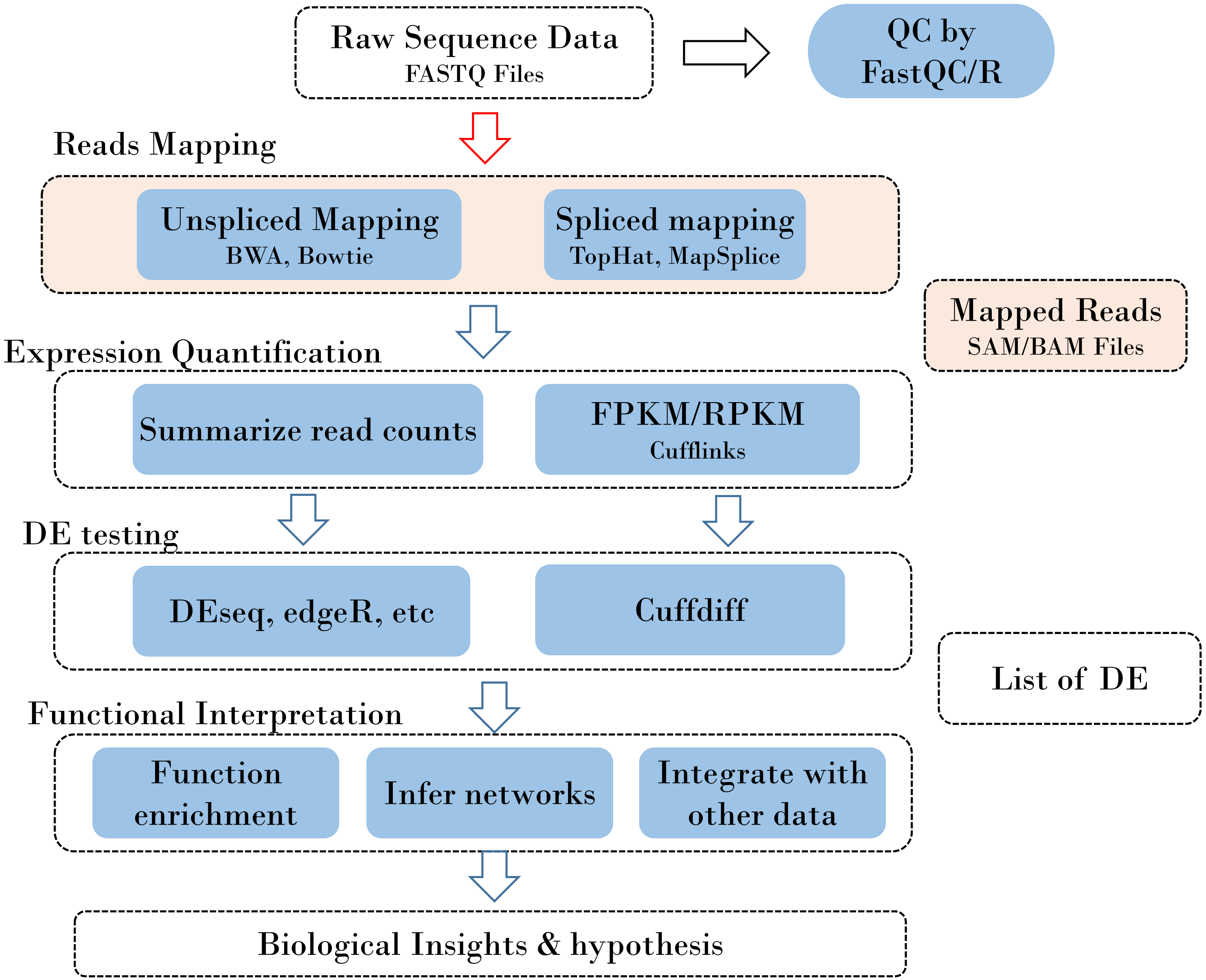
ANALYSIS OF RAW DATSETS AND DIFFERENTIAL EXPRESSION

Presented by – Shikha Roy

Senior Research Fellow, ICGEB

From reads to differential expression



Raw Sequence Data
FASTQ Files

QC by FastQC/R

Reads Mapping

Unspliced Mapping
BWA, Bowtie

Spliced mapping
TopHat, MapSplice

Mapped Reads
SAM/BAM Files

Expression Quantification

Summarize read counts

FPKM/RPKM
Cufflinks

DE testing

DEseq, edgeR, etc

Cuffdiff

List of DE

Functional Interpretation

Function enrichment

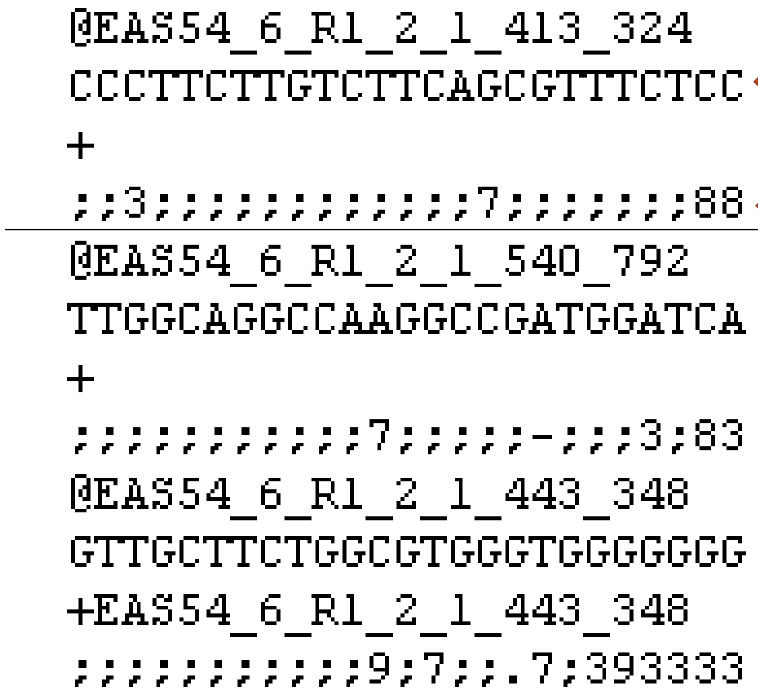
Infer networks

Integrate with other data

Biological Insights & hypothesis

FASTQ format

The FASTQ format stores DNA sequence data as well as associated Phred quality scores of each base.



```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

DNA read

Base quality score

FASTQC

- FastQC is a quality control application that allows users to perform numerous quality control checks on raw sequence data generated by high throughput sequencing pipelines such as Illumina and ABI SOLiD platforms in FASTQ format.

- It generates as output a comprehensive multi-page report on the composition and quality of reads in HTML format, with one page for each of the reads (e.g. Single End, Paired End: forward, Paired End: reverse). The modules included in the report are as follows:

➢ Per Base Sequence Quality

➢ Per Base Sequence Content

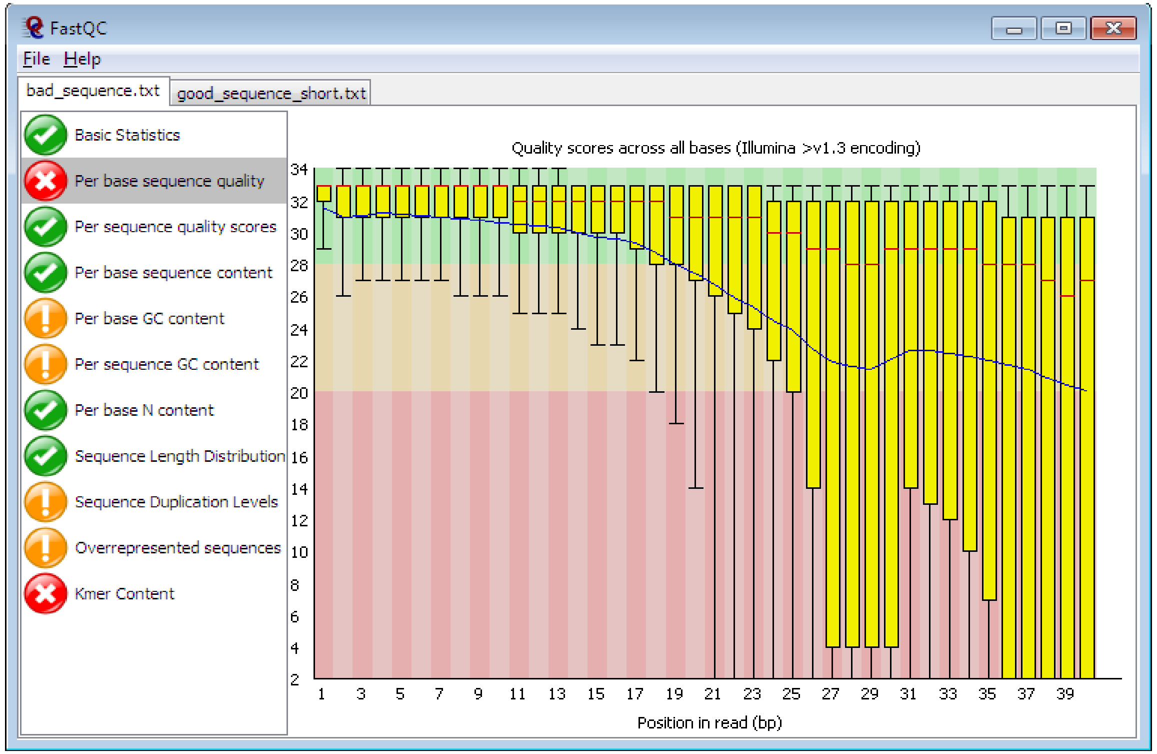
➢ Per Sequence GC content

➢ Per Base N Content

➢ Sequence Length Distribution

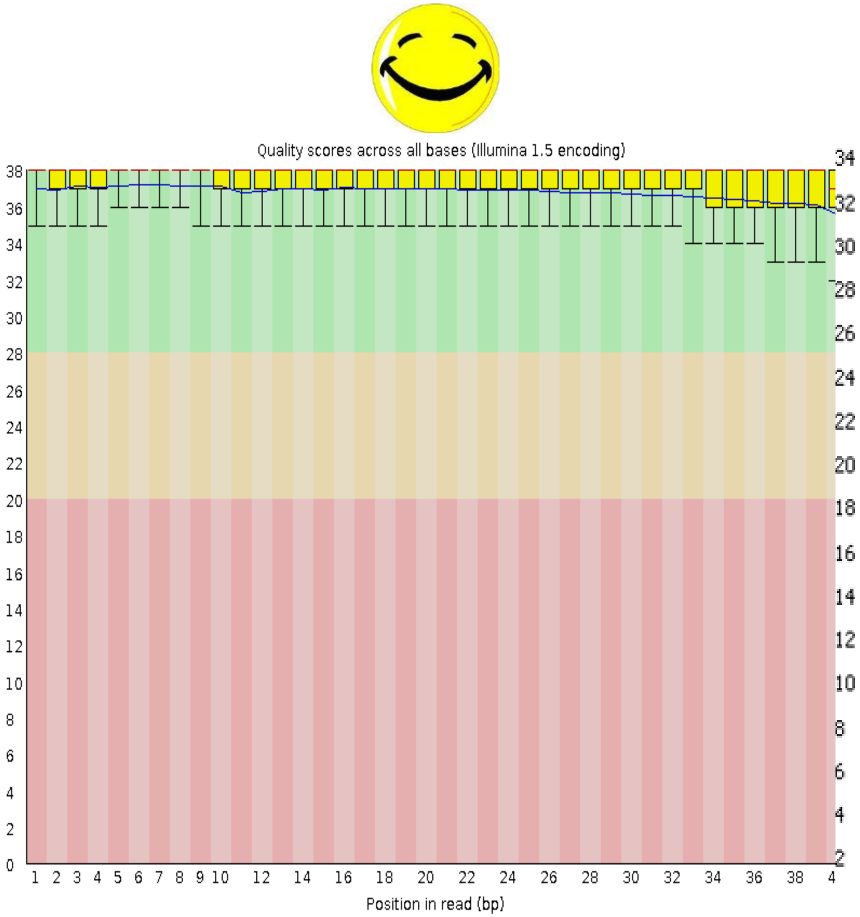
➢ Sequence Duplication Levels

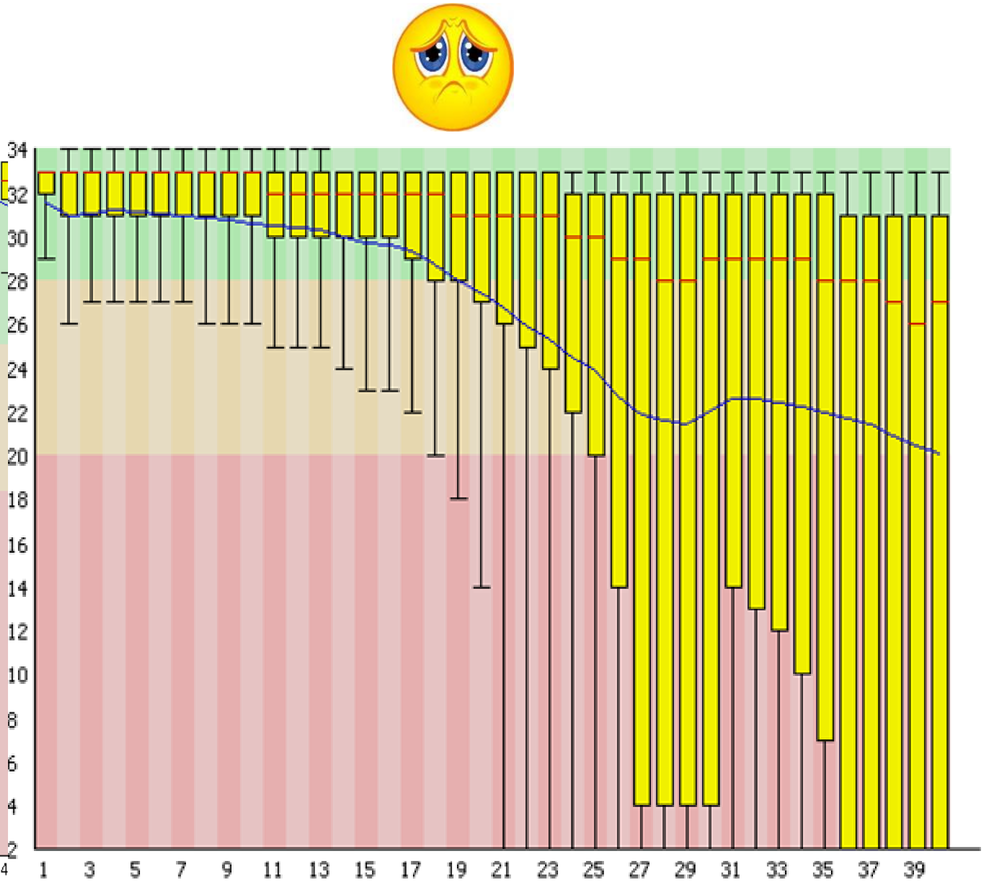
➢ Adapter Content



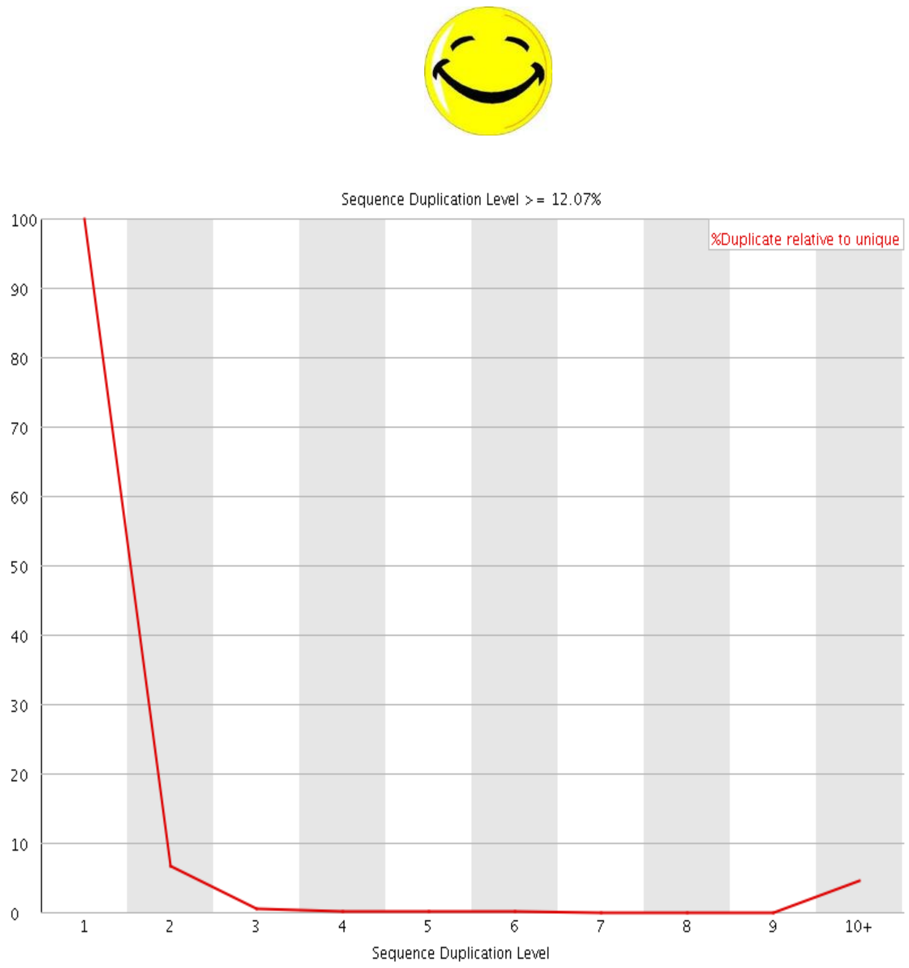
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content

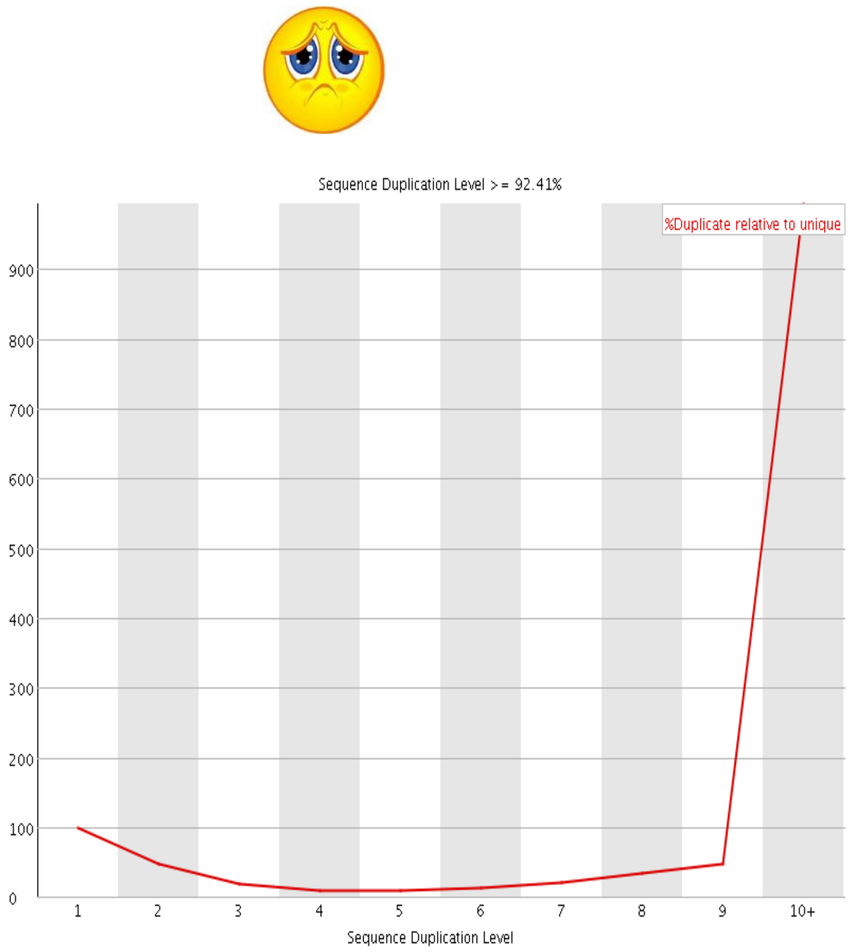
Per base sequence quality



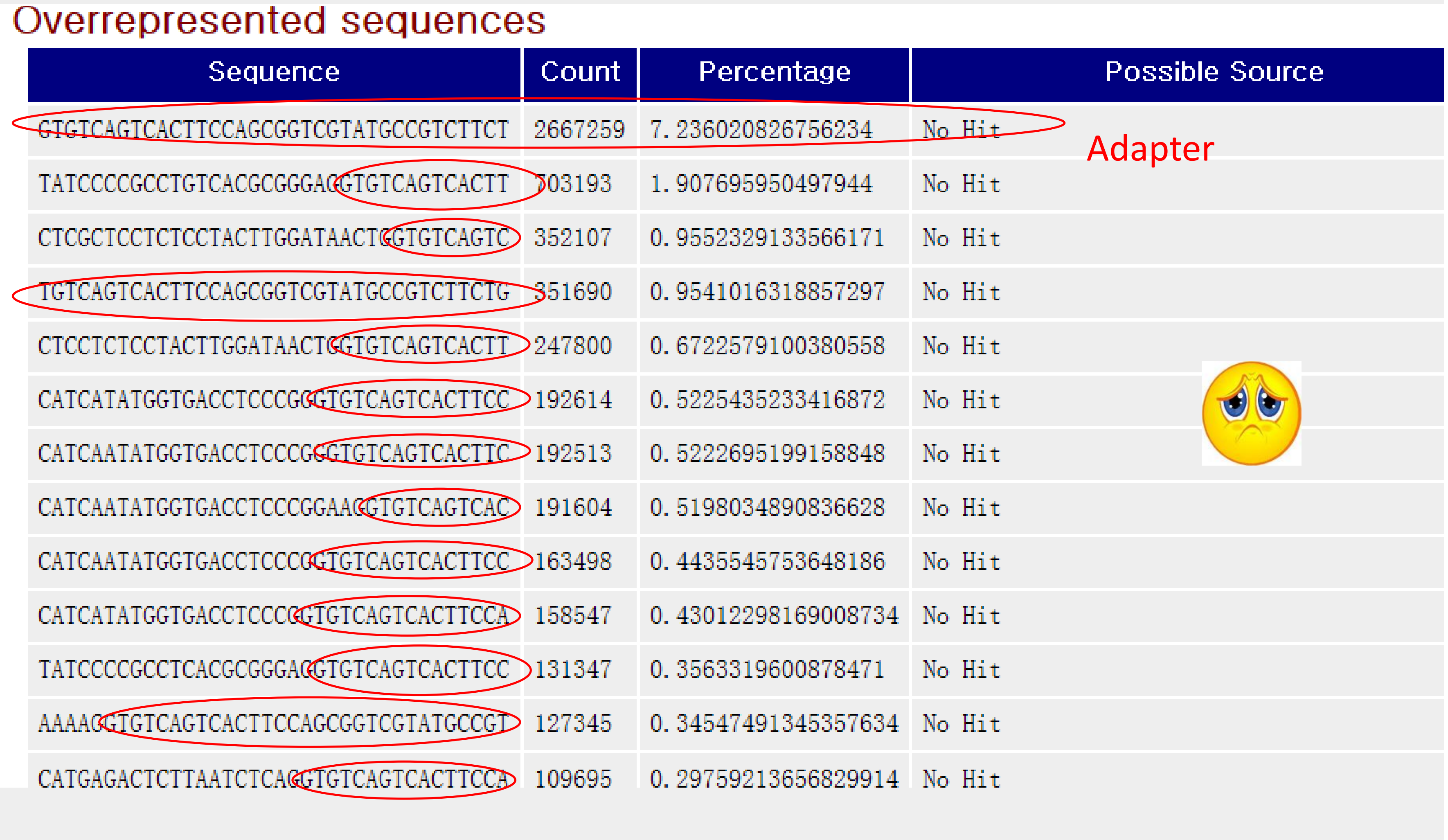


Duplication level





Overrepresented Sequences



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GTGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCT	2667259	7.236020826756234	No Hit
TATCCCCGCCTGTCACGCGGGAGGTGTCAGTCACTT	703193	1.907695950497944	No Hit
CTCGCTCCTCTCCTACTTGGATAACTGGTGTCAGTC	352107	0.9552329133566171	No Hit
TGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCTG	351690	0.9541016318857297	No Hit
CTCCTCTCCTACTTGGATAACTGGTGTCAGTCACTT	247800	0.6722579100380558	No Hit
CATCATATGGTGACCTCCCGGGTGTCAGTCACTTCC	192614	0.5225435233416872	No Hit
CATCAATATGGTGACCTCCCGGGTGTCAGTCACTTC	192513	0.5222695199158848	No Hit
CATCAATATGGTGACCTCCCGGAAGGTGTCAGTCAC	191604	0.5198034890836628	No Hit
CATCAATATGGTGACCTCCCGGTGTCAGTCACTTCC	163498	0.4435545753648186	No Hit
CATCATATGGTGACCTCCCGGTGTCAGTCACTTCCA	158547	0.43012298169008734	No Hit
TATCCCCGCCTCACGCGGGAGGTGTCAGTCACTTCC	131347	0.3563319600878471	No Hit
AAAAGGTGTCAGTCACTTCCAGCGGTCGTATGCCGT	127345	0.34547491345357634	No Hit
CATGAGACTCTTAATCTCAGGTGTCAGTCACTTCCA	109695	0.29759213656829914	No Hit

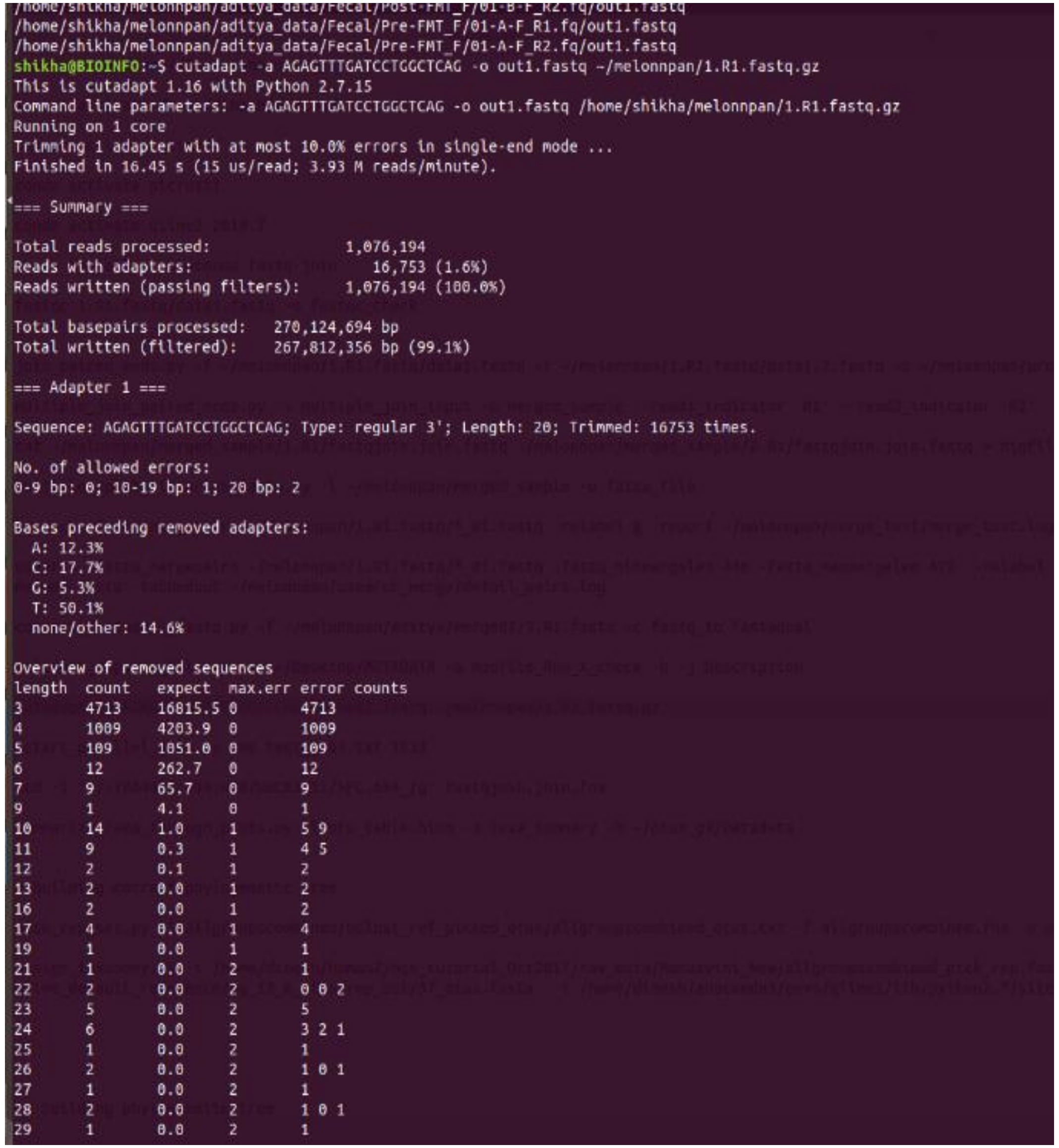
Adapter

CUTADAPT

- Reads from small-RNA sequencing contain the 3’ sequencing adapter because the read is longer than the molecule that is sequenced.
- Poly-A tails are useful for pulling out RNA from your sample, but often you don’t want them to be in your reads.
- Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

- ***sudo apt install cutadapt***

- ***cutadapt –a adaptor sequence –o output.fastq input.fastq***



```
/home/shikha/melonnpan/aditya_data/Fecal/Post-FMT_F/01-B-F_R2.fq/out1.fastq
/home/shikha/melonnpan/aditya_data/Fecal/Pre-FMT_F/01-A-F_R1.fq/out1.fastq
/home/shikha/melonnpan/aditya_data/Fecal/Pre-FMT_F/01-A-F_R2.fq/out1.fastq
shikha@BIOINFO:~$ cutadapt -a AGAGTTTGATCCTGGCTCAG -o out1.fastq ~/melonnpan/1.R1.fastq.gz
This is cutadapt 1.16 with Python 2.7.15
Command line parameters: -a AGAGTTTGATCCTGGCTCAG -o out1.fastq /home/shikha/melonnpan/1.R1.fastq.gz
Running on 1 core
Trimming 1 adapter with at most 10.0% errors in single-end mode ...
Finished in 16.45 s (15 us/read; 3.93 M reads/minute).

=== Summary ===

Total reads processed:               1,076,194
Reads with adapters:                    16,753 (1.6%)
Reads written (passing filters):     1,076,194 (100.0%)

Total basepairs processed:   270,124,694 bp
Total written (filtered):    267,812,356 bp (99.1%)

=== Adapter 1 ===

Sequence: AGAGTTTGATCCTGGCTCAG; Type: regular 3'; Length: 20; Trimmed: 16753 times.

No. of allowed errors:
0-9 bp: 0; 10-19 bp: 1; 20 bp: 2

Bases preceding removed adapters:
  A: 12.3%
  C: 17.7%
  G: 5.3%
  T: 50.1%
  none/other: 14.6%

Overview of removed sequences
length  count   expect  max.err error counts
3       4713    16815.5 0       4713
4       1009    4203.9  0       1009
5       109     1051.0  0       109
6       12      262.7   0       12
7       9       65.7    0       9
9       1       4.1     0       1
10      14      1.0     1       5 9
11      9       0.3     1       4 5
12      2       0.1     1       2
13      2       0.0     1       2
16      2       0.0     1       2
17      4       0.0     1       4
19      1       0.0     1       1
21      1       0.0     2       1
22      2       0.0     2       0 0 2
23      5       0.0     2       5
24      6       0.0     2       3 2 1
25      1       0.0     2       1
26      2       0.0     2       1 0 1
27      1       0.0     2       1
28      2       0.0     2       1 0 1
29      1       0.0     2       1
```

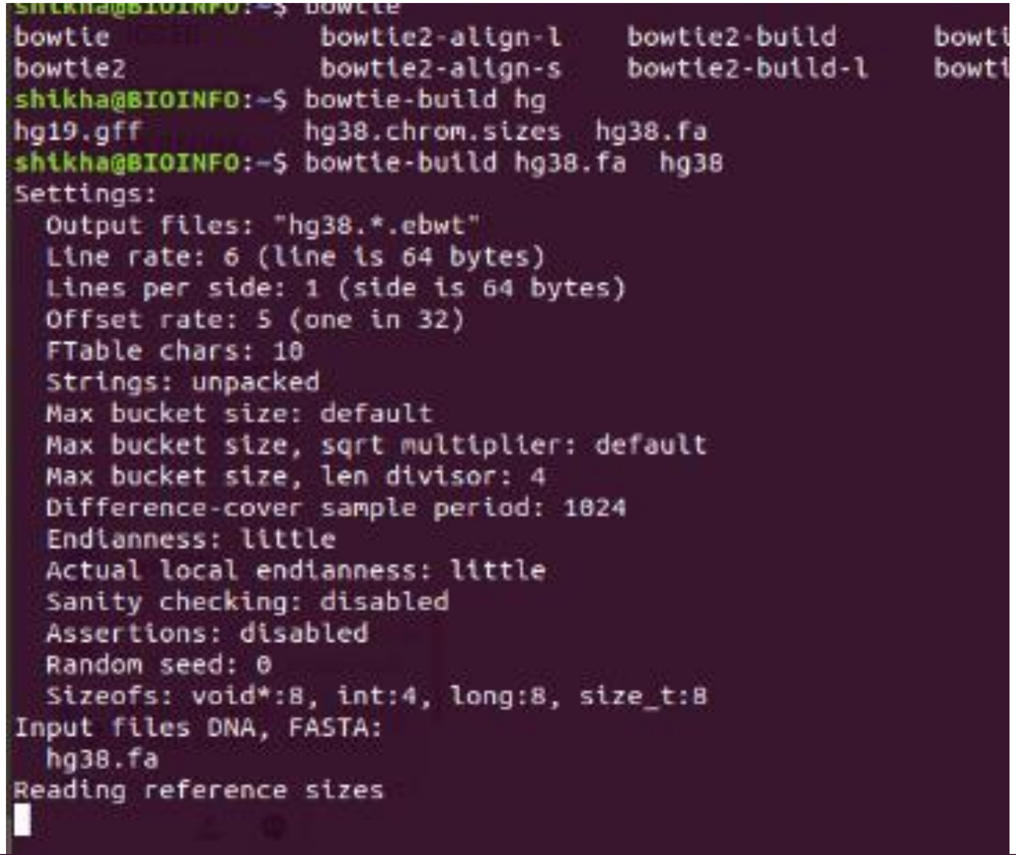
BOWTIE

- Recent software tools allow the mapping (alignment) of millions or billions of short reads to a reference genome.

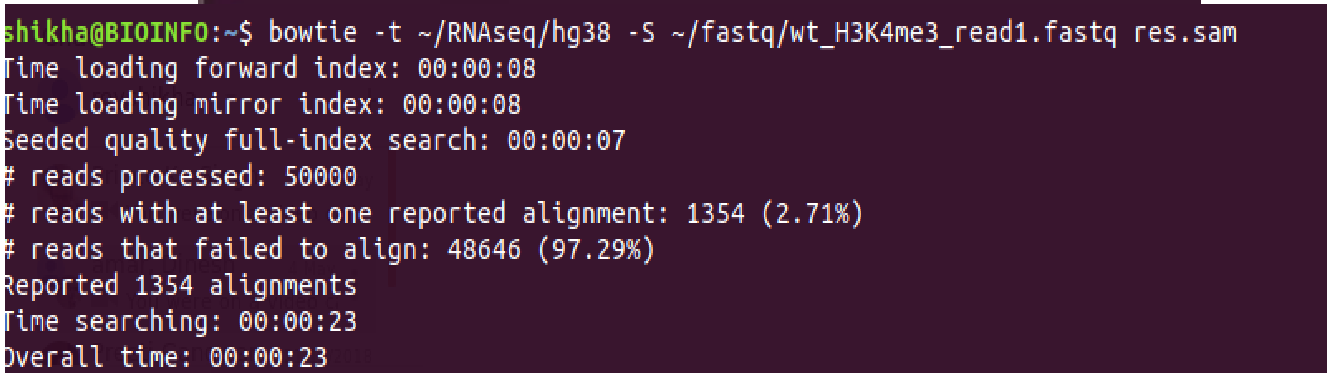
- For the human genome, this would take thousands of hours using BLAST.
- Indexing a genome can be explained similar to indexing a book. If you want to know on which page a certain word appears or a chapter begins, it is much more efficient/faster to look it up in a pre-built index than going through every page of the book until you found it.

bowtie-build ~/hg38.fa hg38

bowtie -t hg38 -S ~/fastq/wt_H3K4me3_read1.fastq res.sam

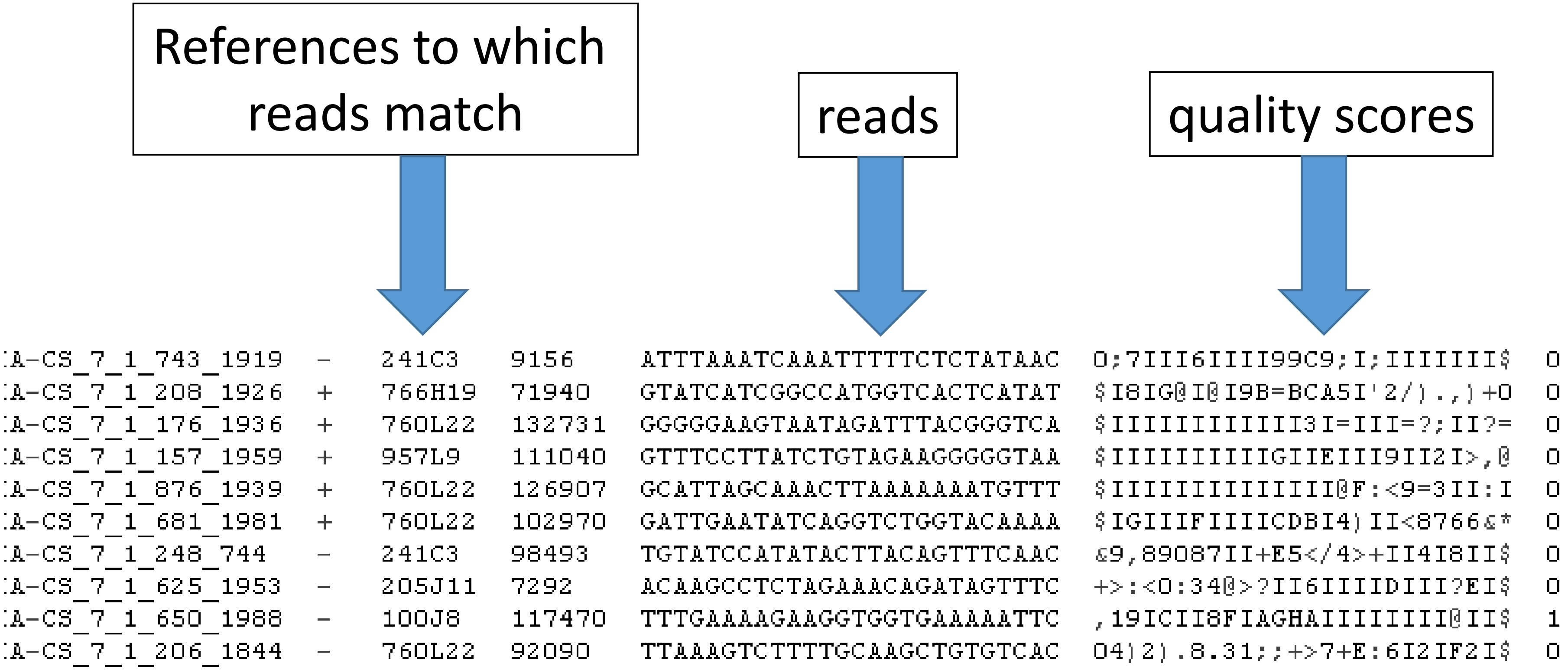


```
shikha@BIOINFO:~$ bowtie
bowtie               bowtie2-align-l    bowtie2-build      bowti
bowtie2              bowtie2-align-s    bowtie2-build-l    bowti
shikha@BIOINFO:~$ bowtie-build hg
hg19.gff            hg38.chrom.sizes  hg38.fa
shikha@BIOINFO:~$ bowtie-build hg38.fa  hg38
Settings:
  Output files: "hg38.*.ebwt"
  Line rate: 6 (line is 64 bytes)
  Lines per side: 1 (side is 64 bytes)
  Offset rate: 5 (one in 32)
  FTable chars: 10
  Strings: unpacked
  Max bucket size: default
  Max bucket size, sqrt multiplier: default
  Max bucket size, len divisor: 4
  Difference-cover sample period: 1024
  Endianness: little
  Actual local endianness: little
  Sanity checking: disabled
  Assertions: disabled
  Random seed: 0
  Sizeofs: void*:8, int:4, long:8, size_t:8
Input files DNA, FASTA:
  hg38.fa
Reading reference sizes
█
```



```
shikha@BIOINFO:~$ bowtie -t ~/RNAseq/hg38 -S ~/fastq/wt_H3K4me3_read1.fastq res.sam
Time loading forward index: 00:00:08
Time loading mirror index: 00:00:08
Seeded quality full-index search: 00:00:07
# reads processed: 50000
# reads with at least one reported alignment: 1354 (2.71%)
# reads that failed to align: 48646 (97.29%)
Reported 1354 alignments
Time searching: 00:00:23
Overall time: 00:00:23
```

Alignment to a reference genome: example of short-read alignment (Bowtie) results



```
A-CS_7_1_743_1919   -    241C3   9156     ATTTAAATCAAATTTTTCTCTATAAC  0;7III6IIII99C9;I;IIIIIII$  0
A-CS_7_1_208_1926   +    766H19  71940    GTATCATCGGCCATGGTCACTCATAT  $I8IG@I@I9B=BCA5I'2/).,)+0  0
A-CS_7_1_176_1936   +    760L22  132731   GGGGGAAGTAATAGATTTACGGGTCA  $IIIIIIIIIII3I=III=?;II?=  0
A-CS_7_1_157_1959   +    957L9   111040   GTTTCCTTATCTGTAGAAGGGGGTAA  $IIIIIIIIIGIIEIII9II2I>,@  0
A-CS_7_1_876_1939   +    760L22  126907   GCATTAGCAAACTTAAAAAAATGTTT  $IIIIIIIIIIIII@F:<9=3II:I  0
A-CS_7_1_681_1981   +    760L22  102970   GATTGAATATCAGGTCTGGTACAAAA  $IGIIIFIIIICDBI4)II<8766&*  0
A-CS_7_1_248_744    -    241C3   98493    TGTATCCATATACTTACAGTTTCAAC  &9,89087II+E5</4>+II4I8II$  0
A-CS_7_1_625_1953   -    205J11  7292     ACAAGCCTCTAGAAACAGATAGTTTC  +>:<0:34@>?II6IIIIDIII?EI$  0
A-CS_7_1_650_1988   -    100J8   117470   TTTGAAAAGAAGGTGGTGAAAAATTC  ,19ICII8FIAGHAIIIIIIII@II$  1
A-CS_7_1_206_1844   -    760L22  92090    TTAAAGTCTTTTGCAAGCTGTGTCAC  04)2).8.31;;+>7+E:6I2IF2I$  0
```

SAMTOOLS

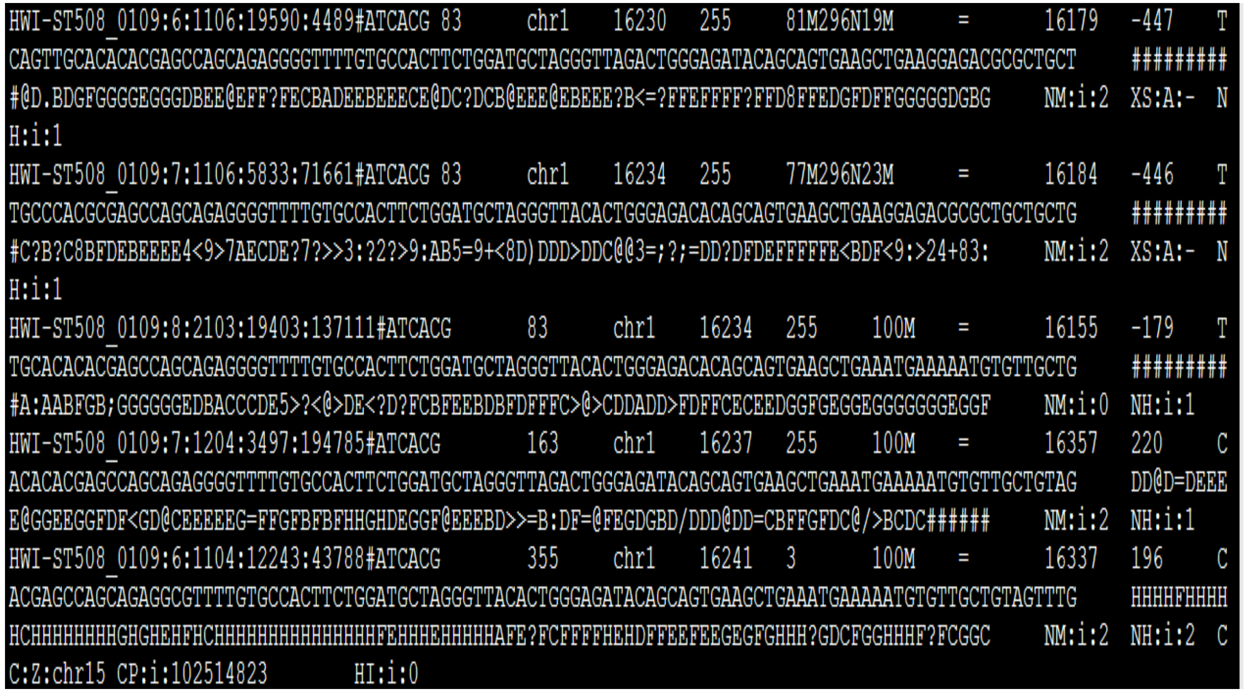
- ✧ **SAM – Sequence Alignment/Map format**
  - ✧ SAM file format stores alignment information
- ✧ **Plain text**
- ✧ **Specification**: http://samtools.sourceforge.net/SAM1.pdf
- ✧ Contains quality information, meta data, alignment information, sequence etc.
- ✧ **Files can be very large:** Many 100's of GB or more
- ✧ Normally converted into **BAM** to save space (and text format is mostly useless for downstream analyses)

SAM is a common format having sequence reads and their alignment to a reference genome.

BAM is the binary form of a SAM file.

SAMTools is a software package commonly used to analyze SAM/BAM files.

samtools view -bS -o res.bam res.sam



```
HWI-ST508_0109:6:1106:19590:4489#ATCACG 83      chr1    16230   255     81M296N19M      =       16179   -447    T
CAGTTGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTAGACTGGGAGATACAGCAGTGAAGCTGAAGGAGACGCGCTGCT     ########
#@D.BDGFGGGGEGGGDBEE@EFF?FECBADEEBEEECE@DC?DCB@EEE@EBEEE?B<=?FFEFFFF?FFD8FFEDGFDFFGGGGGDGBG     NM:i:2  XS:A:-  N
H:i:1
HWI-ST508_0109:7:1106:5833:71661#ATCACG 83      chr1    16234   255     77M296N23M      =       16184   -446    T
TGCCCACGCGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGACACAGCAGTGAAGCTGAAGGAGACGCGCTGCTGCTG     ########
#C?B?C8BFDEBEEEE4<9>7AECDE?7?>>3:?2?>9:AB5=9+<8D)DDD>DDC@@3=;?;=DD?DFDEFFFFFE<BDF<9:>24+83:     NM:i:2  XS:A:-  N
H:i:1
HWI-ST508_0109:8:2103:19403:137111#ATCACG       83      chr1    16234   255     100M    =       16155   -179    T
TGCACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGACACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTG     ########
#A:AABFGB;GGGGGGEDBACCCDE5>?<@>DE<?D?FCBFEEBDBFDFFFC>@>CDDADD>FDFFCECEEDGGFGEGGEGGGGGGGEGGF     NM:i:0  NH:i:1
HWI-ST508_0109:7:1204:3497:194785#ATCACG        163     chr1    16237   255     100M    =       16357   220     C
ACACACGAGCCAGCAGAGGGGTTTTGTGCCACTTCTGGATGCTAGGGTTAGACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAG     DD@D=DEEE
E@GGEEGGFDF<GD@CEEEEEG=FFGFBFBFHHGHDEGGF@EEEBD>>=B:DF=@FEGDGBD/DDD@DD=CBFFGFDC@/>BCDC######     NM:i:2  NH:i:1
HWI-ST508_0109:6:1104:12243:43788#ATCACG        355     chr1    16241   3       100M    =       16337   196     C
ACGAGCCAGCAGAGGCGTTTTGTGCCACTTCTGGATGCTAGGGTTACACTGGGAGATACAGCAGTGAAGCTGAAATGAAAAATGTGTTGCTGTAGTTTG     HHHHFHHHH
HCHHHHHHHHGHGHEHFHCHHHHHHHHHHHHHHHHFEHHHEHHHHHAFE?FCFFFFHEHDFFEEFEEGEGFGHHH?GDCFGGHHHF?FCGGC     NM:i:2  NH:i:2  C
C:Z:chr15 CP:i:102514823        HI:i:0
```

Formats : **BAM**

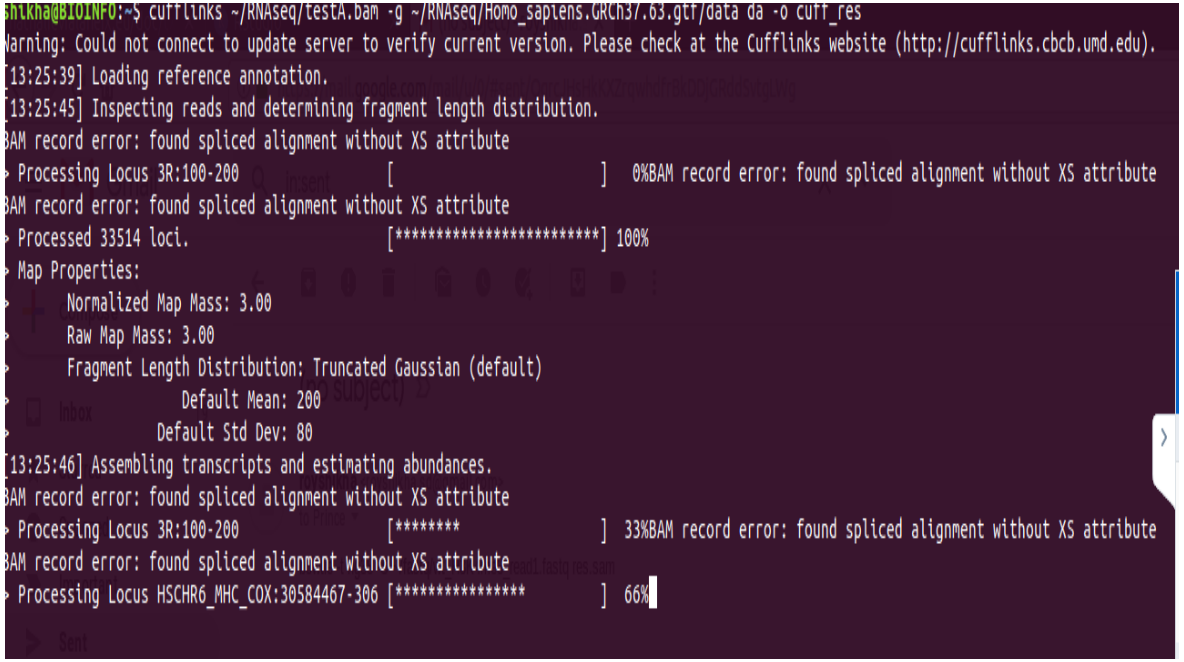
- ✧**BAM – BGZF compressed SAM format**
  - ✧Compressed/binary version of SAM and is **not human readable.** Uses a specialized compression algorithm optimized for indexing and record retrieval (bgzip)
  - ✧Makes the alignment information easily accessible to downstream applications (large genome file not necessary)
  - ✧Unsorted, sorted by sequence name, **sorted by genome coordinates**
  - ✧May be accompanied by an index file (.bai) (only if coordinate sorted)

- ✧**Files are typically very large:** ~ 1/5 of SAM, but still very large

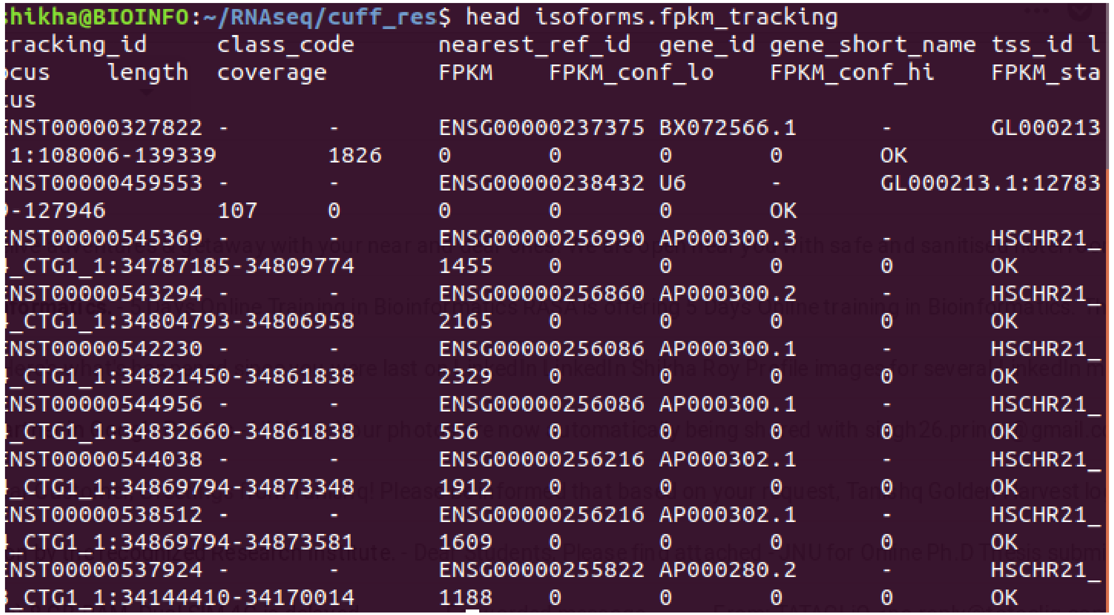
CUFFLINKS

- Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples.
- It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts.
- Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.
- Output tracks of Cufflinks is the Assembled transcripts track, output tables of Cufflinks are Gene expression and Transcript expression tables.

cufflinks testA.bam -g Homo_sapiens.GRCh37.63.gtf/data da -o cuff_res



```
shikha@BIOINFO:~$ cufflinks ~/RNAseq/testA.bam -g ~/RNAseq/Homo_sapiens.GRCh37.63.gtf/data da -o cuff_res
Warning: Could not connect to update server to verify current version. Please check at the Cufflinks website (http://cufflinks.cbcb.umd.edu).
[13:25:39] Loading reference annotation.
[13:25:45] Inspecting reads and determining fragment length distribution.
BAM record error: found spliced alignment without XS attribute
> Processing Locus 3R:100-200          [                         ]   0%BAM record error: found spliced alignment without XS attribute
BAM record error: found spliced alignment without XS attribute
> Processed 33514 loci.                [*************************] 100%
> Map Properties:
>       Normalized Map Mass: 3.00
>       Raw Map Mass: 3.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[13:25:46] Assembling transcripts and estimating abundances.
BAM record error: found spliced alignment without XS attribute
> Processing Locus 3R:100-200          [********                 ]  33%BAM record error: found spliced alignment without XS attribute
BAM record error: found spliced alignment without XS attribute
> Processing Locus HSCHR6_MHC_COX:30584467-306 [***************          ]  66%
```

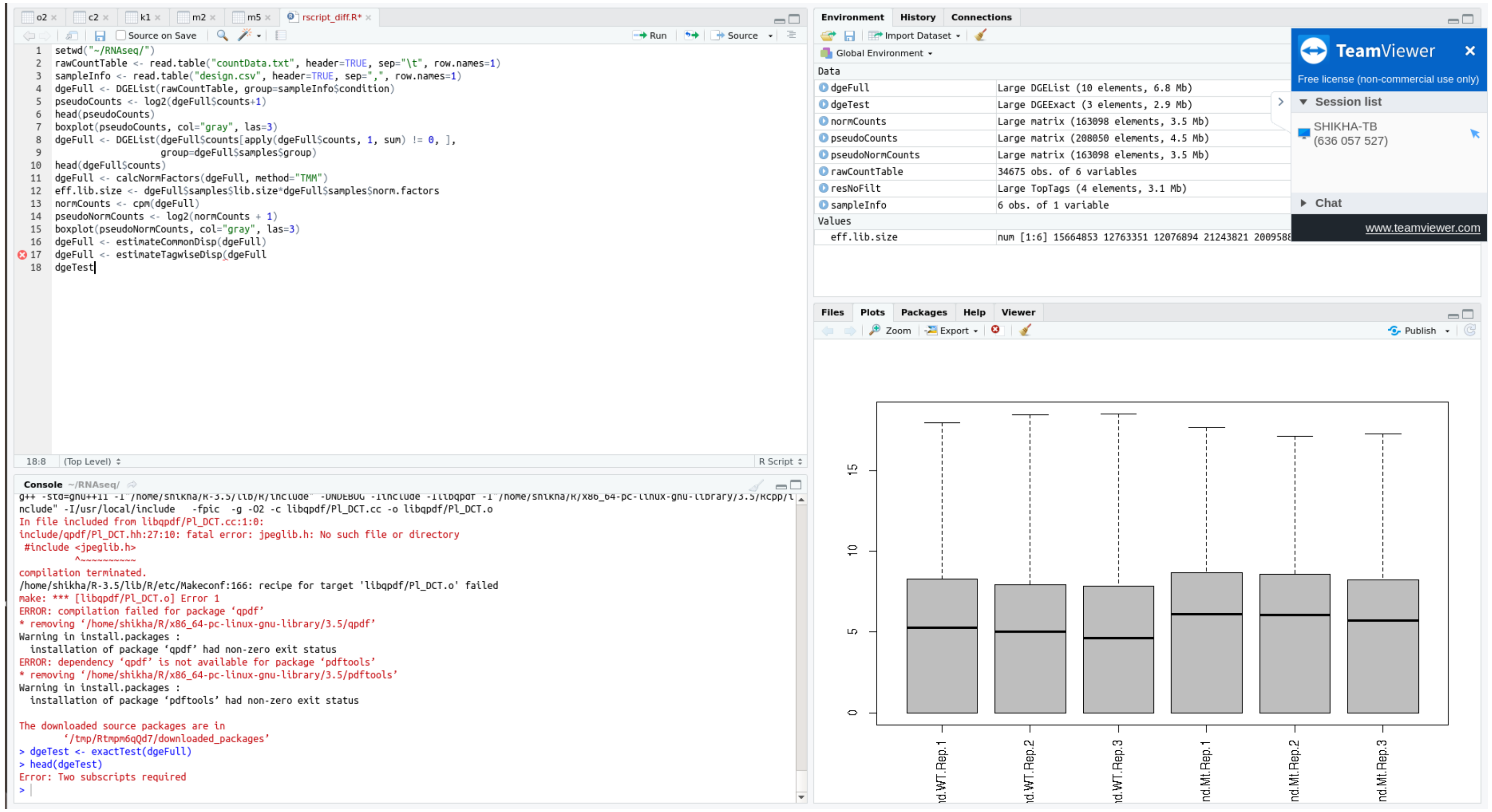


```
hikha@BIOINFO:~/RNAseq/cuff_res$ head isoforms.fpkm_tracking
racking_id      class_code      nearest_ref_id  gene_id gene_short_name tss_id l
cus     length  coverage        FPKM    FPKM_conf_lo    FPKM_conf_hi    FPKM_sta
us
NST00000327822 -       -       ENSG00000237375 BX072566.1      -       GL000213
 1:108006-139339        1826    0       0       0       0       OK
NST00000459553 -       -       ENSG00000238432 U6      -       GL000213.1:12783
-127946 107     0       0       0       0       OK
NST00000545369 -       -       ENSG00000256990 AP000300.3      -       HSCHR21_
_CTG1_1:34787185-34809774       1455    0       0       0       0       OK
NST00000543294 -       -       ENSG00000256860 AP000300.2      -       HSCHR21_
_CTG1_1:34804793-34806958       2165    0       0       0       0       OK
NST00000542230 -       -       ENSG00000256086 AP000300.1      -       HSCHR21_
_CTG1_1:34821450-34861838       2329    0       0       0       0       OK
NST00000544956 -       -       ENSG00000256086 AP000300.1      -       HSCHR21_
_CTG1_1:34832660-34861838       556     0       0       0       0       OK
NST00000544038 -       -       ENSG00000256216 AP000302.1      -       HSCHR21_
_CTG1_1:34869794-34873348       1912    0       0       0       0       OK
NST00000538512 -       -       ENSG00000256216 AP000302.1      -       HSCHR21_
_CTG1_1:34869794-34873581       1609    0       0       0       0       OK
NST00000537924 -       -       ENSG00000255822 AP000280.2      -       HSCHR21_
_CTG1_1:34144410-34170014       1188    0       0       0       0       OK
```

Metrics for quantifying gene expression levels

- RPKM
  - **R**eads **P**er **K**ilobase per **M**illion mapped reads
  - Normalize relative to sequencing depth and gene length
- FPKM
  - Similar to RPKM but count **DNA fragments** instead of reads
  - Used in paired end RNA-Seq experiments to avoid bias
- TPM
  - **T**ranscripts **P**er **M**illion
  - Normalize for gene length, then normalize by sequencing depth

DIFFERENTIAL EXPRESSION USING RSTUDIO

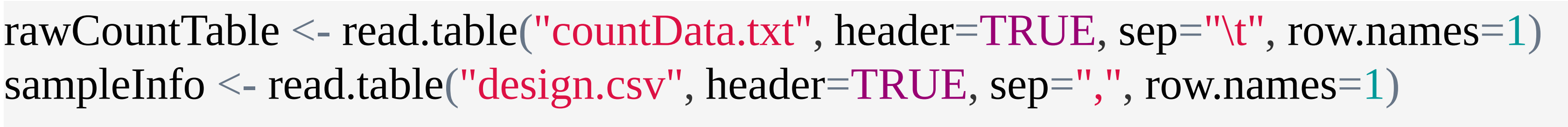


```r
setwd("~/RNAseq/")
rawCountTable <- read.table("countData.txt", header=TRUE, sep="\t", row.names=1)
sampleInfo <- read.table("design.csv", header=TRUE, sep=",", row.names=1)
dgeFull <- DGEList(rawCountTable, group=sampleInfo$condition)
pseudoCounts <- log2(dgeFull$counts+1)
head(pseudoCounts)
boxplot(pseudoCounts, col="gray", las=3)
dgeFull <- DGEList(dgeFull$counts[apply(dgeFull$counts, 1, sum) != 0, ],
                   group=dgeFull$samples$group)
head(dgeFull$counts)
dgeFull <- calcNormFactors(dgeFull, method="TMM")
eff.lib.size <- dgeFull$samples$lib.size*dgeFull$samples$norm.factors
normCounts <- cpm(dgeFull)
pseudoNormCounts <- log2(normCounts + 1)
boxplot(pseudoNormCounts, col="gray", las=3)
dgeFull <- estimateCommonDisp(dgeFull)
dgeFull <- estimateTagwiseDisp(dgeFull
dgeTest
```

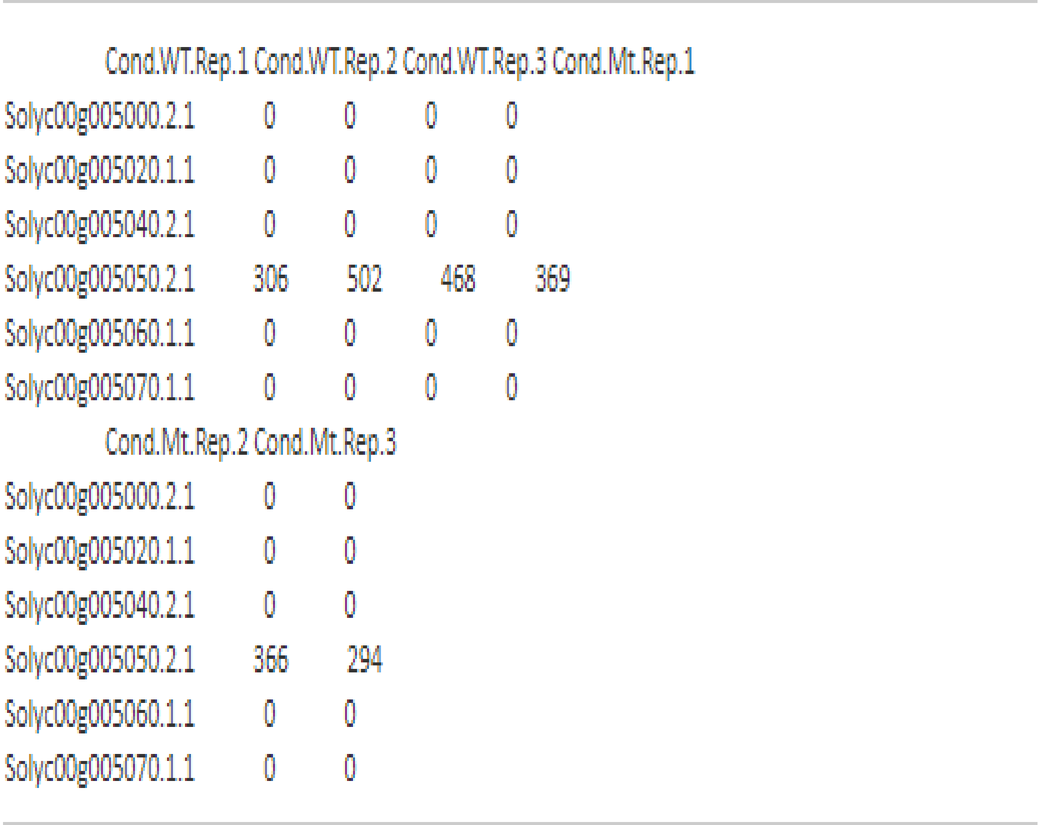
```
g++ -std=gnu++11 -I"/home/shikha/R-3.5/lib/R/include" -DNDEBUG -Iinclude -Ilibqpdf -I"/home/shikha/R/x86_64-pc-linux-gnu-library/3.5/Rcpp/i
nclude" -I/usr/local/include   -fpic  -g -O2 -c libqpdf/Pl_DCT.cc -o libqpdf/Pl_DCT.o
In file included from libqpdf/Pl_DCT.cc:1:0:
include/qpdf/Pl_DCT.hh:27:10: fatal error: jpeglib.h: No such file or directory
 #include <jpeglib.h>
          ^~~~~~~~~~~
compilation terminated.
/home/shikha/R-3.5/lib/R/etc/Makeconf:166: recipe for target 'libqpdf/Pl_DCT.o' failed
make: *** [libqpdf/Pl_DCT.o] Error 1
ERROR: compilation failed for package ‘qpdf’
* removing ‘/home/shikha/R/x86_64-pc-linux-gnu-library/3.5/qpdf’
Warning in install.packages :
  installation of package ‘qpdf’ had non-zero exit status
ERROR: dependency ‘qpdf’ is not available for package ‘pdftools’
* removing ‘/home/shikha/R/x86_64-pc-linux-gnu-library/3.5/pdftools’
Warning in install.packages :
  installation of package ‘pdftools’ had non-zero exit status

The downloaded source packages are in
        ‘/tmp/Rtmpm6qQd7/downloaded_packages’
> dgeTest <- exactTest(dgeFull)
> head(dgeTest)
Error: Two subscripts required
> 
```

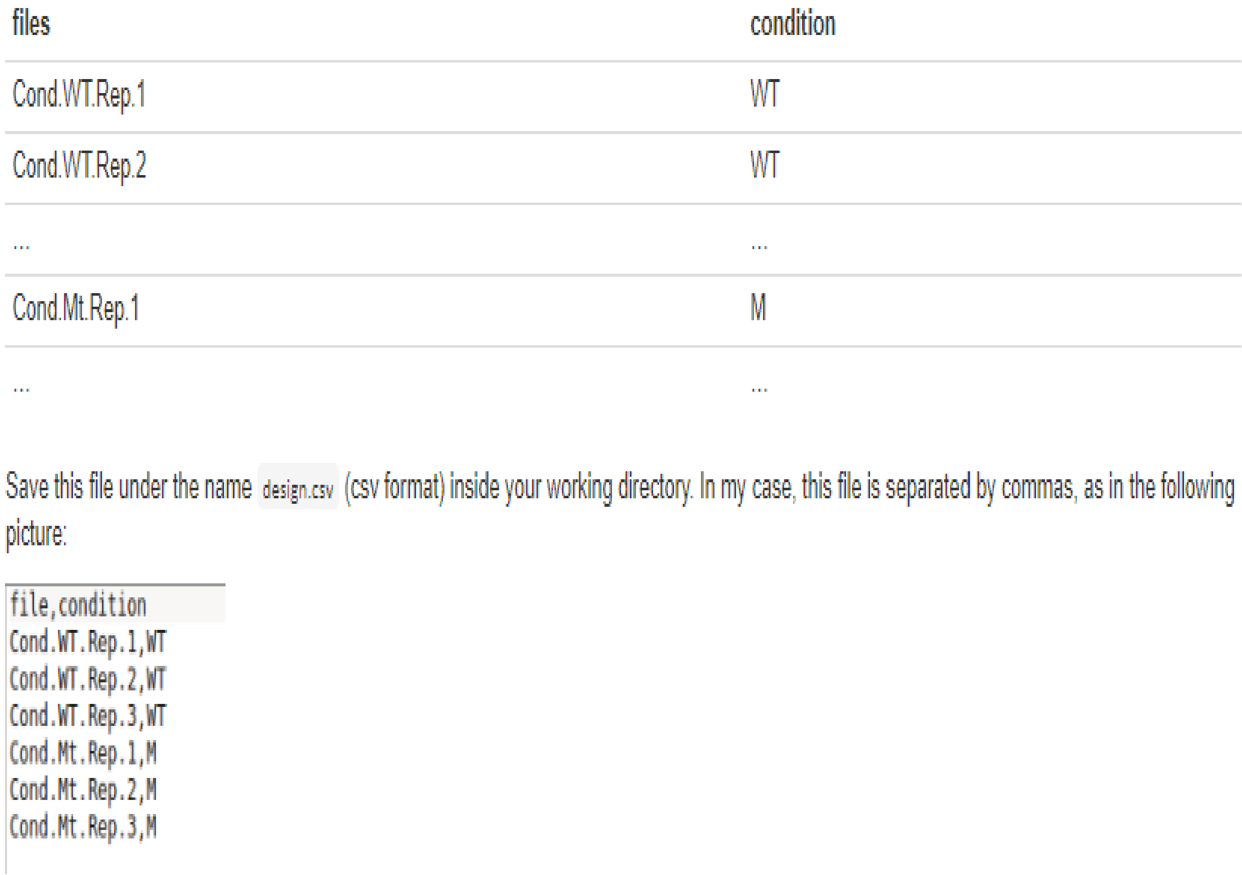
Data	
dgeFull	Large DGEList (10 elements, 6.8 Mb)
dgeTest	Large DGEExact (3 elements, 2.9 Mb)
normCounts	Large matrix (163098 elements, 3.5 Mb)
pseudoCounts	Large matrix (208050 elements, 4.5 Mb)
pseudoNormCounts	Large matrix (163098 elements, 3.5 Mb)
rawCountTable	34675 obs. of 6 variables
resNoFilt	Large TopTags (4 elements, 3.1 Mb)
sampleInfo	6 obs. of 1 variable
Values	
eff.lib.size	num [1:6] 15664853 12763351 12076894 21243821 2009588



```
rawCountTable <- read.table("countData.txt", header=TRUE, sep="\t", row.names=1)
sampleInfo <- read.table("design.csv", header=TRUE, sep=",", row.names=1)
```



```
             Cond.WT.Rep.1 Cond.WT.Rep.2 Cond.WT.Rep.3 Cond.Mt.Rep.1
Solyc00g005000.2.1     0       0       0       0
Solyc00g005020.1.1     0       0       0       0
Solyc00g005040.2.1     0       0       0       0
Solyc00g005050.2.1    306     502     468     369
Solyc00g005060.1.1     0       0       0       0
Solyc00g005070.1.1     0       0       0       0
             Cond.Mt.Rep.2 Cond.Mt.Rep.3
Solyc00g005000.2.1     0       0
Solyc00g005020.1.1     0       0
Solyc00g005040.2.1     0       0
Solyc00g005050.2.1    366     294
Solyc00g005060.1.1     0       0
Solyc00g005070.1.1     0       0
```



files	condition
Cond.WT.Rep.1	WT
Cond.WT.Rep.2	WT
...	...
Cond.Mt.Rep.1	M
...	...

Save this file under the name `design.csv` (csv format) inside your working directory. In my case, this file is separated by commas, as in the following picture:

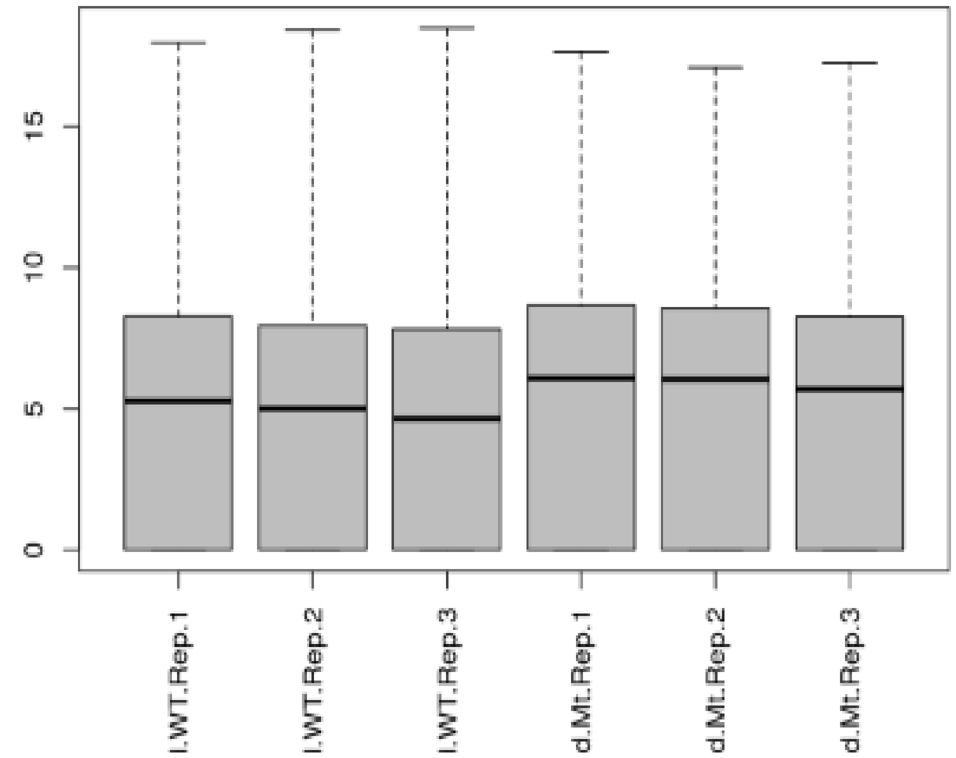
```
file,condition
Cond.WT.Rep.1,WT
Cond.WT.Rep.2,WT
Cond.WT.Rep.3,WT
Cond.Mt.Rep.1,M
Cond.Mt.Rep.2,M
Cond.Mt.Rep.3,M
```

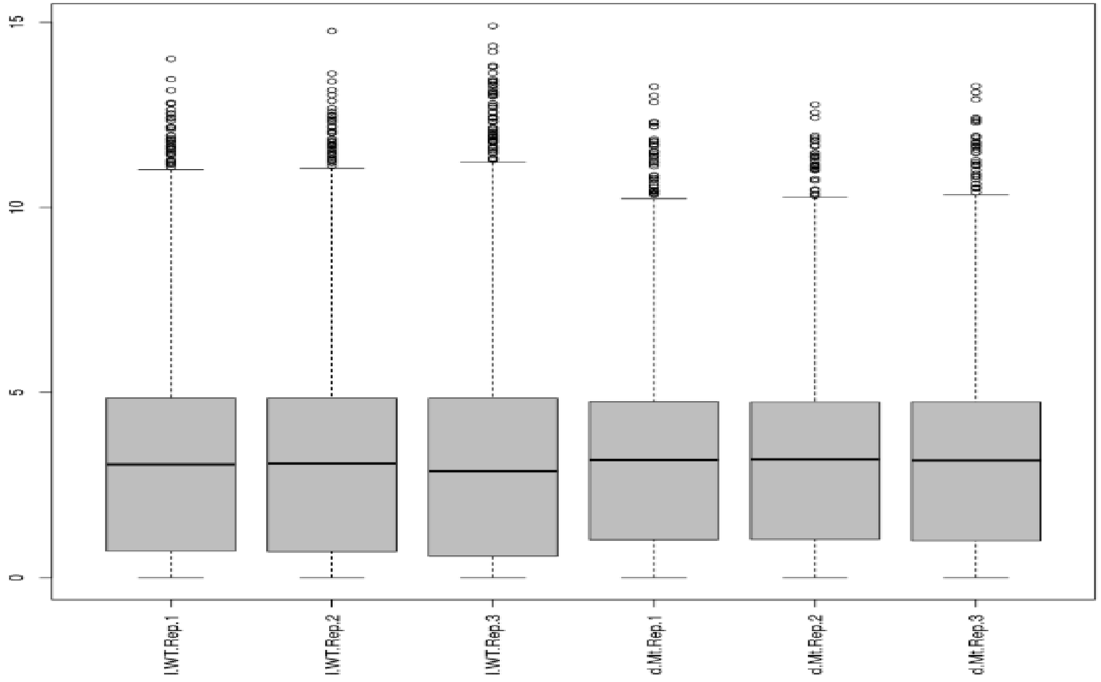
Create a DGEList data object

```
dgeFull <- DGEList(rawCountTable, group=sampleInfo$condition)
pseudoCounts <- log2(dgeFull$counts+1)
boxplot(pseudoCounts, col="gray", las=3)
```

estimate the normalization factors

```
dgeFull <- calcNormFactors(dgeFull, method="TMM")
eff.lib.size <- dgeFull$samples$lib.size*dgeFull$samples$norm.factors
normCounts <- cpm(dgeFull)
pseudoNormCounts <- log2(normCounts + 1)
boxplot(pseudoNormCounts, col="gray", las=3)
```





Differential Gene Expression overview

④ Set up to do differential gene expression (DGE)

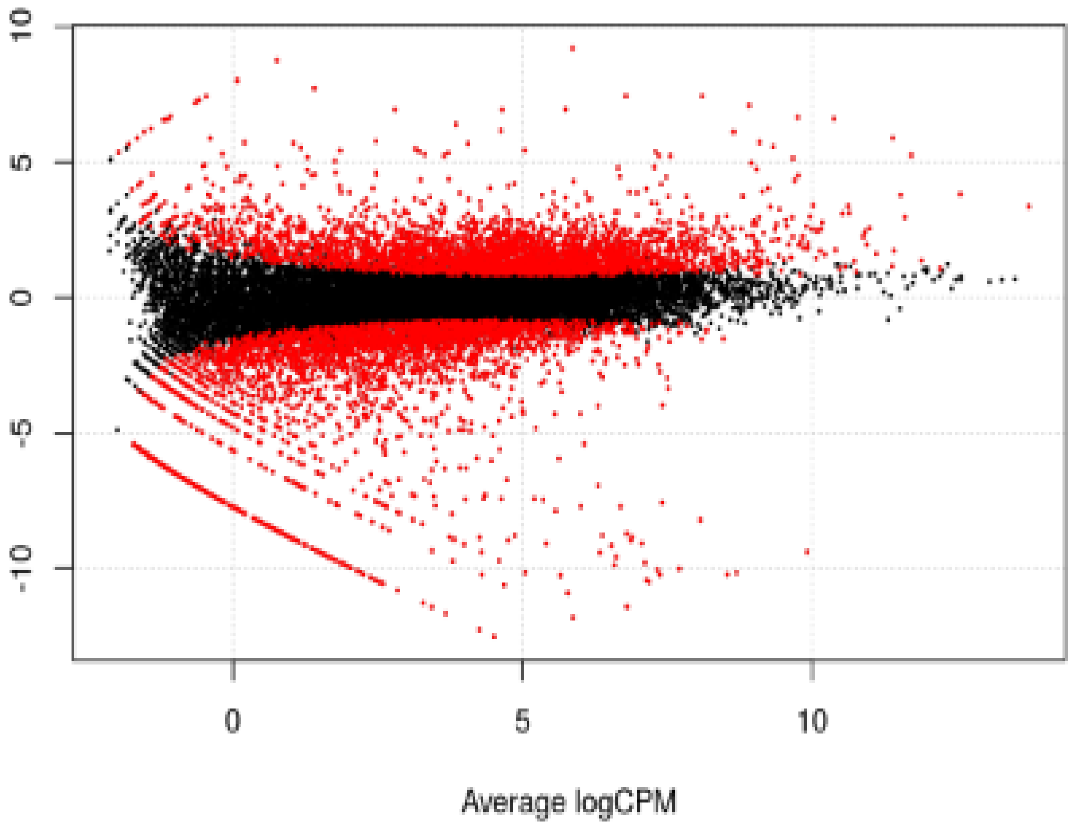
Identify read counts associated with genes

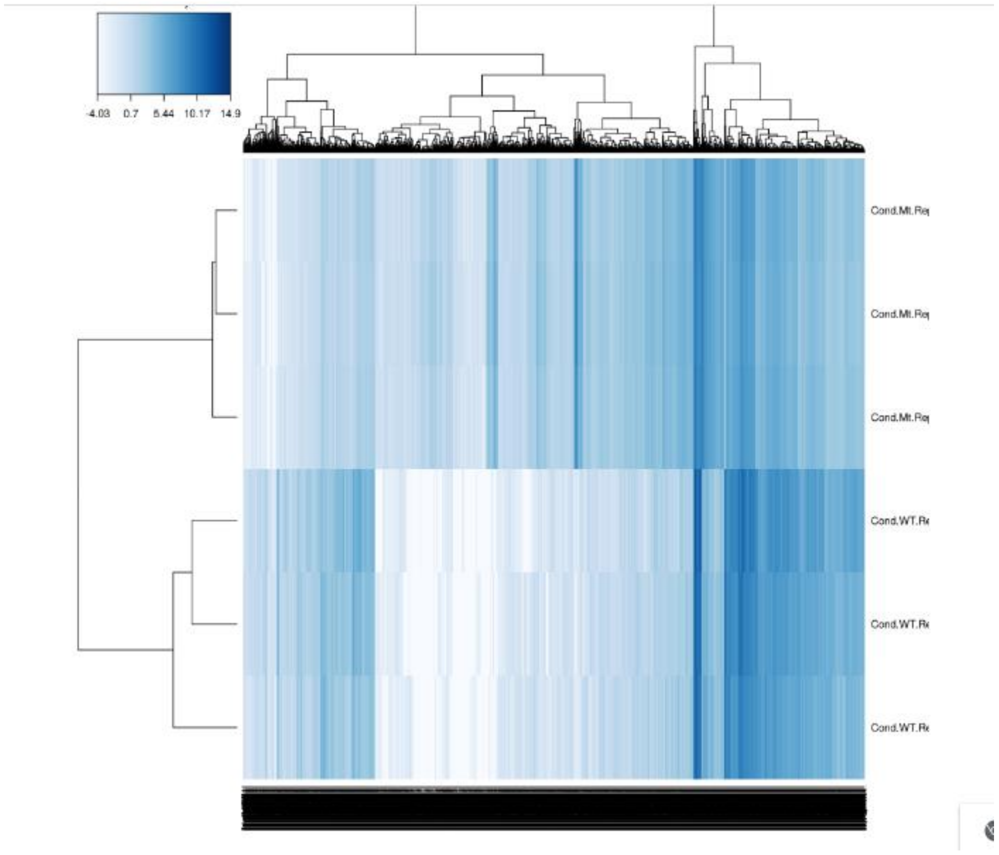
a. Do you want to obtain raw read counts or normalized read counts? This will depend on the statistical analysis you wish to perform downstream

   - ✧ htseq & feature-counts return raw read counts
     - ✧ Required for R programs like DESeq & EdgeR
   - ✧ Ballgown & Cufflinks return FPKM normalized counts for each gene

```
dgeFull <- DGEList(dgeFull$counts
[apply(dgeFull$counts, 1, sum) != 0,
],group=dgeFull$samples$group)
dgeFull <- estimateCommonDisp(dgeFull)
dgeFull <- estimateTagwiseDisp(dgeFull)
dgeTest <- exactTest(dgeFull)
remove low expressed genes
filtData <- HTSFilter(dgeFull)filteredData
dgeTestFilt <- exactTest(filtData)
resFilt <- topTags(dgeTestFilt,
n=nrow(dgeTest$table))
sigReg <- resFilt$table(resFilt$table$FDR<0.01,]
sigReg <- resFilt$table[order(sigReg$logFC),]
```

```
plotSmear(dgeTestFilt, de.tags = rownames(resFilt$table)[which(resFilt$table$FDR<0.01)])
```





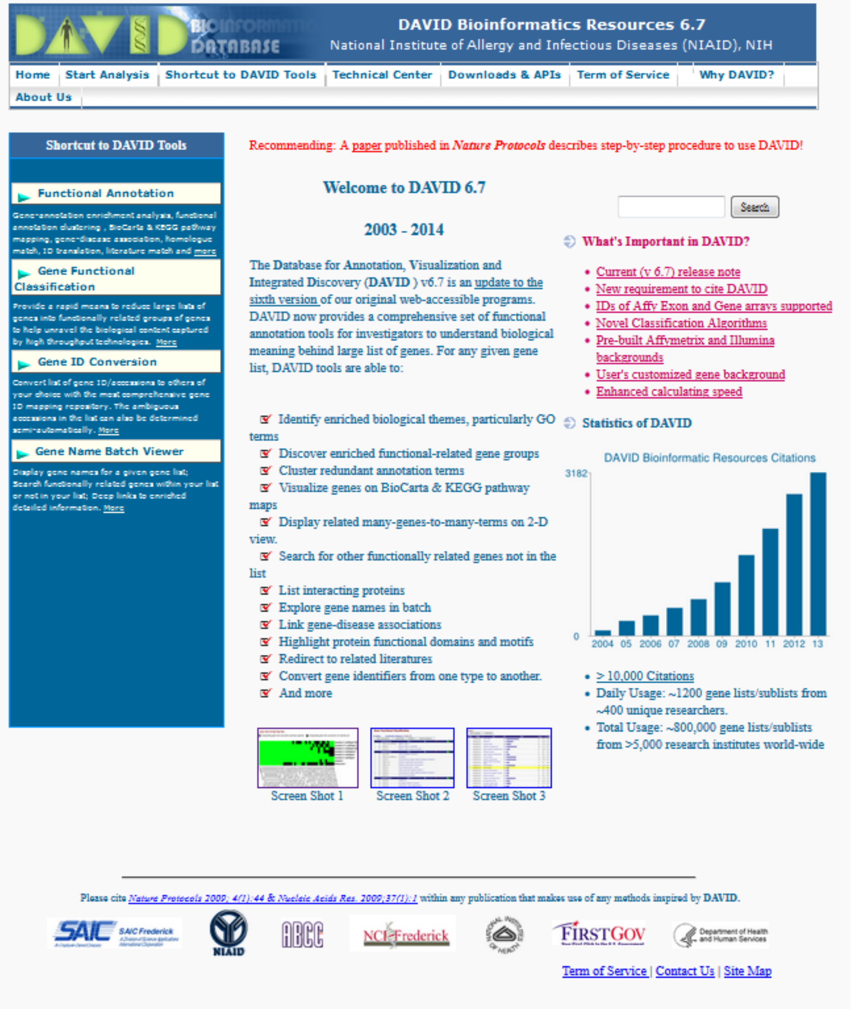
```
selY <- y[rownames(resFilt$table)[resFilt$table$FDR<0.01 & abs(resFilt$table$logFC)>1.5],]
```

```
cimColor <- colorRampPalette(rev(brewer.pal(9, "Blues")))(255)[255:1] finalHM <- cim(t(selY),
 color=cimColor, symkey=FALSE)
```

Tools for analyzing differentially expressed genes

- Gene Ontology (GO) terms enrichment:
  - topGO (https://bioconductor.org/packages/release/bioc/html/topGO.html)
  - goSTAG (https://bioconductor.org/packages/release/bioc/html/goSTAG.html)
  - DAVID (https://david.ncifcrf.gov/)
- Pathway analysis:
  - GAGE (http://bioconductor.org/packages/release/bioc/html/gage.html)
  - Reactome (http://www.reactome.org/)
- Sample walkthrough:
  - From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline
    - https://www.bioconductor.org/help/workflows/RnaSeqGeneEdgeRQL/

GENE ONTOLOGY ENRICHMENT USING DAVID



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2014

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- ☑ Identify enriched biological themes, particularly GO terms
- ☑ Discover enriched functional-related gene groups
- ☑ Cluster redundant annotation terms
- ☑ Visualize genes on BioCarta & KEGG pathway maps
- ☑ Display related many-genes-to-many-terms on 2-D view.
- ☑ Search for other functionally related genes not in the list
- ☑ List interacting proteins
- ☑ Explore gene names in batch
- ☑ Link gene-disease associations
- ☑ Highlight protein functional domains and motifs
- ☑ Redirect to related literatures
- ☑ Convert gene identifiers from one type to another.
- ☑ And more

What's Important in DAVID?

- Current (v 6.7) release note
- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID

DAVID Bioinformatic Resources Citations

- > 10,000 Citations
- Daily Usage: ~1200 gene lists/sublists from ~400 unique researchers.
- Total Usage: ~800,000 gene lists/sublists from >5,000 research institutes world-wide

Screen Shot 1 Screen Shot 2 Screen Shot 3

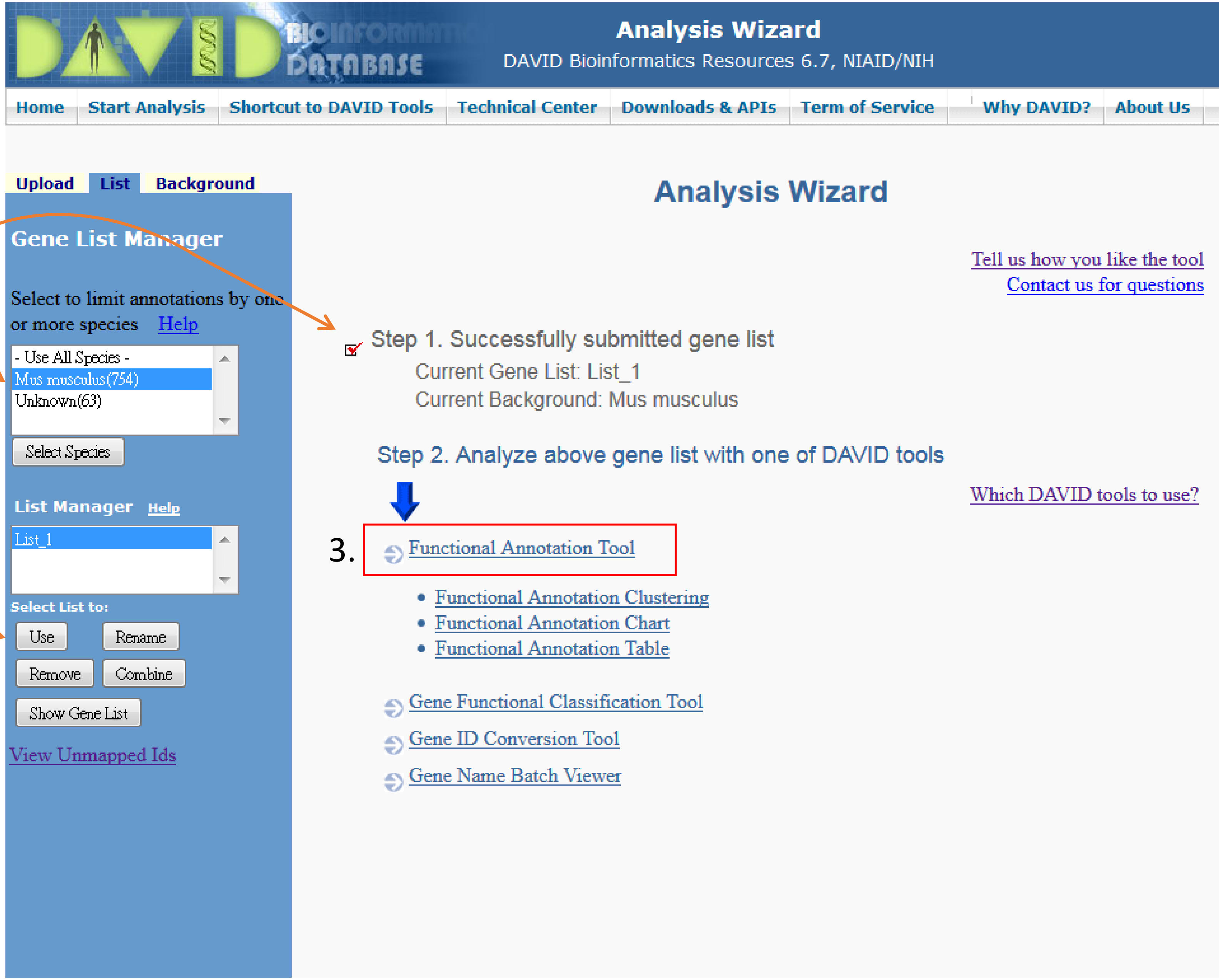
Please cite *Nature Protocols 2009; 4(1):44* & *Nucleic Acids Res. 2009;37(1):1* within any publication that makes use of any methods inspired by DAVID.

Term of Service | Contact Us | Site Map

- ➢ Functional Annotation Tool
  - Gene Ontology
  - Protein interaction
  - Protein domain
  - Pathway
  - Disease
- ➢ Gene ID Conversion
- ➢ Gene Functional Classification

STEP 1

STEP 2



Analysis Wizard

DAVID Bioinformatics Resources 6.7, NIAID/NIH

Gene List Manager

Select to limit annotations by one or more species Help

- Use All Species -
Mus musculus(754)
Unknown(63)

List Manager

List_1

Select List to:

View Unmapped Ids

Analysis Wizard

Tell us how you like the tool
Contact us for questions

Step 1. Successfully submitted gene list
Current Gene List: List_1
Current Background: Mus musculus

Step 2. Analyze above gene list with one of DAVID tools

Which DAVID tools to use?

Functional Annotation Tool

- Functional Annotation Clustering
- Functional Annotation Chart
- Functional Annotation Table

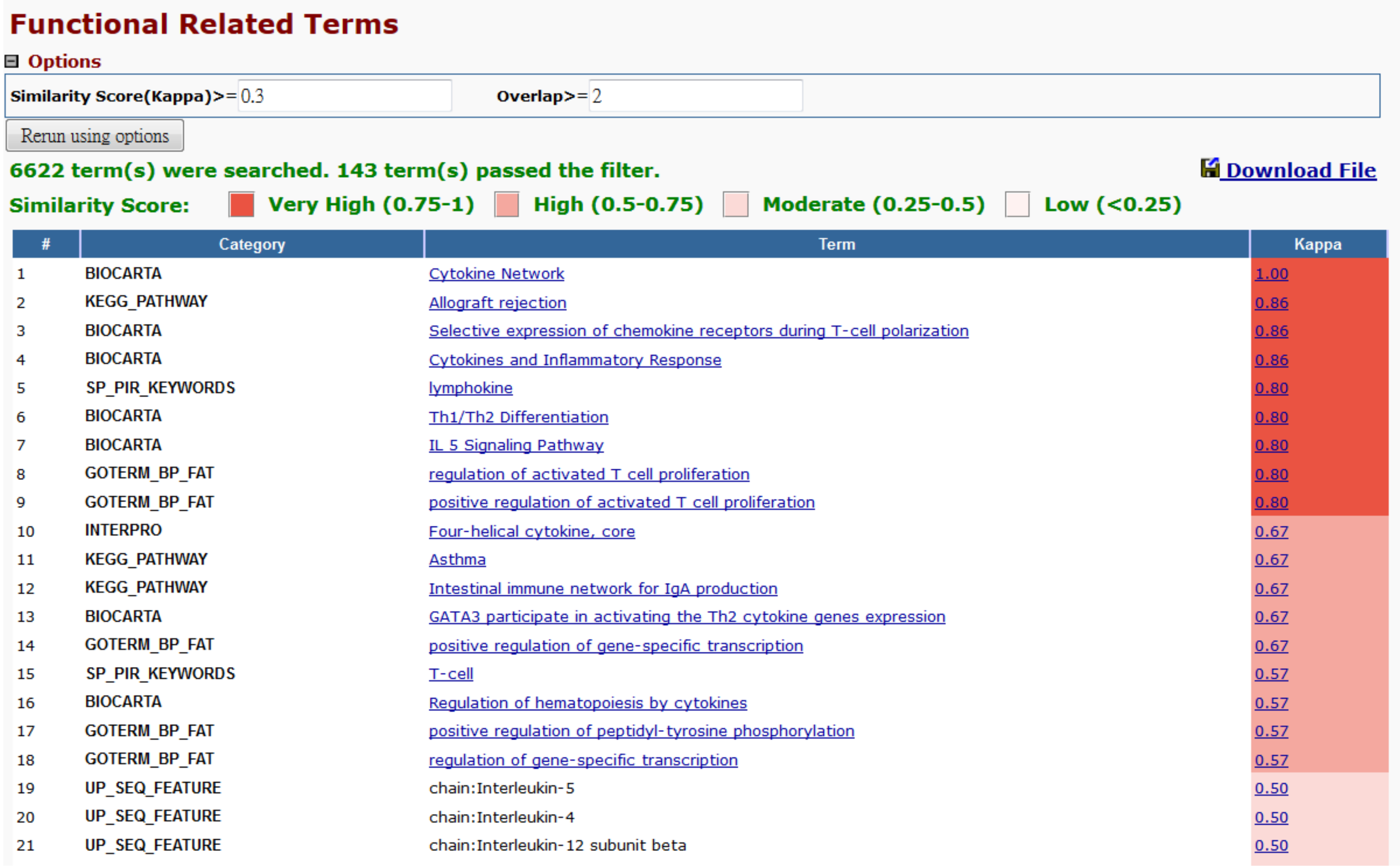
Gene Functional Classification Tool

Gene ID Conversion Tool

Gene Name Batch Viewer

RT (Related Term)

Any given gene is associating with a set of annotation terms. If genes share similar set of those terms, they are most likely involved in similar biological mechanisms. The algorithm adopts kappa statistics to quantitatively measure the degree of the agreement how genes share the similar annotation terms. Kappa result ranges from 0 to 1. The higher the value of Kappa, the stronger the agreement.
Any a biological process/term coming from all functional categories listed in DAVID.



Functional Related Terms

⊟ Options

Similarity Score(Kappa)>= 0.3 Overlap>= 2

6622 term(s) were searched. 143 term(s) passed the filter.

Similarity Score: 🟥 Very High (0.75-1) ▨ High (0.5-0.75) □ Moderate (0.25-0.5) □ Low (<0.25)

#	Category	Term	Kappa
1	BIOCARTA	Cytokine Network	1.00
2	KEGG_PATHWAY	Allograft rejection	0.86
3	BIOCARTA	Selective expression of chemokine receptors during T-cell polarization	0.86
4	BIOCARTA	Cytokines and Inflammatory Response	0.86
5	SP_PIR_KEYWORDS	lymphokine	0.80
6	BIOCARTA	Th1/Th2 Differentiation	0.80
7	BIOCARTA	IL 5 Signaling Pathway	0.80
8	GOTERM_BP_FAT	regulation of activated T cell proliferation	0.80
9	GOTERM_BP_FAT	positive regulation of activated T cell proliferation	0.80
10	INTERPRO	Four-helical cytokine, core	0.67
11	KEGG_PATHWAY	Asthma	0.67
12	KEGG_PATHWAY	Intestinal immune network for IgA production	0.67
13	BIOCARTA	GATA3 participate in activating the Th2 cytokine genes expression	0.67
14	GOTERM_BP_FAT	positive regulation of gene-specific transcription	0.67
15	SP_PIR_KEYWORDS	T-cell	0.57
16	BIOCARTA	Regulation of hematopoiesis by cytokines	0.57
17	GOTERM_BP_FAT	positive regulation of peptidyl-tyrosine phosphorylation	0.57
18	GOTERM_BP_FAT	regulation of gene-specific transcription	0.57
19	UP_SEQ_FEATURE	chain:Interleukin-5	0.50
20	UP_SEQ_FEATURE	chain:Interleukin-4	0.50
21	UP_SEQ_FEATURE	chain:Interleukin-12 subunit beta	0.50

THANK YOU

ANY QUESTIONS?