

## Identification and Removal of Outlier Samples

### 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 each of the four datasets used for gene coexpression network construction in our paper.

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 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 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. Samples with low mean IACs and/or samples that exhibited divergent clustering were excluded, and the mean IACs for all datasets, after removing all outlier samples and performing quantile and batch normalization, were as follows: 0.970 (CTX), 0.975 (CTX\_95), 0.972 (CN), and 0.975 (CB). These values indicate that the overall consistency of gene expression among samples in each dataset used for network construction was very comparable.

Prior to removal of outlier and duplicate samples, dataset 1 ("CTX") consisted of 104 samples from various cortical areas<sup>1-3</sup>, dataset 2 ("CTX\_95") consisted of 82 samples from various cortical areas<sup>4-9</sup>, dataset 3 ("CN") consisted of 32 samples from the head of the caudate nucleus<sup>1</sup>, and dataset 4 ("CB") consisted of 27 samples from cerebellar hemisphere<sup>1</sup>. To eliminate non-specific and mis-targeted probes prior to generating expression values, mask files were obtained for both microarrays (<http://masker.nci.nih.gov/ev/>)<sup>10</sup> and applied to the raw data using GCOSv1.2 or the R package "ProbeFilter" (<http://arrayanalysis.mbni.med.umich.edu/MBNIUM.html#ProbeFilter>)<sup>11</sup>. After applying the mask files, only probe sets with at least seven (HG-U133A) or ten (HG-U95A/v2) remaining probes were retained for further analysis (n = 18,631 and n = 10,553, respectively).

Expression values for CTX, CN, and CB were generated using GCOSv1.2. All arrays were scaled to the same average intensity (200). For CTX\_95, expression values were generated in R using the "expresso" function of the "affy" package (<http://www.bioconductor.org/>)<sup>12</sup> with "mas" settings and no normalization, followed by scaling of arrays to the same average intensity (200). Scaled expression values were imported into R for outlier detection and removal prior to normalization.

Following outlier removal, quantile normalization<sup>13</sup> was performed for each dataset in R. Average linkage hierarchical clustering using 1-IAC as a distance metric revealed that most samples clustered by study (data not shown), indicating the presence of significant batch effects in the data. To eliminate batch effects, additional normalization was performed using the R package "ComBat" (<http://statistics.byu.edu/johnson/ComBat/>)<sup>14</sup>. Within each dataset, each study was assigned a single batch number with the exception of ref. 1, which was assigned two batch numbers (samples from this study exhibited a batch effect that reflected country of origin [U.S. vs. New Zealand; data not shown]). ComBat successfully eliminated batch effects in each dataset as evidenced by hierarchical clustering and significant improvement of mean IAC (data not shown). Negative expression values introduced by ComBat (~0.01% of all expression values) were replaced with the median for the corresponding probe set.

#### Data Description

Four separate datasets consisting of microarray data generated from human brain samples were assembled to construct four gene coexpression networks:

Dataset	Arrays	# samples before pre-processing	# samples after pre-processing	Sample description*
CTX	Affy U133A	104	67	cerebral cortex
CTX_95	Affy U95A/v2	82	42	cerebral cortex
CN	Affy U133A	32	27	caudate nucleus
CB	Affy U133A	27	24	cerebellum

\* For additional sample information, see Supplementary Table 1 from the journal article.

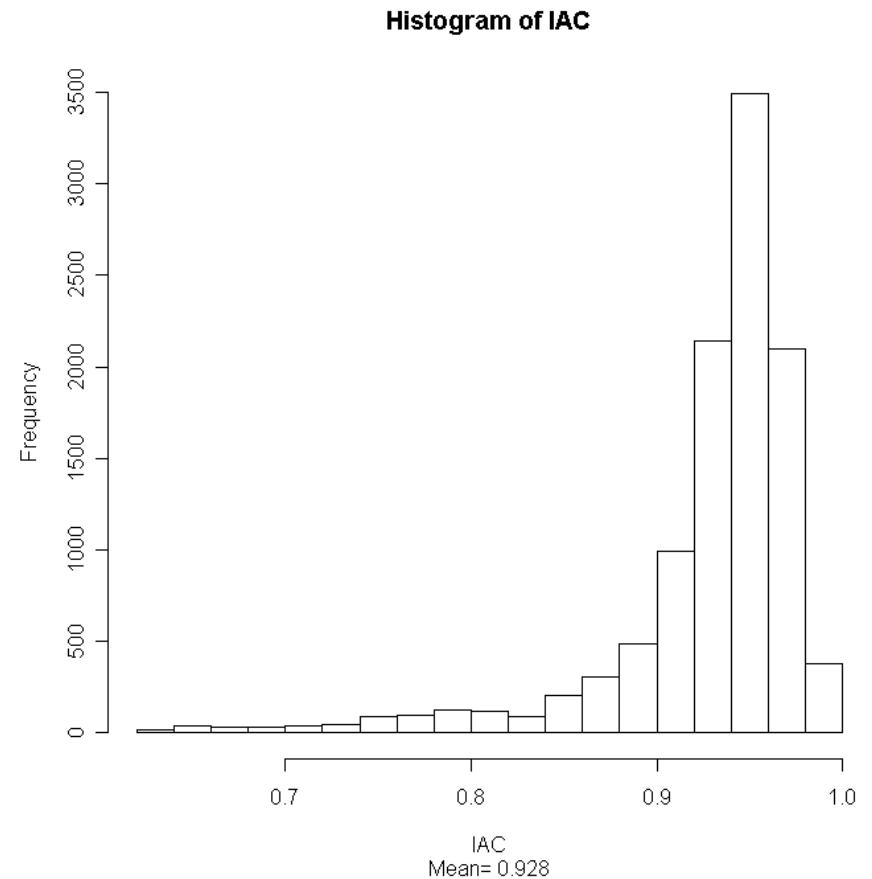
## CTX

## Reading in the unnormalized, scaled expression data (18,631 probe sets, 104 samples; column 1 contains probe set IDs):

```
dat1=read.csv("CTX_104samples_unnormalized_expression_data.csv",header=T)
dim(dat1)
# [1] 18631 105
```

```
dat2=dat1[,2:105]
## Calculating IACs for all pairs of samples and examining the distribution of IACs in the dataset:
```

```
IAC=cor(dat2,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

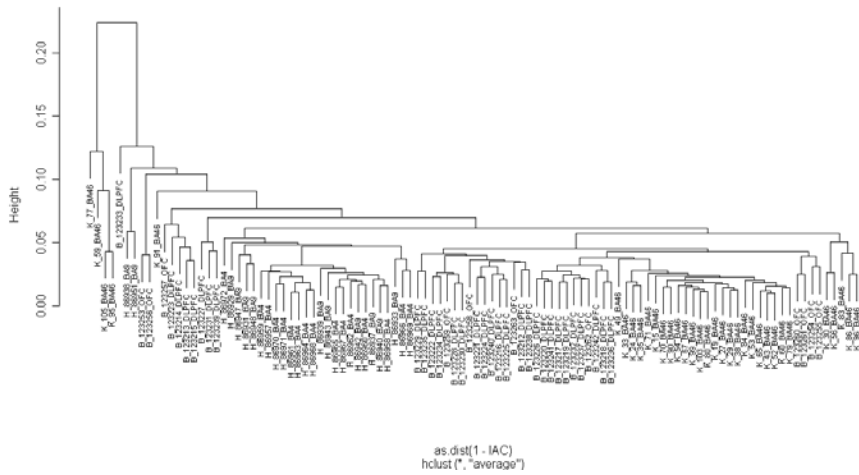


## Here we see that the mean IAC in the unnormalized CTX dataset, with no outlier samples removed, is 0.928. There is a long tail to the left of the distribution, indicating the presence of possible outliers.

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

```
library(cluster)
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat2)[[2]])
```

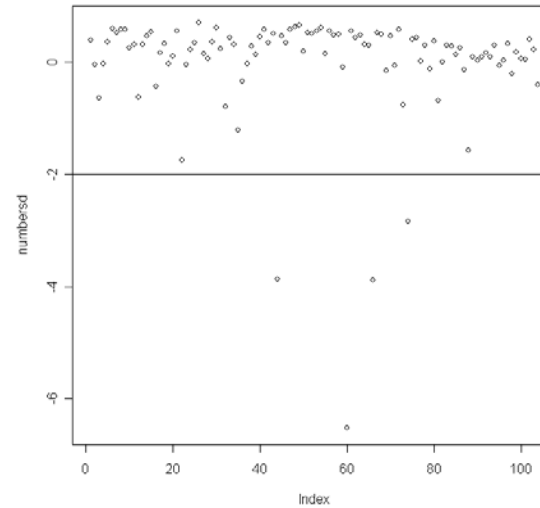
Cluster Dendrogram



## Note that samples are labeled in the dendrogram above in the form study\_sampleID\_brain region. There are four clear outliers from study "K"<sup>2</sup> at left (77, 59, 105, and 95).

## 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 the four outliers depicted above are the same outliers identified in the initial dendrogram:

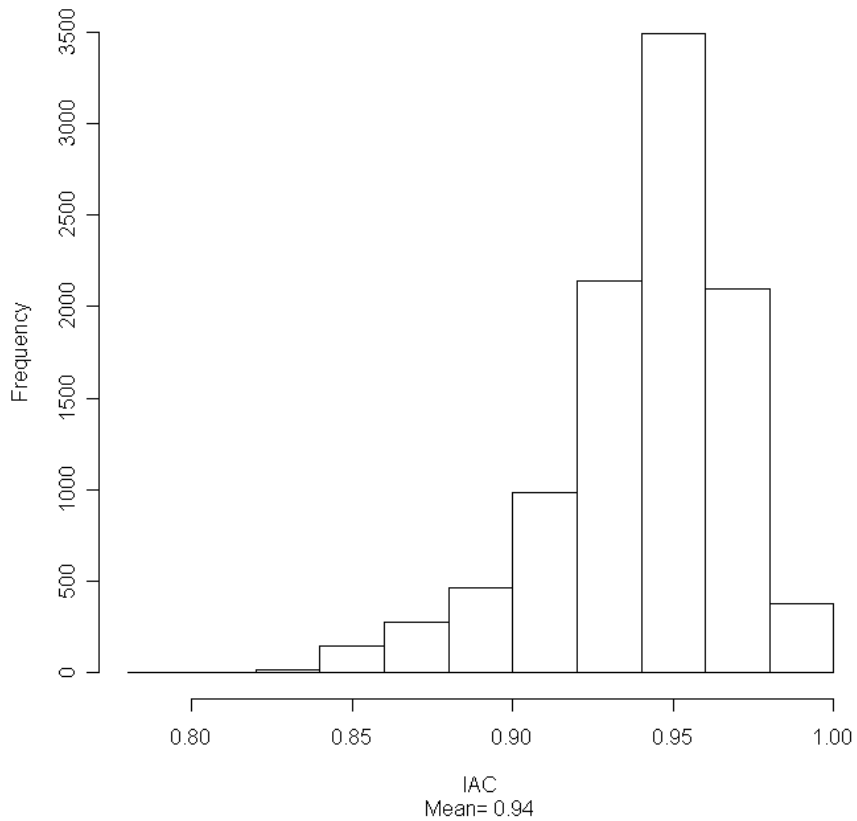
```
sdout=-2
outliers=dimnames(dat2)[[2]][numbersd<sdout]
outliers
[1] "K_105_BA46" "K_59_BA46" "K_77_BA46" "K_95_BA46"
```

## Now, we remove these four outliers and calculate a new IAC matrix:

```
dat3=dat2[,numbersd>sdout]
dim(dat3)
# [1] 18631 100

IAC=cor(dat3,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)])),digits=3))
```

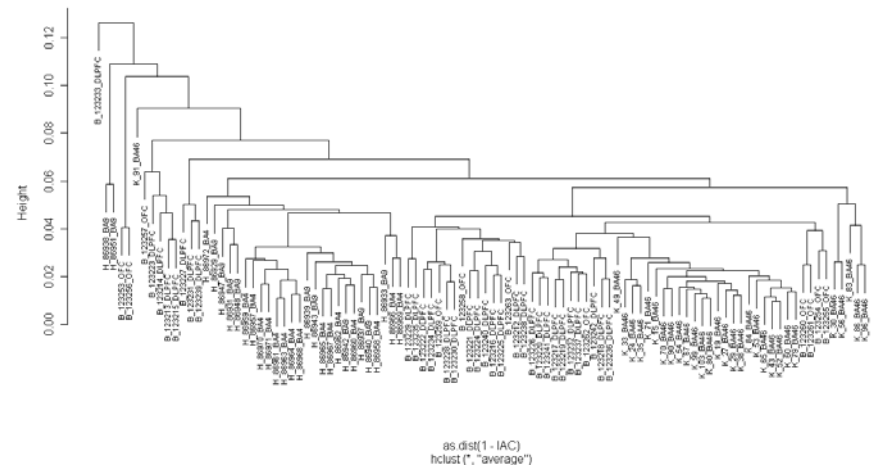
**Histogram of IAC**



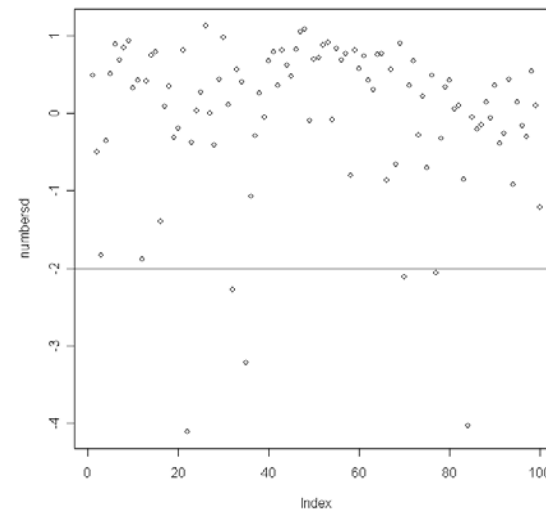
## Here we see that both the distribution of IACs and the mean have improved considerably after removing the four outliers. Clustering:

```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat3)[[2]])
```

**Cluster Dendrogram**



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



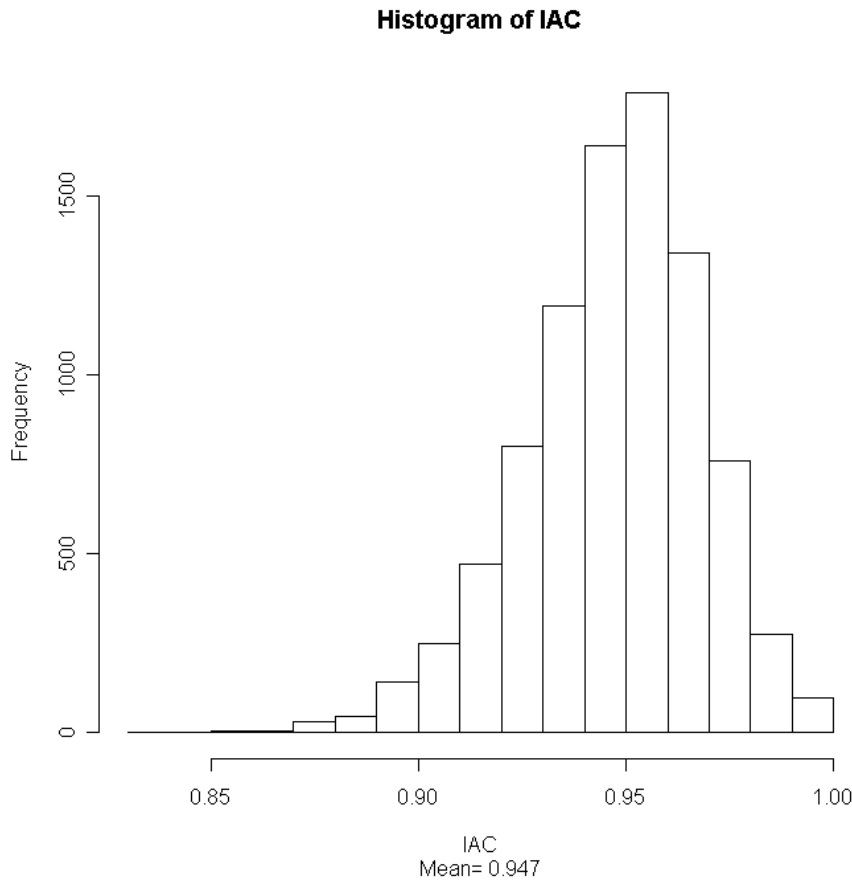
```
sdout=-2
outliers=dimnames(dat3)[[2]][numbersd<sdout]
outliers
```

```
# [1] "B_123233_DLPFC" "B_123253_OFPC" "B_123256_OFPC"
"K_91_BA46"
[5] "H_86938_BA9" "H_86951_BA9"
```

## These six outliers correspond to the leftmost samples in the dendrogram above. They will be removed:

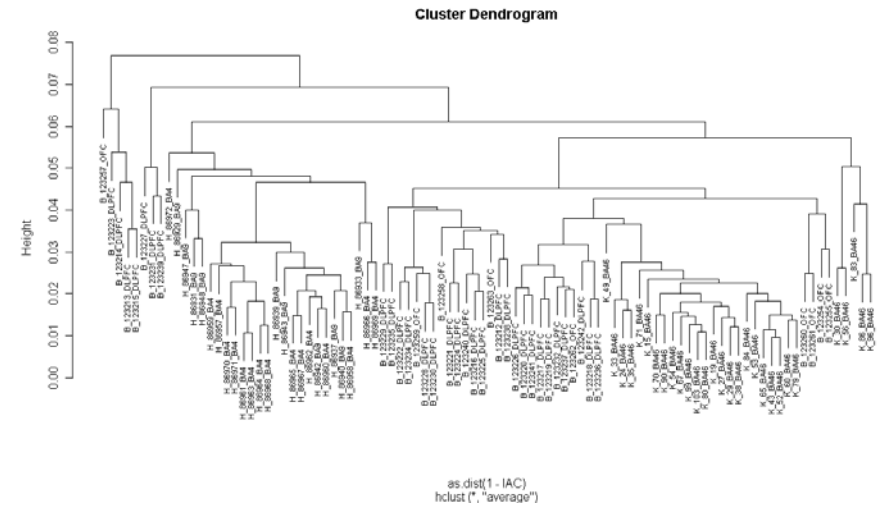
```
dat4=dat3[,numbersd>sdout]
dim(dat4)
#[1] 18631 94
```

```
IAC=cor(dat4,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

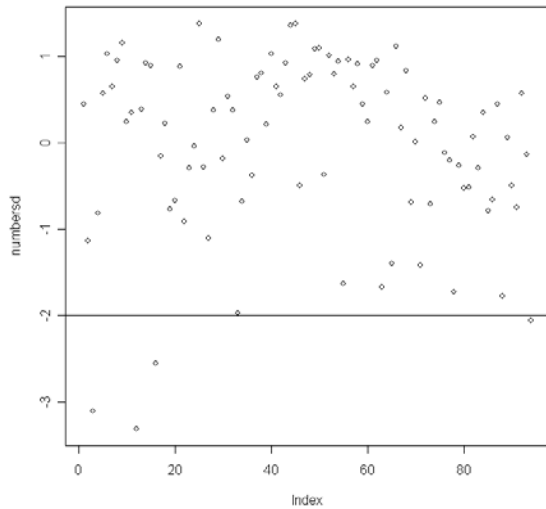


## Again, both the mean and distribution of IACs have improved. Clustering:

```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat4)[[2]])
```



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



```

sdout=-2
outliers=dimnames(dat4)[[2]][numbersd<sdout]

## Here we remove the three outliers depicted in the
bottom left of the above plot:

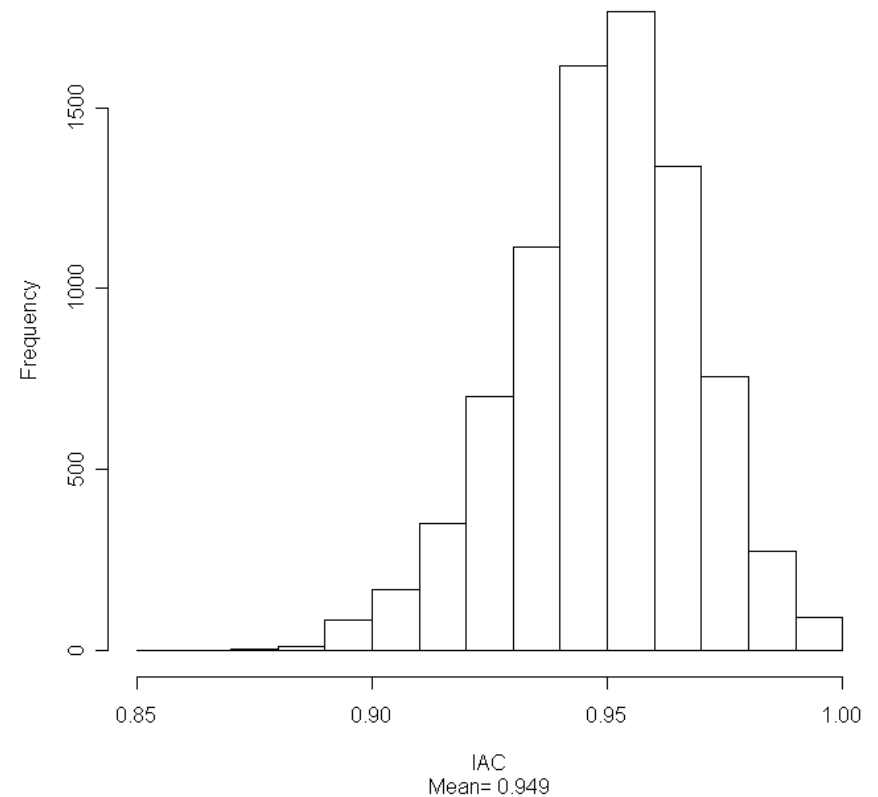
outliers=outliers[1:3]
overlap=!is.element(as.character(dimnames(dat4)[[2]]),as
.character(outliers))

dat5=dat4[,overlap]
dim(dat5)
# [1] 18631    91

IAC=cor(dat5,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC
)]),digits=3)))

```

**Histogram of IAC**

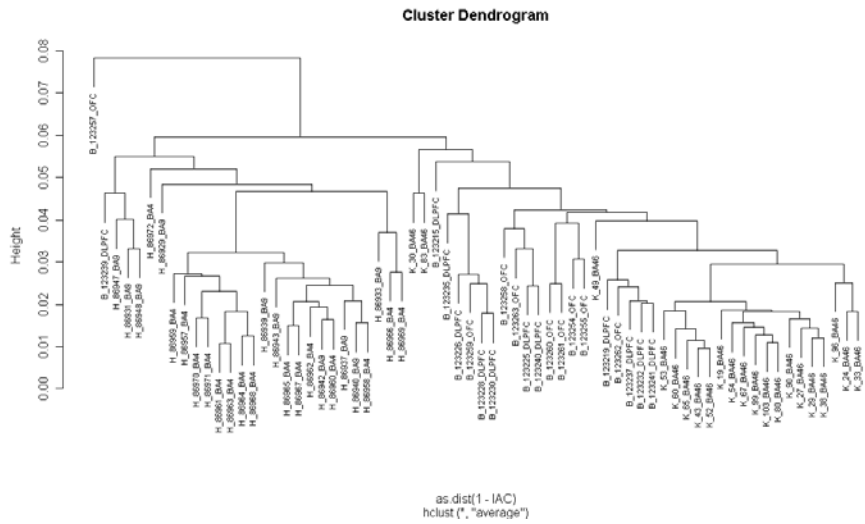


```

## As outliers are successively removed with each round,
the mean IAC begins to stabilize.
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat5)[[2]])

```





## Examining the tree, we see that one clear outlier remains (B\_123257\_OF1C). Removing this individual:

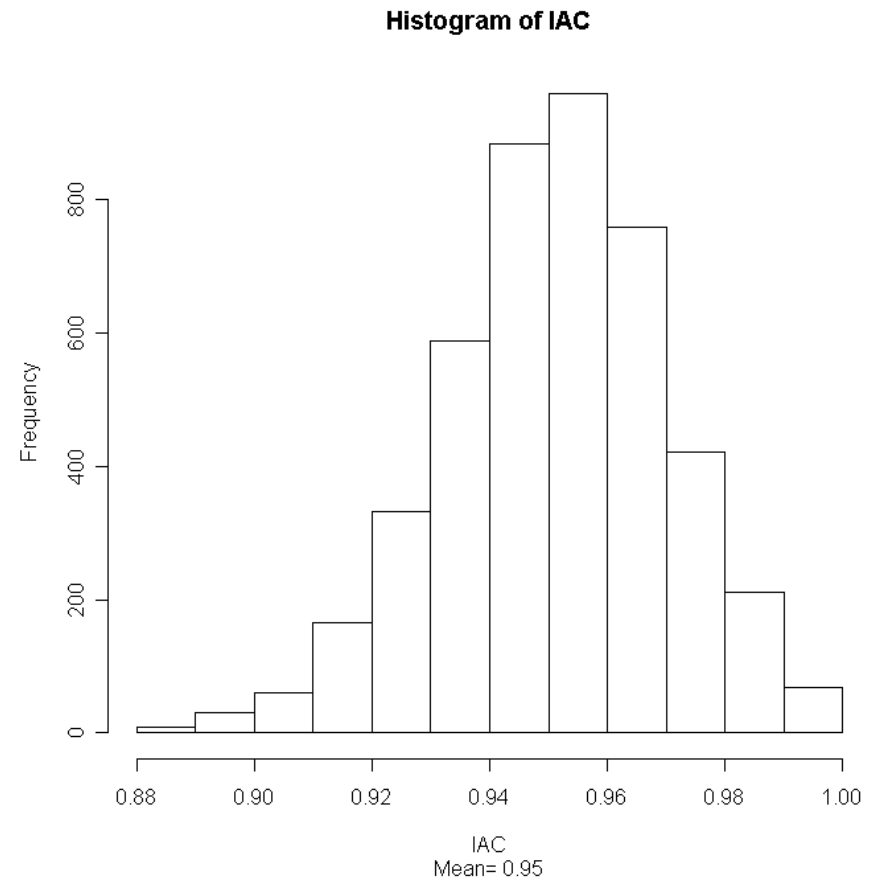
```

meanIAC=apply(IAC,2,mean)
sdCorr=sd(meanIAC)
numbersd=(meanIAC-mean(meanIAC))/sdCorr
sdout=-2
outliers=dimnames(dat6)[[2]][numbersd<sdout]
outliers
[1] "B_123257_OF1C" "K_83_BA46" "H_86972_BA4"
outliers=outliers[1]
overlap=!is.element(as.character(dimnames(dat6)[[2]]),as
.character(outliers))

dat7=dat6[,overlap]
dim(dat7)
#[1] 18631 67

IAC=cor(dat7,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)
])),digits=3))

```



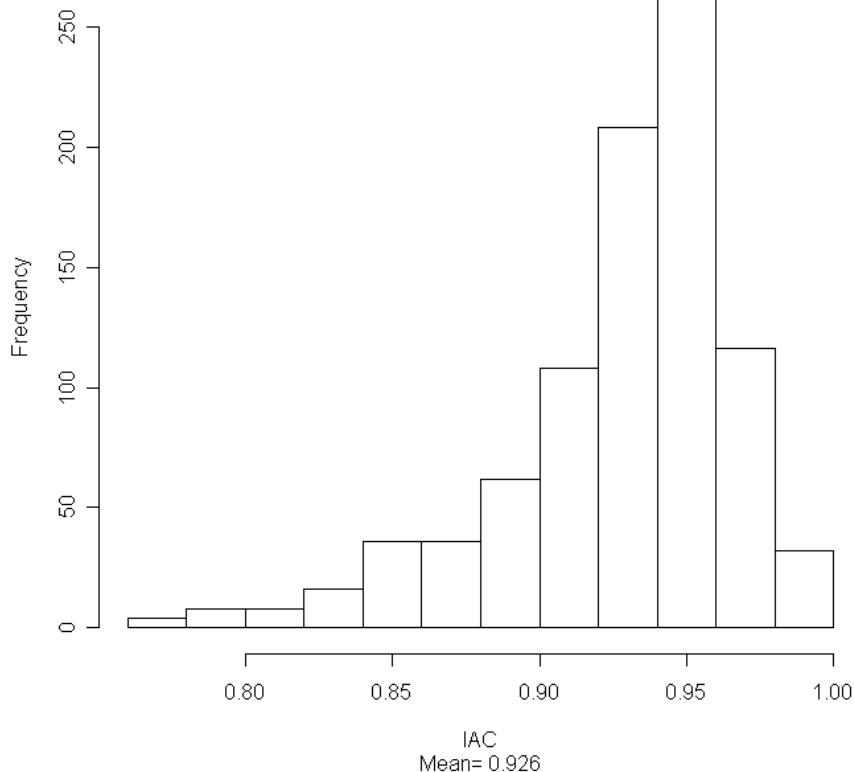
```

cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat7)[[2]])

```



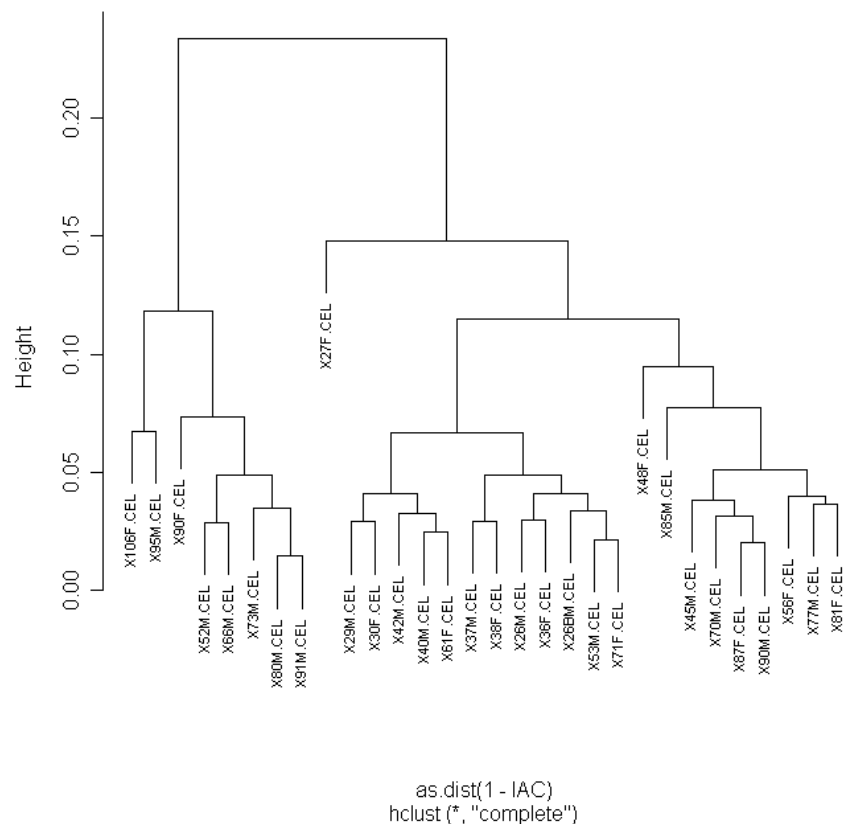
**Histogram of IAC**



## Again we see a long tail at the left of the distribution that suggests the presence of outliers amongst the arrays. Clustering:

```
cluster1=hclust(as.dist(1-IAC))
plot(cluster1,main="Lu",cex=0.7)
```

**Lu**



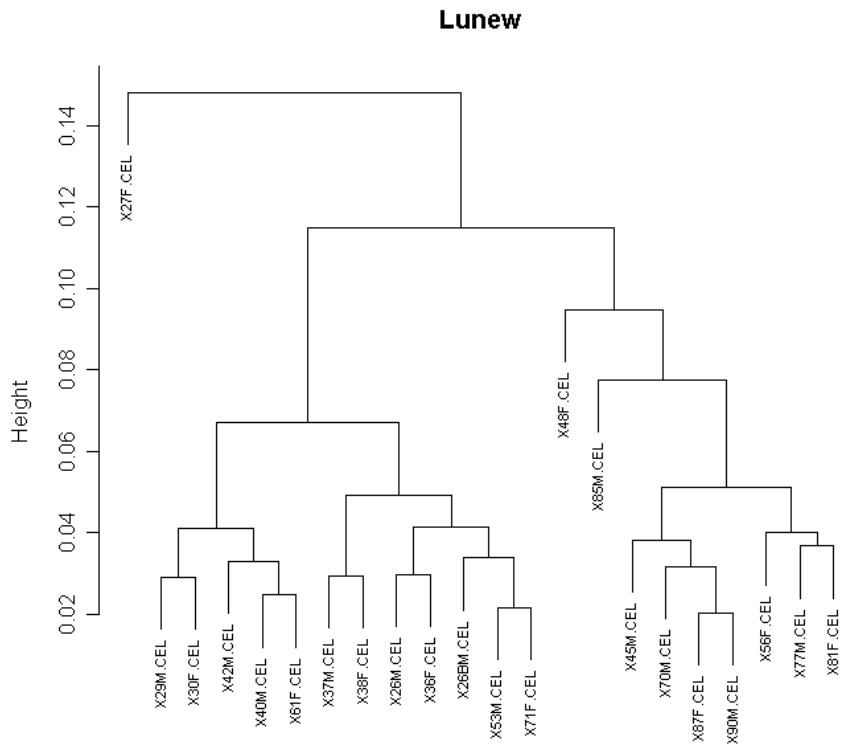
## The cluster at the left forms with a very high branch point (>0.2); we will remove these 8 samples:

```
dimnames(datLu)[[2]]
## [1] "X106F.CEL" "X26BM.CEL" "X26M.CEL" "X27F.CEL" "X29M.CEL"
"X30F.CEL" "X36F.CEL" "X37M.CEL" "X38F.CEL"
##[10] "X40M.CEL" "X42M.CEL" "X45M.CEL" "X48F.CEL" "X52M.CEL"
"X53M.CEL" "X56F.CEL" "X61F.CEL" "X66M.CEL"
##[19] "X70M.CEL" "X71F.CEL" "X73M.CEL" "X77M.CEL" "X80M.CEL"
"X81F.CEL" "X85M.CEL" "X87F.CEL" "X90F.CEL"
##[28] "X90M.CEL" "X91M.CEL" "X95M.CEL"
```

```
allsamples=c(1:length(datLu[1,]))
prunevec=c(1,14,18,21,23,27,29,30)
overlap1=is.element(allsamples,prunevec)
```

```
datLunew=datLu[,!overlap1]
dim(datLunew)
# [1] 10553    22
```

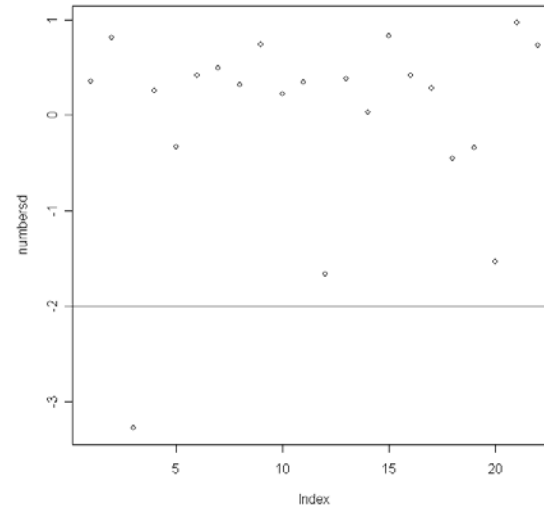
```
IAC=cor(datLunew,use="p")
cluster1=hclust(as.dist(1-IAC))
plot(cluster1,main="Lunew",cex=0.7)
```



```
as.dist(1 - IAC)
hclust(*, "complete")
```

## An additional outlier (X27F) is evident. Removing this individual:

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

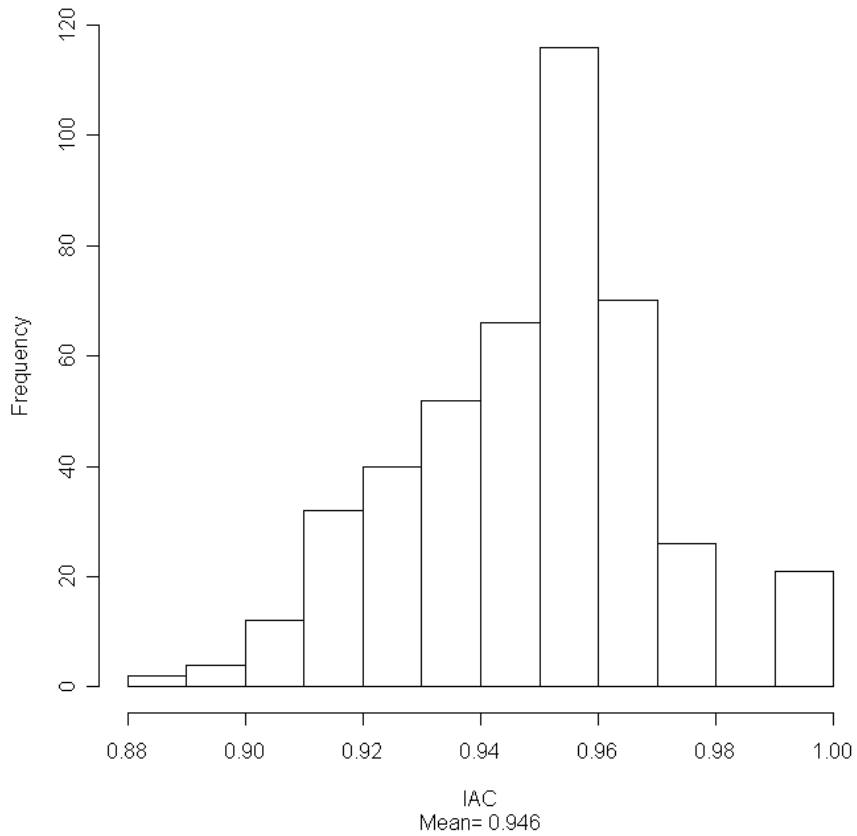


```
sdout=-2
outliers=dimnames(datLunew)[[2]][numbersd<sdout]
outliers
# [1] "X27F.CEL"
```

```
datLunew=datLunew[,numbersd>sdout]
dim(datLunew)
# [1] 10553    21
```

```
IAC=cor(datLunew,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

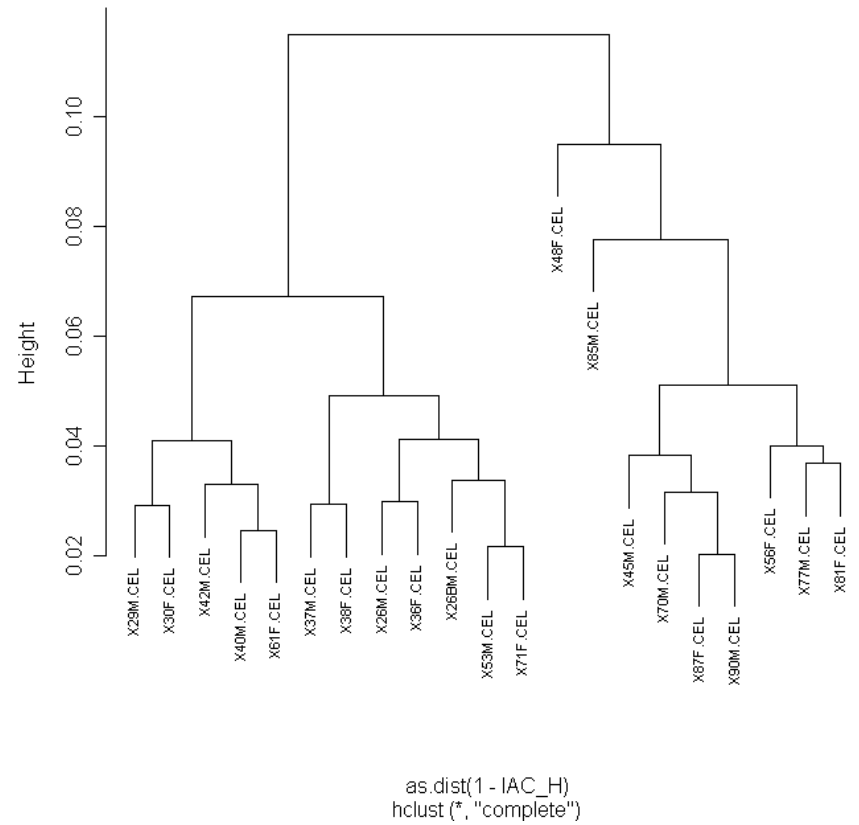
**Histogram of IAC**



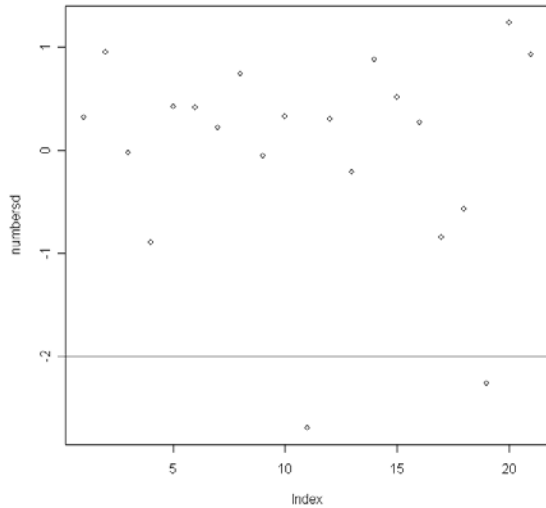
## The distribution of IACs and mean IAC have improved noticeably.

```
cluster1=hclust(as.dist(1-IAC))
plot(cluster1,main="Datnew",cex=0.7)
```

**Datnew**



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

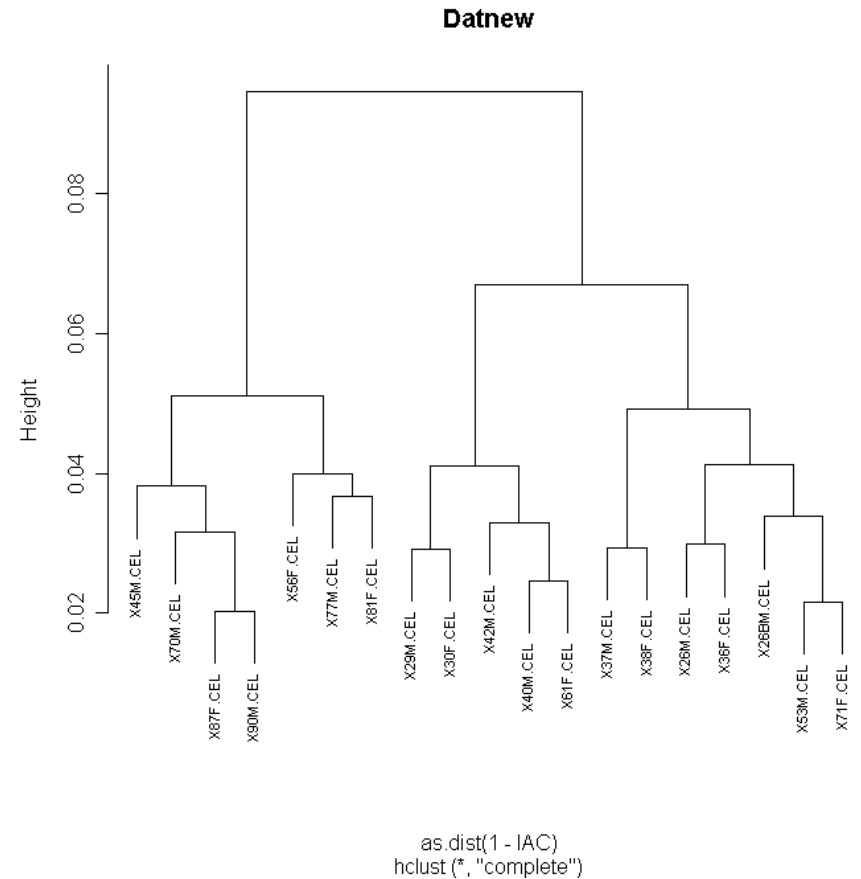


```
sdout=-2
outliers=dimnames(datLunew)[[2]][numbersd<sdout]
outliers
# [1] "X48F.CEL" "X85M.CEL"
```

```
## Removing these two samples:
```

```
datLunew=datLunew[,numbersd>sdout]
dim(datLunew)
# [1] 10553 19
```

```
IAC=cor(datLunew,use="p")
cluster1=hclust(as.dist(1-IAC))
plot(cluster1,main="Datnew",cex=0.7)
```



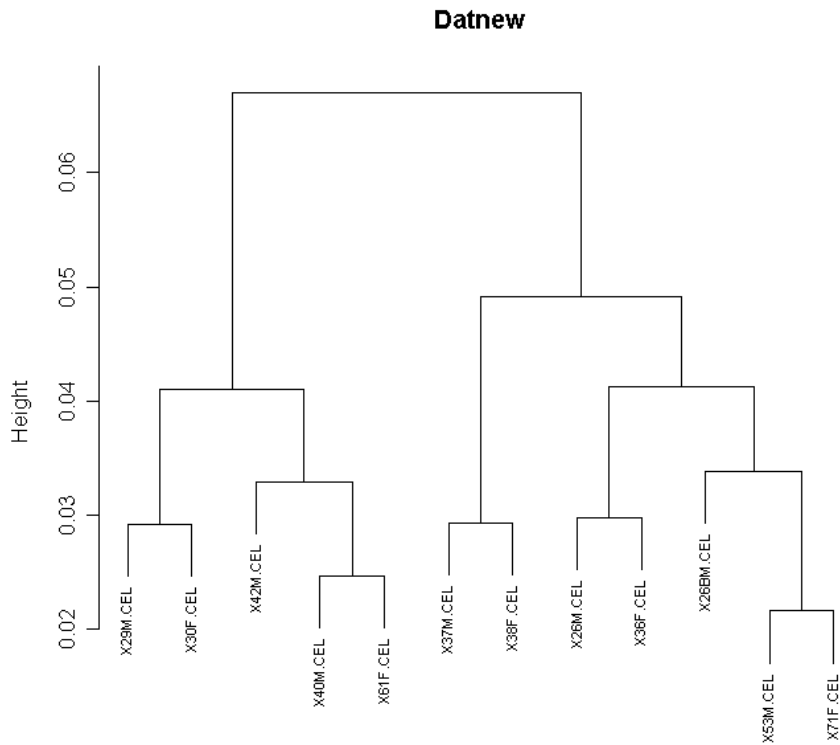
```
## There is still a relatively high branch point between
the group of 7 at left and the remainder at right. To
be conservative, we will remove the 7 samples at left:
```

```
dimnames(datLunew)[[2]]
#[1] "X26BM.CEL" "X26M.CEL" "X29M.CEL" "X30F.CEL" "X36F.CEL"
"X37M.CEL" "X38F.CEL" "X40M.CEL" "X42M.CEL" "X45M.CEL" "X53M.CEL"
#[12] "X56F.CEL" "X61F.CEL" "X70M.CEL" "X71F.CEL" "X77M.CEL"
"X81F.CEL" "X87F.CEL" "X90M.CEL"
```

```
allsamples=c(1:length(datLunew[1,]))
prunevec=c(10,12,14,16,17,18,19)
overlap1=is.element(allsamples,prunevec)
```

```
datLunew=datLunew[,!overlap1]
dim(datLunew)
```

```
# [1] 10553    12
IAC=cor(datLunew,use="p")
cluster1=hclust(as.dist(1-IAC))
plot(cluster1,main="Datnew",cex=0.7)
```



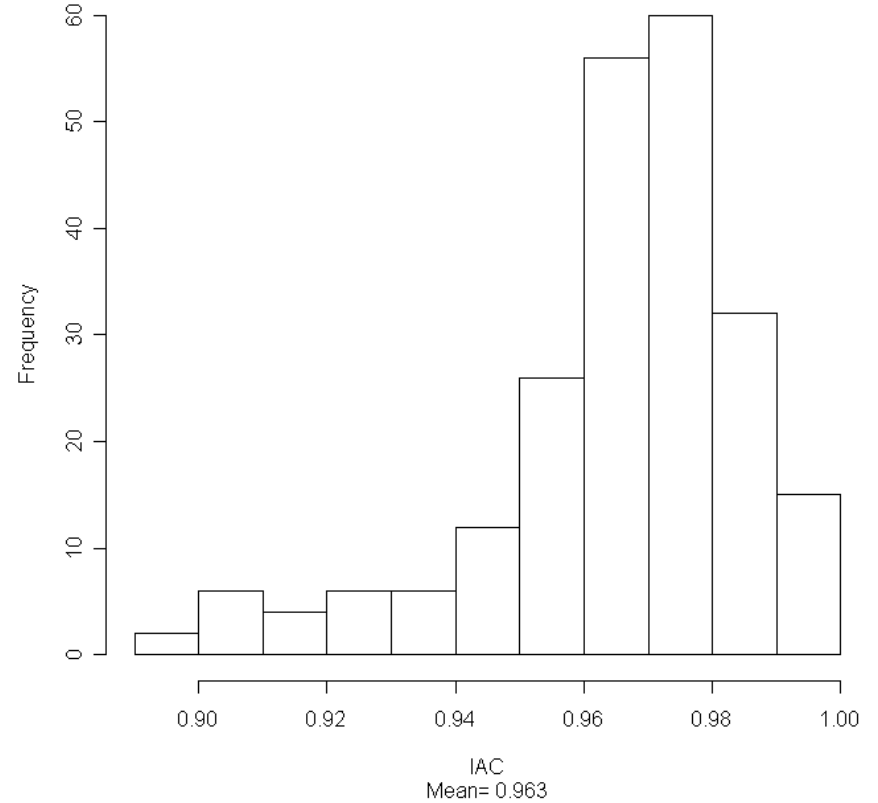
```
as.dist(1-IAC_H)
hclust(*,"complete")
```

## We will stop here and move on to the next study.

```
dimnames(dat2)[[2]]
indexKaz=c(8:22)
datKaz=dat2[,indexKaz]
dim(datKaz)
# [1] 10553    15
```

```
IAC=cor(datKaz,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)
])),digits=3))
```

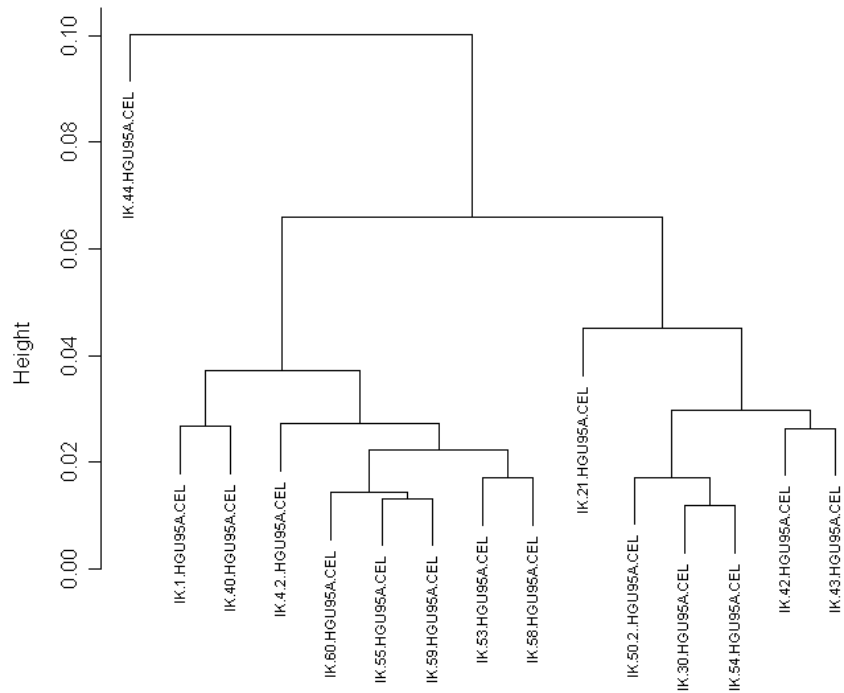
**Histogram of IAC**



## The mean IAC for this study is already very good. However, there may still be outliers.

```
cluster1=hclust(as.dist(1-IAC))
plot(cluster1,main="Kaz",cex=0.7)
```

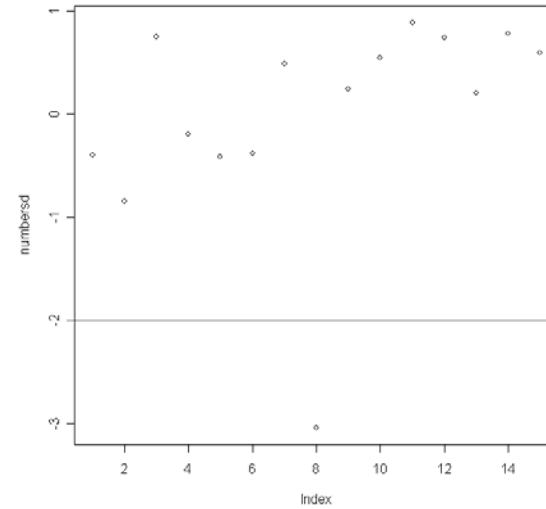
## Kaz



```
as.dist(1 - IAC)
hclust("complete")
```

```
## One clear outlier (44)
```

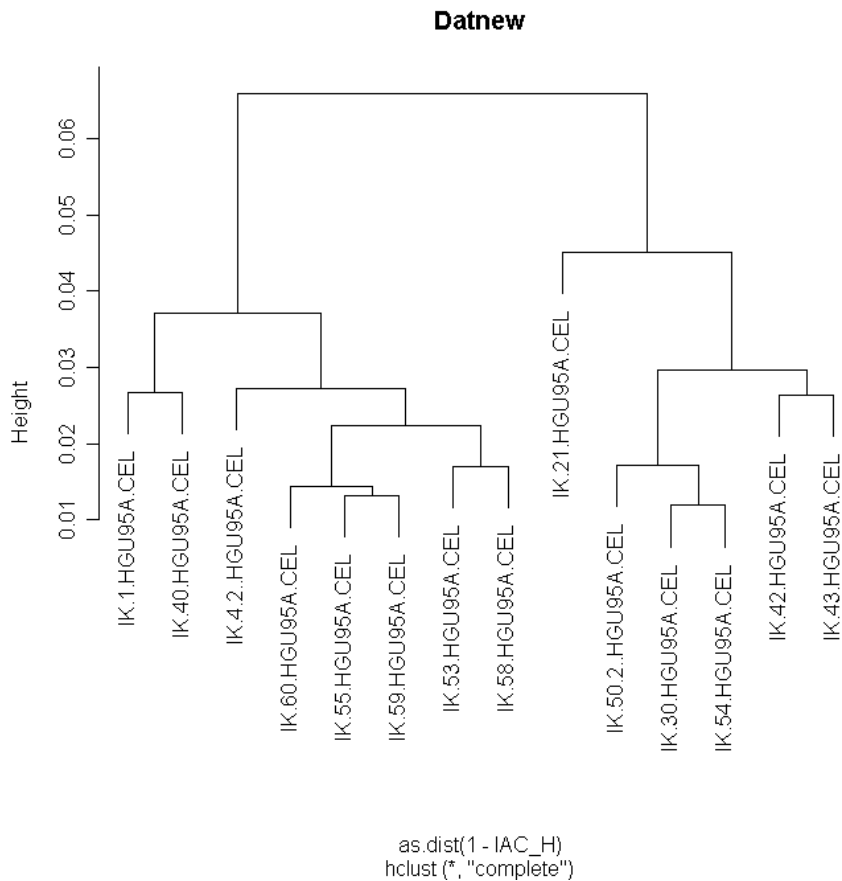
```
meanIAC=apply(IAC, 2, mean)
sdCorr=sd(meanIAC)
numbersd=(meanIAC-mean(meanIAC))/sdCorr
plot(numbersd)
abline(h=-2)
```



```
sdout=-2
outliers=dimnames(datKaz)[[2]][numbersd<sdout]
outliers
# [1] "IK.44.HGU95A.CEL"
```

```
datKaznew=datKaz[, numbersd>sdout]
dim(datKaznew)
# [1] 10553 14
```

```
IAC=cor(datKaznew, use="p")
clusterH=hclust(as.dist(1-IAC))
plot(clusterH, main="Datnew")
```

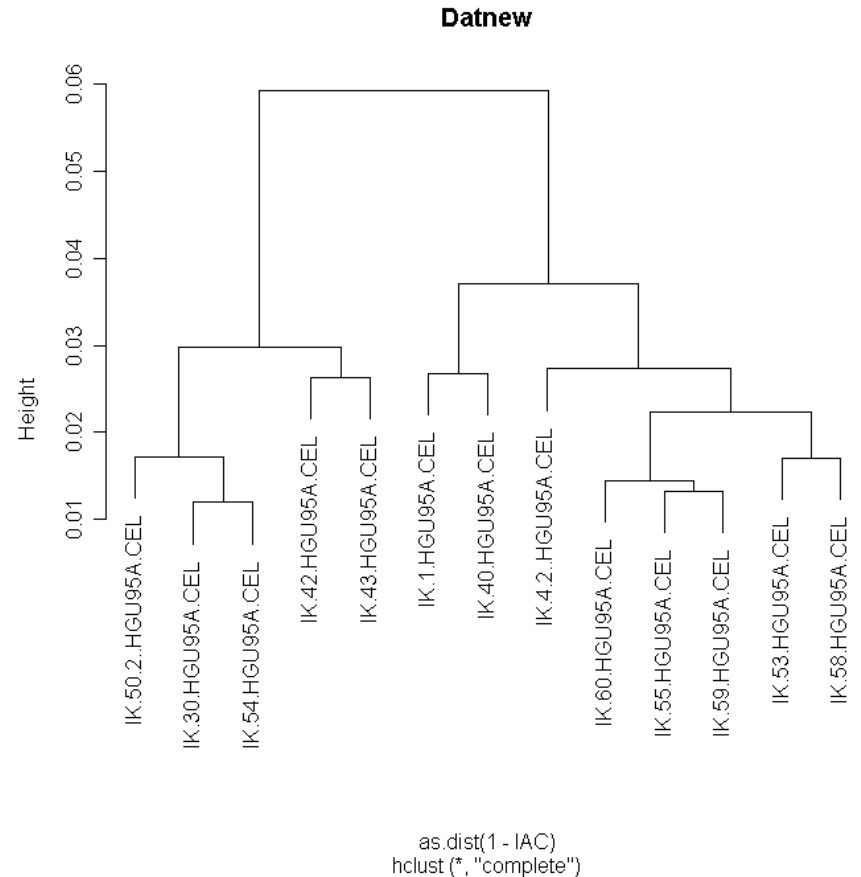


## From the structure of the dendrogram, it appears there may be one more outlier (21). Removing this sample:

```
dimnames(datKaznew)[[2]]
#[1] "IK.1.HGU95A.CEL"      "IK.21.HGU95A.CEL"
"IK.30.HGU95A.CEL"      "IK.4.2.HGU95A.CEL"
"IK.40.HGU95A.CEL"      "IK.42.HGU95A.CEL"
#[7] "IK.43.HGU95A.CEL"      "IK.50.2.HGU95A.CEL"
"IK.53.HGU95A.CEL"      "IK.54.HGU95A.CEL"
"IK.55.HGU95A.CEL"      "IK.58.HGU95A.CEL"
#[13] "IK.59.HGU95A.CEL"      "IK.60.HGU95A.CEL"
keepvec=c(1,3:14)
datKaznew=datKaznew[,keepvec]
dim(datKaznew)
```

```
# [1] 10553 13
```

```
IAC=cor(datKaznew,use="p")
clusterH=hclust(as.dist(1-IAC))
plot(clusterH,main="Datnew")
```

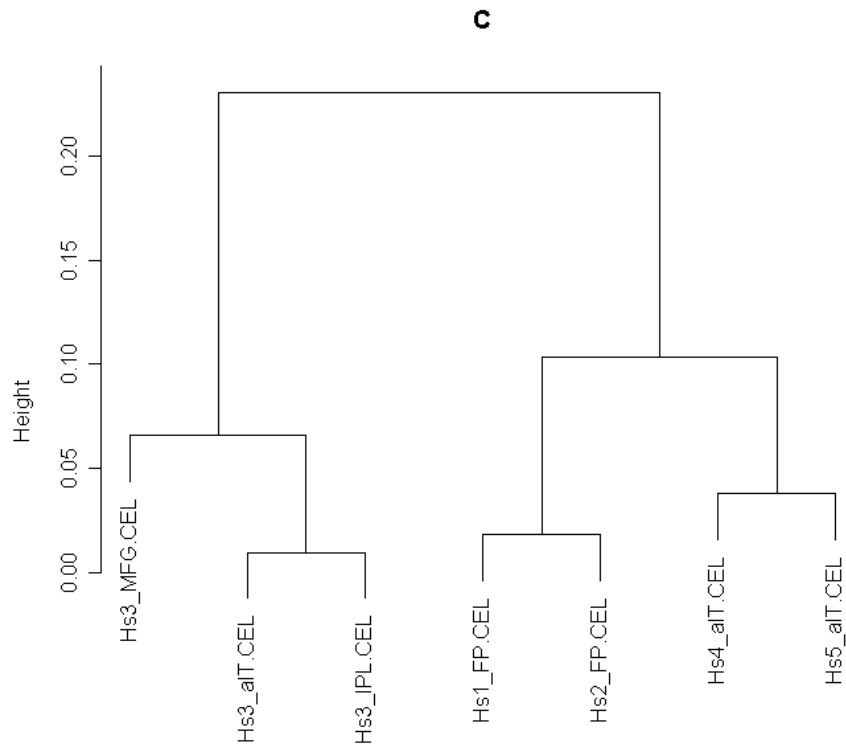


## We will stop there and move on to the next study:

```
dimnames(dat2)[[2]]
indexC=c(1:7)
datC=dat2[,indexC]
dim(datC)
# [1] 10553 7

IAC=cor(datC,use="p")
clusterH=hclust(as.dist(1-IAC))
```

```
plot(clusterH,main="C")
```



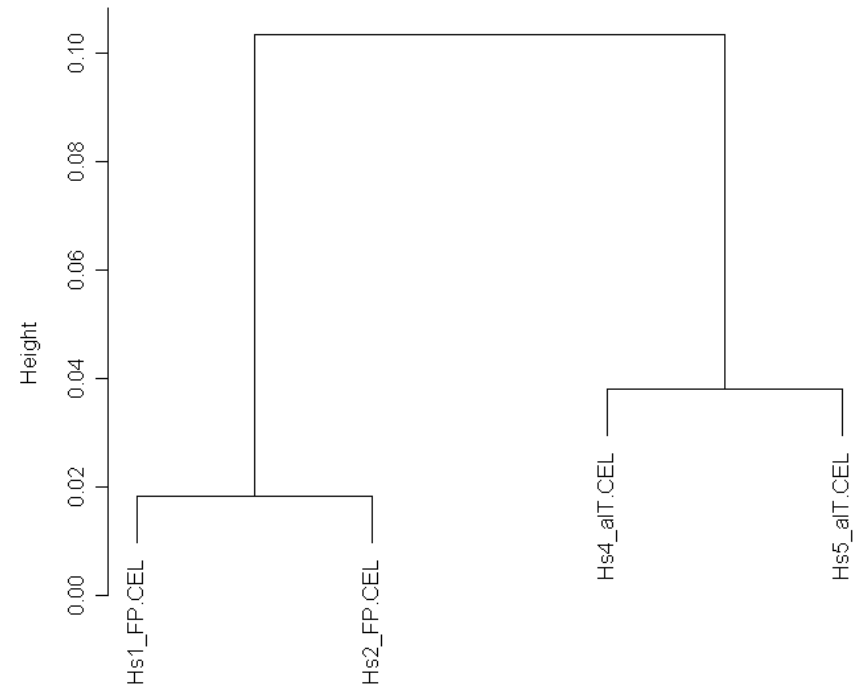
```
as.dist(1 - IAC)
hclust (*, "complete")
```

## There is a high branchpoint that separates three samples from one individual (Hs3) in the tree. We will remove these samples:

```
dimnames(datC)[[2]]
keepvec=c(1,2,6,7)
datCnew=datC[,keepvec]
dim(datCnew)
# [1] 10553      4
```

```
IAC=cor(datCnew,use="p")
clusterH=hclust(as.dist(1-IAC))
plot(clusterH,main="C")
```

**c**



```
as.dist(1 - IAC)
hclust (*, "complete")
```

## There is still a high branch point between these two pairs of samples. We will keep the pair that are more highly correlated with one another (i.e. Hs1 and Hs2):

```
dimnames(datCnew)[[2]]
rm(keepvec)
keepvec=c(1,2)
datCnew=datC[,keepvec]
dim(datCnew)
# [1] 10553      2
```

## Proceeding to the next study:

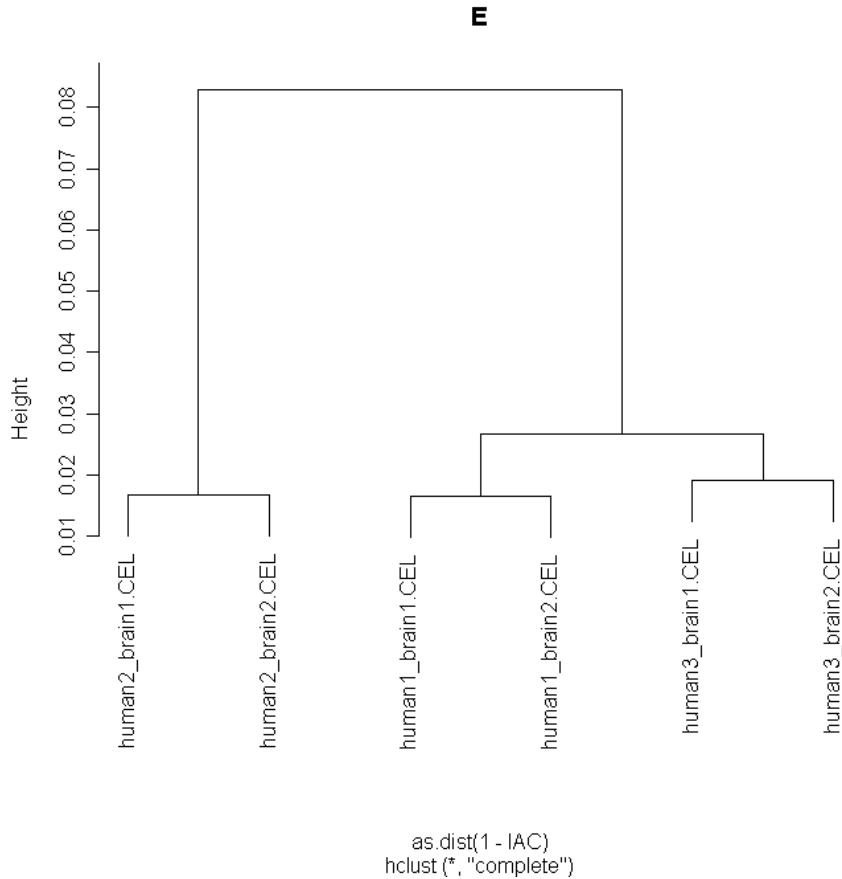
```
dimnames(dat2)[[2]]
indexE=c(77:82)
```

```

datE=dat2[,indexE]
dim(datE)
# [1] 10553    6

IAC=cor(datE,use="p")
clusterH=hclust(as.dist(1-IAC))
plot(clusterH,main="E")

```



## Note that in this study there are two technical replicates for each of the three human individuals. Clearly one individual (human2) is more different than the others; we will therefore remove these samples:

```

dimnames(datE)[[2]]
keepvec=c(1,2,5,6)

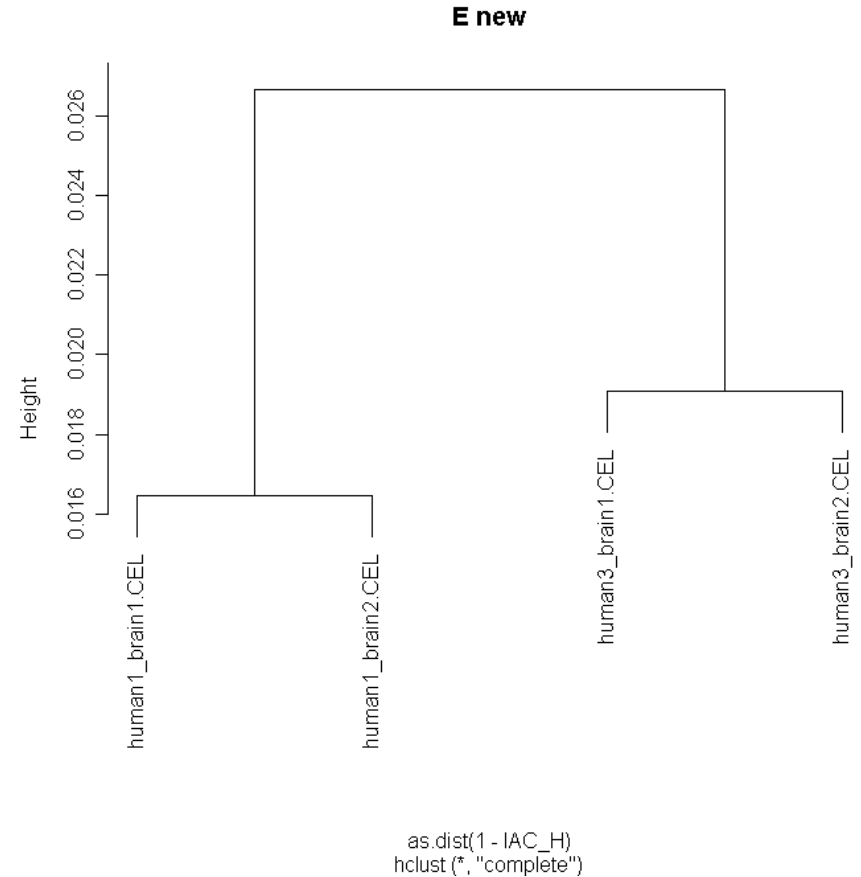
```

```

datEnew=datE[,keepvec]
dimnames(datEnew)[[2]]

IAC=cor(datEnew,use="p")
clusterH=hclust(as.dist(1-IAC))
plot(clusterH,main="E new")

```



## We will stop there and proceed to the next study. Note that samples from refs. 7,8 were combined for this analysis, since a) there were a small number of samples from ref. 8 (n=3) and b) these studies were performed by the same investigator.

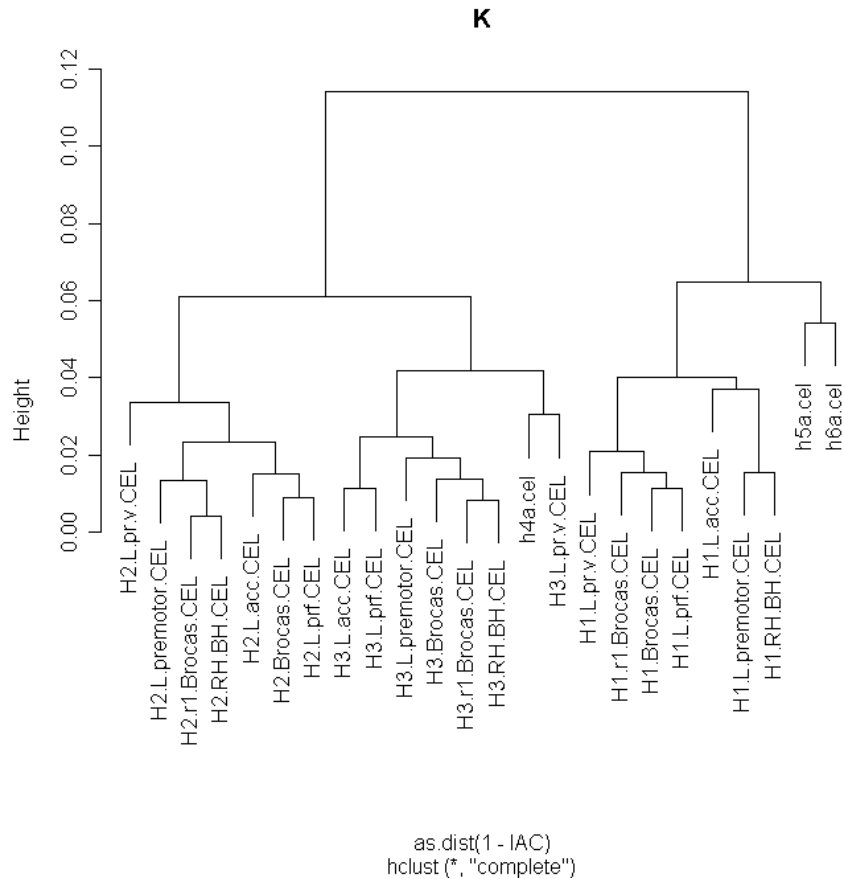
```

dimnames(dat2)[[2]]
indexK=c(23:46)
datK=dat2[,indexK]

```

```
dim(datK)
# [1] 10553 24
```

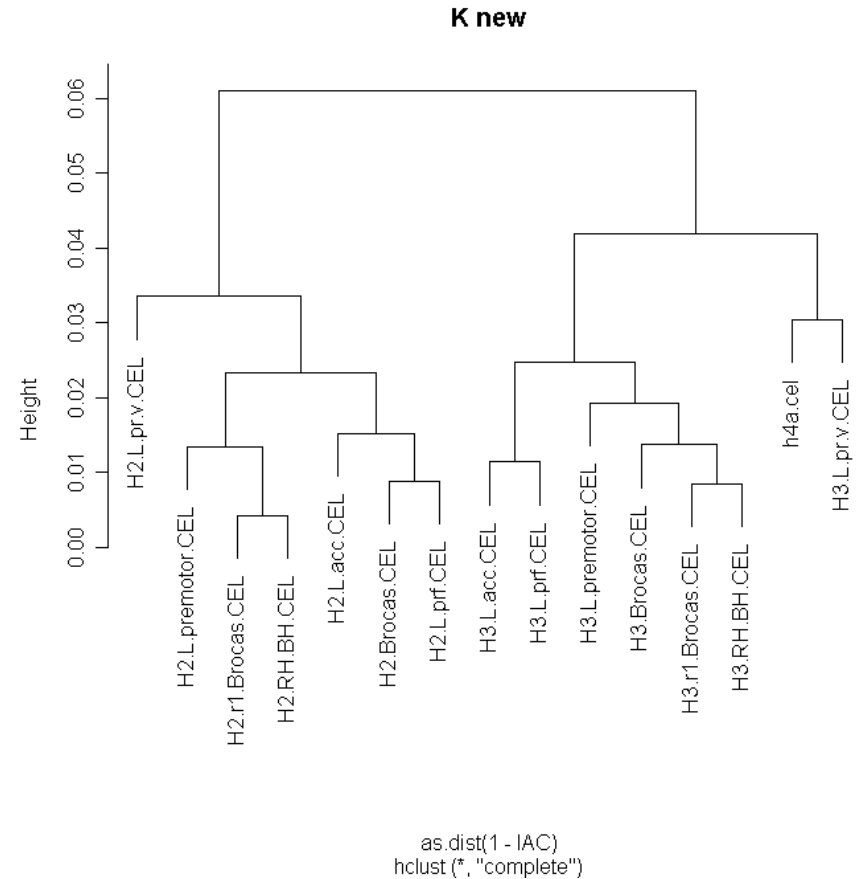
```
IAC=cor(datK,use="p")
clusterH=hclust(as.dist(1-IAC))
plot(clusterH,main="K")
```



## There is a high branch point between the nine samples at right and the remainder of the tree. Note that seven of these 9 samples were taken from a single individual (H1). We will remove these 9 samples:

```
dimnames(datK)[[2]]
keepvec=c(1,11:24)
datKnew=datK[,keepvec]
dimnames(datKnew)[[2]]
```

```
IAC=cor(datKnew,use="p")
clusterH=hclust(as.dist(1-IAC))
plot(clusterH,main="K new")
```

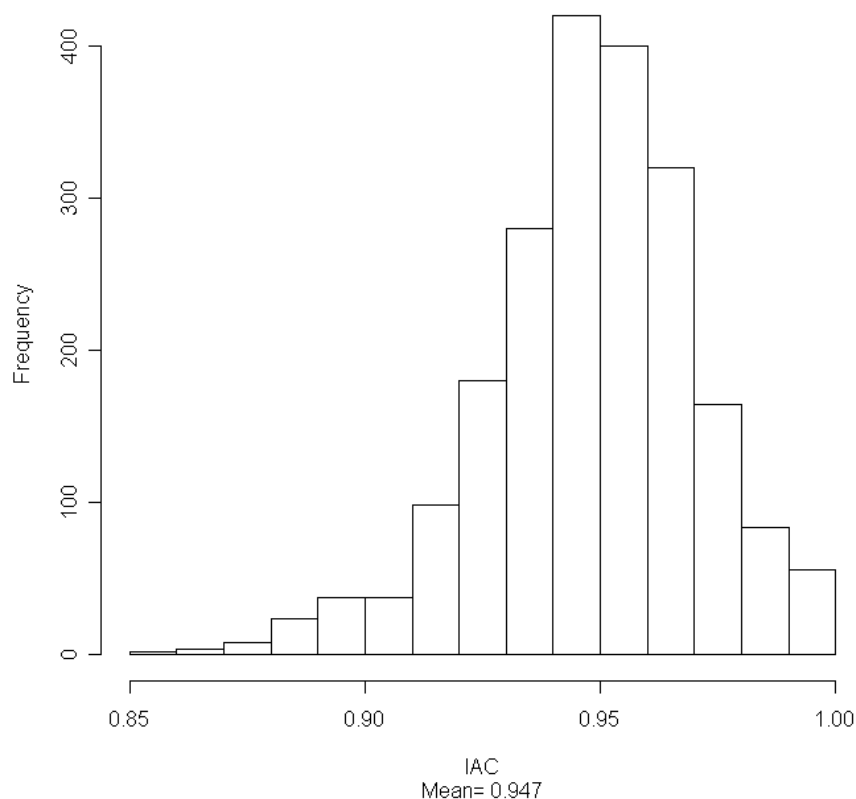


## We will stop here and combine the remaining samples from each of the individual studies:

```
datHumanNew=data.frame(datLunew,datKaznew,datCnew,datEnw,datKnew)
dim(datHumanNew)
# [1] 10553 46
```

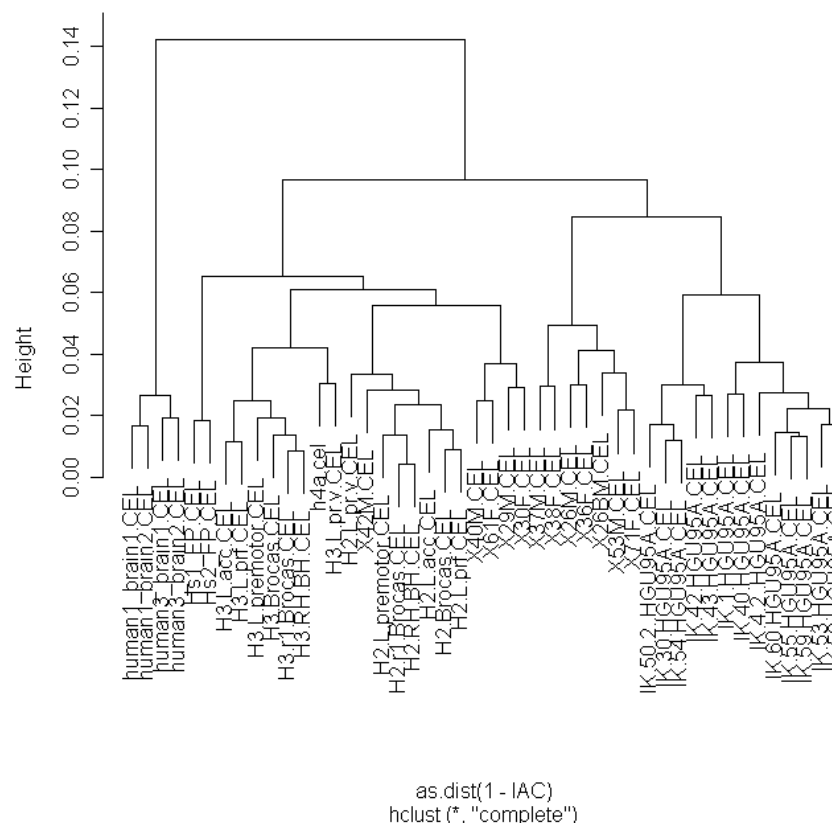
```
IAC=cor(datHumanNew,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)]),digits=3)))
```

**Histogram of IAC**



```
clusterH=hclust(as.dist(1-IAC))
plot(clusterH,main="Human New")
```

**Human New**



## Here the distribution of IACs and the dendrogram can be slightly misleading, as there are pronounced batch effects in the data. After performing quantile and batch normalization (not shown here), the mean IAC for CTX\_95 is considerably improved.

## The characteristics of the samples excluded from CTX\_95 suggest specific biological bases for the large number of outliers identified in this dataset (n=36). Half of the excluded samples were originally produced for a study on the effects of ageing on gene expression in human cerebral cortex<sup>9</sup>. This study identified significant changes in gene expression in human frontal cortex for a subset of genes as a function of age<sup>9</sup>. Our

pre-processing analysis of the samples from this study recapitulated this finding, as evidenced by divergent clustering and reduced inter-array correlations (IACs) for samples representing the most elderly individuals in this dataset. The mean age of these 18 excluded samples (representing 18 unique individuals) was 73 years old (y.o.), compared with 41 y.o. for the 12 samples from this study that were retained and 45 y.o. for all samples comprising the CTX\_95 dataset (see Supplementary Table 1 from the journal article; compare with mean ages of 48 y.o. [CTX], 59 y.o. [CN], and 58 y.o. [CB]). 12 of the remaining 18 samples identified as outliers in CTX\_95 corresponded to only three unique individuals from the original studies (seven from Human 1 of ref. 7, three from Hs3 of ref. 4, and two from human2 of ref. 5). Two of these individuals were more than 70 y.o.<sup>4,7</sup>. While the age of human2 is not known, this individual exhibited significant divergence in gene expression from the other human individuals in its original study<sup>5</sup>. Among the remaining outlier samples, two corresponded to surgical biopsies<sup>4</sup>; all other samples in our study were derived postmortem.

Although many of the samples excluded from CTX\_95 represented from elderly individuals, we reiterate that neither age nor any other sample characteristic was used in an *a priori* fashion to exclude samples from any of the datasets assembled for our analysis. Instead, samples were excluded purely on the basis of IACs, which represents an unbiased method for the identification and removal of microarray samples with divergent gene expression levels.

At this point, technical replicates from refs. 5,7 were averaged, followed by quantile and batch normalization, as described above.

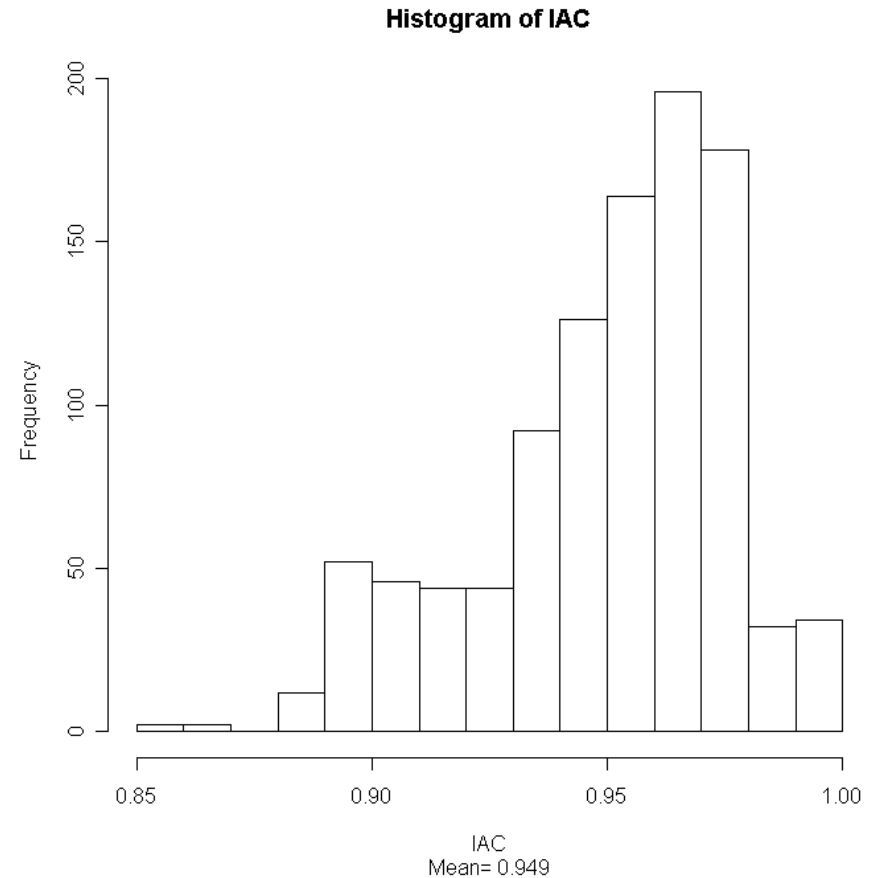
## CN

```
## Reading in the unnormalized, scaled expression data
(18,631 probe sets, 32 samples; column 1 contains probe
set IDs):
```

```
dat1=read.csv("CN_32samples_unnormalized_expression_data
.csv",header=T)
dim(dat1)
# [1] 18631 33
```

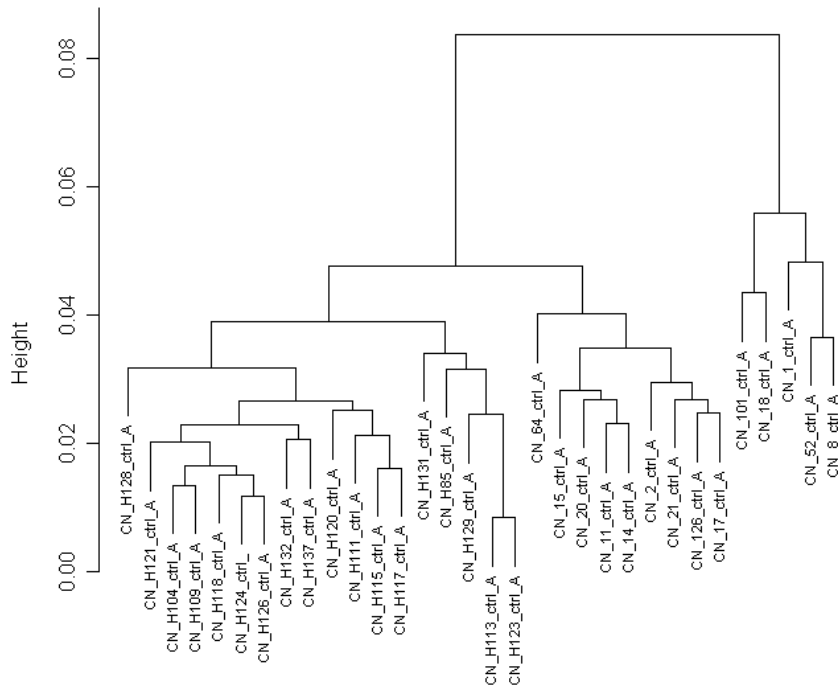
```
dat2=dat1[,2:33]
## Calculating IACs for all pairs of samples and
examining the distribution of IACs in the dataset:

IAC=cor(dat2,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC
)]),digits=3)))
```



```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat2)[[2]],main="C
audate")
```

### Caudate



```
as.dist(1 - IAC)
hclust(*, "average")
```

## There is a branch of five samples at right that are clearly divergent from the rest (101,18,1,52, and 8). We will remove these:

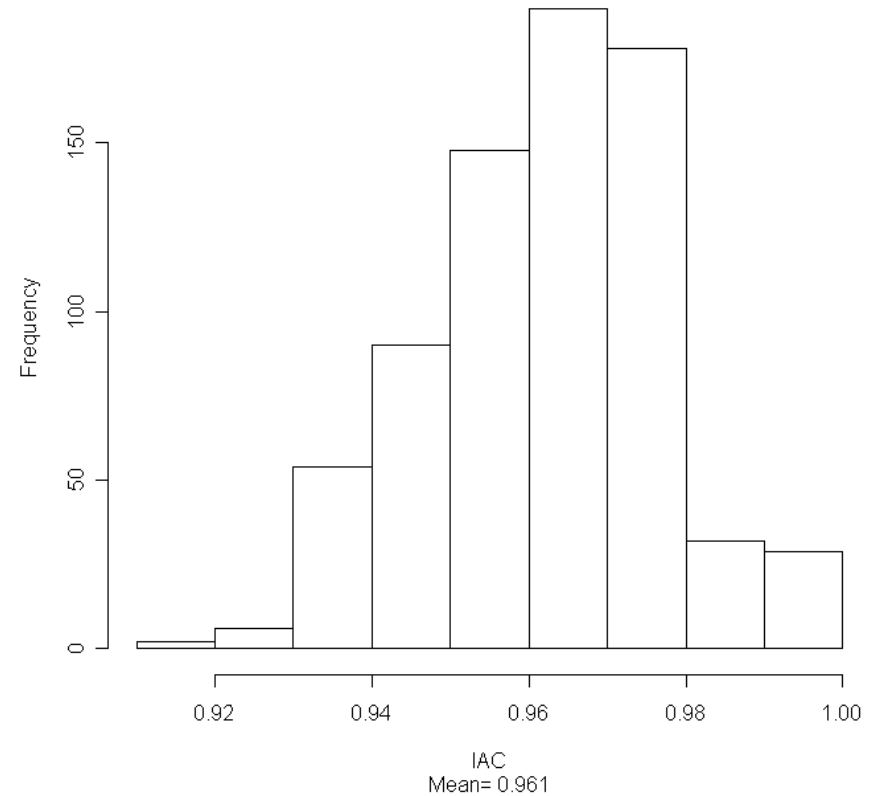
```
dimnames(dat2)[[2]]
#[1] "CN_1_ctrl_A" "CN_101_ctrl_A" "CN_11_ctrl_A"
"CN_126_ctrl_A" "CN_14_ctrl_A" "CN_15_ctrl_A"
"CN_17_ctrl_A" "CN_18_ctrl_A"
#[9] "CN_2_ctrl_A" "CN_20_ctrl_A" "CN_21_ctrl_A"
"CN_52_ctrl_A" "CN_64_ctrl_A" "CN_8_ctrl_A"
"CN_H104_ctrl_A" "CN_H109_ctrl_A"
#[17] "CN_H111_ctrl_A" "CN_H113_ctrl_A" "CN_H115_ctrl_A"
"CN_H117_ctrl_A" "CN_H118_ctrl_A" "CN_H120_ctrl_A"
"CN_H121_ctrl_A" "CN_H123_ctrl_A"
```

```
#[25] "CN_H124_ctrl_" "CN_H126_ctrl_A" "CN_H128_ctrl_A"
"CN_H129_ctrl_A" "CN_H131_ctrl_A" "CN_H132_ctrl_A"
"CN_H137_ctrl_A" "CN_H85_ctrl_A"
```

```
excludevec=c(1,2,8,12,14)
samplevec=c(1:32)
overlap=is.element(samplevec,excludevec)
dat3=dat2[,!overlap]
dim(dat3)
# [1] 18631 27
```

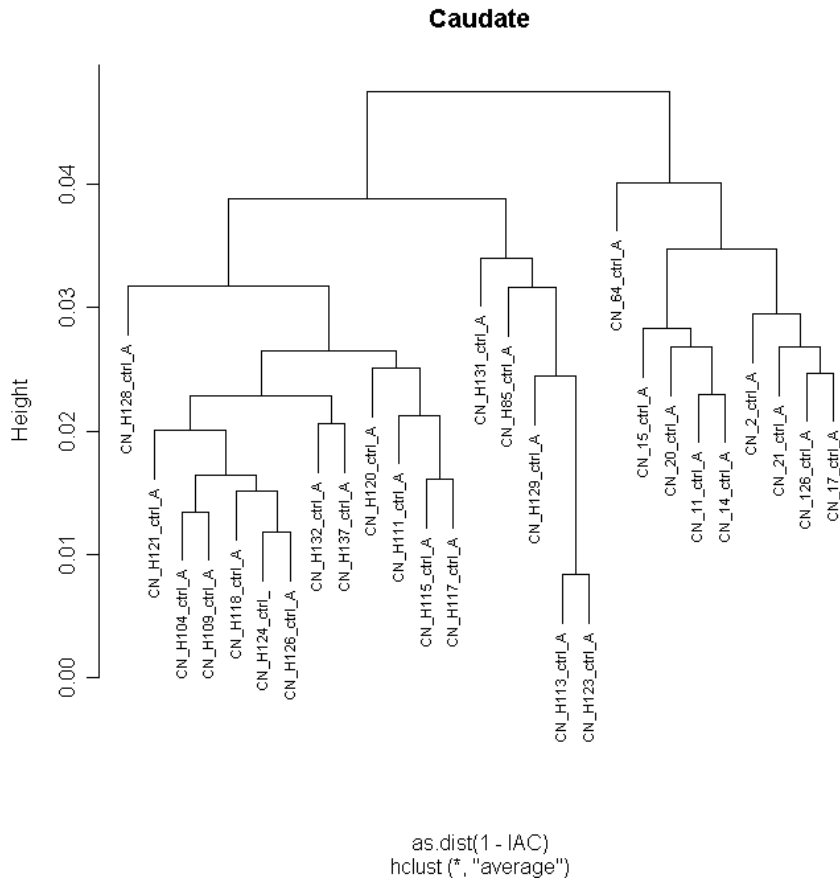
```
IAC=cor(dat3,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC)])),digits=3))
```

### Histogram of IAC



```
## Both the mean IAC and distribution look good.
Clustering:
```

```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat3)[[2]],main="C
audeate")
```



```
## The branch points are quite low. We will keep all of
these samples for quantile and batch normalization.
```

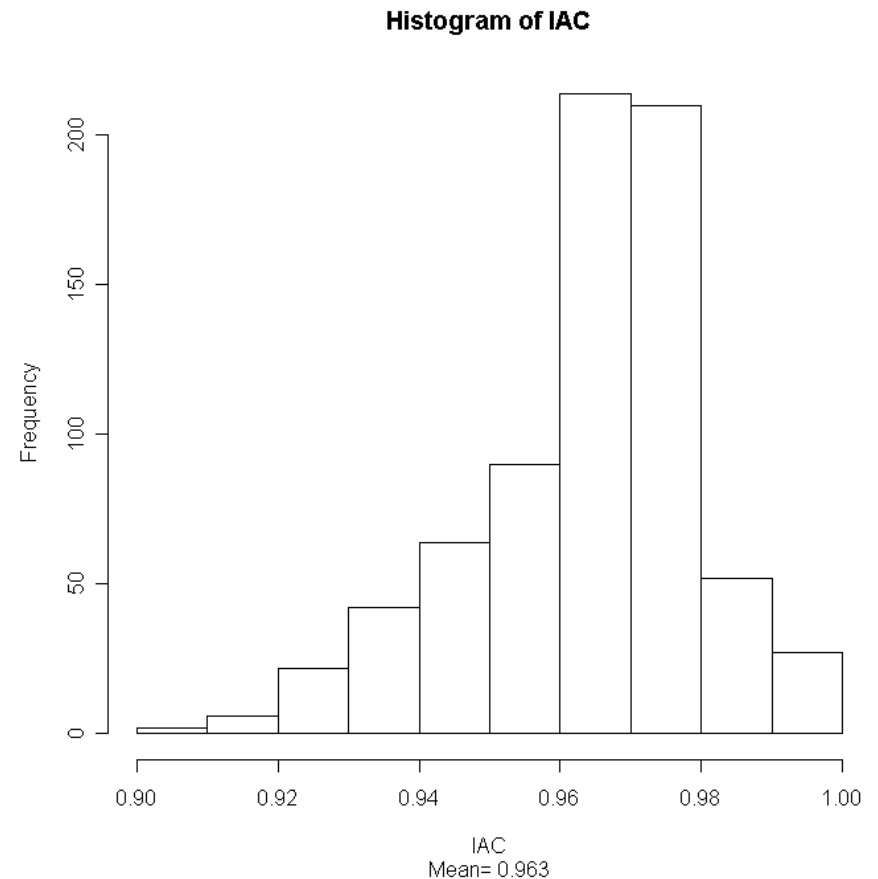
## CB

```
## Reading in the unnormalized, scaled expression data
(18,631 probe sets, 27 samples; column 1 contains probe
set IDs):
```

```
dat1=read.csv("CB_27samples_unnormalized_expression_data
.csv",header=T)
dim(dat1)
# [1] 18631 28
```

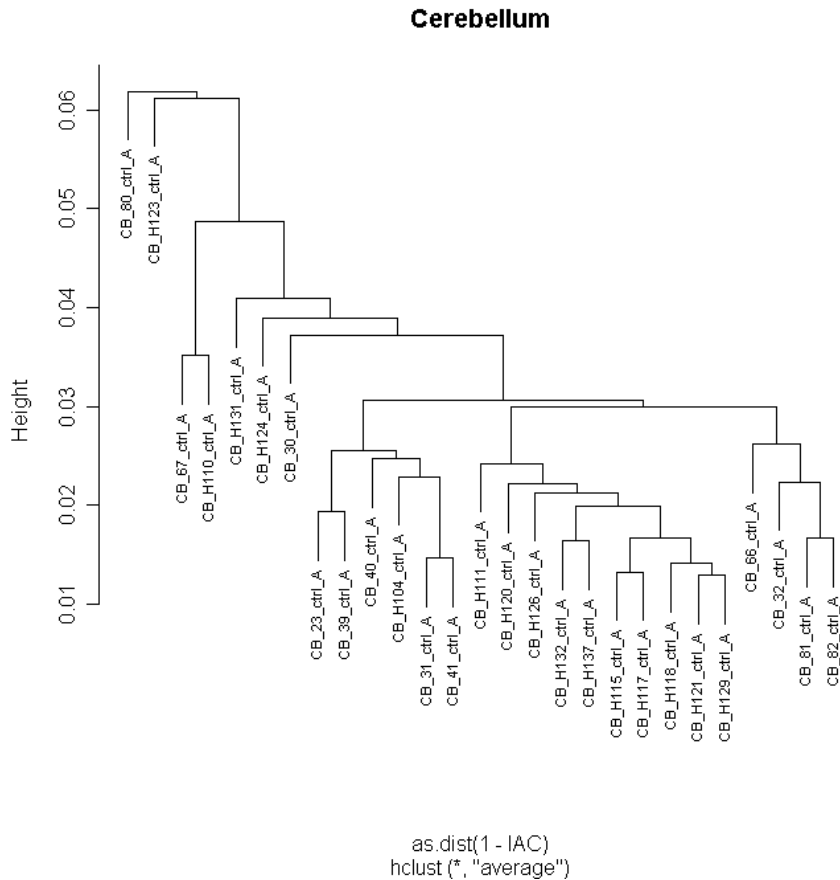
```
dat2=dat1[,2:28]
## Calculating IACs for all pairs of samples and
examining the distribution of IACs in the dataset:
```

```
IAC=cor(dat2,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC
)]),digits=3)))
```

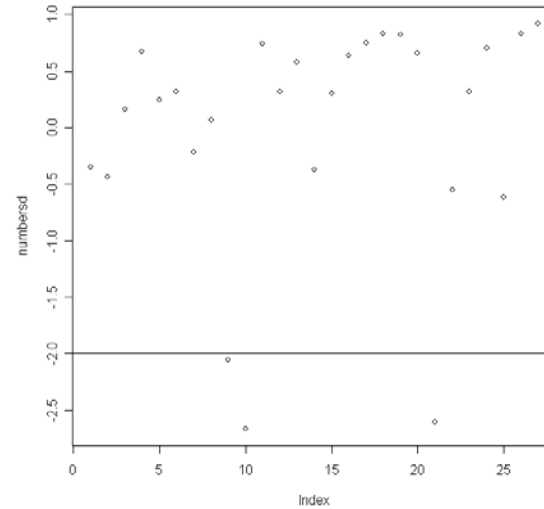


```
## The mean IAC is very good, but there may be outlier
samples.
```

```
cluster1=hclust(as.dist(1-IAC),method="average")
plot(cluster1,cex=0.7,labels=dimnames(dat2)[[2]],main="Cerebellum")
```



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



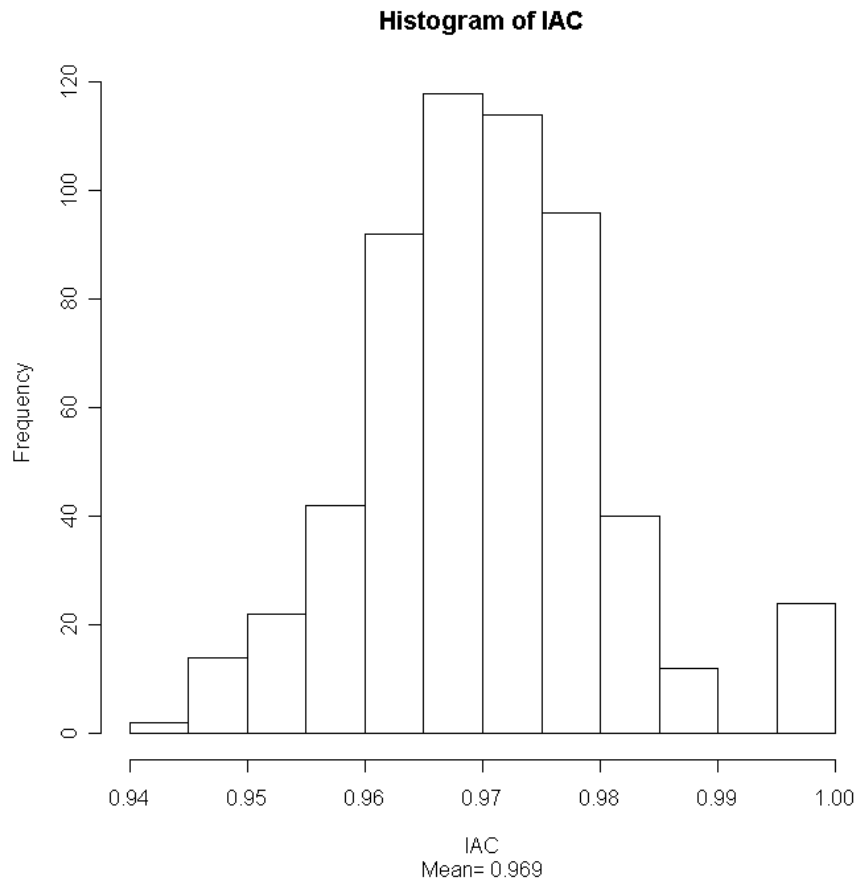
```
sdout=-2
outliers=dimnames(dat2)[[2]][numbersd<sdout]
outliers
# [1] "CB_67_ctrl_A" "CB_80_ctrl_A" "CB_H123_ctrl_A"

## The dendrogram suggests two clear outliers (80 and
123), while the above plot suggests three (80, 123, and
67). We will remove these three samples and recalculate
IACs:
```

```
dimnames(dat2)[[2]]
#[1] "CB_23_ctrl_A" "CB_30_ctrl_A" "CB_31_ctrl_A"
"CB_32_ctrl_A" "CB_39_ctrl_A" "CB_40_ctrl_A"
"CB_41_ctrl_A" "CB_66_ctrl_A"
#[9] "CB_67_ctrl_A" "CB_80_ctrl_A" "CB_81_ctrl_A"
"CB_82_ctrl_A" "CB_H104_ctrl_A" "CB_H110_ctrl_A"
"CB_H111_ctrl_A" "CB_H115_ctrl_A"
#[17] "CB_H117_ctrl_A" "CB_H118_ctrl_A" "CB_H120_ctrl_A"
"CB_H121_ctrl_A" "CB_H123_ctrl_A" "CB_H124_ctrl_A"
"CB_H126_ctrl_A" "CB_H129_ctrl_A"
#[25] "CB_H131_ctrl_A" "CB_H132_ctrl_A" "CB_H137_ctrl_A"

cbsamples=c(1:8,11:20,22:27)
dat3=dat2[,cbsamples]
dim(dat3)
# [1] 18631 24
```

```
IAC=cor(dat3,use="p")
hist(IAC,sub=paste("Mean=",format(mean(IAC[upper.tri(IAC
)]),digits=3)))
```



```
## The mean unnormalized IAC is nearly 0.97 and the
distribution looks approximately normal. We will keep
these 24 samples for quantile and batch normalization.
```

## References

1. Hodges, A. et al. Regional and cellular gene expression changes in human Huntington's disease brain. *Hum Mol Genet* **15**, 965-77 (2006).
2. Iwamoto, K., Bundo, M. & Kato, T. Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. *Hum Mol Genet* **14**, 241-53 (2005).
3. Ryan, M.M. et al. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Mol Psychiatry* **11**, 965-78 (2006).
4. Caceres, M. et al. Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A* **100**, 13030-5 (2003).
5. Enard, W. et al. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340-3 (2002).
6. Iwamoto, K., Kakiuchi, C., Bundo, M., Ikeda, K. & Kato, T. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Mol Psychiatry* **9**, 406-16 (2004).
7. Khaitovich, P. et al. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res* **14**, 1462-73 (2004).
8. Khaitovich, P. et al. A neutral model of transcriptome evolution. *PLoS Biol* **2**, E132 (2004).
9. Lu, T. et al. Gene regulation and DNA damage in the ageing human brain. *Nature* **429**, 883-91 (2004).
10. Zhang, J., Finney, R.P., Clifford, R.J., Derr, L.K. & Buetow, K.H. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics* **85**, 297-308 (2005).
11. Dai, M. et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**, e175 (2005).
12. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
13. 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).
14. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).