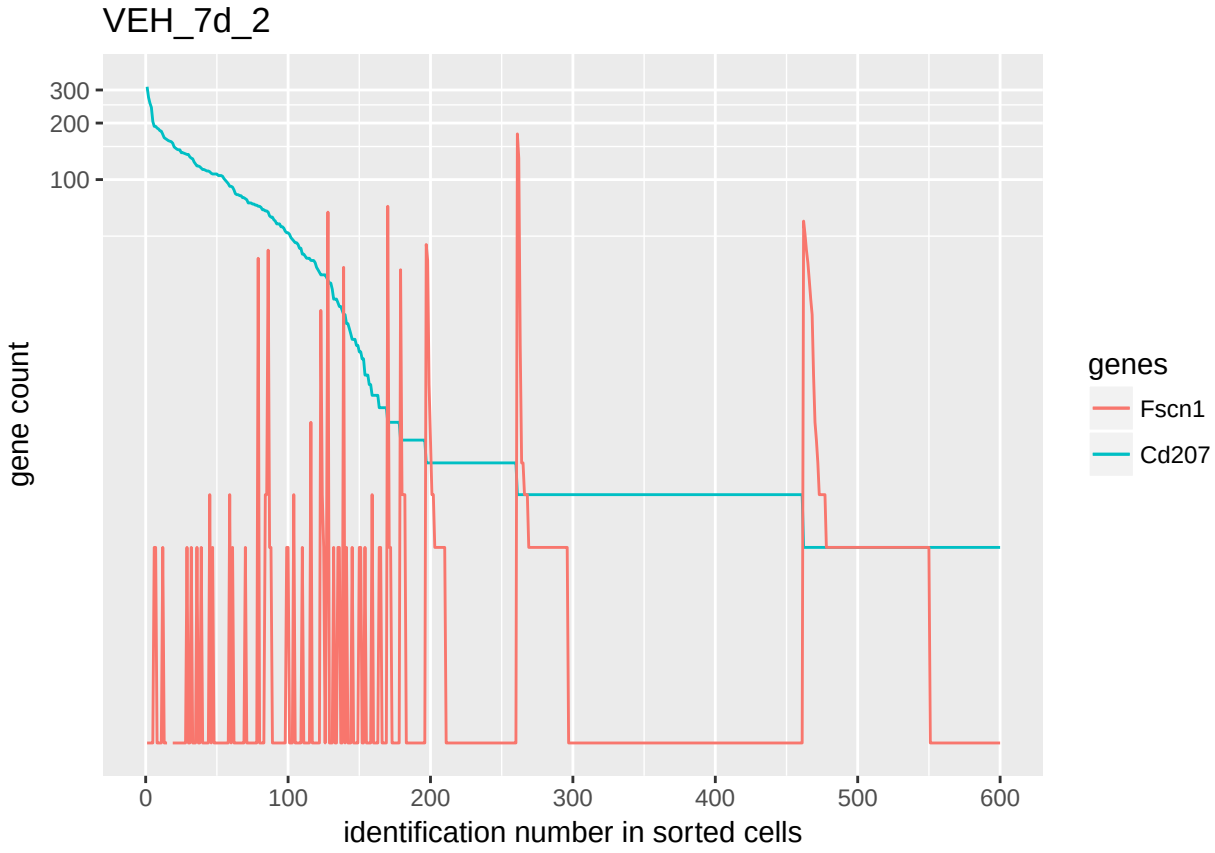
old and new markers

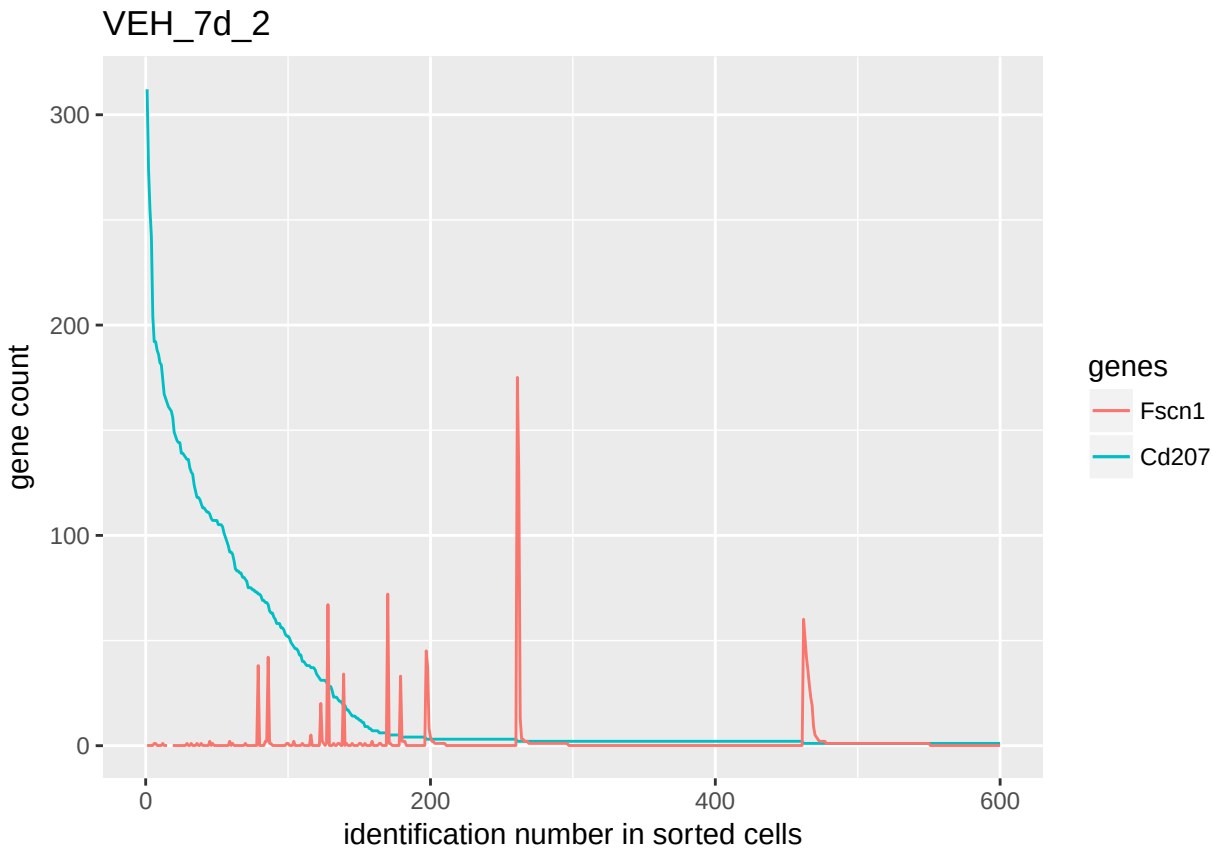
Christoph Pahl

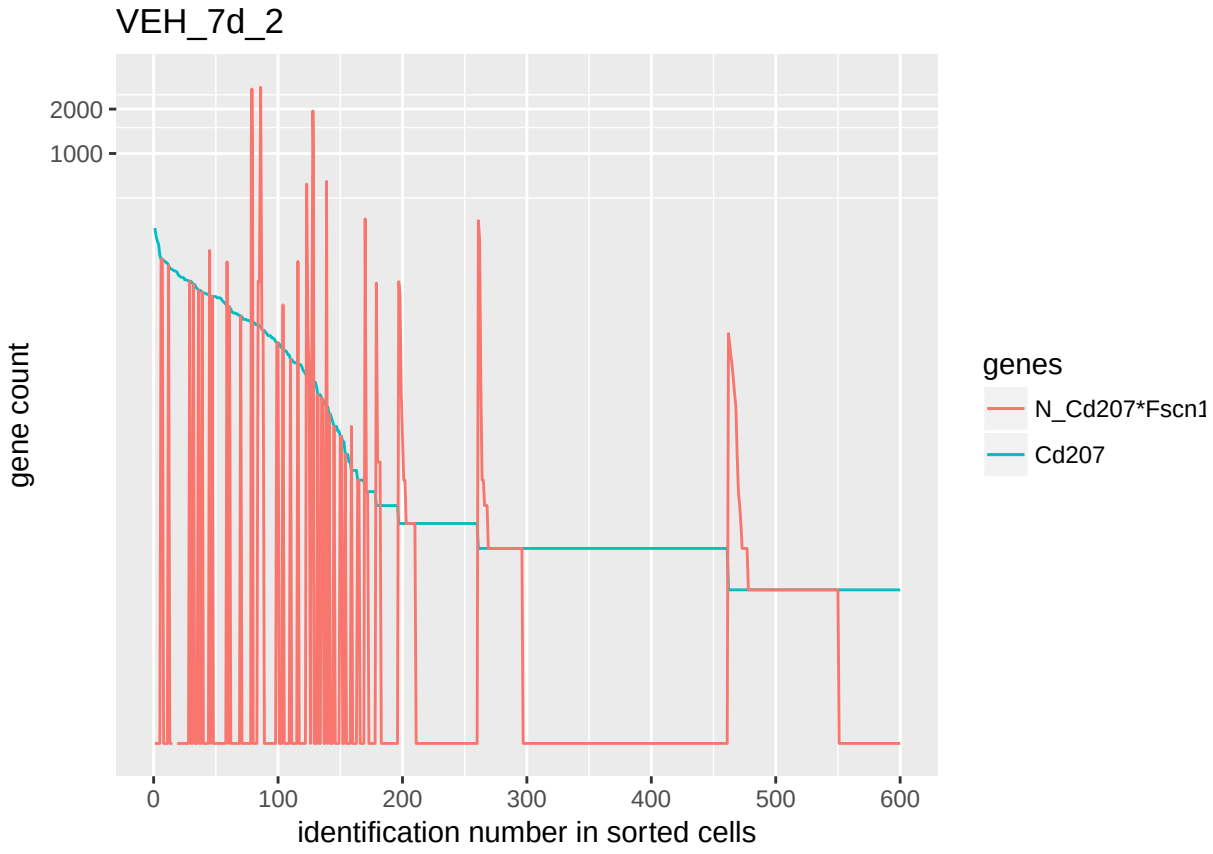
October 12 2018

Contents

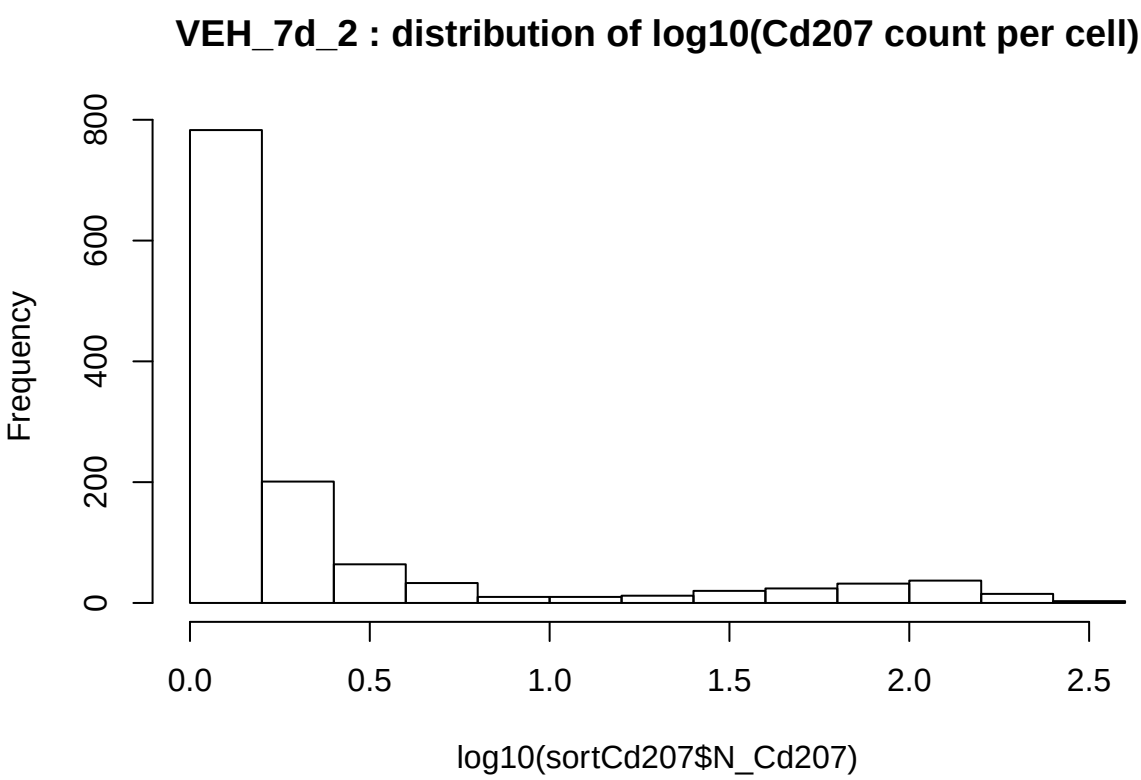
Port last Week's Python Analysis to R: Reproduce Plots




VEH_7d_2
300
200
100
0
gene count
0
200
400
600
identification number in sorted cells
genes
Fscn1
Cd207


VEH_7d_2
2000
1000
gene count
0
100
200
300
400
500
600
identification number in sorted cells
genes
N_Cd207*Fscn1
Cd207

Histogramming Gene Counts per Cell



A linear x-axis gives not enough detail, so I'm histogramming the logarithm of the respective gene count (a histogram with logarithmic axis would need much hacking).

Mean Values of Cd207 and Fscn1 Counts, their Correlation:

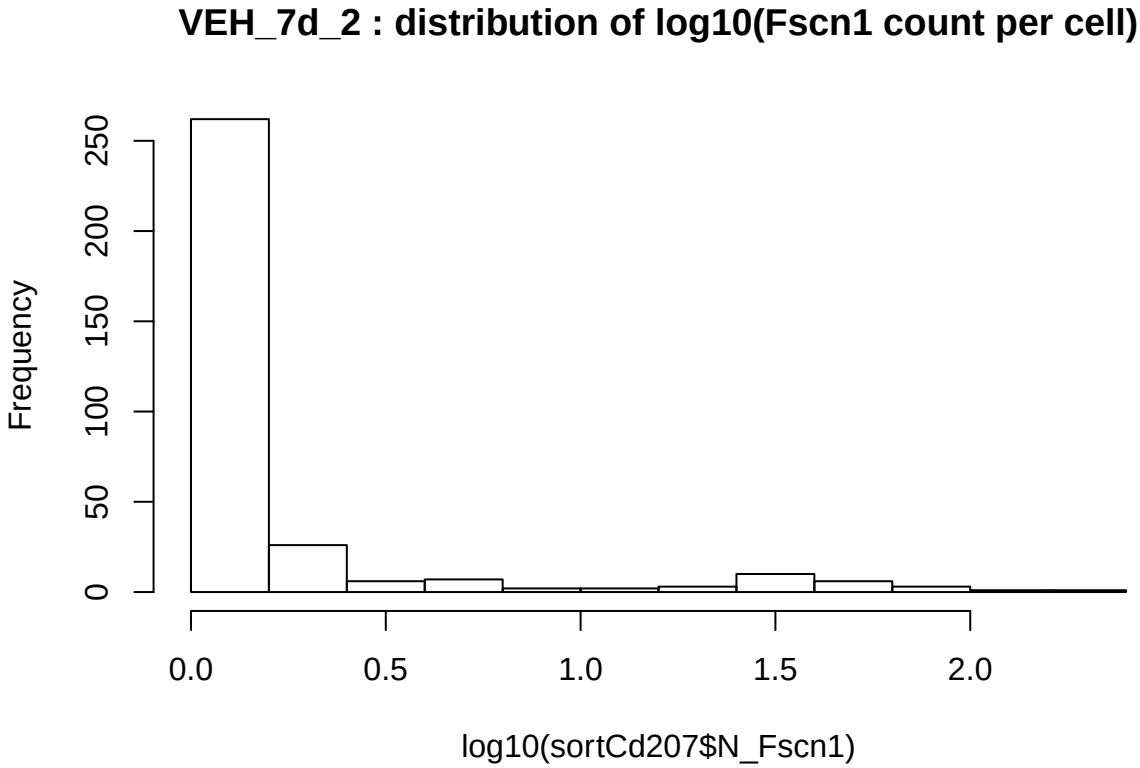
```
## [1] 4.875966
```

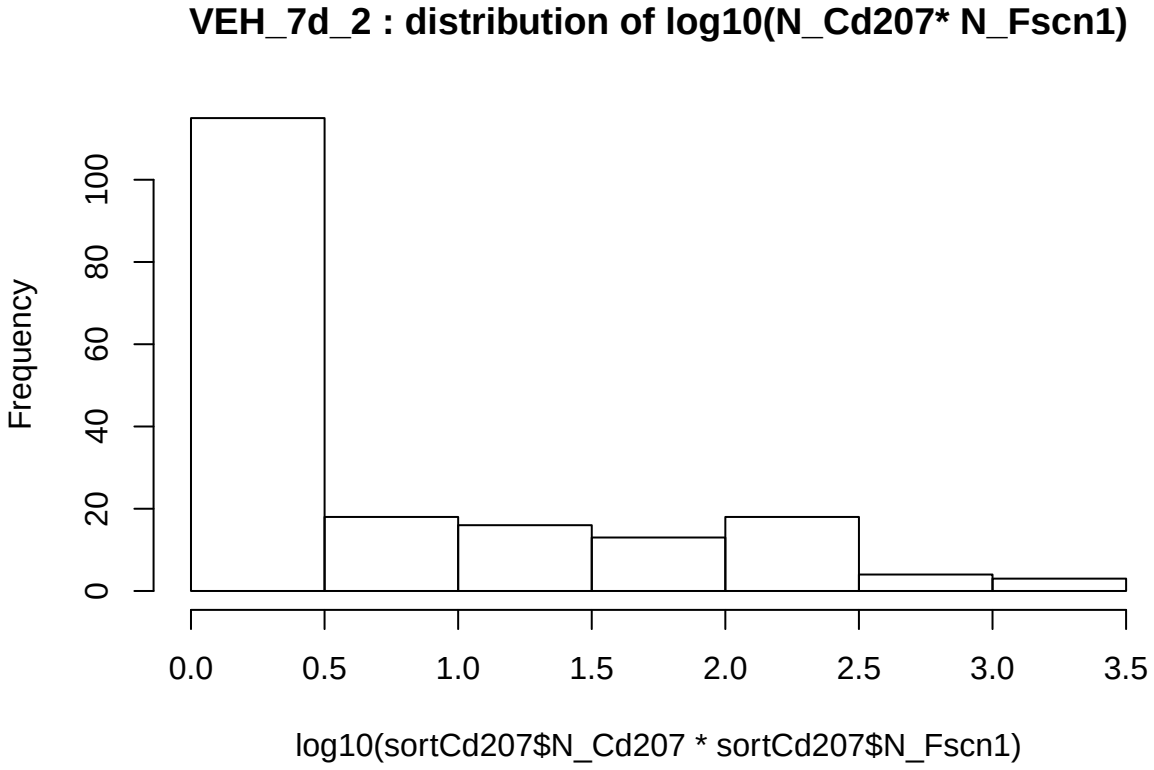
```
## [1] 0.5470172
```

```
##            N_Cd207    N_Fscn1
## N_Cd207 1.00000000 0.01615472
## N_Fscn1 0.01615472 1.00000000
```

Raymond asked for Distribution of 1's

$10^{0.2} \approx 1.6$, so the values of one constitute the first bin. From the first plots (with logarithmic y axis !) he also proposed a cut at #Cd207 ~ 200 to eliminate Langerhans cells, this corresponds to a log10 value of 2.3 at the far right end of the histogram. The above histogram might rather propose a cut value of $10^1 = 10$?





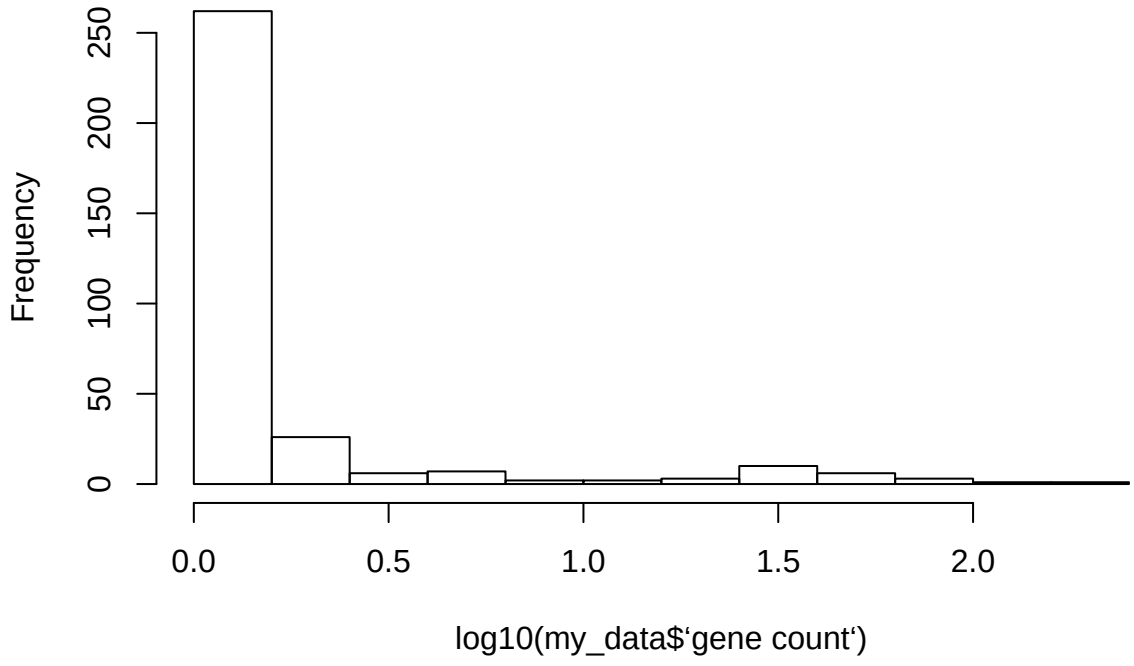
Histogramming New Markers

Raymond: Here is what else I think should be excluded.

1. Any cells with high KRTDAP or high KRT10, KRT1, KRT5, or KRT14 (these are keratinocyte contaminants and not immune cells)

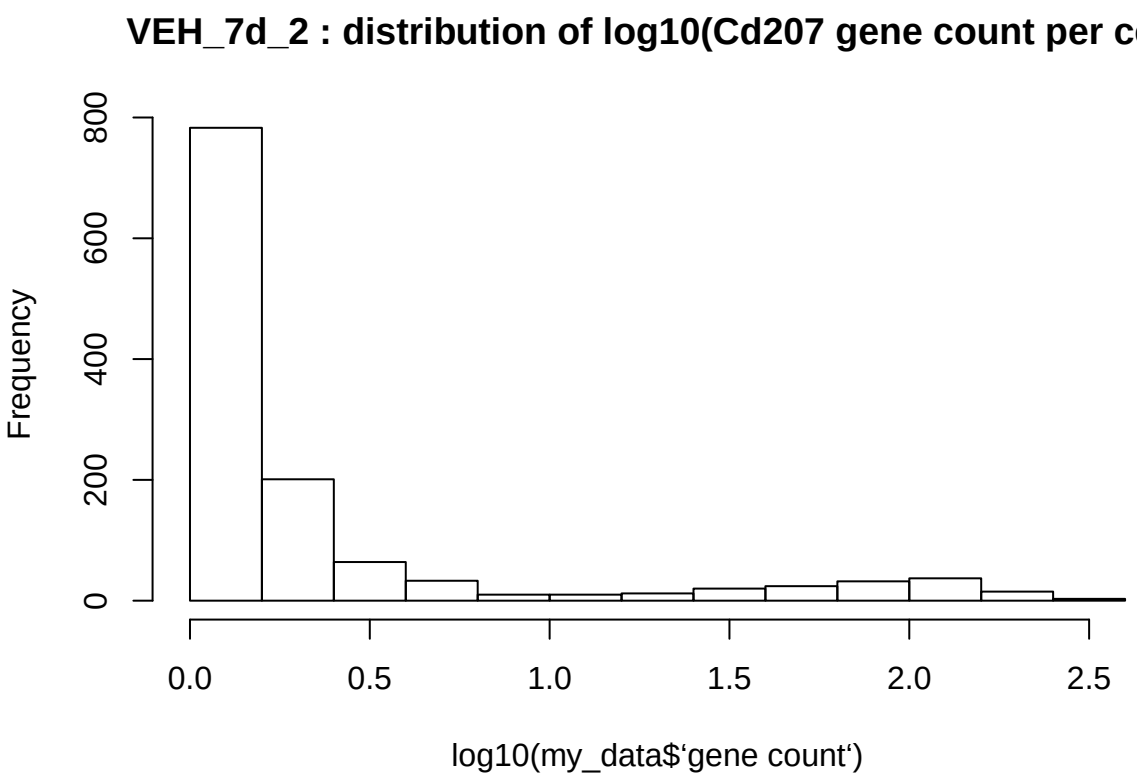
2. For a groups of cells known as macrophages, I think it is worth looking at CCL17, CD14, and LYZ2. I am not sure how these will overlap in real life, as we were not for CD207 and FSCN1, so cross analysis of these markers as in that case will help.

3. The third and messiest class are dendritic cells, where the markers that might not distinguish that are not "canonically' known. Can we look at DCN, C1qa, CCL2, and CD7 as co-markers for a group? A GROUP FOR EXAMPLE IS LANGERHANS CELLS. It would also be interesting to see if there is any CD141 or CD11c detectable in any cells. The latter two are the markers that"should" be in

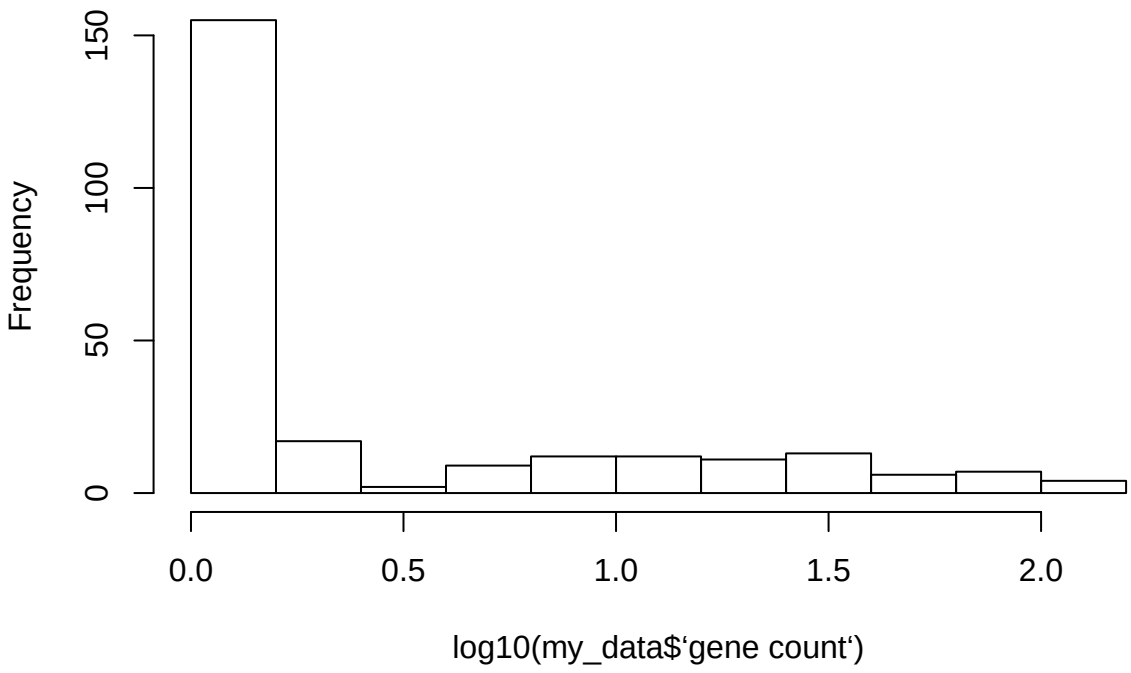


dendritic cells.

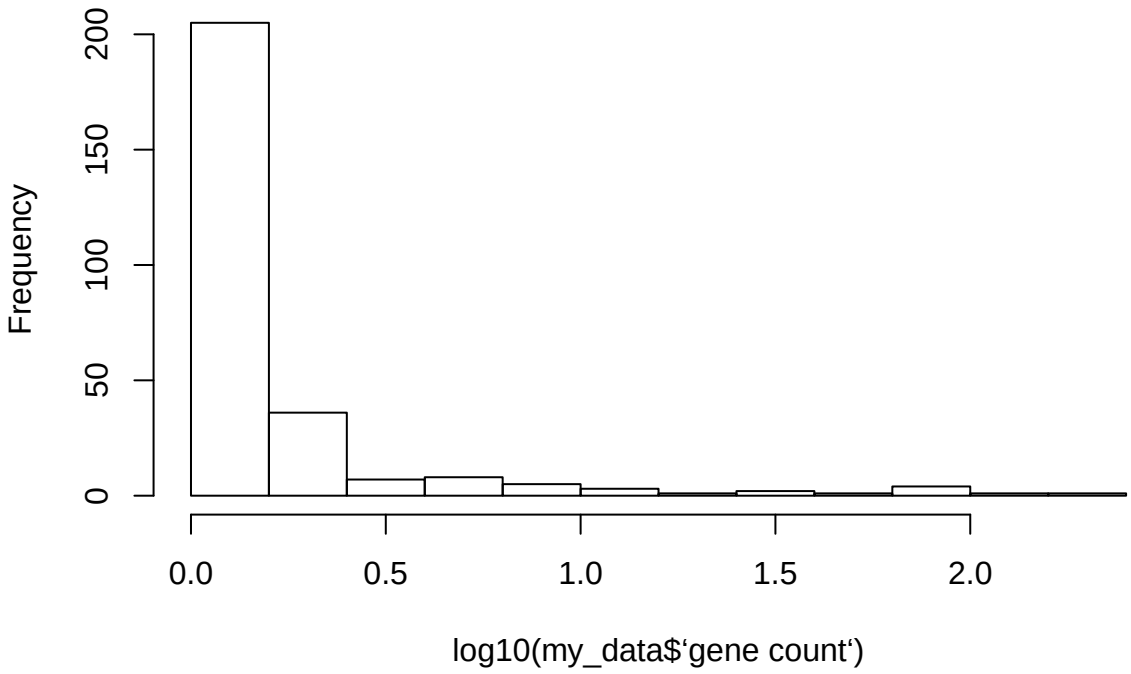
```
## [1] ""
```



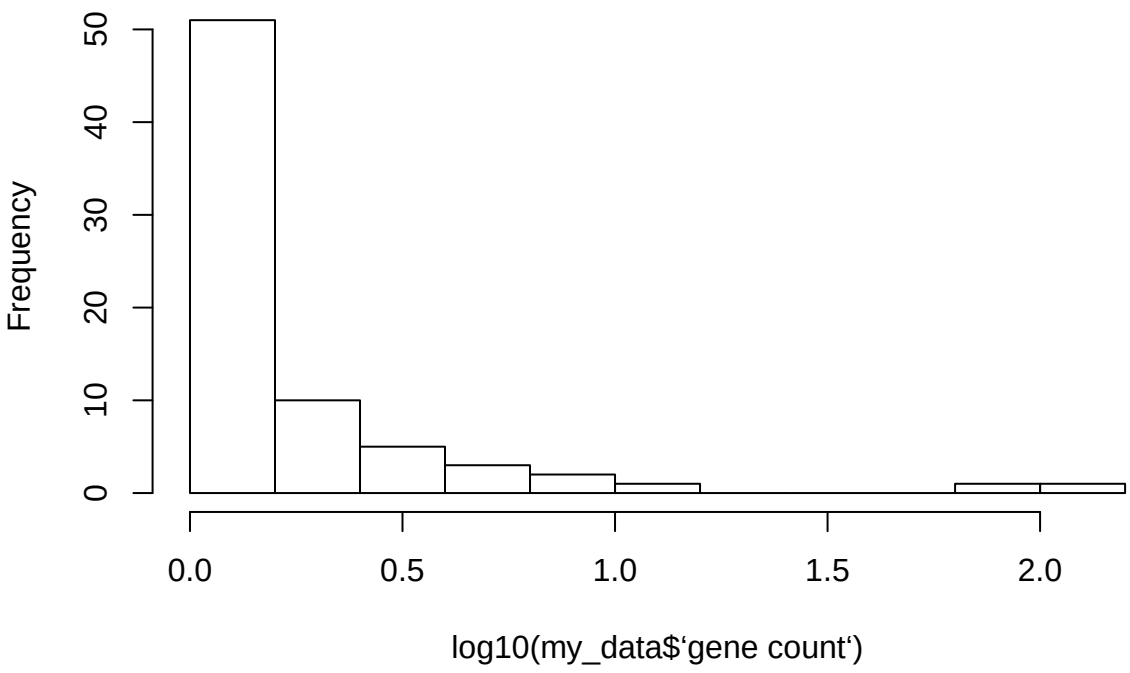

```
## [1] ""
```


VEH_7d_2 : distribution of log10(Krtdap gene count per cell)
Frequency
0
50
100
150
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

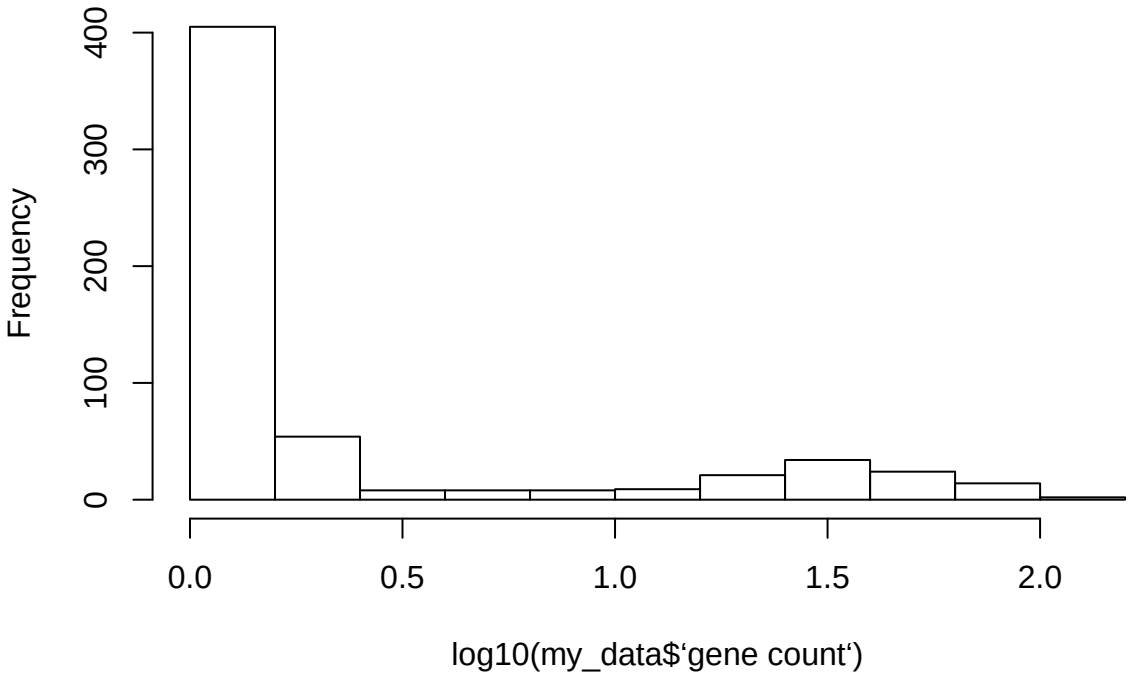
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Krt10 gene count per cell)
Frequency
0
50
100
150
200
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

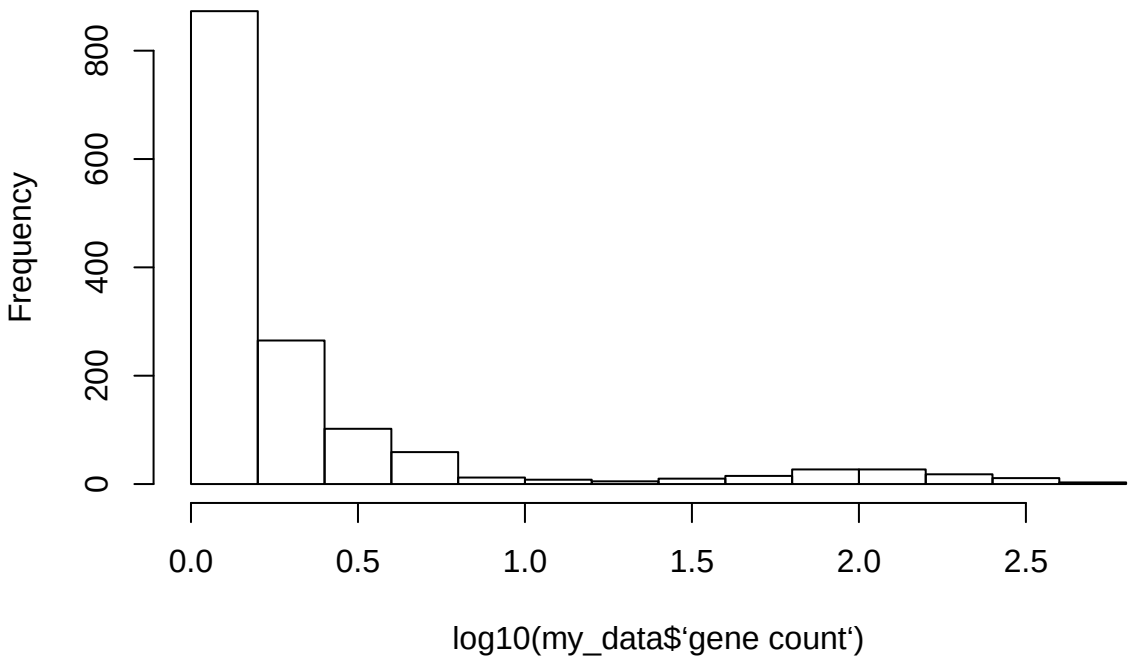
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Krt1 gene count per cell)
Frequency
0
10
20
30
40
50
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

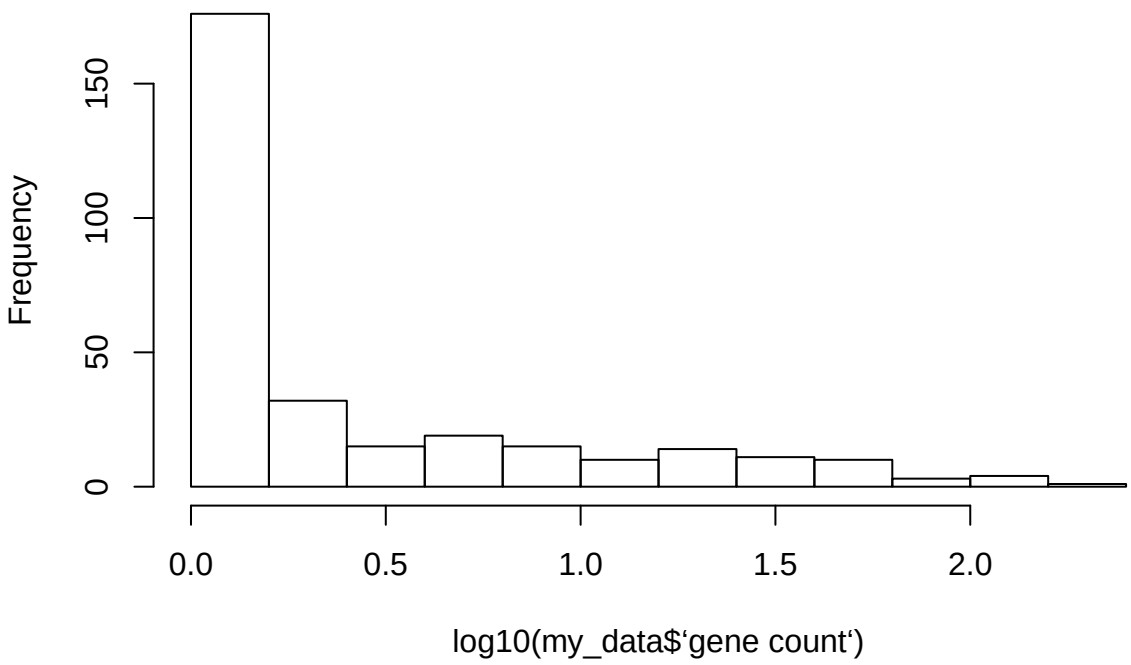
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Krt5 gene count per cell)
Frequency
0
100
200
300
400
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

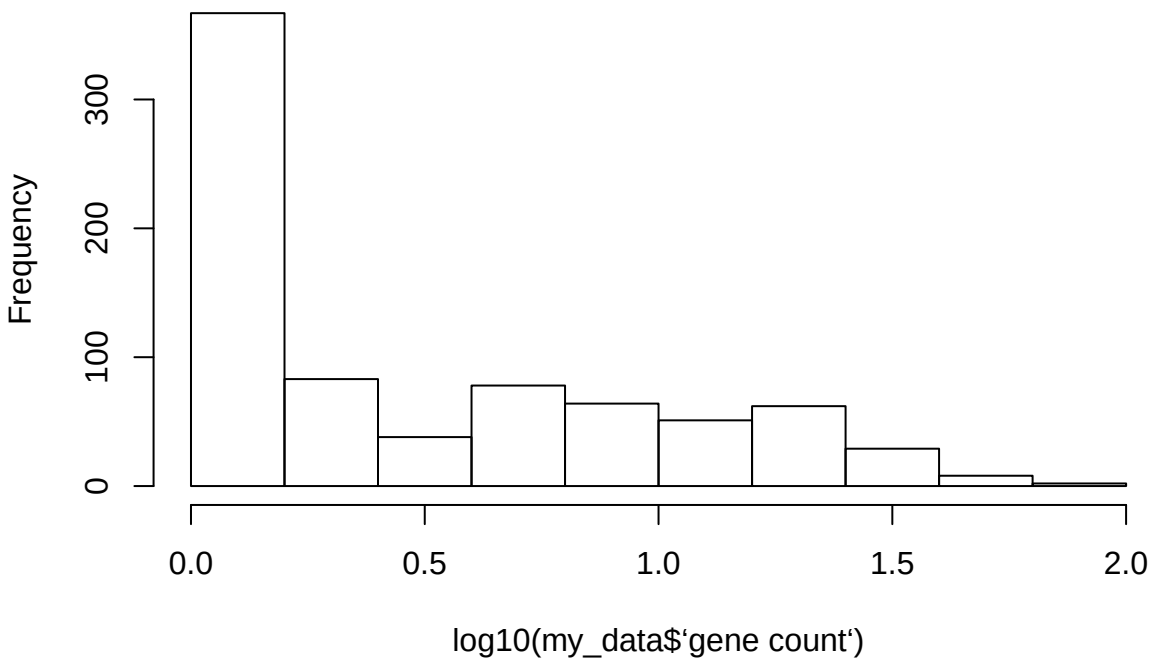
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Krt14 gene count per cell)
Frequency
0
200
400
600
800
0.0
0.5
1.0
1.5
2.0
2.5
log10(my_data$‘gene count‘)

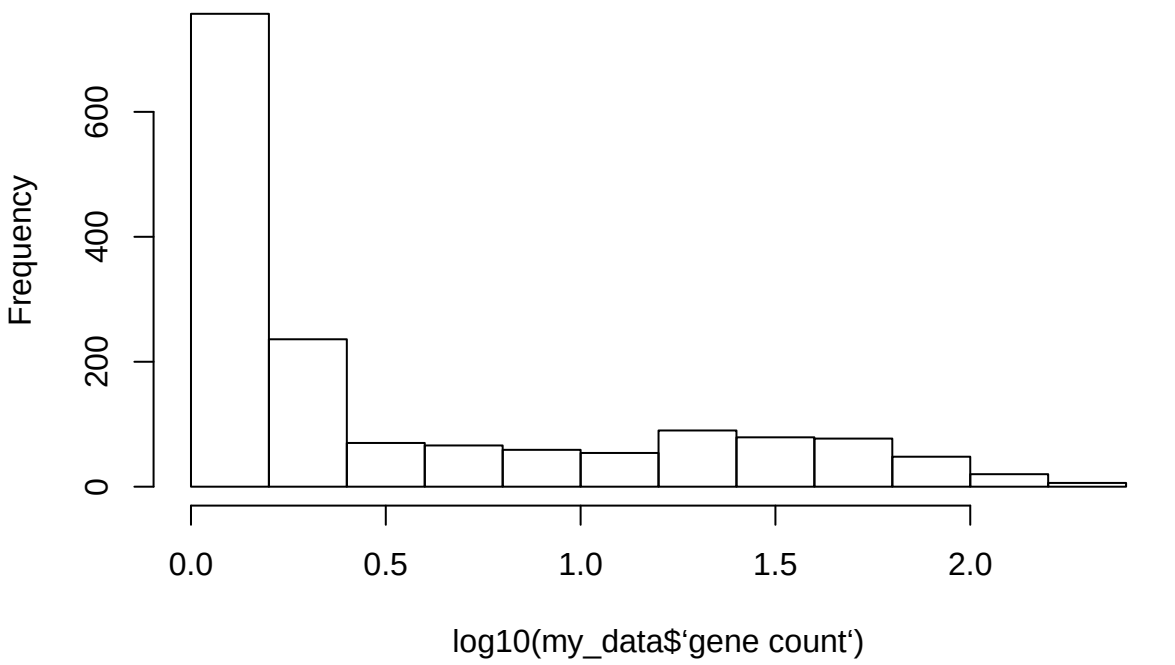
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Ccl17 gene count per cell)
Frequency
0
50
100
150
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

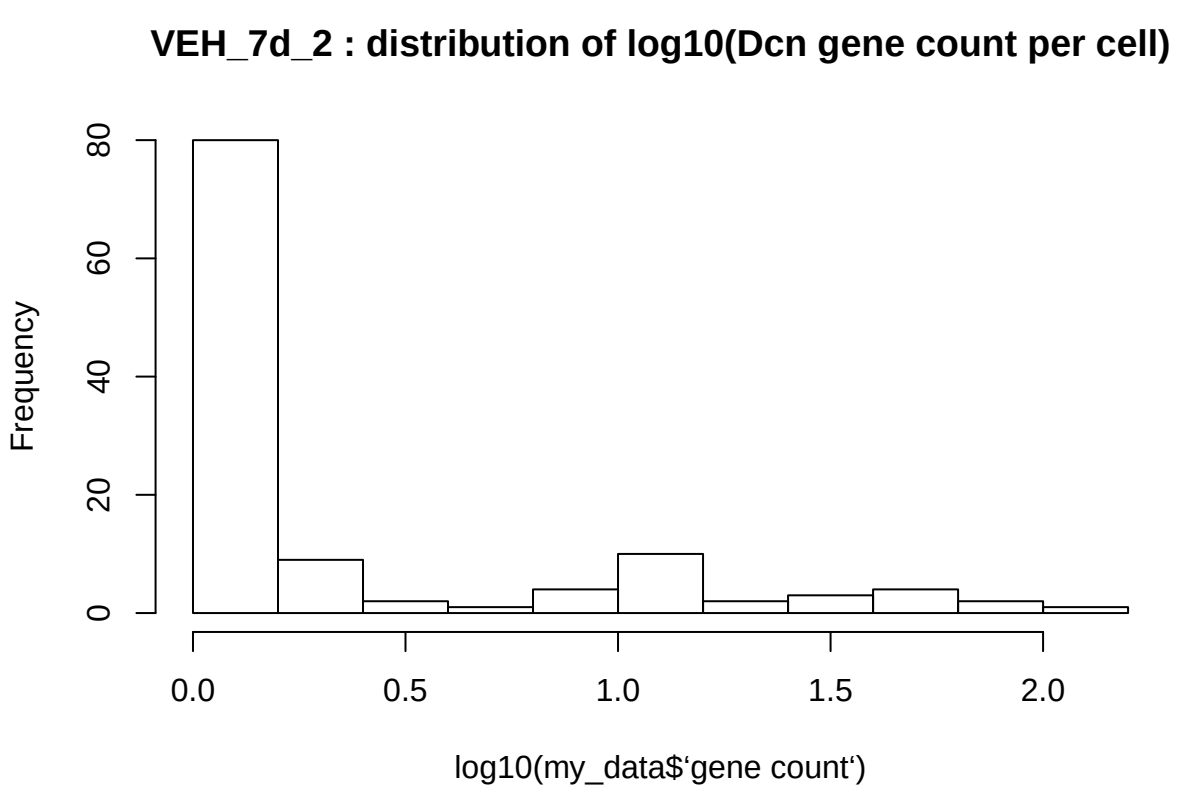
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Cd14 gene count per cell)
Frequency
0
100
200
300
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

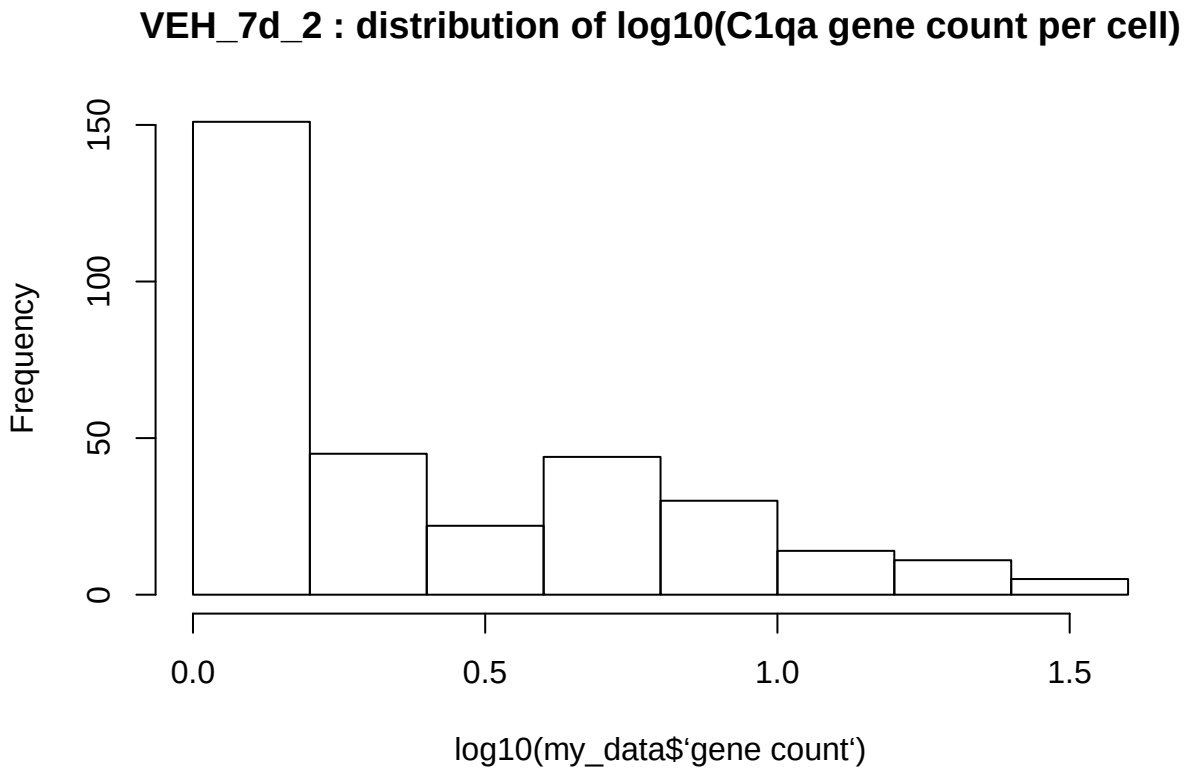
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Lyz2 gene count per cell)
Frequency
0
200
400
600
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

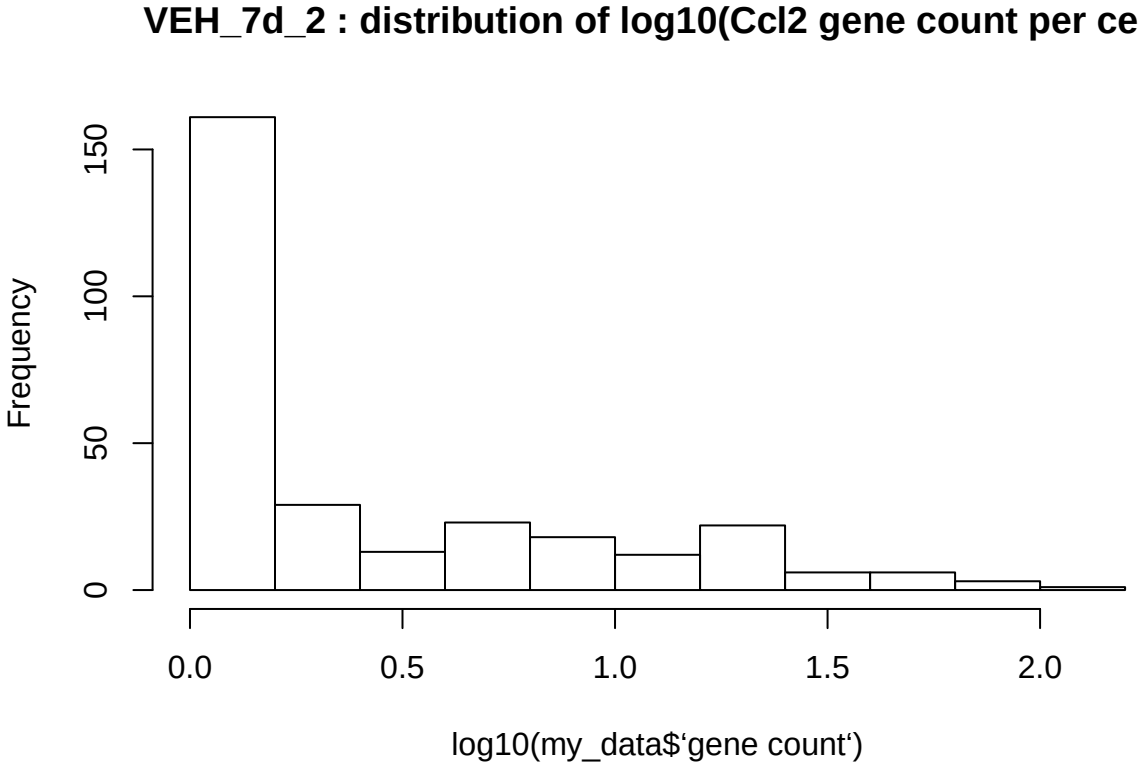
```
## [1] ""
```


VEH_7d_2 : distribution of log10(Dcn gene count per cell)
Frequency
0
20
40
60
80
0.0
0.5
1.0
1.5
2.0
log10(my_data$‘gene count‘)

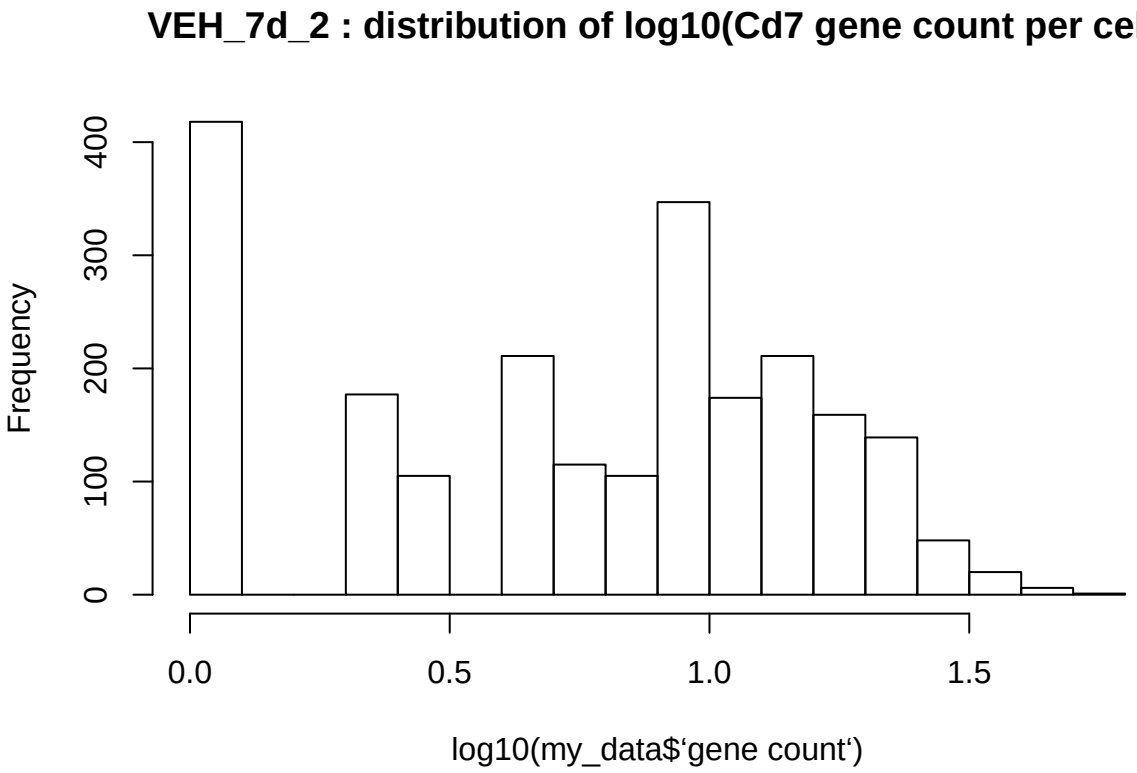
```
## [1] ""
```


VEH_7d_2 : distribution of log10(C1qa gene count per cell)
Frequency
0
50
100
150
0.0
0.5
1.0
1.5
log10(my_data$‘gene count‘)

```
## [1] ""
```




```
## [1] ""
```




```
## [1] ""
```

From where on can we call the counts *high* ?

It would be informative to add the uncertainties of the histogram entries, but I'm not quite sure about modelling their underlying distribution. Do you know any studies showing uncertainties? My first guess would be a binomial distribution...

Cd141 or Cd11c do not exist in this sample.

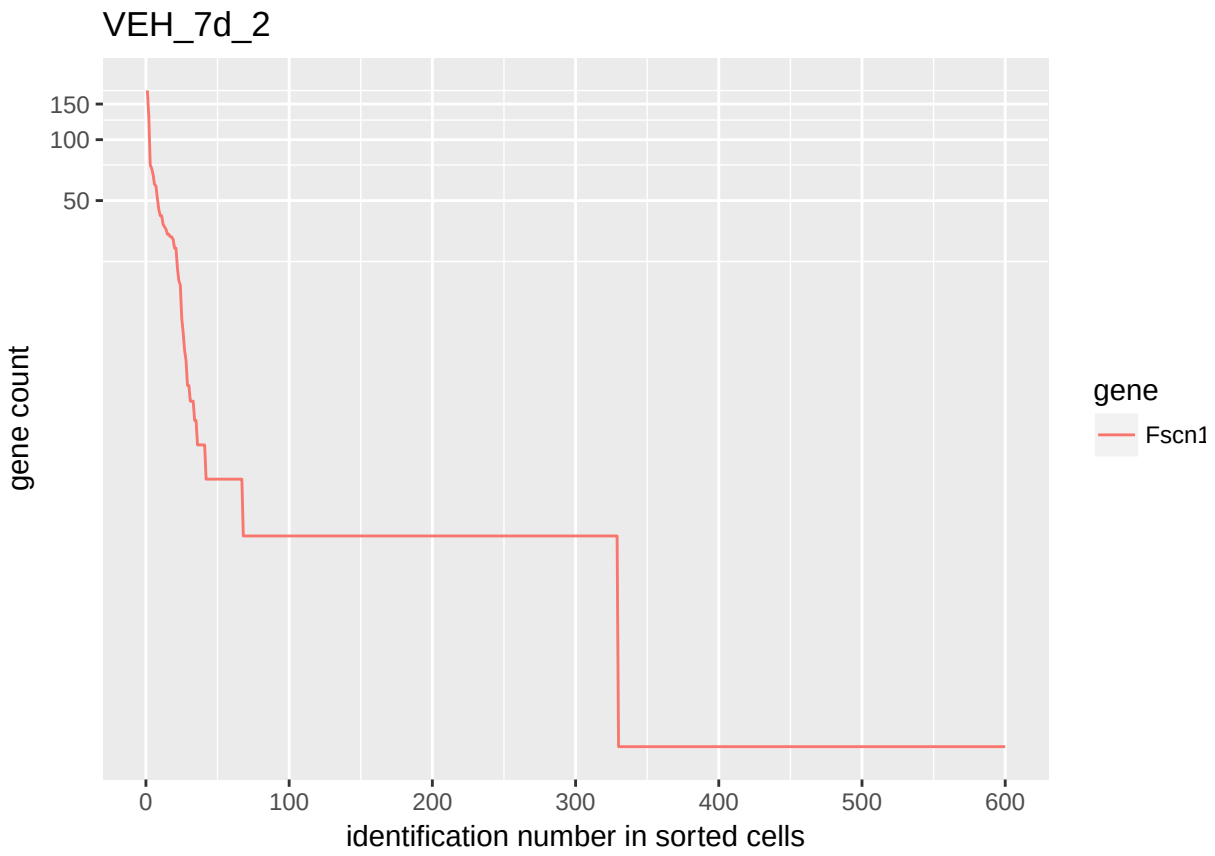
14 Genes into one Data Frame for homogeneous Analysis & Correlation Calculation

Due to limited line length, the correlation matrix is broken two times. It is symmetric, scientific number format used in the lower triangle.

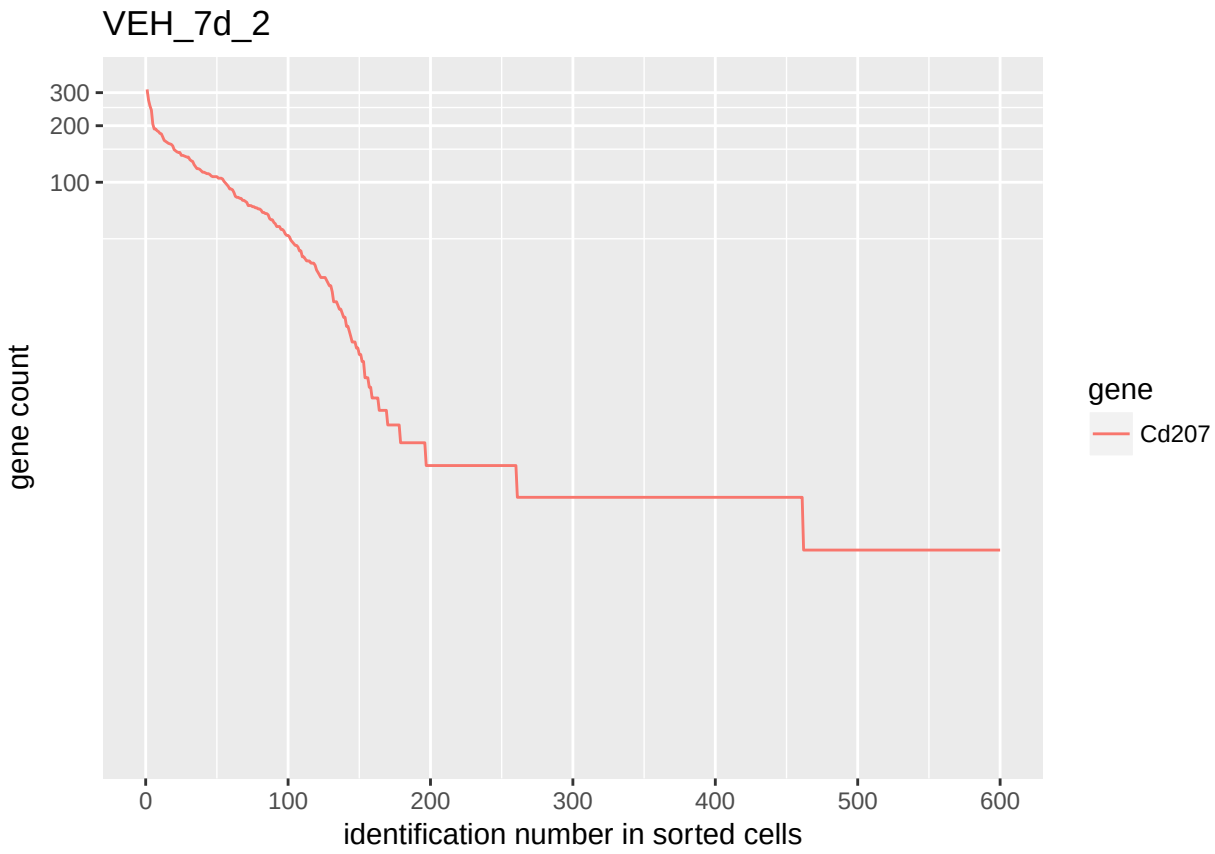
The group of "Kr. . . " genes shows a high correlation, also "Ccl17, Cd14, Lyz2" and to a lesser extend "Dcn, C1qa, Ccl2, Cd7". The correlation af the latter group with "Cd207" is not very pronounced, in the case of "Dcn" even slightly negative.

```
##                Fscn1        Cd207       Krtdap         Krt10         Krt1
## Fscn1   1.000000e+00  0.016154723 -0.002236564 -4.244128e-05  0.004843167
## Cd207   1.615472e-02  1.000000000  0.056462501  3.707376e-02  0.010084748
## Krtdap -2.236564e-03  0.056462501  1.000000000  6.244489e-01  0.519920880
## Krt10  -4.244128e-05  0.037073762  0.624448859  1.000000e+00  0.572402576
## Krt1    4.843167e-03  0.010084748  0.519920880  5.724026e-01  1.000000000
## Krt5   -3.268622e-03  0.098574858  0.445050241  1.196076e-01  0.075687205
## Krt14  -1.078317e-03  0.074276080  0.473166949  1.562014e-01  0.086940891
## Ccl17   4.864883e-02  0.013077307 -0.010258561 -2.874591e-03 -0.003016427
## Cd14   -4.930844e-03  0.074201740 -0.005280944  8.481275e-03 -0.004512958
## Lyz2   -1.364099e-02  0.001023127 -0.018375959 -4.489094e-03 -0.007434644
## Dcn     5.761081e-03 -0.009245424 -0.006381983 -2.611775e-04 -0.002789213
## C1qa   -2.093166e-03  0.118774300  0.003746823  1.151407e-02 -0.004984274
## Ccl2   -5.628612e-03  0.015168057  0.001750705  1.073559e-03 -0.001899169
## Cd7    -3.669397e-03  0.068572910  0.044555928  5.674663e-02  0.023622745
##                Krt5        Krt14        Ccl17         Cd14         Lyz2
## Fscn1  -0.003268622 -0.001078317  0.048648828 -0.004930844 -0.013640988
## Cd207   0.098574858  0.074276080  0.013077307  0.074201740  0.001023127
## Krtdap  0.445050241  0.473166949 -0.010258561 -0.005280944 -0.018375959
## Krt10   0.119607613  0.156201364 -0.002874591  0.008481275 -0.004489094
## Krt1    0.075687205  0.086940891 -0.003016427 -0.004512958 -0.007434644
## Krt5    1.000000000  0.933663753 -0.010266229 -0.018362640 -0.024598086
## Krt14   0.933663753  1.000000000 -0.007764096 -0.021607434 -0.025483481
## Ccl17  -0.010266229 -0.007764096  1.000000000  0.288765456  0.201883630
## Cd14   -0.018362640 -0.021607434  0.288765456  1.000000000  0.585771970
## Lyz2   -0.024598086 -0.025483481  0.201883630  0.585771970  1.000000000
## Dcn    -0.010048375 -0.007541935 -0.005100559 -0.008521356 -0.010217576
## C1qa    0.005888823 -0.001002526  0.012270691  0.348011446  0.338764976
## Ccl2   -0.006024612 -0.007321350  0.008123263  0.155844488  0.145210596
## Cd7     0.050173570  0.051878705  0.007011011  0.005235398  0.014689606
##                 Dcn         C1qa         Ccl2          Cd7
## Fscn1   0.0057610805 -0.002093166 -0.005628612 -0.003669397
## Cd207  -0.0092454241  0.118774300  0.015168057  0.068572910
## Krtdap -0.0063819832  0.003746823  0.001750705  0.044555928
## Krt10  -0.0002611775  0.011514069  0.001073559  0.056746632
## Krt1   -0.0027892134 -0.004984274 -0.001899169  0.023622745
## Krt5   -0.0100483753  0.005888823 -0.006024612  0.050173570
## Krt14  -0.0075419351 -0.001002526 -0.007321350  0.051878705
## Ccl17  -0.0051005589  0.012270691  0.008123263  0.007011011
## Cd14   -0.0085213564  0.348011446  0.155844488  0.005235398
## Lyz2   -0.0102175761  0.338764976  0.145210596  0.014689606
## Dcn     1.0000000000 -0.008890711  0.157914370  0.017626614
## C1qa   -0.0088907112  1.000000000  0.241321108  0.137426272
## Ccl2    0.1579143699  0.241321108  1.000000000  0.078608339
## Cd7     0.0176266143  0.137426272  0.078608339  1.000000000
```

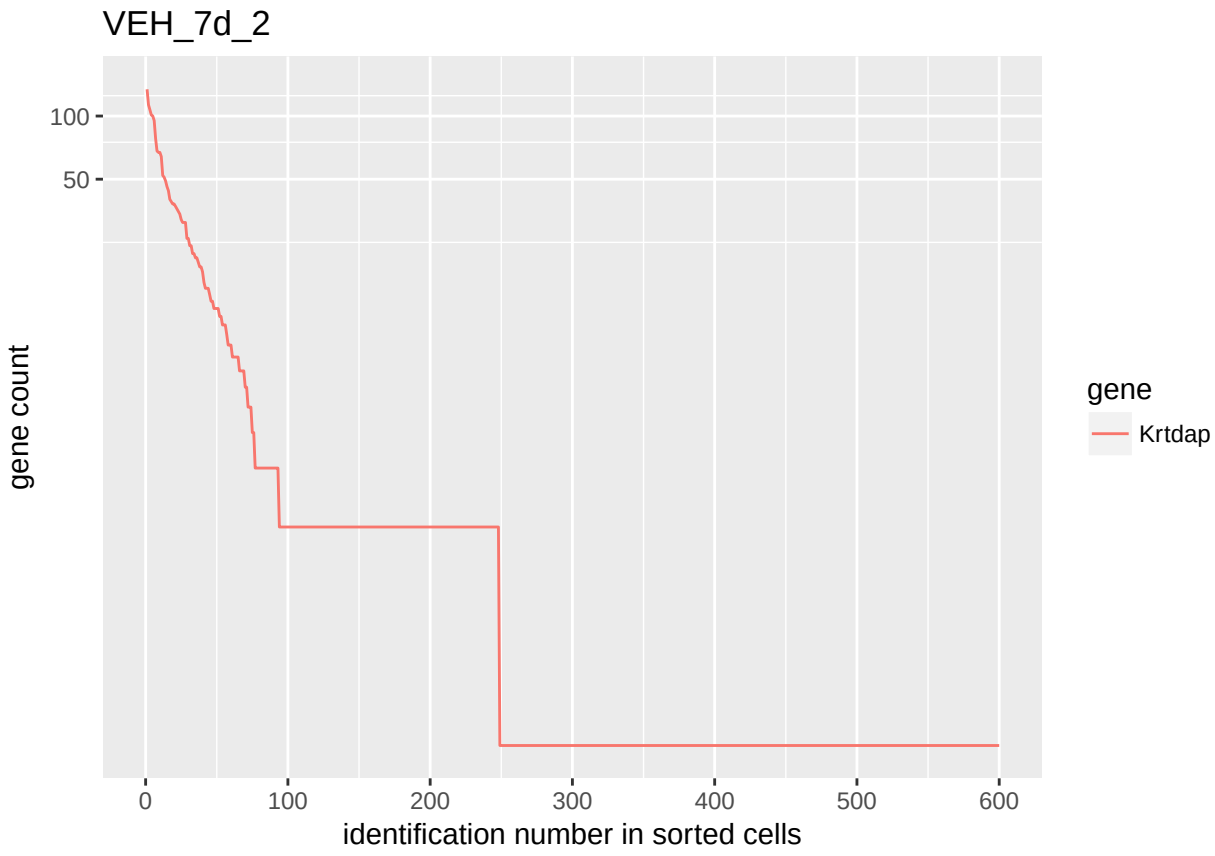
14 Cell / Gene-Count Plots



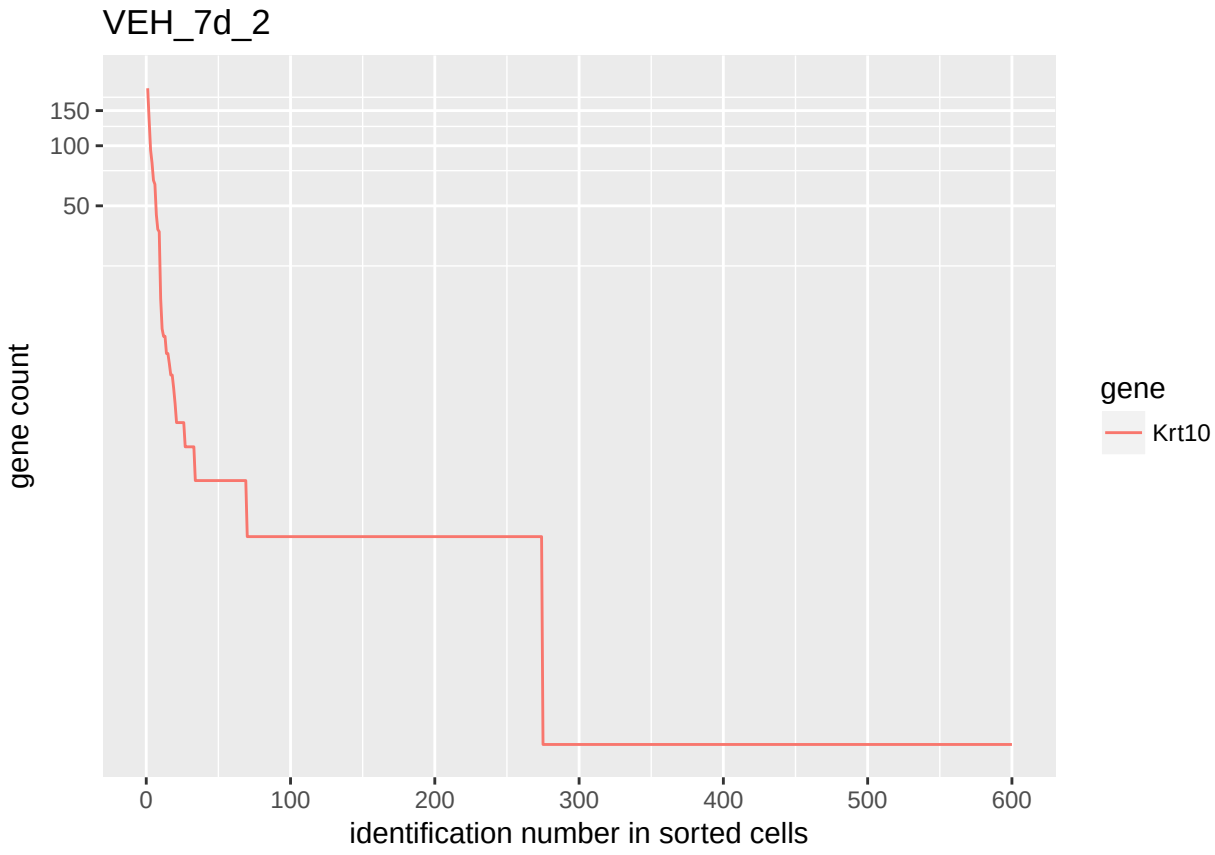

```
## [1] ""
```



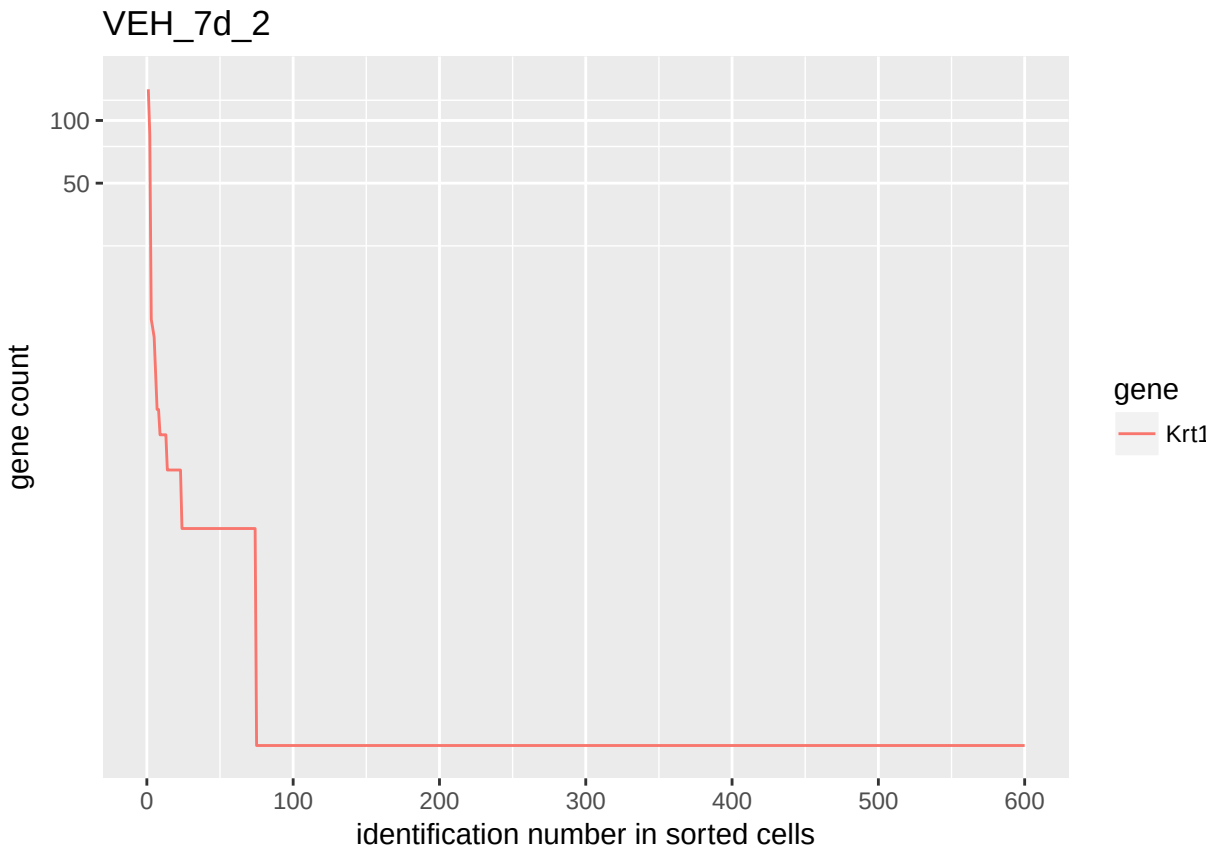

```
## [1] ""
```


VEH_7d_2
100
50
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Krtdap

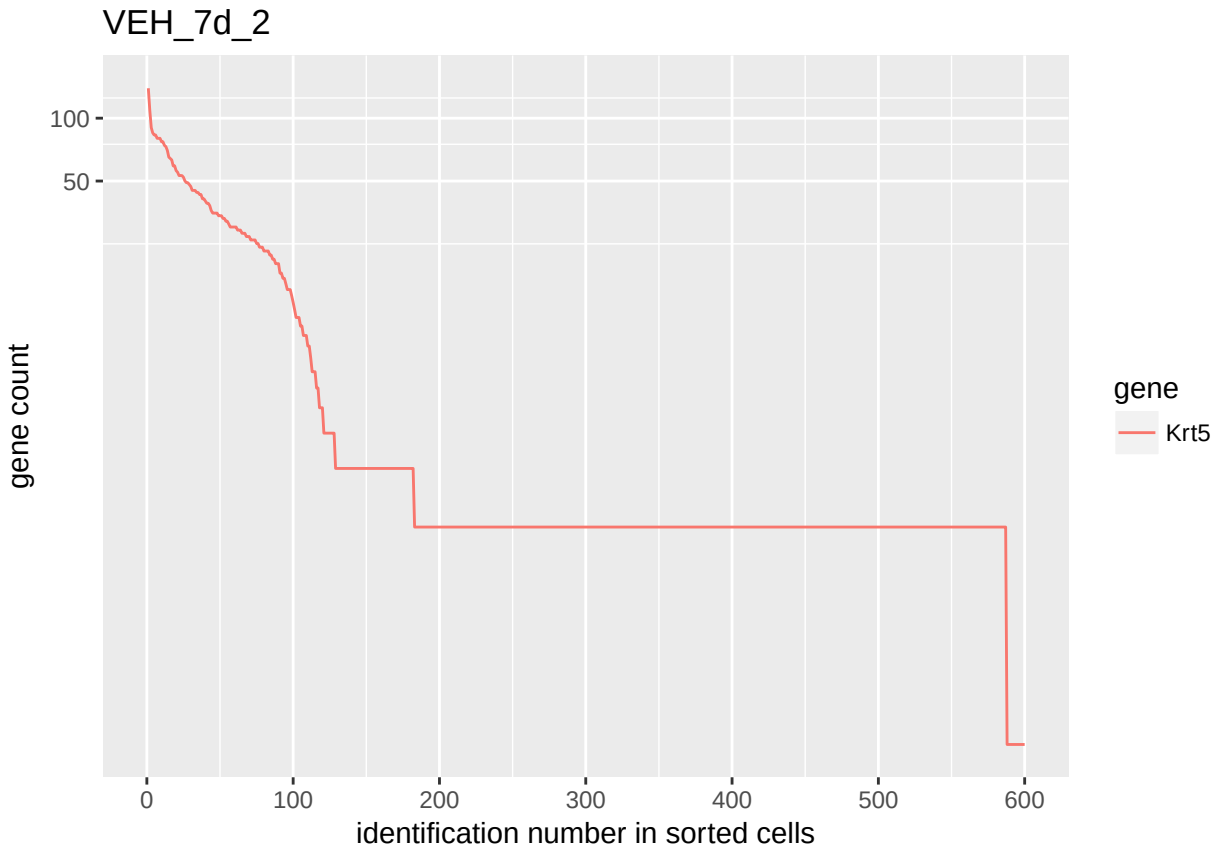
```
## [1] ""
```


VEH_7d_2
150
100
50
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Krt10

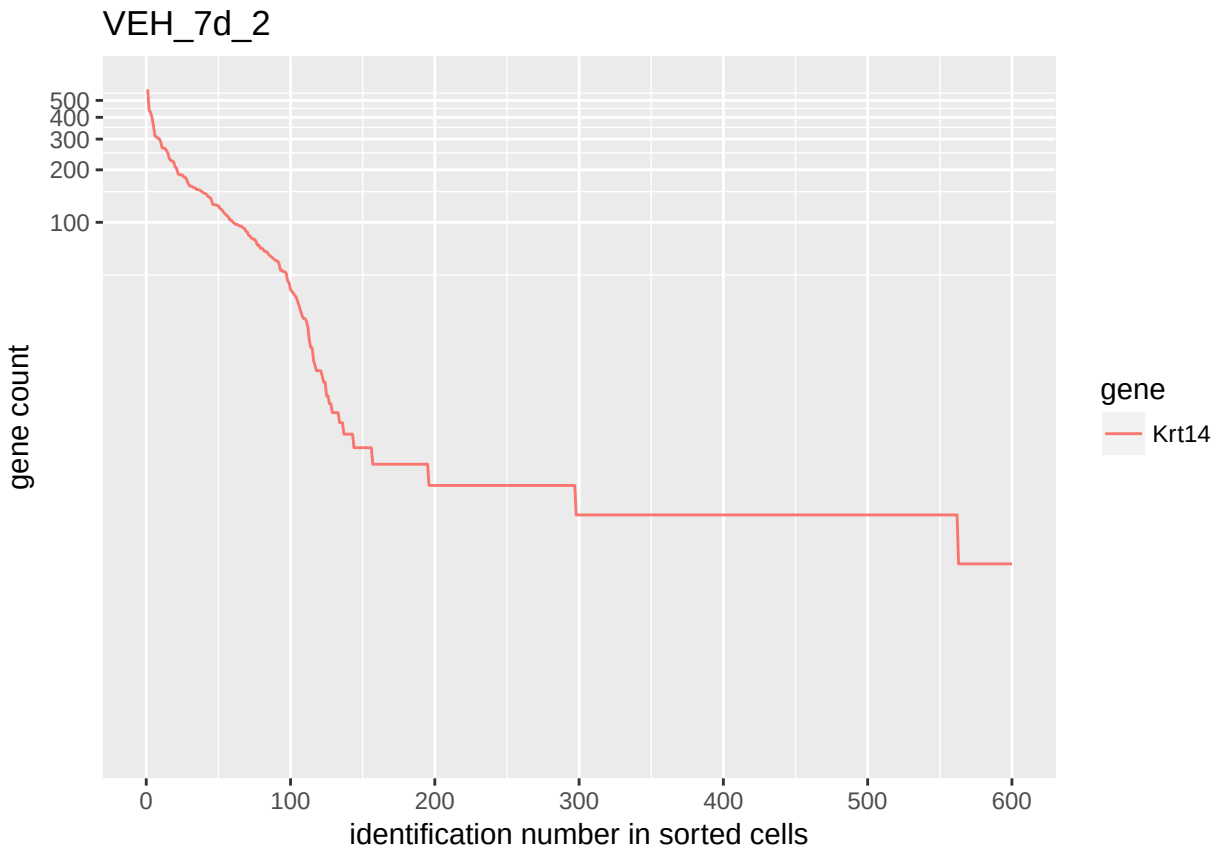
```
## [1] ""
```


VEH_7d_2
100
50
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Krt1

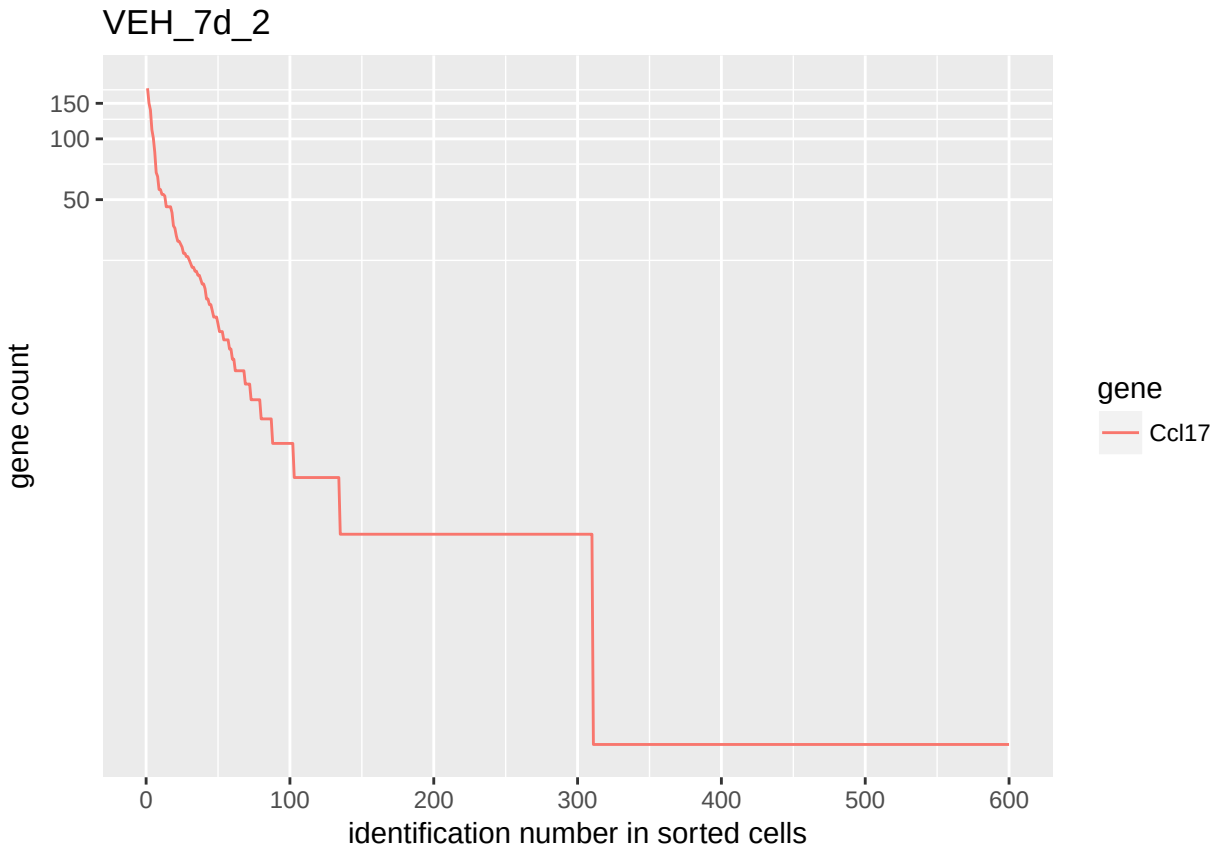
```
## [1] ""
```


VEH_7d_2
100
50
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Krt5

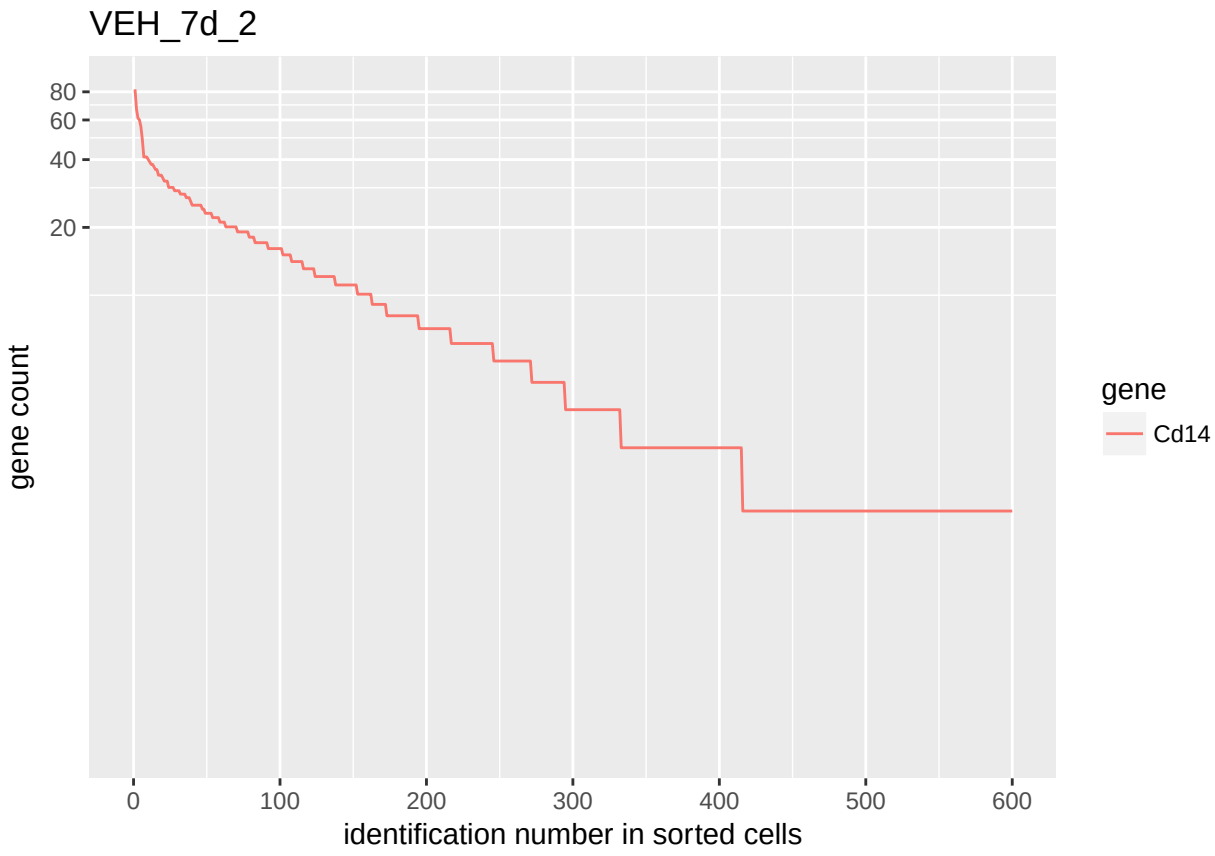
```
## [1] ""
```


VEH_7d_2
500
400
300
200
100
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Krt14

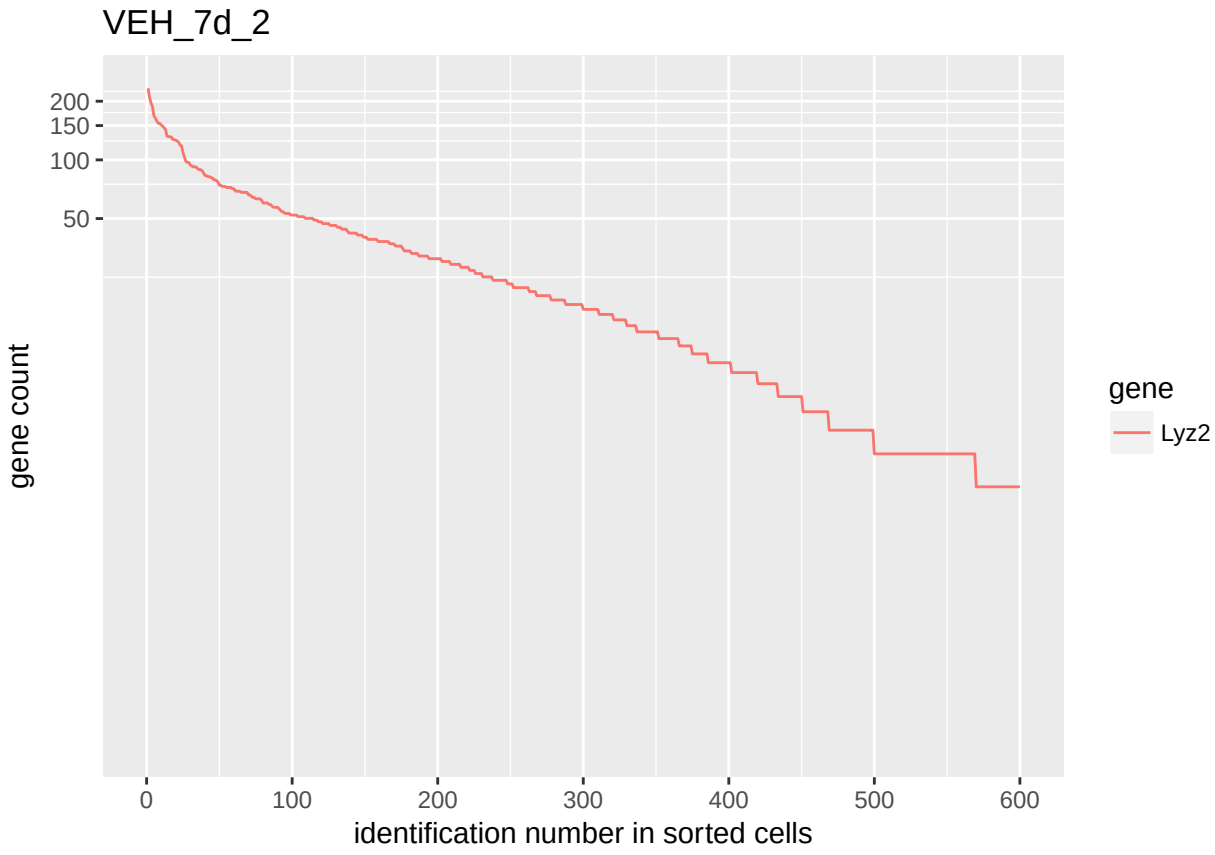
```
## [1] ""
```


VEH_7d_2
150
100
50
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Ccl17

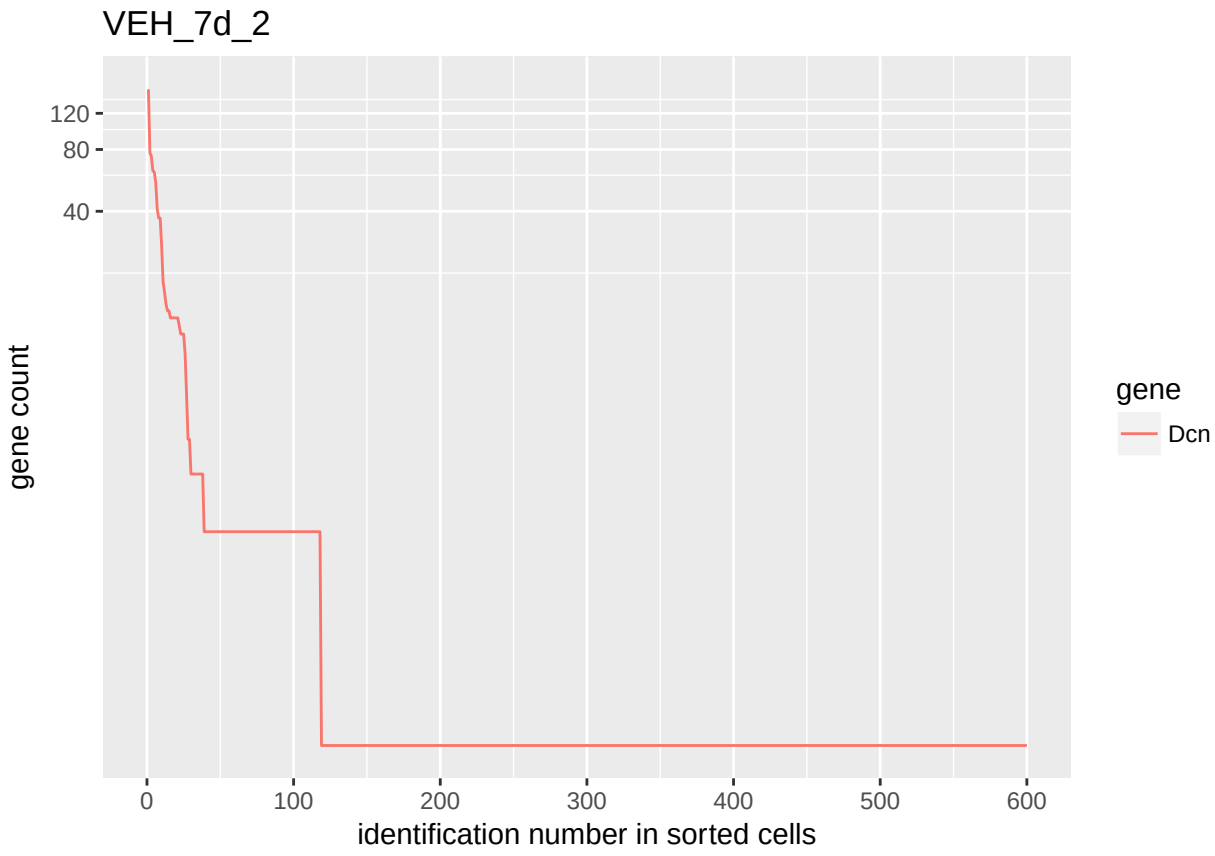
```
## [1] ""
```


VEH_7d_2
80
60
40
20
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Cd14

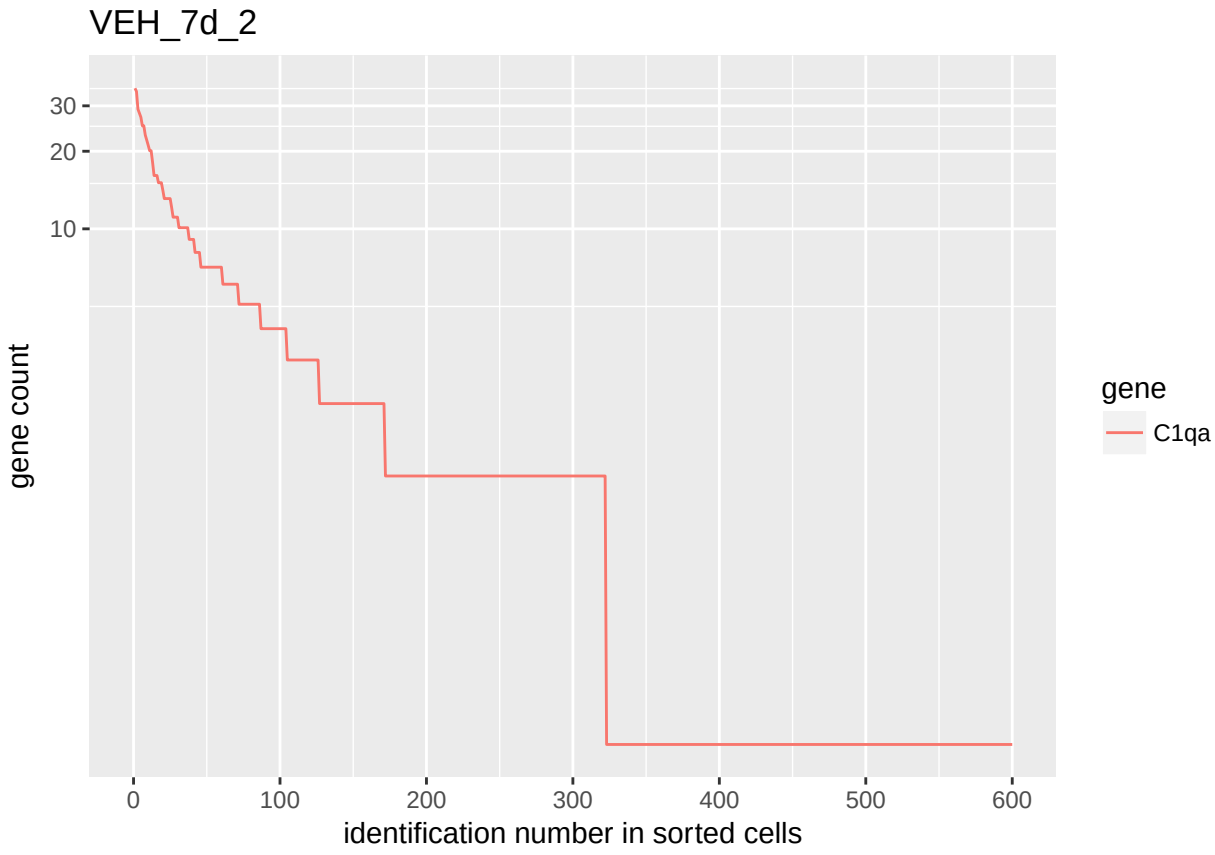
```
## [1] ""
```


VEH_7d_2
200
150
100
50
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Lyz2

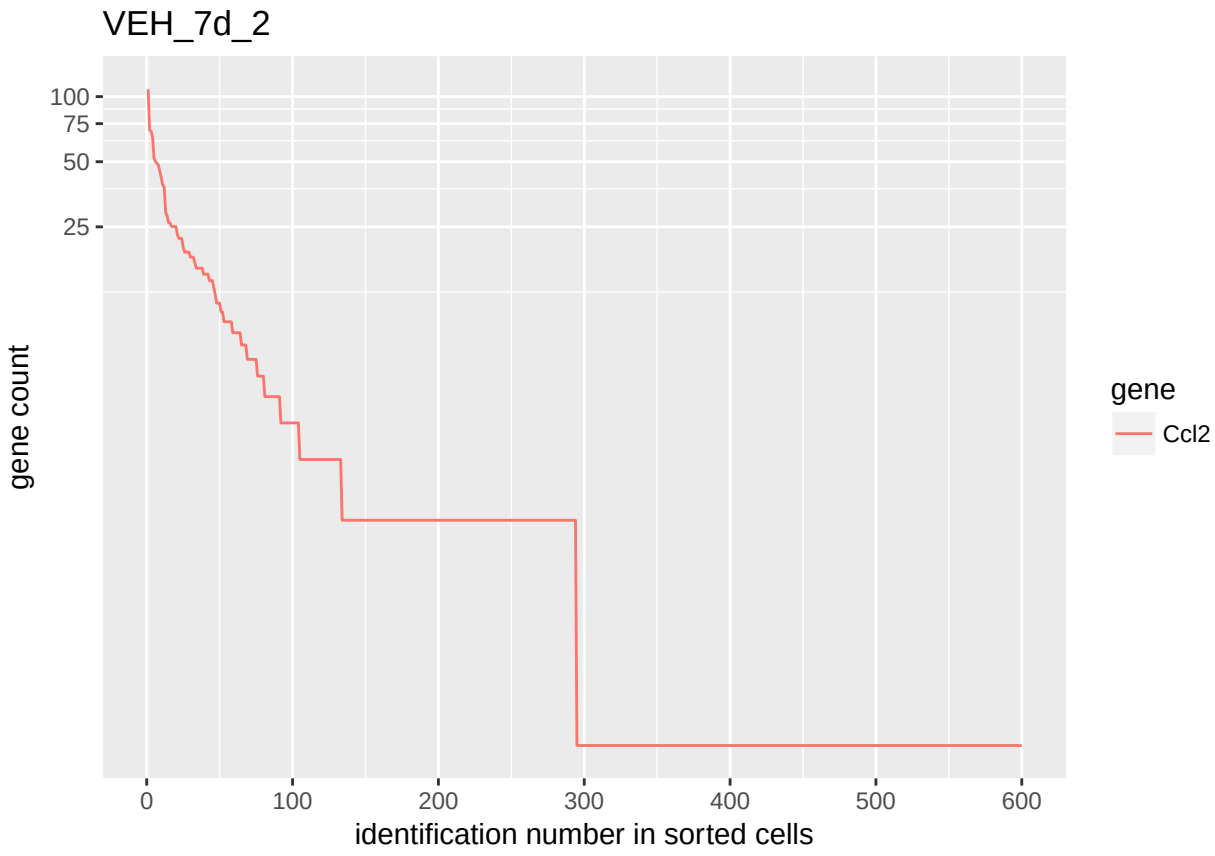
```
## [1] ""
```


VEH_7d_2
120
80
40
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
Dcn

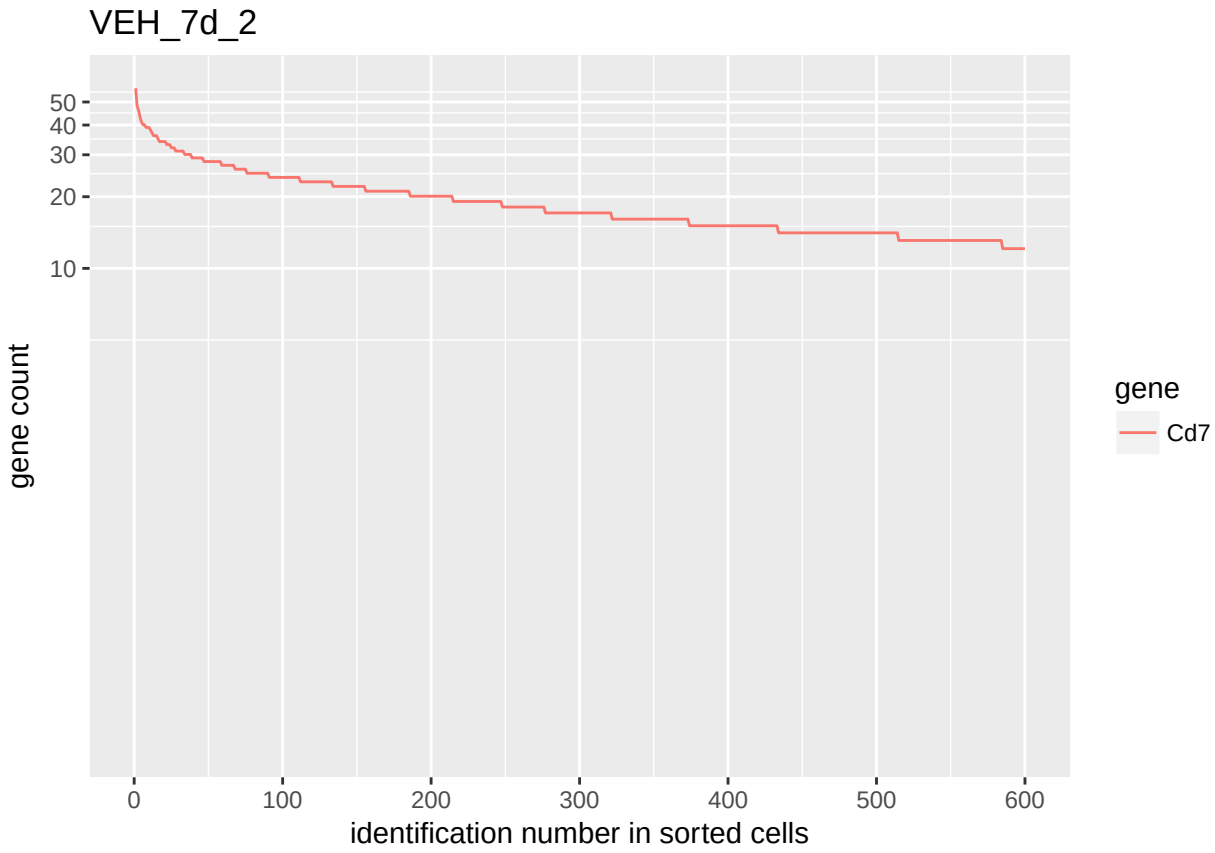
```
## [1] ""
```


VEH_7d_2
30
20
10
gene count
0
100
200
300
400
500
600
identification number in sorted cells
gene
C1qa

```
## [1] ""
```




```
## [1] ""
```




```
## [1] ""
```

Next Steps

- Plots combining genes: 14 genes give 91 pairs, which could in principle be ordered for both gene counts, I need some tactics to not drown in data and propose:

  - I draw all pairs of “Krtdap”, “Krt10”, “Krt1”, “Krt5”, “Krt14” and perhaps triples etc.
  - then all pairs of “Ccl17”,“Cd14”,“Lyz2”
  - as Raymond mentionned “Dcn”, “C1qa”, “Ccl2”, “Cd7” as co-markers for Langerhans cells, I pair them with “Cd207” and perhaps “Fscn1”
  - I pair “Cd141” with “Cd11c” when studying samples where they exist
  - I sort for the gene with higher mean value
- More samples:
  - *Raymond: A few other notes - the samples you started looking at are mouse (IMQ or VEH in name) rather than the remaining 4 which are human. I assume we are sticking with mouse right now as that's where we started. The (3) category above will probably only be detectable in IMQ samples.*
  - QC for mouse samples gave problems (Fscn1 and Cd207 passed but none of the cells), so human samples as crosscheck + **all** mouse samples, then trace the mouse problem
  - *Raymond: Finally after removal of these cells (one we can do at high confidence), we will cluster all 6 mouse samples together. I assume removing cells before normalization is not really an issue.*