpful for further functional, structural, proteomic, genomic and transcriptomic analyses. In this study, we present an improved annotation of the *C.thermophilum* genome based on deep RNA sequencing and establish pipeline tools for the analysis of sequencing data. Our annotation identified 7044 expressed protein-coding genes and 4567 long noncoding RNAs (lncRNAs). Moreover, our pipeline detected UTRs and intron-exon boundaries as well as transcript isoforms. Sequence homology studies revealed that *C.thermophilum* and *Thermothelomyces thermophila* share close sequence similarity of coding transcripts. The downstream analysis of genomic and transcriptomic sequence data is often used to predict the function of genes, identification of biomarkers, grouping and classifying gene expression patterns. Therefore, we present an extended Gene Ontology (GO) and Enzyme Commission (EC) numbers that are associated with protein coding genes of *C.thermophilum*.

MATERIALS AND METHODS

RNA Isolation and sequencing

The *C.thermophilum* strain was received from DSMZ, Braunschweig, Germany (No. 1495). The mycelium scraped off a freshly grown agar plate were grown in 250 ml baffled Erlenmeyer flasks at 52ºC for 6 h. Mycelium was grown in modified CCM medium as described in (29): 0.5 g NaCl, 0.65 g $K_2HPO_4 \cdot 3H_2O$, 0.5 g $MgSO_4 \cdot 7H_2O$, 0.01 g Fe(III)-sulfatehydrate, 8 g D-glucose and 1 g each of peptone and yeast extract per liter H_2O, pH 7.0. *C.thermophilum* mycelia were harvested through a sieve and grounded to a fine powder in liquid nitrogen. About 100 mg mycelia from three independent biological replicates were used directly for total RNA-extraction using the SV total RNA isolation system (Promega). The libraries were prepared with the NEBNext Ultra II Directional RNA Preparation Kit for Illumina in combination with NEBNext Poly A selection Module, plus the NEBNext Multiplex Oligos for Illumina, and sequencing was performed by CellNetworks Deep Sequencing Core Facility (Heidelberg, Germany) on Illumina NextSeq 500 platform (single-end).

RNA sequencing data analysis

The data quality assessment of raw sequence data was performed by FastQC (Version: FastQC 0.11.5, (http://www.bioinformatics. babraham.ac.uk/projects/fastqc/). No samples were discarded from the analysis. *C.thermophilum* reference genome and gene annotation files were downloaded from ENSEMBLE (https://fungi.ensembl.org) and a pipeline was developed to annotate UTR region and to identify the putative novel transcripts (Figure -1A). The raw reads were mapped to *C.thermophilum* genome using

HISAT2 with the following parameters (Version: 2.1.0; [hisat2 -p 8 -x -max-intronlen 2000 -dta-U])(30). The mapped reads from HISAT2 for each sample were assembled separately using StringTie with parameter settings (Version: 1.3.3b; [stringtie -o -m 50 -p 8 -j 3 -c 5 -g 15])(31). The multiple transcript assembly files from the different samples were used together to produce a distinctive transcriptome set using gffcompare with parameter settings (Version: v0.10.1; [gffcompare-merge -K -o gffcomp -i]) (25). Based on the previous assembly results, transcripts shorter than 200 nt were excluded to identify transcripts from the merged transcript assembly. According to the gffcompare class codes "i", "u","y" and "x" are considered as novel transcriptional loci. The coding potential calculator (CPC2) was used to evaluate the coding potential of all transcripts (32).

Sequence conservation and GO annotation

The sequence conservation analysis was performed using dc-mega BLAST (Version: 2.7.1+) (33). All coding and noncoding transcripts including the identified novel transcripts sequences were used for this analysis. *Sordaria macrospora* (NCBI taxid-5147), *Neurospora crassa* (NCBI taxid-5141), *Aspergillus niger* (NCBI taxid-5061), *Saccharomyces cerevisiae* (NCBI taxid-4932), *Takifugu rubripes* (NCBI taxid-31033) and *Thermothelomyces thermophila* (NCBI taxid-78579) were chosen to study the sequence similarity analysis using BLAST (E value,1e^{-3}). phyloT (https://phylot.biobyte.de) was used for the construction and visualization of phylogenetic tree of the above mentioned species. Additionally, functional annotation of *C.thermophilum* transcripts were analyzed using Blast2GO [Version 5.1.1] (34) as described in the manual. The annotated GO terms from *Thermothelomyces thermophila*, *Neurospora crassa* and *Sordaria macrospora* were used as an input for the Blast2GO analysis, based on local blastx. Enzyme Commission numbers were obtained using the same method. The data visualization was carried out using R (Version 3.3.3; http://www.R-project.org/) (35).

Isoform annotation

To create the isoform annotation, the 9755 coding transcripts were analyzed. The longest non-reverse ORF, in search order of [blastx-hit - frame1 - frame2 - frame3], using ATG only as start codon and obligatory stop codon was obtained from Blast2GO. Sequences were grouped by the transcription loci tags pyfaidx python package (36). All groups were aligned [-output=aln] and similarities were calculated [-other_pg seq_reformat -output sim] using T-Coffee software (37). Pairwise similarity scores formed two distinct groups. Based on this, the score cutoff level was set at 49 for designating isoforms. All low scoring accepted hits (score 49-60) were manually checked and visualized for correctness. The accepted protein sequence pairs were merged into isoform groups by connectivity calculation in R. Noncoding transcripts were grouped by overlapping features.

RESULTS

Transcripts reassembly and Identification of novel transcripts

The single end RNA seq data of *C.thermophilum* was obtained in triplicates with a length of 85bp. We performed short read gapped alignment using HISAT2 (30) and recovered more than 95% mapped reads as shown in Table-1. Further, we used StringTie (31) to de novo assemble all three samples separately. The assembled transcript files from these three samples were merged into a combined set of transcripts using the gffcompare utility provided by Cufflinks. After manual curation (17 transcripts) a total of 15363 reliable transcripts were obtained after filtering the transcript length (>200nt) (Figure-1C). In total, the transcripts correspond to 7044 coding genes, represented by 9772 transcripts and transcript isoforms, and 4567 noncoding genes, represented by 5591 transcripts and transcript isoforms. Gffcompare statistics on the comparison of the found transcripts compared to the previous annotation are shown in Figure-1B, and the transcripts annotated class codes are listed in Table-2, Figure-1D represents the increased lengths of the newly annotated transcripts compared to the earlier annotation, which did not include UTRs. Transcripts annotated to gffcompare classes u (no overlap, n = 2744) and x (opposite strand, n = 1754), as well as i (contained in reference intron, n = 28) and y (contains a reference gene within intron, 5 transcripts) are called as a novel transcript in our analysis. By these criteria, we identified 679 novel coding genes represented by 892 transcripts and isoforms as well as 2878 novel noncoding genes represented by 3639 transcripts and isoforms. 749 genes that were classified as a coding in the previous annotation were not detected/expressed under our experimental condition. These non-expressed putative genes are less conserved in related fungi (below 40%), suggesting that this group may contain falsely annotated genes. The total number of newly identified coding and noncoding transcripts is shown in Figure-1C. Finally, we created a gene annotation database containing all above information in TxDb framework in R package for *C.thermophilum* (38). The gene annotation file in gtf format can be found in supplementary file-1.

Annotation of the novel noncoding RNAs in *C.thermophilum*

We identified 5591 noncoding RNA transcripts based on CPC2 analysis. These include highly expressed contaminant RNA species such as ribosomal RNAs, t-RNAs, snoRNAs, RNase RNAs and snRNAs. The remaining non-coding transcripts were classified as intronic, intregenic, sense overlap with coding gene and antisense based on association with annotated protein-coding genes as shown in Figure-2D. In total we identified 2188 lincRNA genes as intregenic, 1949 antisense and 530 genes overlapping in sense direction. 166 genes are in both sense and antisense, these could represent truncated mRNAs or functional RNAs involved in gene regulation as shown in Figure-2C.

Sequence conservation and functional annotation

A phylogenetic tree including 6 other species indicates that *C.thermophilum* is closely related to *Thermothelomyces thermophila* and more distantly to *Neurospora crassa* and *Sordaria macrospora* (Figure-2F). Local dc-megaBLAST similarity search was carried out for the newly annotated coding, anti-sense and lincRNAs of *C.thermophilum*. We found that 90% of the coding and 45% of the non-coding sequences resulted in a significant hit with *Thermothelomyces thermophila* (Figure-2E). We also observed that both *Sordaria macrospora* and *Neurospora crassa* appear to be sharing 80% coding and 25% noncoding sequence similarity (Figure-2E). The similarity search against the other three species revealed lower sequence similarities as shown in Figure-2A. The gene ontology (GO) facilitates the functional characterization of genes; therefore, Blast2GO was used to associate the transcripts with a functional annotation. We found altogether 4283 GO terms, 1428 belong to Molecular Function (MF), 2140 to Biological Process (BP), and 715 to Cellular Component (CC)(Figure-2B). With these, we could annotate 4336 coding genes. All GO terms are listed in supplementary file-2 as well as top 10 GO-slim terms for each category (MF, CC, BP) are shown in supplementary Figure-1. Further, we created a GO annotation R package for *C.thermophilum* using function makeOrgPackage (38). To facilitate functional gene finding for industrial applications, we also retrieved Enzyme Commission numbers (E.C) from the Blast2GO analysis. We could associate 1366 coding genes with 643 E.C. numbers. The main EC classes distribution is shown in Figure-3.

DISCUSSION

C.thermophilum is an economically important species in pharmaceutical and food processing industries, and a useful model organism in basic research, particularly in the field of ribosome biogenesis (12–14, 22, 39, 40). In this study an analysis pipeline was developed to characterize the transcriptome of *C.thermophilum*. The main aim of the present study was to provide a comprehensive gene annotation, containing high fidelity coding and noncoding transcripts and isoforms with both 3' and 5' UTRs. We evaluated all transcript class codes from our transcriptome assembly and compared it with the previous annotation. We observed substantial discrepancies with the previous annotation of *C.thermophilum*, as only 2935 coding and 143 noncoding transcripts displayed a complete intron match (class code "="). The majority of coding transcripts showed at least one intron mismatch with the old annotation (class code "j"). We observed that 749 coding and 254 noncoding transcripts of the previous annotation were not expressed (no overlaps with any identified transcript) in our analysis. 104 of these genes were potentially expressed in other growth conditions (> 100 read counts, unpublished RNA sequencing data), while the remaining genes were not expressed. Furthermore, most of these genes had no homology in related species, suggesting that they might have been wrongly annotated. These unexpressed transcripts are listed in supplementary files-4.

Our transcript assembly revealed that 1640 genes express at least 2-3 transcript isoforms (4368 transcripts, 44.7%), indicating rather complex alternative splicing in *C.thermophilum*. A hand full of genes may express even higher numbers of isoforms. However, due to theoretical limitations in the analysis of the single end sequencing data such potentially complex isoforms cannot be reliably predicted. Here we present all potential transcript isoforms, and the corresponding predicted protein sequences are listed in supplementary file-3. Our analysis also revealed a surprisingly high number of noncoding transcripts, both lincRNA and antisense RNAs. Taken together with the alternative splicing of coding genes our data suggests an astonishing complexity of the genome architecture and transcriptional network of *C.thermophilum*. Such complexity is usually found in higher eukaryotes where it is supposed to permit complex life style. Comparative analysis uncovered that *Thermothelomyces thermophila* shows 90% genome sequence similarity with *C.thermophilum* as shown in Figure-2E). The high sequence similarity of *Thermothelomyces thermophila* and *C.thermophilum* is likely to reflect related functions, potentially associated with their thermophilic lifestyle. Functional annotation through Gene Ontology (GO) associations facilitates the interpretation of genomic and transcriptomic sequence data. Functional annotation achieved by the integration of several databases such as KEGG (43), UniProt (44), InterPro(45), Pfam(46), NCBI (47), SEED (48), ConsensusPathDB (49) and Reactome (50) may help addressing this question. Our Blast2GO analysis substantially expands the GO term annotations in *C.thermophilum*. Hence, our annotation and transcript assembly opens opportunities for the systematic analysis of *C.thermophilum*. An example would be a known protein fragment of a laccase enzyme from *C.thermophilum* (Uniprot ID: Q692I0) that was so far not annotated in the genome and the full protein sequence was unknown. We found the gene expressed in our conditions and have annotated the complete transcript and protein sequence (Gene001724). *C.thermophilum* genes responsible for the formation conidia are another example for an extended functional annotation. The reproduction cycle of *C.thermophilum* is not well understood, and previously only a single gene had been associated with the formation of asexual spores (conidia). Our GO term annotation revealed that several additional genes, such as Gene000283; Gene003743; Gene006754; Gene006902, are associated with the conidiation process (GO:0048315, GO:0030437, GO:0030435).

Our analysis is solely based on a single growth condition with single-end RNA seq data. Hence, different growth conditions, as well as paired-end sequencing data could further improve the gene annotation. RNA degradation seems to be a general problem in *C.thermophilum*. Therefore, improvement of the RNA isolation procedure could decrease the number of truncated transcripts. In summary, we detected a large number of novel coding and noncoding transcripts and discovered a high diversity of alternative splicing in *C.thermophilum*. Thus, our study provides useful resources for functional genomics and proteomics research on *C.thermophilum* and facilitates the analysis of biological and biochemical processes. Moreover, our analysis pipeline can be used for the genome annotation of other organisms.

AVAILABILITY

The GO Term and gene annotation R packages of *C.thermophilum* are available in the following link. The gene annotation R package can be found in https://www.bzh.uni-heidelberg.de/brunner/Chaetomium_thermophilum/TxDb.Chaetomium.ct39.knownGene_1.0.0.tar.gz and GO annotation can be found in https:/www.bzh.uni heidelberg.de/brunner/Chaetomium_thermophilum/org.Cthermophilum.eg.db_1.0.0.tar.gz

ACCESSION NUMBERS

The RNA seq data are deposited in the Gene Expression Omnibus database under accession no GSE116834.

SUPPLEMENTARY DATA

We provide the following files for both scientific and industrial applications.

1. The C.thermophilum_new.gtf file containing the essential attributes/annotation features that can be used further in genomics and transcriptomics data analysis of *C.thermophilum*
2. The C.thermophilum.xlsx file containing all genes associated with GO annotation and E.C. numbers.
3. The C.thermophilum protein.fasta file contains all the translated protein sequences, ORF with both start and stop codons in the forward strand, with at least 50 AA lengths.
4. Missing.genes.xlsx contains the 749 coding genes from the old annotations where we found no expression in excel format.

ACKNOWLEDGEMENTS

We thank David Ibberson at the CellNetworks Deep Sequencing Core Facility (Heidelberg, Germany) for performing library preparation and HiSeq sequencing. AS thanks Prof. Dr. Thomas Höfer and Dr. Congxin Li (Bioqunat, Heidelberg, Germany) for allowing processing the sequencing data in their computer cluster. As thanks to PD Dr. Jochen Baßler for fruitful discussion.

FUNDING

The work was supported by ERC ADG No.741781 GLOWSOME to EH and the Collaborative Research Centre TRR 186 of the Deutsche Forschungsgemeinschaft (DFG) to MB. EH and MB are members of CellNetworks.

CONFLICT OF INTEREST

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

REFERENCES

1. La Touche,G. (1948) A chaetomium-like thermophile fungus. *Nature*, **161**, 320.

2. Ganju,R.K., Vithayathil,P.J. and Murthy,S. (1989) Purification and characterization of two xylanases from chaetomium thermophile var. Coprophile. *Canadian journal of microbiology*, **35**, 836–842.

3. Chefetz,B., Chen,Y. and Hadar,Y. (1998) Purification and characterization of laccase fromChaetomium thermophilium and its role in humification. *Applied and Environmental Microbiology*, **64**, 3175–3179.

4. Li,D.-C., Lu,M., Li,Y.-L. and Lu,J. (2003) Purification and characterization of an endocellulase from the thermophilic fungus chaetomium thermophilum ct2. *Enzyme and microbial technology*, **33**, 932–937.

5. Li,A.-N., Ding,A.-Y., Chen,J., Liu,S.-A., Zhang,M. and Li,D.-C. (2007) Purification and characterization of two thermostable proteases from the thermophilic fungus chaetomium thermophilum. *Journal of microbiology and biotechnology*, **17**, 624–631.

6. Li,A.-N., Yu,K., Liu,H.-Q., Zhang,J., Li,H. and Li,D.-C. (2010) Two novel thermostable chitinase genes from thermophilic fungi: Cloning, expression and characterization. *Bioresource technology*, **101**, 5546–5551.

7. Hakulinen,N., Turunen,O., Jänis,J., Leisola,M. and Rouvinen,J. (2003) Three-dimensional structures of thermophilic β-1, 4-xylanases from chaetomium thermophilum and nonomuraea flexuosa. *The FEBS Journal*, **270**, 1399–1412.

8. Rosgaard,L., Pedersen,S., Cherry,J.R., Harris,P. and Meyer,A.S. (2006) Efficiency of new fungal cellulase systems in boosting enzymatic degradation of barley straw lignocellulose. *Biotechnology progress*, **22**, 493–498.

9. Voutilainen,S.P., Puranen,T., Siika-aho,M., Lappalainen,A., Alapuranen,M., Kallio,J., Hooman,S., Viikari,L., Vehmaanperä,J. and Koivula,A. (2008) Cloning, expression, and characterization of novel thermostable family 7 cellobiohydrolases. *Biotechnology and bioengineering*, **101**, 515–528.

10. Elleuche,S., Schäfers,C., Blank,S., Schröder,C. and Antranikian,G. (2015) Exploration of extremophiles for high temperature biotechnological processes. *Current opinion in microbiology*, **25**, 113–119.

11. Sriyapai,T., Somyoonsap,P., Matsui,K., Kawai,F. and Chansiri,K. (2011) Cloning of a thermostable xylanase from actinomadura sp. S14 and its expression in escherichia coli and pichia pastoris. *Journal of bioscience and bioengineering*, **111**, 528–536.

12. Chen,X., Li,W., Ji,P., Zhao,Y., Hua,C. and Han,C. (2018) Engineering the conserved and noncatalytic residues of a thermostable β-1, 4-endoglucanase to improve specific activity and thermostability. *Scientific reports*, **8**, 2954.

13. Zhou,Q., Ji,P., Zhang,J., Li,X. and Han,C. (2017) Characterization of a novel thermostable gh45 endoglucanase from chaetomium thermophilum and its biodegradation of pectin. *Journal of bioscience and bioengineering*, **124**, 271–276.

14. Li,W., Ji,P., Zhou,Q., Hua,C. and Han,C. (2018) Insights into the synergistic biodegradation of waste papers using a combination of thermostable endoglucanase and cellobiohydrolase from chaetomium thermophilum. *Molecular biotechnology*, **60**, 49–54.

15. Monecke,T., Haselbach,D., Voß,B., Russek,A., Neumann,P., Thomson,E., Hurt,E., Zachariae,U., Stark,H. and Grubmüller,H. *et al.* (2013) Structural basis for cooperativity of crm1 export complex formation. *Proceedings of the National Academy of Sciences*, **110**, 960–965.

16. Ulrich,A. and Wahl,M.C. (2014) Structure and evolution of the spliceosomal peptidyl-prolyl cis–trans isomerase cwc27. *Acta Crystallographica Section D: Biological Crystallography*, **70**, 3110–3123.

17. Aibara,S., Valkov,E., Lamers,M.H., Dimitrova,L., Hurt,E. and Stewart,M. (2015) Structural characterization of the principal mRNA-export factor mex67–mtr2 from chaetomium thermophilum. *Acta Crystallographica Section F: Structural Biology Communications*, **71**, 876–888.

18. Hondele,M., Stuwe,T., Hassler,M., Halbach,F., Bowman,A., Zhang,E.T., Nijmeijer,B., Kotthoff,C., Rybin,V. and Amlacher,S. *et al.* (2013) Structural basis of histone h2a–h2b recognition by the essential chaperone fact. *Nature*, **499**, 111.

19. Leidig,C., Bange,G., Kopp,J., Amlacher,S., Aravind,A., Wickles,S., Witte,G., Hurt,E., Beckmann,R. and Sinning,I. (2013) Structural characterization of a eukaryotic chaperone?The ribosome-associated complex. *Nature Structural and Molecular Biology*, **20**, 23.

20. Baker,R.W., Jeffrey,P.D., Zick,M., Phillips,B.P., Wickner,W.T. and Hughson,F.M. (2015) A direct role for the sec1/munc18-family protein vps33 as a template for snare assembly. *Science*, **349**, 1111–1114.

21. Stuwe,T., Bley,C.J., Thierbach,K., Petrovic,S., Schilbach,S., Mayo,D.J., Perriches,T., Rundlet,E.J., Jeon,Y.E. and Collins,L.N. *et al.* (2015) Architecture of the fungal nuclear pore inner ring complex. *Science*, **350**, 56–64.

22. Kornprobst,M., Turk,M., Kellner,N., Cheng,J., Flemming,D., Koš-Braun,I., Koš,M., Thoms,M., Berninghausen,O. and Beckmann,R. *et al.* (2016) Architecture of the 90S pre-ribosome: A structural view on the birth of the eukaryotic ribosome. *Cell*, **166**, 380–393.

23. Amlacher,S., Sarges,P., Flemming,D., Noort,V. van, Kunze,R., Devos,D.P., Arumugam,M., Bork,P. and Hurt,E. (2011) Insight into structure and assembly of the nuclear pore complex by utilizing the genome of a eukaryotic thermophile. *Cell*, **146**, 277–289.

24. Bock,T., Chen,W.-H., Ori,A., Malik,N., Silva-Martin,N., Huerta-Cepas,J., Powell,S.T., Kastritis,P.L., Smyshlyaev,G. and Vonkova,I. *et al.* (2014) An integrated approach for genome annotation of the eukaryotic thermophile chaetomium thermophilum. *Nucleic acids research*, **42**, 13525–13533.

25. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., Van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**, 511.

26. Robertson,G., Schein,J., Chiu,R., Corbett,R., Field,M., Jackman,S.D., Mungall,K., Lee,S., Okada,H.M. and Qian,J.Q. *et al.* (2010) De novo assembly and analysis of rna-seq data. *Nature methods*, **7**, 909.

27. Griffith,M., Griffith,O.L., Mwenifumbo,J., Goya,R., Morrissy,A.S., Morin,R.D., Corbett,R., Tang,M.J., Hou,Y.-C. and Pugh,T.J. *et al.* (2010) Alternative expression analysis by rna sequencing. *Nature methods*, **7**, 843.

28. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470.

29. Walz,M. and Kück,U. (1991) Polymorphic karyotypes in related acremonium strains. *Current genetics*, **19**, 73–76.

30. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: A fast spliced aligner with low memory requirements. *Nature methods*, **12**, 357.

31. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C.,

dell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, **33**, 290.

32. Kang,Y.-J., Yang,D.-C., Kong,L., Hou,M., Meng,Y.-Q., Wei,L. and Gao,G. (2017) CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, **45**, W12–W16.

33. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. BLAST+: Architecture and applications. *BMC Bioinformatics*, **10**, 421.

34. Conesa,A., Götz,S., García-Gómez,J.M., Terol,J., Talón,M. and Robles,M. (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

35. R Core Team (2014) R: A language and environment for statistical computing. Vienna, austria: R foundation for statistical computing; 2014.

36. Shirley,M.D., Ma,Z., Pedersen,B.S. and Wheelan,S.J. (2015) Efficient" pythonic" access to fasta files using pyfaidx PeerJ PrePrints.

37. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-coffee: A novel method for fast and accurate multiple sequence alignment1. *Journal of molecular biology*, **302**, 205–217.

38. Carlson,M. and Pages,H. AnnotationForge: Code for building annotation database packages. R package version 1.4. 4.

39. Mäntylä,A., Paloheimo,M., Hakola,S., Lindberg,E., Leskinen,S., Kallio,J., Vehmaanperä,J., Lantto,R. and Suominen,P. (2007) Production in trichoderma reesei of three xylanases from chaetomium thermophilum: A recombinant thermoxylanase for biobleaching of kraft pulp. *Applied microbiology and biotechnology*, **76**, 377–386.

40. Li,A.-N. and Li,D.-C. (2009) Cloning, expression and characterization of the serine protease gene from chaetomium thermophilum. *Journal of applied microbiology*, **106**, 369–380.

41. Garre,V., Müller,U. and Tudzynski,P. (1998) Cloning, characterization, and targeted disruption of cpcat1, coding for an in planta secreted catalase of claviceps purpurea. *Molecular plant-microbe interactions*, **11**, 772–783.

42. Zámocky,M., Tafer,H., Chovanová,K., Lopandic,K., Kamlárová,A. and Obinger,C. (2016) Genome sequence of the filamentous soil fungus chaetomium cochliodes reveals abundance of genes for heme enzymes from all peroxidase and catalase superfamilies. *BMC genomics*, **17**, 763.

43. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2016) KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, **45**, D353–D361.

44. Consortium,U. (2016) UniProt: The universal protein knowledgebase. *Nucleic acids research*, **45**, D158–D169.

45. Finn,R.D., Attwood,T.K., Babbitt,P.C., Bateman,A., Bork,P., Bridge,A.J., Chang,H.-Y., Dosztányi,Z., El-Gebali,S. and Fraser,M. *et al.* (2016) InterPro in 2017?Beyond protein family and domain annotations. *Nucleic acids research*, **45**, D190–D199.

46. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. and Durbin,R. *et al.* (2006) Pfam: Clans, web tools and services. *Nucleic acids research*, **34**, D247–D251.

47. Coordinators,N.R. (2016) Database resources of the national center for biotechnology information. *Nucleic acids research*, **44**, D7.

48. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B. and Shukla,M. *et al.* (2013) The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic acids research*, **42**, D206–D214.

49. Kamburov,A., Wierling,C., Lehrach,H. and Herwig,R. (2008) ConsensusPathDB?A database for integrating human functional interaction networks. *Nucleic acids research*, **37**, D623–D628.

50. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., Bono,B. de, Jassal,B., Gopinath,G., Wu,G. and Matthews,L. *et al.* (2005) Reactome: A knowledgebase of biological pathways. *Nucleic acids research*, **33**, D428–D432.

51. Chen, X.-F., Lehmann, L., Lin, J. J., Vashisht, A., Schmidt, R., Ferrari, R., Huang, C., McKee, R., Mosley, A., Plath, K., et al. (2012) Mediator and SAGA have distinct roles in Pol II preinitiation complex assembly and function. *Cell reports,* 2(5), 1061–1067. □

52. Baßler, J., Ahmed, Y. L., Kallas, M., Kornprobst, M., Calviño, F. R., Gnädig, M., ... & Castillo, N. (2017). Interaction network of the ribosome assembly machinery from a eukaryotic thermophile. *Protein Science*, *26*(2), 327-342.

53. Brighenti, E., Trere´, D., and Derenzini, M. (2015) Targeted cancer therapy with ribosome biogenesis inhibitors: a real possibility. *Oncotarget,* 6(36), 38617.

54. Quin, J. E., Devlin, J. R., Cameron, D., Hannan, K. M., Pearson, □ R. B., and Hannan, R. D. (2014) Targeting the nucleolus for cancer intervention. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease,* 1842(6), 802–816.

55. Burger, K., Mu □ehl, B., Harasim, T., Rohrmoser, M., Malamoussi, A., Orban, M., Kellner, M., Gruber-Eber, A., Kremmer, E., Ho □elzel, M., et al. (2010) Chemotherapeutic drugs inhibit ribosome biogenesis at various levels. *Journal of Biological Chemistry,* pp. jbc–M109

TABLE AND FIGURES LEGENDS

Table- 1: RNA-seq data alignment results for reads of different samples.

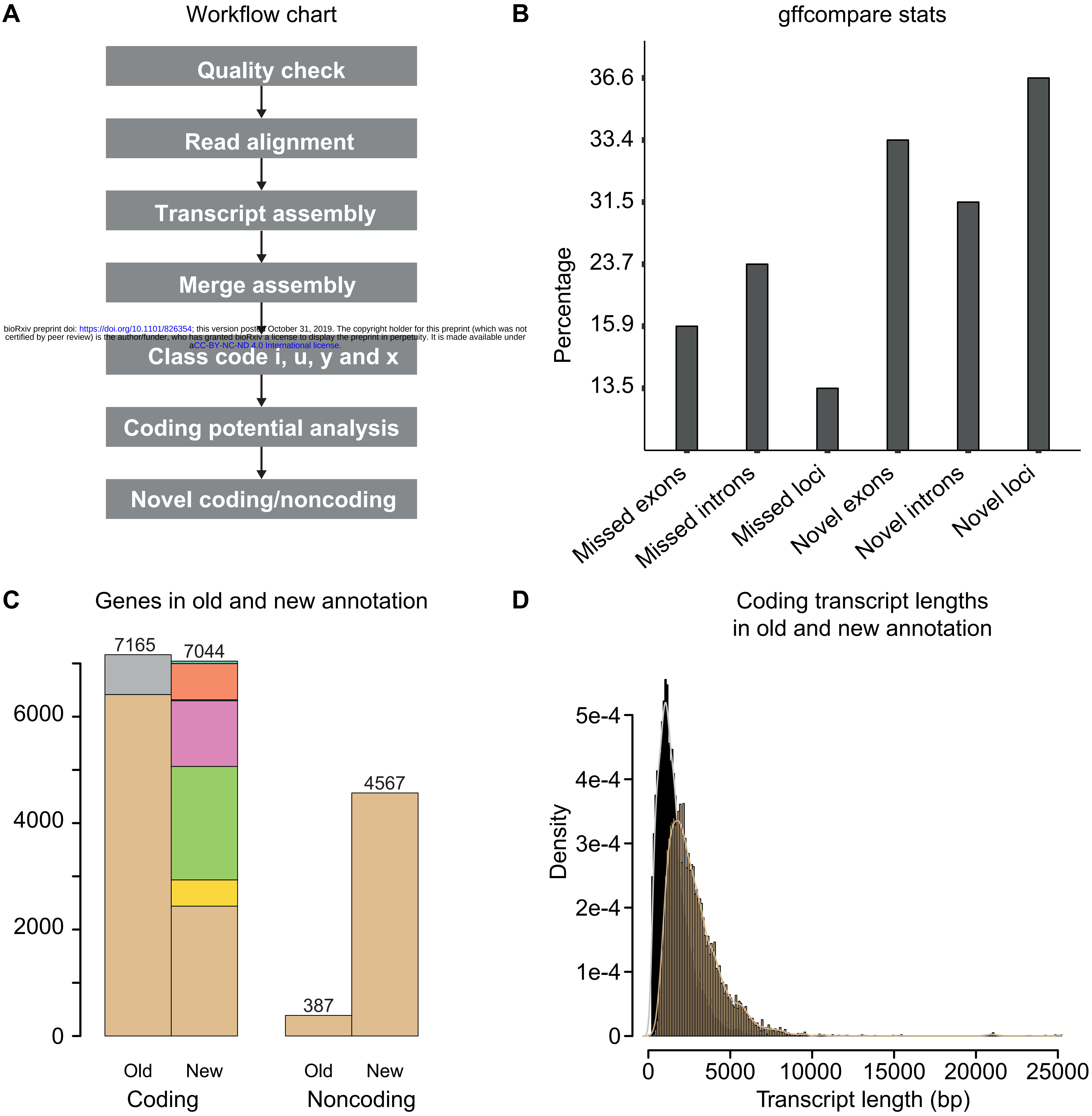
Table- 2: Different classes of assembled transcripts

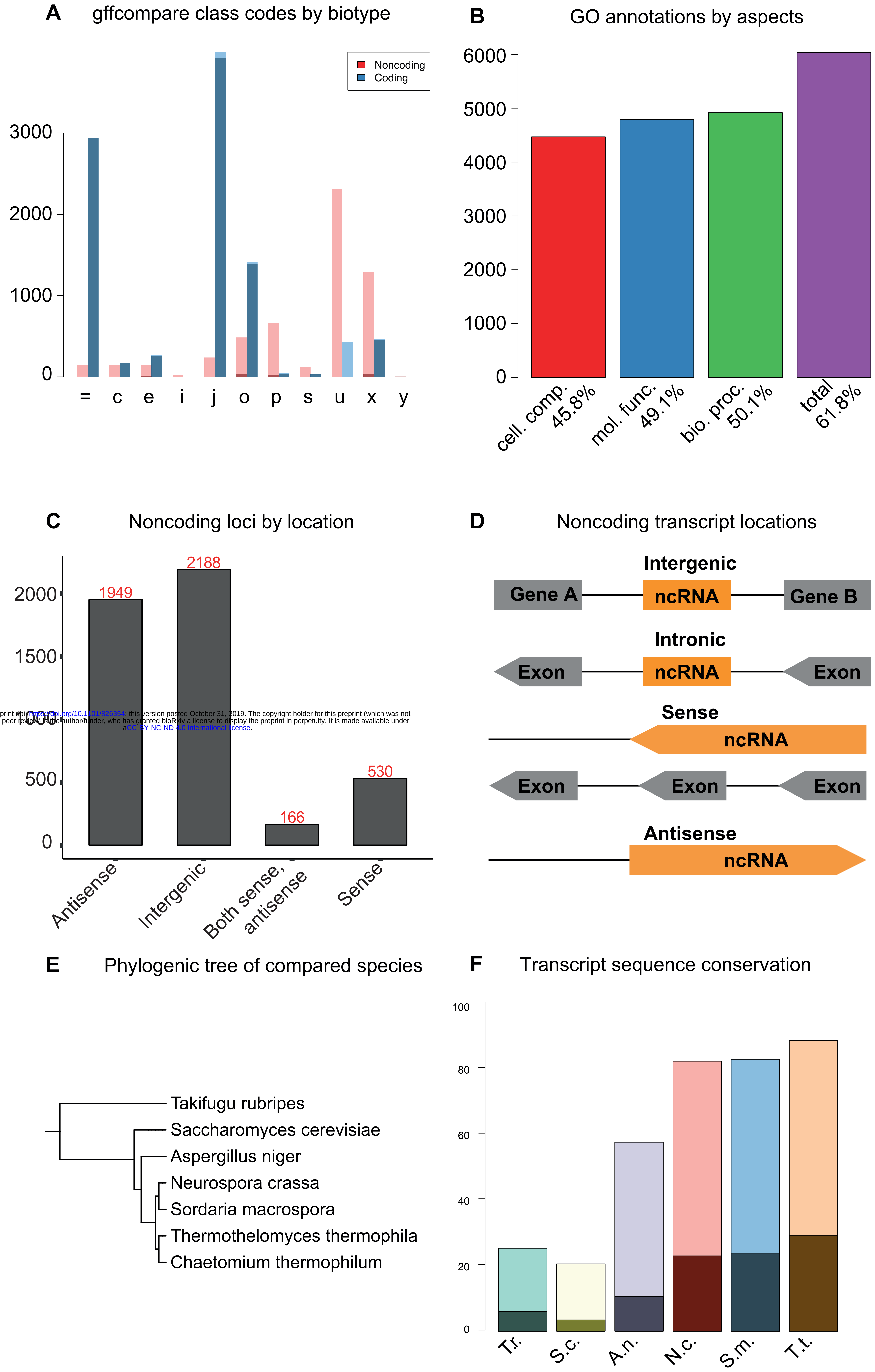
Figure-1: Transcriptome-based annotation of the genome of *C.thermophilum*. A) Systematic overview of the analysis pipeline. B) Identification of the novel intron and exon loci in percentage from newly annotated genome. C) Stack bar plot of genes of the old and new annotation. The stacked grey bar for the coding genes in the old annotation represents genes that were not expressed in our conditions. In the stacked bar of the new annotation, the colors represent differences compared to the previous annotation. The individual colors indicate for the following features: Beige; genes without change in the intron structure (n=2440). Yellow: genes where at least one isoform has a novel splicing variant (n=493). Green: genes where all transcripts have at least one intron-exon junction different than in the previous annotation (n=2132). Pink; genes where all junctions are different (n=1234). Blue; genes that are flipped (opposite strand with same or similar splice junctions) (n=19). Orange; completely novel genes (n=679). On the top the aquamarine represents the uncategorized genes (n=47). D) Histograms representing the coding transcript lengths from the new *C.thermophilum* genome annotation (beige) compared to the length of the previous annotation of ORFs (black). A detailed description of the methods is given in the respective section of the main text.

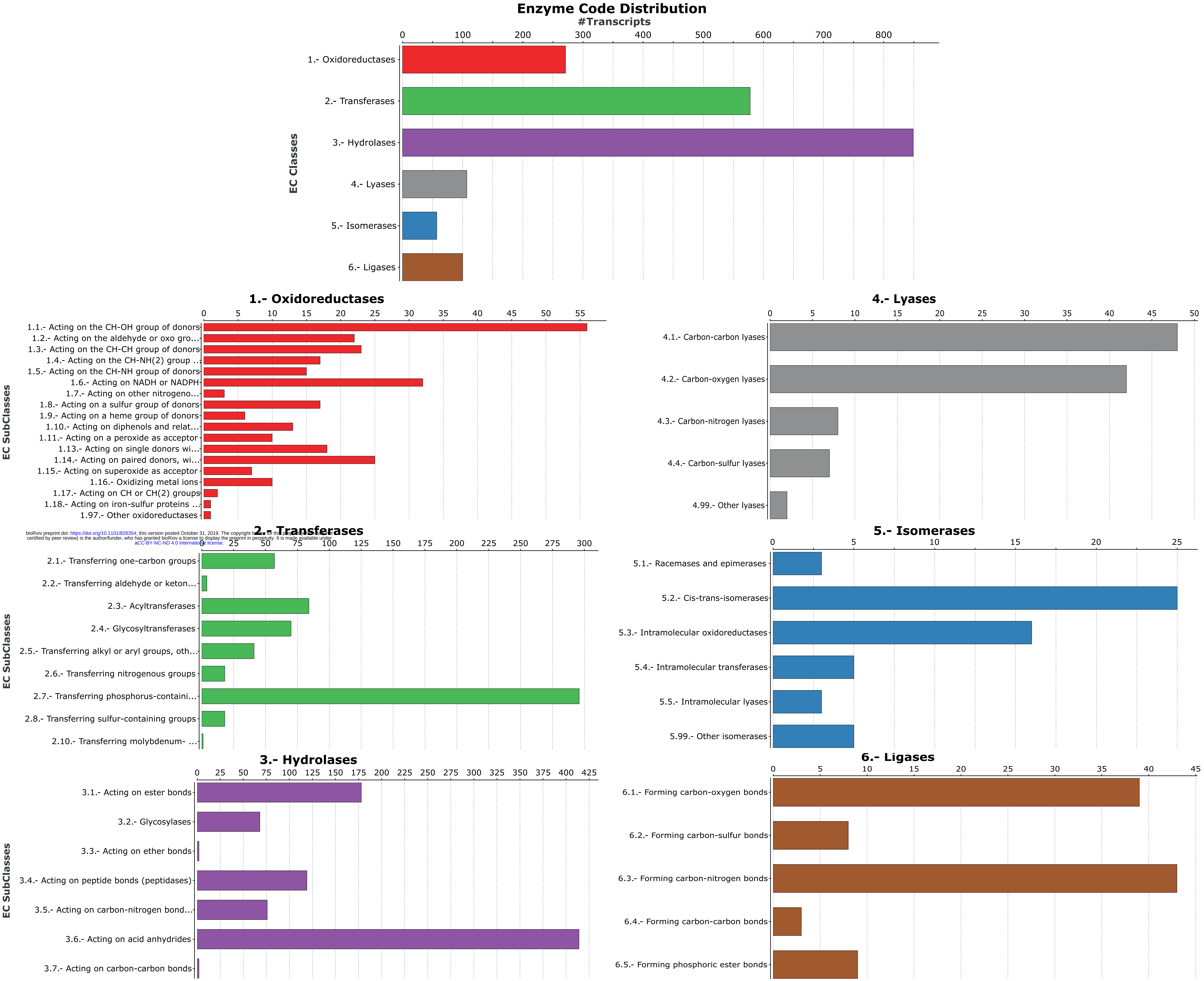
Figure-2*:* A) Transcript assemblies and gffcompare class codes. Stack bar plot represents the gffcompare class codes containing coding and noncoding transcripts. Darker shading represents the part where the corresponding transcript in the old annotation has the same biotype. B) GO annotations associated with genes, representing biological process (BP), molecular function (MF) and cellular compartment (CC). C) Bar plot representing different types of noncoding genes. D) Graphical notation of different types of noncoding transcripts by location. E) Phylogenic tree of 6 compared species. F) Stack bar plot representing the sequence similarity of transcripts of 6 compared species. Darker shade for the non-coding transcripts, light shade for the coding.

Figure-3: Bar plot enzyme class code distribution with transcripts. 6 different enzyme classes are distributed throughout the transcripts. Each bar plot (1-6) containing top 5 enzymes with specific enzyme class code.

Supplementary Figure-1: Bar plot of top abundant GO slim terms obtained from Blast2GO analysis. Y- axis represents the GO term and x- axis represents the total number of transcripts are involved in particular term. Different color code such as light-yellow corresponds to the cellular component and light green to biological process and light blue to molecular functions.


A
Workflow chart
Quality check
Read alignment
Transcript assembly
Merge assembly
Class code i, u, y and x
Coding potential analysis
Novel coding/noncoding
B
gffcompare stats
Percentage
36.6
33.4
31.5
23.7
15.9
13.5
Missed exons
Missed introns
Missed loci
Novel exons
Novel introns
Novel loci
C
Genes in old and new annotation
7165
7044
4567
387
6000
4000
2000
0
Old
New
Old
New
Coding
Noncoding
D
Coding transcript lengths
in old and new annotation
Density
5e-4
4e-4
3e-4
2e-4
1e-4
0
0
5000
10000
15000
20000
25000
Transcript length (bp)


A
gffcompare class codes by biotype
Noncoding
Coding
3000
2000
1000
0
=
c
e
i
j
o
p
s
u
x
y
B
GO annotations by aspects
6000
5000
4000
3000
2000
1000
0
cell. comp. 45.8%
mol. func. 49.1%
bio. proc. 50.1%
total 61.8%
C
Noncoding loci by location
2000
1500
1000
500
0
1949
2188
166
530
Antisense
Intergenic
Both sense, antisense
Sense
D
Noncoding transcript locations
Intergenic
Gene A
ncRNA
Gene B
Intronic
Exon
ncRNA
Exon
Sense
ncRNA
Exon
Exon
Exon
Antisense
ncRNA
E
Phylogenic tree of compared species
Takifugu rubripes
Saccharomyces cerevisiae
Aspergillus niger
Neurospora crassa
Sordaria macrospora
Thermothelomyces thermophila
Chaetomium thermophilum
F
Transcript sequence conservation
100
80
60
40
20
0
T.r.
S.c.
A.n.
N.c.
S.m.
T.t.



Enzyme Code Distribution
#Transcripts
0
100
200
300
400
500
600
700
800
EC Classes
1.- Oxidoreductases
2.- Transferases
3.- Hydrolases
4.- Lyases
5.- Isomerases
6.- Ligases
1.- Oxidoreductases
0
5
10
15
20
25
30
35
40
45
50
55
EC SubClasses
1.1.- Acting on the CH-OH group of donors
1.2.- Acting on the aldehyde or oxo gro...
1.3.- Acting on the CH-CH group of donors
1.4.- Acting on the CH-NH(2) group ...
1.5.- Acting on the CH-NH group of donors
1.6.- Acting on NADH or NADPH
1.7.- Acting on other nitrogeno...
1.8.- Acting on a sulfur group of donors
1.9.- Acting on a heme group of donors
1.10.- Acting on diphenols and relat...
1.11.- Acting on a peroxide as acceptor
1.13.- Acting on single donors wi...
1.14.- Acting on paired donors, wi...
1.15.- Acting on superoxide as acceptor
1.16.- Oxidizing metal ions
1.17.- Acting on CH or CH(2) groups
1.18.- Acting on iron-sulfur proteins ...
1.97.- Other oxidoreductases
4.- Lyases
0
5
10
15
20
25
30
35
40
45
50
4.1.- Carbon-carbon lyases
4.2.- Carbon-oxygen lyases
4.3.- Carbon-nitrogen lyases
4.4.- Carbon-sulfur lyases
4.99.- Other lyases
2.- Transferases
0
25
50
75
100
125
150
175
200
225
250
275
300
EC SubClasses
2.1.- Transferring one-carbon groups
2.2.- Transferring aldehyde or keton...
2.3.- Acyltransferases
2.4.- Glycosyltransferases
2.5.- Transferring alkyl or aryl groups, oth...
2.6.- Transferring nitrogenous groups
2.7.- Transferring phosphorus-containi...
2.8.- Transferring sulfur-containing groups
2.10.- Transferring molybdenum- ...
5.- Isomerases
0
5
10
15
20
25
5.1.- Racemases and epimerases
5.2.- Cis-trans-isomerases
5.3.- Intramolecular oxidoreductases
5.4.- Intramolecular transferases
5.5.- Intramolecular lyases
5.99.- Other isomerases
3.- Hydrolases
0
25
50
75
100
125
150
175
200
225
250
275
300
325
350
375
400
425
EC SubClasses
3.1.- Acting on ester bonds
3.2.- Glycosylases
3.3.- Acting on ether bonds
3.4.- Acting on peptide bonds (peptidases)
3.5.- Acting on carbon-nitrogen bond...
3.6.- Acting on acid anhydrides
3.7.- Acting on carbon-carbon bonds
6.- Ligases
0
5
10
15
20
25
30
35
40
45
6.1.- Forming carbon-oxygen bonds
6.2.- Forming carbon-sulfur bonds
6.3.- Forming carbon-nitrogen bonds
6.4.- Forming carbon-carbon bonds
6.5.- Forming phosphoric ester bonds

Most abundant Go-slim terms

#Transcripts
0
250
500
750
1,000
1,250
1,500
protein-containing complex
cytoplasm
nucleus
cellular_component
mitochondrion
cytosol
endoplasmic reticulum
intracellular
plasma membrane
ribosome
Cellular Component
biosynthetic process
cellular nitrogen compound metabolic process
cellular protein modification process
response to stress
small molecule metabolic process
catabolic process
chromosome organization
biological_process
DNA metabolic process
vesicle-mediated transport
Biological Process
ion binding
molecular_function
oxidoreductase activity
DNA binding
transmembrane transporter activity
RNA binding
kinase activity
ATPase activity
DNA binding transcription factor activity
peptidase activity
Molecular Function