Icahn School of Medicine at Mount Sinai

Department of Pharmacology and System Therapeutics

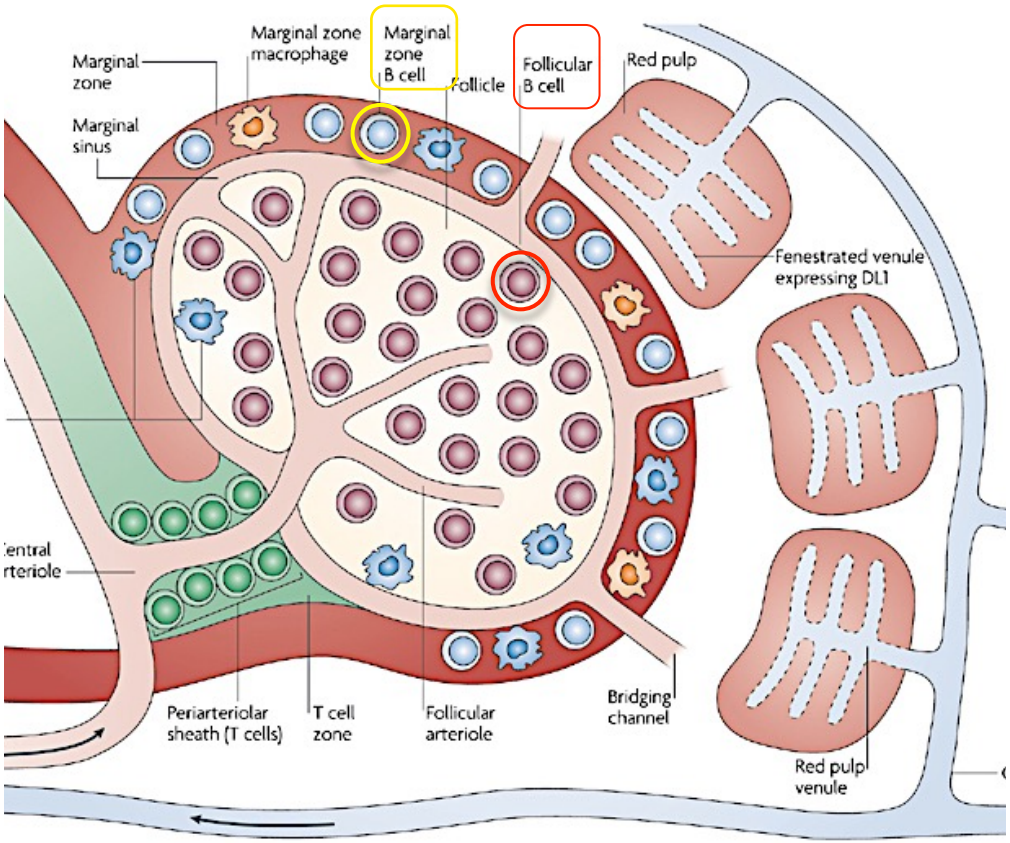
RNA-Seq Analysis

From FASTQ to Differential Expression

Caroline D. Monteiro

Ma'ayan Lab

Background

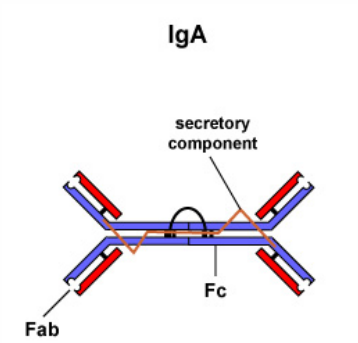


	Follicular B cells (FOB)	Marginal Zone B cells (MZB)
Type	Mature B cells	Mature B cells
Circulate	Freely	Noncirculating
Location	Primary follicles of B cell zones, in the white pulp of the	Interface between the non-lymphoid red pulp and the lymphoid white-pulp.
Express high levels of:	IgM	IgM
	IgD	CD21
	CD23	CD1
		CD9

The follicular versus marginal zone B lymphocyte cell fate decision
Shiv Pillai1 & Annaiah Cariappa1

Background

Follicular B cells (FOB)	Marginal Zone B cells (MZB)
Wild Type	Wild Type
IgA Knockout	IgA Knockout

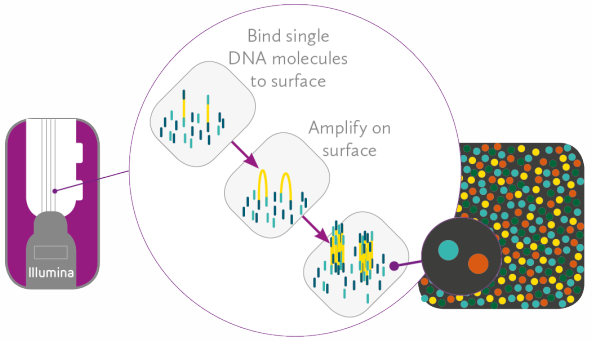


Mucous Membranes
Increase Infections

Single-end sequencing

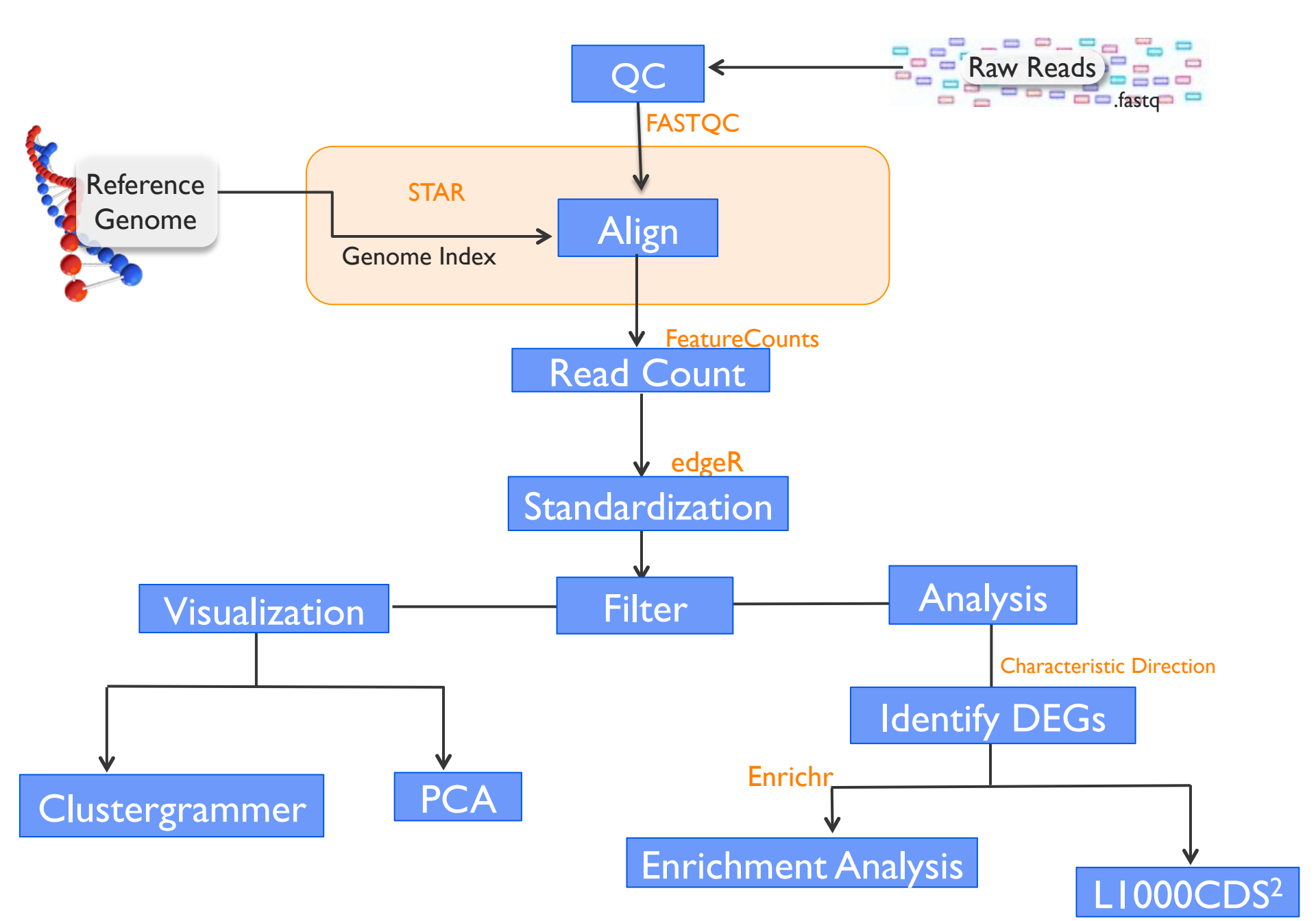
illumina HiSeq 2500

19 samples
3 replicas



Reference Genome :

UCSC mm9 - Mus musculus


QC
Raw Reads
.fastq
FASTQC
Reference Genome
STAR
Genome Index
Align
FeatureCounts
Read Count
edgeR
Standardization
Visualization
Filter
Analysis
Characteristic Direction
Identify DEGs
Clustergrammer
PCA
Enrichr
Enrichment Analysis
L1000CDS²

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

Installation

Star Aligner

I) Get latest STAR source from releases

```
$ wget https://github.com/alexdobin/STAR/archive/STAR_2.4.2a.tar.gz
$ tar -xzf STAR_2.4.2a.tar.gz
$ cd STAR-STAR_2.4.2a
```

```
# or get STAR source using git
$ git clone https://github.com/alexdobin/STAR.git
$ cd STAR-STAR_2.4.2a
```

II) Change file permission

```
$ sudo chmod  755 STAR
```

III) Build STAR

```
$ cd source
$ make STAR
```

FeatureCounts

I) Download the Subread source package

```
$ tar zxvf subread-1.x.x.tar.gz
```

II) Change permissions:

```
$ sudo chmod  755
```

III) Build

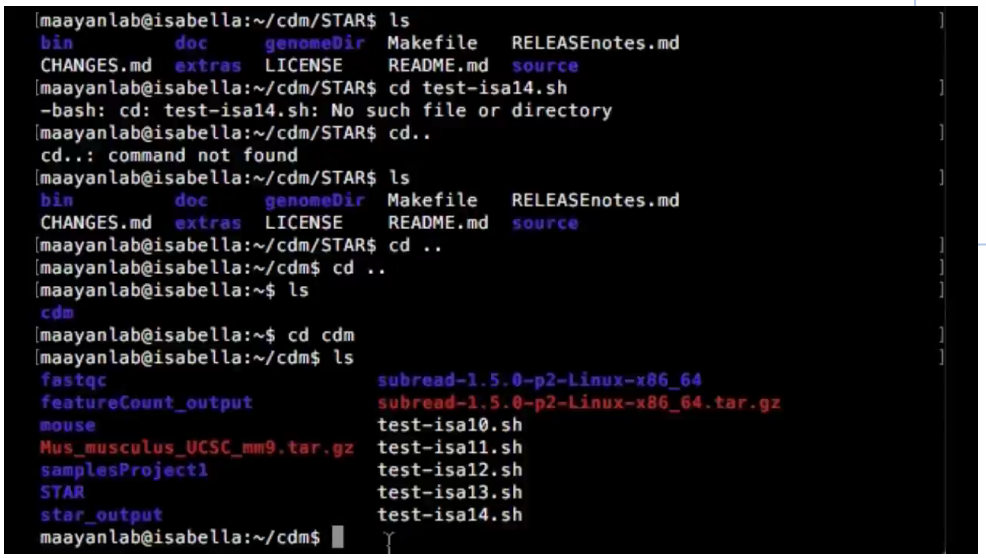
```
$ cd src
$ make -f Makefile.Linux
```

* 755 is the directory permission

Alignment

Create Genome Index

```
STAR \
    --runThreadN 16 \
    --runMode genomeGenerate \
    --genomeDir $STAR_INDEX \
    --genomeFastaFiles $GENOME_FA \
    --sjdbGTFfile $GENOME_GTF \
    --sjdbOverhang 100
```



Star Run

```
STAR \
    --genomeDir $STAR_INDEX \
    --sjdbGTFfile $GENOME_GTF \
    --runThreadN 16 \
    --outSAMstrandField intronMotif \
    --outFilterIntronMotifs RemoveNoncanonical \
    --outFileNamePrefix $WORDIR/star_output/$basename\
    --readFilesIn $fq1,$fq2,$fq3 \
    --outSAMtype BAM SortedByCoordinate \
    --outReadsUnmapped Fastx \
    --outSAMmode Full
```

Star output

Aligned.sortedByCoord.out.bam: output sorted by coordinate file, similar to samtools sort command.

Log.out: main log file with a lot of detailed information about the run.

Log.progress.out: reports job progress statistics (number of processed reads, % of mapped reads,etc). Updated in 1 min intervals.

Log.final.out: summary mapping statistics after mapping job is complete.

Unmapped.out.mate1: unmapped reads.

SJ.out.tab: contains high confidence collapsed splice junctions in txt.

```
maayanlab@isabella:~/cdm$ cd star_output
maayanlab@isabella:~/cdm/star_output$ ls
11-1163-Wt-P8Aligned.sortedByCoord.out.bam
11-1163-Wt-P8Log.final.out
11-1163-Wt-P8Log.out
11-1163-Wt-P8Log.progress.out
11-1163-Wt-P8SJ.out.tab
11-1163-Wt-P8_STARgenome
11-1163-Wt-P8Unmapped.out.mate1
12-1163-Wt-P9Aligned.sortedByCoord.out.bam
12-1163-Wt-P9Log.final.out
12-1163-Wt-P9Log.out
12-1163-Wt-P9Log.progress.out
12-1163-Wt-P9SJ.out.tab
12-1163-Wt-P9_STARgenome
12-1163-Wt-P9_STARtmp
12-1163-Wt-P9Unmapped.out.mate1
16-1164-IgAKO-P8Aligned.sortedByCoord.out.bam
16-1164-IgAKO-P8Log.final.out
16-1164-IgAKO-P8Log.out
16-1164-IgAKO-P8Log.progress.out
16-1164-IgAKO-P8SJ.out.tab
16-1164-IgAKO-P8_STARgenome
16-1164-IgAKO-P8Unmapped.out.mate1
```

QC

```
$ cat 16-1164-IgAKO-P8Log.final.out
```

```
                                 Started job on |   May 02 21:45:33
                             Started mapping on |   May 02 21:46:59
                                    Finished on |   May 02 21:49:53
        Mapping speed, Million of reads per hour |   758.84

                           Number of input reads |   36677079
                       Average input read length |   51
                                     UNIQUE READS:
          Uniquely mapped reads number |   25860357
                Uniquely mapped reads % |   70.51%
                       Average mapped length |   50.76
                    Number of splices: Total |   4748840
         Number of splices: Annotated (sjdb) |   4644685
                    Number of splices: GT/AG |   4711313
                    Number of splices: GC/AG |   32340
                    Number of splices: AT/AC |   5187
                Number of splices: Non-canonical |   0
                    Mismatch rate per base, % |   0.19%
                       Deletion rate per base |   0.00%
                      Deletion average length |   1.89
                      Insertion rate per base |   0.00%
                     Insertion average length |   1.16
                           MULTI-MAPPING READS:
     Number of reads mapped to multiple loci |   9687601
          % of reads mapped to multiple loci |   26.41%
     Number of reads mapped to too many loci |   615328
          % of reads mapped to too many loci |   1.68%
                               UNMAPPED READS:
    % of reads unmapped: too many mismatches |   0.00%
              % of reads unmapped: too short |   1.21%
                  % of reads unmapped: other |   0.19%
                               CHIMERIC READS:
                    Number of chimeric reads |   0
                         % of chimeric reads |   0.00%
```

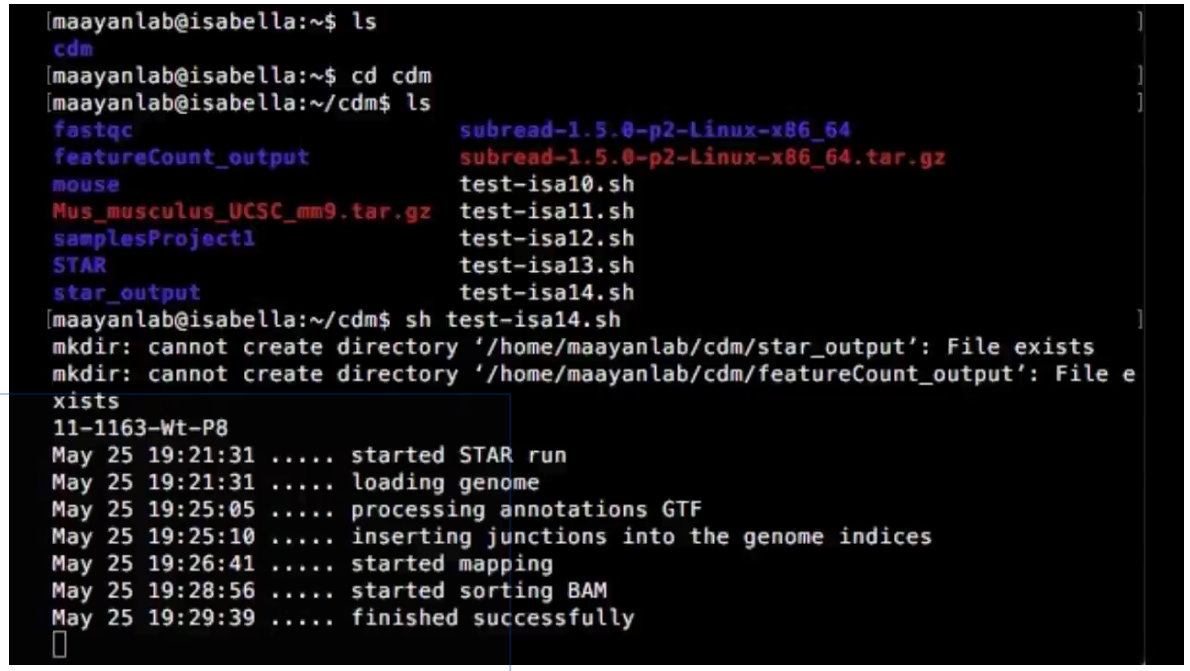
```
cat 11-1163-Wt-P8Log.final.out
```

```
                                 Started job on |   May 25 19:38:04
                             Started mapping on |   May 25 19:39:49
                                    Finished on |   May 25 19:42:27
        Mapping speed, Million of reads per hour |   804.81

                           Number of input reads |   35322160
                       Average input read length |   51
                                     UNIQUE READS:
                Uniquely mapped reads number |   24946248
                      Uniquely mapped reads % |   70.62%
                       Average mapped length |   50.76
                    Number of splices: Total |   4741855
         Number of splices: Annotated (sjdb) |   4637043
                    Number of splices: GT/AG |   4703421
                    Number of splices: GC/AG |   32893
                    Number of splices: AT/AC |   5541
                Number of splices: Non-canonical |   0
                    Mismatch rate per base, % |   0.18%
                       Deletion rate per base |   0.00%
                      Deletion average length |   1.54
                      Insertion rate per base |   0.00%
                     Insertion average length |   1.18
                           MULTI-MAPPING READS:
     Number of reads mapped to multiple loci |   9374842
          % of reads mapped to multiple loci |   26.54%
     Number of reads mapped to too many loci |   628579
          % of reads mapped to too many loci |   1.78%
                               UNMAPPED READS:
    % of reads unmapped: too many mismatches |   0.00%
              % of reads unmapped: too short |   0.89%
                  % of reads unmapped: other |   0.16%
                               CHIMERIC READS:
                    Number of chimeric reads |   0
                         % of chimeric reads |   0.00%
```

Read Counts

```
featureCounts \
        -T 16 \
        -t exon \
        -g gene_id \
        -a $GENOME_GTF \
        -o $WORDIR/featureCount_output/$outname \
        $bam
```



QC

```
## Check the featureCount_output summary files for the alignment stats.
## This will output the first 10 lines of all summary files from the featureCounts folder
!head ../featureCount_output/*.summary
```

```
maayanlab@isabella:~/cdm/featureCount_output$ head /home/maayanlab/cdm/featureCount_output/*.summary,<br>
==> /home/maayanlab/cdm/featureCount_output/11-1163-Wt-P8.count.txt.summary <==<br>
Status  /home/maayanlab/cdm/star_output/11-1163-Wt-P8Aligned.sortedByCoord.out.bam<br>
Assigned        21801137<br>
Unassigned_Ambiguity    341618<br>
Unassigned_MultiMapping 31929676<br>
Unassigned_NoFeatures   2803493<br>
Unassigned_Unmapped     0<br>
Unassigned_MappingQuality       0<br>
Unassigned_FragmentLength       0<br>
Unassigned_Chimera      0<br>
Unassigned_Secondary    0<br>

==> /home/maayanlab/cdm/featureCount_output/12-1163-Wt-P9.count.txt.summary <==<br>
Status  /home/maayanlab/cdm/star_output/12-1163-Wt-P9Aligned.sortedByCoord.out.bam<br>
Assigned        23105766<br>
Unassigned_Ambiguity    363702<br>
Unassigned_MultiMapping 26166037<br>
Unassigned_NoFeatures   3506987<br>
Unassigned_Unmapped     0<br>
Unassigned_MappingQuality       0<br>
Unassigned_FragmentLength       0<br>
Unassigned_Chimera      0<br>
Unassigned_Secondary    0<br>

==> /home/maayanlab/cdm/featureCount_output/16-1164-IgAKO-P8.count.txt.summary <==<br>
Status  /home/maayanlab/cdm/star_output/16-1164-IgAKO-P8Aligned.sortedByCoord.out.bam<br>
Assigned        21976115<br>
Unassigned_Ambiguity    331259<br>
Unassigned_MultiMapping 32706166<br>
Unassigned_NoFeatures   3552983<br>
Unassigned_Unmapped     0<br>
Unassigned_MappingQuality       0<br>
Unassigned_FragmentLength       0<br>
Unassigned_Chimera      0<br>
Unassigned_Secondary    0<br>
```

Full file

Standardization

Counts Per Million (CPM)

```
setwd('/../')
library(edgeR)

fns <- system2('ls', args = '~/featureCount_output_no_outliers/*.txt', stdout = T)
counts.df <- NULL
lengths.df <- NULL
for (fn in fns) {


  df <- read.table(fn, check.names=F, sep='\t', header=T)
  df <- subset(df, select=-c(Chr, Start, End, Strand))
  colnames(df)[3] <- strsplit(basename(fn), '\\.')[[1]][1]
  if (is.null(counts.df)) {
    lengths.df <- subset(df, select=c(Geneid, Length))
    df <- subset(df, select=-c(Length))
    counts.df <- df
  } else {
    df <- subset(df, select=-c(Length))
    counts.df <- merge(counts.df, df, by="Geneid", all.x=T, sort=F)
  }
}

write.csv(counts.df, file='featureCounts_matrix.csv', row.names=F)
write.csv(lengths.df, file='featureCounts_gene_lengths.csv', row.names=F)

genes <- as.vector(counts.df[[1]])
gene.lengths <- lengths.df[[2]]
# make a matrix of counts
count.mat <- as.matrix(counts.df[2:length(counts.df)])

## Calculate RPKM
d <- DGEList(counts=count.mat)
d$genes$Length <- gene.lengths
rpkm.mat <- rpkm(d)
rownames(rpkm.mat) <- genes
write.csv(rpkm.mat, file="repRpkmMatrix_featureCounts.csv", row.names=T)

## Calculate CPM
cpm.mat <- cpm(d)
rownames(cpm.mat) <- genes
write.csv(cpm.mat, file="repCpmMatrix_featureCounts.csv", row.names=T)
```

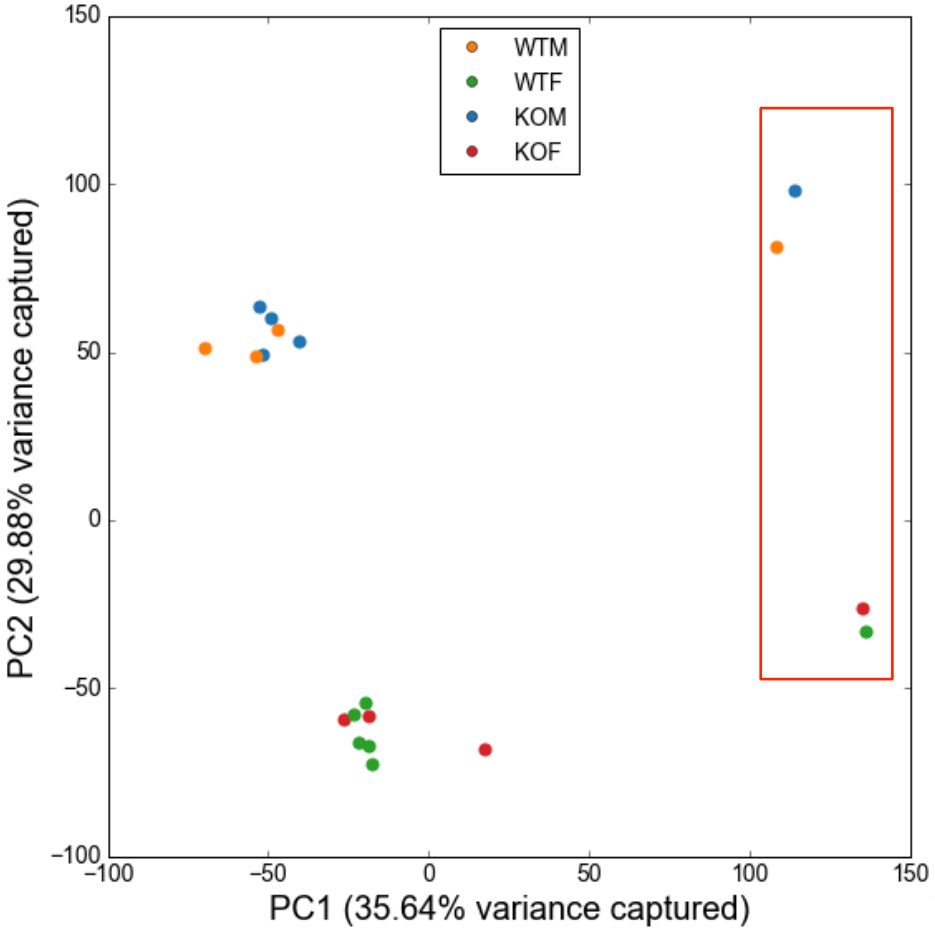
Filtering

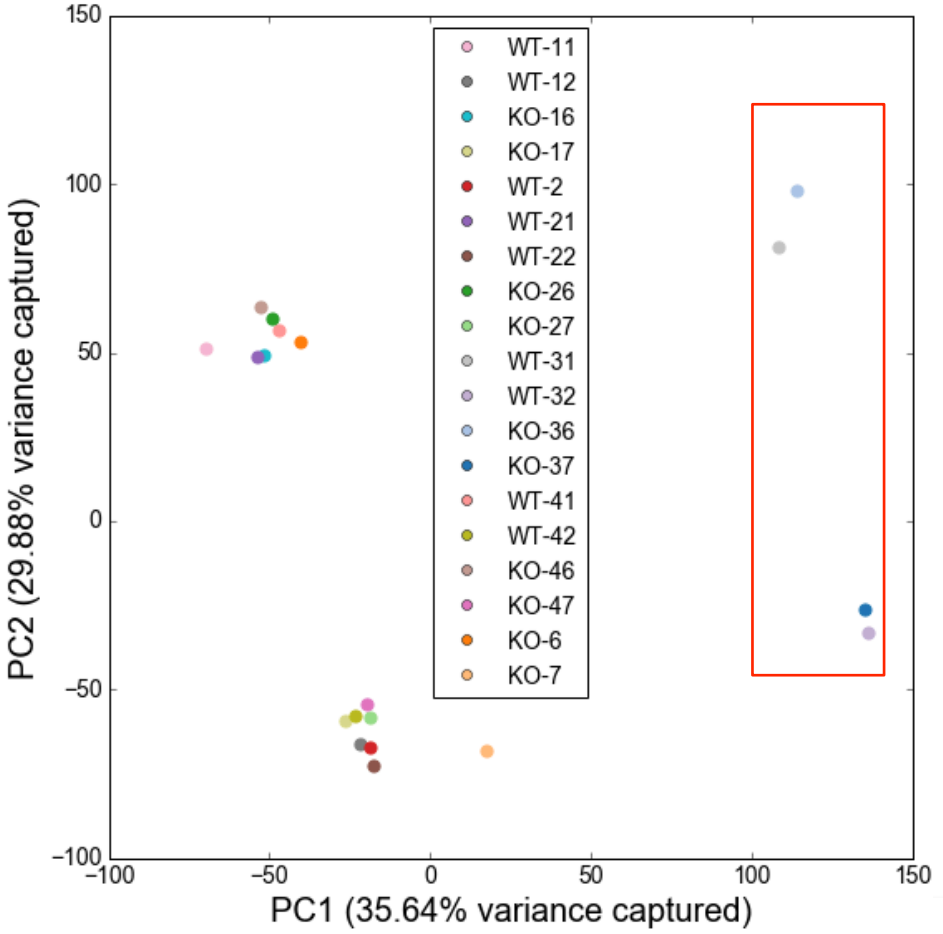
```
## Filter out non-expressed genes
expr_df = expr_df.loc[expr_df.sum(axis=1) > 0, :]
print (expr_df.shape)

## Filter out lowly expressed genes
mask_low_vals = (expr_df > 0.3).sum(axis=1) > 2
expr_df = expr_df.loc[mask_low_vals, :]

print (expr_df.shape)
```

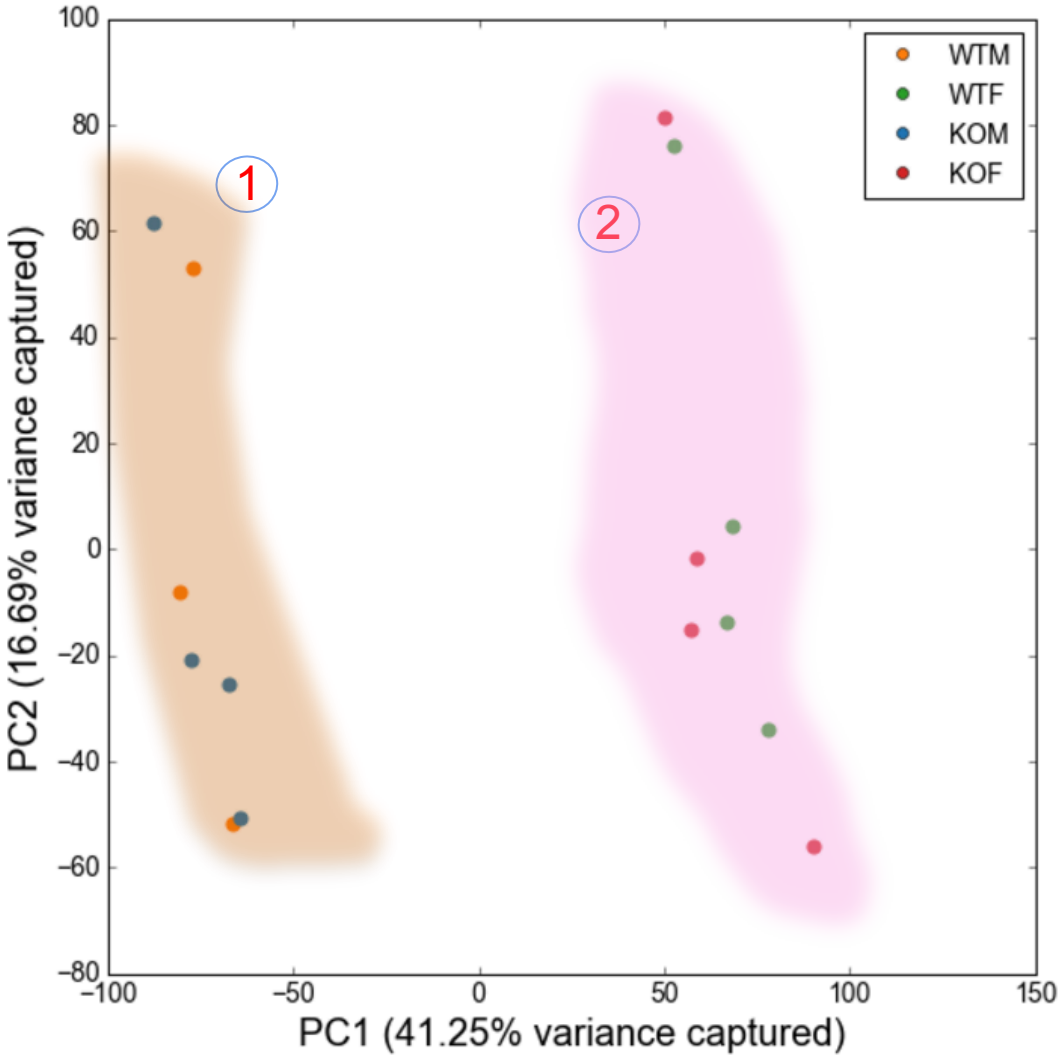
PCA





* Outliers

Removed from the analysis.



Outliers removed

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

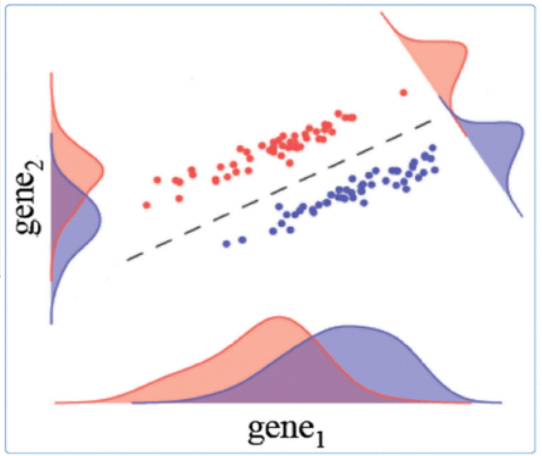
Differentially expressed genes

```
import geode
cd_results = pd.DataFrame(index=expr_df.index)

expr_df=expr_df.rename(columns = {'Wt-P8':1,'Wt-P9':1,'IgAKO-P8':2, 'IgAKO-P9':2})
```

Up List	
Sell	0.013721303141
Otud4	0.016151196404
Lyl1	0.012064258106
Csde1	0.016992975498
Setd5	0.010809564506
Med13l	0.014161874538
Smek2	0.010458106765
AW549877	0.010936138289
Ccr7	0.017973715650
Hnrnpu	0.018505879222
Mfhas1	0.015914145478
Zdhhc20	0.011254588028
Entpd4	0.013707767396
Zfp318	0.011803867763
Prdm2	0.016584782897
Nufip2	0.012044621426
Actb	0.058448714858
2410006H16Rik	0.018414134452
Tgif1	0.014321462854
Ubr4	0.031516174213
Herc1	0.010333135017
Mga	0.018168731984
Pnrc1	0.024273572128
Tob2	0.026627274383
Cdk12	0.011326196577

Down List	
Sorl1	-0.044350606207
Capzb	-0.015111540429
Cyp4f18	-0.025986026582
Apoe	-0.015506247242
Chordc1	-0.010971547875
Lgmn	-0.020847414502
Cpne1	-0.013714579099
Psap	-0.099490900566
Lynx1	-0.013201990124
Sdcbp	-0.015016617933
Ctnnb1	-0.024503311580
Arhgef6	-0.016431547892
Fcrl5	-0.015774198311
Tmbim6	-0.022416944511
Pik3ap1	-0.013740767407
Pld4	-0.010343932979
Ptprcap	-0.043940651031
Cdk19	-0.011675684079
Dph5	-0.011116727858
Tmod3	-0.027223587210
Anxa6	-0.026809787983
Sun2	-0.027184934001
Ctsa	-0.013445641261
Rnf187	-0.014716619658



https://github.com/wangz10/geode.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

REFERENCES

RNA seq analysis

An Open RNA-Seq Data Analysis Pipeline Tutorial with an Example of Reprocessing Data from a Recent Zika Virus Study
Zichen Wang, Avi Ma'ayan

A survey of best practices for RNA-seq data analysis
Conesa A, et al

Star

Star Manual
http://labshare.cshl.edu/shares/gingeraslab/www-data/dobin/STAR/STAR.posix/doc/STARmanual.pdf

FeatureCounts

http://bioinf.wehi.edu.au/featureCounts

Lab Tools & Methods

Characteristic Direction

Clustergrammer

Enrichr

L1000CDS^2


THANK YOU

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.