FINAL ANALYSIS REPORT

Whole Transcriptome Sequencing & Analysis

On Illumina Novaseq X Plus Platform

Submitted to:

Client Name

Client Organization

1 Project Details

This report details the whole transcriptome sequencing and bioinformatics analysis for the project 240540. Sequencing was carried out using high-throughput technology to provide deep coverage of the transcriptome.

Table 1.1: Project Specifications

SERVICE TYPE	PLATFORM	READ LENGTH	DATA THROUGHPUT
Transcriptome Sequencing	Illumina Novaseq X Plus	2 X 150 PE	~05 GB / Sample

2 Sample Information

A total of 0 samples were received for library preparation and sequencing. The samples were collected from plant root tissues and maintained under strict temperature controls during transport.

Table 2.1: Sample Overview

TYPE	NO. OF SAMPLES	SHIPPING CONDITION	LIBRARIES PREPARED
Plant Root	0	-20°C	0

Sample Nomenclature:

3 Methods

3.1 Isolation and Quantitative analysis of RNA:

RNA samples were extracted from leaf sample using PVP extraction method. RNA quantity was measured using Qubit® 4.0 fluorometer and quality were analyzed by using 1% agarose gel.

3.2 Preparation of Library

The paired-end sequencing library was prepared using KAPA mRNA HyperPrep Kit for Illumina. (CAT #KK8581). The library preparation process was initiated with 500-1000 ng input. mRNA enrichment was performed as per user manual and mRNA was subjected to fragmentation, first & second-strand cDNA synthesis, end-repair, 3´ adenylating, adapter ligation, selective enrichment of adapter-ligated DNA fragments through PCR amplification, followed by validation of Library on Agilent 4150 tape station. The final library was pooled with other-samples, denatured & loaded on to flow cell. On the flow cell, cluster generation & sequencing was performed using Illumina Novaseq X plus platform to generate 2×150bp paired-end (PE) read

3.3 Quantification using Qubit 4.0 Fluorometer

Table 3.1: RNA Quantification (Qubit)

S.N	SAMPLE ID	CONC. (NG/ML)	VOLUME (ML)	YIELD (MG)	REMARKS
No Qubit data available					

3.4 Cluster Generation and Sequencing

After obtaining the Qubit concentration for the library and the mean peak size from Tape Station profile, library will be loaded onto illumine Novaseq X Plus for cluster generation and sequencing. Paired-End sequencing allows the template fragments to be sequenced in both the forward and reverse directions. The library molecules will bind to complementary adapter oligos on paired-end flow cell. The adapters are designed to allow selective cleavage of the forward strands after re-synthesis of the reverse strand during sequencing. The copied reverse strand is then used to sequence from the opposite end of the fragment.

4 Library Quality Control

Amplified libraries were analyzed on a TapeStation 4150 (Agilent Technologies) using High Sensitivity D1000 ScreenTape® to determine mean peak sizes and ensure proper fragment distribution.

Figure 4.1: Mean library peak size distribution (bp) validated via TapeStation 4150.

4.1 Wet Lab Inferences

The average sizes of the generated libraries were determined to be respectively for the 0 samples. All libraries met the concentration and size requirements for high-quality sequencing on the Novaseq X Plus platform.

5 Bioinformatics Analysis

5.1 Bioinformatics Workflow

Raw data was filtered to remove adapters and low-quality bases. Clean reads thus produced were used for analysis. Reference guided transcript assembly was performed for samples first by mapping clean reads on reference genome using STAR (v2.7.10a) aligner and then performing transcript assembly by StringTie (v 2.2.1). A consensus set of transcripts was obtained using SringTie merge function which merges together all the gene structures found in any of the samples. Transcript abundance was then estimated for individual sample using merged transcript consensus again using StringTie and read counts thus obtained for each transcript were taken as input for differential expression analysis using edgeR package. Gene Ontology and pathway analysis of the significantly differentially expressed transcripts were performed using Blast2go cli and KEGG-KAAS server respectively. Overall bioinformatics workflow is graphically represented in figure above and software version along with its use in table below.

5.2 Detail of Software used for analysis

Table 5.1: Software Stack

SOFTWARE	VERSION	APPLICATION
Trim Galore	0.6.4	Adapter and low-quality data removal
STAR	2.7.10	Mapping of reads to reference genome
Stringtie	2.2.1	Reference based assembly
gffcompare	0.12.6	Comparing reference genome gff with merged gff file generated from stringtie
gffread	0.12.7	Fetching fasta sequences corresponding to gff file generated after reference-based assembly by stringtie
Blastx	2.30.0+	Similarity search against NCBI NR database
Blast2go cli	1.5	GO mapping and annotation
edgeR	3.6.2	Differential expression analysis and its visualization
KEGG KAAS	Webserver	Pathway analysis
clusterProfiler	4.12.6	Over-representation (or enrichment) analysis

6 Data Statistics

Next-generation sequencing statistics for raw data generated on the Novaseq platform are provided in the table below.

Table 6.1: Sequencing Data Statistics

SAMPLE	TOTAL READS	TOTAL BASES	DATA (GB)

7 Reference Genome

Based on information received from client, the reference genome of **Organism Name** (GCF_000499845.2) was used for alignment and annotation. The data was downloaded from ICRISAT Pearl Millet resources. The corresponding GFF file contains locus and structural information regarding 0 genes.

Table 7.1: Reference Genome Statistics

DESCRIPTION	STATS
Total Chromosomes/Scaffolds	7
Total Genome Length (bp)	1,564,537,551
Mean Scaffold Size	223,505,364
Max Scaffold Size (bp)	300,907,625
Total Genes Annotated	0

8 Mapping & Assembly

Alignment to the reference genome allows the identification of reads originating from specific loci. Mapping results show high percentages of uniquely mapped reads for most samples.

Table 8.1: Mapping Statistics

SAMPLE NAME	TOTAL CLEAN READS	NO. OF MAPPED READS	% MAPPED	# UNIQUELY MAPPED	% UNIQUELY MAPPED

A total of 66758 transcripts were identified in the merged assembly with a mean size of 1,392 bp.

Table 8.2: Assembly Statistics

SAMPLE NAME	# ASSEMBLED TRANSCRIPTS	TOTAL (BP)	MEAN SIZE (BP)
merged.fasta	66,758	92,931,869	1,392

9 Differential Expression

Differential gene expression was inferred between samples by applying the R package **edgeR**. This provides statistical significance based on a negative binomial distribution model.

Table 9.1: Differential Expression Statistics

COMPARISON	TOTAL DEGS	SIG. DOWN	SIG. UP	TOTAL SIGNIFICANT
Comparison2	27819	1349	1220	**2569**
Comparison1	28070	966	803	**1769**

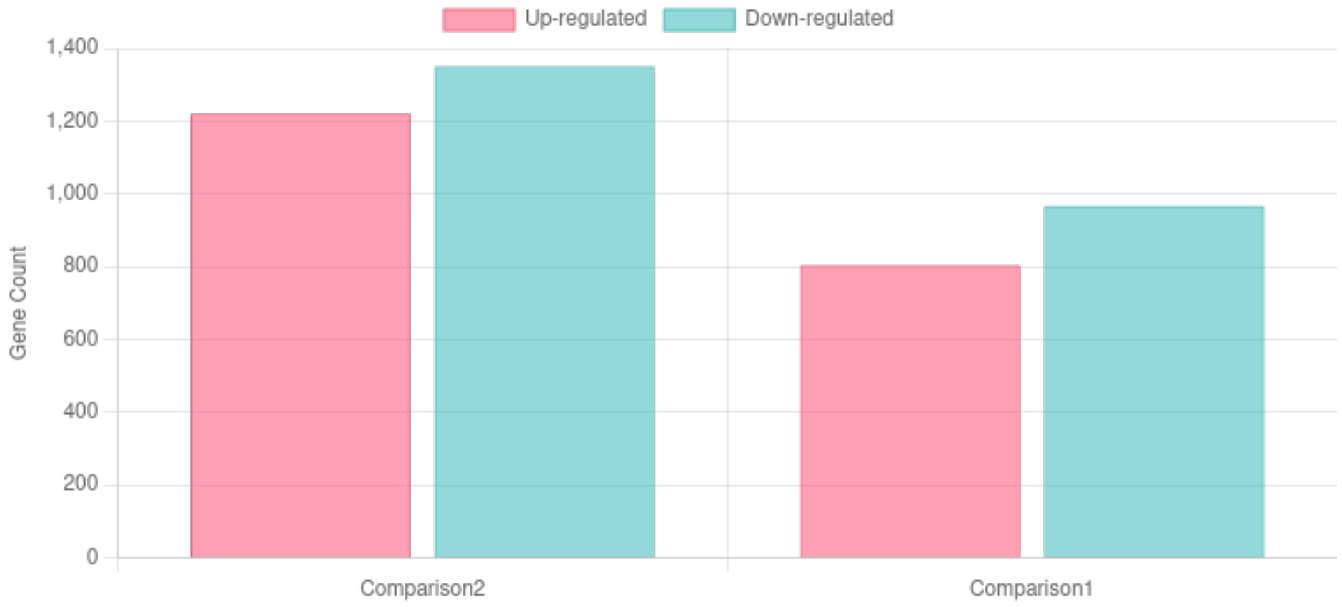

Figure 9.1: Summary of Significantly Upregulated and Downregulated genes per comparison.

10 Functional Annotation

Mapping of significantly differentially expressed transcripts to biological pathways was performed using **KEGG-KAAS**. Results represent metabolic pathways of major biomolecules.

Table 10.1: KEGG Pathway Statistics

LEVEL 1 CATEGORY	LEVEL 2 SUB-CATEGORY	COUNTS

11 Deliverables

Final deliverables are organized in the following directory structure:

```
Deliverables
├── 01_Raw_Data
│   ├── NGS_240540_md5sum.txt
│   ├── NGS_240540_Raw_Stats.txt
│   └── (Fastq files)
├── 02_reference_genome_and_gff
│   ├── (Reference Genome FASTA)
│   └── (Reference GFF/GTF)
├── 03_transcript_assembly_gtf
│   ├── merged_transcripts.gtf
│   ├── gffcompare_results.gtf
│   └── (Individual Sample GTFs)
├── 04_transcript_sequences_fasta
│   ├── Merged.fasta
│   └── (Sample Aligned Transcripts)
├── 05_differential_expression_analysis
│   ├── Comparison*_DGE.xlsx
│   ├── Comparison*_Heatmap.png
│   └── Comparison*_MA_and_Volcano.pdf
├── 06_Significant_DGE_GO
│   └── Comparison*_Significant_DGE_with_GO.xlsx
├── 07_Significant_DGE_pathways
│   ├── Comparison*_Significant_DGE_Pathways.xlsx
│   └── (Pathway Images)
├── 08_Significant_DGE_Enrichment
│   ├── (GO Enrichment Plots & Tables)
│   └── (KEGG Enrichment Plots & Tables)
└── Readme.txt
```

[END OF REPORT]