Optimizing of pre-processing analysis for Illumina RNA-Seq data in *Arabidopsis thaliana*

1st Jana Schwarzerová
Department of Biomedical Engineering
Faculty of Electrical Engineering and Communication,
Brno University of Technology
Brno, Czech Republic

2nd Patricia Janigová
Department of Biomedical Engineering
Faculty of Electrical Engineering and Communication,
Brno University of Technology
Brno, Czech Republic

3rd Martina Dvořáčková
Mendel Centre for Plant Genomics and Proteomics,
CEITEC, Masaryk University,
Brno, Czech Republic
martina.dvorackova@ceitec.muni.cz

4th Wolfram Weckwerth
Molecular Systems Biology (MOSYS), University of Vienna.
Vienna, Austria
Vienna Metabolomics Center (VIME),
University of Vienna,
Vienna, Austria
wolfram.weckwerth@univie.ac.at

**Abstract** — Gene expression analysis through RNA sequencing (RNA-Seq) has revolutionized molecular biology, providing profound insights into the intricate transcriptional landscapes of organisms. *Arabidopsis thaliana*, a widely studied model plant, serves as a cornerstone for investigating fundamental biological and ecology processes. However, accurate interpretation of RNA-Seq data hinges on meticulous pre-processing methods to ensure data integrity and trustworthiness, especially in the context of Illumina sequencing. In this research, we present a comprehensive framework for optimizing pre-processing analysis tailored specifically for *Arabidopsis thaliana* RNA-Seq datasets generated through Illumina sequencing. Our approach encompasses rigorous quality control, precise read alignment, transcript quantification, and normalization procedures crucial for subsequent differential expression analysis. Additionally, we address unique considerations and challenges inherent to *Arabidopsis thaliana* datasets, providing valuable insights for researchers in the field.

**Keywords** — Gene expression analysis, Quality control, *Arabidopsis thaliana*, Transcriptomics

I. INTRODUCTION

Gene expression analysis has become a cornerstone of modern molecular biology research, offering invaluable insights into the complex regulatory mechanisms governing cellular processes [1]. Among the various techniques employed for gene expression profiling, RNA sequencing (RNA-Seq) has emerged as a powerful tool due to its high sensitivity, wide dynamic range, and ability to provide comprehensive transcriptome information. With Illumina sequencing having established itself as one of the primary RNA sequencing platforms in genomics, it continues to be a dominant force. There is an increasing imperative to consistently optimize and discuss its outputs, especially for non-traditional organisms such as various plants utilized in ecological research, which are gaining prominence. Plant science, primarily focused on the study of *Arabidopsis thaliana*, is particularly prominent within the realm of ecology.

*Arabidopsis thaliana* [2], a small flowering plant native to Eurasia, has long served as a model organism for studying plant genomics. Its relatively simple genome, short life cycle, and amenability to genetic manipulation make it an ideal system for investigating fundamental biological processes [3].

Currently, there is a continuous push for advancements in sequencing methodologies to outpace techniques like microarrays. Despite the availability of numerous algorithms for processing microarray data, this method proves to be more costly and complex to perform in wet lab processes, potentially leading to inaccuracies in the data obtained. However, with precise and effective pre-processing analysis, RNA-Seq data has the potential to provide equal or even superior informational value compared to microarray data [4].

Accurate interpretation of RNA-Seq data necessitates rigorous pre-processing steps to ensure data quality and reliability. Quality control measures are crucial for identifying and mitigating potential biases and artifacts introduced during library preparation and sequencing. Furthermore, read alignment to the reference genome, transcript quantification, and normalization are essential steps in the pre-processing pipeline, enabling accurate and reproducible analysis of gene expression patterns.

In this study, we focus on elucidating the pre-processing methods tailored specifically for RNA-Seq data analysis in *Arabidopsis thaliana*. We aim to provide a comprehensive overview of the key steps involved in ensuring the robustness and accuracy of gene expression analysis. Additionally, we discuss specific considerations and challenges unique to *Arabidopsis thaliana* datasets, particularly in the context of systems biology approaches aimed at unraveling the intricacies...
of plant molecular networks based on our previous studies [5], [6]. By elucidating the gold standard of pre-processing protocols, this study aims to contribute to the advancement of research in Arabidopsis thaliana and provide valuable insights into the broader field of plant systems biology.

II. MATERIALS AND METHODS

A. Dataset – Arabidopsis thaliana

Our study analysed on the utilization of Arabidopsis thaliana, a model plant species renowned for its pivotal role in molecular biology research. Specifically, we focused on the Columbia (Col-0) ecotype, a widely studied genetic background, along with specific mutants obtained from the Arabidopsis Biological Resource Center. These mutant lines were selected based on their relevance to the biological processes under investigation, thus enriching the diversity of our experimental material.

The RNA-seq data utilized in our research were sourced from publicly available repositories, specifically the NCBI Gene Expression Omnibus (GEO) database. The dataset was deposited under the GEO Series accession number GSE188493 [7], facilitating transparency and reproducibility in our analyses. The data originated from a study conducted by Zhong et al [7]. All Arabidopsis samples utilized in this investigation belong to the Columbia ecotype (Col-0) and were cultivated at 22°C under LD conditions (16 hours of light, 8 hours of darkness) [7].

The analysed Arabidopsis samples include Col-0 (SRR16892914, SRR16892916, SRR16892917); and mutants: asf1a (SRR16892918, SRR16892919, SRR16892920); hira-1 (SRR16892921, SRR16892922, SRR16892923); fas1 (SRR16892924, SRR16892925, SRR16892926); fas2 (SRR16892927, SRR16892928, SRR16892929); asf1b (SRR16892930, SRR16892931, SRR16892932) and fwa (SRR16892933, SRR16892934, SRR16892935). [7]

B. Pre-processed RNA-Seq pipeline

The gold standard preprocessing workflow [8] was applied to RNA-Seq data from Arabidopsis thaliana. FastQC [9] was utilized to evaluate overall sequence quality, encompassing GC percentage distribution and the presence of overrepresented sequences. The initial step involves merging data for pair-end sequencing, which is optional depending on the data type. In our case, as we utilized pair-end data, we utilized the PEAR tool [10] for the “Merged data” step. Subsequently, conducting additional rRNA filtering is advisable, despite standard wet lab protocols typically including rRNA removal. For this purpose, the SortMeRNA tool [11] proved to be useful (see Fig 1. part 1*).

Quality in base pairs is influenced by their position in the read, leading to lower average quality in later cycles of the sequencing process. To enhance read mapping rates, a common strategy involves removing low-quality bases through quality trimming. The Trimmomatic tool [12] is commonly employed for this task. Following quality trimming and adapter removal, FastQC is rerun to validate the improvements.

The last step involves aligning reads (see Fig 1. part 2*) to a reference genome – TAIR10 [13]. The choice of aligner depends on the type of reference available, with STAR [14] being recommended for genome-based alignment of RNA-Seq data. This alignment process requires a prepared index genome. Once prepared, sample reads can be aligned to it, resulting in the creation of a SAM and BAM format.

Fig. 1. Visualization of RNA-Seq Data Pre-processing Pipeline: This visual guide illustrates the pre-processing pipeline implemented in the study. The pipeline is divided into three color-coded sections. The orange section represents the foundational steps of the pipeline, the blue section indicates potential follow-up analyses, and the yellow section focuses on the main aspect of this study, which is the differential expression analysis. In general, the entire pipeline includes steps such as quality control, read trimming, alignment to a reference genome, transcript quantification, and normalization.

The Fig 1 outlines the sequential steps involved in preparing raw RNA-Seq data for downstream analysis. The main focus of the presented pipeline is the generation of a count table containing information on differential gene expression.

III. RESULTS AND DISCUSSION

In this section, we delve into a pivotal stage within our pre-processing pipeline, which significantly contributes to the robustness and trustworthiness of our RNA-Seq data analysis in Arabidopsis thaliana. This intermediate step plays a crucial role in ensuring the accuracy and integrity of the subsequent analytical procedures.

Our pre-processing pipeline meticulously handles a comprehensive set of 21 raw datasets, each comprising a total of 7 samples, with each sample having 3 replicates. To fortify our analyses and account for variability, each sample is meticulously replicated three times. This approach not only enhances the statistical power of our study but also enables us to confidently discern genuine biological signals from potential artifacts or noise. By meticulously curating and standardizing our datasets through this pre-processing pipeline, we establish a solid foundation for subsequent analyses, ensuring that our findings...
are robust, reproducible, and reflective of the true biological phenomena underlying gene expression in *Arabidopsis thaliana*.

A. Quality control before and after pre-processing

Firstly, the study present rigorous quality control measures before and after pre-processing to ensure the reliability and integrity of the RNA-Seq data. Quality control assessments, such as those conducted with FastQC, enabled us to identify potential issues such as overrepresented sequences and sequence quality of GC distribution, see Fig 2 (before), and Fig 3 (after).

![Fig. 2. Quality Control (QC) report represented GC distribution before pre-processing pipeline](image1)

![Fig. 3. Quality Control (QC) report represented GC distribution after pre-processing pipeline](image2)

After the application of SortMeRNA and Trimmomatic tool, we observed an approximation of the GC distribution curve towards a normal distribution. This shift suggests an improvement in the quality of the RNA-Seq data following pre-processing.

The initial GC distribution, as depicted in Fig. 2, exhibited deviations from the expected normal distribution, indicating potential biases or artifacts in the sequencing data. However, post-processing, as illustrated in Fig. 3, the GC distribution curve appears to align more closely with the expected normal distribution, reflecting a reduction in sequencing errors and an enhancement in data quality. This normalization of the GC distribution is indicative of the effectiveness of the pre-processing steps, particularly in mitigating biases and improving the reliability of the RNA-Seq data for downstream analyses.

B. Comparison of filtering rRNA using default database and special *A. thaliana* database

Subsequently, we applied various pre-processing techniques, including rRNA filtering using SortMeRNA (1.0.1. based on default rRNA eukaryote database offers SortMeRNA and 1.2. based on special rRNA database for *A. thaliana* available on RNAcentral database [15]) and quality trimming with Trimmomatic, to enhance the quality of the sequencing data.

These steps resulted in improved sequence quality metrics and enhanced the overall reliability of the data for downstream analyses.

C. Final Count Table: Transcript Abundance Analysis

The final output was generated using featureCount [16], and R\DESeq normalization [17] was applied to it. In our analysis of transcript abundance, we focused on the raw count data for three specific samples associated with Col-0, as depicted in Table I. These counts, obtained from replicates 1, 2, and 3, offer insight into the variability of transcript abundance across different samples.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Merged data</th>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR16892914</td>
<td>33207850</td>
<td>17151179</td>
<td>16056671</td>
</tr>
<tr>
<td>SRR16892916</td>
<td>43639467</td>
<td>22380486</td>
<td>21258981</td>
</tr>
<tr>
<td>SRR16892917</td>
<td>31546058</td>
<td>16114941</td>
<td>15431117</td>
</tr>
</tbody>
</table>

Following this, we proceeded to assess transcript abundance before and after normalization for three selected samples – see Table II and Table III. In Table II, we present the transcript abundance after normalization for the first replicates. The normalized counts provide a more accurate representation of gene expression levels, facilitating robust downstream analyses.

<table>
<thead>
<tr>
<th>Transcript ID</th>
<th>SRR16892916</th>
<th>SRR16892917</th>
<th>SRR16892918</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G01010</td>
<td>148</td>
<td>146</td>
<td>104</td>
</tr>
<tr>
<td>AT1G01020</td>
<td>358</td>
<td>340</td>
<td>225</td>
</tr>
<tr>
<td>AT1G03987</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>AT1G01030</td>
<td>126</td>
<td>115</td>
<td>82</td>
</tr>
<tr>
<td>AT1G01040</td>
<td>1944</td>
<td>1835</td>
<td>1932</td>
</tr>
</tbody>
</table>

Following normalization, which is crucial for removing systematic biases and enabling fair comparisons across samples,
we proceeded to assess transcript abundance for three selected samples. Normalization adjusts the raw count data to account for differences in sequencing depth and other technical factors, ensuring that gene expression levels are accurately represented. In Table II, we present the transcript abundance before normalization for the first replicates. These raw counts provide an initial insight into gene expression levels but may be influenced by technical variation.

Subsequently, in Table III, we present example of the transcript abundance after normalization for the first replicates. The normalized counts provide a more accurate representation of gene expression levels, as they have been adjusted to account for differences in sequencing depth and other technical factors. This normalization step enables robust downstream analyses, allowing for meaningful comparisons of gene expression levels between samples.

<table>
<thead>
<tr>
<th>Table III. Example of results: transcript abundance after normalization – first replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SRR16892916</strong></td>
</tr>
<tr>
<td>ATIG01010</td>
</tr>
<tr>
<td>ATIG01020</td>
</tr>
<tr>
<td>ATIG03987</td>
</tr>
<tr>
<td>ATIG01030</td>
</tr>
<tr>
<td>ATIG01040</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

In conclusion, our study has successfully implemented a robust pre-processing pipeline specifically tailored for RNA-Seq data analysis in *Arabidopsis thaliana*. By incorporating stringent quality control measures and leveraging specialized databases for rRNA filtering, we have fortified the reliability and fidelity of the sequencing data. These outcomes underscore the pivotal importance of effective pre-processing methodologies in facilitating precise interpretation of RNA-Seq data and propelling advancements in plant molecular biology research.

Moreover, our application of this gold standard approach to transcriptomic data holds significant promise in bridging the gaps between genomic and metabolomic analyses within the panOMICs platform. This integrated approach not only enhances our understanding of the complex regulatory networks underlying plant biology, but also offers a novel insight into the interconnectedness of various molecular processes.

Ultimately, our research endeavors to contribute to the broader landscape of systems biology and pave the way for transformative discoveries in plant science. By elucidating the intricate molecular mechanisms at play in *Arabidopsis thaliana*, we aim to provide valuable insights that will inform future studies and drive innovation in the field of plant molecular biology.

**ACKNOWLEDGMENT**

This work has been supported by grant project FEKT/FIT-J23-8274.

LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures.

**REFERENCES**