007-Sensitivity Analysis of Many Paramters

Clustering Gene Expression Data

May 27, 2015

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`.

Contents

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

```
dim(dat.filter)
```

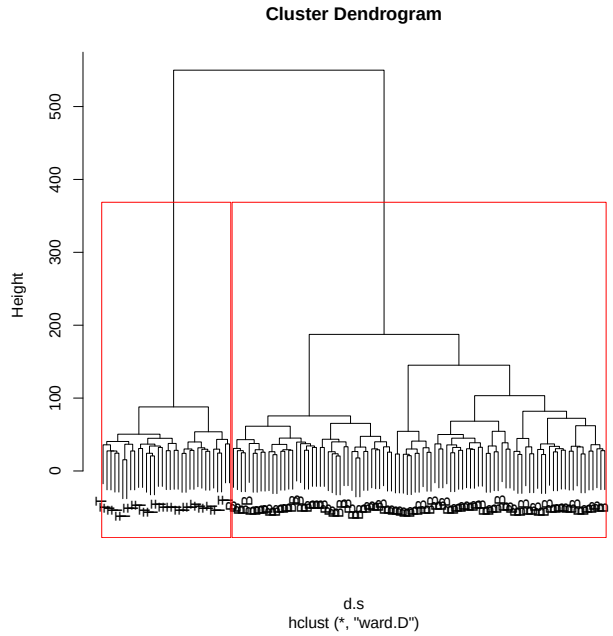
```
## [1] 632 128
```

```
table(groups, cl)
```

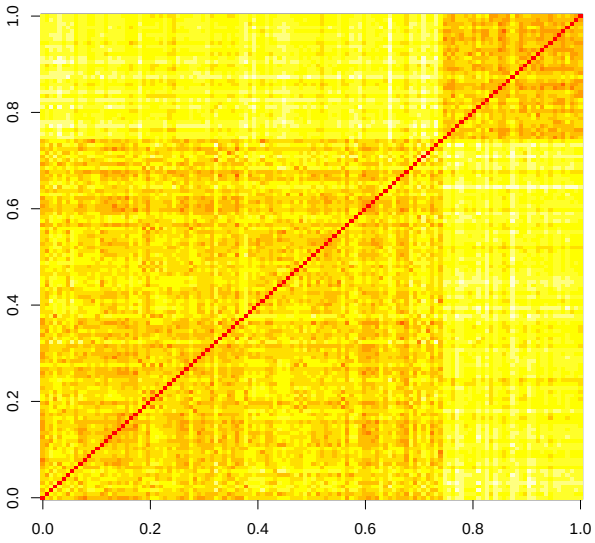
```
##        cl
## groups  B  T
##      1 95  0
##      2  0 33
```

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

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



(a) Dendogram



(b) Distance Matrix

Figure 1: based on Method: ward.D, Filter: 95%, Groups: 2

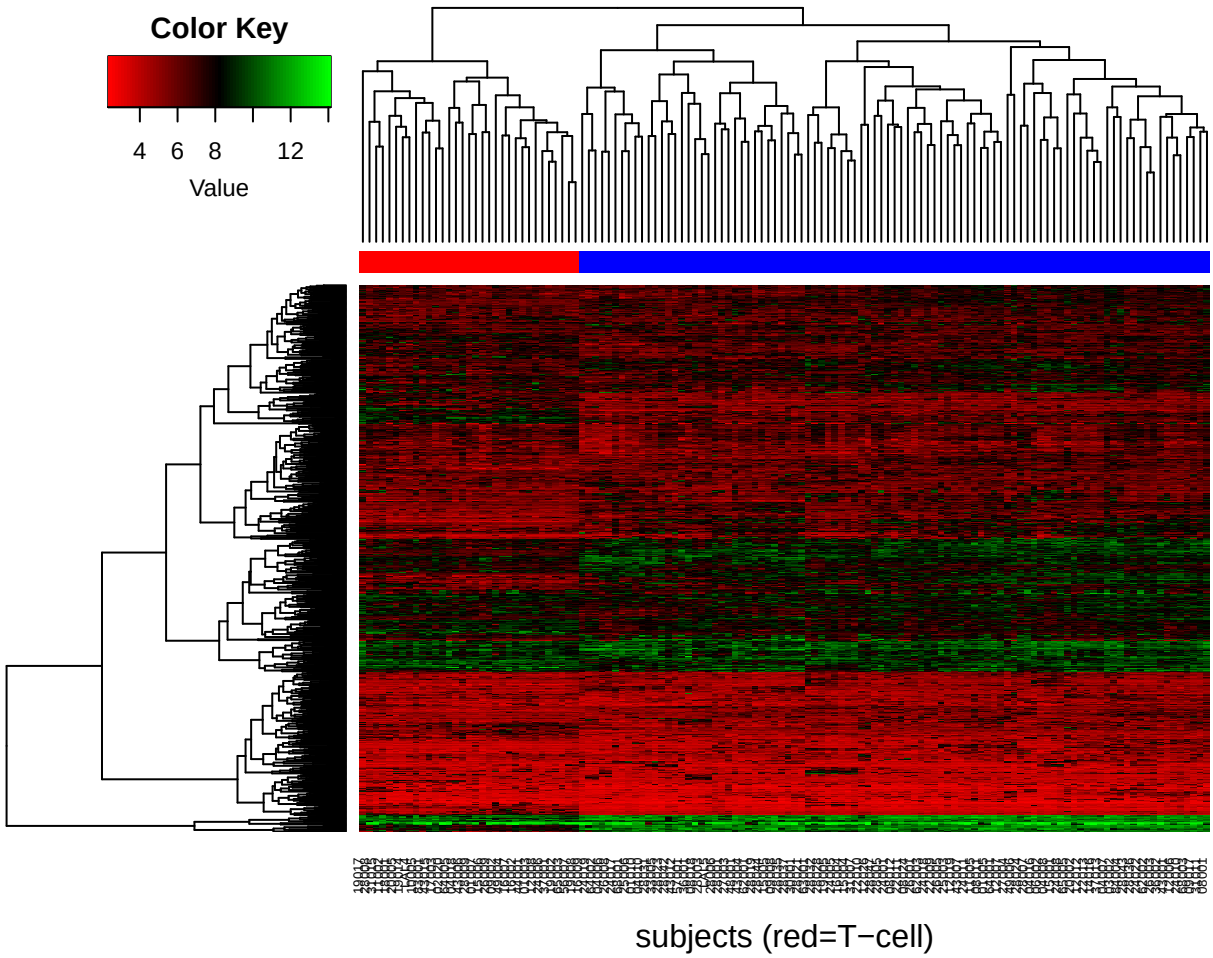
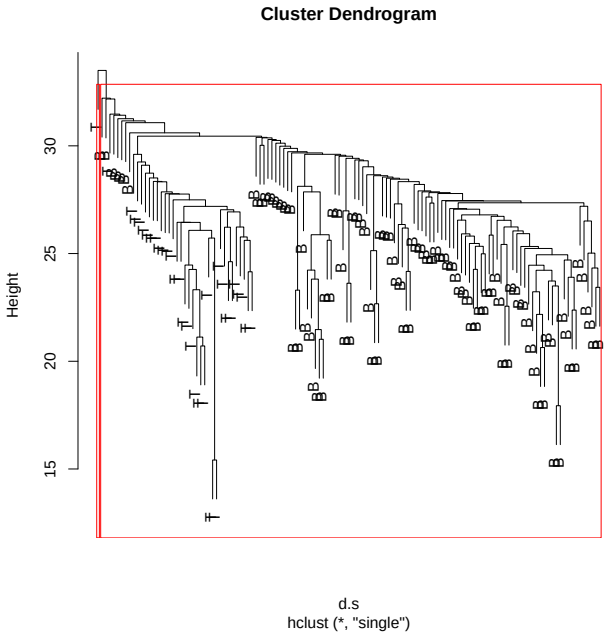


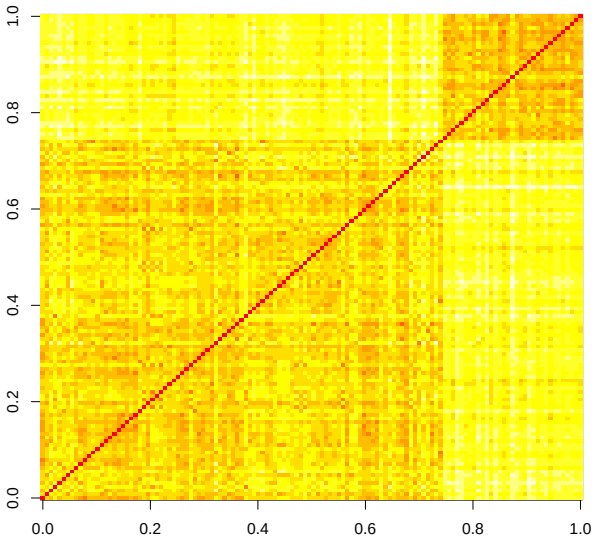
Figure 2: Heatmap of Gene expression values for genes that survived a filter of 95%

2 Method: single, 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
```



(a) Dendogram



(b) Distance Matrix

Figure 3: based on Method: single, Filter: 95%, Groups: 2

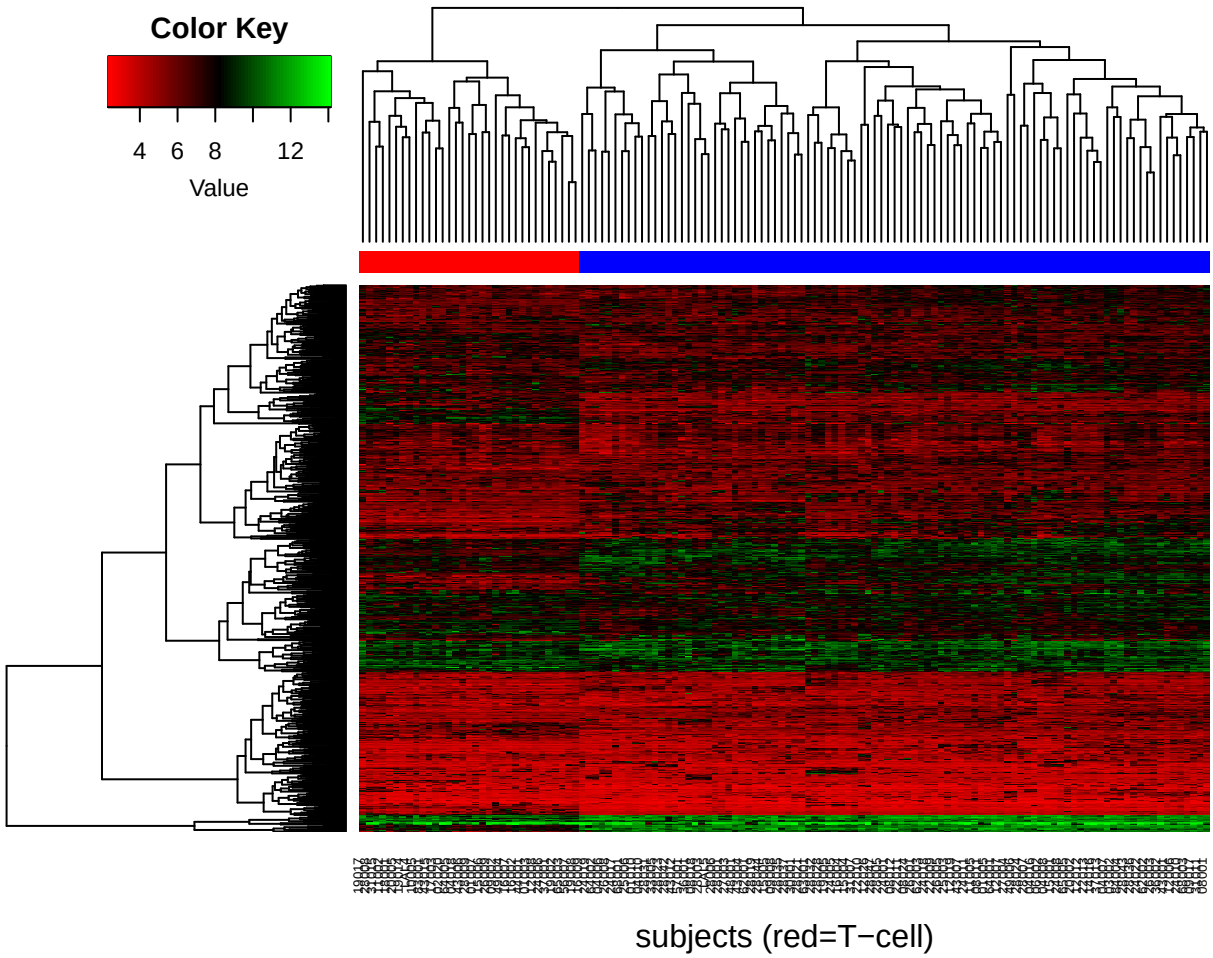
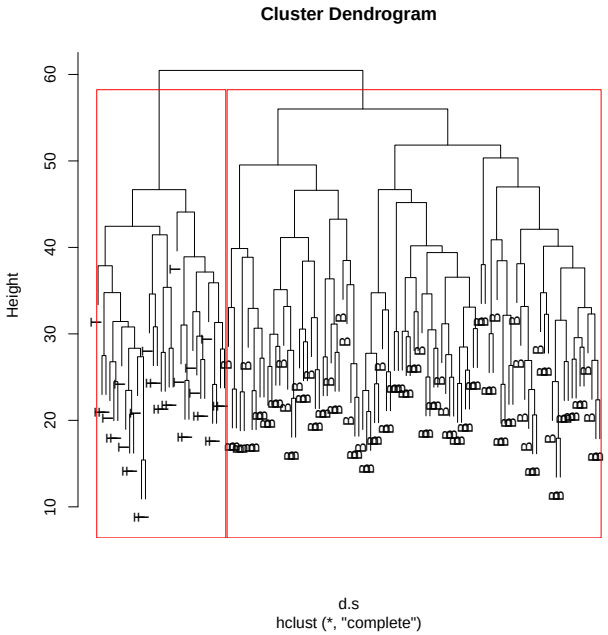


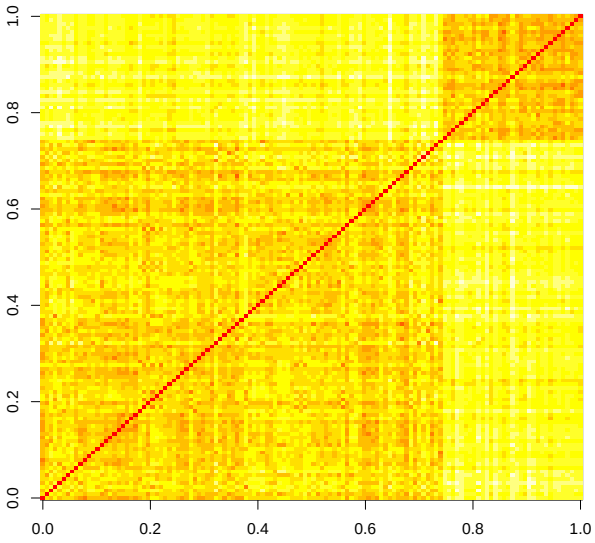
Figure 4: Heatmap of Gene expression values for genes that survived a filter of 95%

3 Method: complete, 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
```



(a) Dendogram



(b) Distance Matrix

Figure 5: based on Method: complete, Filter: 95%, Groups: 2

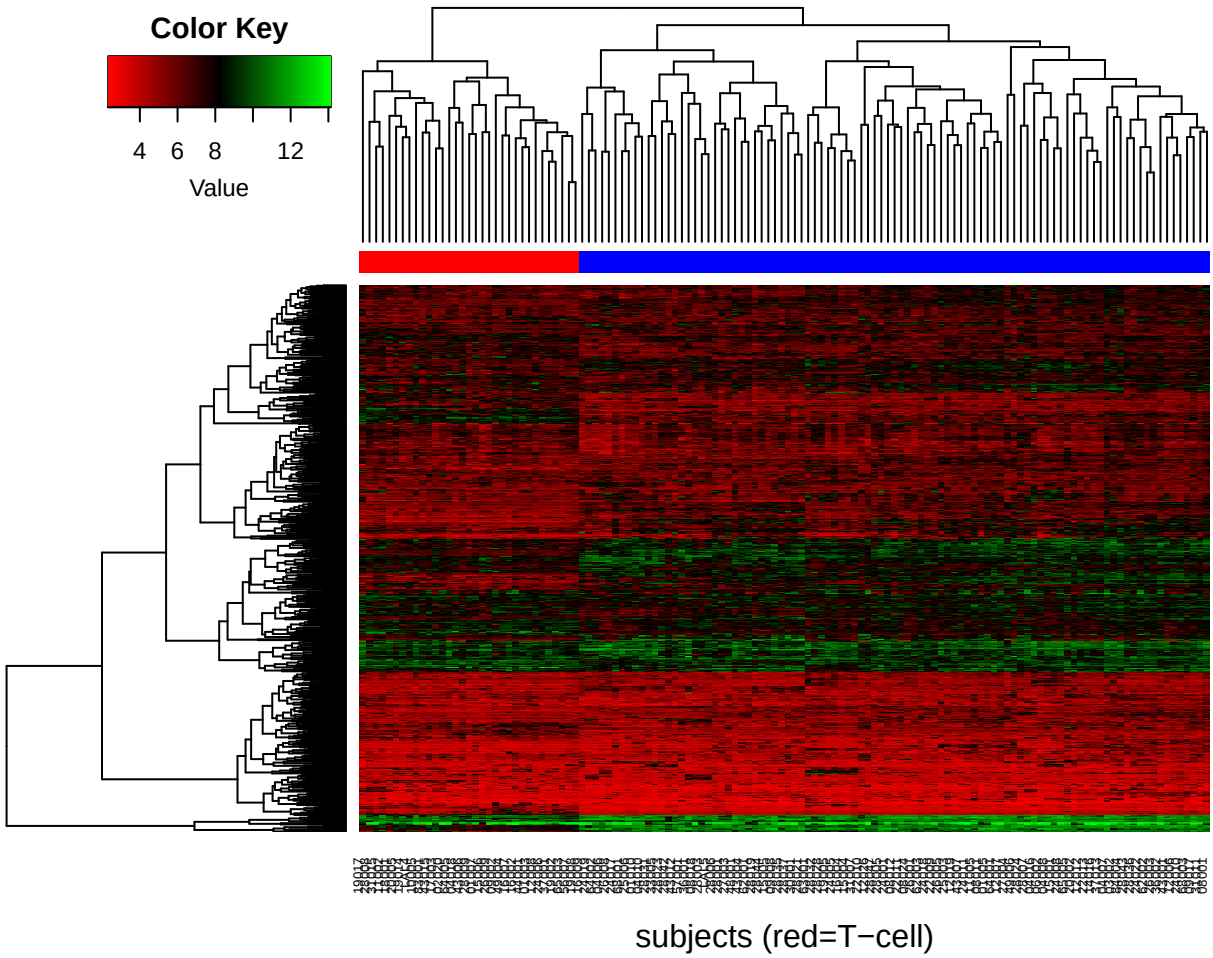
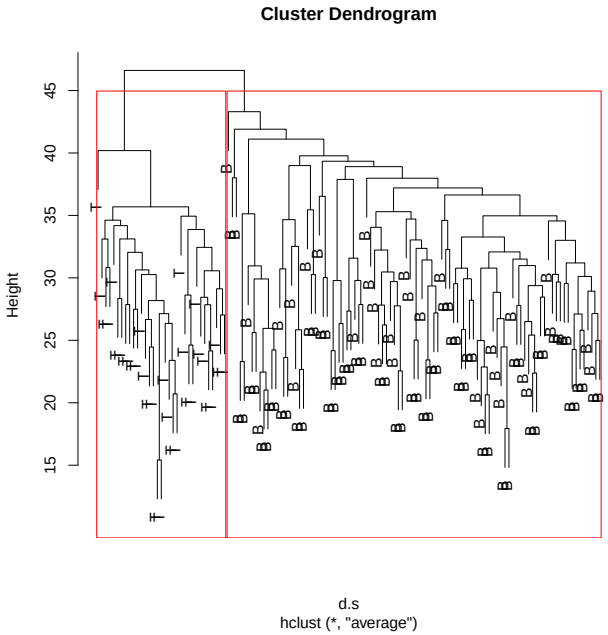


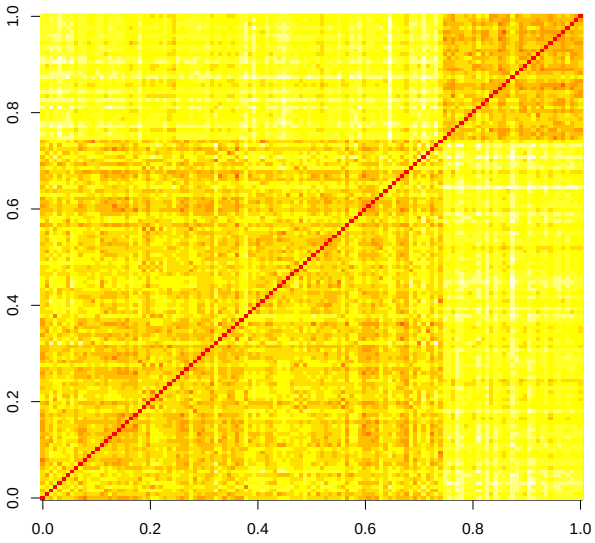
Figure 6: Heatmap of Gene expression values for genes that survived a filter of 95%

4 Method: average, 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
```



(a) Dendogram



(b) Distance Matrix

Figure 7: based on Method: average, Filter: 95%, Groups: 2

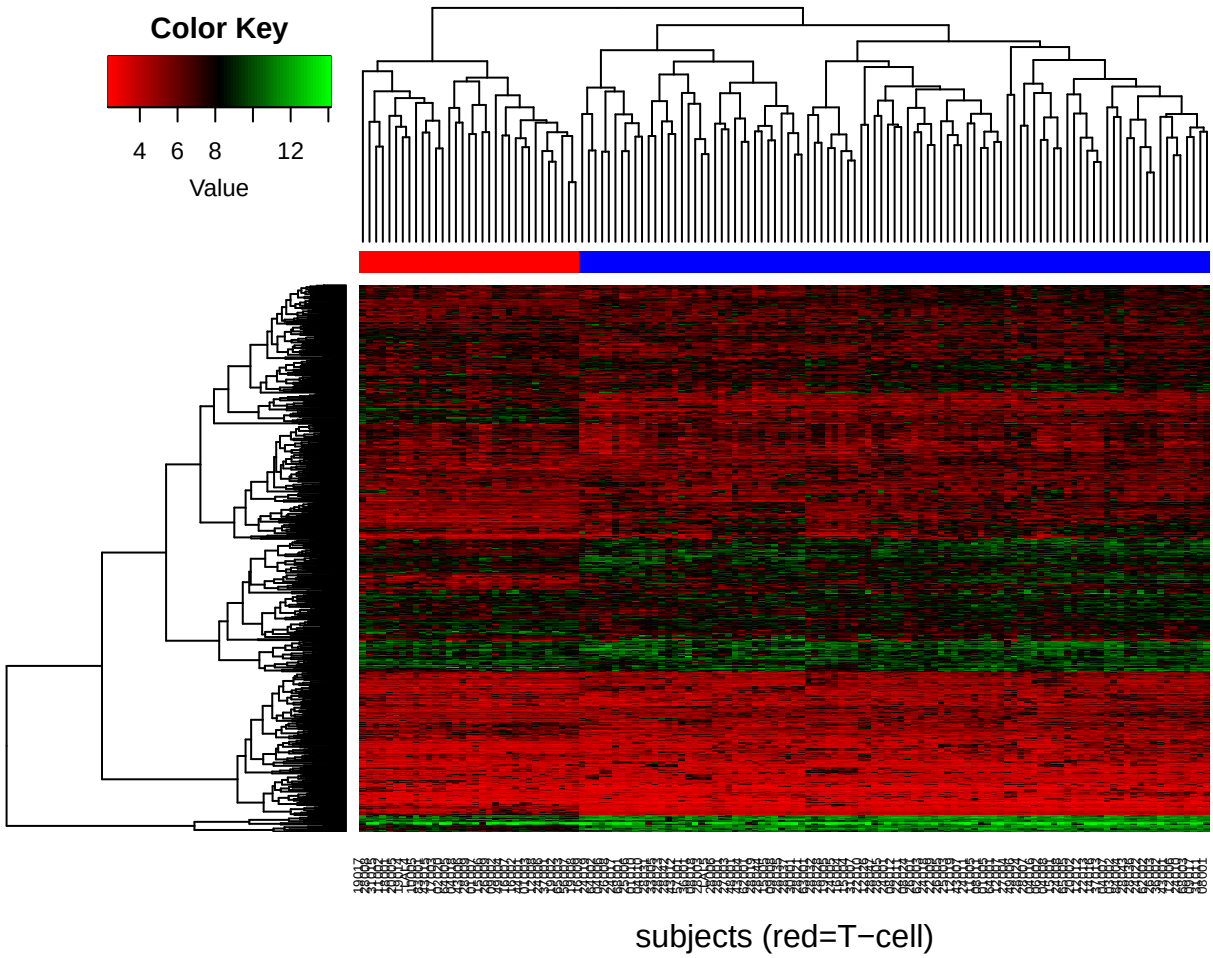
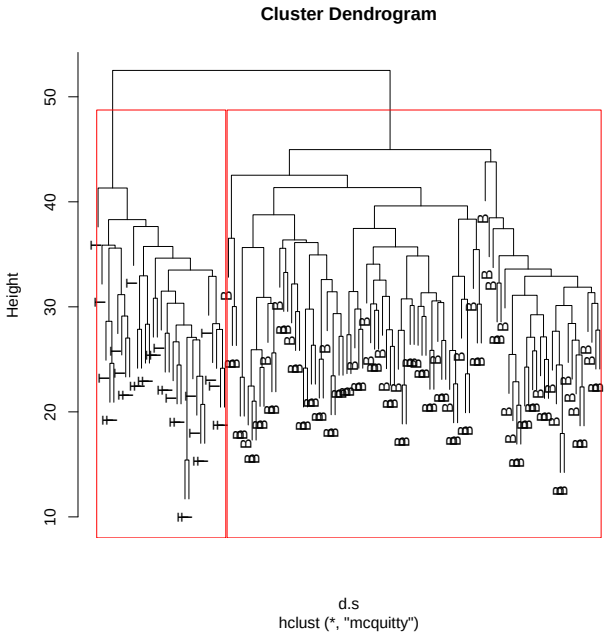


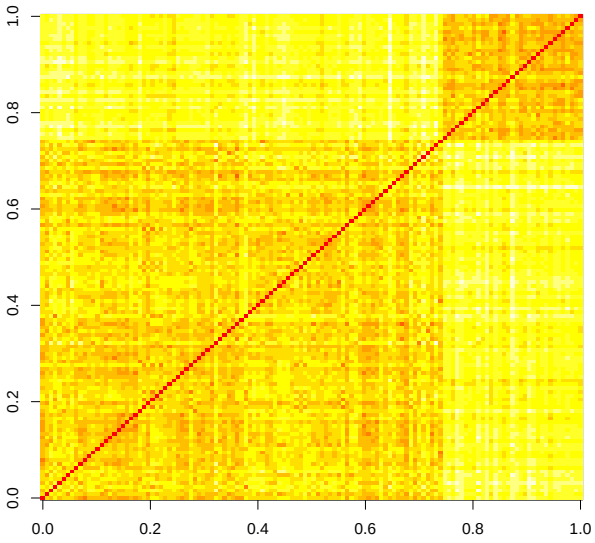
Figure 8: Heatmap of Gene expression values for genes that survived a filter of 95%

5 Method: mcquitty, 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
```



(a) Dendogram



(b) Distance Matrix

Figure 9: based on Method: mcquitty, Filter: 95%, Groups: 2

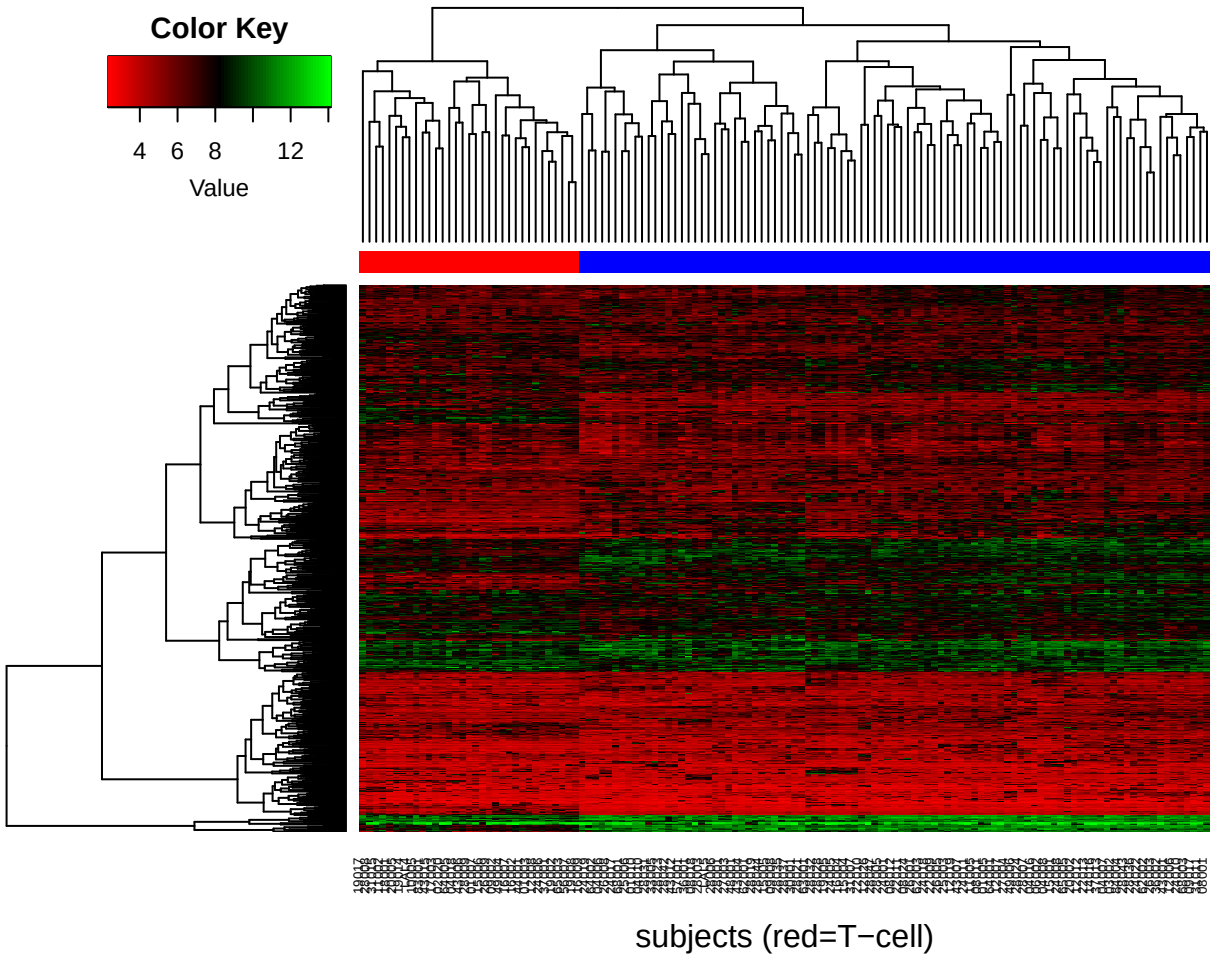
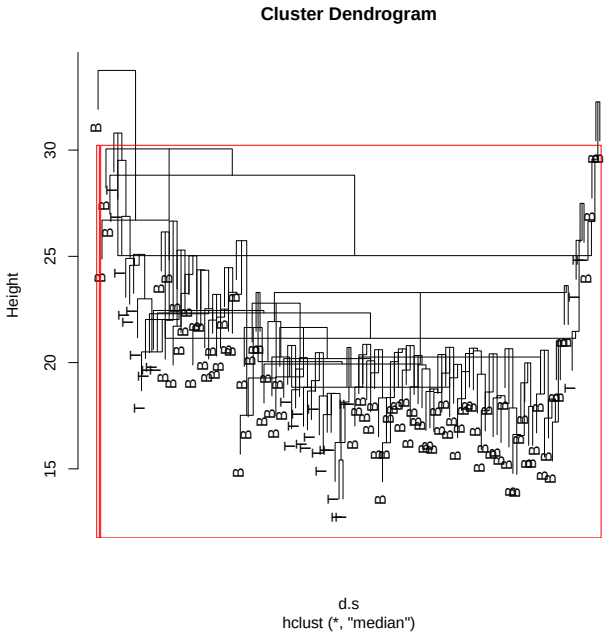


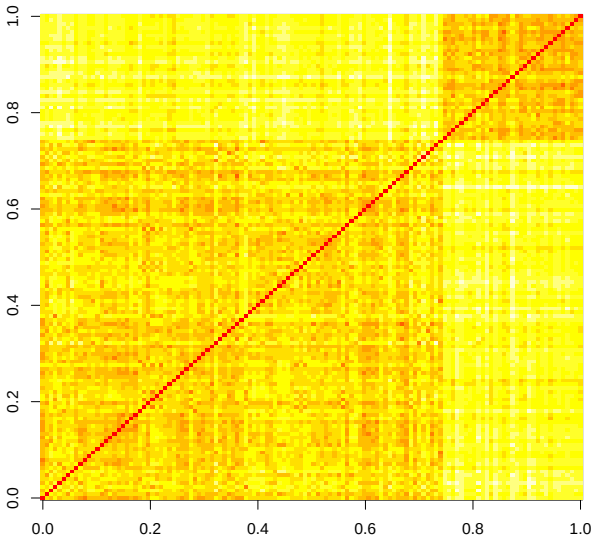
Figure 10: Heatmap of Gene expression values for genes that survived a filter of 95%

6 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
```



(a) Dendogram



(b) Distance Matrix

Figure 11: based on Method: median, Filter: 95%, Groups: 2

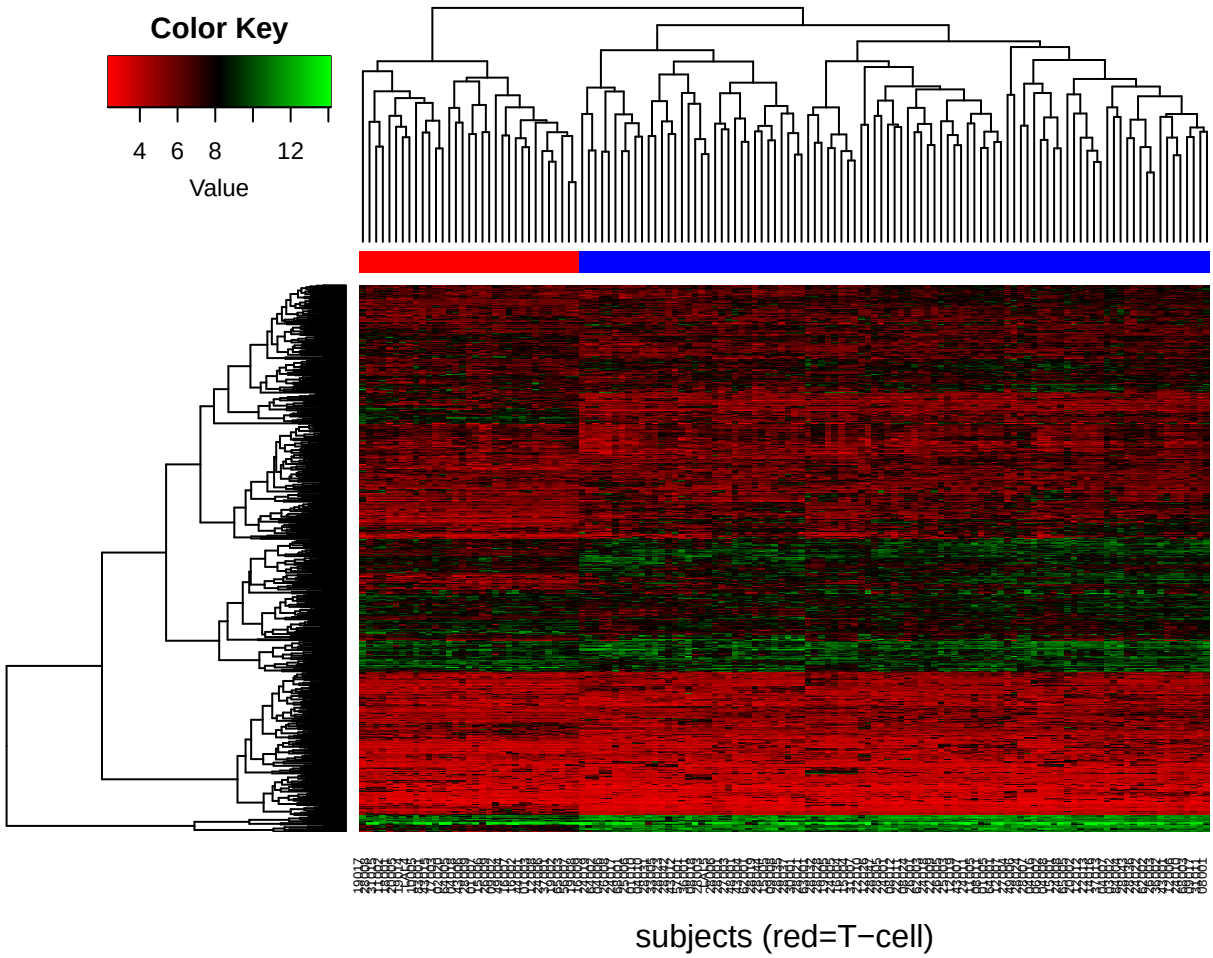
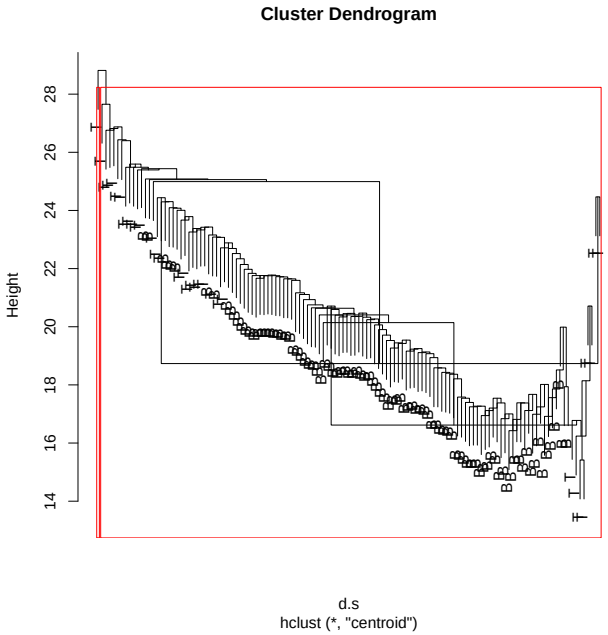


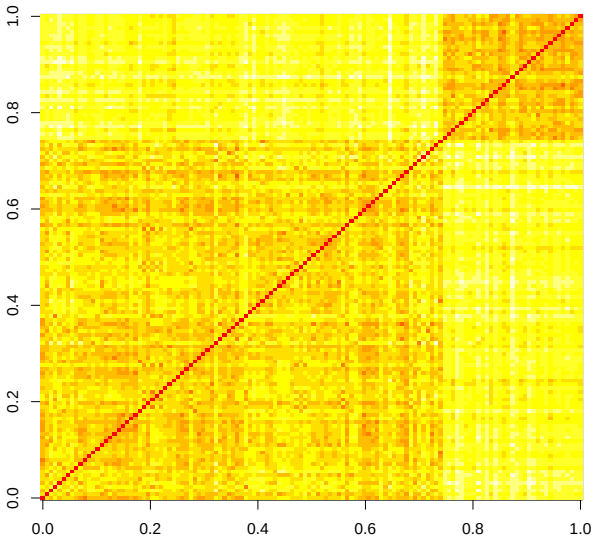
Figure 12: Heatmap of Gene expression values for genes that survived a filter of 95%

7 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
```



(a) Dendogram



(b) Distance Matrix

Figure 13: based on Method: centroid, Filter: 95%, Groups: 2

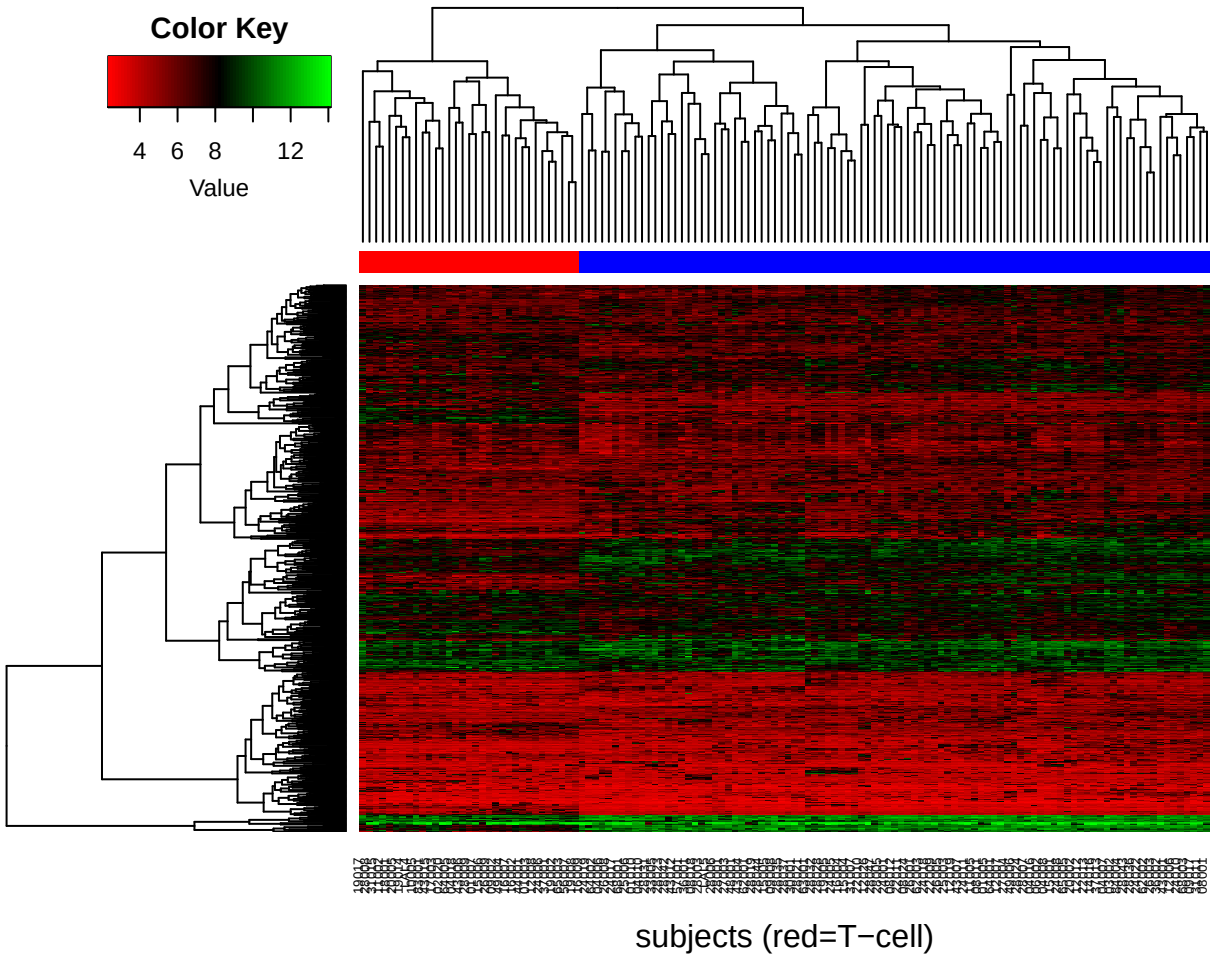


Figure 14: Heatmap of Gene expression values for genes that survived a filter of 95%

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