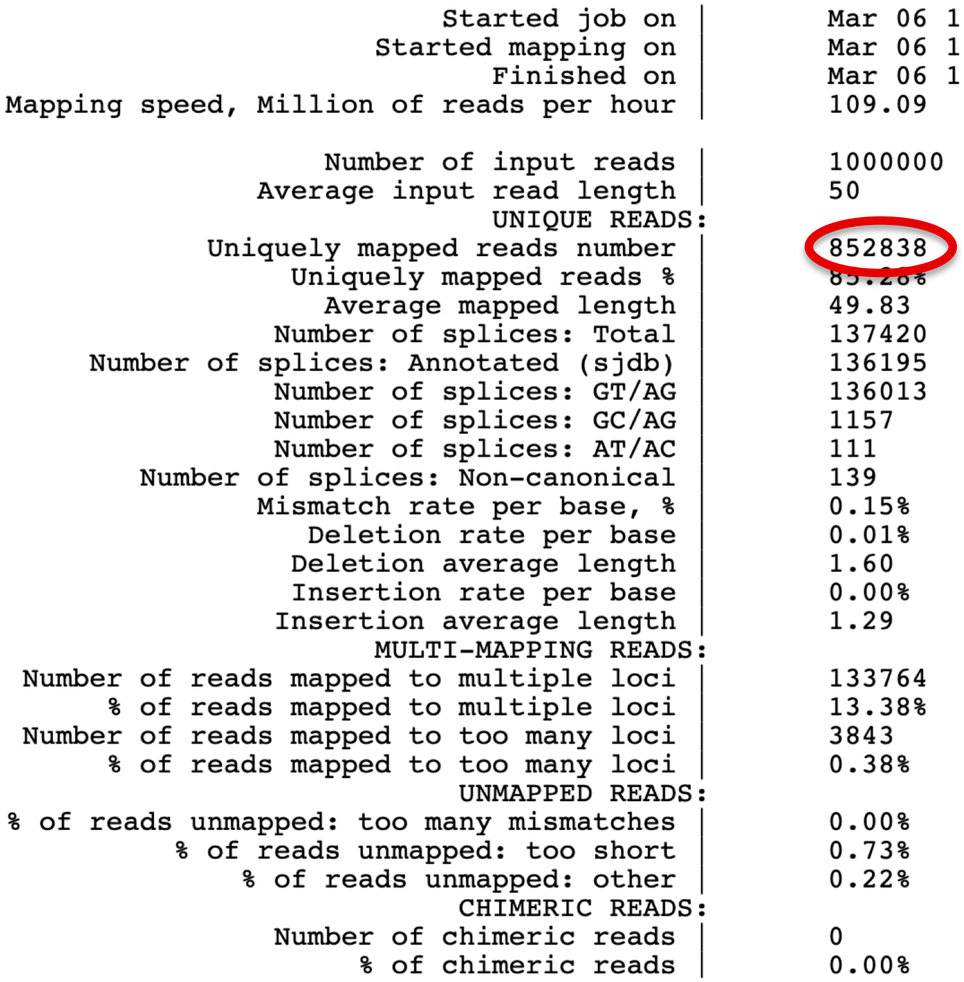
Analysis of RNA-seq data : answers to questions

Céline Keime
keime@igbmc.fr

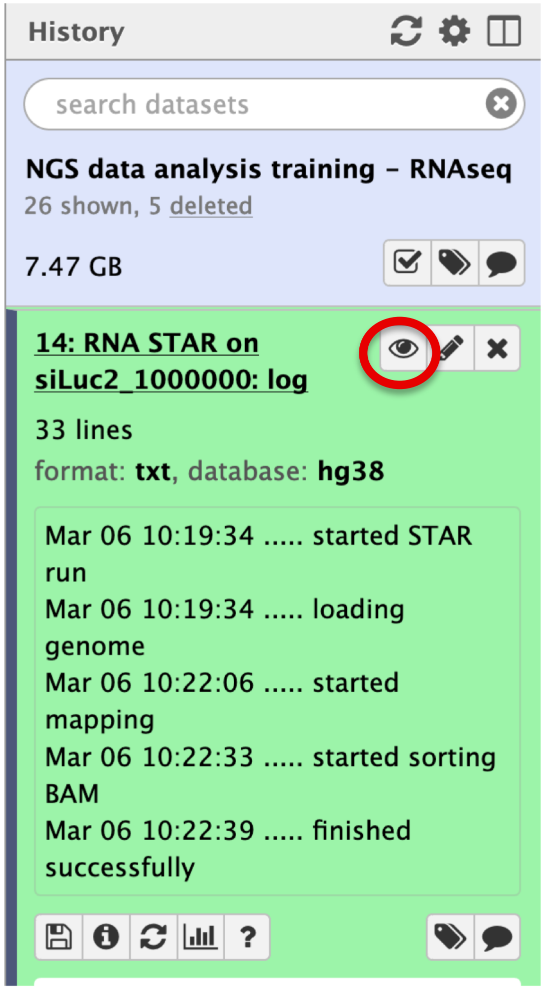
Question 1

- Number of uniquely mapped reads



```
                                 Started job on |	Mar 06 10:19:34
                             Started mapping on |	Mar 06 10:22:06
                                    Finished on |	Mar 06 10:22:39
       Mapping speed, Million of reads per hour |	109.09

                          Number of input reads |	1000000
                      Average input read length |	50
                                    UNIQUE READS:
                   Uniquely mapped reads number |	852838
                        Uniquely mapped reads % |	85.28%
                          Average mapped length |	49.83
                       Number of splices: Total |	137420
            Number of splices: Annotated (sjdb) |	136195
                       Number of splices: GT/AG |	136013
                       Number of splices: GC/AG |	1157
                       Number of splices: AT/AC |	111
               Number of splices: Non-canonical |	139
                      Mismatch rate per base, % |	0.15%
                         Deletion rate per base |	0.01%
                        Deletion average length |	1.60
                        Insertion rate per base |	0.00%
                       Insertion average length |	1.29
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |	133764
             % of reads mapped to multiple loci |	13.38%
        Number of reads mapped to too many loci |	3843
             % of reads mapped to too many loci |	0.38%
                                  UNMAPPED READS:
       % of reads unmapped: too many mismatches |	0.00%
                 % of reads unmapped: too short |	0.73%
                     % of reads unmapped: other |	0.22%
                                  CHIMERIC READS:
                       Number of chimeric reads |	0
                            % of chimeric reads |	0.00%
```



NGS data analysis training – RNAseq

26 shown, 5 deleted

7.47 GB

14: RNA STAR on siLuc2_1000000: log

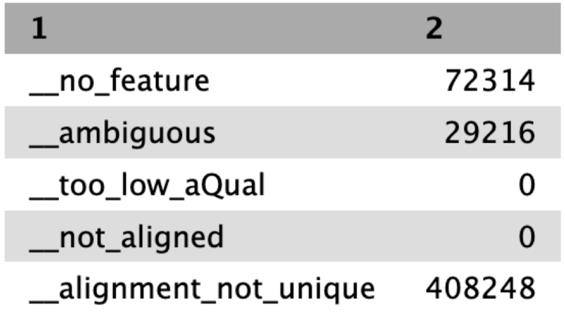
33 lines

format: **txt**, database: **hg38**

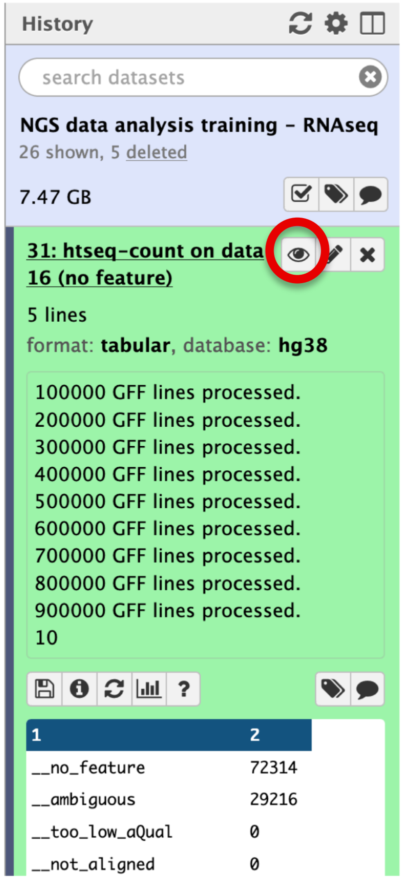
Mar 06 10:19:34 ..... started STAR run
Mar 06 10:19:34 ..... loading genome
Mar 06 10:22:06 ..... started mapping
Mar 06 10:22:33 ..... started sorting BAM
Mar 06 10:22:39 ..... finished successfully

Question 1

- No feature reads
  - Number
    - 72314
  - Proportion :
    - 72314*100/852838 = 8.48



1	2
__no_feature	72314
__ambiguous	29216
__too_low_aQual	0
__not_aligned	0
__alignment_not_unique	408248



NGS data analysis training – RNAseq

26 shown, 5 deleted

7.47 GB

31: htseq-count on data 16 (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	72314
__ambiguous	29216
__too_low_aQual	0
__not_aligned	0

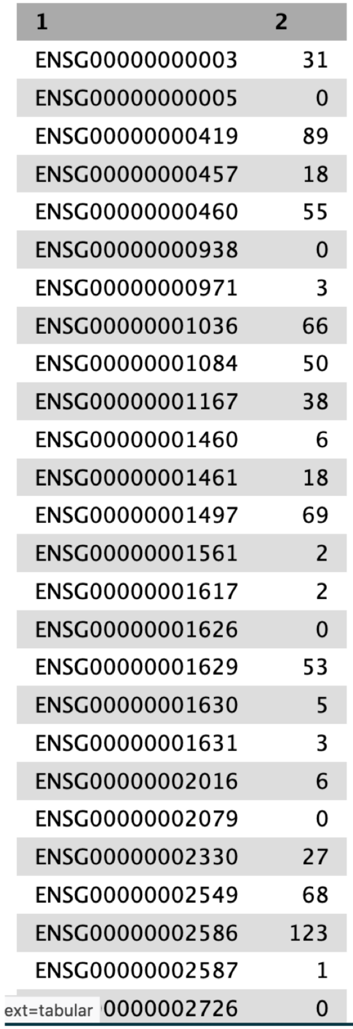
- Ambiguous reads
  - Number
    - 29216
  - Proportion
    - 29216*100/852838 = 3.43

Question 1

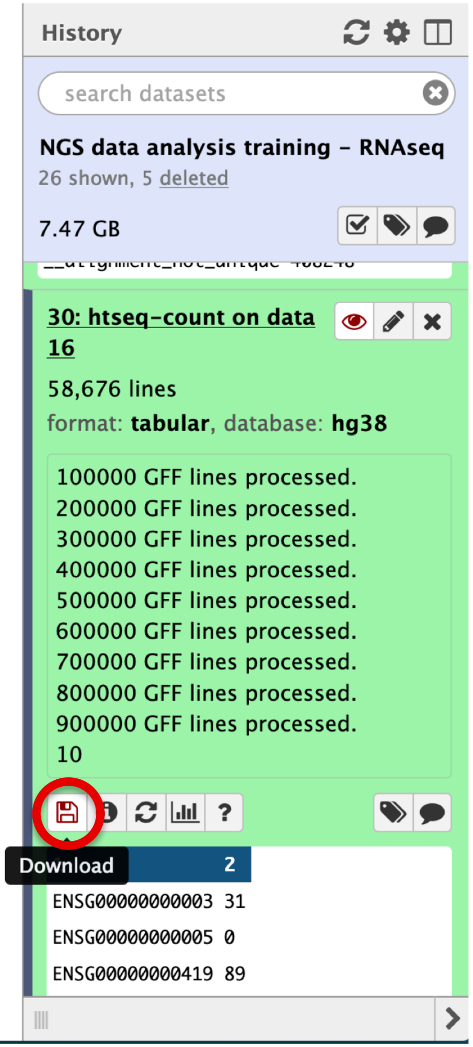
- Proportion of reads among uniquely aligned reads
  - Assigned : 100-8.48-3.43= 88.09 %
  - No feature : 8.48 %
  - Ambiguous : 3.43 %

Question 1

- Number of assigned reads



1	2
ENSG00000000003	31
ENSG00000000005	0
ENSG00000000419	89
ENSG00000000457	18
ENSG00000000460	55
ENSG00000000938	0
ENSG00000000971	3
ENSG00000001036	66
ENSG00000001084	50
ENSG00000001167	38
ENSG00000001460	6
ENSG00000001461	18
ENSG00000001497	69
ENSG00000001561	2
ENSG00000001617	2
ENSG00000001626	0
ENSG00000001629	53
ENSG00000001630	5
ENSG00000001631	3
ENSG00000002016	6
ENSG00000002079	0
ENSG00000002330	27
ENSG00000002549	68
ENSG00000002586	123
ENSG00000002587	1
ext=tabular 0000002726	0



NGS data analysis training – RNAseq
26 shown, 5 deleted
7.47 GB

__alignment_not_unique 408248

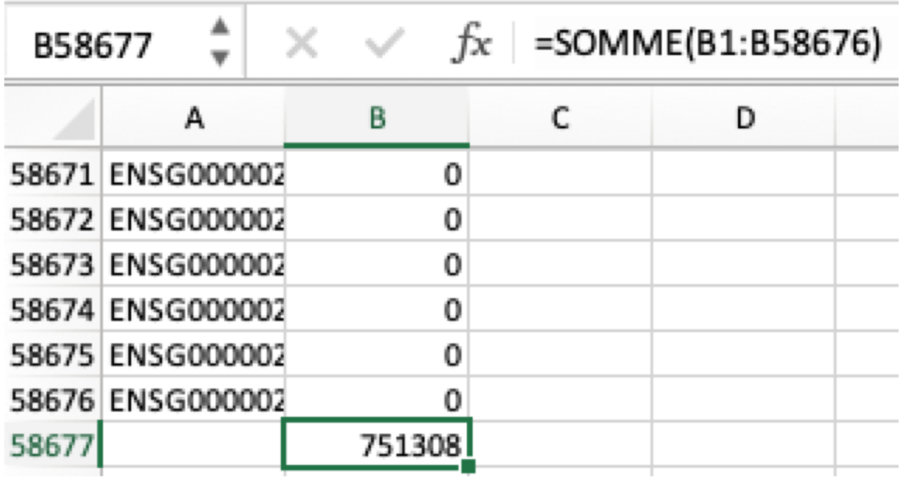
30: htseq-count on data 16
58,676 lines
format: tabular, database: hg38

```
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
```

```
              2
ENSG00000000003 31
ENSG00000000005 0
ENSG00000000419 89
```

Question 1

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



B58677 *fx* =SOMME(B1:B58676)

	A	B	C	D
58671	ENSG000002	0		
58672	ENSG000002	0		
58673	ENSG000002	0		
58674	ENSG000002	0		
58675	ENSG000002	0		
58676	ENSG000002	0		
58677		751308		

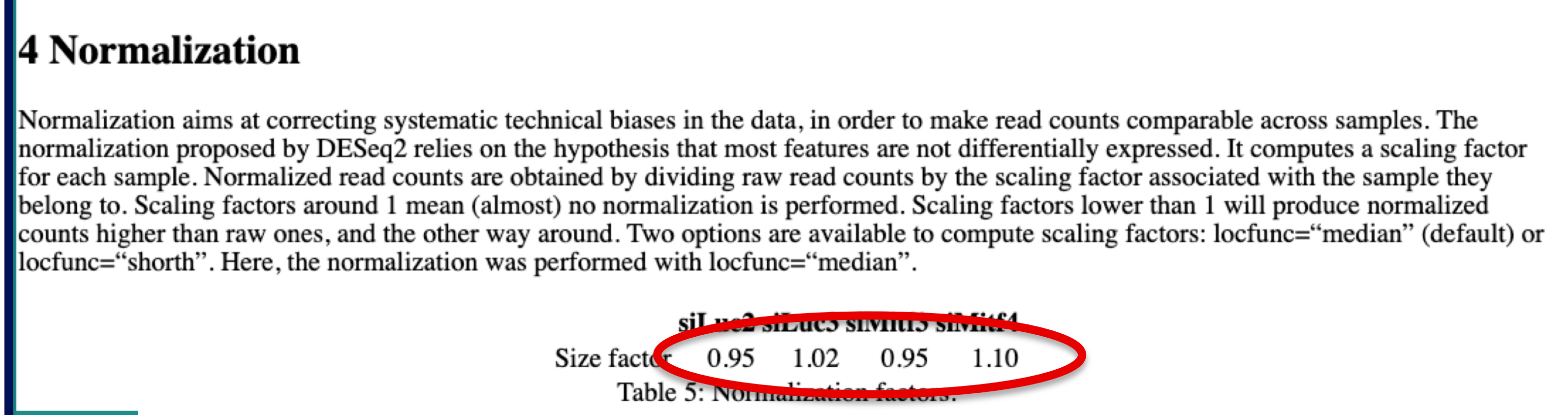
→ Number of assigned reads = 751308

→ Proportion of assigned reads = 751308 *100/852838 = 88.09

Number of assigned reads
= number of uniquely aligned reads – number of no feature reads – number of ambiguous reads
= 852838 – 72314 – 29216 = 751308

Question 2

- Values of normalization factors for Mitf dataset

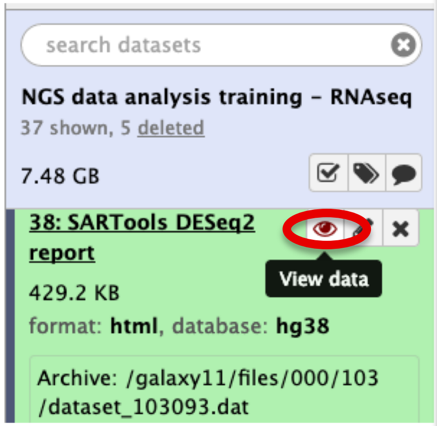


4 Normalization

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.



NGS data analysis training – RNAseq

37 shown, 5 deleted

7.48 GB

38: SARTools DESeq2 report

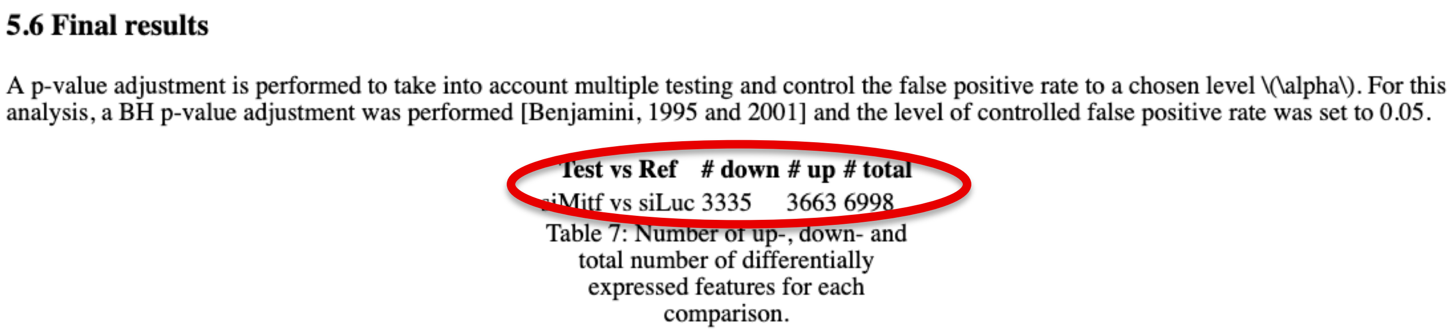
429.2 KB

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

Archive: /galaxy11/files/000/103
/dataset_103093.dat

Question 3

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

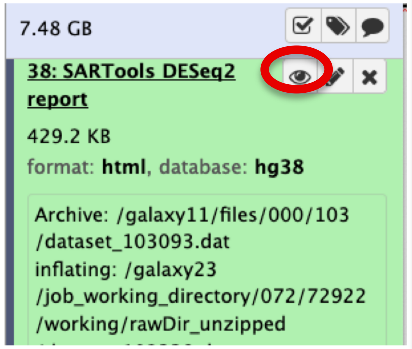


5.6 Final results

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	3335	3663	6998

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



7.48 GB

38: SARTools DESeq2 report

429.2 KB

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

Archive: /galaxy11/files/000/103/dataset_103093.dat
inflating: /galaxy23/job_working_directory/072/72922/working/rawDir_unzipped

→ 6998 significantly differentially expressed genes

→ 3335 genes significantly under-exressed in siMitf vs siLuc

→ 3663 genes significantly over-expressed in siMitf vs siLuc