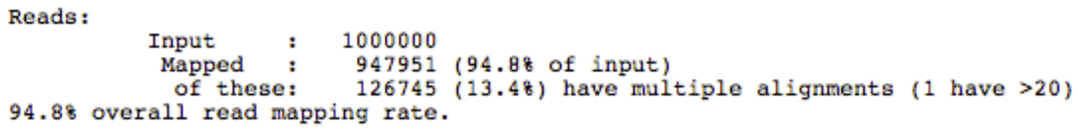
# Analysis of RNA-seq data : answers to questions

Céline Keime
keime@igbmc.fr

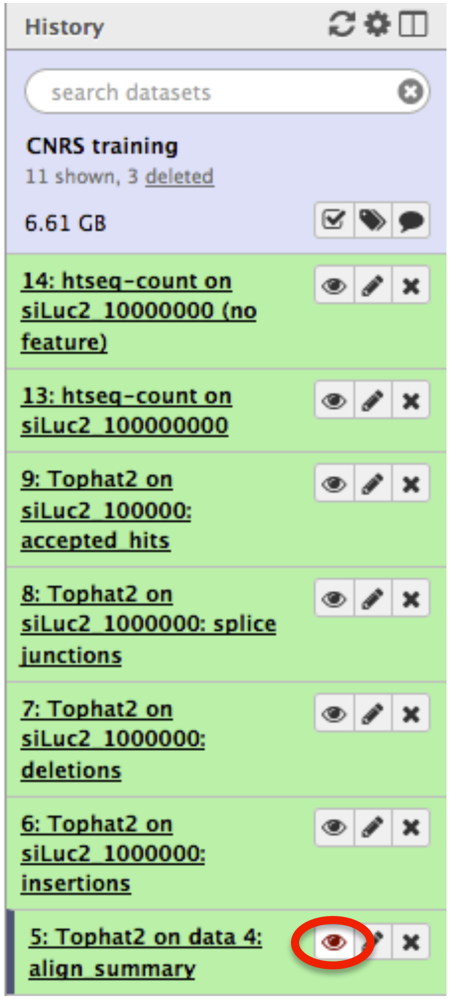
# Question 1

- Number of uniquely aligned reads



```
Reads:
          Input     :   1000000
           Mapped   :    947951 (94.8% of input)
            of these:    126745 (13.4%) have multiple alignments (1 have >20)
94.8% overall read mapping rate.
```

Number of uniquely mapped reads
= Number of mapped reads –
number of reads with multiple alignments
= 947951 – 126745 = 821206



History

search datasets

CNRS training

11 shown, 3 deleted

6.61 GB

14: htseq-count on siLuc2_10000000 (no feature)

13: htseq-count on siLuc2_100000000

9: Tophat2 on siLuc2_100000: accepted_hits

8: Tophat2 on siLuc2_1000000: splice junctions

7: Tophat2 on siLuc2_1000000: deletions

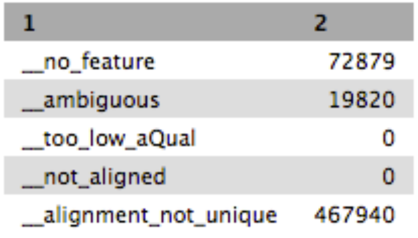
6: Tophat2 on siLuc2_1000000: insertions

5: Tophat2 on data 4: align_summary

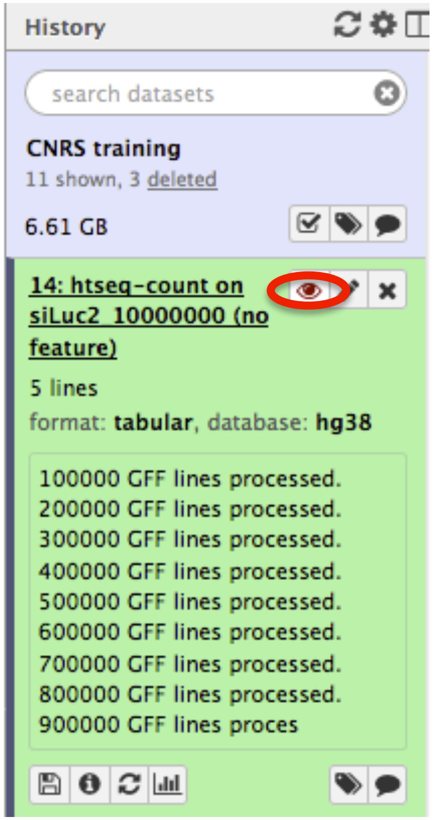
# Question 1

- No feature reads
  - Number
    - 72879
  - Proportion :
    - 72879*100/821206 = 8.87

- Ambiguous reads
  - Number
    - 19820
  - Proportion
    - 19820*100/821206 = 2.41



| 1 | 2 |
|---|---|
| __no_feature | 72879 |
| __ambiguous | 19820 |
| __too_low_aQual | 0 |
| __not_aligned | 0 |
| __alignment_not_unique | 467940 |



History

**CNRS training**

11 shown, 3 deleted

6.61 GB

**14: htseq-count on siLuc2_10000000 (no feature)**

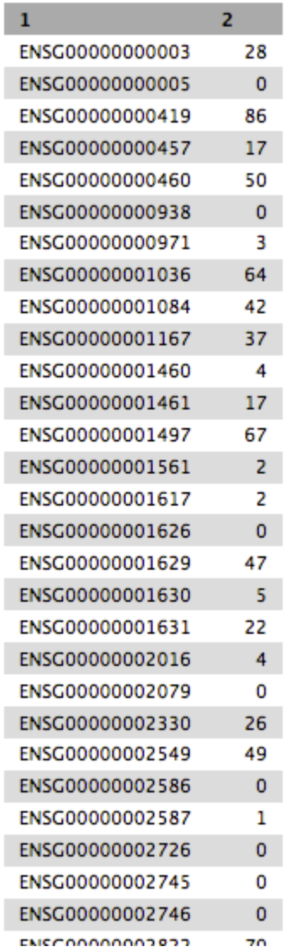
5 lines

format: **tabular**, database: **hg38**

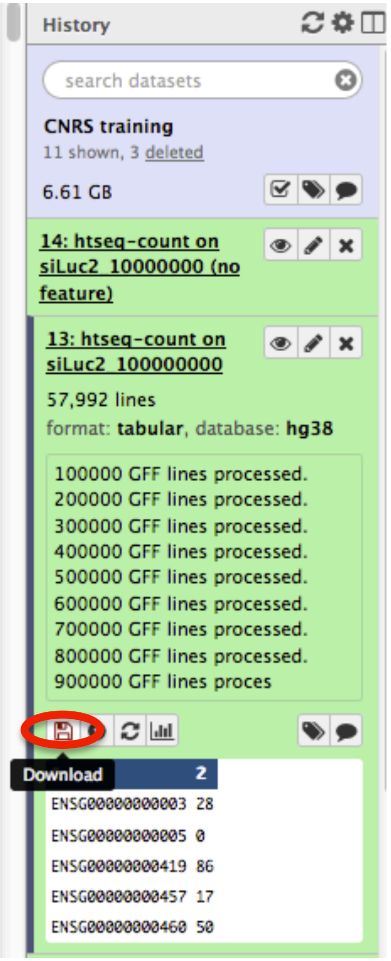
```
100000 GFF lines processed.
200000 GFF lines processed.
300000 GFF lines processed.
400000 GFF lines processed.
500000 GFF lines processed.
600000 GFF lines processed.
700000 GFF lines processed.
800000 GFF lines processed.
900000 GFF lines proces
```

# Question 1

- Number of assigned reads



| 1 | 2 |
|---|---|
| ENSG00000000003 | 28 |
| ENSG00000000005 | 0 |
| ENSG00000000419 | 86 |
| ENSG00000000457 | 17 |
| ENSG00000000460 | 50 |
| ENSG00000000938 | 0 |
| ENSG00000000971 | 3 |
| ENSG00000001036 | 64 |
| ENSG00000001084 | 42 |
| ENSG00000001167 | 37 |
| ENSG00000001460 | 4 |
| ENSG00000001461 | 17 |
| ENSG00000001497 | 67 |
| ENSG00000001561 | 2 |
| ENSG00000001617 | 2 |
| ENSG00000001626 | 0 |
| ENSG00000001629 | 47 |
| ENSG00000001630 | 5 |
| ENSG00000001631 | 22 |
| ENSG00000002016 | 4 |
| ENSG00000002079 | 0 |
| ENSG00000002330 | 26 |
| ENSG00000002549 | 49 |
| ENSG00000002586 | 0 |
| ENSG00000002587 | 1 |
| ENSG00000002726 | 0 |
| ENSG00000002745 | 0 |
| ENSG00000002746 | 0 |



CNRS training

11 shown, 3 deleted

6.61 GB

14: htseq-count on siLuc2_10000000 (no feature)

13: htseq-count on siLuc2_100000000

57,992 lines

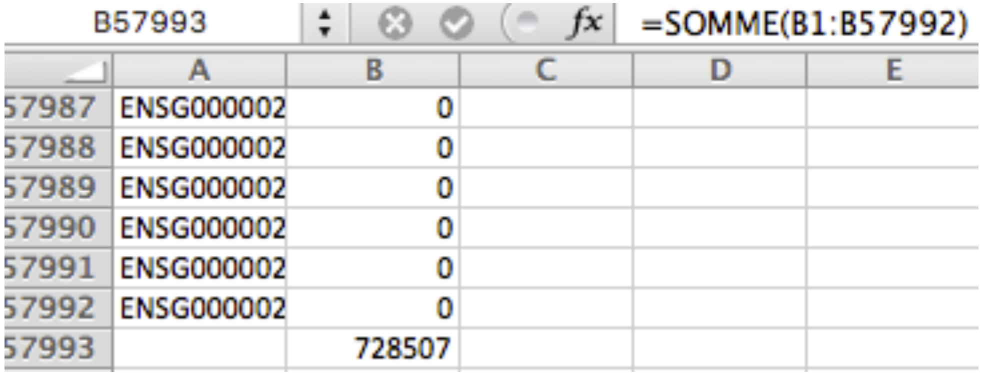
format: **tabular**, database: **hg38**

```
100000 GFF lines processed.
200000 GFF lines processed.
300000 GFF lines processed.
400000 GFF lines processed.
500000 GFF lines processed.
600000 GFF lines processed.
700000 GFF lines processed.
800000 GFF lines processed.
900000 GFF lines proces
```

```
2
ENSG00000000003 28
ENSG00000000005 0
ENSG00000000419 86
ENSG00000000457 17
ENSG00000000460 50
```

# Question 1

- Number of assigned reads
  - Open the downloaded file with excel
  - Calculate the total number of reads in the second column



B57993 =SOMME(B1:B57992)

|  | A | B | C | D | E |
|---|---|---|---|---|---|
| 57987 | ENSG000002 | 0 |  |  |  |
| 57988 | ENSG000002 | 0 |  |  |  |
| 57989 | ENSG000002 | 0 |  |  |  |
| 57990 | ENSG000002 | 0 |  |  |  |
| 57991 | ENSG000002 | 0 |  |  |  |
| 57992 | ENSG000002 | 0 |  |  |  |
| 57993 |  | 728507 |  |  |  |

→ Number of assigned reads = 728507

→ Proportion of assigned reads = 728507*100/821206 = 88.71

Or

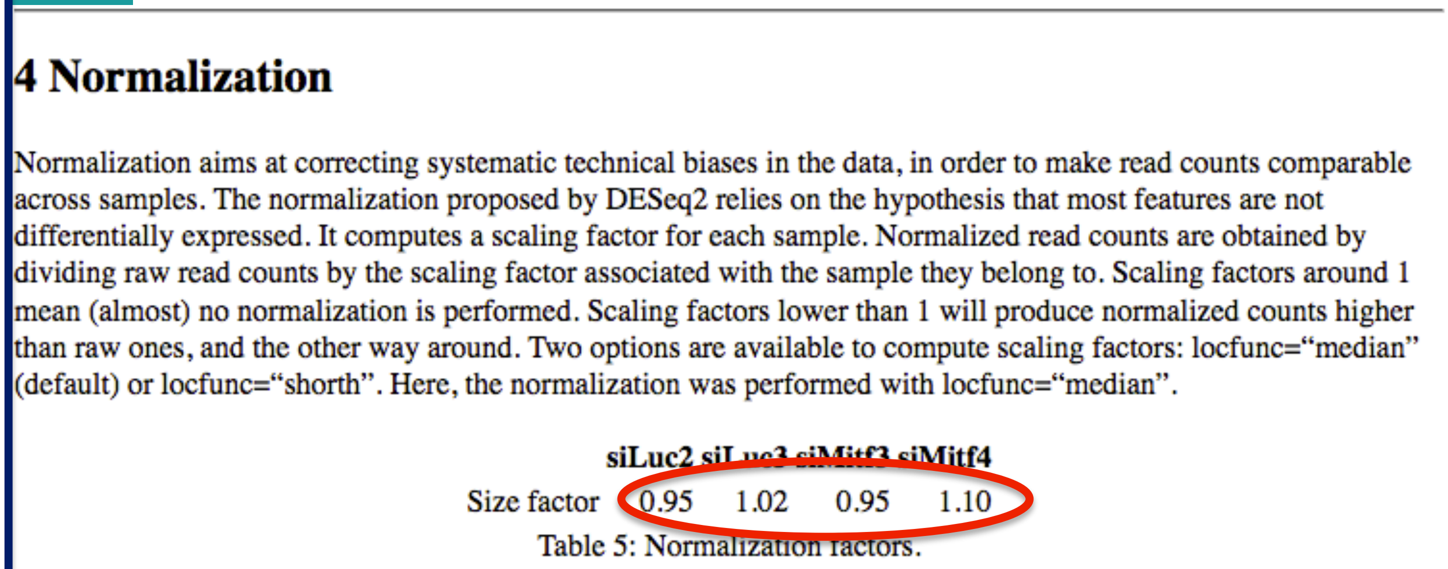
Number of assigned reads
= number of uniquely aligned reads – number of no feature reads – number of ambiguous reads
= 821206 – 72879 – 19820 = 728507

# Question 1

- Proportion of reads among uniquely aligned reads
  - Assigned : 88.71%
  - No feature : 8.87%
  - Ambiguous : 2.41%

# Question 2

- Values of normalization factors for Mitf dataset

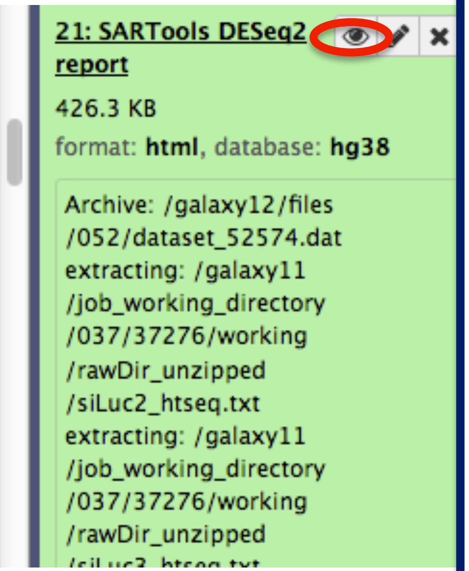


## 4 Normalization

Normalization aims at correcting systematic technical biases in the data, in order to make read counts comparable across samples. The normalization proposed by DESeq2 relies on the hypothesis that most features are not differentially expressed. It computes a scaling factor for each sample. Normalized read counts are obtained by dividing raw read counts by the scaling factor associated with the sample they belong to. Scaling factors around 1 mean (almost) no normalization is performed. Scaling factors lower than 1 will produce normalized counts higher than raw ones, and the other way around. Two options are available to compute scaling factors: locfunc="median" (default) or locfunc="shorth". Here, the normalization was performed with locfunc="median".

| | siLuc2 | siLuc3 | siMitf3 | siMitf4 |
|---|---|---|---|---|
| Size factor | 0.95 | 1.02 | 0.95 | 1.10 |

Table 5: Normalization factors.



21: SARTools DESeq2 report

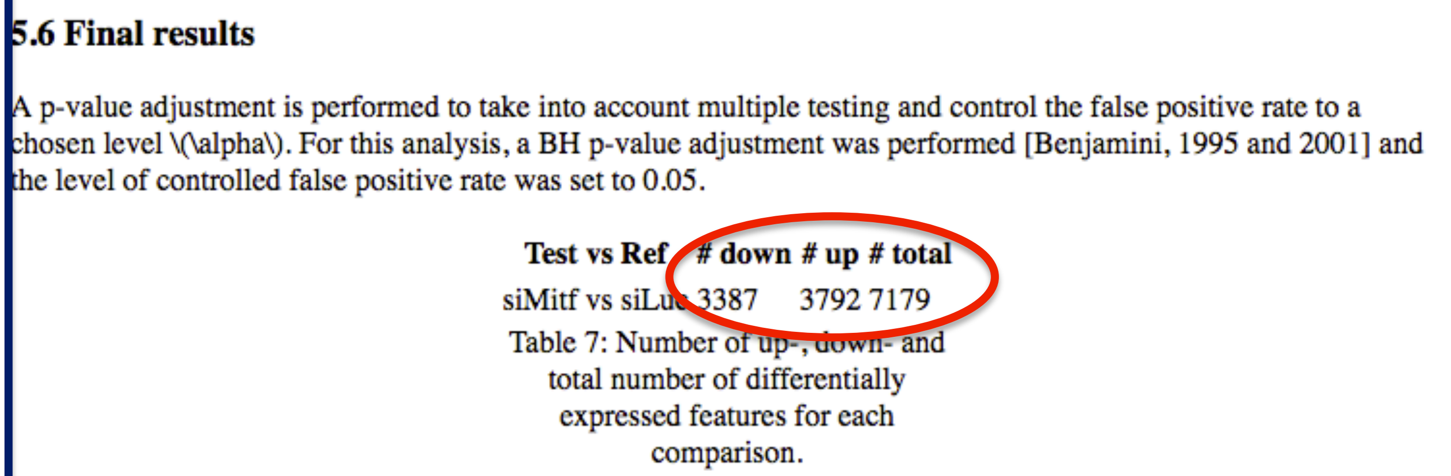
426.3 KB

format: html, database: hg38

```
Archive: /galaxy12/files
/052/dataset_52574.dat
extracting: /galaxy11
/job_working_directory
/037/37276/working
/rawDir_unzipped
/siLuc2_htseq.txt
extracting: /galaxy11
/job_working_directory
/037/37276/working
/rawDir_unzipped
/siLuc3_htseq.txt
```

# Question 3

- Number of significantly differentially expressed genes between siMitf and siLuc (FDR<0.05)

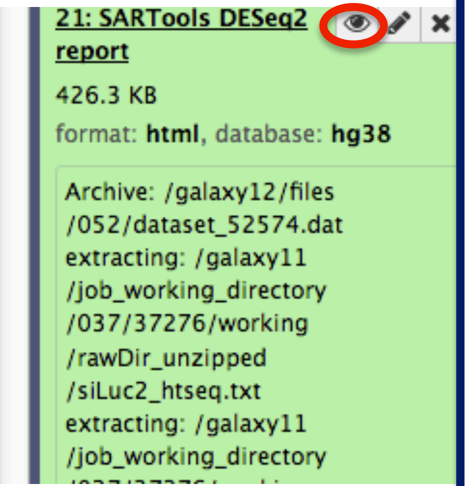


### 5.6 Final results

A p-value adjustment is performed to take into account multiple testing and control the false positive rate to a chosen level \(\alpha\). For this analysis, a BH p-value adjustment was performed [Benjamini, 1995 and 2001] and the level of controlled false positive rate was set to 0.05.

| Test vs Ref | # down | # up | # total |
|---|---|---|---|
| siMitf vs siLuc | 3387 | 3792 | 7179 |

Table 7: Number of up-, down- and total number of differentially expressed features for each comparison.



21: SARTools DESeq2 report

426.3 KB

format: **html**, database: **hg38**

```
Archive: /galaxy12/files
/052/dataset_52574.dat
extracting: /galaxy11
/job_working_directory
/037/37276/working
/rawDir_unzipped
/siLuc2_htseq.txt
extracting: /galaxy11
/job_working_directory
```

→ 7179 significantly differentially expressed genes

→ 3387 genes significantly under-exressed in siMitf vs siLuc

→ 3792 genes significantly over-expressed in siMitf vs siLuc