An Introduction to RNA-sequencing

GSND 5340Q, BMDA

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School

2025-05-20

Installing R Packages:

Install the following tools: `Rsubread`, `Rsamtools`, `edgeR`, `DESeq2`, `sva` `SummarizedExperiment`, `ComplexHeatmap`, `umap`, and the `TBSignatureProfiler`. We will also need help from the `tidyverse`.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install(c("Rsubread","Rsamtools","tidyverse",
  "SummarizedExperiment","edgeR","DESeq2","sva",
  "ComplexHeatmap", "TBSignatureProfiler","umap"))
```

Installing and using the SCTK

```
install.packages("devtools")
devtools::install_github("wevanjohnson/singleCellTK")
library(singleCellTK)
singleCellTK()

### Example: open downstream_analysis/
### features_combined.txt and meta_data.txt
```

Load Packages for RNA-Seq

We will be using the following packages for our RNA-seq lecture:

```
library(tidyverse) ## tools for data wranging
library(Rsubread) ## alignment and feature counts
library(Rsamtools) ## managing .sam and .bam files
library(SummarizedExperiment) ## managing counts data
library(edgeR) ## differential expression
library(DESeq2) ## differential expression
library(ComplexHeatmap) ## Heatmap visualization
library(TBSignatureProfiler) ## TB signature analysis
library(umap) ## dimenstion reduction and plotting data
```

Objective

- Disclaimer: non-comprehensive introduction to RNA-sequencing
- Introduce preprocessing steps
- Visualization
- Analytical methods
- Common software tools

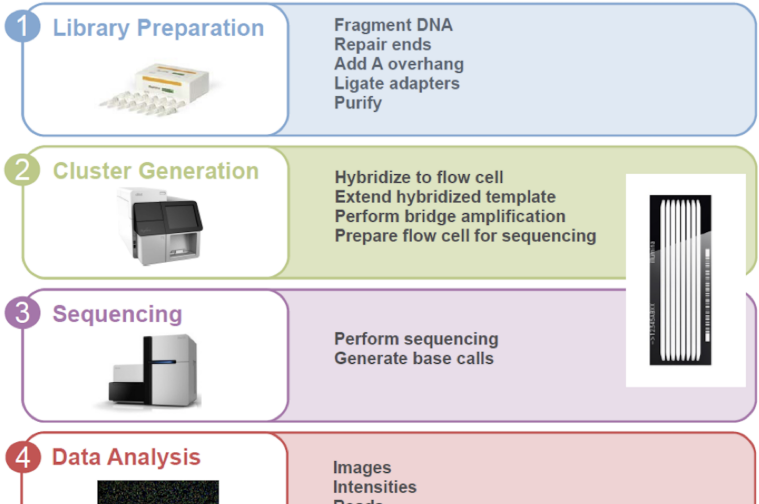
Steps to an RNA-seq Analysis (Literacy)

1. Preprocessing and QC:
   - Fasta and Fastq files
   - FastQC: good vs. bad examples
   - Visualization
2. Alignment
   - Obtaining genome sequence and annotation
   - Software: Bowtie, TopHat, STAR, Subread/Rsubread
3. Expression Quantification
   - Count reads hitting genes, etc
   - Approaches/software: HT-Seq, STAR, Cufflinks, RPKM FPKM or CPM, RSEM, edgeR, findOverlaps (GenomicRanges). featureCounts (Rsubread)

Steps to an RNA-seq Analysis (Literacy)

4. More visualization
   - Heatmaps, boxplots, PCA, t-SNE, UMAP
5. Differential Expression
   - Batch correction
   - Overdispersion
   - General Workflow
   - Available tools: edgeR, DESeq, Limma/voom
   - Even more visualization!!

Illumina Sequencing Workflow



Sequencing Data Formats

Genome sequcencing data is often stored in one of two formats, FASTA and FASTQ text files. For example a FASTA files looks like the following:

```
>chrX
ttgaactcctgacctcaggtgatccgcccgccttgacctcccaaagcgct
cctgcctcagcctcccgagtagctgggactacaggtgcctgccaccatgc
....
caggctaatttttgtatttttagtagagacggggtttcaccatgttagc
caggatggtctcaatctcctgacctcatgatccgcctgcctcggcctccc
>chrY
tgtacactttaaatgggtgaattttatggaatgtgaattataCGTGTTGG
CTTGTAAAAAAAAATGATGGAGATGGAGACGTGACTCTAGCGTGAAGGGG
```

FASTQ Files

We can also store confidence or quality scores using a FASTQ format:

```
@HWI-EAS240_0001:2:1:1142:17571#0/1
CTCTCTTTTCTCCCCANGTCTCCTCATGACCATATCCNTGTTGTCCATTGTGTANGGNNNCTTGCGCTGCAACATC
+HWI-EAS240_0001:2:1:1142:17571#0/1
bababb`abbbbabbYB^[^^`_`_bbbbbbbbbb`_B`_``^`^bbbba^`^^BOOBBBNOOOObY^^bbbbbbb
@HWI-EAS240_0001:2:1:1142:6453#0/1
CAGGACGTGCACTATGNCATCCACGATGCAAGTCTTANCATTATTCAGGATACANATANNGGCAAACGTGTAATTA
+HWI-EAS240_0001:2:1:1142:6453#0/1
_U][aa_aa^^```^[B][][]]]S```Y``a`aaZ]B]]^^^[^```[OVY]XBNNNBB]]]]]L^^^^```[V`
@HWI-EAS240_0001:2:1:1142:19443#0/1
TCAGAAAACAGAAAGGTCATTTTCTTACTTCTTTGCANGATGCCACCCTCCAGANCAGNNAATTGCTACTTTGCAC
+HWI-EAS240_0001:2:1:1142:19443#0/1
bbbabb^^bbb^bbabb]bb]^]`]]]bbbbbbbbYYBYXOOOOObb`bbVY^HBOOOBBOJOLO_bbbb]b_b`a
```

FASTQ Encoding

In order to translate FASTQ quality scores:

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

FASTQ Probability

And now converting to confidence probabilities:

> Sanger

	ASCII	Quality
!	33	0.0000
"	34	0.2057
#	35	0.3690
$	36	0.4988
%	37	0.6019
&	38	0.6838
'	39	0.7488
(	40	0.8005
)	41	0.8415
*	42	0.8741
+	43	0.9000

> Solexa

	ASCII	Quality
;	59	0.2403
<	60	0.2847
=	61	0.3339
>	62	0.3869
?	63	0.4427
@	64	0.5000
A	65	0.5573
B	66	0.6131
C	67	0.6661
D	68	0.7153
E	69	0.7597
F	70	0.7992
G	71	0.8337

> Illumina1.3

	ASCII	Quality
@	64	0.5000
A	65	0.5573
B	66	0.6131
C	67	0.6661
D	68	0.7153
E	69	0.7597
F	70	0.7992
G	71	0.8337
H	72	0.8632
I	73	0.8882
J	74	0.9091
K	75	0.9264
L	76	0.9406

> Illumina1.5

	ASCII	Quality
B	66	0.3690
C	67	0.4988
D	68	0.6019
E	69	0.6838
F	70	0.7488
G	71	0.8005
H	72	0.8415
I	73	0.8741
J	74	0.9000
K	75	0.9206
L	76	0.9369
M	77	0.9499
N	78	0.9602

Preprocessing and QC using FASTQC

FastQC provides a simple way to do QC checks on raw sequence data:

- Import of data from BAM, SAM or FastQ files
- Quick overview and summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

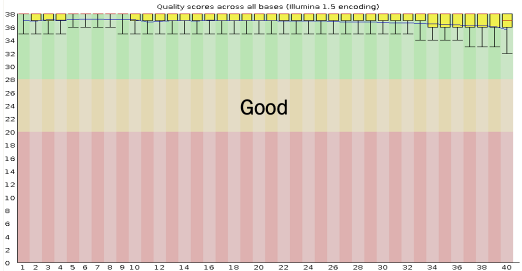
Preprocessing and QC using FASTQC

To run FastQC you can launch the GUI app, or run form the command line:

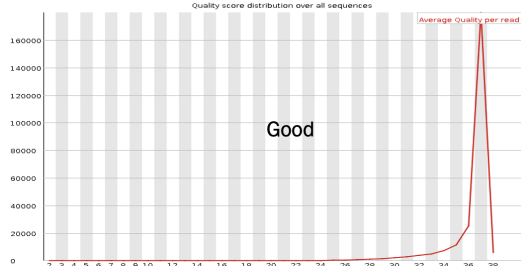
```
rna_seq/FastQC/./fastqc \
  rna_seq/reads/R01_10_short500K.fq.gz
```

FastQC Score Distribution

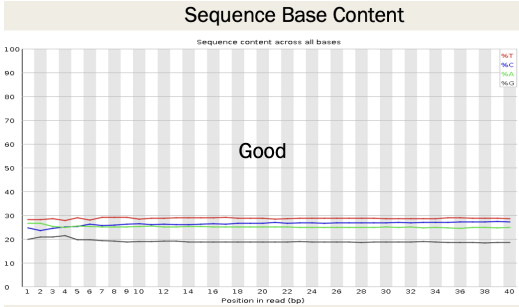
Quality scores

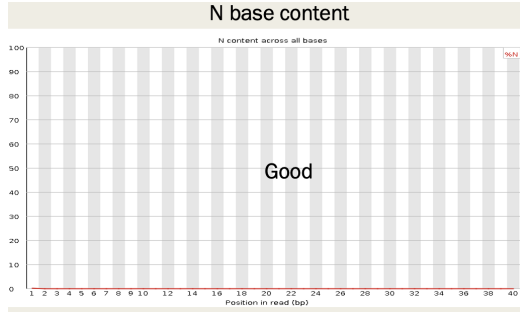


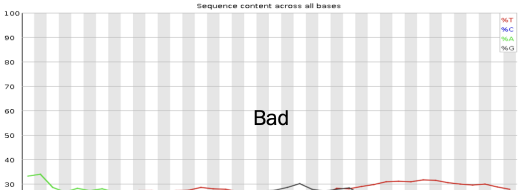
Quality score distribution

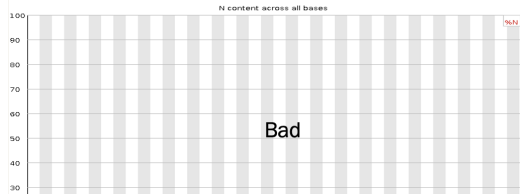


FastQC Base and N Distribution



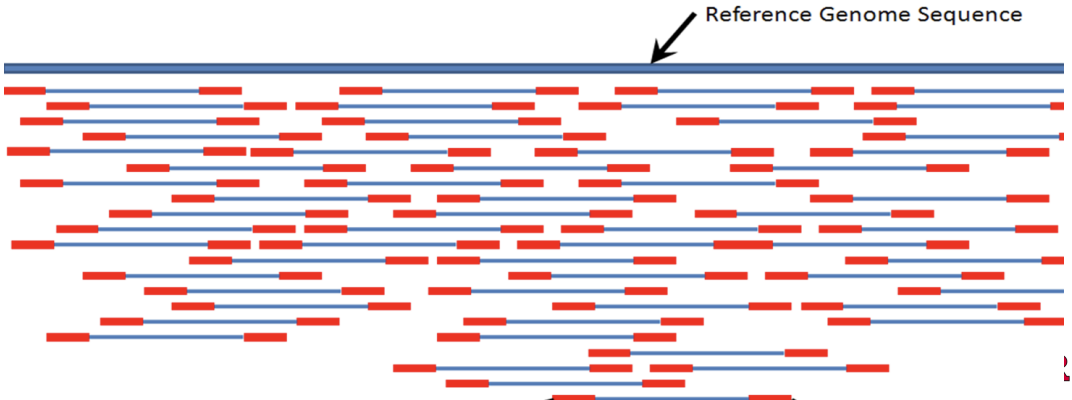






Alignment to the Reference Genome

Goal: Find the genomic Location of origin for the sequencing read. Software: `Bowtie2`, `TopHat`, `STAR`, `Subread/Rsubread`, many others!



Indexing your genome

Abraham Lincoln: "Give me six hours to chop down a tree and I will spend the first four sharpening the axe."

(4 minutes indexing the genome, 2 minutes aligning the reads)

Indexing your genome

Note that you will rarely do this for human alignment. You will usually download an existing index given to you by others who have already done this work. You will do this often if you are aligning microbial reads, e.g. MTB or some other organism for which others have not already made your index for you.

```
buildindex(basename="rna_seq/genome/ucsc.hg19.chr1_120-150M",
 reference="rna_seq/genome/ucsc.hg19.chr1_120-150M.fasta.gz")
```

Took me ~0.2 minutes!

Aligning your reads:

Note that this outputs results in a .bam file and not a .sam file

```
align(index="rna_seq/genome/ucsc.hg19.chr1_120-150M",
      readfile1="rna_seq/reads/R01_10_short500K.fq.gz",
      output_file="rna_seq/alignments/R01_10_short.bam",
      nthreads=4)
```

My laptop is an Apple M2, which has 8 cores (used 4 cores), 24GB RAM:

- Took 15.7 minutes to align ~60M reads to the 30M bases
- Took 0.7 minutes to align ~6.5M reads to the 30M bases
- Took 0.3 minutes to align ~500K reads to the 30M bases

Aligned Sequencing Data Formats (SAM and BAM)

Note that Rsubread outputs a .bam file (bam = binary alignment map) and not a .sam file (sam = sequence alignment map). Here is some information about a .sam file: https://en.wikipedia.org/wiki/SAM_(file_format)

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Aligned Sequencing Data Formats (SAM and BAM)

To convert .sam to .bam or vice versa, a package called Rsamtools. Using Rsamtools, you can convert bam to sam as follows:

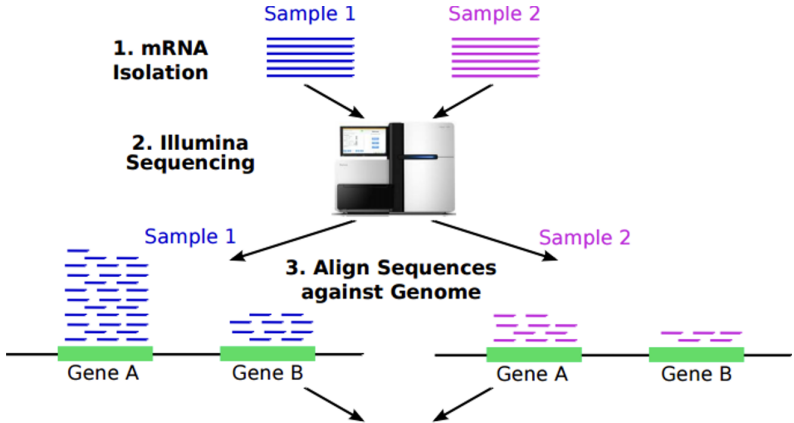
```
asSam("rna_seq/alignments/R01_10_short.bam",
      overwrite=T)

# To convert to bam:
#asBam("rna_seq/alignments/R01_10_short.bam")
```

Feature counts

Now we can count reads hitting genes. Approaches/software:

- HT-Seq
- STAR
- Cufflinks
- RPKM FPKM or CPM
- RSEM
- edgeR
- findOverlaps (GenomicRanges)
- featureCounts (Rsubread)

Feature counts



Feature counts

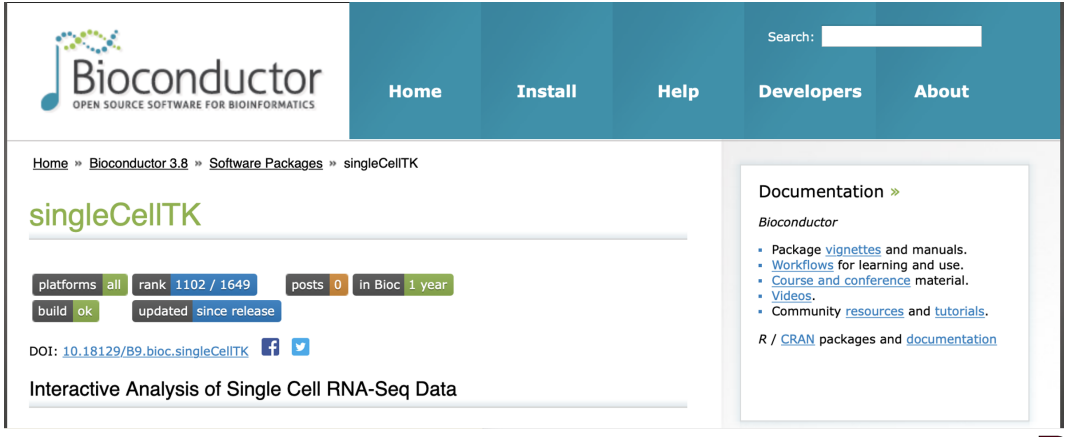
```
fCountsList = featureCounts(
  "rna_seq/alignments/R01_10_short.bam",
  annot.ext="rna_seq_files/genome/genes.chr1_120-150M.gtf",
  isGTFAnnotationFile=TRUE)

featureCounts = cbind(fCountsList$annotation[,1],
                      fCountsList$counts)

write.table(featureCounts,
    "rna_seq/alignments/R01_10_short.features.txt",
    sep="\t", col.names=FALSE, row.names=FALSE, quote=FALSE)
```

SCTK

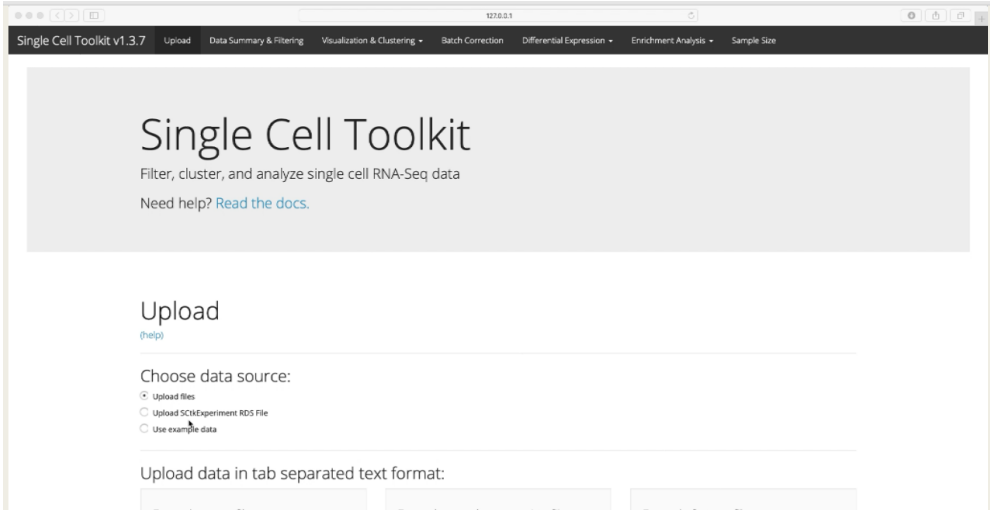
Use the Single Cell Toolkit (SCTK) to analyze your RNA-seq data!

- Inputs: RNA-seq, Nanostring, Proteomic, immunological assay data
- Interactive analyses and visualization of data
- Save results, figures, etc
- Sophisticated data structures
- R/Bioconductor package

SCTK


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Search:
Home
Install
Help
Developers
About
Home » Bioconductor 3.8 » Software Packages » singleCellTK
singleCellTK
platforms all
rank 1102 / 1649
posts 0
in Bioc 1 year
build ok
updated since release
DOI: 10.18129/B9.bioc.singleCellTK
Interactive Analysis of Single Cell RNA-Seq Data
Documentation »
Bioconductor
Package vignettes and manuals.
Workflows for learning and use.
Course and conference material.
Videos.
Community resources and tutorials.
R / CRAN packages and documentation

SCTK


127.0.0.1
Single Cell Toolkit v1.3.7
Upload
Data Summary & Filtering
Visualization & Clustering
Batch Correction
Differential Expression
Enrichment Analysis
Sample Size
Single Cell Toolkit
Filter, cluster, and analyze single cell RNA-Seq data
Need help? Read the docs.
Upload
(help)
Choose data source:
Upload files
Upload SCtkExperiment RDS File
Use example data
Upload data in tab separated text format:

Installing and using the SCTK

```
install.packages("devtools")
devtools::install_github("wevanjohnson/singleCellTK")
library(singleCellTK)
singleCellTK()

### Example: open downstream_analysis/
### features_combined.txt and meta_data.txt
```

Data Structures

A data structure is a particular way of organizing data in a computer so that it can be used effectively. The idea is to reduce the space and time complexities of different tasks.

Data Structures

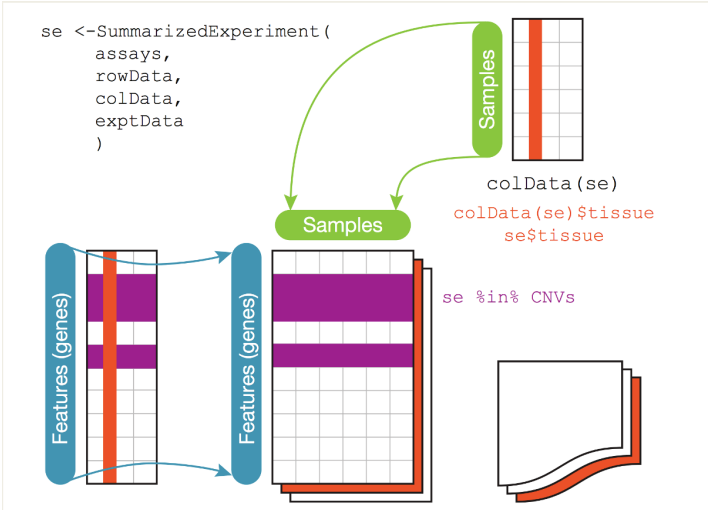
Data structures in R programming are tools for holding multiple values, variables, and sometimes functions

Please think very carefully about the way you manage and store your data! This can make your life much easier and make your code and data cleaner and more portable!

Data Structures

There are advanced R data structures, **S3** and **S4** class objects, that can facilitate object orientated programming. One useful example of an S4 class data structure is the **SummarizedExperiment** object.

Data Structures

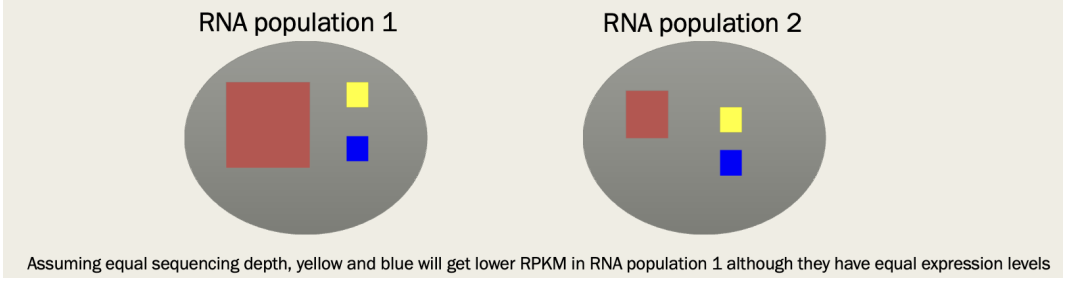


Normalization

Need to normalize data because of:

- Sequencing depth difference in each RNA sample
- RNA composition differences
- Highly expressed genes can consume a substantial proportion of RNA-Seq reads, causing other genes to be under-sampled
- Different methods
  - Log counts
  - Counts per million (CPM and logCPM; RPKM, FPKM)
  - Trimmed mean of M-values (edgeR/limma)
  - Median of Ratios method (DESeq)

Normalization

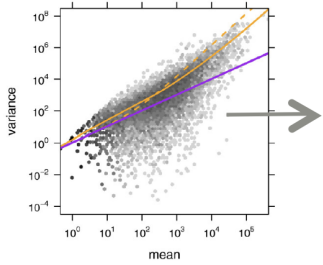


Problem of overdispersion:

Alignment and feature counting result in discrete count data (i.e. the number of reads to each gene). A first thought might be to use a Poisson distribution to model the counts. However, the Poisson makes a strict mean-variance assumption (i.e. they are the same). Studies have demonstrated that a negative binomial fits data better.

Problem of overdispersion:

Many studies have shown that the variance grows faster than the mean in RNAseq data. This is known as **overdispersion**.



Software use:
Negative binomial distribution
Non-parametric methods
Voom transformation

- Mean count vs variance of RNA seq data. Orange line: the fitted observed curve. Purple: the variance implied by

Batch effects

> *Batch Effect: Non-biological variation due to differences in batches of data that confound the relationships between covariates of interest.*

Batch effects are caused by differences in:

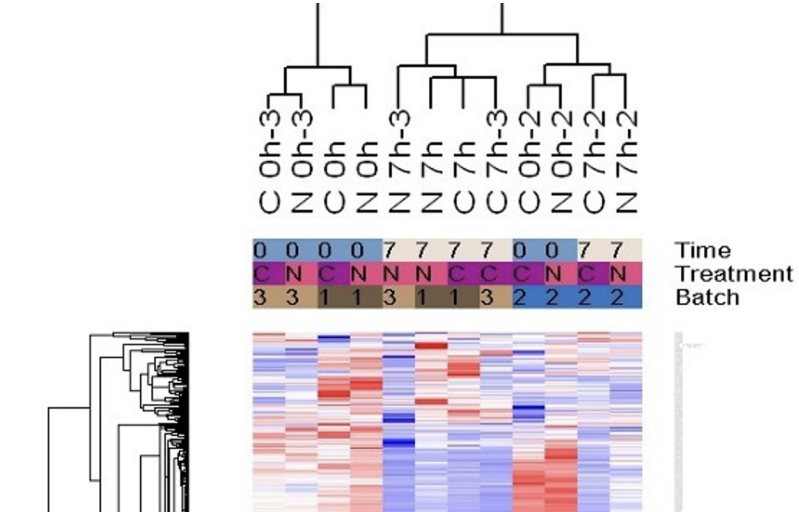
- Gene expression profiling platform
- Lab protocol or experimenter
- Time of day or processing
- Atmospheric ozone level (Rhodes et al. 2004)

Batch Effect Example #1: Nirtic Oxide

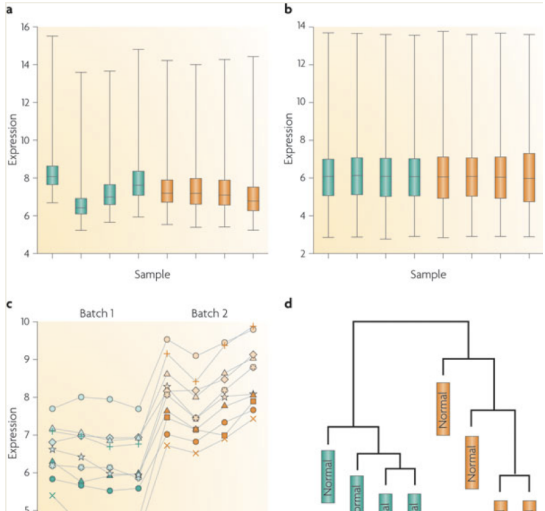
This Example is an oligonucleotide microarray (Affymetrix HG-U133A) experiment on human lung fibroblast cells (IMR90) designed to reveal whether exposing mammalian cells to nitric oxide (NO) stabilizes mRNAs.

Microarray data were collected at baseline (0 h, just before transcription inhibition) and at the end of the experiment (after 7.5 h) for both the control and the NO-treated group.

Batch Effect Example #1: Nirtic Oxide



Batch Effect Example #2: Control Gene Expression



Batch Effect Example #3: Proteomic markers

Proteomic markers to predict endometriosis (39 total):

Single peptide predictors of disease (AUC): 0.82, 0.76, 0.74, 0.74, 0.70 (+12 more >0.6)

Single peptide predictors of batch (AUC): 0.99, 0.94, 0.91, 0.86, 0.86, 0.84, 0.84, 0.84, 0.83, 0.82 (+7 more >0.6)

Predict batch better than disease!

ComBat Batch Adjustment

Consider the following model:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$$

where:

- α_g is the overall gene expression
- X is a design matrix
- β_g contains the regression coefficients
- The error terms $\epsilon_{ijg} \sim N(0, \sigma_g^2)$

ComBat Batch Adjustment

Adjust for batch effects:

$$Y^*_{ijg} = \frac{Y_{ig} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + X\hat{\beta}_g$$

Answer: Empirical Bayes!

BatchQC Example

We can use BatchQC to evaluate and correct for batch effects:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("BatchQC")

library(BatchQC)
BatchQC()
```

Visualization and Dimension reduction

Using an example dataset from: Verma, et al., 2018

```
## read in data
counts <- read.table(
  "rna_seq/downstream_analysis/features_combined.txt",
  sep="\t", header=T, row.names=1)
meta_data <- read.table(
  "rna_seq/downstream_analysis/meta_data.txt",
  sep="\t", header=T, row.names=1)
group <- meta_data$Disease
```

Visualization and dimension reduction

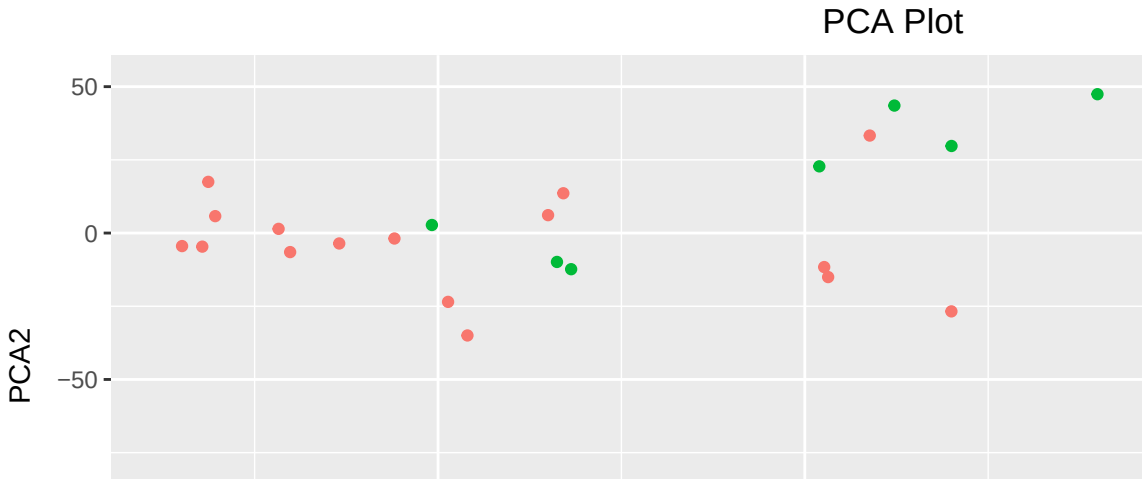
```
## Make SummarizedExperiment
se_hivtb <- SummarizedExperiment(assays=list(counts=counts),
                      colData = meta_data)

## Make log counts, counts per million (cpm), logcpm
se_hivtb <- mkAssay(se_hivtb, log = TRUE,
                      counts_to_CPM = TRUE)
assays(se_hivtb)
```

```
## List of length 4
## names(4): counts log_counts counts_cpm log_counts_cpm
```

Principal Components Analysis (PCA)

```
set.seed(1)
pca_out <- prcomp(t(assay(se_hivtb,"log_counts_cpm")))

pca_plot <- as.data.frame(pca_out$x)
pca_plot$Disease <- as.factor(se_hivtb$Disease)

g <- pca_plot %>% ggplot(aes(x=PC1, y=PC2, color=Disease)) +
  geom_point(size=1.5) + xlab("PCA1") + ylab("PCA2") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("PCA Plot")

plot(g)
```

Principal Components Analysis (PCA)



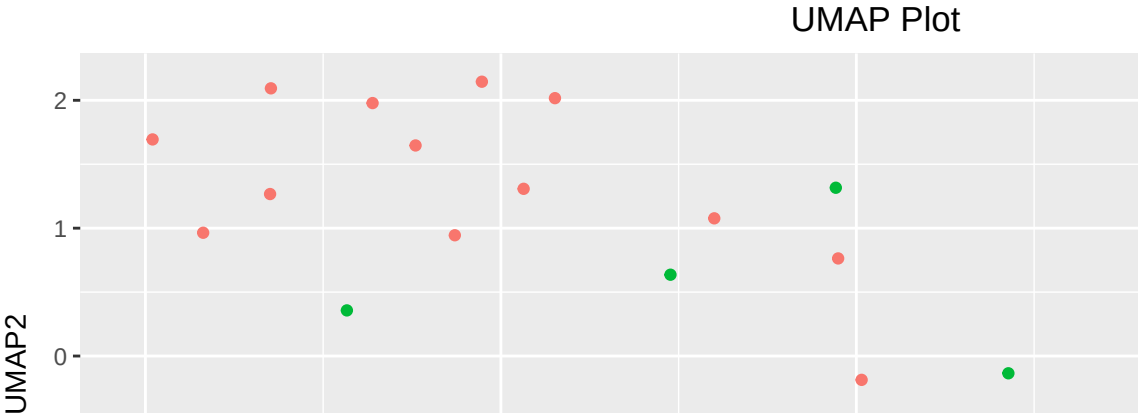
Uniform Manifold Approximation and Projection (UMAP)

For more on UMAP, please visit the following excellent tutorial: https://pair-code.github.io/understanding-umap/

Uniform Manifold Approximation and Projection (UMAP)

```
set.seed(1)
umap_out <- umap(t(assay(se_hivtb,"log_counts_cpm")))

umap_plot <- as.data.frame(umap_out$layout)
umap_plot$Disease <- as.factor(se_hivtb$Disease)

g <- umap_plot %>% ggplot(aes(x=V1, y=V2, color=Disease)) +
  geom_point(size=1.5) + xlab("UMAP1") + ylab("UMAP2") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("UMAP Plot")

plot(g)
```

Uniform Manifold Approximation and Projection (UMAP)



Differential Expression

Table 1: Software packages for detecting differential expression

Method	Version	Reference	Normalization[a]	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors (quantile/TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	RPKM/TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[10]	Geometric (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	*t*-test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially

EdgeR Example

Implements statistical methods for DE analysis based on the negative binomial model:

```
#Gene Filtering
counts<-counts[which(rowSums(counts)>1),]
#Computes library size
dge <- DGEList(counts=counts, group=group)
#TMM normalization
dge <- calcNormFactors(dge)
# Design matrix
design<-model.matrix(~Disease, data=meta_data)
#Estimates common, trended and tagwise dispersion
dge<-estimateDisp(counts,design)
```

EdgeR Example

In negative binomial models, each gene is given a dispersion parameter. Dispersions control the variances of the gene counts and underestimation will lead to false discovery and overestimation may lead to a lower rate of true discovery.

EdgeR Example

```
# Neg Bin GLM with the dispersion estimates
fit<-glmFit(counts,design,
            dispersion=dge$tagwise.dispersion)
# Performs likelihood ratio test
# Compares full versus reduced model
lrt<-glmLRT(fit, coef=2)
```

EdgeR Example

```r
# Prints the top results
topTags(lrt)
```

```
## Coefficient:  Diseasetb_hiv
##              logFC   logCPM        LR        PValue           FDR
## IL1R2     4.334193 8.207857 100.85736 9.885183e-24 2.246012e-19
## AP3B2     5.759328 2.952737  72.84396 1.403164e-17 1.594065e-13
## FCGR1C    2.818495 4.536083  65.19665 6.778459e-16 4.625233e-12
## VNN1      3.150173 8.071712  64.83532 8.142658e-16 4.625233e-12
## CYP1B1    3.135462 6.872932  63.45886 1.637513e-15 7.441186e-12
## IL18R1    2.726129 6.487399  60.92141 5.939939e-15 2.249356e-11
## SLC29A1 -4.135206 3.969652  60.49932 7.360393e-15 2.389078e-11
## CACNG8  -4.134323 3.189711  59.74578 1.079363e-14 3.065527e-11
## SOCS3     2.665234 6.475857  57.88478 2.779298e-14 7.016493e-11
## ZAK       2.601742 6.126677  56.21456 6.497856e-14 1.476378e-10
```

EdgeR Example

```
# Perform quasi-likelihood F-tests
## Replace the chisquare approximation to the likelihood
## ratio statistic with a quasi-likelihood F-test,
## more control of error rate
fit<-glmQLFit(counts, design,
        dispersion=dge$tagwise.dispersion)
## use for small dataset, uncertainty in estimating
## control when number of replicates is small
## dispersion for each gene, more robust and reliable
qlf<-glmQLFTest(fit, coef=2)
```

EdgeR Example

```
# Prints the top results
topTags(qlf)
```

```
## Coefficient:  Diseasetb_hiv
##              logFC   logCPM         F       PValue          FDR
## IL1R2     4.334308 8.207857 106.81157 4.916658e-25 1.117114e-20
## AP3B2     5.770618 2.952737  75.67435 3.352473e-18 3.808577e-14
## VNN1      3.150232 8.071712  68.45477 1.300513e-16 9.849655e-13
## FCGR1C    2.818873 4.536083  66.52246 3.465087e-16 1.968256e-12
## CYP1B1    3.135299 6.872932  65.52904 5.735679e-16 2.606407e-12
## IL18R1    2.726092 6.487399  62.59660 2.540673e-15 9.621105e-12
## SLC29A1 -4.134474 3.969652  62.02406 3.397808e-15 1.037900e-11
## CACNG8  -4.133832 3.189711  61.88068 3.654416e-15 1.037900e-11
## SOCS3     2.665301 6.475857  59.47323 1.241356e-14 3.133873e-11
## ZAK       2.601740 6.126677  57.55635 3.288465e-14 6.798238e-11
```

EdgeR Example

```
#For visualization, heatmaps/PCA
Logcpm<-cpm(counts,log=TRUE)
```

DESeq2 Example

```
#colData is a data frame of demographic/phenotypic data
dds <- DESeqDataSetFromMatrix(countData = counts,
                              colData=meta_data,
                              design=~Disease)
#Gene Filtering
dds<-dds[rowSums(counts(dds))>1,]
```

DESeq2 Example

```
#Performs estimation of size factors,
#dispersion, and negative binomial GLM fitting
dds<-DESeq(dds)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 182 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

DESeq2 Example

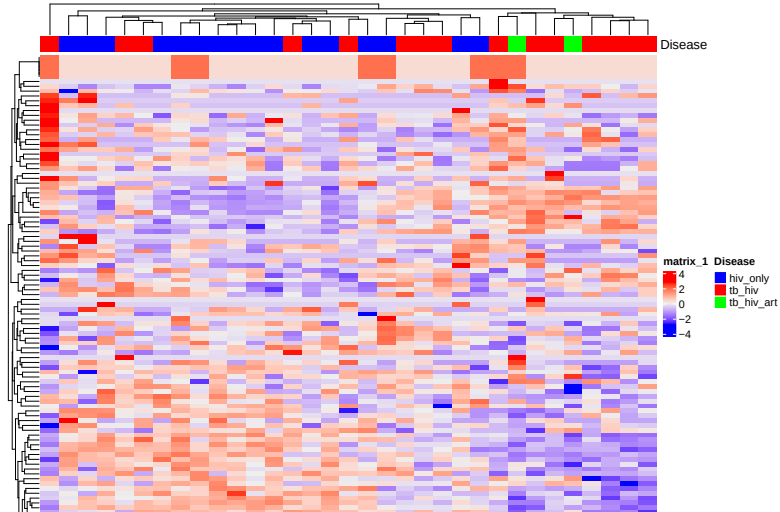
```
res <- results(dds)[order(results(dds)[,6]),]
res[1:10,]
```

```
## log2 fold change (MLE): Disease tb hiv art vs hiv only
## Wald test p-value: Disease tb hiv art vs hiv only
## DataFrame with 10 rows and 6 columns
##          baseMean log2FoldChange     lfcSE      stat      pvalue        padj
##         <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
## ZEB2    1702.8164        1.75542  0.217004   8.08937 5.99766e-16 9.66403e-12
## DCUN1D3   48.7029        2.52497  0.361809   6.97874 2.97846e-12 2.39960e-08
## TIPARP   724.9429        2.22978  0.333470   6.68658 2.28449e-11 1.22700e-07
## ITPKC    196.8723        3.04144  0.465468   6.53416 6.39684e-11 2.57681e-07
## LAIR1   1528.6852        3.01391  0.471741   6.38892 1.67061e-10 3.77242e-07
## COPS4    559.7527        2.67612  0.419494   6.37940 1.77785e-10 3.77242e-07
## IGF2BP3  360.1066        3.29761  0.517563   6.37141 1.87298e-10 3.77242e-07
## GSAP     964.2702        1.26505  0.198550   6.37144 1.87256e-10 3.77242e-07
## RBMS1   3297.7791        2.51612  0.399893   6.29197 3.13466e-10 5.61209e-07
## ZDHHC20  554.9810        2.51988  0.403043   6.25214 4.04856e-10 6.52344e-07
```

Heatmap of DEGs

```
# Make a Heatmap of DEGs
mat = as.matrix(assay(se_hivtb,"log_counts_cpm")
                )[order(results(dds)[,6])[1:100],]
                # Using first 1000 genes to simplify
mat = t(scale(t(mat)))
df=data.frame(Disease=colData(se_hivtb)$Disease)
ha = HeatmapAnnotation(df = df,
                       col = list(Disease=c(
                         "tb_hiv"="Red",
                         "hiv_only"="Blue",
                         "tb_hiv_art"="Green")))
Heatmap(mat,show_row_names=F, show_column_names = F,
        top_annotation = ha)
```

Heatmap of DEGs



Limma Example

- Most similar to microarray data flow
- Reads counts are converted to log2 counts per million (logCPM) and the mean-variance relationship is modeled with precision weights (voom transform)

Limma Example

```
#From edgeR, Computes library size
dge <- DGEList(counts=counts, group=group)
#Gene Filtering
counts<-counts[which(rowSums(cpm(counts))>1),]
dge <- calcNormFactors(dge) #TMM normalization
```

Limma Example

```
design<-model.matrix(~group)
#voom transform to calculate weights to
#eliminate mean-variance relationship
v<-voom(dge, design)
#use usual limma pipelines
fit<-lmFit(v,design)
fit<-eBayes(fit)
```

Limma Example

```
topTable(fit, coef=ncol(design))
```

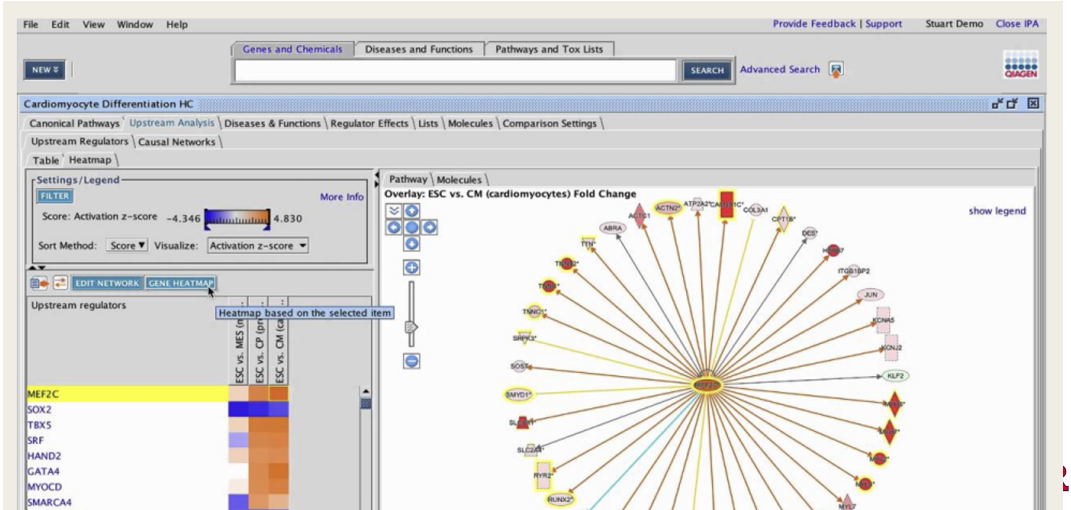
```
##                logFC    AveExpr        t      P.Value    adj.P.Val        B
## LINC01093  5.339162  0.8709573 7.319084 1.086635e-08 7.509193e-05 9.727414
## FAM151B    3.583038  1.2893582 7.254113 1.323872e-08 7.509193e-05 9.575599
## ZEB2       1.558822  6.4207702 7.181311 1.652479e-08 7.509193e-05 9.472646
## CYP19A1    6.733757 -0.9440469 7.284435 1.207258e-08 7.509193e-05 9.456297
## DCUN1D3    2.286360  1.1212630 7.186582 1.626140e-08 7.509193e-05 9.258885
## C7orf61    2.481334  2.7347171 6.989601 2.969015e-08 1.124317e-04 8.895030
## IGF2BP3    3.219039  3.6438771 6.702223 7.183335e-08 2.040157e-04 8.072570
## TIPARP     1.995087  5.0815463 6.646489 8.531423e-08 2.153805e-04 7.915650
## COL4A2-AS1 5.267064 -3.3142663 6.844693 4.632151e-08 1.503530e-04 7.461298
## ZDHHC20    2.329418  4.6107854 6.314821 2.382575e-07 4.556152e-04 6.939199
```

Pathway analysis

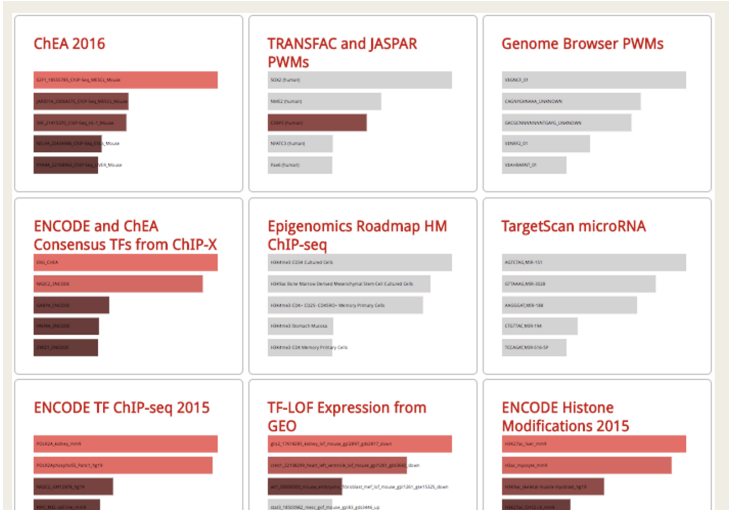
After finding DEGs, look for correlated genes/networks and enriched pathway sets in the gene set using:

- Weighted gene coexpression network analysis (WGCNA)
- GSEA, GSVA, EnrichR, many more!!
- Qiagen Ingenuity Pathway Analysis (IPA)

Pathway analysis



Pathway analysis



TBSignatureProfiler Analysis

The TBSignatureProfiler was developed in the Johnson Lab in 2021 to profile new and existing TB gene expression signatures:

https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-020-05598-z

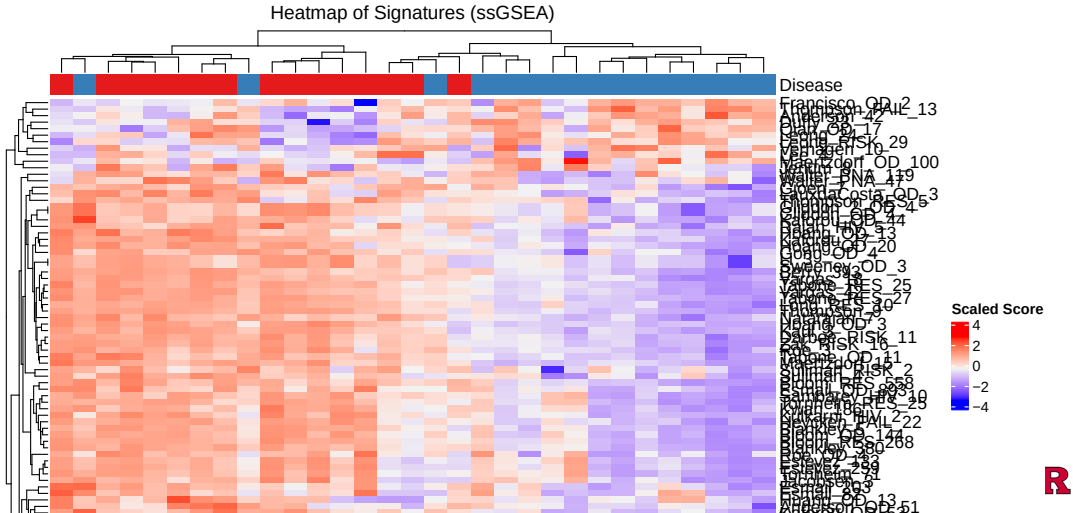
TBSignatureProfiler Analysis

```
se_hivtb_2 <- se_hivtb[,
      colData(se_hivtb)$Disease != "tb_hiv_art"]
TBsigs <- TBsignatures[-12]
ssgsea_res <- runTBsigProfiler(se_hivtb_2,
                  useAssay = "log_counts_cpm",
                  signatures = TBsigs,
                  algorithm = "ssGSEA",
                  combineSigAndAlgorithm = TRUE,
                  parallel.sz = 1)
```

Signature Heatmap:

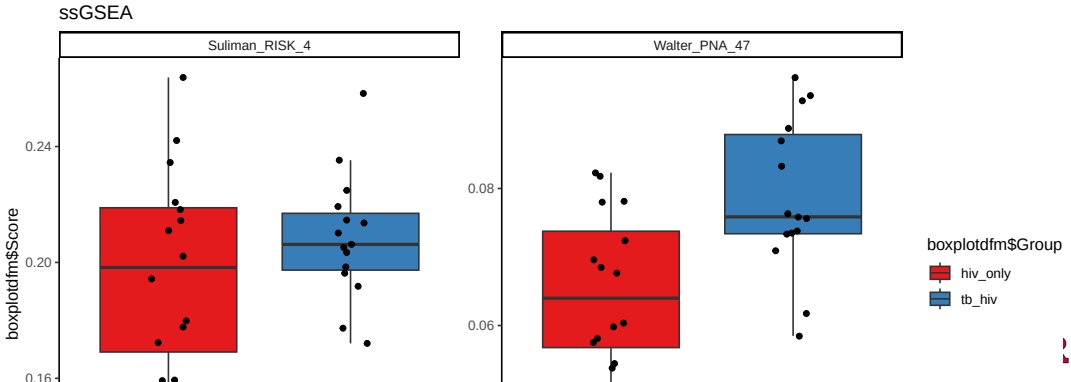
```
# Colors for gradient
signatureHeatmap(ssgsea_res,
        name = "Heatmap of Signatures (ssGSEA)",
        signatureColNames = names(TBsigs),
        annotationColNames = c("Disease"),
        scale = TRUE,
        split_heatmap = "none",
        showColumnNames = FALSE)
```

Signature Heatmap:



Signature Boxplots

```
signatureBoxplot(ssgsea_res, name="ssGSEA",
          signatureColNames = names(TBsigs)[c(62,77)],
          annotationColName = c("Disease"))
```



Session info

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.4.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] grid      stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] umap_0.2.10.0               TBSignatureProfiler_1.17.2
##  [3] ComplexHeatmap_2.20.0       DESeq2_1.44.0
##  [5] edgeR_4.2.2                 limma_3.60.6
##  [7] SummarizedExperiment_1.34.0 Biobase_2.64.0
##  [9] MatrixGenerics_1.16.0       matrixStats_1.5.0
## [11] Rsamtools_2.20.0            Biostrings_2.72.1
## [13] XVector_0.44.0              GenomicRanges_1.56.2
```