doi 10.18699/PlantGen-2025-267

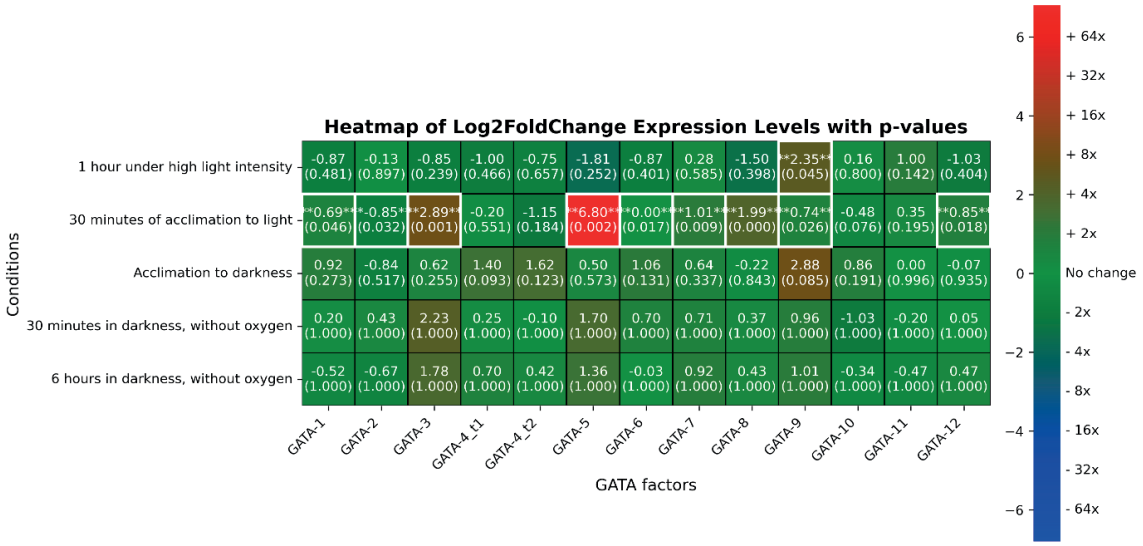
Processing of raw RNA-seq data of *Chlamydomonas reinhardtii* cells using the newly developed PipeSeq program

A.M. Nerezenko*, P.A. Virolainen, E.M. Chekunova
Saint Petersburg State University, Saint Petersburg, Russia
** alexnerezenko@gmail.com*

Motivation and aim: RNA sequencing is a common highly sensitive method for the study of the transcriptome, which allows to measure the level of gene expression. The aim of our work is to analyze the expression of genes encoding GATA family transcription factors (GATA TFs) in green alga *Chlamydomonas reinhardtii*, a model object of photosynthesis genetics. To determine the direction of our further studies, we analyzed the mass of raw RNA-seq data available in public databases. The development of a software tool allowed us to automate the time-consuming process of RNA-seq data processing and analysis.

Materials and methods: Available RNA-seq data from the NCBI database were used to analyze the expression of 12 GATA-coding genes in *C. reinhardtii* in response to various culturing conditions. Pipeline was written in Python programming language using Pandas, Matplotlib, PyQt libraries and bioinformatics tools HISAT2, SAMtools, StringTie, DESeq2. The result of the analysis is the calculation of TPM and FPKM (normalized gene or transcript expression values), and log2FoldChange (effect size estimation – the degree of change in gene or transcript expression in the experimental group compared to the control group). To assess the reliability of the obtained results, p-value is calculated using Student's t-test.

Results: To optimize data processing and reduce analysis time from 21 days to 1 day, an algorithm was developed that includes: 1) data acquisition, 2) alignment to a reference genome, 3) transcript assembly and expression counting, 4) statistical processing, 5) results visualization. Data quality checks are performed at each step of the program: reads alignment evaluation, low-quality data filtering, exclusion of duplicates and control of annotated transcripts. Output data: 1) tables with all calculated values for selected genes and conditions, 2) tables with average expression values by group, 3) log2 tables with FDR-corrected p-values, 4) heat maps illustrating gene expression by condition. An example of the resulting heat map is shown in Figure. The expression level of almost all *C. reinhardtii* GATA-coding genes changes significantly (p-value < 0.05) in response to light acclimation. Most of the publicly available packages of raw RNA-seq data of *C. reinhardtii* either did not contain data for the control group or had an insufficient number of repeats (less than three).



Heat map of *C. reinhardtii* GATA-coding genes expression in response to light conditions

Conclusions: The newly developed PipeSeq program provides full automation of RNA-seq data analysis and demonstrates reliability in the study of the *C. reinhardtii* transcriptome. The obtained results confirmed and supplemented the previously predicted functions of *C. reinhardtii* GATA TFs by other *in silico* methods. Our next step will be to complete the information on the expression of GATA-coding genes in *C. reinhardtii* in the dark conditions with our *in vivo* experiments. The successful application of our PipeSeq program opens up the possibility of its further use in the analysis of raw RNA-seq data available for other species.

Funding: The authors acknowledge Saint Petersburg State University for a research project 124032000041-1.