15 datasets of 3 cancer types namely bone cancer (SRR25192668, SRR25192669, SRR25192670, SRR25192671, SRR25192672), melanoma (SRR29269580, SRR29269581, SRR29269582, SRR29269583, SRR29269584) and uterine cancer (SRR14146030, SRR14146031, SRR14146032, SRR14146033, SRR14146034) were collected from SRA ncbi using the command

***prefetch (accession\_no.sra)***

then the fastq files were generated using the command

***fastq-dump –split-files (accession\_no.sra)***

the quality of the reads were checked using fastqc

***fastqc \*.fastq***

then the fastq files were trimmed using trimmomatic tool

***java -jar trimmomatic-0.39.jar PE -threads 6 \*\_1.fastq \*\_2.fastq forward\_paired.fastq forward\_unpaired.fastq reverse\_paired.fastq reverse\_unpaired.fastq TRAILING:3 -phred33*** (for paired ends)

***java -jar trimmomatic-0.39.jar SE -threads 6 \*.fastq trimmed.fastq LEADING:3 TRAILING:3 -phred33*** (for single reads)

the quality of the trimmed fastq files were checked again

***fastqc forward\_paired.fastq reverse\_paired.fastq*** (for paired end)

***fastqc trimmed.fastq*** (for single ends)

then the multiqc was done for comparison of the untrimmed and trimmed fastq files

***multiqc .***

The grch38 which is the reference genome was downloaded from the hisat2 website and Homo\_sapiens.GRCh38.113.gtf file from the ensemble. After the extracting both the files alignment was done using hisat2

***hisat2 -x grch38/genome -1 forward\_paired.fastq -2 reverse\_paired.fastq -S aligned.sam --rna-strandness RF -p 8*** (for paired ends)

***hisat2 -x grch38/genome -U trimmed.fastq -S aligned.sam --rna-strandness R -p 8*** (for single ends)

the overall alignment rate for all the fastq files were above 91% and the strandness specificity was checked by infer experiment tool in the galaxy which gave RF for paired ends and R for single ends.

The sam tool was used to convert sam file to bam file and to sort and index the bam file

***samtools view -@ 8 -bS aligned.sam > aligned.bam***

***samtools sort -@ 8 -o sorted.bam aligned.bam***

***samtools index sorted.bam***

for extraction of counts from the bam file, featureCounts was used given by the command

***featureCounts -a Homo\_sapiens.GRCh38.110.gtf/Homo\_sapiens.GRCh38.110.gtf -o gene\_counts.txt -T 8 -p -B -C sorted.bam*** (for PE)

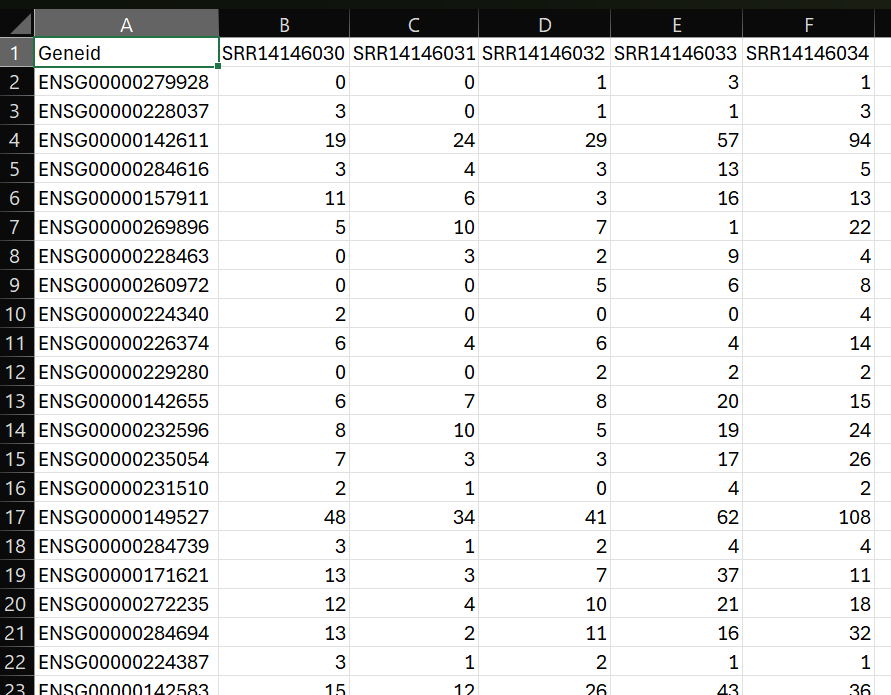
***featureCounts -a Homo\_sapiens.GRCh38.110.gtf/Homo\_sapiens.GRCh38.110.gtf -o gene\_counts.txt -T 8 sorted.bam*** (for SE)

The gene\_counts.text was opened in excel sheet and saved as CSV format.

**DESeq2 Analysis**

**Uterine Cancer:**

Based on gene counts CSV file of SRR14146030, SRR14146031, SRR14146032, SRR14146033 and SRR14146034, a counts csv file was created as follows:



Metadata for the uterine cancer was as follows:

A screenshot of a computer

Description automatically generated

The codes for DESeq2 analysis were as follows

*setwd("D:/project/uterine\_cancer")*

*library(DESeq2)*

*counts <- as.matrix(read.csv("counts.csv", row.names=1))*

*coldata <- read.csv("metadata.csv", row.names=1)*

*dds <- DESeqDataSetFromMatrix(countData = counts,*

*colData = coldata,*

*design = ~treatment)*

*dds <- DESeq(dds)*

*results <- results(dds)*

*head(results)*

*A screenshot of a computer

Description automatically generated*

*plotMA(results, main="DESeq2 MA Plot")*

*A graph with numbers and lines

Description automatically generated*

*library(ggplot2)*

*results$log2FoldChange <- as.numeric(results$log2FoldChange)*

*ggplot(results, aes(x=log2FoldChange, y=-log10(padj))) +*

*geom\_point(alpha=0.4) + xlim(c(-5, 5)) + ylim(c(0, 10))*

38513 rows containing missing values or values outside the scale range were removed

*A graph with a line drawn on it

Description automatically generated*

*library(pheatmap)*

*topGenes <- head(order(results$padj), 20)*

*normalizedCounts <- counts(dds, normalized=TRUE)*

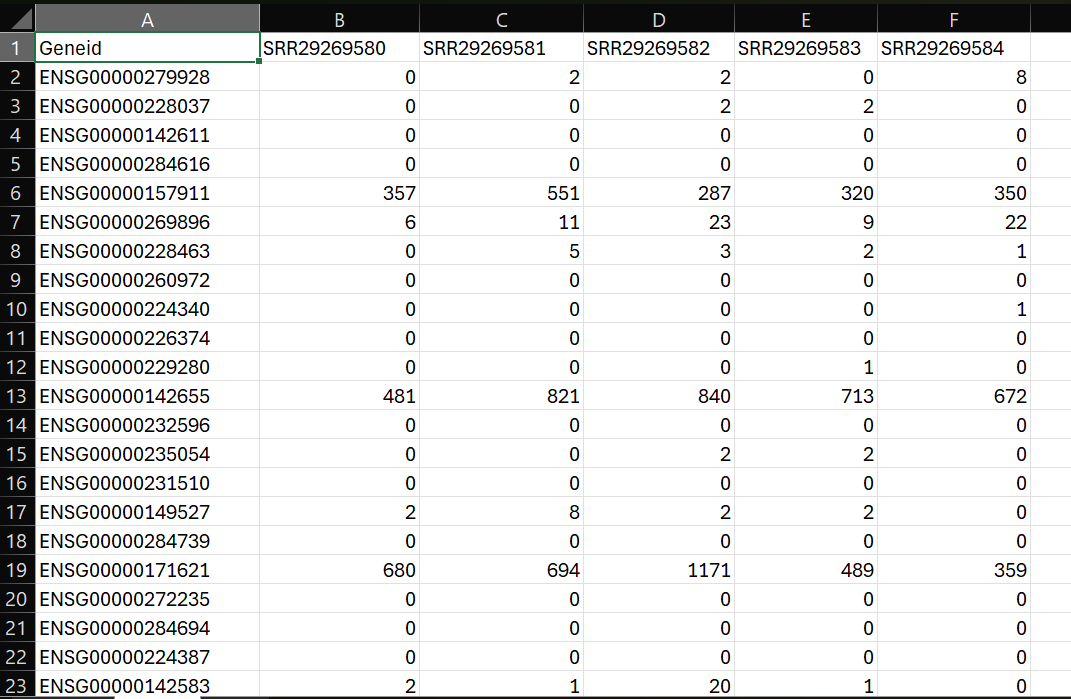
*pheatmap(normalizedCounts[topGenes,])*

*A diagram of a number and a number

Description automatically generated with medium confidence*

**Melanoma Cancer:**

Based on gene counts CSV file of SRR29269580, SRR29269581, SRR29269582, SRR29269583 and SRR29269584, a counts csv file was created as follows:



Metadata for the melanoma cancer was as follows:

A screenshot of a computer

Description automatically generated

The DESeq2 analysis were as follows:

*setwd("D:/project/melanoma\_cancer")*

*library(DESeq2)*

*counts <- as.matrix(read.csv("counts.csv", row.names=1))*

*coldata <- read.csv("metadata.csv", row.names=1)*

*dds <- DESeqDataSetFromMatrix(countData = counts,*

*colData = coldata,*

*design = ~treatment)*

*dds <- DESeq(dds)*

*results <- results(dds)*

*head(results)*

*A screenshot of a computer screen

Description automatically generated*

*plotMA(results, main="DESeq2 MA Plot")*

*A graph of a number of numbers

Description automatically generated with medium confidence*

*library(ggplot2)*

*results$log2FoldChange <- as.numeric(results$log2FoldChange)*

*ggplot(results, aes(x=log2FoldChange, y=-log10(padj))) +*

*geom\_point(alpha=0.4) + xlim(c(-5, 5)) + ylim(c(0, 10))*

44897 rows containing missing values or values outside the scale range were removed

*A graph of a graph showing a graph of a graph

Description automatically generated with medium confidence*

*library(pheatmap)*

*topGenes <- head(order(results$padj), 20)*

*normalizedCounts <- counts(dds, normalized=TRUE)*

*pheatmap(normalizedCounts[topGenes,])*

*A diagram of a number

Description automatically generated with medium confidence*

**Bone Cancer:**

Based on gene counts CSV file of SRR25192668, SRR25192669, SRR25192670, SRR25192671 and SRR25192672, a counts csv file was created as follows:

A screenshot of a computer

Description automatically generated

Metadata for the bone cancer was as follows:

A screenshot of a computer

Description automatically generated

The DESeq2 analysis were as follows:

*setwd("D:/project/bone\_cancer")*

*library(DESeq2)*

*counts <- as.matrix(read.csv("counts.csv", row.names=1))*

*coldata <- read.csv("metadata.csv", row.names=1)*

*dds <- DESeqDataSetFromMatrix(countData = counts,*

*colData = coldata,*

*design = ~treatment)*

*dds <- DESeq(dds)*

*results <- results(dds)*

*head(results)*

*A screenshot of a computer program

Description automatically generated*

*plotMA(results, main="DESeq2 MA Plot")*

*A graph of a number of numbers

Description automatically generated*

*library(ggplot2)*

*results$log2FoldChange <- as.numeric(results$log2FoldChange)*

*ggplot(results, aes(x=log2FoldChange, y=-log10(padj))) +*

*geom\_point(alpha=0.4) + xlim(c(-5, 5)) + ylim(c(0, 10))*

28798 rows containing missing values or values outside the scale range were removed

*A graph with numbers and a graph

Description automatically generated*

*library(pheatmap)*

*topGenes <- head(order(results$padj), 20)*

*normalizedCounts <- counts(dds, normalized=TRUE)*

*pheatmap(normalizedCounts[topGenes,])*

*A diagram of a number

Description automatically generated with medium confidence*