Feb 05, 2020

Materials and Methods - Transcriptomic profiling identifies novel transcripts, isomorphs and lncRNAs in Paracoccidioides brasiliensis

Fabiano Menegidio[1], David Aciole Barbosa[1], Regina Costa de Oliveira[1], Daniela L. Jabes[1], Luiz R. Nunes[2]

[1]Núcleo Integrado de Biotecnologia, Universidade de Mogi das Cruzes (UMC), Brazil., [2]Centro de Ciências Naturais e Humanas, Universidade Federal do ABC (UFABC), Brazil

In Development dx.doi.org/10.17504/protocols.io.bbwtipen

Fabiano Menegidio

Culture conditions

P. brasiliensis isolate 18 (Pb18), (provided by Dr. R. Puccia, UNIFESP, Brazil) was used throughout this work. Fungal yeasts, isolated from infected mice, were initially grown on solid modified YPD medium (0.5% yeast extract, 0.5% casein peptone, 1.5% glucose, pH 6.3, 1.5% agar) at 36°C. Samples from these plates were transferred to 10 ml modified YPD liquid cultures and incubated at 120 rpm, in a rotary shaker, at 36°C for 5 days, until mid-exponential growth phase (OD600 = 2.0) was reached. At this point, cells were harvested by centrifugation for total RNA extraction.

RNA extraction, library construction and NGS Sequencing

RNA was extracted as described by [1] and further purified with Qiagen RNeasy columns. Next, RNA samples were treated with RNAse-free RQ1 DNAse (1 unit/µg of RNA), for 2 h, at 37°C and purified again in RNeasy columns. Total RNA was then quantified using a Quantus Fluorometer (Promega) and RNA integrity was assessed in a Bioanalyzer 2100 using the Agilent RNA 6000 Nano chip. Samples showing RNA Integrity Number (RIN) ≥ 8.0, were selected for preparation of sequencing libraries, with the TruSeq Stranded mRNA LT Sample Preparation kit, following the manufacturer's instructions (Illumina, Inc.). Quantification of libraries using qPCR was done using a commercial kit (KAPA Library Quant kit, KAPA Biosystems) and a 7500 fast Real Time qPCR System (Applied Biosystems, Inc.). The standard curve was used to calculate the PCR efficiency, as well as molarity of the libraries, using KAPA's recommended curve to scale the quantification, based on average fragment size, previously determined in the Bioanalyzer run. In total, 10 different RNA libraries were prepared, with RNA samples derived from 10 independent cultures. Sequencing was performed using an Illumina MiSeq platform (75 X 75 bp paired-end reads).

Bioinformatics analyses

Raw sequencing data, in FASTQ format, was processed by a Galaxy Public Server, available at https://usegalaxy.eu. Initially, the libraries were submitted to quality control checking, using FastQC [2] and Trimmomatic was used to remove low-quality (Q<30) reads, adapters and other contaminant sequences [3]. Quality-filtered reads were then mapped to the latest version of the Pb18 reference genome, using HISAT2 [4]. StringTie [5] was then used to assemble the mapped readings into transcripts, using the de novo transcriptome reconstruction method, allowing identification of all transcripts present in each sample (including currently annotated genes, as well as newly identified elements and isomorphs). SringTie Merge was next used to combine redundant transcription structures, providing a non-redundant reference transcriptome, with unique identifiers. Cufflinks [6] was next used to estimate expression values (FPKM) for each element in the StringTie-generated reference transcriptome. Transcriptome completeness was assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software, version 3.0 [7]. Finally, transcripts were classified into different Transcription Class Codes (TCCs), reflecting their respective nature/origin, with the aid of Cuffcompare [6], using a GFF3 reference annotation file for the Pb18 genome, downloaded from NCBI (see Table S1, spreadsheet 1 for details). The final dataset obtained from the analyses with StringTie/Cufflinks/Cuffcompare was filtered by expression level and by TCC, as suggested elsewhere [8, 9, 10, 11], so that only elements belonging to TCCs "=", "j", "u" and "x", and displaying FPKM ≥ 1 were considered real transcripts and used in subsequent analyses.

Transcripts belonging to TCCs "=" and "j" were annotated using BLAST against the Paracoccidioides-specific database ParaDB [12], using an E-value ≤ e-10 as cutoff for identification. Transcripts classified as "u" and "x" were evaluated for their respective coding potential with the aid of four tools: (i) Coding Potential Predictor (CPPred) [13], Coding Potential Calculator (CPC2) [14], Coding-Non-Coding Index (CNCI) [15] and Coding-Potential Assessment Tool (CPAT) [16]. Transcripts identified as having coding potential by any of these tools were tentatively identified using BLAST against the latest version of the NCBI nt database (V.20180122), using an E-value ≤ e-10 as cut-off for identification. The remaining transcripts, identified as non-coding by all four prediction tools, were submitted to Infernal [17], for classification into the different families of non-coding RNAs defined in the Rfam database [18]. Finally, these putative ncRNAs were submitted to lncRNA extraction with the aid of FEElnc [19], which evaluated their potential nature as lncRNAs and identified their respective partner genes.

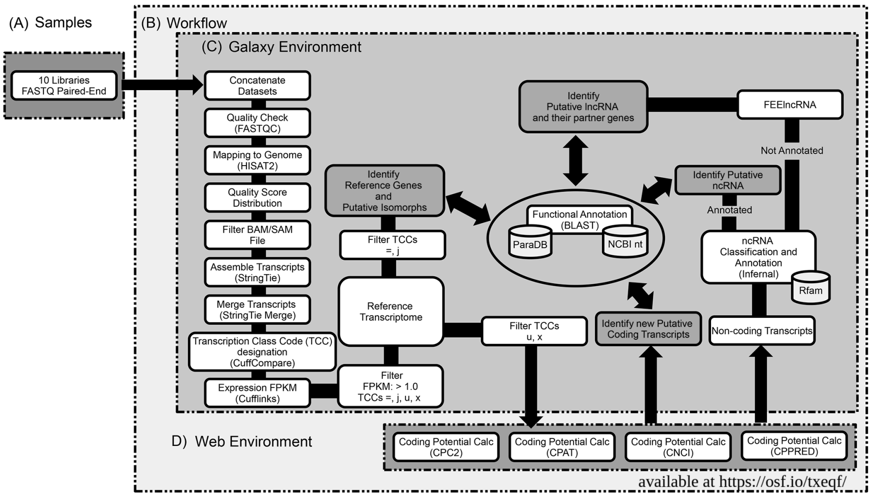
(A) Samples
10 Libraries FASTQ Paired-End
(B) Workflow
(C) Galaxy Environment
Concatenate Datasets
Quality Check (FASTQC)
Mapping to Genome (HISAT2)
Quality Score Distribution
Filter BAM/SAM File
Assemble Transcripts (StringTie)
Merge Transcripts (StringTie Merge)
Transcription Class Code (TCC) designation (CuffCompare)
Expression FPKM (Cufflinks)
Identify Reference Genes and Putative Isomorphs
Filter TCCs =, j
Reference Transcriptome
Filter FPKM: > 1.0 TCCs =, j, u, x
Identify Putative lncRNA and their partner genes
FEElncRNA
Not Annotated
Identify Putative ncRNA
Annotated
Functional Annotation (BLAST)
ParaDB
NCBI nt
ncRNA Classification and Annotation (Infernal)
Rfam
Filter TCCs u, x
Identify new Putative Coding Transcripts
Non-coding Transcripts
D) Web Environment
Coding Potential Calc (CPC2)
Coding Potential Calc (CPAT)
Coding Potential Calc (CNCI)
Coding Potential Calc (CPPRED)
available at https://osf.io/txeqf/

Figure 1. Schematic representation of the pipeline employed to assemble and characterize the Pb18 transcriptome. Initially, libraries were submitted to quality control checking/filtering, mapped to the Pb18 reference genome, and used to assemble a non-redundant (merged) transcriptome. Elements in this transcriptome were distributed into different Transcript Class Codes and had their normalized expression values (FPKM) evaluated. Transcripts that displayed low FPKM values (< 1) and/or did not belong to Transcription Class Codes (TCCs) "=", "j", "x" or "u" were considered potential artifacts and excluded from further analyses. The remaining transcripts belonging to TCCs "=" and "j" were used to identify putative isomorphs, while elements belonging to TCCs "u" and "x" were evaluated for their coding Potential (CP), using a series of CP calculators. Elements that displayed CP by any of these tools were considered new putative coding transcripts, while elements that displayed no CP by all four tools were evaluated for their potential role as ncRNAs using Infernal. Elements not recognized by Infernal were checked for their potential role as lncRNAs with the aid of FEElncRNA, which also identified the potential partner (or target) genes for each putative lncRNA identified by the software. All elements in the transcriptome were also tentatively annotated, by BLAST, using the NCBI nt database and the Paracoccidioides-specific database ParaDB [16], using an E-value ≤ e-10 as cutoff for identification. See text for details.

References:

1. Jabes DL, de Freitas Oliveira AC, Alencar VC, et al. Thioridazine inhibits gene expression control of the cell wall signaling pathway (CWI) in the human pathogenic fungus Paracoccidioides brasiliensis. Mol Genet Genomics. 2016; 91(3): 1347-62.
2. Andrew, S. FASTQC. A quality control tool for high throughput sequence data (2010).
3. Bolger, A.M. &Lohse, M., Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics**30**, 2114–2120, (2014).
4. Kim, D. Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. Nature methods**12**, 357–360, (2015).
5. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol.**33**, 290–295, (2015).
6. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol.**28**, 511–515 (2010).
7. Waterhouse, R.M. et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. **35**, 543–548 (2017).
8. Hutchins, A.P., Poulain, S., Fujii, H. &Miranda-Saavedra, D. Discovery and characterization of new transcripts from RNAseq data in mouse CD4(+) T cells. Genomics**100**, 303 – 313 (2012).
9. Mohammadin, S., Edger, P.P., Pires, J.C. &Schranz, M.E. Positionally-conserved but sequence-diverged: identification of long

non-coding RNAs in the Brassicaceae and Cleomaceae. BMC Plant Biol. **15**, 217 (2015).

10. Guo, R. et al. First identification of long non-coding RNAs in fungal parasite Nosema ceranae. Apidologie**49,** 660–670 (2018).
11. Severing, E. et al. Arabidopsis thaliana ambient temperature responsive lncRNAs. BMC Plant Biol. **18**, 145–145 (2018).
12. Aciole Barbosa, D. et al. ParaDB: A manually curated database containing genomic annotation for the human pathogenic fungi Paracoccidioides spp. PLoS Negl. Trop. Dis. **13**, e0007576– e0007576 (2019).
13. Tong, X. & Liu, S. CPPred: coding potential prediction based on the global description of RNA sequence. Nucleic Acids Res.**47**, e43–e43 (2019).
14. Kang, Y.J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. **45**, W12–W16 (2017).
15. Sun, L. et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res. **41**, e166–e166 (2013).
16. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment free logistic regression model. Nucleic Acids Res. **41**, e74–e74 (2013).
17. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics**22**, 2933–2935 (2013).
18. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. **46**, D335–D342 (2018).
19. Wucher, V. et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. **45**, e57–e57 (2017).