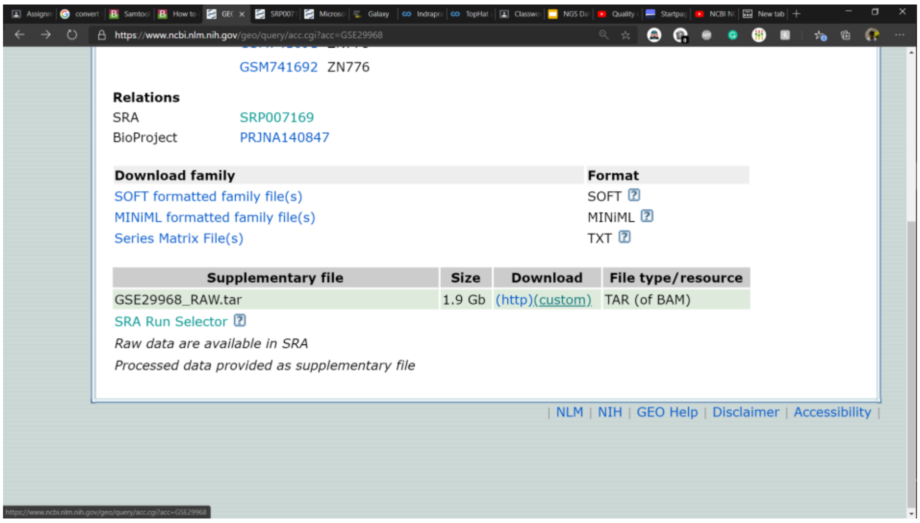
Assignment 3

1. First, I headed to NCBI website to get info about the data and get SRA ids so that I can import then to galaxy portal.
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29968
   https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA140847&o=acc_s%3Aa



GSM741692 ZN776

Relations

SRA	SRP007169
BioProject	PRJNA140847

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

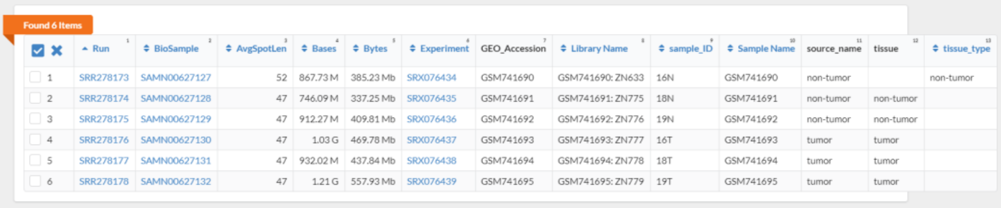
Supplementary file	Size	Download	File type/resource
GSE29968_RAW.tar	1.9 Gb	(http)(custom)	TAR (of BAM)

SRA Run Selector

Raw data are available in SRA

Processed data provided as supplementary file

NCBI Data base page

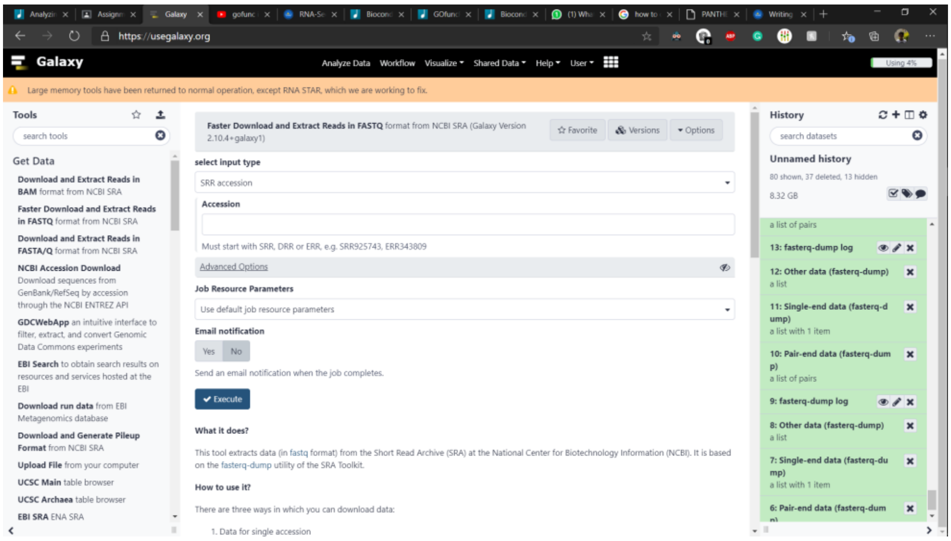


Found 6 Items

	Run	BioSample	AvgSpotLen	Bases	Bytes	Experiment	GEO_Accession	Library Name	sample_ID	Sample Name	source_name	tissue	tissue_type
1	SRR278173	SAMN00627127	52	867.73 M	385.23 Mb	SRX076434	GSM741690	GSM741690: ZN633	16N	GSM741690	non-tumor		non-tumor
2	SRR278174	SAMN00627128	47	746.09 M	337.25 Mb	SRX076435	GSM741691	GSM741691: ZN775	18N	GSM741691	non-tumor	non-tumor	
3	SRR278175	SAMN00627129	47	912.27 M	409.81 Mb	SRX076436	GSM741692	GSM741692: ZN776	19N	GSM741692	non-tumor	non-tumor	
4	SRR278176	SAMN00627130	47	1.03 G	469.78 Mb	SRX076437	GSM741693	GSM741693: ZN777	16T	GSM741693	tumor	tumor	
5	SRR278177	SAMN00627131	47	932.02 M	437.84 Mb	SRX076438	GSM741694	GSM741694: ZN778	18T	GSM741694	tumor	tumor	
6	SRR278178	SAMN00627132	47	1.21 G	557.93 Mb	SRX076439	GSM741695	GSM741695: ZN779	19T	GSM741695	tumor	tumor	

SRA page with annotation of dataset

2. After I headed to galaxy portal at https://usegalaxy.org/ to import data from option **Faster Download and Extract Reads in FASTQ**.



Large memory tools have been returned to normal operation, except RNA STAR, which we are working to fix.

Tools

Get Data

- **Download and Extract Reads in BAM** format from NCBI SRA
- **Faster Download and Extract Reads in FASTQ** format from NCBI SRA
- **Download and Extract Reads in FASTA/Q** format from NCBI SRA
- **NCBI Accession Download** Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API
- **GDCWebApp** an intuitive interface to filter, extract, and convert Genomic Data Commons experiments
- **EBI Search** to obtain search results on resources and services hosted at the EBI
- **Download run data** from EBI Metagenomics database
- **Download and Generate Pileup Format** from NCBI SRA
- **Upload File** from your computer
- **UCSC Main** table browser
- **UCSC Archaea** table browser
- **EBI SRA** ENA SRA

Faster Download and Extract Reads in FASTQ format from NCBI SRA (Galaxy Version 2.10.4+galaxy1)

select input type: SRR accession

Accession

Must start with SRR, DRR or ERR, e.g. SRR925743, ERR343809

Advanced Options

Job Resource Parameters: Use default job resource parameters

Email notification: Yes No

Send an email notification when the job completes.

✓ Execute

What it does?

This tool extracts data (in fastq format) from the Short Read Archive (SRA) at the National Center for Biotechnology Information (NCBI). It is based on the fasterq-dump utility of the SRA Toolkit.

How to use it?

There are three ways in which you can download data:

1. Data for single accession

History

Unnamed history

80 shown, 37 deleted, 13 hidden

8.32 GB

a list of pairs

13: fasterq-dump log

12: Other data (fasterq-dump)
a list

11: Single-end data (fasterq-dump)
a list with 1 item

10: Pair-end data (fasterq-dump)
a list of pairs

9: fasterq-dump log

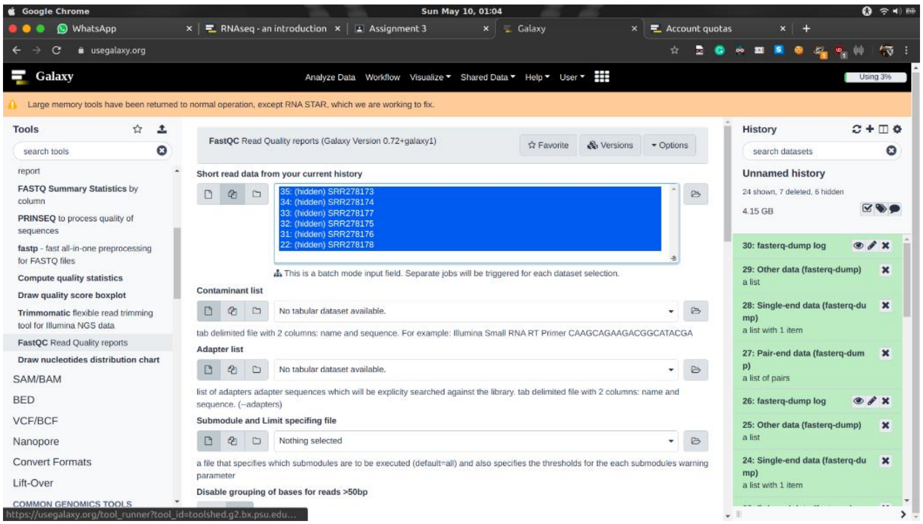
8: Other data (fasterq-dump)
a list

7: Single-end data (fasterq-dump)
a list with 1 item

6: Pair-end data (fasterq-dump)

As you can see in the history data is downloaded after typing SRR ids from SRA page in NCBI. Approx. 4 GB data was downloaded from NCBI which took around 3 hours

3. After downloading datasets in fastq files, quality report is generated using **FastQC.**


Google Chrome
Sun May 10, 01:04
WhatsApp
RNAseq - an introduction
Assignment 3
Galaxy
Account quotas
usegalaxy.org
Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 3%
Tools
search tools
report
FASTQ Summary Statistics by column
PRINSEQ to process quality of sequences
fastp - fast all-in-one preprocessing for FASTQ files
Compute quality statistics
Draw quality score boxplot
Trimmomatic flexible read trimming tool for Illumina NGS data
FastQC Read Quality reports
Draw nucleotides distribution chart
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over
COMMON GENOMICS TOOLS
Favorite
Versions
Options
History
search datasets
https://usegalaxy.org/tool_runner?tool_id=toolshed.g2.bx.psu.edu...

Large memory tools have been returned to normal operation, except RNA STAR, which we are working to fix.

FastQC Read Quality reports (Galaxy Version 0.72+galaxy1)

Short read data from your current history

35: (hidden) SRR278173
34: (hidden) SRR278174
33: (hidden) SRR278177
32: (hidden) SRR278175
31: (hidden) SRR278176
22: (hidden) SRR278178

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Contaminant list

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list

No tabular dataset available.

list of adapters adapter sequences which will be explicity searched against the library. tab delimited file with 2 columns: name and sequence. (--adapters)

Submodule and Limit specifing file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads >50bp

Unnamed history

24 shown, 7 deleted, 6 hidden

4.15 GB

30: fasterq-dump log

29: Other data (fasterq-dump)
a list

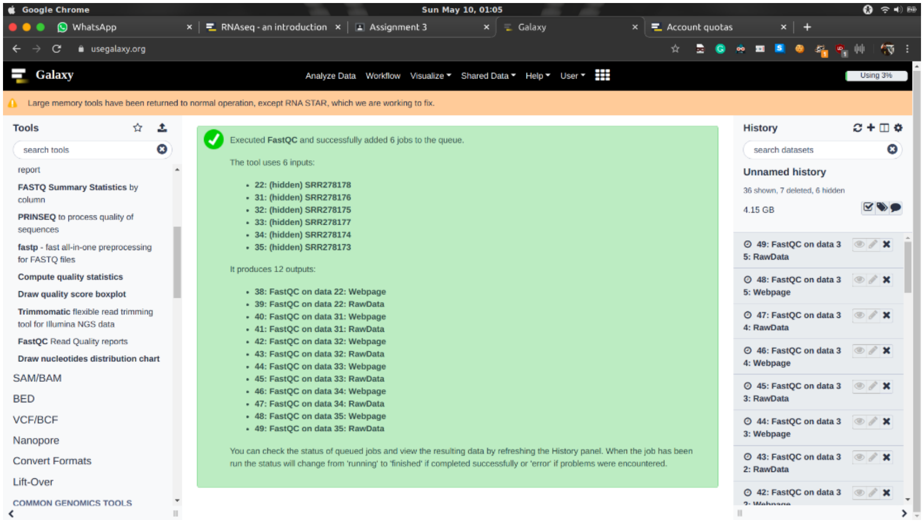
28: Single-end data (fasterq-dump)
a list with 1 item

27: Pair-end data (fasterq-dump)
a list of pairs

26: fasterq-dump log

25: Other data (fasterq-dump)
a list

24: Single-end data (fasterq-dump)
a list with 1 item


Google Chrome
Sun May 10, 01:05
WhatsApp
RNAseq - an introduction
Assignment 3
Galaxy
Account quotas
usegalaxy.org
Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 3%
Tools
search tools
report
FASTQ Summary Statistics by column
PRINSEQ to process quality of sequences
fastp - fast all-in-one preprocessing for FASTQ files
Compute quality statistics
Draw quality score boxplot
Trimmomatic flexible read trimming tool for Illumina NGS data
FastQC Read Quality reports
Draw nucleotides distribution chart
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over
COMMON GENOMICS TOOLS
History
search datasets

Large memory tools have been returned to normal operation, except RNA STAR, which we are working to fix.

Executed **FastQC** and successfully added 6 jobs to the queue.

The tool uses 6 inputs:

- **22: (hidden) SRR278178**
- **31: (hidden) SRR278176**
- **32: (hidden) SRR278175**
- **33: (hidden) SRR278177**
- **34: (hidden) SRR278174**
- **35: (hidden) SRR278173**

It produces 12 outputs:

- **38: FastQC on data 22: Webpage**
- **39: FastQC on data 22: RawData**
- **40: FastQC on data 31: Webpage**
- **41: FastQC on data 31: RawData**
- **42: FastQC on data 32: Webpage**
- **43: FastQC on data 32: RawData**
- **44: FastQC on data 33: Webpage**
- **45: FastQC on data 33: RawData**
- **46: FastQC on data 34: Webpage**
- **47: FastQC on data 34: RawData**
- **48: FastQC on data 35: Webpage**
- **49: FastQC on data 35: RawData**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Unnamed history

36 shown, 7 deleted, 6 hidden

4.15 GB

49: FastQC on data 35: RawData

48: FastQC on data 35: Webpage

47: FastQC on data 34: RawData

46: FastQC on data 34: Webpage

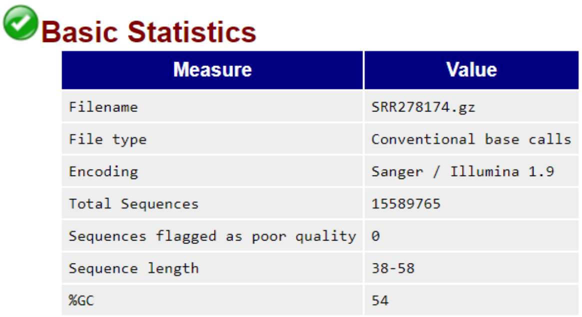
45: FastQC on data 33: RawData

44: FastQC on data 33: Webpage

43: FastQC on data 32: RawData

42: FastQC on data 3

Inference

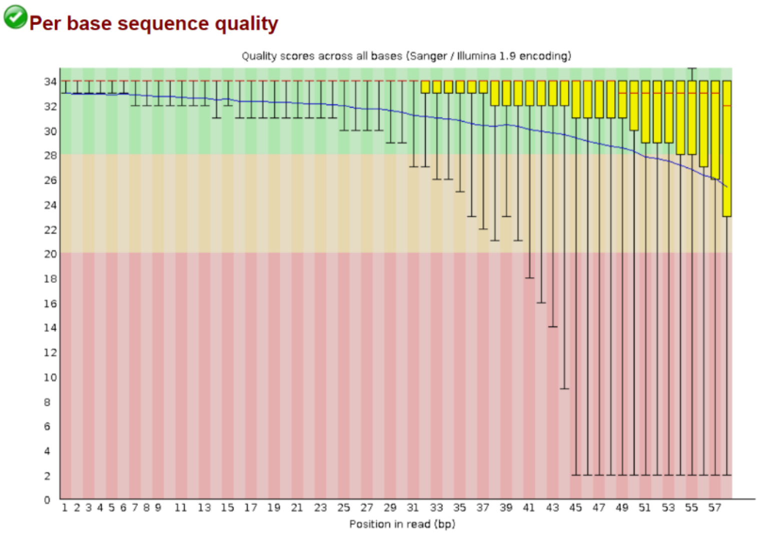


✓ Basic Statistics

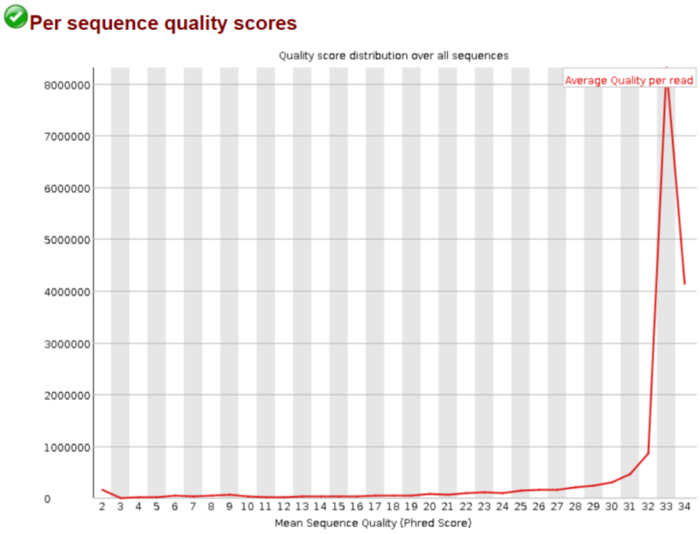
Measure	Value
Filename	SRR278174.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	15589765
Sequences flagged as poor quality	0
Sequence length	38-58
%GC	54

6 reports were made as our dataset contain 3 dataset of normal tissue and 3 for infected tissue.

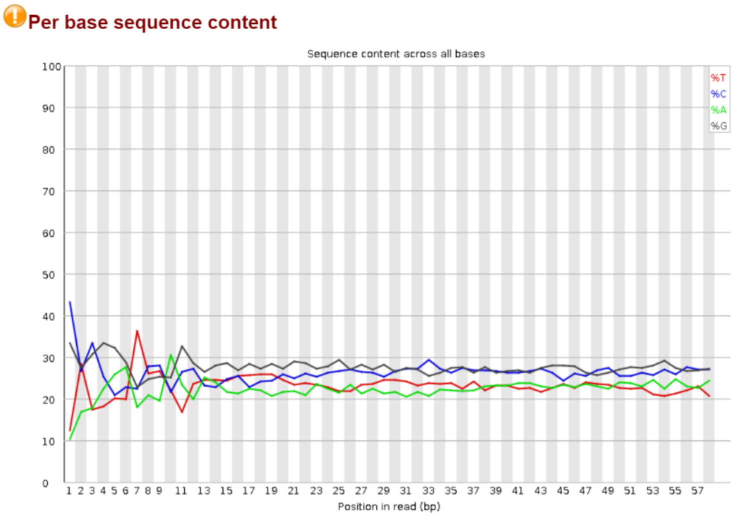
Basic Statistics part of fastqc report contains basic info like what is the size of reads, GC percentage in all reads, how reads were recorded etc. Here in our example there were about 15million read of size from 38 to 58 which has 54% of GC content as it is recorded with illumina it also contains quality score and 0 is of poor quality.



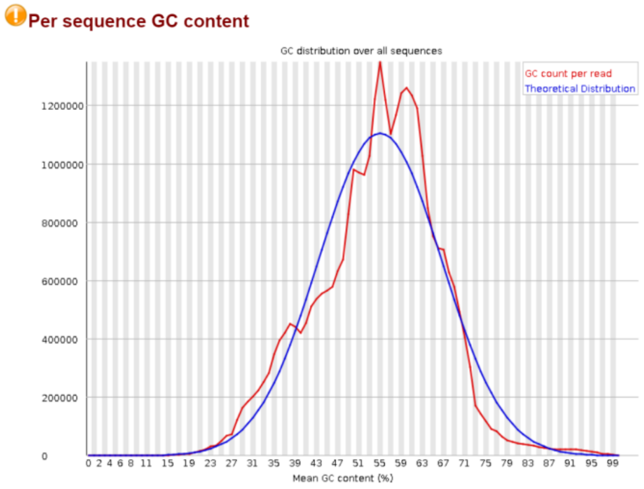
Per base sequence quality part of fastqc report contains info of box plot of quality score at every base position of read and our read length is max 58 so x axis of 58 length, there is a trend in all report that per base quality score decreases as position of read increases



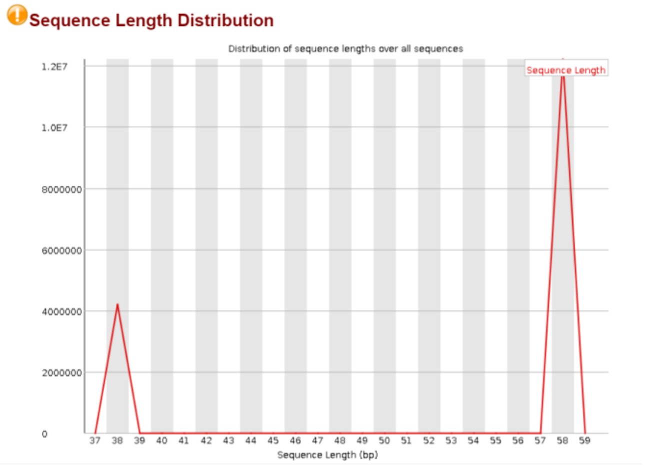
Per sequence quality score part of fastqc report contains frequency of mean sequence phred score i.e. average of phred score is taken and then these scores are grouped by intervals to get frequency of each grp. In our example we can see that about 8M sequence or mostly has phred score around 33.



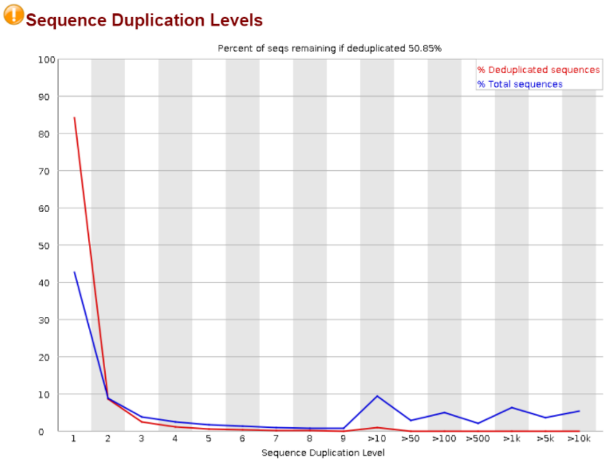
Per base sequence content part of fastqc report contains info of base at position of read. It gives us percentage of base at every position of sequence or reads. We can see that in the initial content are mixed up which also shows that there is bias that why there is warning sign before heading in graph.



Per sequecnce GC content part of fastqc report shows that what should be distribution of GC across all sequence and what is in the data. Blue line represent theoretical data line or we can ideal situation and red line represent our data from both line it shown that most of the sequence should contain about 50% percent of GC content. In our example about 1.2 M sequence have 50% GC content.

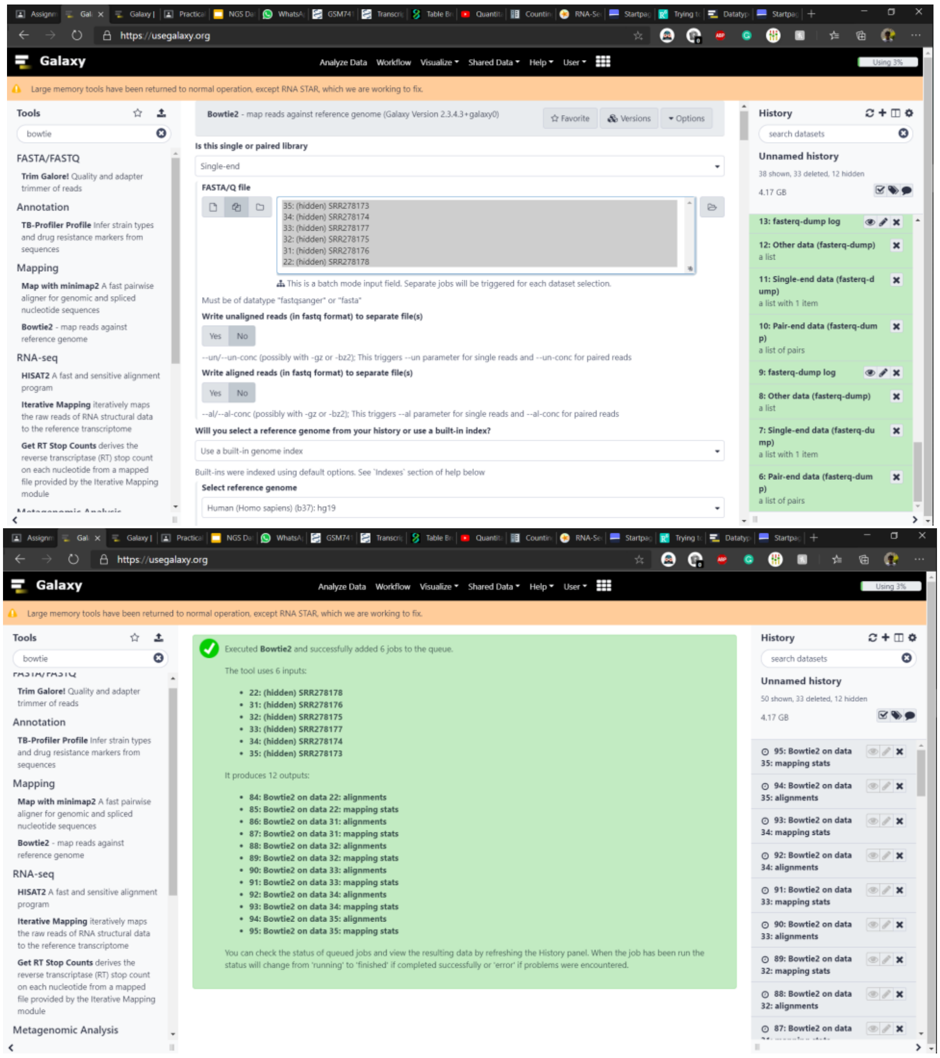


Sequence length distribution part of fastqc contain info of length of reads or sequence. It shows frequency of length of read in data. Here data either has length of 38 or 58. This may be because of primer selection.



Sequence duplication levels contains Info of duplication in the data, red line represent number of distinct sequence that are duplicated similarly blue line represent count of all sequence.

4. After downloading datasets in fastq files, reads were aligned using **Bowtie2** using inbuild reference genome hg19 version as it is suitable for the study.



Large memory tools have been returned to normal operation, except RNA STAR, which we are working to fix.

Bowtie2 - map reads against reference genome (Galaxy Version 2.3.4.3+galaxy0)

Is this single or paired library

Single-end

FASTA/Q file

35: (hidden) SRR278173
34: (hidden) SRR278174
33: (hidden) SRR278177
32: (hidden) SRR278175
31: (hidden) SRR278176
22: (hidden) SRR278178

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Must be of datatype "fastqsanger" or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

Yes No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

Yes No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome

Human (Homo sapiens) (b37): hg19

Executed **Bowtie2** and successfully added 6 jobs to the queue.

The tool uses 6 inputs:

- **22: (hidden) SRR278178**
- **31: (hidden) SRR278176**
- **32: (hidden) SRR278175**
- **33: (hidden) SRR278177**
- **34: (hidden) SRR278174**
- **35: (hidden) SRR278173**

It produces 12 outputs:

- **84: Bowtie2 on data 22: alignments**
- **85: Bowtie2 on data 22: mapping stats**
- **86: Bowtie2 on data 31: alignments**
- **87: Bowtie2 on data 31: mapping stats**
- **88: Bowtie2 on data 32: alignments**
- **89: Bowtie2 on data 32: mapping stats**
- **90: Bowtie2 on data 33: alignments**
- **91: Bowtie2 on data 33: mapping stats**
- **92: Bowtie2 on data 34: alignments**
- **93: Bowtie2 on data 34: mapping stats**
- **94: Bowtie2 on data 35: alignments**
- **95: Bowtie2 on data 35: mapping stats**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

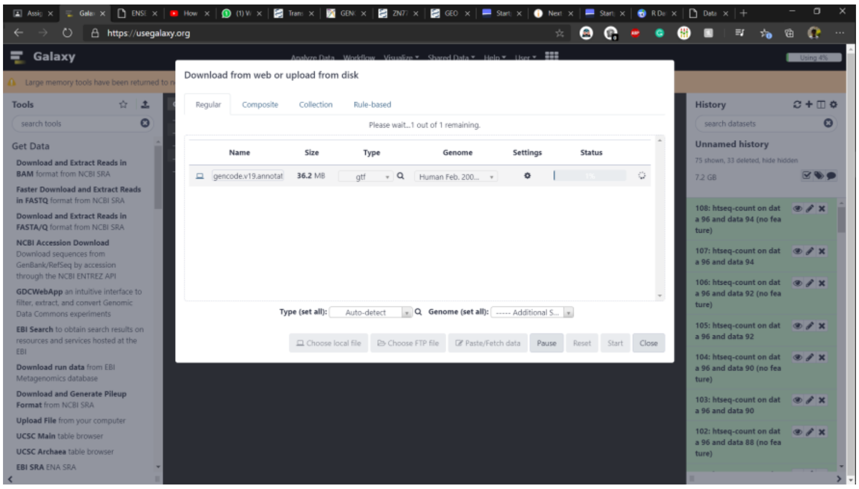
Bowtie2 took around 6 hours to complete

Default parameters were selected in the bowtie2 except reference genome and history visibility. Bowtie2 is fast and take low memory as compared to other alignment tools like hisat2 etc. In these reads were align to reference genome to get count matrix in the subsequent steps.

```
25541142 reads; of these:
  25541142 (100.00%) were unpaired; of these:
    2254424 (8.83%) aligned 0 times
    5827495 (22.82%) aligned exactly 1 time
    17459223 (68.36%) aligned >1 times
91.17% overall alignment rate
```

This is stat of one of the data after alignment it shows that around 68% of read aligned more than 1 time and 22% of sequence exactly one time and 8% of sequence are not aligned, these sequences can be adapter sequence which is used in illumina.

5. Download annotation file from ensembl.
   Data from https://www.gencodegenes.org/human/release_19.html
   First data was download into computer then on galaxy portal



Download from web or upload from disk

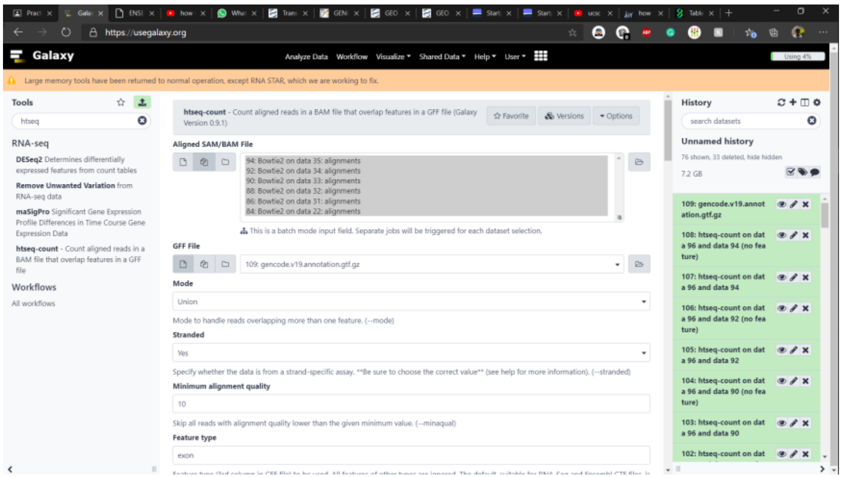
Regular | Composite | Collection | Rule-based

Please wait...1 out of 1 remaining.

Name	Size	Type	Genome	Settings	Status
gencode.v19.annotat	36.2 MB	gtf	Human Feb. 200...		

Type (set all): Auto-detect Genome (set all): ----- Additional S...

6. Using htseq-count tool count matrix was created and annotation or gtf file was upload in the previous step.
   Union mode was selected and min 10 quality was selected. BAM file from bowtie2 result was selected as input.



htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file (Galaxy Version 0.9.1)

Aligned SAM/BAM File
94: Bowtie2 on data 35: alignments
92: Bowtie2 on data 34: alignments
90: Bowtie2 on data 33: alignments
88: Bowtie2 on data 32: alignments
86: Bowtie2 on data 31: alignments
84: Bowtie2 on data 22: alignments

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

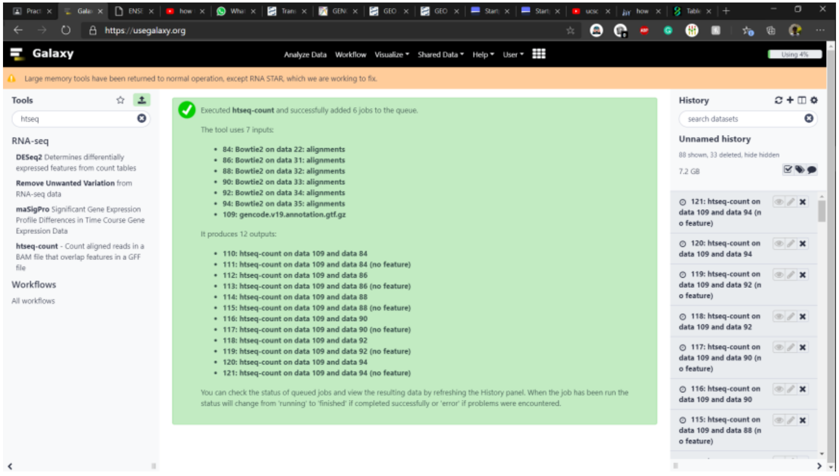
GFF File
109: gencode.v19.annotation.gtf.gz

Mode
Union
Mode to handle reads overlapping more than one feature. (--mode)

Stranded
Yes
Specify whether the data is from a strand-specific assay. **Be sure to choose the correct value** (see help for more information). (--stranded)

Minimum alignment quality
10
Skip all reads with alignment quality lower than the given minimum value. (--minaqual)

Feature type
exon



Executed **htseq-count** and successfully added 6 jobs to the queue.

The tool uses 7 inputs:

- **84: Bowtie2 on data 22: alignments**
- **86: Bowtie2 on data 31: alignments**
- **88: Bowtie2 on data 32: alignments**
- **90: Bowtie2 on data 33: alignments**
- **92: Bowtie2 on data 34: alignments**
- **94: Bowtie2 on data 35: alignments**
- **109: gencode.v19.annotation.gtf.gz**

It produces 12 outputs:

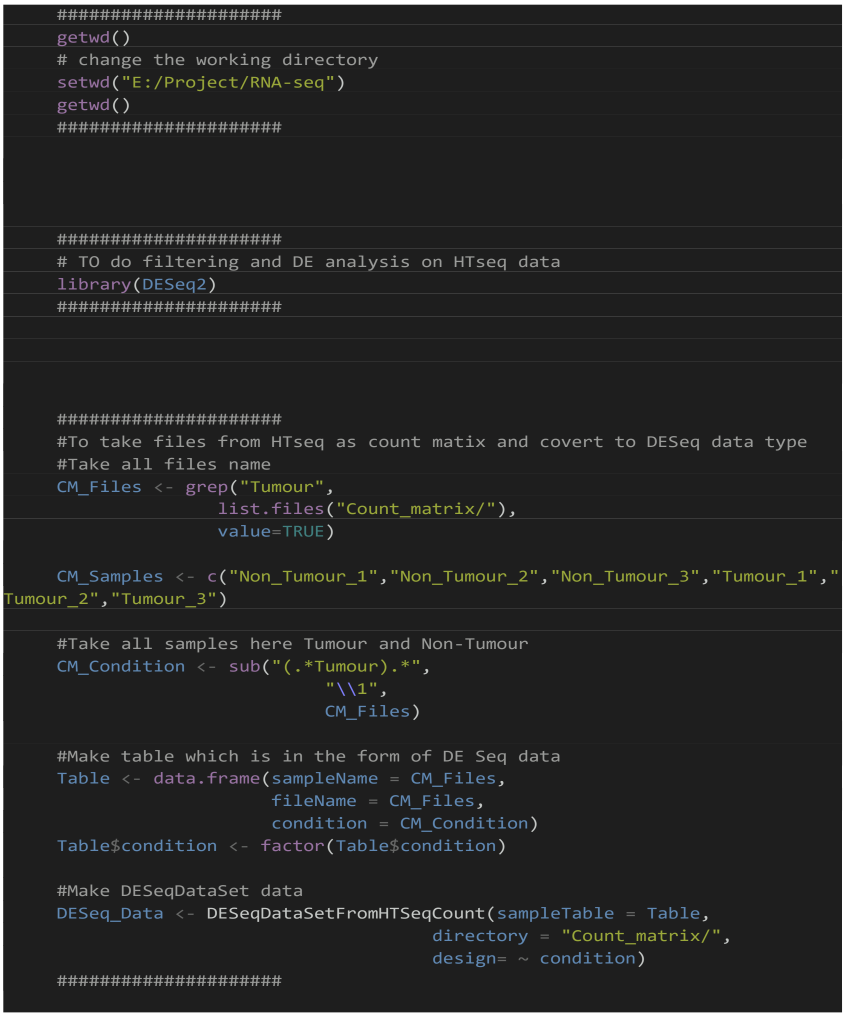
- **110: htseq-count on data 109 and data 84**
- **111: htseq-count on data 109 and data 84 (no feature)**
- **112: htseq-count on data 109 and data 86**
- **113: htseq-count on data 109 and data 86 (no feature)**
- **114: htseq-count on data 109 and data 88**
- **115: htseq-count on data 109 and data 88 (no feature)**
- **116: htseq-count on data 109 and data 90**
- **117: htseq-count on data 109 and data 90 (no feature)**
- **118: htseq-count on data 109 and data 92**
- **119: htseq-count on data 109 and data 92 (no feature)**
- **120: htseq-count on data 109 and data 94**
- **121: htseq-count on data 109 and data 94 (no feature)**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

7. **Metadata checklist**
   - **a)** Reads originate from **Homo sapiens.**
   - **b)** Yes, reference genome is available from Homo sapiens that is download from ensembl which was hg19 version.
   - **c)** GPL10999 Illumina Genome Analyzer IIx (Homo sapiens) platform was used.
   - **d)** These are short reads
   - **e)** **ID list**

SRA ID	GEO ID	Sample ID	Type
SRR278173	GSM741690	16N	Non-Tumour
SRR278174	GSM741691	17N	Non-Tumour
SRR278175	GSM741692	19N	Non-Tumour
SRR278176	GSM741693	16T	Tumour
SRR278177	GSM741694	17T	Tumour
SRR278178	GSM741695	19T	Tumour

8. From this step and onward all steps were taken in R code. For Differential Gene Expression analysis DESeq2 was used in R. Before coding in R I downloaded count file from galaxy in tabular form in my PC.



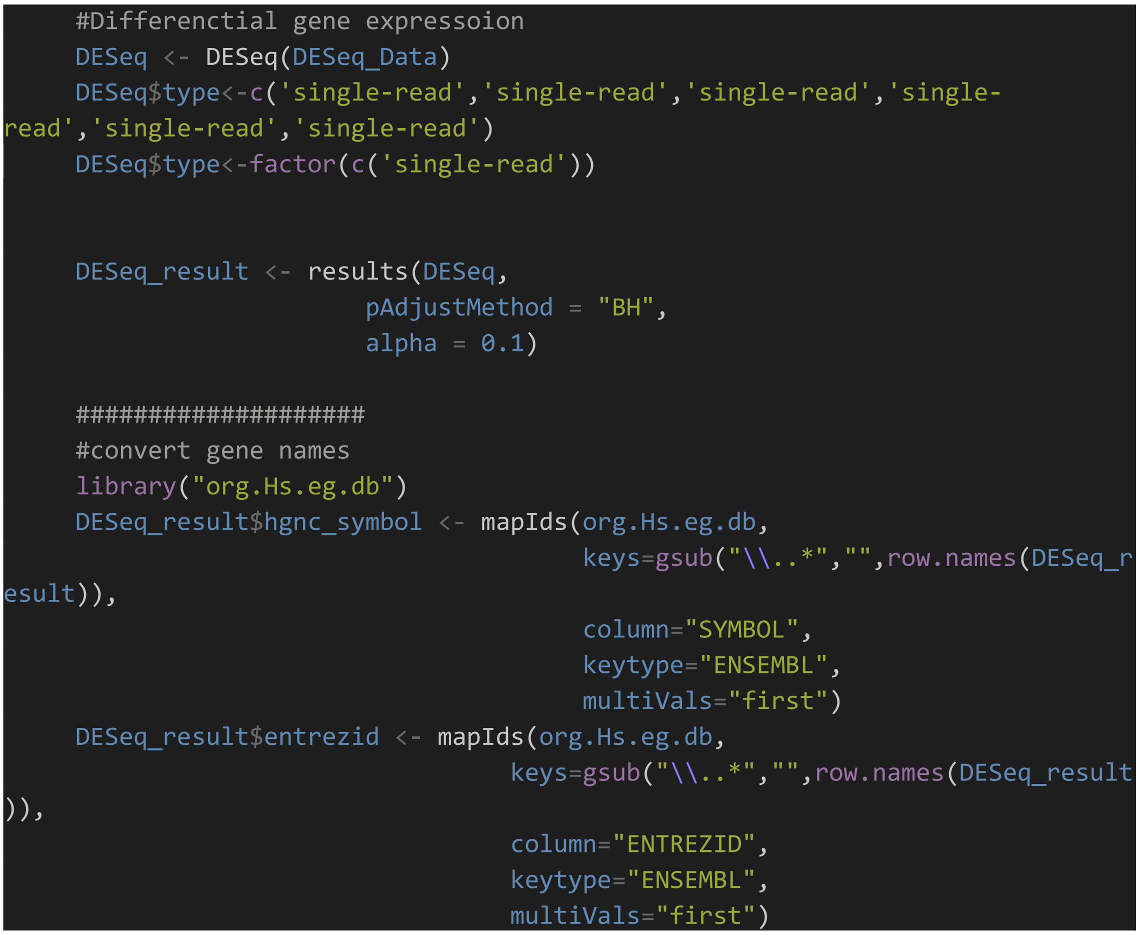
```
    #####################
    getwd()
    # change the working directory
    setwd("E:/Project/RNA-seq")
    getwd()
    #####################




    #####################
    # TO do filtering and DE analysis on HTseq data
    library(DESeq2)
    #####################




    #####################
    #To take files from HTseq as count matix and covert to DESeq data type
    #Take all files name
    CM_Files <- grep("Tumour",
                     list.files("Count_matrix/"),
                     value=TRUE)

    CM_Samples <- c("Non_Tumour_1","Non_Tumour_2","Non_Tumour_3","Tumour_1","
Tumour_2","Tumour_3")

    #Take all samples here Tumour and Non-Tumour
    CM_Condition <- sub("(.*Tumour).*",
                        "\\1",
                        CM_Files)

    #Make table which is in the form of DE Seq data
    Table <- data.frame(sampleName = CM_Files,
                        fileName = CM_Files,
                        condition = CM_Condition)
    Table$condition <- factor(Table$condition)

    #Make DESeqDataSet data
    DESeq_Data <- DESeqDataSetFromHTSeqCount(sampleTable = Table,
                                             directory = "Count_matrix/",
                                             design= ~ condition)
    #####################
```

Directory E:/Project/RNA-seq working site
Directory E:/Project/RNA-seq/Count_matrix count tables were stored

Files were download and renamed as Tumour_1 that represent tumour count matrix in first patient similarly Non-Tumout_1 represent first patient non tumour rna-seq count matrix from galaxy which is generated after htseq-count.

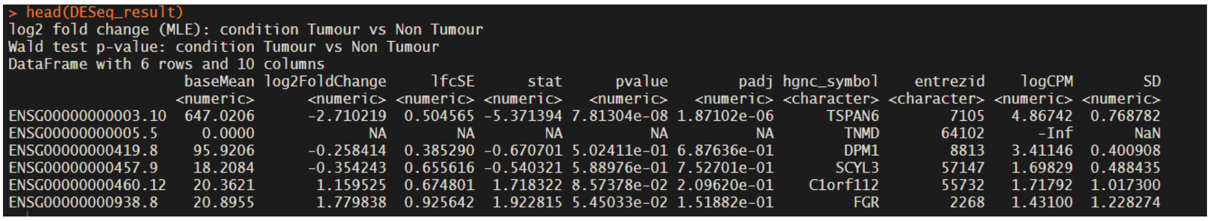
Then in the previous code condition and files were import with htseq function in deseq2

After getting data model from DESeq2 was fitted to get DESeq type file, which also make count file in single file which make it easy to calculate dispersion, size factors etc.
Base mean, LogFC(fold change), Pvalue, FDR value was calculated b passing DESeq type data in results function.
As data contains ID from ensembl we have to convert it into normal gene symbol by using org.Hs.eg.db library.
LogCPM or log of count per million is not calculated using results so it calculated using edgeR library.



```
    #Differenctial gene expressoion
    DESeq <- DESeq(DESeq_Data)
    DESeq$type<-c('single-read','single-read','single-read','single-
read','single-read','single-read')
    DESeq$type<-factor(c('single-read'))


    DESeq_result <- results(DESeq,
                            pAdjustMethod = "BH",
                            alpha = 0.1)

    ####################
    #convert gene names
    library("org.Hs.eg.db")
    DESeq_result$hgnc_symbol <- mapIds(org.Hs.eg.db,
                                       keys=gsub("\\..*","",row.names(DESeq_r
esult)),
                                       column="SYMBOL",
                                       keytype="ENSEMBL",
                                       multiVals="first")
    DESeq_result$entrezid <- mapIds(org.Hs.eg.db,
                                    keys=gsub("\\..*","",row.names(DESeq_result
)),
                                    column="ENTREZID",
                                    keytype="ENSEMBL",
                                    multiVals="first")
```

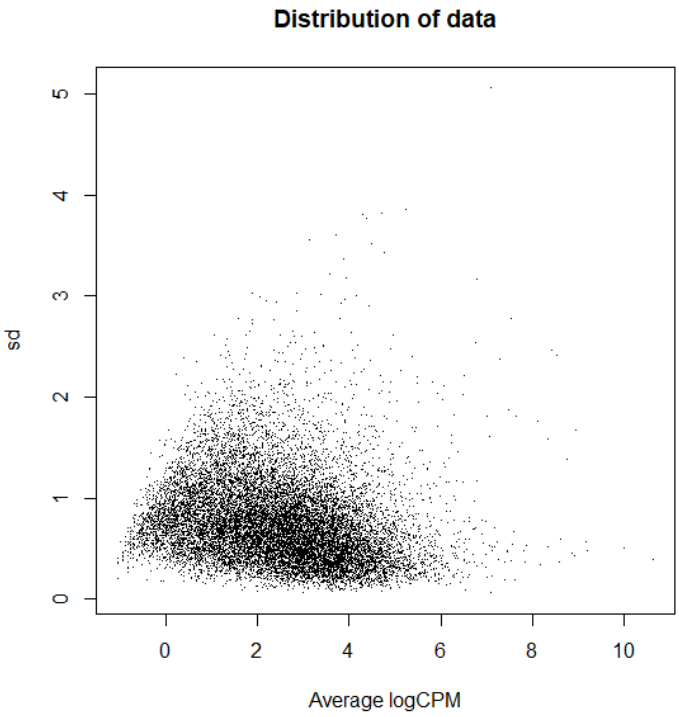
Head of DESeq_result matrix



```
> head(DESeq_result)
log2 fold change (MLE): condition Tumour vs Non Tumour
Wald test p-value: condition Tumour vs Non Tumour
DataFrame with 6 rows and 10 columns
                    baseMean log2FoldChange     lfcSE      stat      pvalue        padj hgnc_symbol    entrezid    logCPM        SD
                   <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric> <character> <character> <numeric> <numeric>
ENSG00000000003.10  647.0206      -2.710219  0.504565 -5.371394 7.81304e-08 1.87102e-06      TSPAN6        7105   4.86742  0.768782
ENSG00000000005.5     0.0000             NA        NA        NA          NA          NA        TNMD       64102      -Inf       NaN
ENSG00000000419.8    95.9206      -0.258414  0.385290 -0.670701 5.02411e-01 6.87636e-01        DPM1        8813   3.41146  0.400908
ENSG00000000457.9    18.2084      -0.354243  0.655616 -0.540321 5.88976e-01 7.52701e-01       SCYL3       57147   1.69829  0.488435
ENSG00000000460.12   20.3621       1.159525  0.674801  1.718322 8.57378e-02 2.09620e-01    C1orf112       55732   1.71792  1.017300
ENSG00000000938.8    20.8955       1.779838  0.925642  1.922815 5.45033e-02 1.51882e-01         FGR        2268   1.43100  1.228274
```

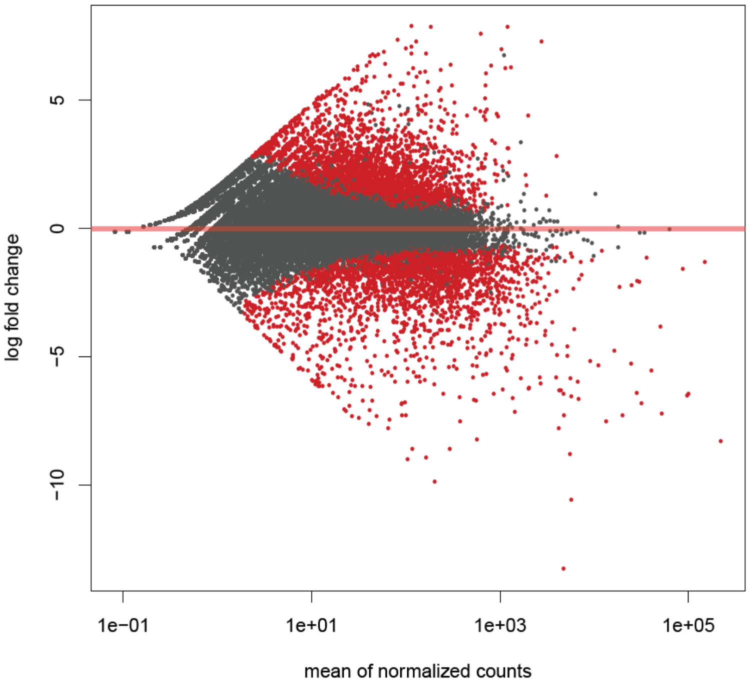
Following graph are created in R and saved in PNG file which are then attached here

Standard deviation vs average logCPM distribution of data



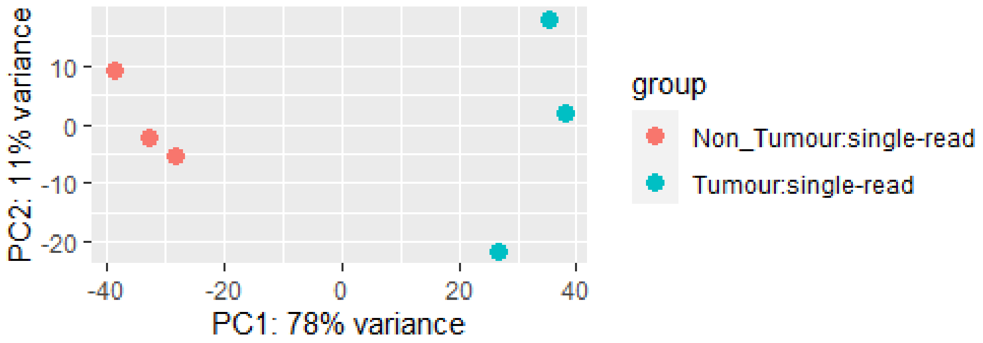
As we can see that most of the gene are highly expressed and logCPM as mostly gene lie around 2 and 4.

MA-plot



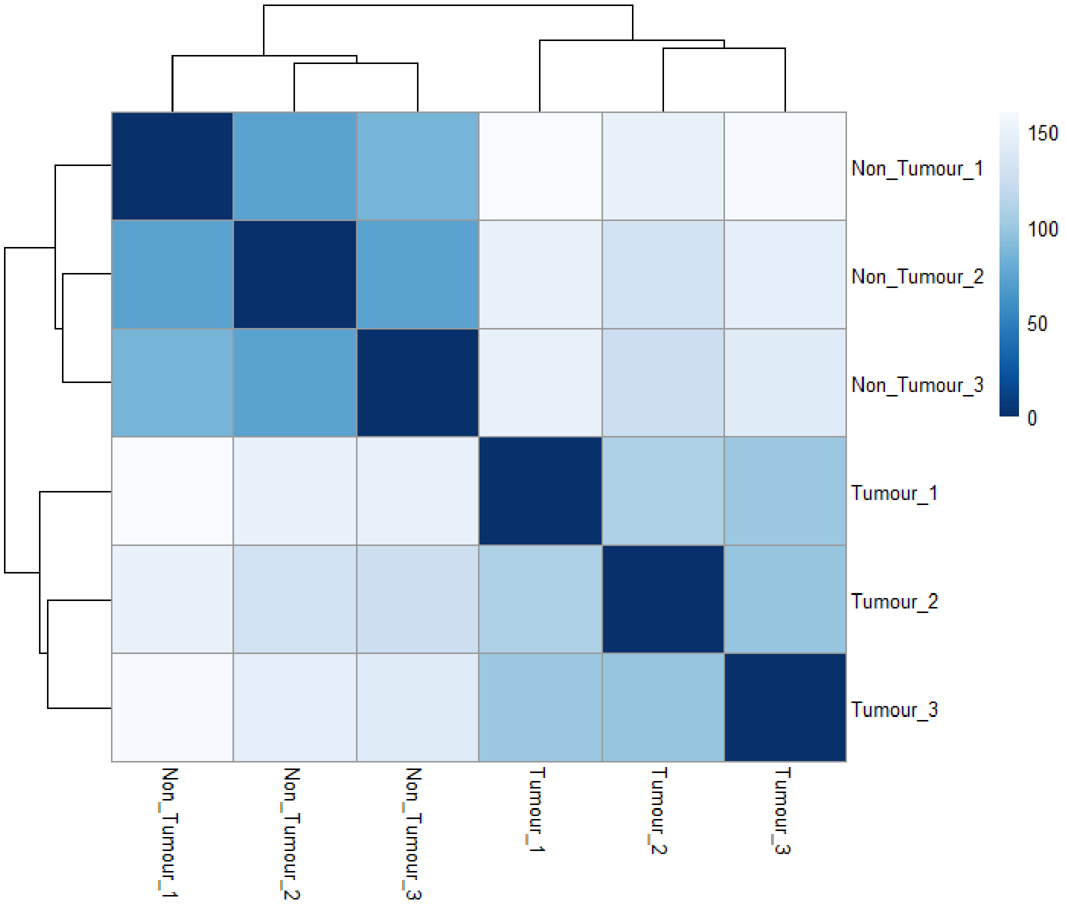
This graph shows us dispersion of fold change ie if logFC is 2 then gene expression is 4 times when condition is changed i.e. from non-tumour to tumour and x axis contain normalised count of count of genes.

PCA plot



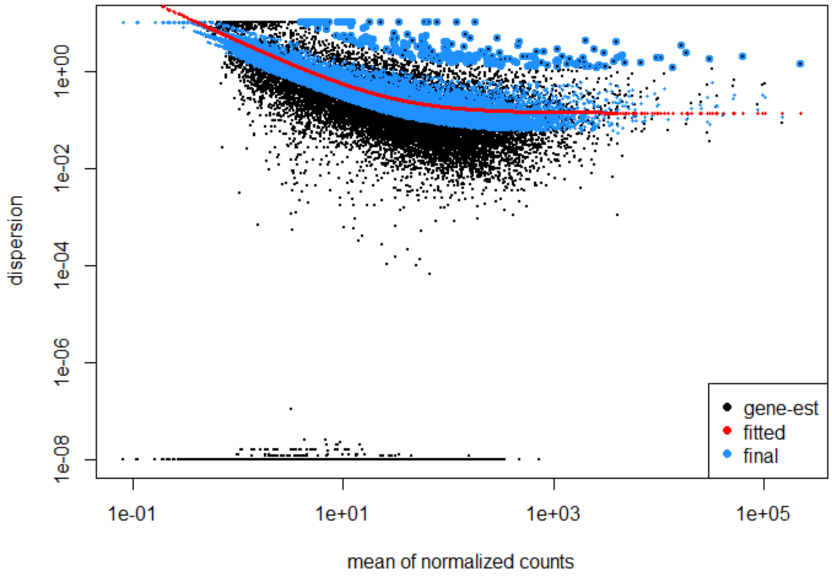
PCA plot is calculated using regularized log transformation is DESeq2 function PC1 is good in clustering group in condition while PC2 is not. As left cluster is of non-tumour read and right is of tumour read, data is separated greatly in x axis but no in y axis.

Sample to sample distance



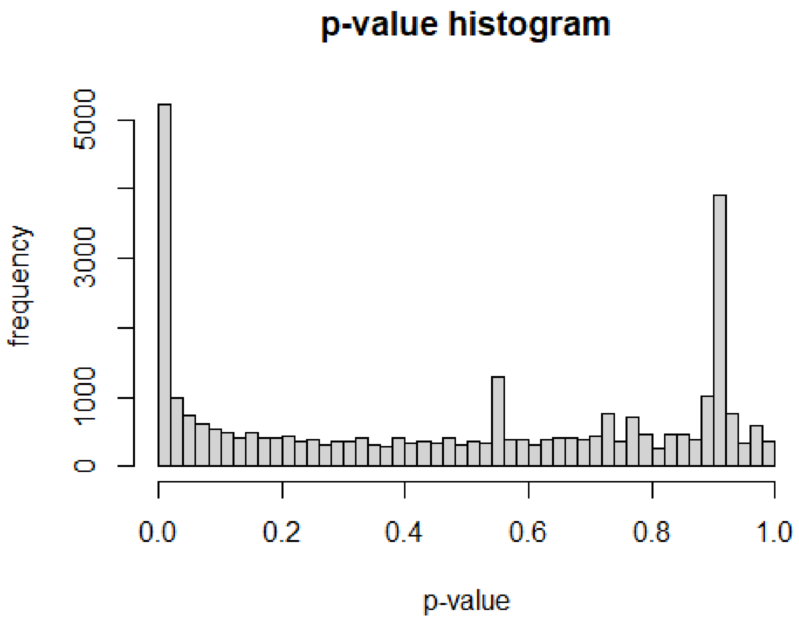
This heatmap shows the relative distance between 2 sample, it show how much 2 sample are similar if square is dark blue then distance is 0 it means they are 100% similar, as similarity decreases square colours become light blue.

Dispersion Estimation

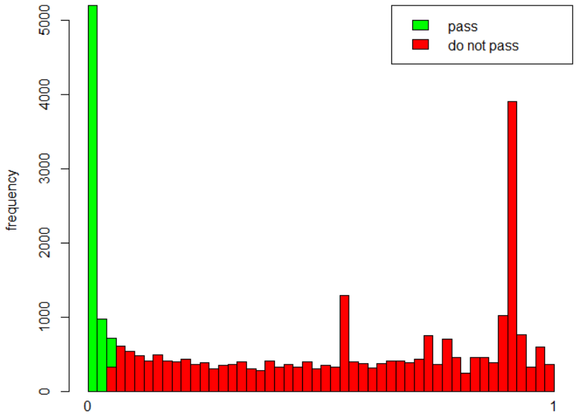


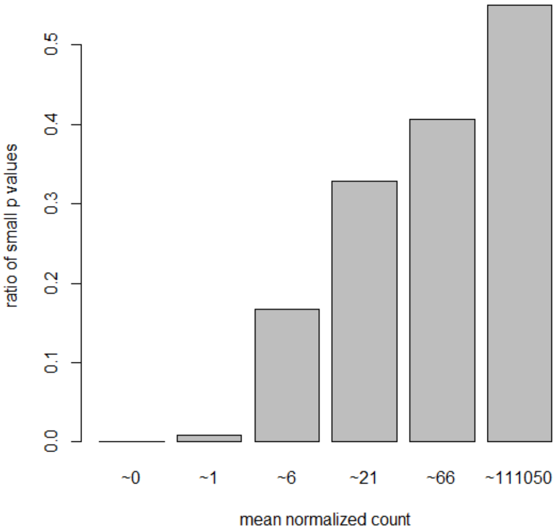
This dispersion graph shows how gene should be align, red line is the expected dispersion value for genes of a given expression strength and black dots is a gene with maximum likelihood estimation (MLE) of dispersion and blue dots are in the expected dispersion value to filter data we can make data so that every dot is red or blue dots.

Histogram of pvalue vs frequency



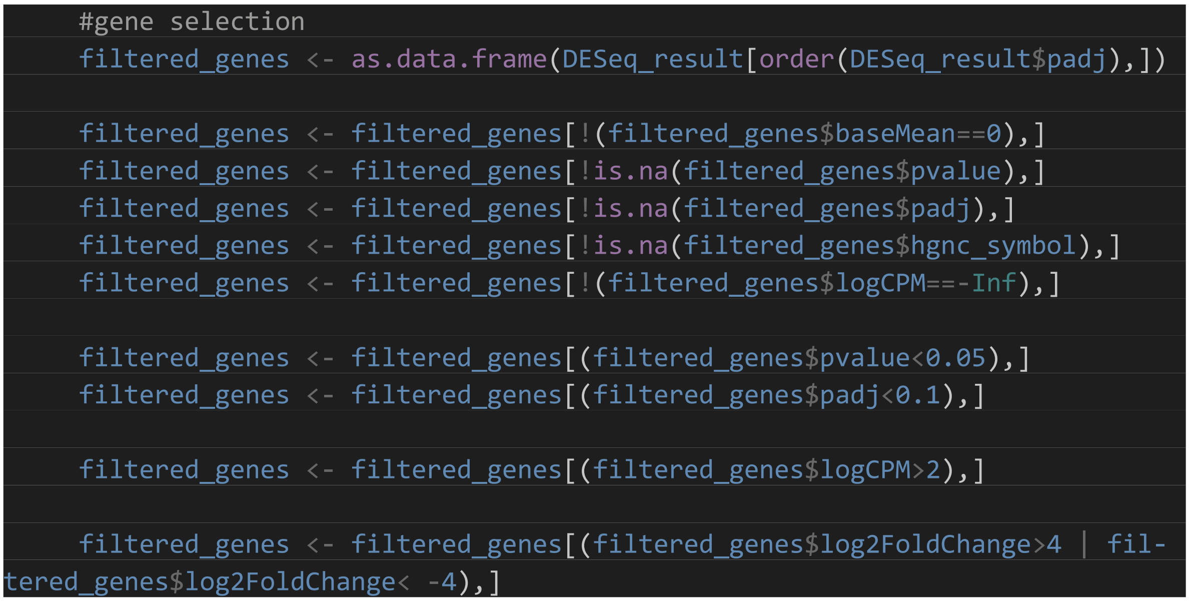
X axis is divided into 50 division and each bar represent frequency of that particular p value and following graph represent data set which are select (p-value < 0.05)





This graph represent mean of normalised of selected dataset i.e. which have p value less than 0.05 against ratio of small p value.

Following code snippet was used to filter out genes



```
#gene selection
filtered_genes <- as.data.frame(DESeq_result[order(DESeq_result$padj),])

filtered_genes <- filtered_genes[!(filtered_genes$baseMean==0),]
filtered_genes <- filtered_genes[!is.na(filtered_genes$pvalue),]
filtered_genes <- filtered_genes[!is.na(filtered_genes$padj),]
filtered_genes <- filtered_genes[!is.na(filtered_genes$hgnc_symbol),]
filtered_genes <- filtered_genes[!(filtered_genes$logCPM==-Inf),]

filtered_genes <- filtered_genes[(filtered_genes$pvalue<0.05),]
filtered_genes <- filtered_genes[(filtered_genes$padj<0.1),]

filtered_genes <- filtered_genes[(filtered_genes$logCPM>2),]

filtered_genes <- filtered_genes[(filtered_genes$log2FoldChange>4 | filtered_genes$log2FoldChange< -4),]
```

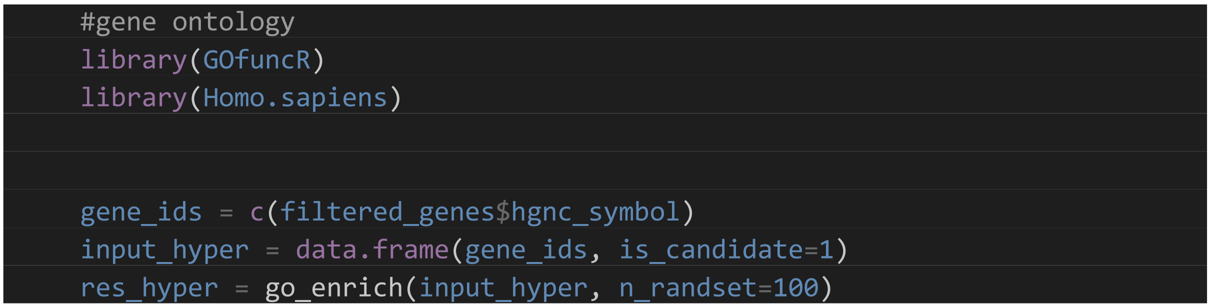
Conditions were

1. P-value < 0.05
2. FDR value using Benjamini Hochberg method < 0.1
3. Log(CPM) >2
4. Log(FC) >4 or <-4

Dataset was sorted by FDR value and row with NA value was removed.

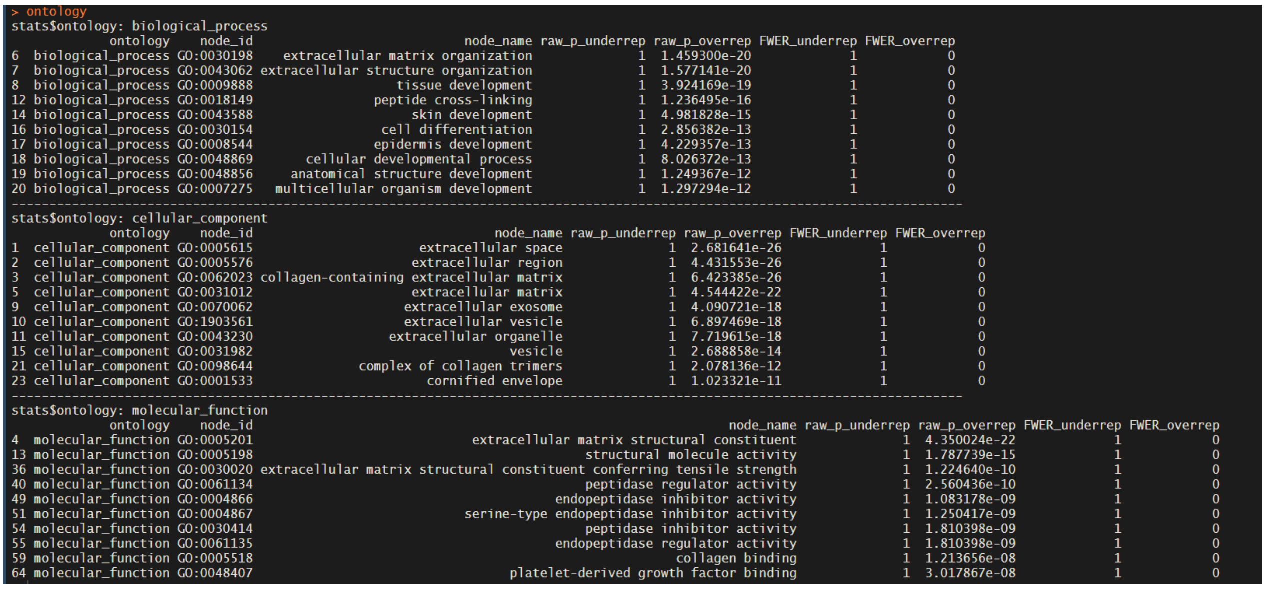
This gives us 184 genes

9. Gene ontology was performed using GoFuncR function.



```
#gene ontology
library(GOfuncR)
library(Homo.sapiens)


gene_ids = c(filtered_genes$hgnc_symbol)
input_hyper = data.frame(gene_ids, is_candidate=1)
res_hyper = go_enrich(input_hyper, n_randset=100)
```

Above code is executed to generate data in GO id terms



```
> ontology
stats$ontology: biological_process
             ontology    node_id                            node_name raw_p_underrep raw_p_overrep FWER_underrep FWER_overrep
6  biological_process GO:0030198    extracellular matrix organization              1  1.459300e-20             1            0
7  biological_process GO:0043062 extracellular structure organization              1  1.577141e-20             1            0
8  biological_process GO:0009888                   tissue development              1  3.924169e-19             1            0
12 biological_process GO:0018149                peptide cross-linking              1  1.236495e-16             1            0
14 biological_process GO:0043588                     skin development              1  4.981828e-15             1            0
16 biological_process GO:0030154                 cell differentiation              1  2.856382e-13             1            0
17 biological_process GO:0008544                epidermis development              1  4.229357e-13             1            0
18 biological_process GO:0048869       cellular developmental process              1  8.026372e-13             1            0
19 biological_process GO:0048856     anatomical structure development              1  1.249367e-12             1            0
20 biological_process GO:0007275   multicellular organism development              1  1.297294e-12             1            0
--------------------------------------------------------------------------------------------------------------------------------
stats$ontology: cellular_component
             ontology    node_id                                node_name raw_p_underrep raw_p_overrep FWER_underrep FWER_overrep
1  cellular_component GO:0005615                      extracellular space              1  2.681641e-26             1            0
2  cellular_component GO:0005576                     extracellular region              1  4.431553e-26             1            0
3  cellular_component GO:0062023 collagen-containing extracellular matrix              1  6.423385e-26             1            0
5  cellular_component GO:0031012                     extracellular matrix              1  4.544422e-22             1            0
9  cellular_component GO:0070062                    extracellular exosome              1  4.090721e-18             1            0
10 cellular_component GO:1903561                    extracellular vesicle              1  6.897469e-18             1            0
11 cellular_component GO:0043230                  extracellular organelle              1  7.719615e-18             1            0
15 cellular_component GO:0031982                                  vesicle              1  2.688858e-14             1            0
21 cellular_component GO:0098644              complex of collagen trimers              1  2.078136e-12             1            0
23 cellular_component GO:0001533                       cornified envelope              1  1.023321e-11             1            0
--------------------------------------------------------------------------------------------------------------------------------
stats$ontology: molecular_function
             ontology    node_id                                                               node_name raw_p_underrep raw_p_overrep FWER_underrep FWER_overrep
4  molecular_function GO:0005201                             extracellular matrix structural constituent              1  4.350024e-22             1            0
13 molecular_function GO:0005198                                            structural molecule activity              1  1.787739e-15             1            0
36 molecular_function GO:0030020 extracellular matrix structural constituent conferring tensile strength              1  1.224640e-10             1            0
40 molecular_function GO:0061134                                            peptidase regulator activity              1  2.560436e-10             1            0
49 molecular_function GO:0004866                                        endopeptidase inhibitor activity              1  1.083178e-09             1            0
51 molecular_function GO:0004867                            serine-type endopeptidase inhibitor activity              1  1.250417e-09             1            0
54 molecular_function GO:0030414                                            peptidase inhibitor activity              1  1.810398e-09             1            0
55 molecular_function GO:0061135                                        endopeptidase regulator activity              1  1.810398e-09             1            0
59 molecular_function GO:0005518                                                        collagen binding              1  1.213656e-08             1            0
64 molecular_function GO:0048407                                  platelet-derived growth factor binding              1  3.017867e-08             1            0
```

Above is the top 10 ontology result which is grouped by gene ontology types

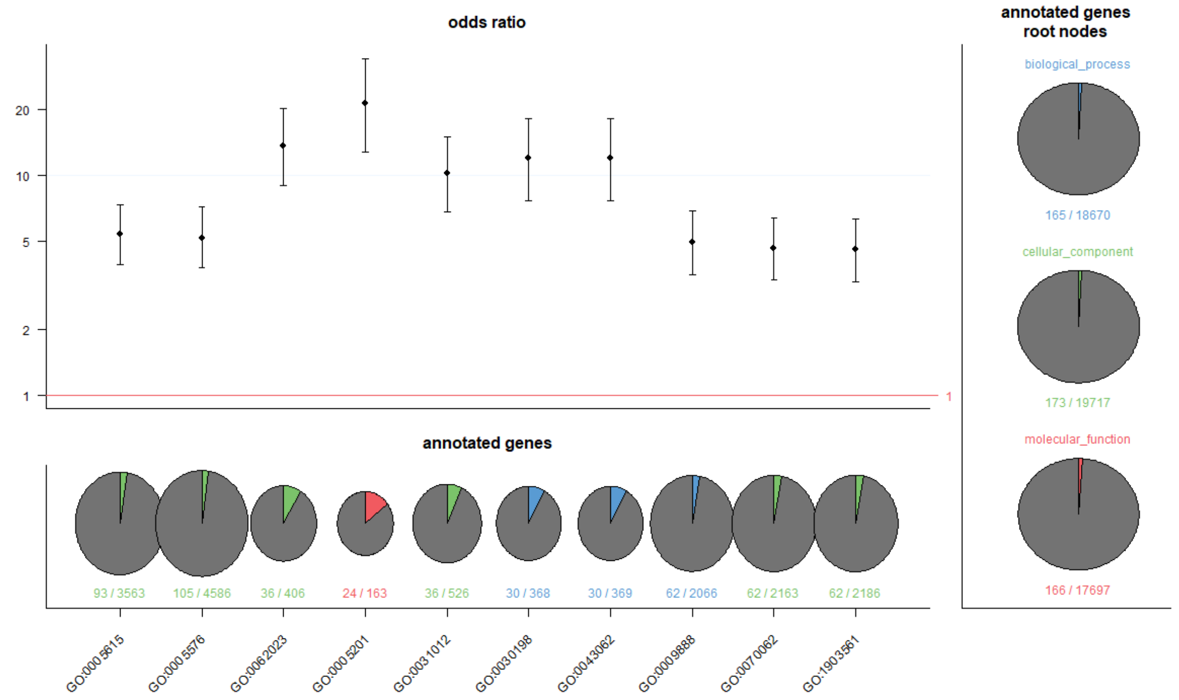
Most of the top gene ontology result shows that genes were in the path of extracellular space and region of skin of vesicle and that is true as we have taken sample from esophageal squamous cell.

Biological processes in this ontology mainly consist of cellular matrix and structure, tissue and skin development processes lead to the reason that these genes can be cancer promoter.

Cellular component of ontology also mainly consists of cellular region and space, as sample is taken from infected cell and normal cell so ontology will be focused of cell area and parts.

Molecular function in these ontologies mainly consist of endopeptidase or peptidase activity which show high activity of breakage of peptide bonds in the cellular area or space.

If we confine all the info then top ontology show that there is high activity of breakage of peptide bond cell area which leads to abnormal development of tissue and cell.

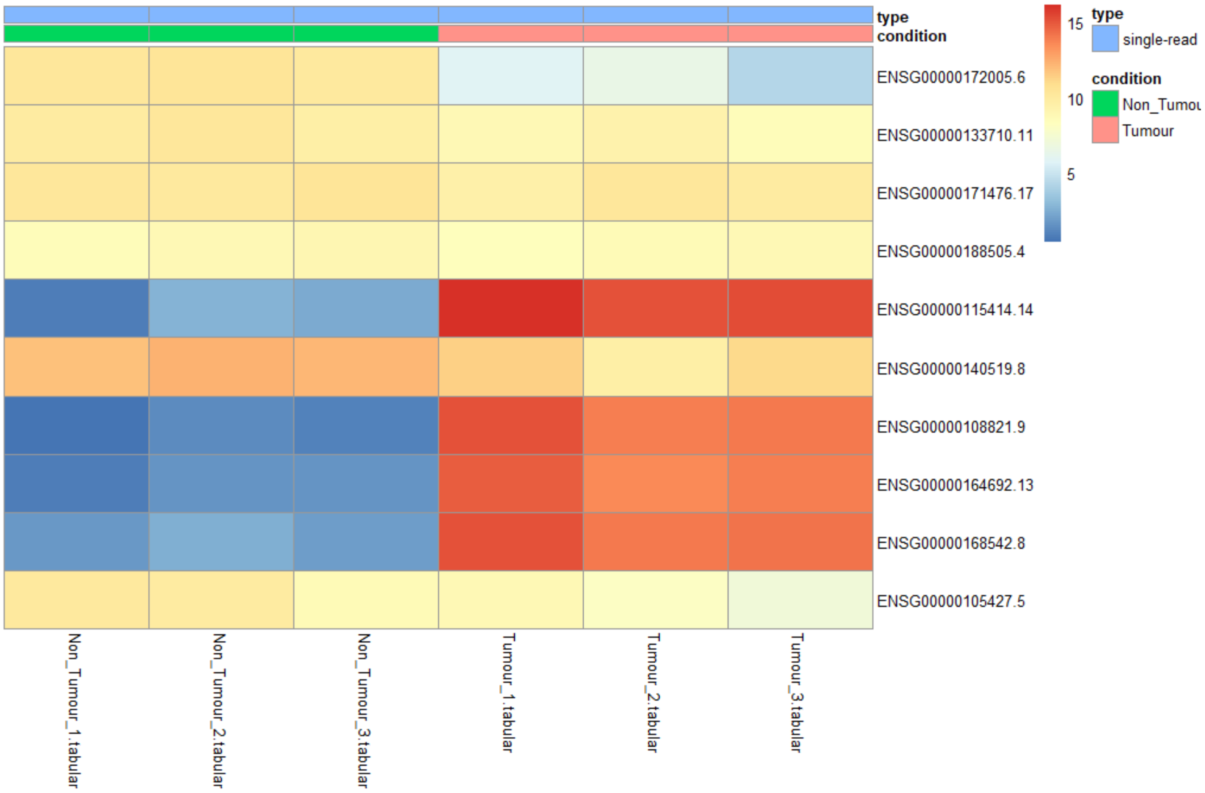


This is graph generated from GoFuncR function, x axis of annotated genes contains name of GO ID which are top of ontology data when sorted by p value.

For eg in first pic chart annotation gene section the amounts of candidate and background genes that are annotated to the GO-categories and the root nodes so 93 candidate gene out of 3563 background genes in cellular component root.

The top panel shows the odds-ratio comparing the GO-categories with their root nodes.

10. Heatmap



This graph was made using top 10 expressing which were calculated when we did ontology process.

From the above heatmap we can clearly see that most of the gene here are highly expressed in tumour cell as compared to non-tumour cell

11. Microarray vs RNA-seq data

Microarray	**RNA-seq**
Old technology, and it contained non discreate data	New technology, and it contains discreate data that is reads have finite number
Most of the old is in the microarray form	Almost all data in the recent is done using RNA-seq
Microarray can be done for sample which can be fluorescently tagged	RNA-seq can be done for many sample
It requires transcript specific probes	It doesn't require any transcript specific probes
It has intensity of light that is captured by camera	It has base with quality score that is captured by camera
For high quality of data high number of RNA are required	High quality data can be generated using even low number of RNA
In this lots of manual work requires so degree of error by human is very large	While most the work is done by computer so degree of error by human is very low