BU Bioinformatics MS Internship Final Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Student Name:** | Lulu Jiang | | **BU login:** | luluj@bu.edu |
| **Company/Org. Name:** | Boston University BMSIP | | | |
| **Group Name (e.g. lab):** | Chao Zhang Lab | | | |
| **Supervisor Name:** | Chao Zhang | **Email:** | | chz2009@bu.edu |
| **Start Date:** | 06/03/2024 | **End Date:** | | 08/09/2024 |

This report should summarize your internship experience. There is no specific length requirement but be sure to include sufficient detail to convey your internship activities to someone unfamiliar with your work. Include material that supports the narrative, including tables, figures, publications, etc. Your report will be read by program faculty and used to assess your internship grade along with your supervisor’s feedback.

# Background

Describe the background of the project and the hypothesis you sought to test, knowledge you sought to gain, or a specific product (e.g. analysis pipeline) you created

PAX3 is a critical transcription factor that plays a significant role in melanocyte development and has been implicated in the progression of melanoma. This gene belongs to the paired box (PAX) family of transcription factors, which are essential regulators of tissue development and cellular differentiation. In the context of melanoma, PAX3 has been found to be overexpressed in a subset of cases, approximately 1-2% of melanomas, suggesting its potential as a therapeutic target.

PAX3's function as a transcription factor allows it to regulate the expression of numerous genes involved in cell proliferation, survival, and migration. However, recent studies have suggested that PAX3's regulatory role may extend beyond direct transcriptional control. There is growing evidence that PAX3 might also influence post-transcriptional processes, including nonsense-mediated decay (NMD).

Nonsense-mediated decay is a crucial quality control mechanism in cells that targets and degrades mRNAs containing premature termination codons (PTCs). This process helps prevent the production of truncated proteins that could have deleterious effects on cellular function. The interplay between transcription factors like PAX3 and NMD is an emerging area of research, with potential implications for understanding gene regulation in both normal and disease states.

Our primary hypothesis is that PAX3 influences nonsense-mediated decay in melanoma cells. We hypothesize that PAX3 may alter the efficiency of NMD for specific transcripts, thereby adding another layer of complexity to its gene regulatory functions. This hypothesis is based on preliminary data suggesting that some PAX3 target genes exhibit expression patterns consistent with NMD regulation.

The main goal of our study is to characterize how PAX3 affects gene expression and splicing patterns, both with and without NMD inhibition. To achieve this, we designed a comprehensive experimental approach:

1. We employed cycloheximide (CHX) treatment to inhibit NMD, allowing us to distinguish between transcripts directly regulated by PAX3 and those affected through NMD modulation.

2. We performed RNA sequencing to obtain a genome-wide view of gene expression changes under different conditions (PAX3 overexpression versus control, with and without NMD inhibition).

3. We complemented our RNA-seq data with quantitative PCR (qPCR) for selected genes of interest to validate our findings.

4. We analyzed both gene-level and isoform-level changes, allowing us to detect subtle alterations in splicing patterns that might be influenced by PAX3 and NMD.

# Activities

Summarize your activities, including writing code, running analyses, and meeting with people at your organization

During my internship, I was deeply involved in analyzing RNA sequencing data to understand how PAX3 affects gene expression and splicing. This process involved several steps, starting with processing the raw sequencing data. I used tools like FastQC for quality control, Cutadapt for trimming adapters, and STAR for aligning reads to the human genome. To quantify gene expression, I used both featureCounts and RSEM, which gave us a comprehensive view of gene activity.

A significant part of my work involved differential expression analysis using DESeq2 in R. This allowed us to compare gene expression between different experimental conditions, such as PAX3 overexpression versus control, both with and without CHX treatment to inhibit nonsense-mediated decay. We focused on genes that showed significant changes in expression, using strict statistical criteria.

I also delved into isoform-level analysis using IsoformSwitchAnalyzeR. This tool helped us detect changes in how genes were spliced under different conditions. We looked at how these splicing changes might affect protein function or make transcripts more or less sensitive to nonsense-mediated decay.

To make our findings more accessible, I created various visualizations. These included PCA plots to show how samples clustered based on their overall gene expression patterns, volcano plots to highlight significantly changed genes, and Venn diagrams to compare sets of differentially expressed genes across conditions.

Throughout the project, I participated in weekly lab meetings where I presented my progress, discussed challenges, and got feedback from colleagues. These discussions were invaluable in shaping our analysis approach and interpreting results. Every two weeks or so, my PIs, Dr. Lang and Dr. Zhang helped us adjust our experimental design and analysis plans as new insights emerged. We also spent time discussing potential follow-up experiments and thinking about future directions for the project.

# Results

Summarize the results of your activities, describing what you learned or created

Principal Component Analysis (PCA) demonstrated distinct clustering of samples based on experimental conditions, with PAX3 overexpression and cycloheximide (CHX) treatment showing separate effects on global gene expression patterns. Sample D2 was identified as an outlier and subsequently removed from further analyses to maintain data integrity.

Differential gene expression analysis revealed a significant number of genes affected by PAX3 overexpression, with a predominance of downregulated genes, indicating a potential role of PAX3 as a transcriptional repressor in this context.

Comparison of gene expression profiles with and without CHX treatment identified a subset of genes downregulated by PAX3 overexpression under normal conditions, but upregulated when nonsense-mediated decay (NMD) was inhibited by CHX. This pattern suggests potential indirect regulation of these genes by PAX3 through modulation of the NMD pathway.

ATF3 exhibited notable expression changes in response to both PAX3 overexpression and NMD inhibition. PAX3 overexpression significantly decreased ATF3 expression under normal conditions, while NMD inhibition by CHX resulted in substantial upregulation of ATF3, independent of PAX3 expression status. These observations indicate high NMD sensitivity of ATF3 and suggest PAX3 may primarily influence its expression through NMD modulation.

Isoform-level analysis using IsoformSwitchAnalyzeR detected alterations in splicing patterns for multiple genes upon PAX3 overexpression. Some of these splicing changes produced isoforms with differential sensitivity to NMD, introducing an additional layer of complexity to PAX3's regulatory function.

Validation of RNA-seq results by qPCR showed consistency in overall expression trends, though discrepancies in the magnitude of changes were observed between the two methods.

# Professional Impact

Describe how your experiences and activities in this internship have helped in your professional development goals, and how you expect the experience will aid you in your next steps

This internship has significantly advanced my professional development in bioinformatics and genomics research. The experience has enhanced my technical skills in RNA-seq analysis, from quality control to advanced statistical analyses. I've gained proficiency in using essential bioinformatics tools such as FastQC, STAR, DESeq2, and IsoformSwitchAnalyzeR, and improved my R programming skills, particularly in the context of large-scale genomics data analysis. This experience has also exposed me to the realities of conducting research in a professional setting, including project management, troubleshooting experimental issues, and adapting analyses based on emerging results. These skills are transferable to many future roles in academia.

Looking ahead, this internship has solidified my interest in genomics research and provided me with a strong foundation for pursuing graduate studies. I look forward to applying my skills and experiences in my PhD studies.

# Challenges & Lessons

Describe any challenges you encountered, and any lessons you learned along the way

A significant challenge I encountered during this internship was the necessity for a more comprehensive approach to quality control and data processing than I had previously employed in academic settings. This was particularly evident in two key areas: the analysis of FastQC results and the interpretation of log files from feature counting

Regarding FastQC, I initially underestimated the importance of thoroughly examining these reports. In school projects, I often performed only a cursory review of FastQC output; however, this internship taught me that subtle quality issues could have substantial downstream effects on the analysis. For instance, I initially overlooked a slight drop in base quality scores towards the end of the reads in one of our samples, which led to inconsistencies in our initial differential expression results.

Another important lesson I learned about was the critical nature of log files and terminal output—particularly from the Subread package—when performing feature counts. I did not initially realize the importance of closely monitoring these outputs for each sample and its replicates. It became clear that ensuring consistency in metrics such as the number of uniquely mapped reads and the assignment rate across all samples was crucial for the validity of our downstream analyses.

In one instance, I failed to notice a significant discrepancy in the assignment rate between replicates of a particular condition. This oversight led to skewed results in our differential expression analysis, which we only caught later in the process. Consequently, this required us to backtrack, reassess our raw data, and rerun our analyses, costing us valuable time. These experiences taught me several valuable lessons:

Attention to detail is crucial in bioinformatics. Even small quality issues or inconsistencies can propagate through the analysis pipeline and affect final results.

# Feedback

Your answers to these questions will help the MS program better understand your internship experiences and enable us to better prepare our students in the future. *Note: your responses won’t be shared with your internship supervisor or organization and will be kept confidential by the program. Please be honest!*

**Besides your supervisor, how many other people did you interact with? What were their roles (e.g. scientist, staff, IT, etc)?**

**Please list the bioinformatics tools, software packages, and programming languages that were the most important for your internship. e.g. STAR, Jupyterlab, bash, python, RShiny, samtools, Nextflow, etc. List as many as you deem appropriate.**

**What were your impressions about the organization where you did your internship overall? What were some things you liked? Was there anything you didn’t like? Would you recommend other MS students do internships there?**

**What aspects of your internship do you feel the program prepared you for?**

**How could the program have prepared you better for your internship experience?**

**Considering the program overall, what were aspects that you liked or thought we did well?**

**Considering the program overall, what are some things we could have done better?**