**ASSIGNMENT 3**

**Taniya Pal (s373094)**

**Quality Control of Reads (FASTQC Visualization and Trimming)**

The fastq reads provided were visualized with the tool FASTQC. The FASTQC tools usually checks the following modules in a fastq file:

* Basic Statistics
* Per Base Sequence quality
* Per Sequence Quality Scores
* Per Base Sequence Quality
* Per Sequence GC Content
* Per Base N Content
* Sequence Length Distribution
* Duplicate Sequence
* Overrepresented Sequence
* Adapter Content
* Kmer Content

In the given fastq files, all the modules pass, except Per Base Sequence Quality and Sequence Duplication Levels.

To improve the quality of the fastq files, Trimmomatic software was used. It could be visualized the FASTQC interface, the bias in the Per Base Sequence Quality lasted for about 10 base pairs. So, if the first 10 base pairs could be removed, the sequence quality may improve. The following command was used to trim the first 10 base pairs in Trimmomatic:

module load Trimmomatic

for file in /scratch/s373094/dataForAssignment/\*r1.fq

do

sample="${file%r1.fq}"

java -jar $EBROOTTRIMMOMATIC/trimmomatic-0.38.jar PE \

-phred33 "$sample"r1.fq "$sample"r2.fq \

"$sample"r1\_paired.fq "$sample"r1\_unpaired.fq \

"$sample"r2\_paired.fq "$sample"r2\_unpaired.fq \

ILLUMINACLIP:/apps/software/Trimmomatic/0.38-Java-1.8.0\_92/adapters\

/TruSeq3-PE.fa:2:30:10 \

HEADCROP:10

The headcrop command in Trimmomatic software removes the specified number of base pairs from the beginning.

The warning in the sequence duplication levels were ignored as duplication levels are usullay seen higher in RNA Sequencing. This can be due to the high expression of the genes during transcription.

**Indexing:**

Thereafter, indexing was done on the trimmed reads. There were two aligner tools used, namely, HISAT2 and STAR aligner, so indexing was done with both the two aligners. HISAT2 aligner was used as Musich et al stated that HISAT2 and STAR are among the aligners with second best alignment rate after Bowtie. The command use for indexing for STAR and HISAT2 are as follows:

module load STAR

STAR --runThreadN 4 --runMode genomeGenerate \

--genomeDir /scratch/s373094/dataForAssignment/STAR\_genome\_index \ -genomeFastaFiles \

/scratch/s373094/dataForAssignment/potato\_chr12.fa \ –sjdbGTFtagExonParentTranscript \ /scratch/s373094/dataForAssignment/potato\_annot\_chr12.gff

module load HISAT2

hisat2-build /scratch/s373094/dataForAssignment/potato\_chr12.fa \

/scratch/s373094/dataForAssignment/genome\_index/genome

**Alignment:**

Alignment was done with both STAR and HISAT2. The command line for STAR and HISAT2 alignment is as follows:

module load STAR

STAR --runThreadN 4 \

--genomeDir /scratch/s373094/dataForAssignment/STAR\_genome\_index \

--sjdbGTFfile /scratch/s373094/dataForAssignment/potato\_annot\_chr12.gff \

--sjdbGTFtagExonParentTranscript Parent \

--outSAMtype BAM SortedByCoordinate \

--readFilesIn /scratch/s373094/dataForAssignment/ST\_w0\_R1.chr12.r1\_paired.fq \

/scratch/s373094/dataForAssignment/ST\_w0\_R1.chr12.r2\_paired.fq \

--outFileNamePrefix ST\_w0\_R1

module load HISAT2

hisat2 -x /scratch/s373094/dataForAssignment/genome\_index/genome \

-1 /scratch/s373094/dataForAssignment/ST\_w0\_R1.chr12.r1\_paired.fq \

-2 /scratch/s373094/dataForAssignment/ST\_w0\_R1.chr12.r2\_paired.fq \

-U /scratch/s373094/dataForAssignment/ST\_w0\_R1.chr12.r1\_unpaired.fq \

-S /scratch/s373094/dataForAssignment/alignment\_1\_output/ST\_w0\_R1.sam

A comparision was made between the uniquely mapped reads of both the aligners with the samtools flagstat command. On running the samtools command, it displayed the following output after aligning one of the fastq paired end reads:

**For HISAT2:**

samtools flagstat alignment\_1\_output/ST\_w0\_R1.sam

[W::sam\_read1] Parse error at line 447704

[bam\_flagstat\_core] Truncated file? Continue anyway.

447700 + 0 in total (QC-passed reads + QC-failed reads)

35714 + 0 secondary

0 + 0 supplementary

0 + 0 duplicates

438452 + 0 mapped (97.93% : N/A)

411986 + 0 paired in sequencing

205993 + 0 read1

205993 + 0 read2

340746 + 0 properly paired (82.71% : N/A)

402476 + 0 with itself and mate mapped

262 + 0 singletons (0.06% : N/A)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)

**For STAR:**

samtools flagstat ST\_w0\_R1Aligned.sortedByCoord.out.bam

3093726 + 0 in total (QC-passed reads + QC-failed reads)

243120 + 0 secondary

0 + 0 supplementary

0 + 0 duplicates

3093726 + 0 mapped (100.00% : N/A)

2850606 + 0 paired in sequencing

1425303 + 0 read1

1425303 + 0 read2

2850606 + 0 properly paired (100.00% : N/A)

2850606 + 0 with itself and mate mapped

0 + 0 singletons (0.00% : N/A)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)

We can see the percentage of uniquely mapped reads for STAR (100 %) is more than HISAT2 (97.93%). So the downstream processing was continued on the STAR alignment ouputs for each replicate of the two samples.

**Counting reads mapping to each gene in reference genome:**

The following step was done with htseq count command where the uniquely, non uniquely or reads with none of the either feature are segregated and counted. It gives back a counts file which can be used further for analysis of differential expression of the genes.

The following command was run for analysis with htseq count:

htseq-count --stranded=no \

-f bam \

ST\_w0\_R1Aligned.sortedByCoord.out.bam \ ST\_w0\_R2Aligned.sortedByCoord.out.bam \

ST\_w0\_R3Aligned.sortedByCoord.out.bam \

ST\_w1\_R1Aligned.sortedByCoord.out.bam \ ST\_w1\_R2Aligned.sortedByCoord.out.bam \ ST\_w1\_R3Aligned.sortedByCoord.out.bam potato\_annot\_chr12.gff >\ HTSeq/HTSeq\_counts.txt

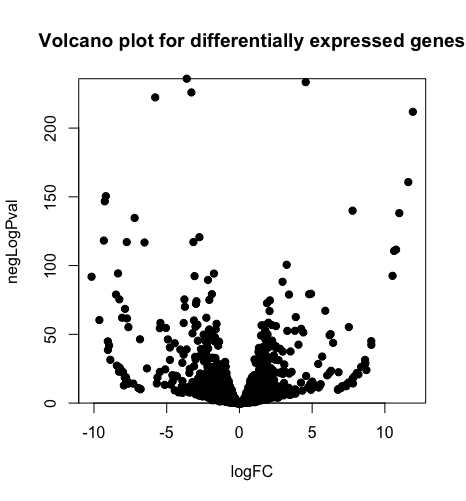
The “HTSeq\_counts.txt” file was used for analysis of differential expression of the genes using R.

**Diffrential Expression ANaysis:**

The package “edgeR” in R was used for the analysis of differential expression of the genes. In the begging of the script, the file “HTSeq\_counts.txt” was imported into the environment of R. Then the DGEList command was used to store the data in a “simple list based data object.” The filterByExpr command was used to filter the lowly expressed genes from the data. Thereafter the library size was normalized using “calcNormFactors”. The common and tagwise dispersion was estimated and finally the differentially expressed genes were tested for using “exactTest”. The down and up regulated genes were filtered and stored in files called “Downregulated\_genes.csv” and “Upregulated\_genes.csv” respectively.

By calculating the log of fold change and negative p values of the top 100 up regulated genes, a volcano plot was generated.

Chart, scatter chart

Description automatically generatedA PCA plot was also generated for viewing how the replicates were clustered and identify any potential outlier.

From the PCA plot, we can conclude, there are no outliers among the replicates, as they are clustered together.

**Annotation:**

The top 100 up regulated genes were annotated using biomart. The database “plants\_mart” was used and dataset “stuberosum\_eg\_gene” was used giving “plants.ensembl.org” as host. The annotation of the top 100 up regulated genes can be found in the file “annot\_biomart\_potato”.

**Variant Calling:**

The variant calling was done with bcftools. The following command was used to generate the bcf files:

bcftools mpileup -f potato\_chr12.fa ST\_w0\_R1Markedduplicates.bam > ST\_w0\_R1.marked.raw.bcf

The marked duplctes bam file was generated using the tool picard. The command MarkDuplicates was used to generate this file.

java -jar $EBROOTPICARD/picard.jar MarkDuplicates I=ST\_w0\_R1Aligned.sortedByCoord.out.bam O=ST\_w0\_R1Markedduplicates.bam M=ST\_w1\_R3Marked\_dup\_metrics.txt

Hence, the bcf files were converted to vcf files using bcftools.

bcftools call -vc --ploidy 1 ST\_w0\_R1.marked.raw.bcf > ST\_w0\_R1.marked.raw.vcf

The variants were viewed using SnpEff tool.

Chart, bar chart

Description automatically generated

Fig 1

Chart, bar chart

Description automatically generated

Fig 2

Fig 1 shows the variations of the genes from the referene genome in week 0 sample. Fig 2 shows the same for week 1 sample. We can see there are downstream variations introduced in the week 1 sample which are not present in the week 0 sample. We can conclude these downstream variations in week 1 sample might be responsible for dormancy break and sprouting in Solanum tuberosum.

Thereafer the gene IDs for the downstream variants were filtered using the following command:

cat ST\_w1\_R1.marked\_anno.vcf | grep -e "downstream\_gene\_variant" | cut -f8 | cut -d'|' -f4 > downstream\_variants.txt

The corresponding gene IDs can be found in the file “downstream variants.txt”.

**References:**

* Musich, R., Cadle-Davidson, L. and Osier, M.V., 2021. Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science*, *12*.
* egdeR: differential analysis of sequence read count data:User’s Guide