007-Sensitivity Analysis of Many Paramters

Clustering Gene Expression Data

May 14, 2019

Abstract

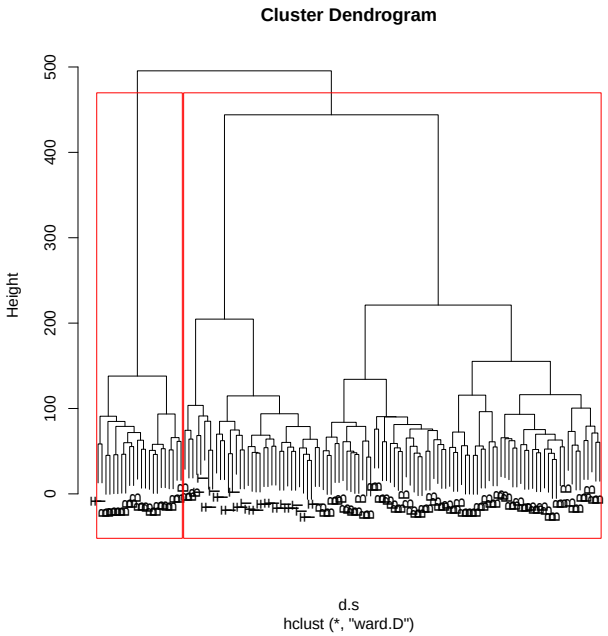
DNA microarrays may be used to characterize the molecular variations among tumors by monitoring gene expression profiles on a genomic scale. This may lead to a finer and more reliable classification of tumors, and to the identification of marker genes that distinguish among these classes. Eventual clinical implications include an improved ability to understand and predict cancer survival (Dudoit and Gentleman, 2002). Therefore, a common task is to determine whether or not gene expression data can reliably identify or classify different types of a disease. We consider gene expression data from patients with acute lymphoblastic leukemia (ALL) that were investigated using HGU95AV2 Affymetrix GeneChip arrays (Chiaretti et al., 2004). The data consist of 128 patients with 12,625 genes. A number of additional covariates are available such as the type and stage of the disease; "B" indicates B-cell ALL, while a "T" indicates T-cell ALL. Several clustering procedures require user imputs such as the type of clustering and the number of clusters. Pre-filtering the data based on the most variable genes can also lead to increased power. We are interested in the effect these parameters have on the clustering results. Here I provide an illustration of performing such a task in an efficient and reproducible way using the function `knitr::knit_expand` (Xie, 2015, 2013, 2014) with the ALL dataset (Li, 2009).

Contents

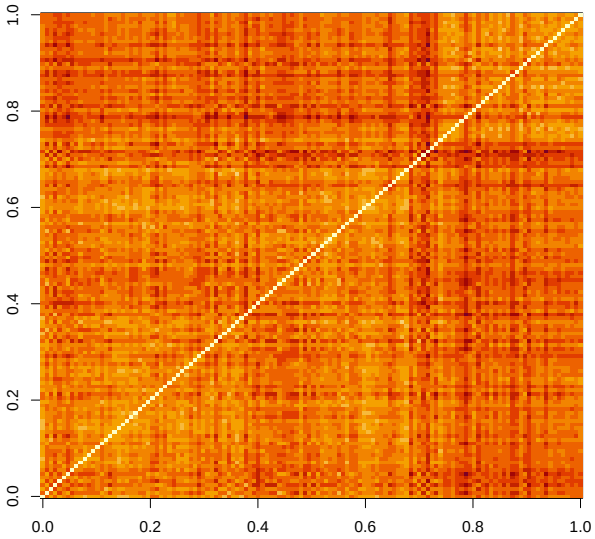
1 Method: ward.D, Filter: 10%, Groups: 2

```
dim(dat.filter)

## [1] 11362   128

table(groups, cl)

##       cl
## groups  B  T
##      1 75 31
##      2 20  2

fisher.test(groups, cl)$p.value

## [1] 0.061
```

1 and 1a and 1b.



(a) Dendogram



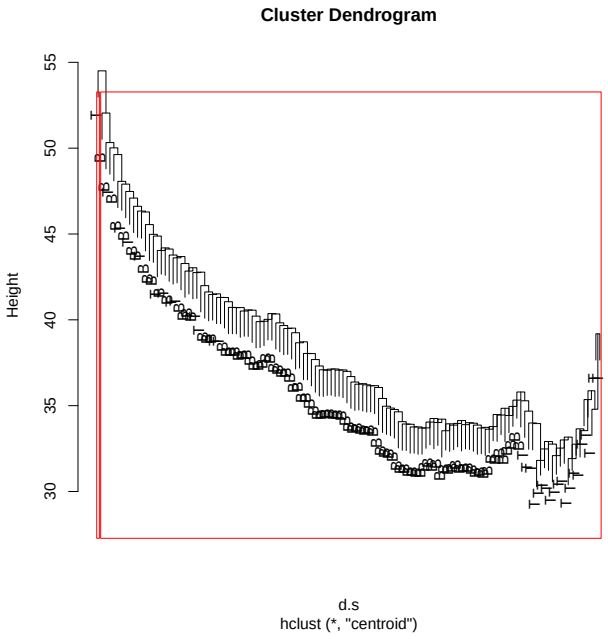
(b) Distance Matrix

Figure 1: based on Method: ward.D, Filter: 10%, Groups: 2

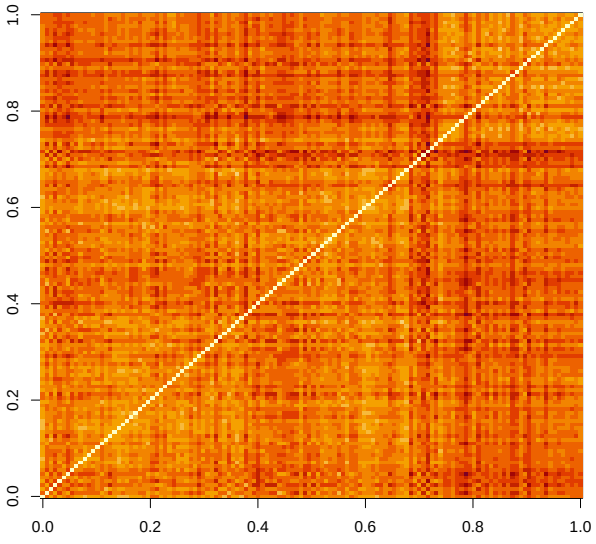
2 Method: centroid, Filter: 10%, Groups: 2

```
dim(dat.filter)

## [1] 11362   128

table(groups, cl)

##       cl
## groups  B  T
##      1 95 32
##      2  0  1

fisher.test(groups, cl)$p.value

## [1] 0.26
```

2 and 2a and 2b.



(a) Dendogram



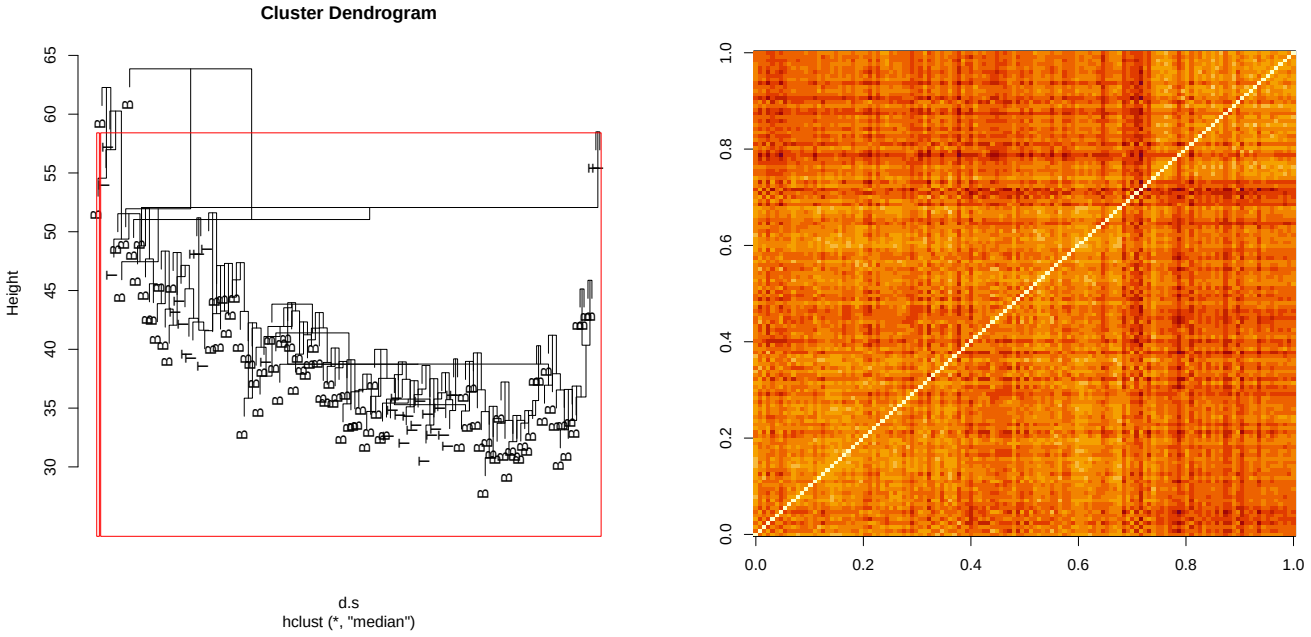
(b) Distance Matrix

Figure 2: based on Method: centroid, Filter: 10%, Groups: 2

3 Method: median, Filter: 10%, Groups: 2

```
dim(dat.filter)

## [1] 11362   128

table(groups, cl)

##       cl
## groups  B  T
##      1 94 33
##      2  1  0

fisher.test(groups, cl)$p.value

## [1] 1
```

3 and 3a and 3b.



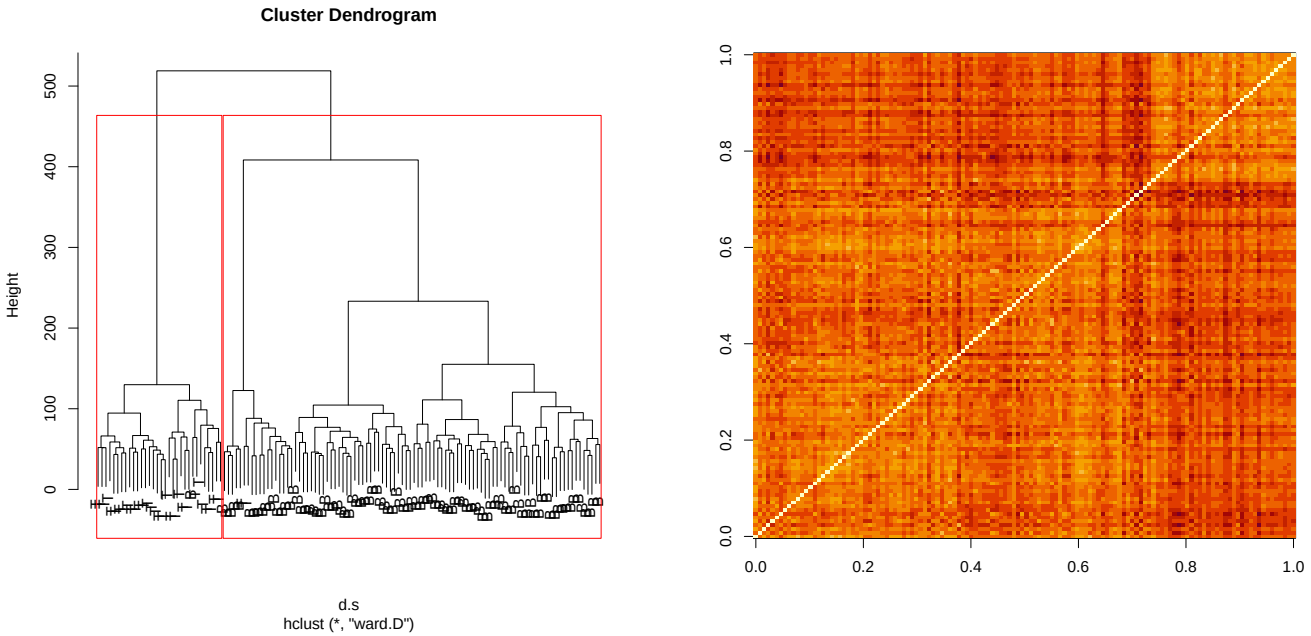
(a) Dendogram (b) Distance Matrix

Figure 3: based on Method: median, Filter: 10%, Groups: 2

4 Method: ward.D, Filter: 50%, Groups: 2

```
dim(dat.filter)

## [1] 6313  128

table(groups, cl)

##       cl
## groups  B  T
##      1 94  2
##      2  1 31

fisher.test(groups, cl)$p.value

## [1] 3.4e-26
```

4 and 4a and 4b.



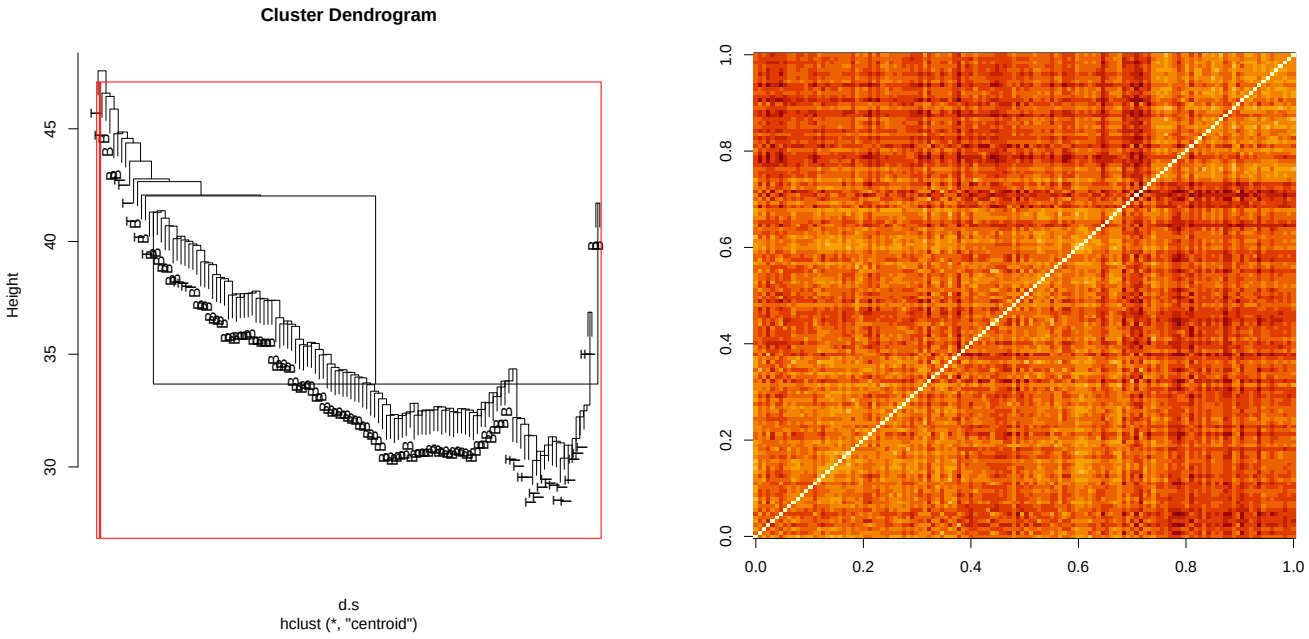
(a) Dendogram (b) Distance Matrix

Figure 4: based on Method: ward.D, Filter: 50%, Groups: 2

5 Method: centroid, Filter: 50%, Groups: 2

```
dim(dat.filter)

## [1] 6313  128

table(groups, cl)

##       cl
## groups  B  T
##      1 95 32
##      2  0  1

fisher.test(groups, cl)$p.value

## [1] 0.26
```

5 and 5a and 5b.



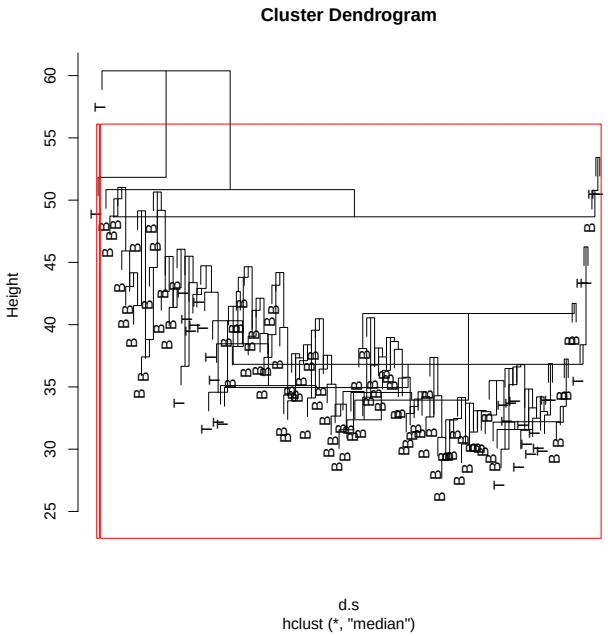
(a) Dendogram (b) Distance Matrix

Figure 5: based on Method: centroid, Filter: 50%, Groups: 2

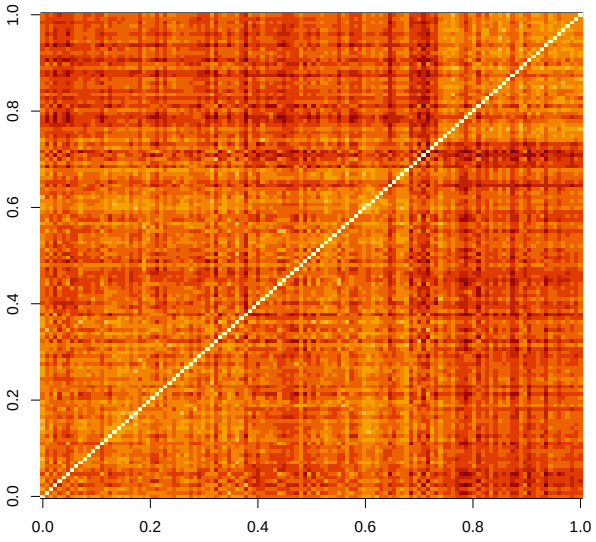
6 Method: median, Filter: 50%, Groups: 2

```
dim(dat.filter)

## [1] 6313  128

table(groups, cl)

##        cl
## groups  B  T
##      1 95 32
##      2  0  1

fisher.test(groups, cl)$p.value

## [1] 0.26
```

6 and 6a and 6b.



(a) Dendogram



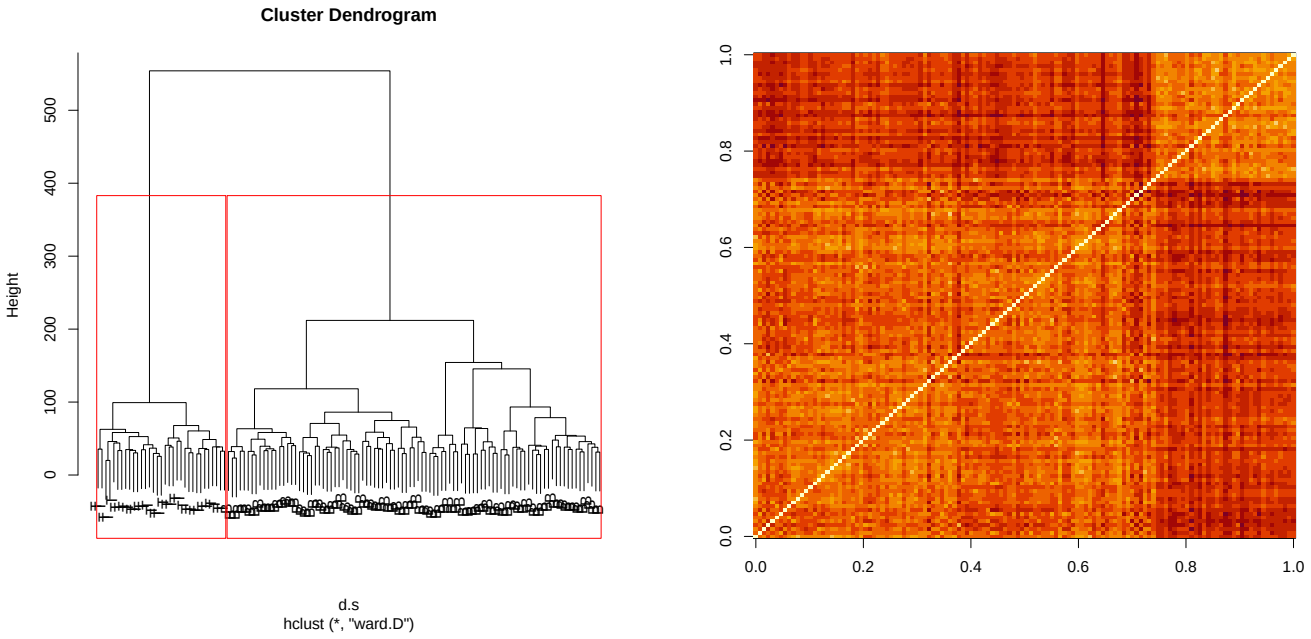
(b) Distance Matrix

Figure 6: based on Method: median, Filter: 50%, Groups: 2

7 Method: ward.D, Filter: 90%, Groups: 2

```
dim(dat.filter)

## [1] 1263  128

table(groups, cl)

##       cl
## groups  B  T
##      1 95  0
##      2  0 33

fisher.test(groups, cl)$p.value

## [1] 2.3e-31
```

7 and 7a and 7b.



(a) Dendogram (b) Distance Matrix

Figure 7: based on Method: ward.D, Filter: 90%, Groups: 2

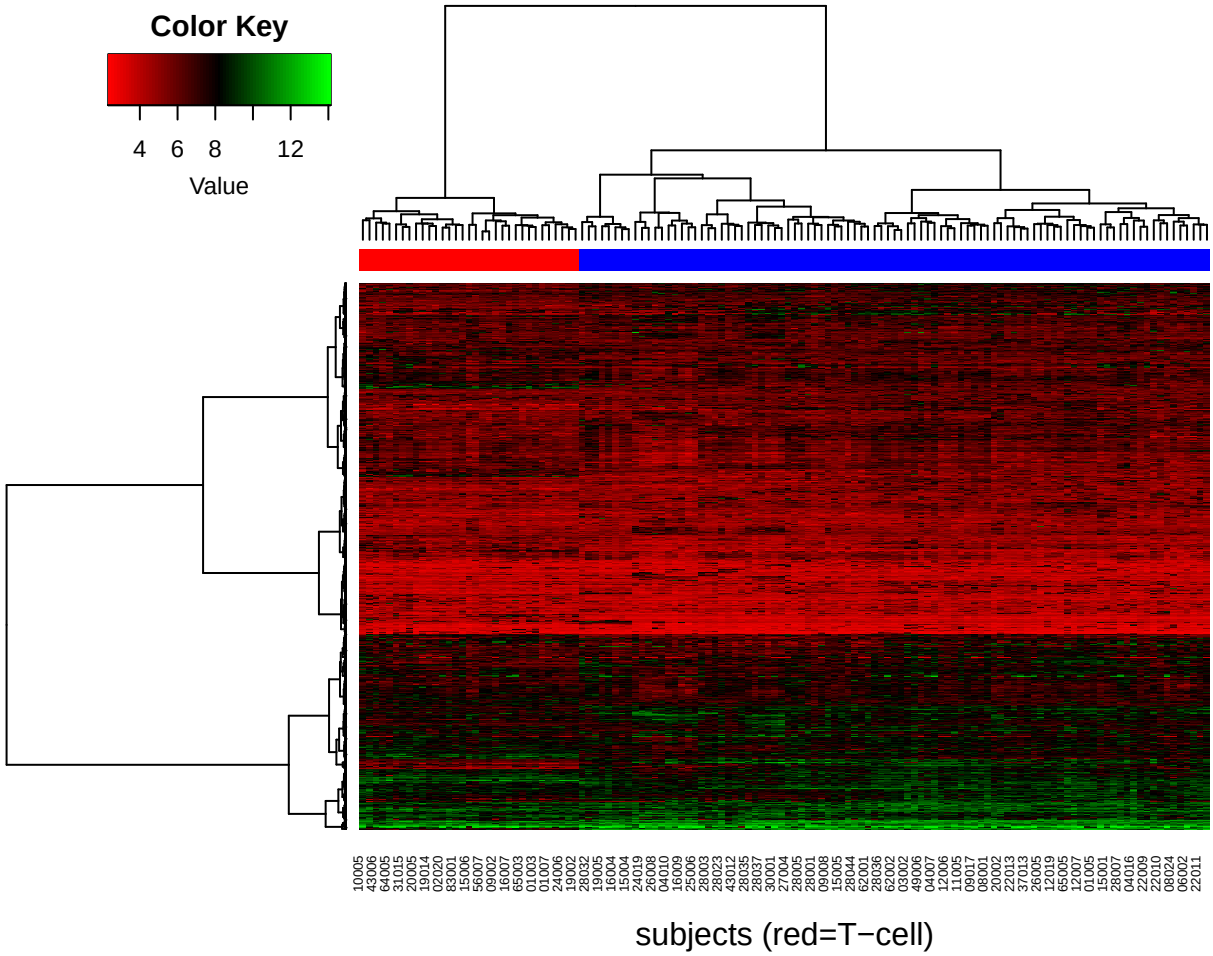
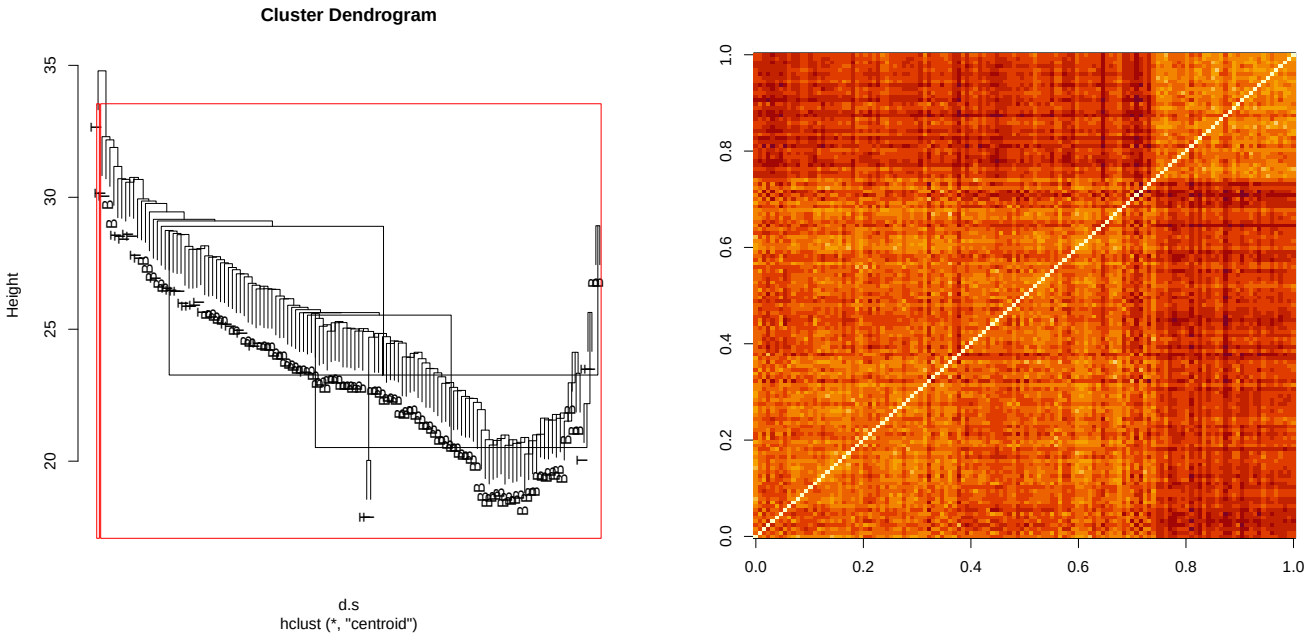


Figure 8: Heatmap of Gene expression values for genes that survived a filter of 90% and the ward.D clustering algoritm. Rows are genes and columns are subjects. There are a total of 1263 genes and 128 subjects in this plot.

8 Method: centroid, Filter: 90%, Groups: 2

```
dim(dat.filter)

## [1] 1263  128

table(groups, cl)

##       cl
## groups  B  T
##      1 95 32
##      2  0  1

fisher.test(groups, cl)$p.value

## [1] 0.26
```

9 and 9a and 9b.



(a) Dendogram (b) Distance Matrix

Figure 9: based on Method: centroid, Filter: 90%, Groups: 2

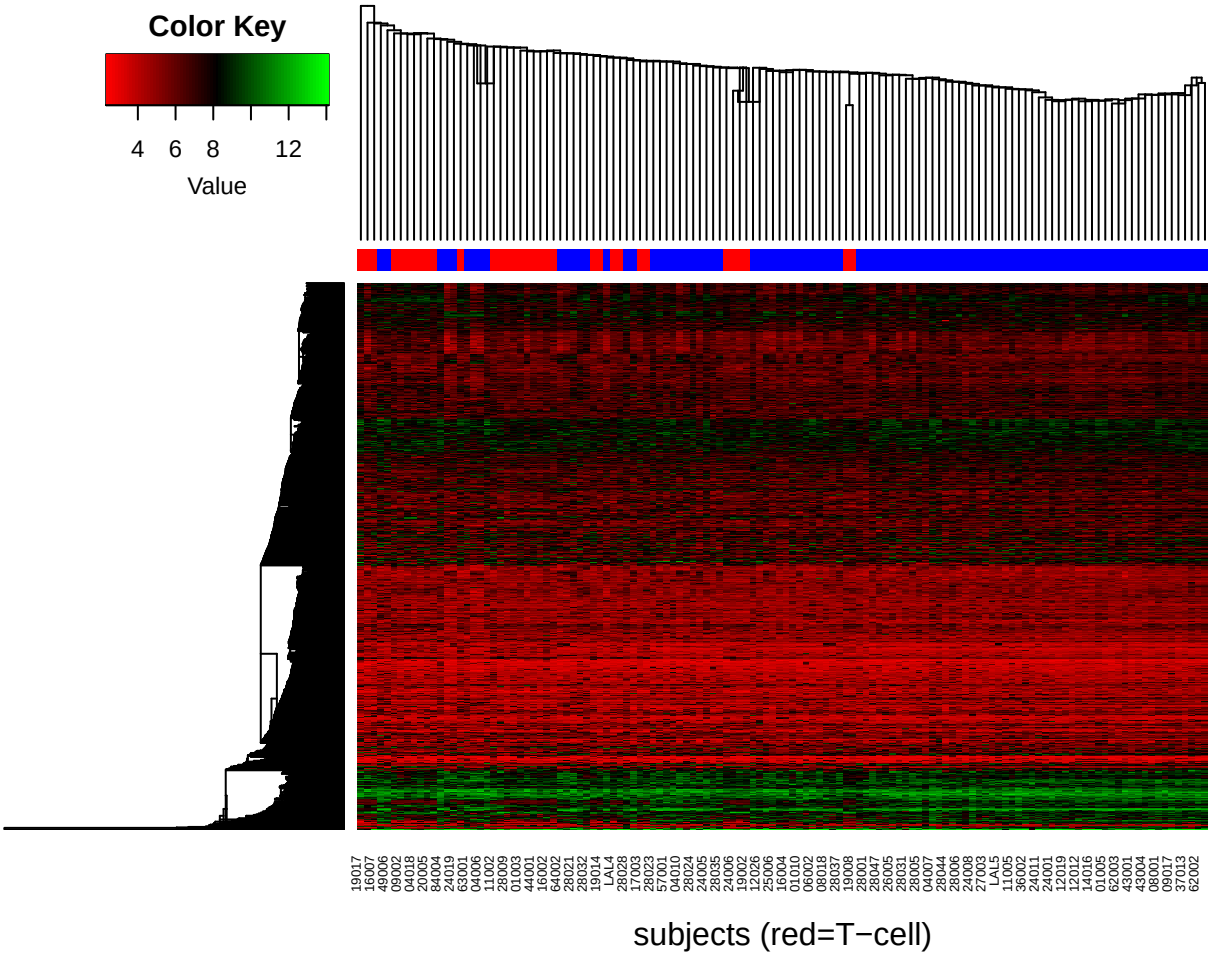
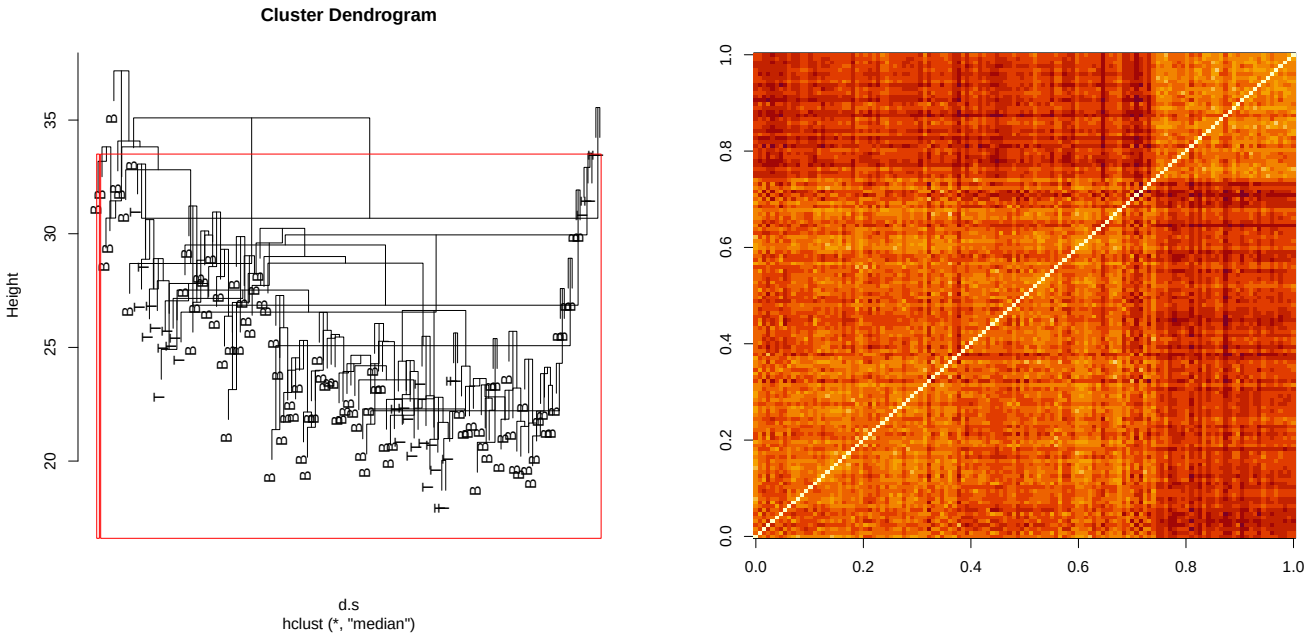


Figure 10: Heatmap of Gene expression values for genes that survived a filter of 90% and the centroid clustering algoritm. Rows are genes and columns are subjects. There are a total of 1263 genes and 128 subjects in this plot.

9 Method: median, Filter: 90%, Groups: 2

```
dim(dat.filter)

## [1] 1263  128

table(groups, cl)

##       cl
## groups  B  T
##      1 94 33
##      2  1  0

fisher.test(groups, cl)$p.value

## [1] 1
```

11 and 11a and 11b.



(a) Dendogram (b) Distance Matrix

Figure 11: based on Method: median, Filter: 90%, Groups: 2

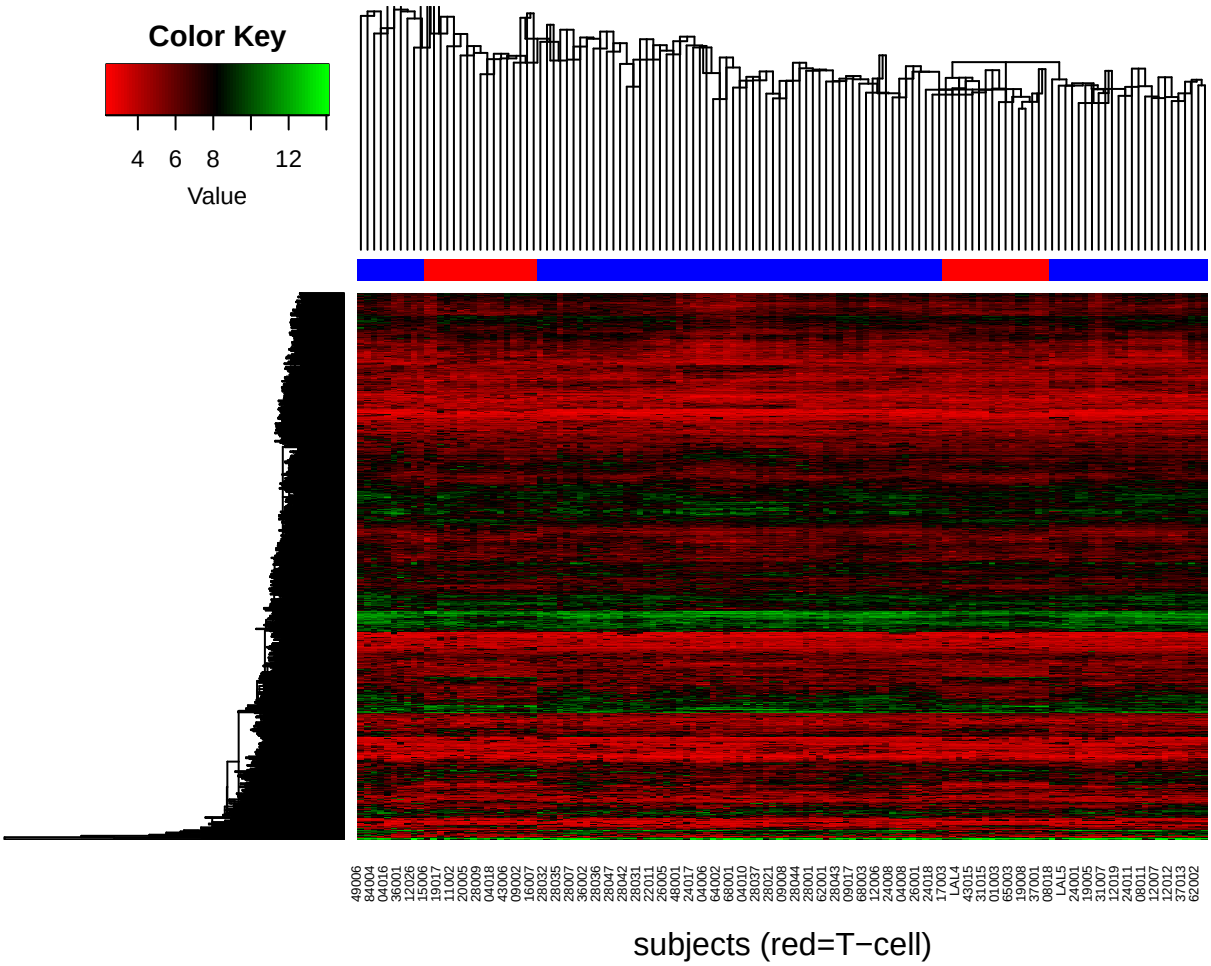
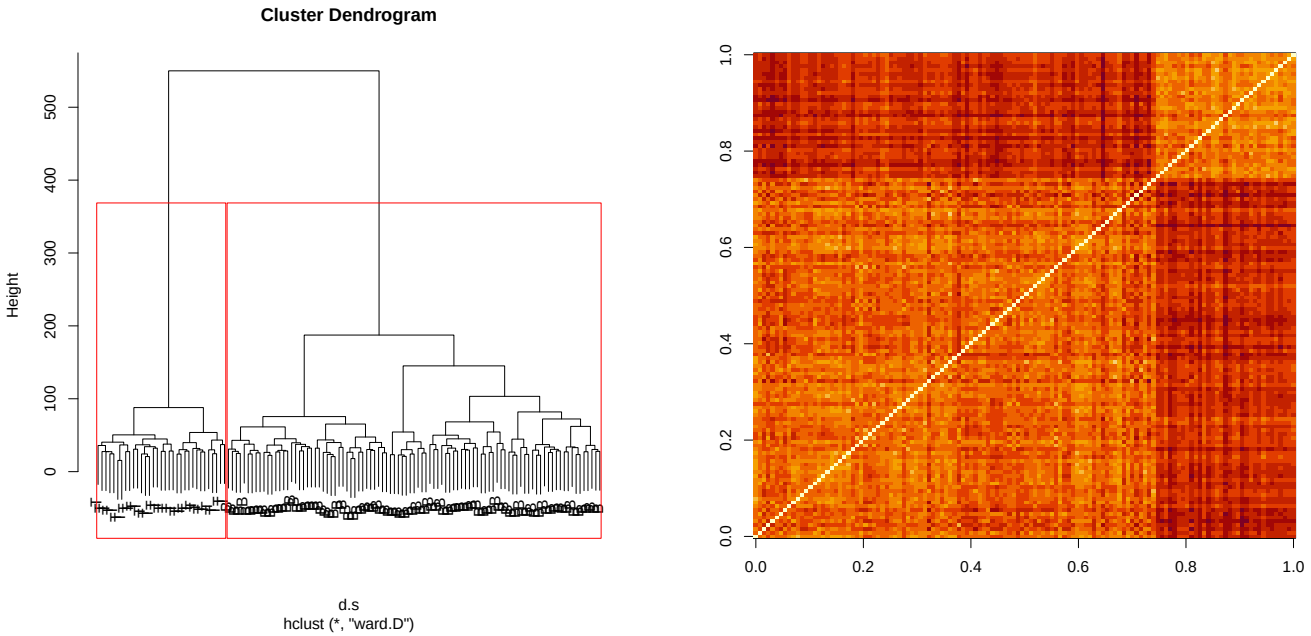


Figure 12: Heatmap of Gene expression values for genes that survived a filter of 90% and the median clustering algoritm. Rows are genes and columns are subjects. There are a total of 1263 genes and 128 subjects in this plot.

10 Method: ward.D, Filter: 95%, Groups: 2

```
dim(dat.filter)

## [1] 632 128

table(groups, cl)

##       cl
## groups  B  T
##      1 95  0
##      2  0 33

fisher.test(groups, cl)$p.value

## [1] 2.3e-31
```

13 and 13a and 13b.



(a) Dendogram (b) Distance Matrix

Figure 13: based on Method: ward.D, Filter: 95%, Groups: 2

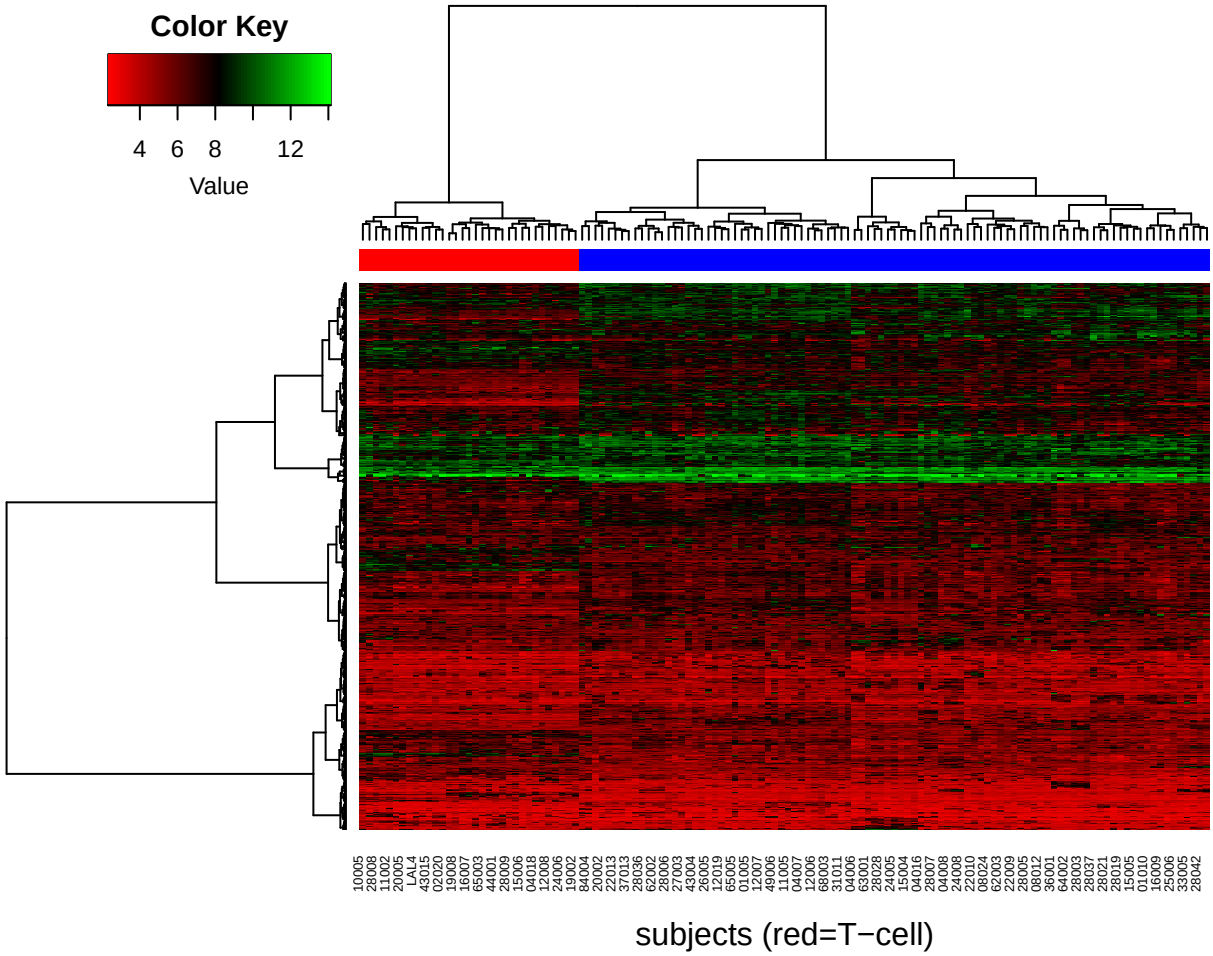
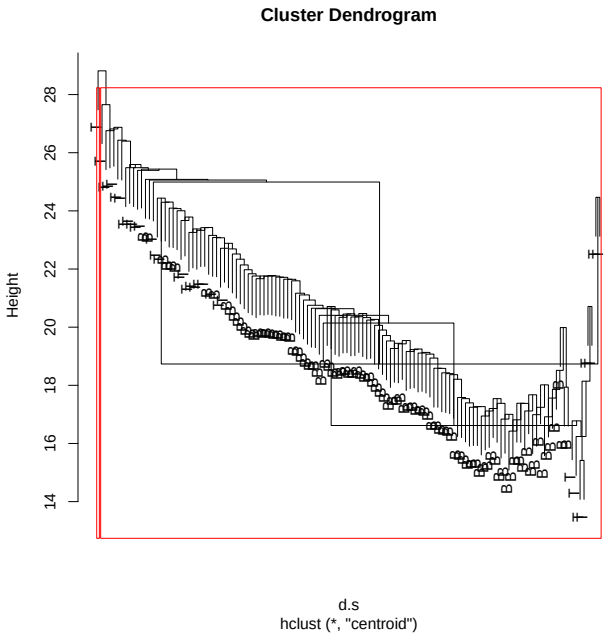


Figure 14: Heatmap of Gene expression values for genes that survived a filter of 95% and the ward.D clustering algoritm. Rows are genes and columns are subjects. There are a total of 632 genes and 128 subjects in this plot.

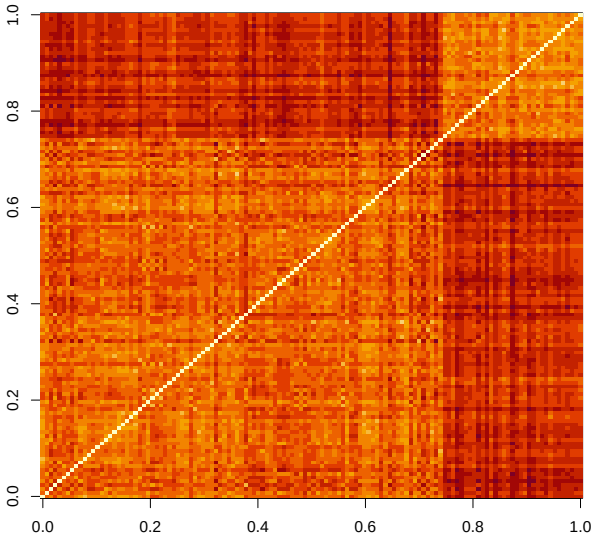
11 Method: centroid, Filter: 95%, Groups: 2

```
dim(dat.filter)

## [1] 632 128

table(groups, cl)

##        cl
## groups  B  T
##      1 95 32
##      2  0  1

fisher.test(groups, cl)$p.value

## [1] 0.26
```

15 and 15a and 15b.



(a) Dendogram



(b) Distance Matrix

Figure 15: based on Method: centroid, Filter: 95%, Groups: 2

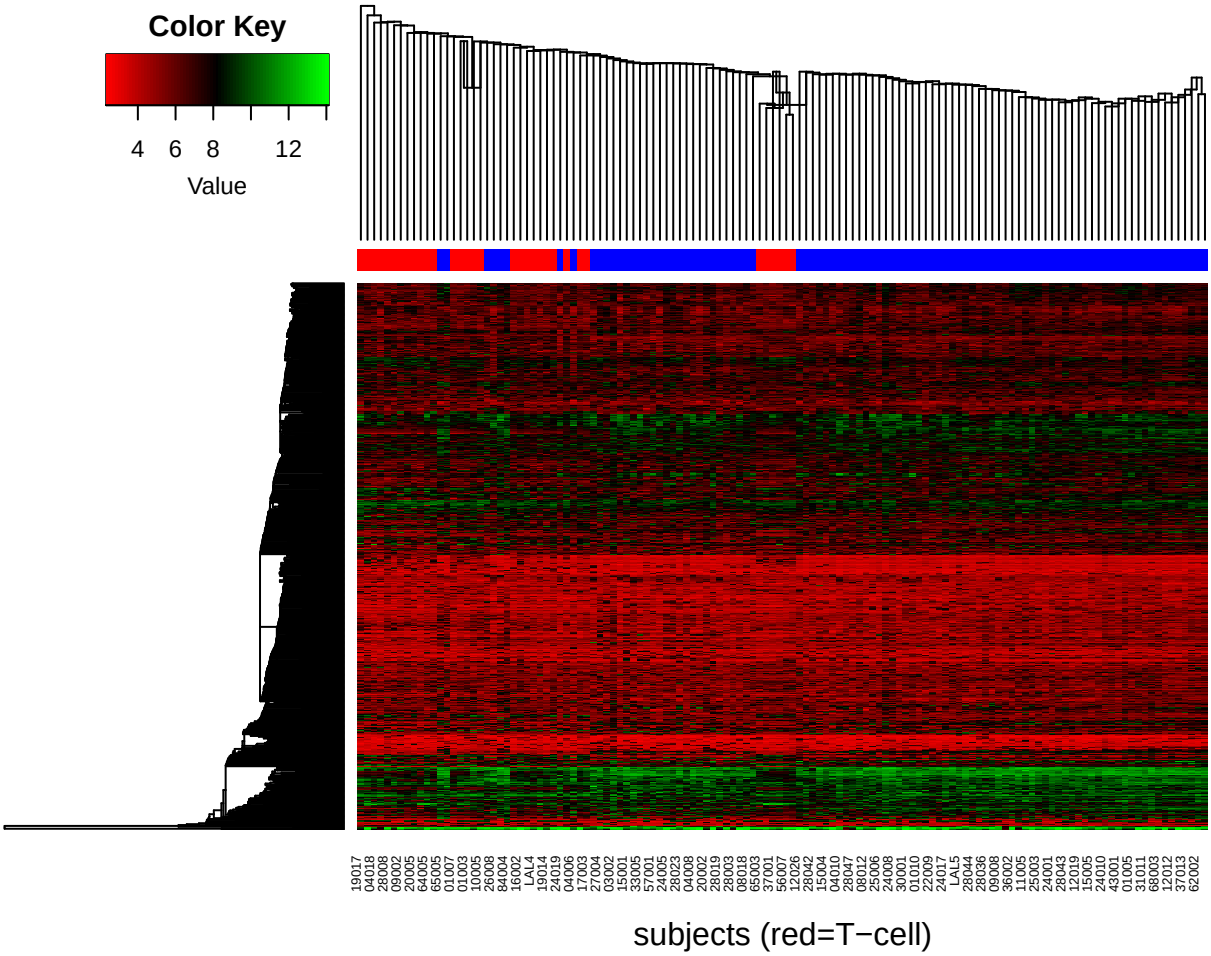
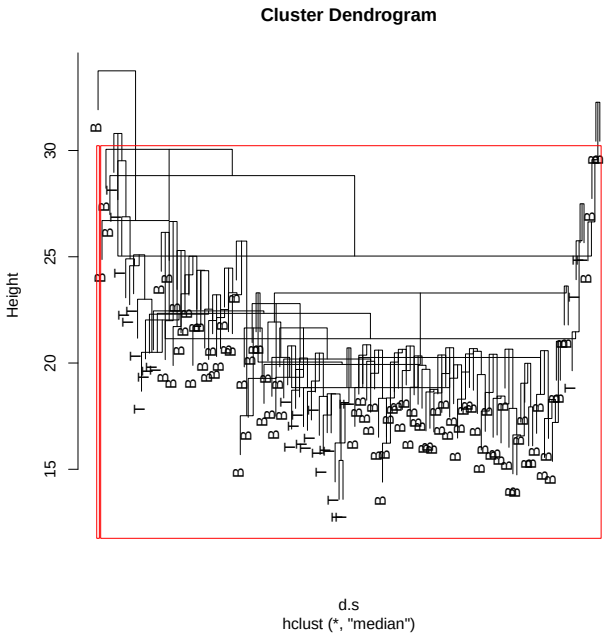


Figure 16: Heatmap of Gene expression values for genes that survived a filter of 95% and the centroid clustering algoritm. Rows are genes and columns are subjects. There are a total of 632 genes and 128 subjects in this plot.

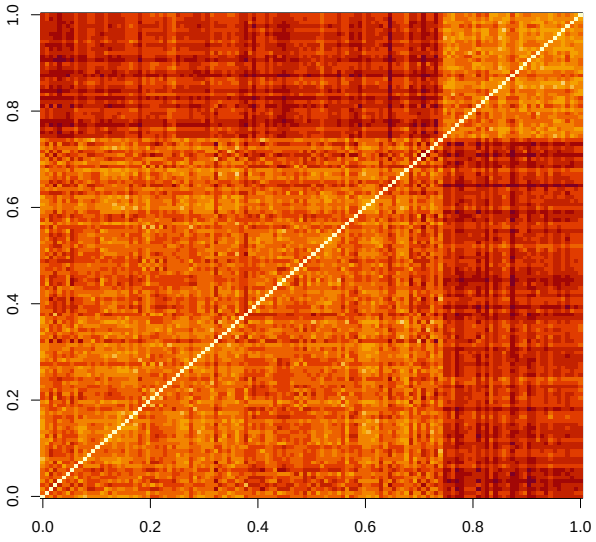
12 Method: median, Filter: 95%, Groups: 2

```
dim(dat.filter)

## [1] 632 128

table(groups, cl)

##        cl
## groups  B  T
##      1 94 33
##      2  1  0

fisher.test(groups, cl)$p.value

## [1] 1
```

17 and 17a and 17b.



(a) Dendogram



(b) Distance Matrix

Figure 17: based on Method: median, Filter: 95%, Groups: 2

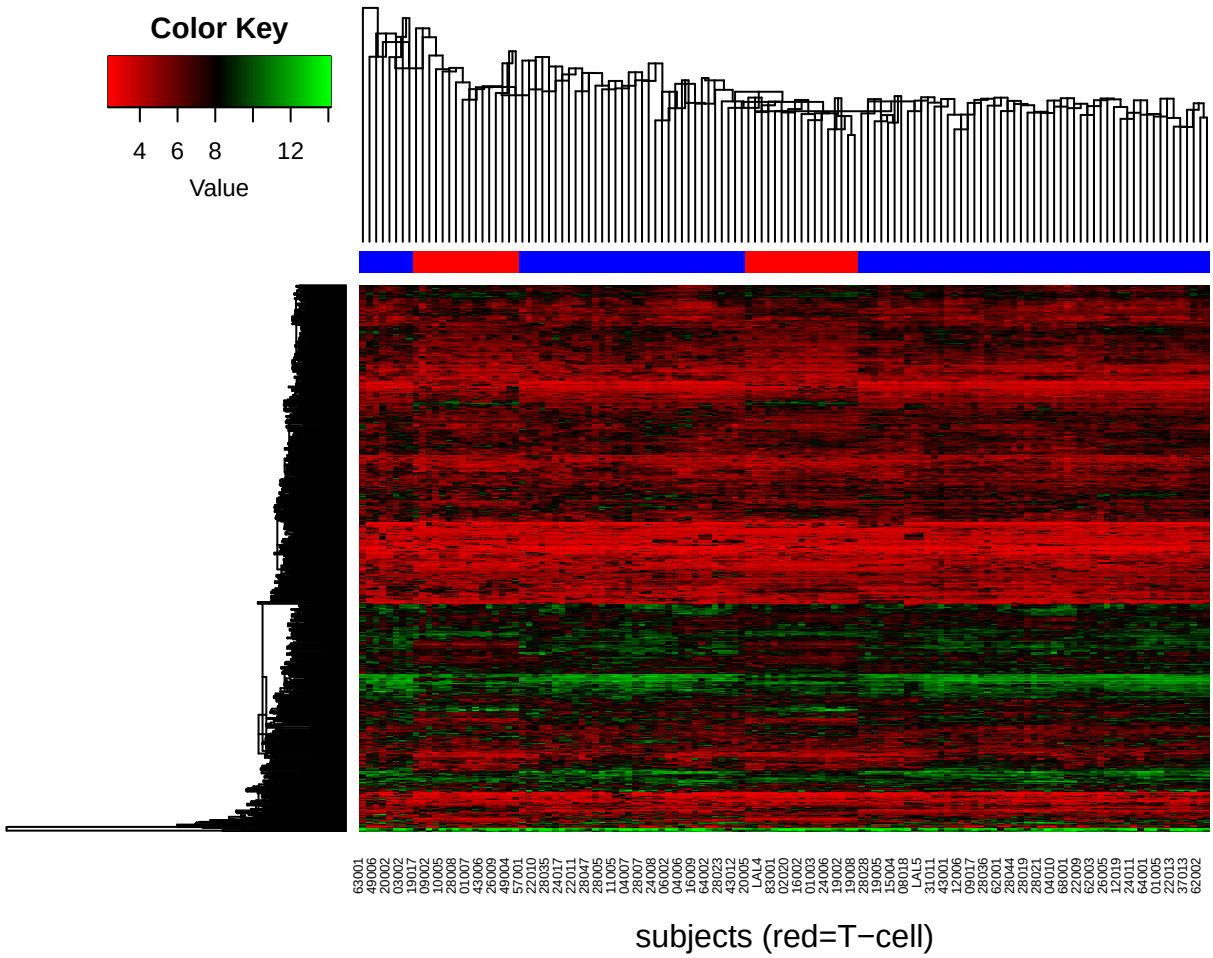


Figure 18: Heatmap of Gene expression values for genes that survived a filter of 95% and the median clustering algoritm. Rows are genes and columns are subjects. There are a total of 632 genes and 128 subjects in this plot.

References

Sabina Chiaretti, Xiaochun Li, Robert Gentleman, Antonella Vitale, Marco Vignetti, Franco Mandelli, Jerome Ritz, and Robin Foa. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778, 2004. 1

Sandrine Dudoit and Robert Gentleman. Cluster Analysis in DNA Microarray Experiments. Technical report, 2002. URL http://www.bioconductor.org/help/course-materials/2002/Seattle02/Cluster/cluster.pdf. 1

Xiaochun Li. *ALL: A data package*, 2009. R package version 1.10.0. 1

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2013. URL http://yihui.name/knitr/. ISBN 978-1482203530. 1

Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. URL http://www.crcpress.com/product/isbn/9781466561595. ISBN 978-1466561595. 1

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2015. URL http://yihui.name/knitr/. R package version 1.10.5. 1

A Session Information

```
print(sessionInfo(), locale = FALSE)
```

```
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Pop!_OS 18.10
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.8.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.8.0
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## other attached packages:
## [1] gplots_3.0.1.1 here_0.1       pacman_0.5.0
## [4] knitr_1.22
##
## loaded via a namespace (and not attached):
##  [1] gtools_3.8.1        rprojroot_1.3-2
##  [3] bitops_1.0-6        backports_1.1.3
##  [5] formatR_1.6         magrittr_1.5
##  [7] evaluate_0.13       highr_0.8
##  [9] KernSmooth_2.23-15  stringi_1.4.3
## [11] gdata_2.18.0        tools_3.6.0
## [13] stringr_1.4.0       Biobase_2.42.0
## [15] parallel_3.6.0      xfun_0.6
## [17] compiler_3.6.0      BiocGenerics_0.28.0
## [19] caTools_1.17.1.1
```