



# Analysis of RNA-seq data : answers to questions

Céline Keime  
keime@igbmc.fr

# Question 1

## ■ Number of uniquely aligned reads

```
Reads:
      Input      : 1000000
      Mapped     : 983595 (98.4% of input)
      of these:  88965 ( 9.0%) have multiple alignments (434 have >20)
98.4% overall read mapping rate.
```

Number of uniquely mapped reads  
= Number of mapped reads –  
number of reads with multiple alignments  
= 983595 – 88965 = 894630

History

search datasets

RNAseq1709  
20 shown  
213.02 MB

- 10: htseq-count on siLuc2
- 9: htseq-count on data 7 (no feature)
- 8: htseq-count on data 7
- 7: TopHat 2 on data 2 and data 1: accepted hits
- 6: TopHat 2 on data 2 and data 1: splice junctions
- 5: TopHat 2 on data 2 and data 1: deletions
- 4: TopHat 2 on data 2 and data 1: insertions
- 3: TopHat 2 on data 2 and data 1: align summary

# Question 1

## ■ No feature reads

- Number
  - 71322
- Proportion :
  - $71322 * 100 / 894630 = 7.97$

1	2
__no_feature	71322
__ambiguous	23562
__too_low_aQual	0
__not_aligned	0
__alignment_not_unique	294859

## ■ Ambiguous reads

- Number
  - 23562
- Proportion
  - $23562 * 100 / 894630 = 2.63$

History

search datasets

RNAseq1709  
20 shown  
213.02 MB

10: htseq-count on siLuc2

9: htseq-count on data 7 (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 processed.  
10

1	2
__no_feature	71322
__ambiguous	23562
__too_low_aQual	0
__not_aligned	0
__alignment_not_unique	294859

# Question 1

## ■ Number of assigned reads

1	2
ENSG00000000003	31
ENSG00000000005	0
ENSG00000000419	89
ENSG00000000457	18
ENSG00000000460	55
ENSG00000000938	0
ENSG00000000971	3
ENSG0000001036	66
ENSG0000001084	48
ENSG0000001167	38
ENSG0000001460	4
ENSG0000001461	17
ENSG0000001497	70
ENSG0000001561	2
ENSG0000001617	2
ENSG0000001626	0
ENSG0000001629	52
ENSG0000001630	5
ENSG0000001631	25
ENSG0000002016	5
ENSG0000002079	0
ENSG0000002330	27
ENSG0000002549	70
ENSG0000002586	122
ENSG0000002587	1
ENSG0000002726	0
ENSG0000002745	0

History

search datasets

**RNAseq1709**  
20 shown  
213.02 MB

**8: htseq-count on data 7**  
57,992 lines  
format: **tabular**, database: **hg38**

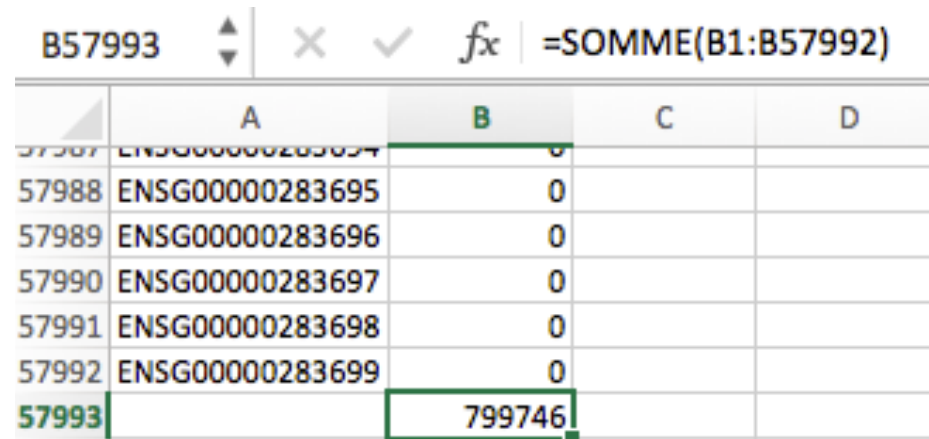
100000 GFF lines processed.  
200000 GFF lines processed.  
300000 GFF lines processed.  
400000 GFF lines processed.  
500000 GFF lines processed.  
600000 GFF lines processed.  
700000 GFF lines processed.  
800000 GFF lines processed.  
900000 GFF lines processed.  
10

Download 2

ENSG00000000003 31  
ENSG00000000005 0  
ENSG00000000419 89  
ENSG00000000457 18  
ENSG00000000460 55

# Question 1

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



	A	B	C	D
57987	ENSG00000283694	0		
57988	ENSG00000283695	0		
57989	ENSG00000283696	0		
57990	ENSG00000283697	0		
57991	ENSG00000283698	0		
57992	ENSG00000283699	0		
57993		799746		

→ Number of assigned reads = 799746

→ Proportion of assigned reads =  $799746 * 100 / 894630 = 89.39$

Or

Number of assigned reads

= number of uniquely aligned reads – number of no feature reads – number of ambiguous reads

=  $894630 - 71322 - 23562 = 799746$

# Question 1

---

- Proportion of reads among uniquely aligned reads
  - Assigned : 89.39%
  - No feature : 7.97%
  - Ambiguous : 2.63%

# 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

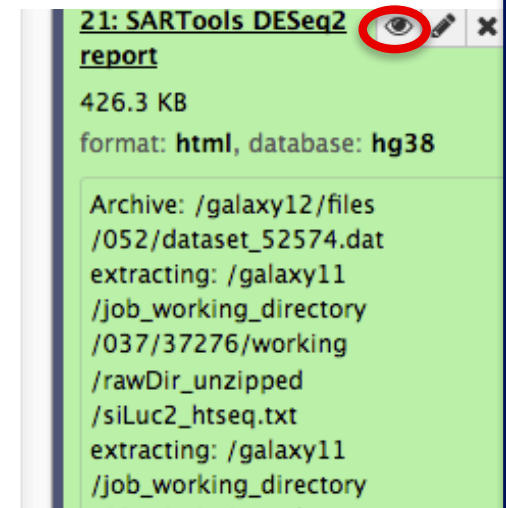
- Number of significantly differentially expressed genes between siMitf and siLuc (FDR<0.05)

## 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-expressed in siMitf vs siLuc
  - 3792 genes significantly over-expressed in siMitf vs siLuc