ISSN 0026-8933, Molecular Biology, 2018, Vol. 52, No. 4, pp. 497–509. © Pleiades Publishing, Inc., 2018.
Original Russian Text © R.F. Gubaev, V.Y. Gorshkov, L.M. Gapa, N.E. Gogoleva, E.P. Vetchinkina, Y.V. Gogolev, 2018, published in Molekulyarnaya Biologiya, 2018, Vol. 52, No. 4, pp. 576–590.

GENOMICS. TRANSCRIPTOMICS

UDC 577.2

Algorithm for Physiological Interpretation of Transcriptome Profiling Data for Non-Model Organisms[1]

R. F. Gubaev[a, b], V. Y. Gorshkov[a, b, *], L. M. Gapa[a, b], N. E. Gogoleva[a, b], E. P. Vetchinkina[c], and Y. V. Gogolev[a, b]

[a] *Kazan Institute of Biochemistry and Biophysics, Federal Research Center "Kazan Scientific Center of RAS", Kazan, 420111 Russia*

[b] *Kazan (Volga Region) Federal University, Kazan, 420008 Russia*

[c] *Institute of Biochemistry and Physiology of Plants and Microorganisms, Russian Academy of Sciences, Saratov, 410049 Russia*

**e-mail: gvy84@mail.ru*

Received June 29, 2017; in final form, July 6, 2017

Abstract—Modern techniques of next-generation sequencing (NGS) allow obtaining expression profile of all genes and provide an essential basis for characterizing metabolism in the organism of interest on a broad scale. An important condition for obtaining a demonstrative physiological picture using high throughput sequencing data is the availability of the genome sequence and its sufficient annotation for the target organism. However, a list of species with properly annotated genomes is limited. Transcriptome profiling is often performed in the so-called non-model organisms, which are those with unknown or poorly assembled and/or annotated genome sequences. The transcriptomes of non-model organisms are possible to investigate using algorithms of *de novo* assembly of the transcripts from sequences obtained as the result of RNA sequencing. A physiological interpretation of the data is difficult in this case because of the absence of annotation of the assembled transcripts and their classification by metabolic pathway and functional category. An algorithm for transcriptome profiling in non-model organisms was developed, and a transcriptome analysis was performed for the basidiomycete *Lentinus edodes*. The algorithm includes open access software and custom scripts and encompasses a complete analysis pipeline from the selection of cDNA reads to the functional classification of differentially expressed genes and the visualization of the results. Based on this algorithm, a comparative transcriptome analysis of the nonpigmented mycelium and brown mycelial mat was performed in *L. edodes*. The comparison revealed physiological differences between the two morphogenetic stages, including an induction of cell wall biogenesis, intercellular communication, ion transport, and melanization in the brown mycelial mat.

Keywords: RNA sequencing, *de novo* transcriptome assembly, transcript annotation, functional classification of expressed genes, visualization of metabolic pathways, morphogenesis of *Lentinus edodes*

DOI: 10.1134/S0026893318040076

INTRODUCTION

Intense development of next-generation sequencing (NGS) methods provides a basis for large-scale genome and transcriptome studies [1]. In particular, NGS is employed in transcriptome profiling, which makes it possible to study expression of the total genome and to characterize the physiological parameters of cells, tissues, or organisms on a global scale rather than by expression of single genes. Describing metabolism on the basis of NGS data is far easier when the genome nucleotide sequence is available and properly annotated for the target organism, including details of particular proteins encoded by the genome. However, annotated complete genome sequences are available for only a limited number of common model organisms, while many studies involve the so-called non-model organisms, whose genomes have not been completely sequenced or annotated. Several problems have to be solved in parallel in transcriptome studies in non-model organisms.

There is no universal algorithm for transcriptome profiling in non-model organisms, and different software have to be selected and combined for the purpose, which is time consuming and requires a high computer performance [2–4]. The programs, calculation parameters, and scripts are usually not described in detail in experimental works aimed at *de novo* transcriptome analysis and identification of differentially expressed genes (DEGs). Moreover, such studies often focus on a limited number of genes, which are

Abbreviations: NGS, next-generation sequencing; ORF, open reading frame; DEG, differentially expressed gene; GO, Gene Ontology; OG, orthologous gene.