

## Identification and Removal of Outlier Samples (Illumina)

### Supplement for:

### "Functional Organization of the Transcriptome in Human Brain"

Michael C. Oldham, Steve Horvath, Genevieve Konopka, Kazuya

Iwamoto, Peter Langfelder, Tadafumi Kato, and Daniel H. Geschwind

#### Summary

Here we present additional details on the microarray data pre-processing steps performed prior to the construction of gene coexpression networks in our study, "*Functional Organization of the Transcriptome in Human Brain*". To ensure full reproducibility of our research findings, below we provide an annotated supplement that contains all of the relevant R code and corresponding figure images that were used to guide our decisions to remove outlier samples in a previously processed Illumina microarray dataset consisting of 193 samples from human cerebral cortex<sup>1</sup>. This dataset ("CTX\_ILMN") was analyzed to provide additional validation across platforms and individuals for the significance of gene coexpression relationships in human cerebral cortex identified in our study.

Since network analysis and module detection can be severely biased by the presence of outlying microarray samples, it is important to carry out pre-processing steps to identify and remove such samples in each dataset prior to network construction. Our main statistical diagnostic for flagging potential outlying samples in this dataset was the inter-array correlation (IAC), which was defined as the Pearson correlation coefficient of the expression levels for a given pair of microarrays (using all probe sets for which complete data were available). The exclusion of samples purely on the basis of IACs represents an unbiased method for the identification and removal of microarray samples with divergent gene expression levels. The distribution of IACs within a dataset can be visualized as a histogram (frequency plot), while the relationships between arrays can be visualized as a dendrogram using average linkage hierarchical clustering with 1-IAC as a distance metric.

Ideally, microarray pre-processing steps begin with the analysis of raw data. Unlike the other microarray datasets analyzed in our study, however, raw data for CTX\_ILMN were unavailable. Therefore, the identification and removal of outlier samples in CTX\_ILMN was performed

using data that had been previously normalized by the authors of the original study via the rank-invariant normalization method offered by Illumina's BeadStudio software<sup>1</sup>. These data consisted of expression levels for 14,078 transcripts that were detected in  $\geq 5\%$  of all 193 samples<sup>1</sup>, with undetected transcripts coded as missing values (detailed sample information can be found in ref. 1). From this list, we selected 5,269 transcripts with no missing values for further pre-processing and network analysis.

Samples in CTX\_ILMN that exhibited divergent clustering and/or samples with low mean IACs ( $n=34/193$  [18%]) were excluded, and the mean IAC after removing all outlier samples and performing quantile normalization<sup>2</sup> was 0.943. We note that this quantity is lower than the mean IACs for the other datasets analyzed in this study (0.970 - 0.975). This discrepancy may reflect uncorrected batch effects or other technical artifacts in the present analysis, increased sample heterogeneity, or other factors. All analyses described below were performed in R.

#### Data Description

Dataset	Arrays	# samples before pre-processing	# samples after pre-processing	Sample description*
CTX_ILMN	Illumina HumanRefseq-8	193	159	cerebral cortex

\* For additional sample information, see Supplementary Table 1 from ref. 1.

#### CTX\_ILMN

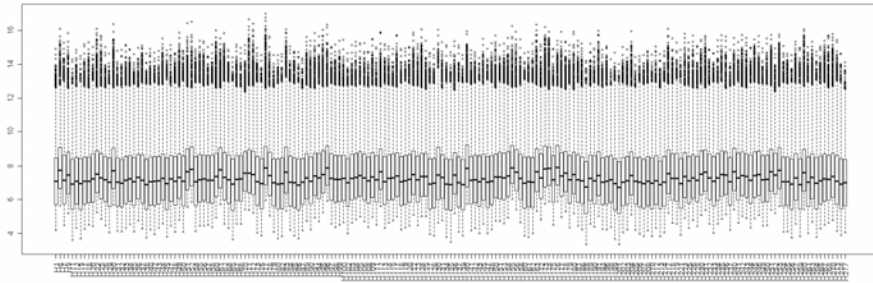
## Reading in the previously normalized expression data (14,078 probe sets, 193 samples; columns 1-3 contain probe set information and column 197 contains the number of missing values for each probe set):

```
library(cluster)
library(affy)
```

```
dat1=read.csv("CTXILMN_193samples_normalized_expression_
data.csv",header=T)
dim(dat1)
# [1] 14078 197
dimnames(dat1)[[2]]
```

```
## First we will examine the overall distribution of
expression values:
```

```
boxplot(log(dat1[,4:196],2),las=3,cex=0.7)
```

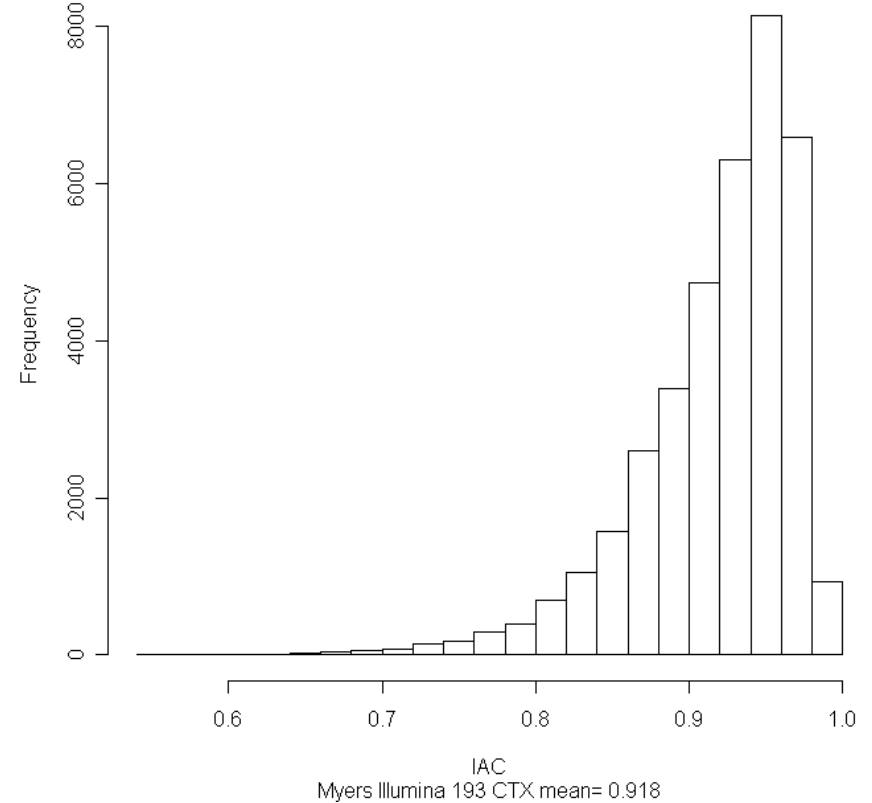


```
## There appears to be a fair amount of variation in the
distribution. Calculating IACs for all pairs of samples
and examining the distribution of IACs in the dataset:
```

```
IAC=cor(dat1[,4:196],use="p")
```

```
hist(IAC,sub=paste("Myers Illumina 193 CTX
mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

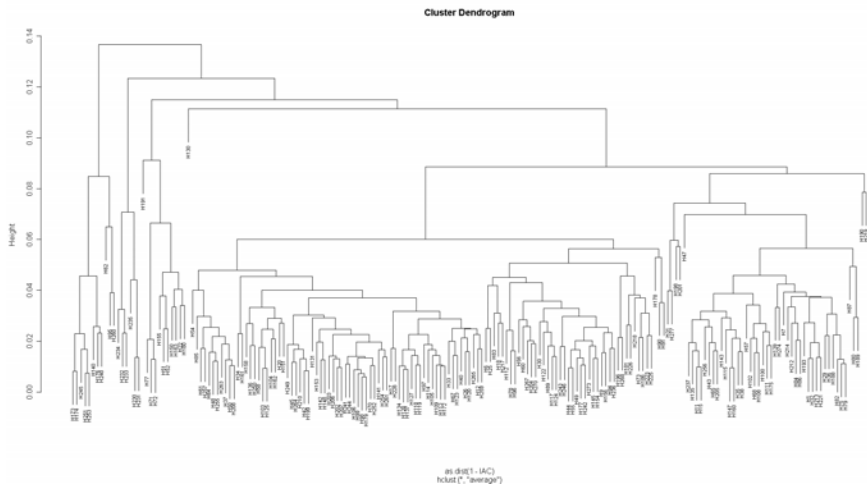
**Histogram of IAC**



```
## Here we see that the mean IAC in the initial CTX_ILMN
dataset, with no outlier samples removed, is 0.918.
There is a long tail to the left of the distribution,
indicating the presence of possible outliers.
```

```
## Performing hierarchical clustering (average linkage)
using 1-IAC as a distance metric:
```

```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat1)[[2]][4:196])
```



```
## This dendrogram suggests that there are outliers in
the dataset (notably at left). However, we would like
to further restrict the dataset to include only
transcripts with no missing values (n=5,269):
```

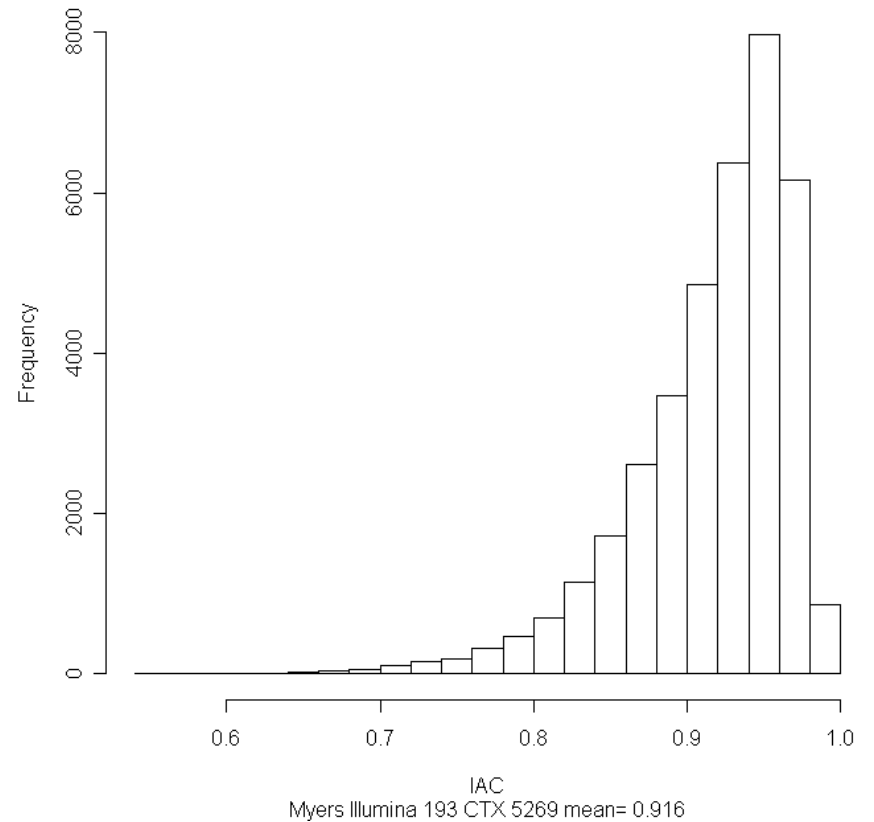
```
datrest=as.matrix(dat1[dat1$No_NaN<1,4:196])
dim(datrest)
# [1] 5269 193
```

```
## Repeating the previous steps:
```

```
IAC=cor(datrest,use="p")
```

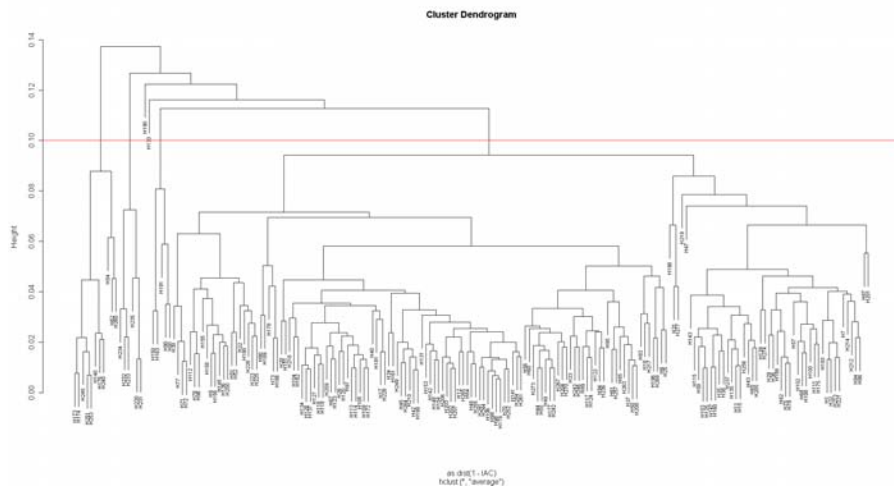
```
hist(IAC,sub=paste("Myers Illumina 193 CTX 5269
mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

**Histogram of IAC**



```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(datrest)[[2]])
abline(h=0.1,col="red")
```

```
## Again, we see a long tail to the left of the
distribution. Clustering:
```



## We will remove the outlying samples at the left of the dendrogram (n=24) by "cutting" the tree at a specified height (red line). These samples were excluded as follows:

```
removevec=c("H54","H62","H90","H91","H130","H140","H150",
,"H163","H173","H174","H176","H190","H191","H230","H231",
,"H232","H233","H234","H235","H240","H245","H251","H253",
,"H260")
```

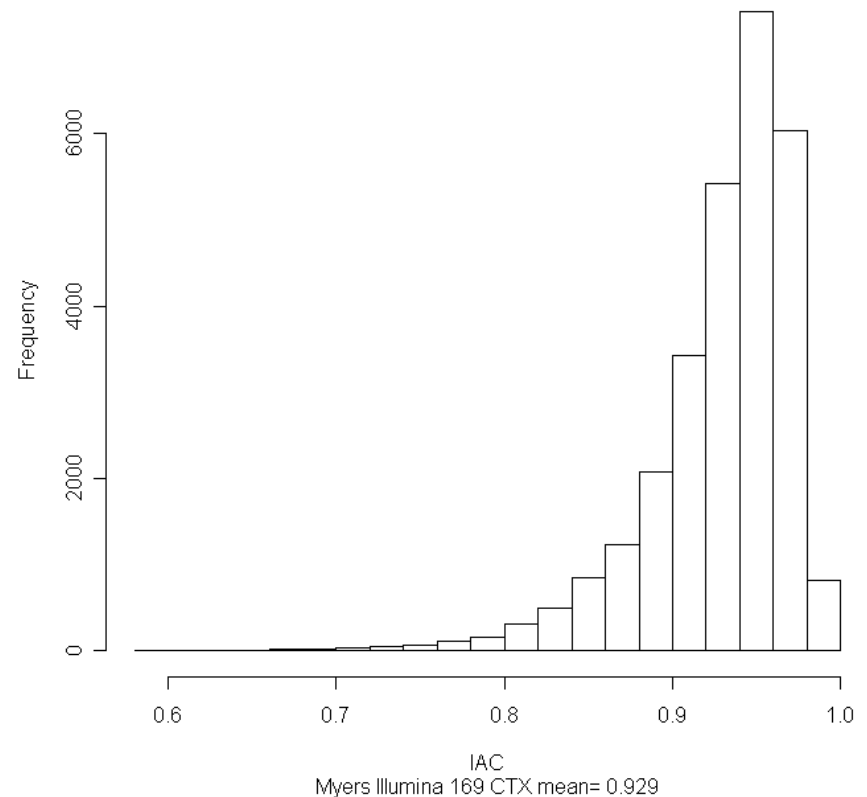
```
dim(datrest)
# [1] 5269 193
```

```
overlap1=is.element(dimnames(datrest)[[2]],removevec)
datrest2=datrest[,!overlap1]
dim(datrest2)
# [1] 5269 169
```

```
IAC=cor(datrest2,use="p")
```

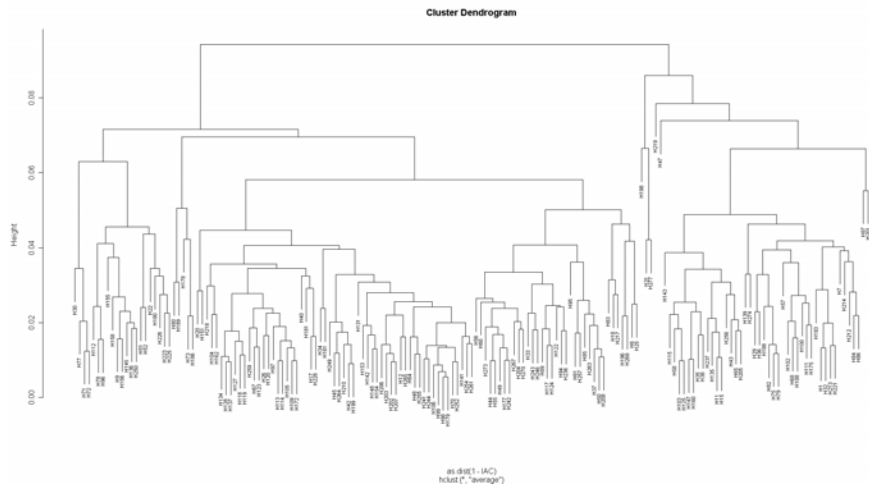
```
hist(IAC,sub=paste("Myers Illumina 169 CTX
mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

**Histogram of IAC**



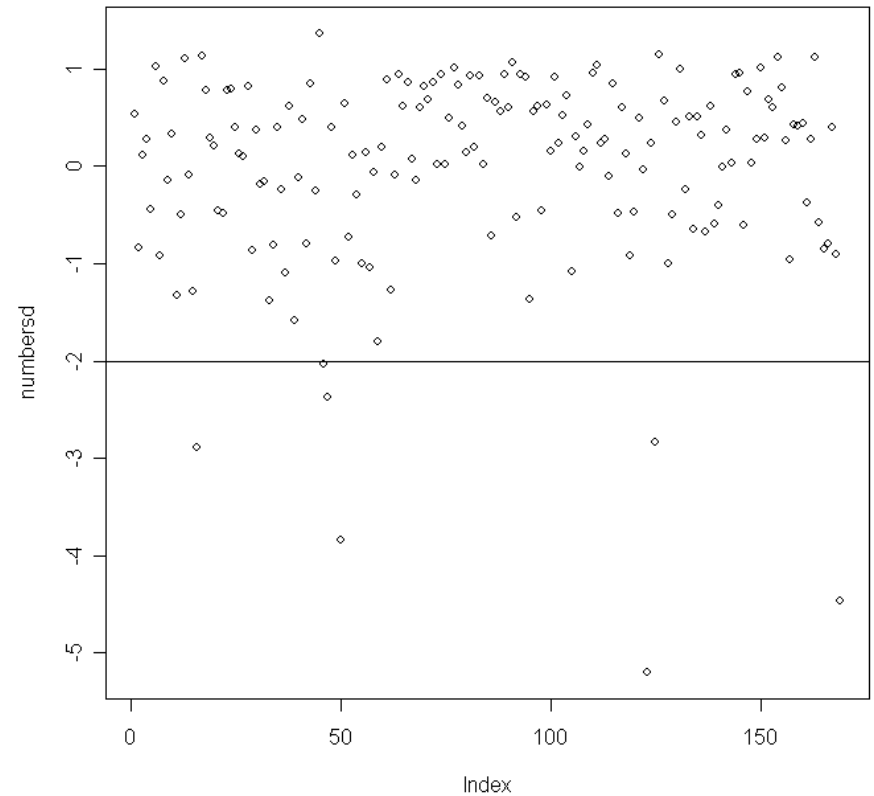
## So after removing those samples, the mean IAC has improved considerably (from 0.916 to 0.929). Clustering again:

```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(datrest2)[[2]])
```



## The dendrogram looks better, but there still appear to be outliers. Another way to visualize outliers is to calculate the mean IAC for each array and examine this distribution:

```
meanIAC=apply(IAC,2,mean)
sdCorr=sd(meanIAC)
numbersd=(meanIAC-mean(meanIAC))/sdCorr
plot(numbersd)
abline(h=-2)
```



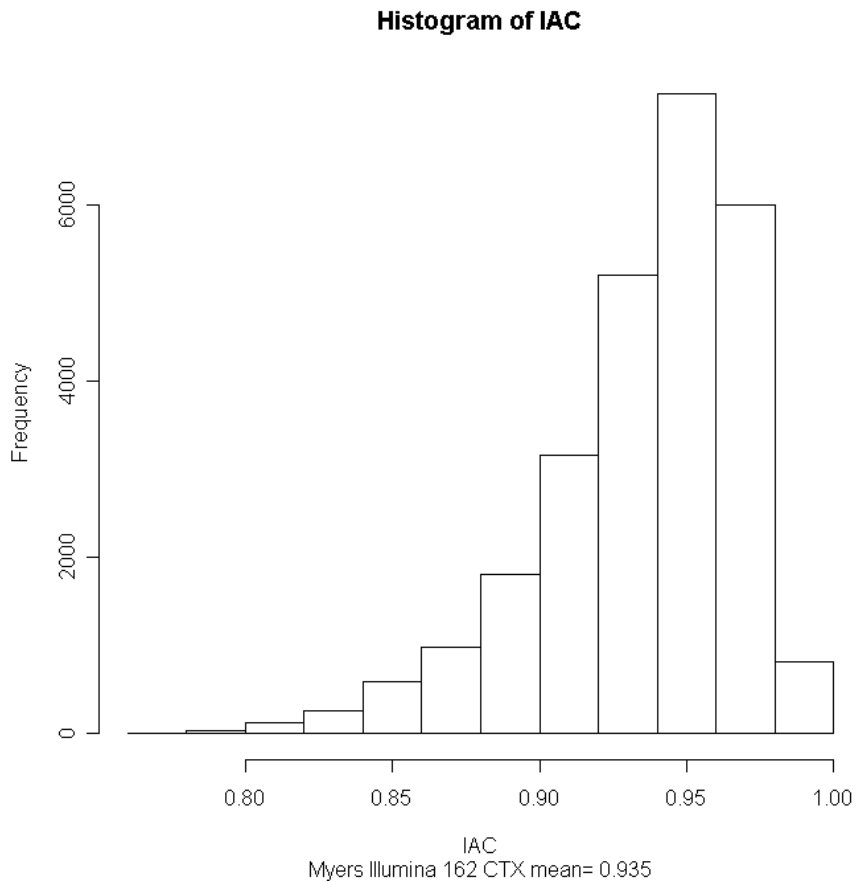
## Here one can see that a small group of arrays have mean IACs that are significantly lower than the rest. Setting -2 SD as the cutoff, we will remove the seven arrays that fall below this threshold:

```
sdout=-2
outliers=dimnames(datrest2)[[2]][numbersd<sdout]
outliers
# [1] "H31" "H71" "H72" "H77" "H198" "H201" "H277"

datrest3=datrest2[,numbersd>sdout]
dim(datrest3)
# [1] 5269 162

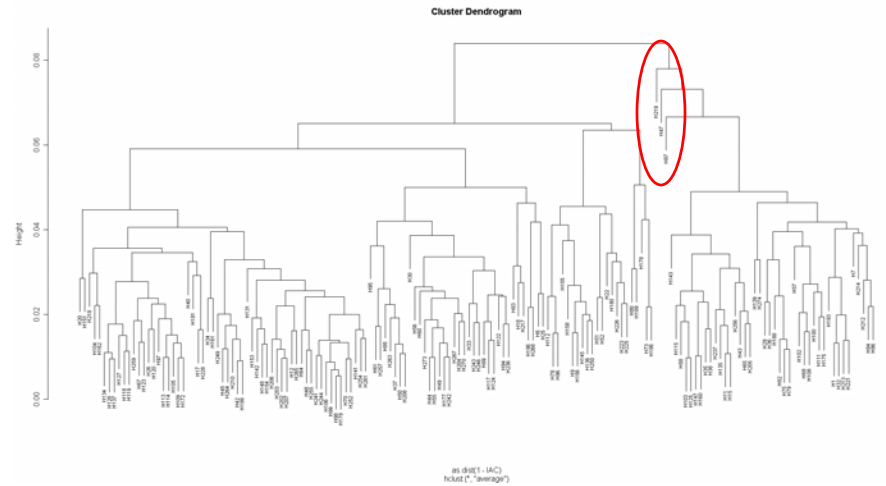
IAC=cor(datrest3,use="p")
```

```
hist(IAC,sub=paste("Myers Illumina 162 CTX
mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```



```
## The mean IAC has improved to 0.935. Clustering:
```

```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(datrest3)[[2]])
```



```
## Going to remove three clear outliers (circled in red):
```

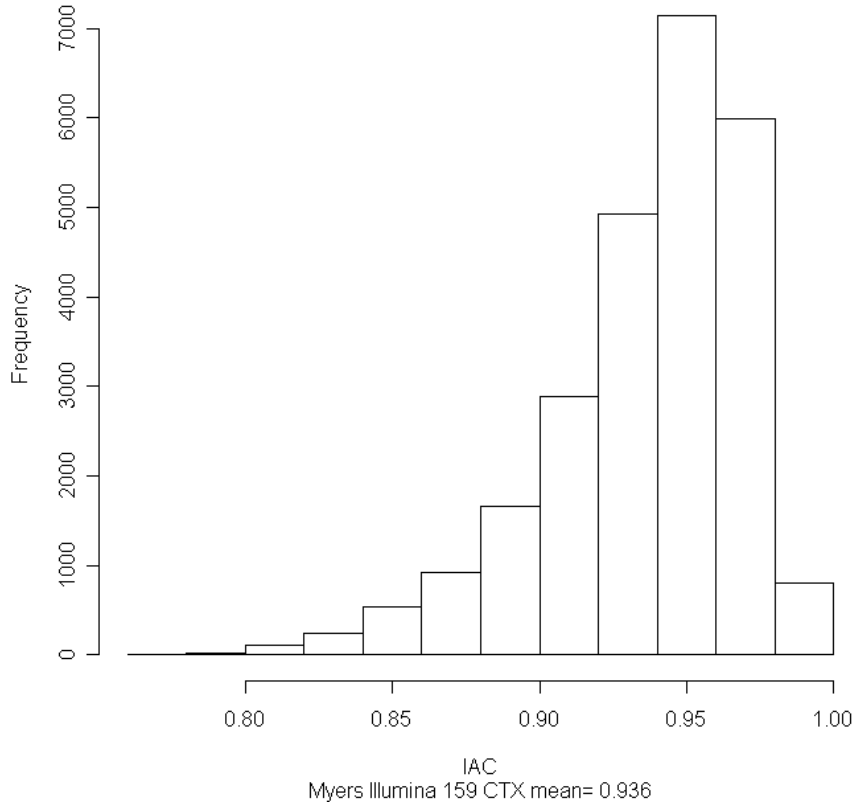
```
rm(removevec)
removevec=c("H47","H87","H219")
```

```
overlap1=is.element(dimnames(datrest3)[[2]],removevec)
datrest4=datrest3[,!overlap1]
dim(datrest4)
# [1] 5269 159
```

```
IAC=cor(datrest4,use="p")
```

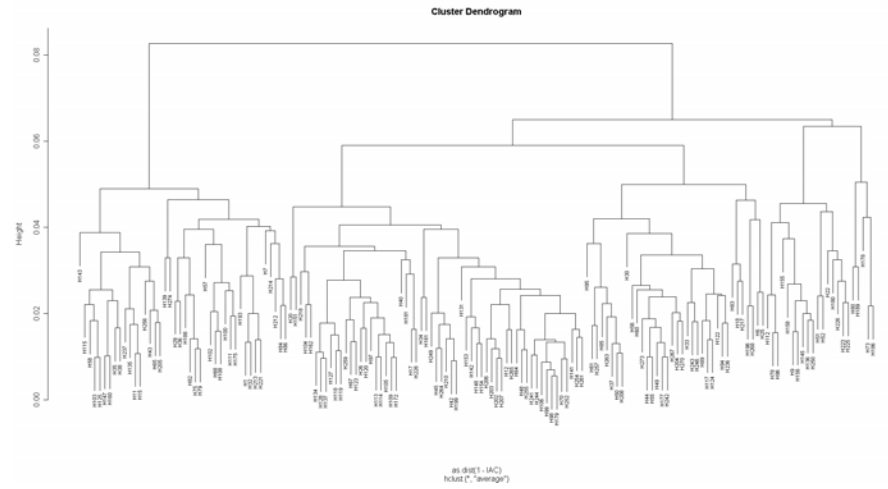
```
hist(IAC,sub=paste("Myers Illumina 159 CTX
mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

**Histogram of IAC**



## The mean IAC has only improved incrementally.

```
cluster1=hclust(as.dist(1-IAC),method="average")  
plot(cluster1,cex=0.7,labels=dimnames(datrest4)[[2]])
```



## At this point, the dendrogram does not reveal any clear outliers. Therefore, we will stop here and proceed to quantile normalization (not shown).

**References**

1. Myers, A.J. et al. A survey of genetic human cortical gene expression. *Nat Genet* **39**, 1494-9 (2007).
2. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).