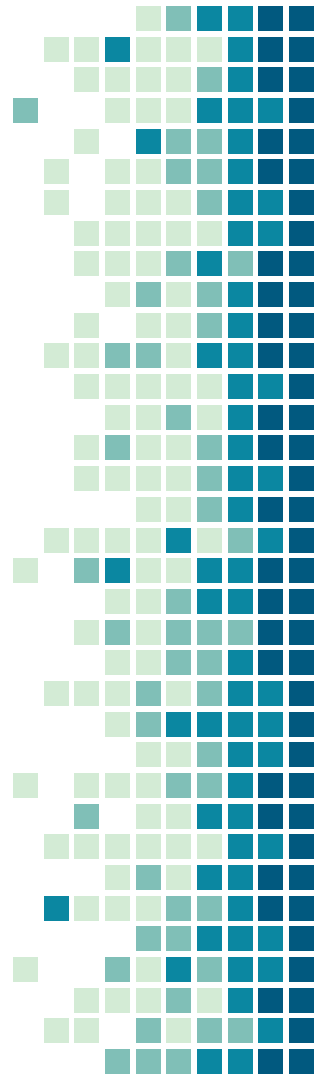


Bioinformatics Summer 2020

Goral, Alex, Isha, Erin, Yves, Annie, Arthur



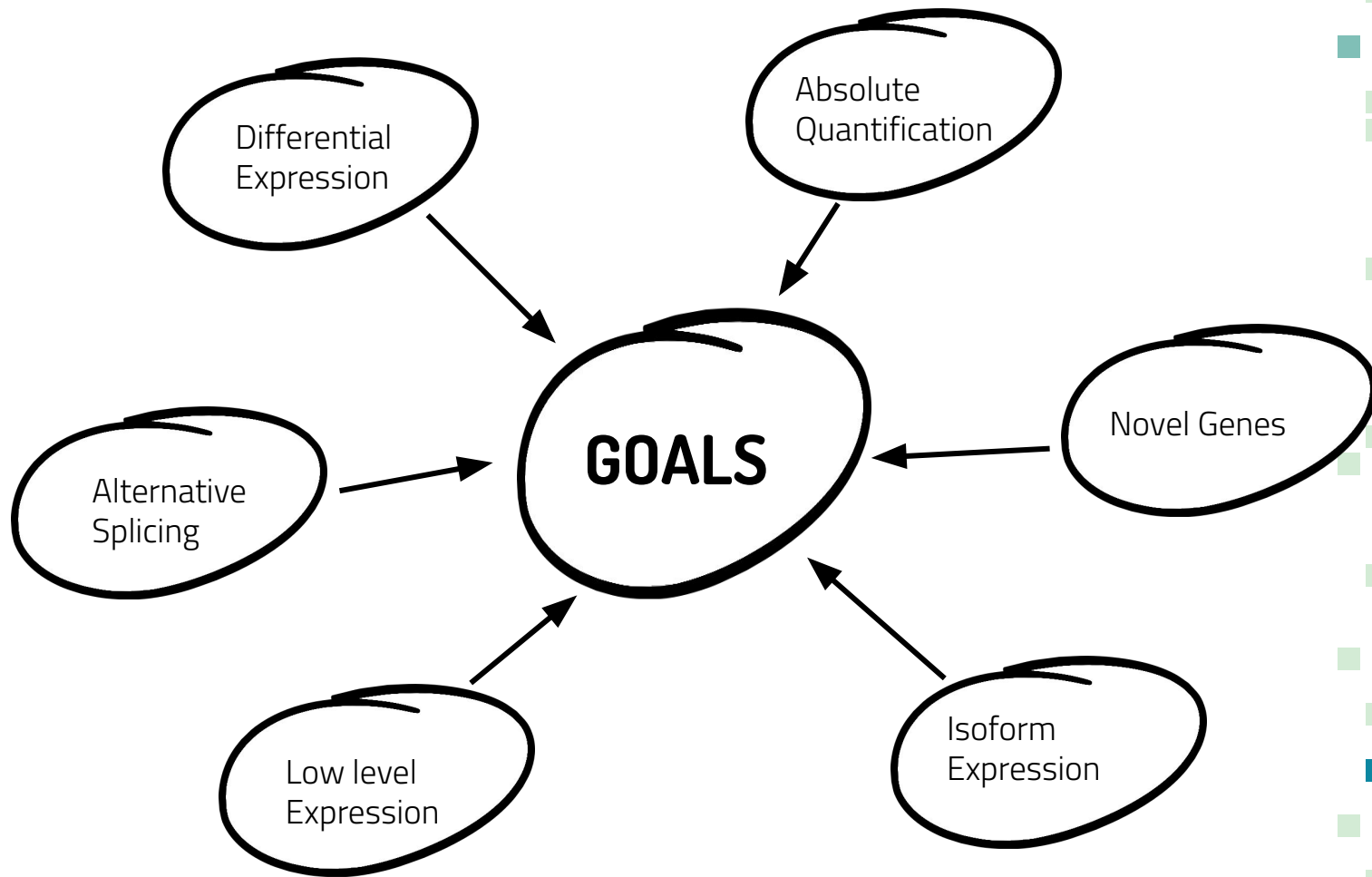
Microarrays vs RNA sequencing



Practical Questions

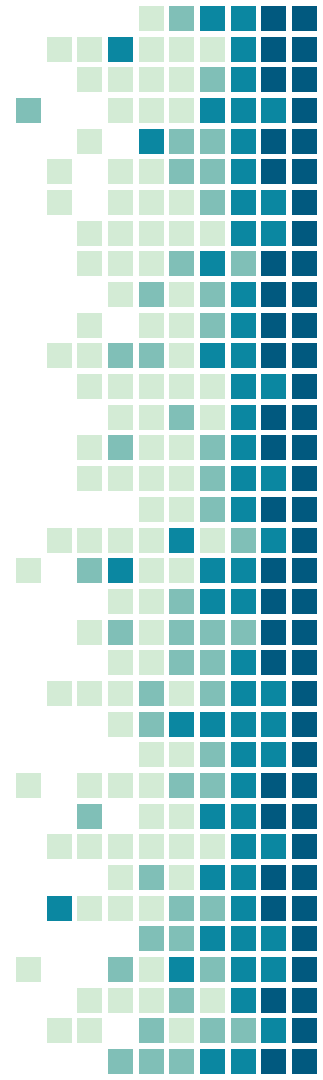
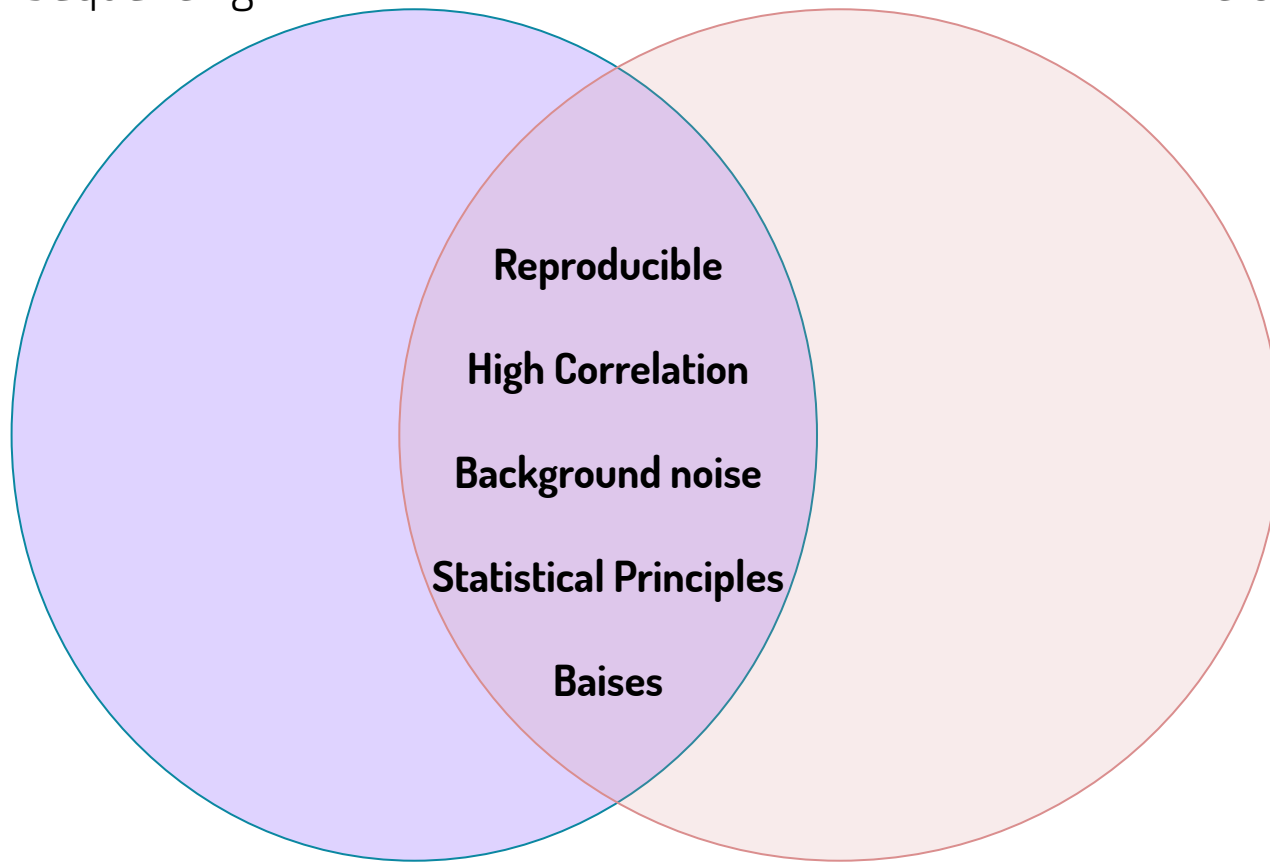
- What genome information is available for my species of interest?
- How much data analysis expertise do I have or have access to?
- How much money do I have to spend?





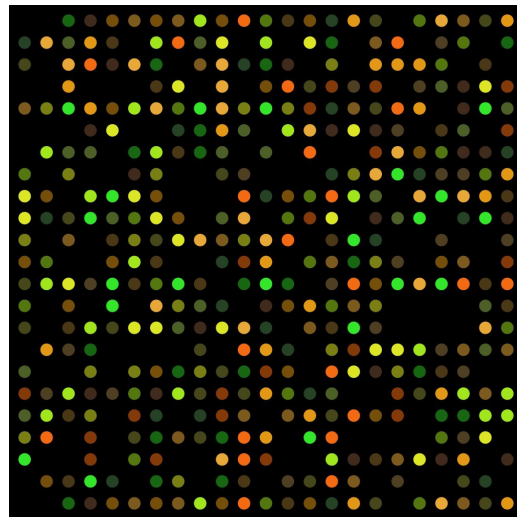
RNA Sequencing

Microarray



Microarrays

- Robust, reliable method, proven over decades of use
- High throughput method - 1000s of samples analysed per month
- Streamlined handling - can easily be automated
- Straightforward data analysis
- Low cost
- Dependent on prior sequence knowledge
- Cannot detect structural variation
- Cannot detect isoforms

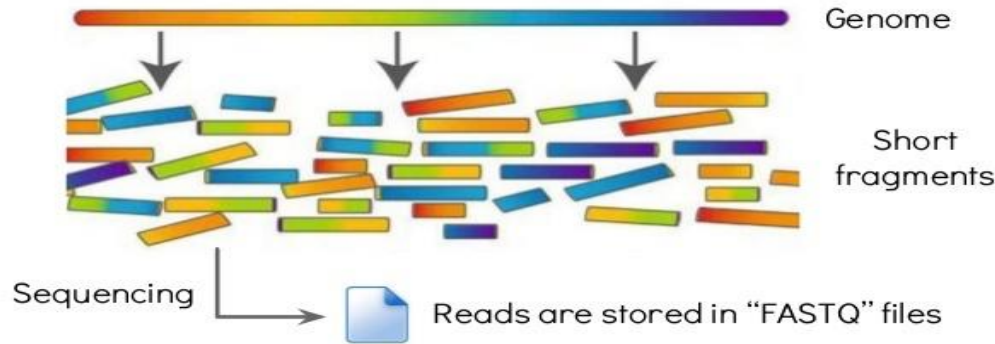


RNA Sequencing

- Provides a comprehensive view of the transcriptome
- Not dependent on any prior sequence knowledge
- Can detect structural variations such as gene fusion and alternative splicing events
- A truly digital solution (absolute abundance vs relative abundance)
- Data storage is more challenging
- Analysis is more complex - no standard protocol
- Specialized computing infrastructure and personnel is required
- More expensive than microarrays



An in-depth look to sequences and RNAseq pipeline



Sequence ID

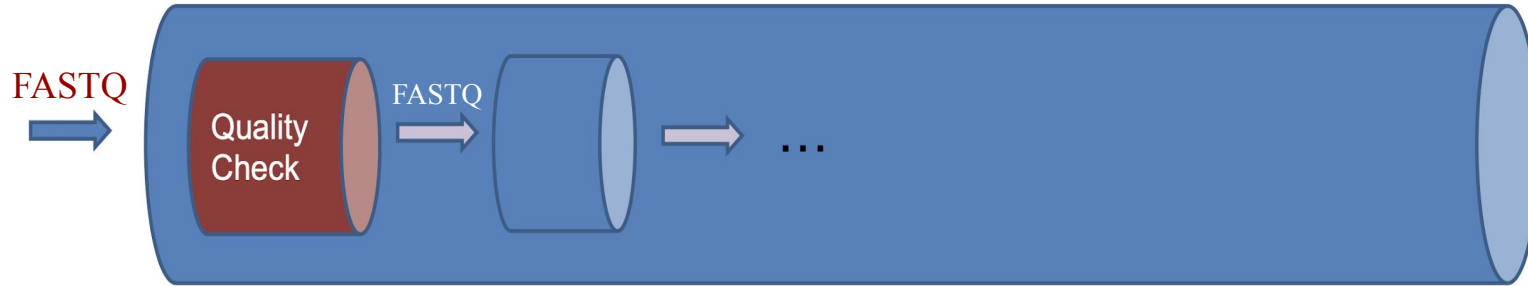
Sequenced Read

```
@SRR081708.237649/1
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
+
!>:<9>>>:=;>?<:@?>;==@@@?>AAA<=>A@?6>4B=<=>.@?<@;?#####
```

Blank

Quality Score

High Level Data Flow Management



The typical first step of a NGS analysis pipeline is the quality check of produced sequences.

FASTQC: free program that reports quality profile of reads

1) Run FASTQC

fastqc sample.fastq

2) Open the output file

sample_fastq.html

Summary



[Basic Statistics](#)



[Per base sequence quality](#)



[Per tile sequence quality](#)



[Per sequence quality scores](#)



[Per base sequence content](#)



[Per sequence GC content](#)



[Per base N content](#)



[Sequence Length Distribution](#)



[Sequence Duplication Levels](#)



[Overrepresented sequences](#)

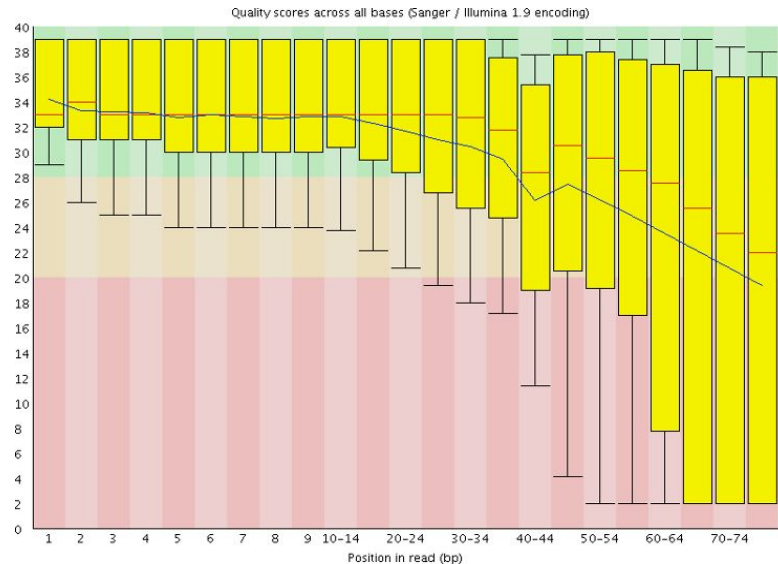
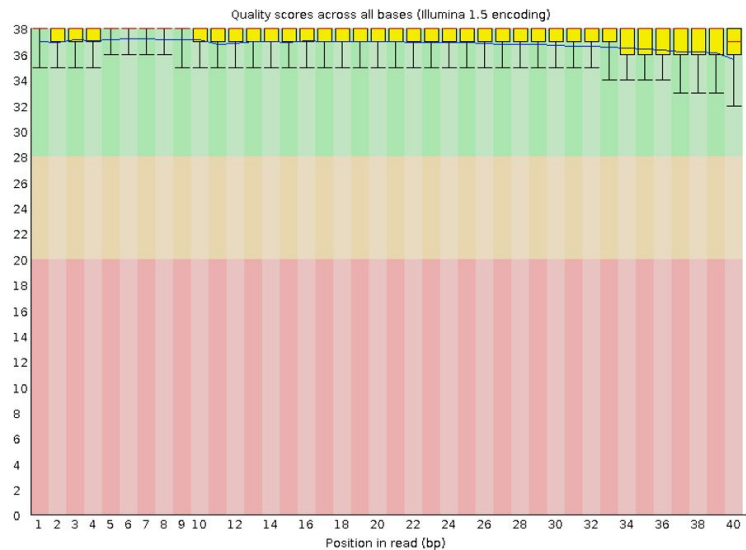


[Adapter Content](#)

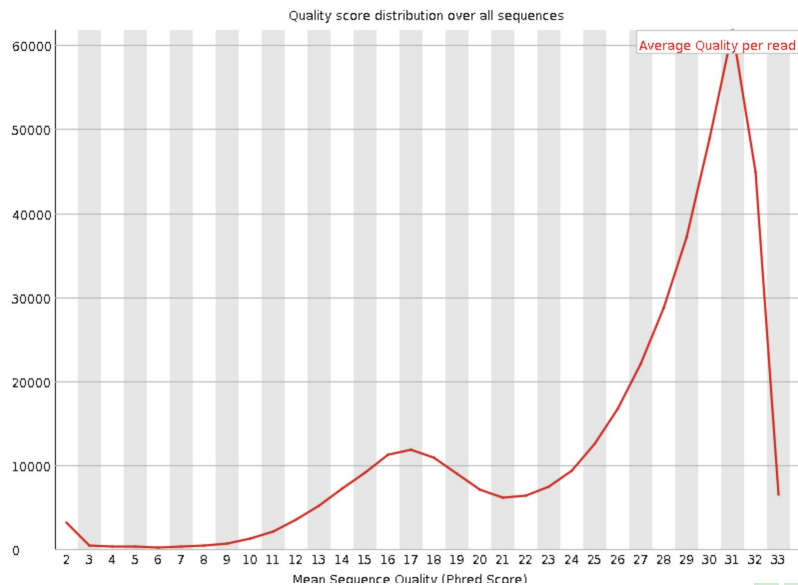
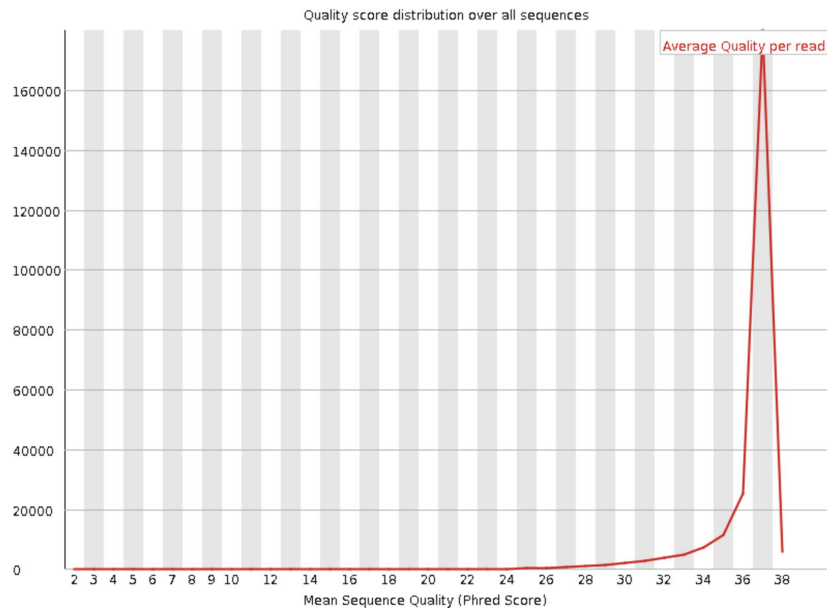


[Kmer Content](#)

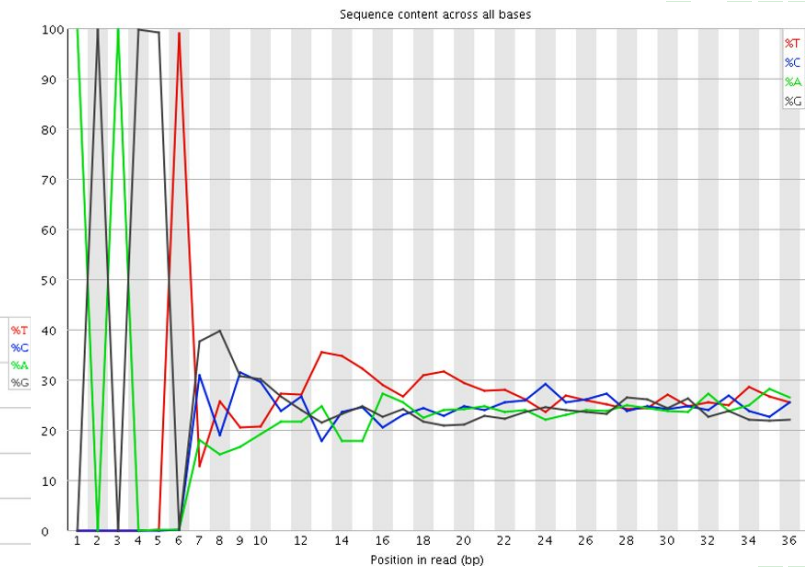
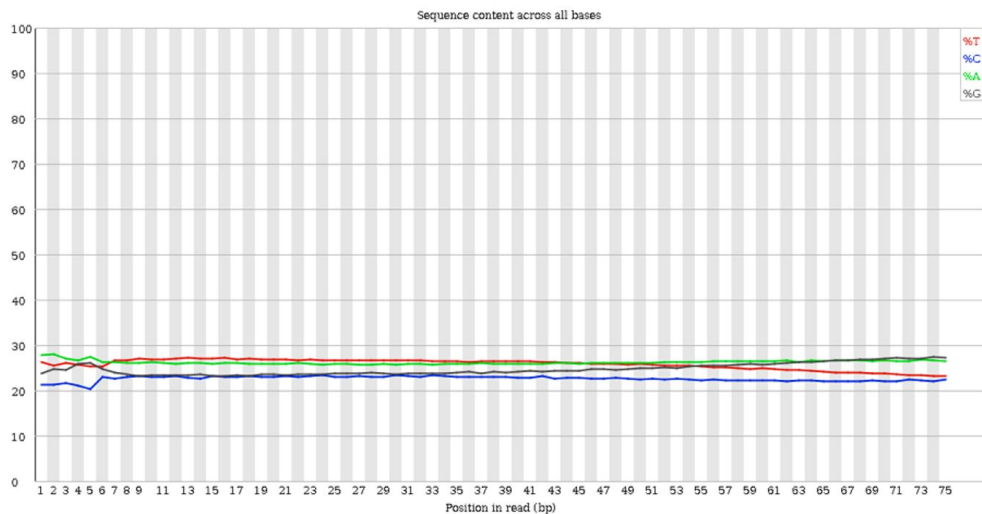
FASTQC: Per base sequence quality



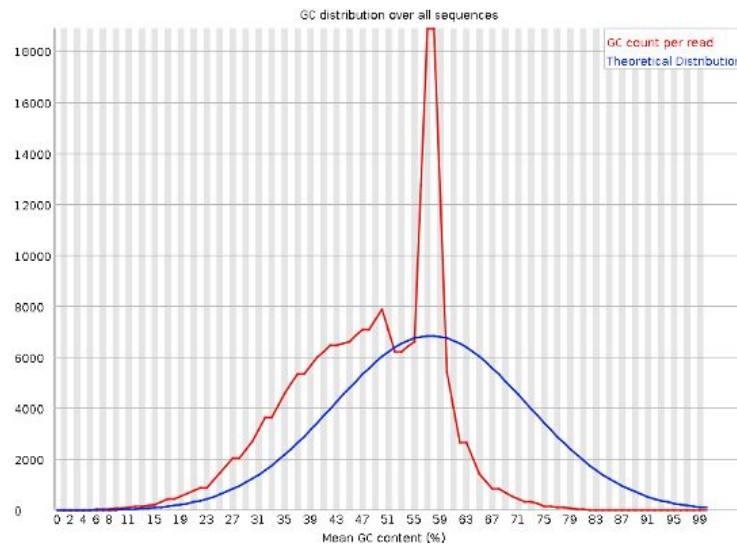
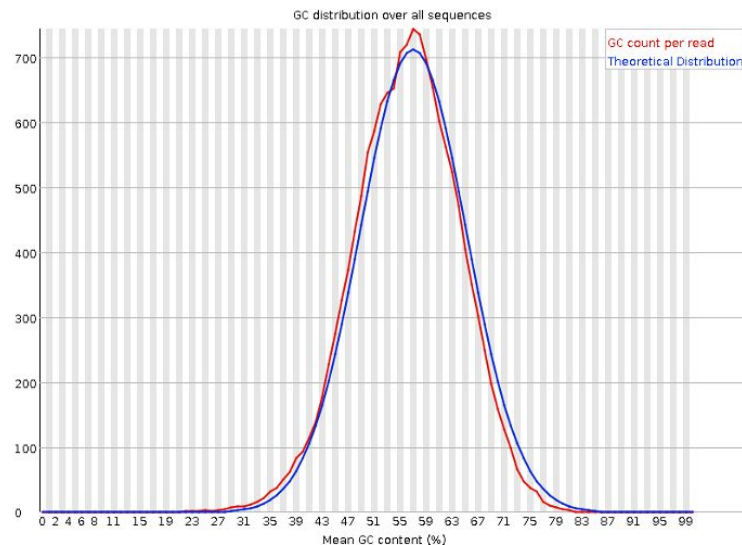
FASTQC: Per sequence quality scores



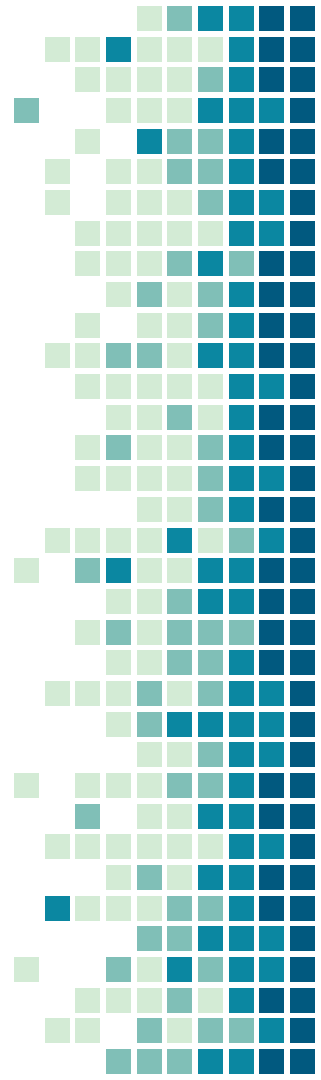
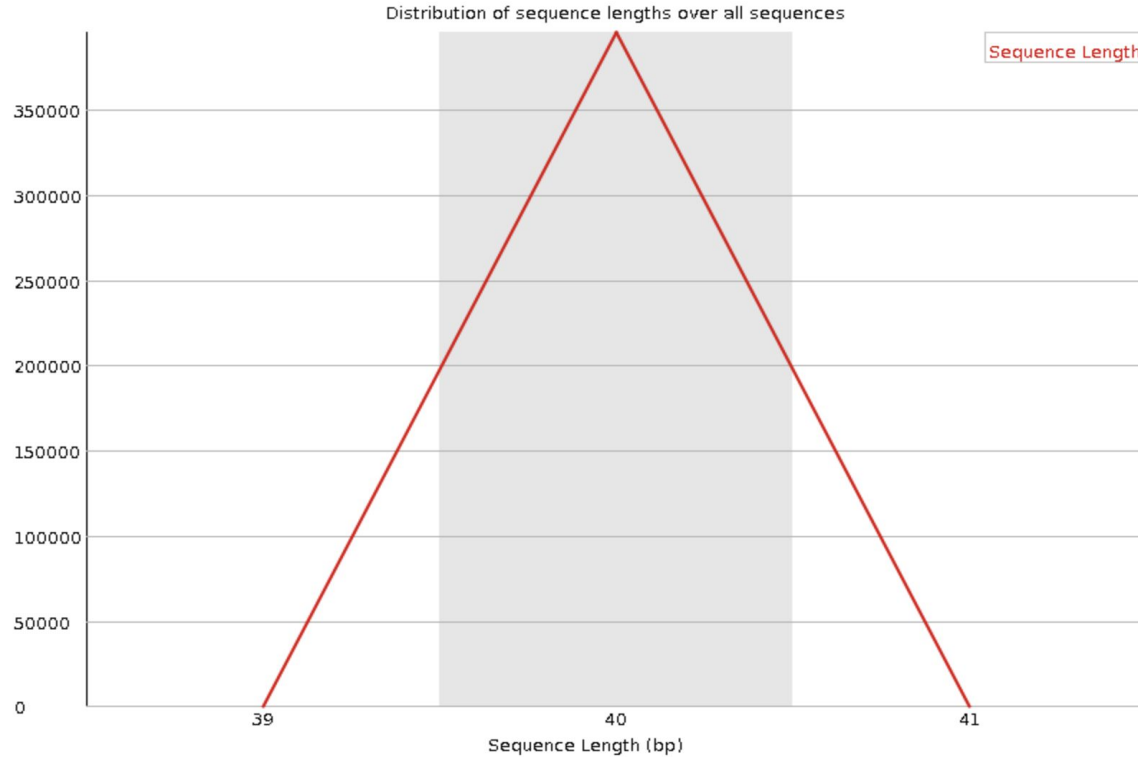
FASTQC: Per base sequence content



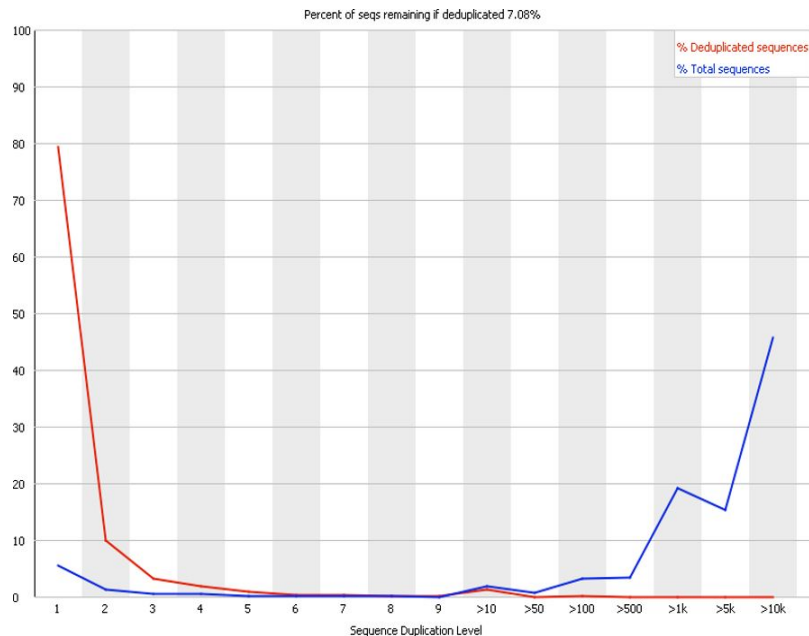
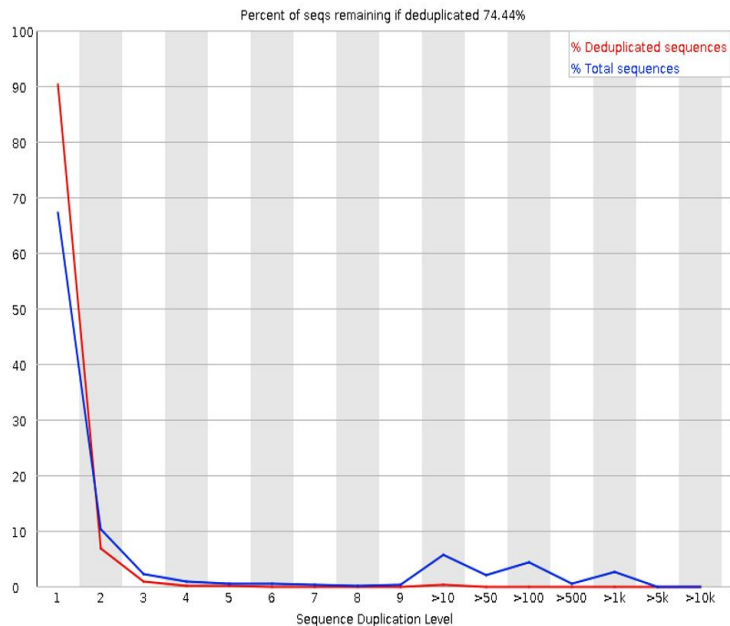
FASTQC: Per sequence GC content



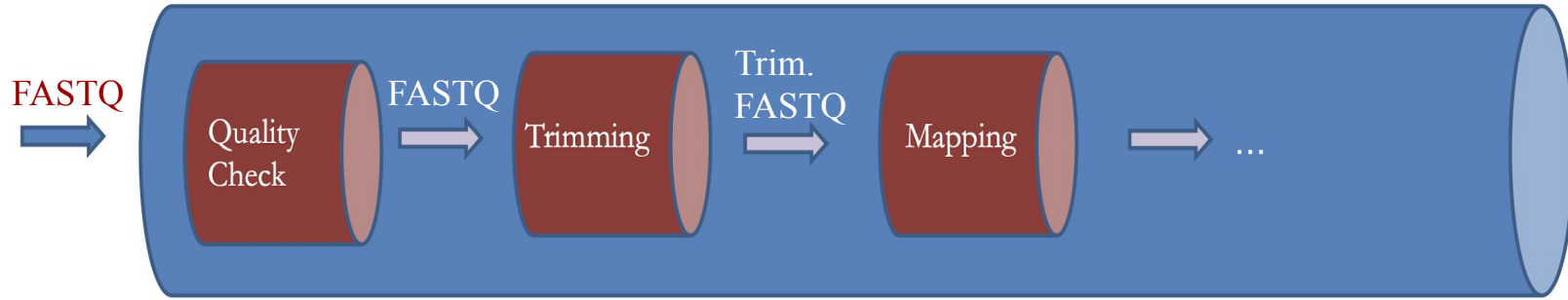
FASTQC: Sequence Length Distribution



FASTQC: Sequence duplication levels

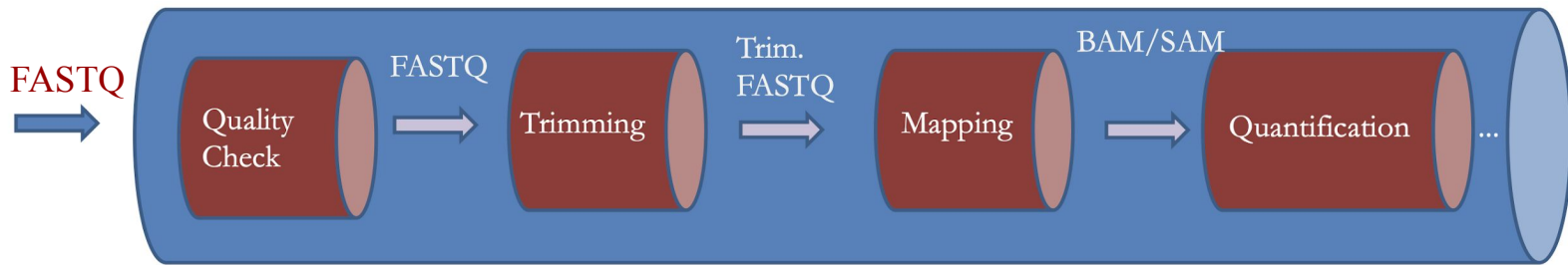


Another step into our pipeline...



Read Mapping

Almost at the end of the tunnel!



Different Measures:

- Raw counts
- Counts per million reads (CPMR)
- RPKM
- FPKM
- Transcripts per million

GEO2R

- Interactive web tool
- Can compare 2+ groups of samples in the GEO series
- **Goal: Identify genes that are differentially expressed across experimental conditions**
- Greater proportion of GEO data can be analyzed in a more time-efficient manner

The screenshot displays the NCBI GEO2R web interface. At the top, the NCBI logo and 'GEO Gene Expression Omnibus' are visible. The navigation bar includes links for HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. The main content area shows the 'Accession Display' for series GSE18388. The search criteria are set to Scope: Self, Format: HTML, Amount: Quick, and GEO accession: GSE18388. The series title is 'Microarray Analysis of Space-flown Murine Thymus Tissue' from the organism 'Mus musculus'. The experiment type is 'Expression profiling by array'. The summary describes a microarray analysis of space-flown murine thymus tissue, identifying 12 genes that were significantly up- or down-regulated by at least 1.5 fold after spaceflight (p<0.05). The interface also lists 8 samples (GSM458594, GSM458595, GSM458596) and a relation to BioProject 118071. At the bottom, there is a section for 'Analyze with GEO2R' and a table for downloading supplementary files.

Supplementary file	Size	Download
...

GEO2R

- “Define Groups” for your samples
- Value of distributions
- Table of results ordered by significance (**lower P value = more significant**)

NCBI

GEO » GEO » GEO2R » GSE18388

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed in a table of genes ordered by significance. Full instructions...

GEO accession: Set Microarray Analysis of Space-flown Murine Thymus Tissue

Samples

Define groups

Enter a group name: List

Group Accession

space flown GSM458594

space flown GSM458595

space flown GSM458596

space flown GSM458597

control GSM458598

control GSM458599

control GSM458600

control GSM458601

Source name

Thymus mRNA extracted from space-flown mi

Thymus mRNA extracted from space-flown mi

Thymus mRNA extracted from space-flown mi

Thymus mRNA extracted from space-flown mi

Thymus mRNA extracted from ground-control

Thymus mRNA extracted from ground-control

Thymus mRNA extracted from ground-control

Thymus mRNA extracted from ground-control

Quick start

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare.
- Assign Samples to each group. Highlight Sample rows then click the group name (or 'Define groups') columns to help determine which Samples belong to which group.
- Click 'Top 250' to perform the calculation with default settings.
- Results are presented as a table of genes ordered by significance. The table may be saved.
- You may change settings in Options tab.

How to use

Top 250 Save all results

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
▶ 10358454	0.00509	1.96e-07	-13.48	4.3773	-1.384	Rbm3	RNA binding motif pr...
▶ 10603469	0.00509	2.86e-07	-12.92	4.24228	-1.304	Rbm3	RNA binding motif pr...
▶ 10556113	0.00546	4.61e-07	-12.24	4.06259	-1.408	Rbm3	RNA binding motif pr...
▶ 10535904	0.00844	9.50e-07	11.28	3.76621	1.894	Hsp1	heat shock 105kDa/1...
▶ 10490946	0.03472	5.44e-06	9.21	2.93118	1.02	Hsp90aa1	heat shock protein 9...
▶ 10514713	0.03472	5.86e-06	9.13	2.89184	0.948	Wdr78	WD repeat domain 78
▶ 10531415	0.03973	7.82e-06	8.83	2.73562	1.499	Cxcl10	chemokine (C-X-C m...
▶ 10391146	0.04368	9.83e-06	8.59	2.60896	1.232	Acly	ATP citrate lyase
▶ 10497079	0.0521	1.32e-05	8.29	2.44138	1.251	Ptger3	prostaglandin E rece...
▶ 10378848	0.05835	1.64e-05	8.08	2.31365	0.882	Hsp90aa1	heat shock protein 9...
▶ 10402615	0.07332	2.27e-05	7.77	2.11964	0.853	Hsp90aa1	heat shock protein 9...
▶ 10465604	0.11029	3.72e-05	7.31	1.81172	0.881	Stip1	stress-induced phos...
▶ 10548194	0.11323	4.14e-05	7.21	1.74391	0.933	Fkbp4	FK506 binding protei...
▶ 10343263	0.12007	4.73e-05	-7.09	1.65832	-1.149		
▶ 10475502	0.12943	5.50e-05	-6.96	1.55967	-1.182		
▶ 10565794	0.12943	5.82e-05	6.91	1.52204	0.869	Serpinh1	serine (or cysteine) p...
▶ 10413542	0.13514	6.46e-05	6.82	1.45337	0.749	Tkt	transketolase
▶ 10342984	0.17386	8.82e-05	-6.56	1.24409	-1.035		
▶ 10580382	0.17386	9.82e-05	6.47	1.17061	0.695	Neto2	neuropilin (NRP) and...
▶ 10517727	0.17386	1.02e-04	6.44	1.14785	0.757	Klhd7a	kelch domain contai...
▶ 10584578	0.17386	1.03e-04	6.42	1.13682	1.008	Hspa8	heat shock protein 8
▶ 10565609	0.17386	1.08e-04	6.39	1.10817	1.211	Thrsp	thyroid hormone resp...
▶ 10481111	0.18898	1.22e-04	6.29	1.01954	0.761		
▶ 10364712	0.19171	1.29e-04	-6.24	0.97985	-1.268	Cirbp	cold inducible RNA b...

R Overview

- R is a language and environment for statistical computing and graphics.
- Includes an effective data handling and storage facility
- Includes graphical facilities for data analysis and display



R - Vector

```
v <- c(1,2,3,4) ## Create a vector v with element 1, 2, 3, 4
```

```
v[3]           # access 3rd element  
v[c(2, 4)]     # access 2nd and 4th element  
v[-1]          # access all but 1st element
```

```
v[2] <- 0      # modify 2nd element  
v[x<3] <- 5    # modify elements less than 0  
v <- x[1:4]     # truncate x to first 4 elements
```

R- Matrix (Creation)

```
> A=matrix(1:9, nrow = 3, ncol = 3) ##Create a 3 by 3 matrix
> A
```

	[,1]	[,2]	[,3]
[1,]	1	4	7
[2,]	2	5	8
[3,]	3	6	9

```
> cbind(c(1,2,3),c(4,5,6))
```

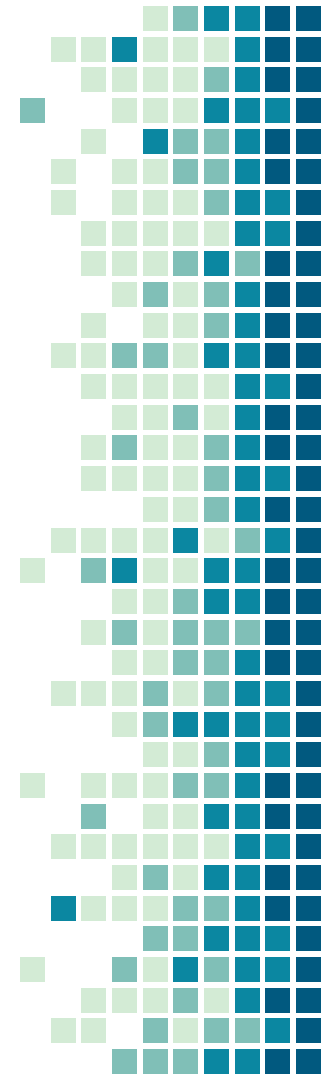
[,1]	[,2]
[1,]	1 4
[2,]	2 5
[3,]	3 6

```
> rbind(c(1,2,3),c(4,5,6))
```

[,1]	[,2]	[,3]	
[1,]	1	2	3
[2,]	4	5	6

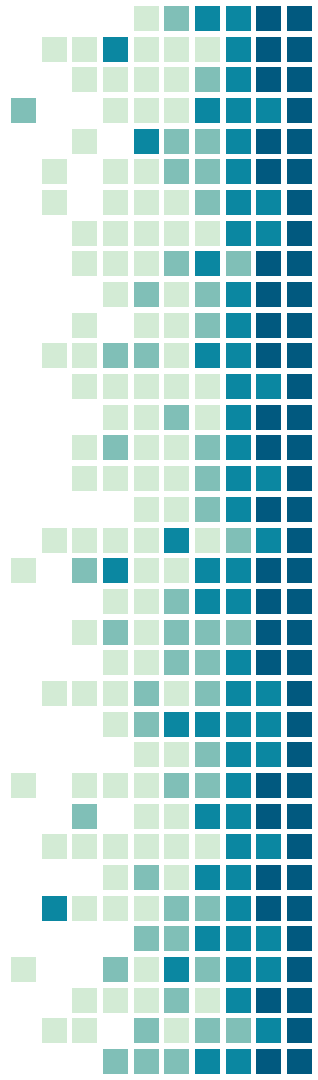
R- Matrix (vector as index)

```
> x
[,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> x[c(1,2),c(2,3)]      # select rows 1 & 2 and columns 2 & 3
[,1] [,2]
[1,]    4    7
[2,]    5    8
> x[c(3,2),]           # leaving column field blank will select entire columns
[,1] [,2] [,3]
[1,]    3    6    9
[2,]    2    5    8
> x[,]                # leaving row as well as column field blank will select entire matrix
[,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> x[-1,]              # select all rows except first
[,1] [,2] [,3]
[1,]    2    5    8
[2,]    3    6    9
```



Volcano Plot

- Volcano plots represent a useful way to visualize the results of differential expression analyses.
- A volcano plot is a type of scatterplot that shows statistical significance (P value) versus magnitude of change (fold change).

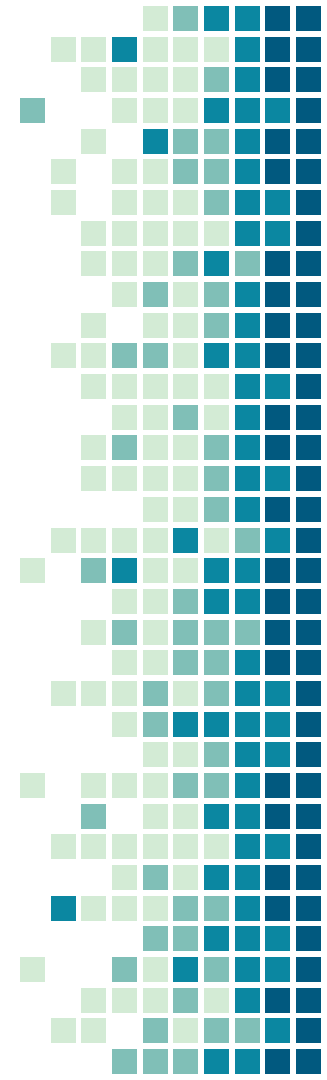


Volcano Plot (Installation and Syntax)

```
BiocManager::install('EnhancedVolcano')
```

```
library(EnhancedVolcano)
```

```
EnhancedVolcano(data, lab = rownames(data),  
x = 'log2FoldChange', y = 'pvalue', xlim = c(..., ...))
```

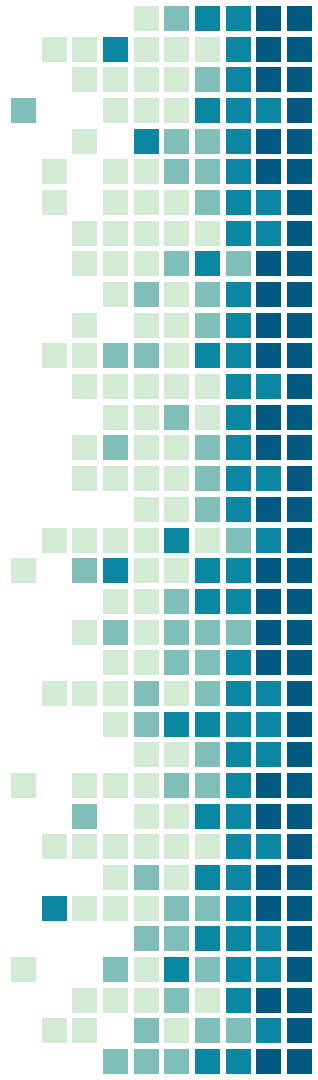


Example: Volcano Plot (Data Input)

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
▶ 10358454	0.00509	1.96e-07	-13.48	4.3773	-1.384	Rbm3	RNA binding motif pr...
▶ 10603469	0.00509	2.86e-07	-12.92	4.24228	-1.304	Rbm3	RNA binding motif pr...
▶ 10556113	0.00546	4.61e-07	-12.24	4.06259	-1.408	Rbm3	RNA binding motif pr...
▶ 10535904	0.00844	9.50e-07	11.28	3.76621	1.894	Hsph1	heat shock 105kDa/1...
▶ 10490946	0.03472	5.44e-06	9.21	2.93118	1.02	Hsp90aa1	heat shock protein 9...
▶ 10514713	0.03472	5.86e-06	9.13	2.89184	0.948	Wdr78	WD repeat domain 78
▶ 10531415	0.03973	7.82e-06	8.83	2.73562	1.499	Cxcl10	chemokine (C-X-C m...
▶ 10391146	0.04368	9.83e-06	8.59	2.50896	1.232	Acl1	ATP citrate lyase
▶ 10497079	0.0521	1.32e-05	8.29	2.44138	1.251	Pltger3	prostaglandin E rece...
▶ 10378848	0.05835	1.64e-05	8.08	2.31365	0.882	Hsp90aa1	heat shock protein 9...
▶ 10402615	0.07332	2.27e-05	7.77	2.11964	0.853	Hsp90aa1	heat shock protein 9...
▶ 10465604	0.11029	3.72e-05	7.31	1.81172	0.881	Stip1	stress-induced phos...
▶ 10548194	0.11323	4.14e-05	7.21	1.74391	0.933	Fkbp4	FK506 binding protei...
▶ 10343263	0.12007	4.73e-05	-7.09	1.65832	-1.149		
▶ 10475502	0.12943	5.50e-05	-6.96	1.55967	-1.182		
▶ 10565794	0.12943	5.82e-05	6.91	1.52204	0.869	Serpinh1	serine (or cysteine) p...
▶ 10413542	0.13514	6.46e-05	6.82	1.45337	0.749	Tkt	transketolase
▶ 10342984	0.17386	8.82e-05	-6.56	1.24409	-1.035		
▶ 10580382	0.17386	9.82e-05	6.47	1.17061	0.695	Neto2	neuropilin (tRNP) and...
▶ 10517727	0.17386	1.02e-04	6.44	1.14785	0.757	Klhd7a	kelch domain contain...
▶ 10584578	0.17386	1.03e-04	6.42	1.13682	1.008	Hspa8	heat shock protein 8...
▶ 10565609	0.17386	1.08e-04	6.39	1.10817	1.211	Thrsp	thyroid hormone resp...
▶ 10481111	0.18898	1.22e-04	6.29	1.01954	0.761		
▶ 10364712	0.19171	1.29e-04	-6.24	0.97985	-1.268	Cirbp	cold inducible RNA b...

Example: Volcano Plot (Creation)

```
EnhancedVolcano(data, lab = rownames(data),  
x = 'log2FoldChange', y = 'pvalue', xlim = c(-5, 5))
```



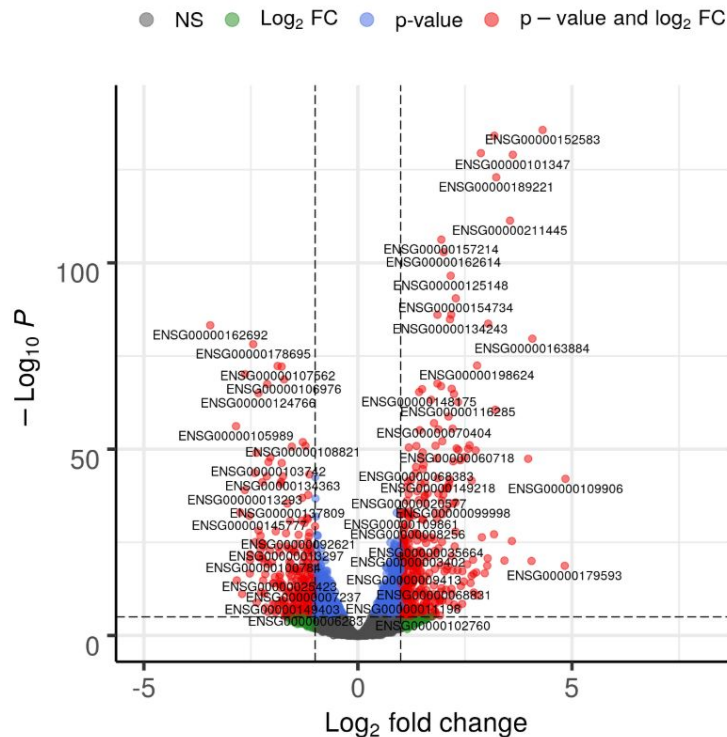
Volcano Plot (Full Code)

```
# install package
BiocManager::install("EnhancedVolcano")
library(EnhancedVolcano)

# load data
genetable <- read.table("geo2r.txt", header = T)
# use the ID to name the rowname
rownames(genetable) <- genetable$ID
# make plot and modify cut-offs for log2FC and P value
# ; specify title; adjust point and label size
EnhancedVolcano_plot<- EnhancedVolcano(genetable,
    lab = rownames(genetable),
    pCutoff = 0.05,
    FCcutoff = 0.5,
    subtitle = NULL,
    x = 'logFC',
    y = 'adj.P.Val',
    pointSize=1,
    legendLabSize = 10,
    xlim = c(-1.5, 1.5),
    labSize = 3.0)
```



Example: Volcano Plot (Plot)



Program Goals

- Learn technical skills
- Work with teams to apply skills



Putting It All Together- Project Overview

- *Construction and Analysis of a ceRNA Network Reveals Potential Prognostic Markers in Colorectal Cancer*
 - Using portion of the dataset from the original paper
- Objective: To find differentially expressed genes that act as prognostic markers for colorectal cancer
- Next week: Training using part of the dataset

