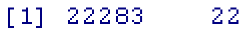
Thomas muamar

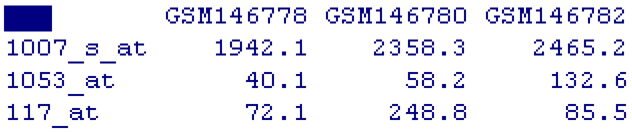
HW#1

1. data <- read.table("renal_cell_carcinoma.txt",header=T,row.names=1)
   dim(data)



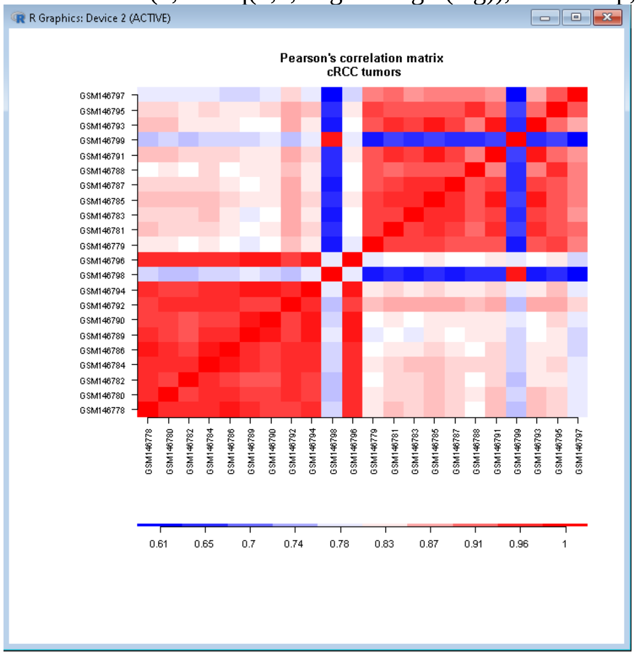
```
[1] 22283     22
```

2.

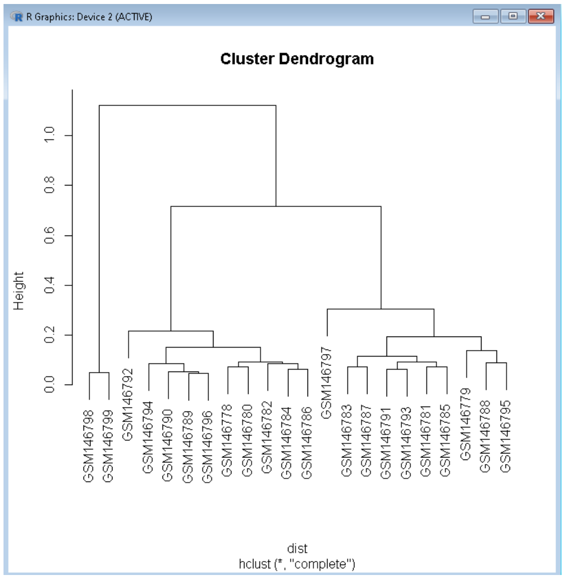


```
           GSM146778 GSM146780 GSM146782
1007_s_at     1942.1    2358.3    2465.2
1053_at         40.1      58.2     132.6
117_at          72.1     248.8      85.5
```

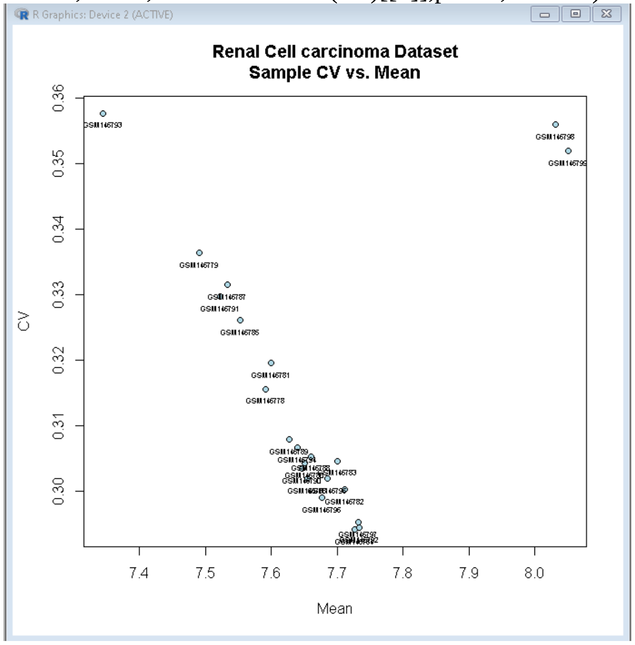
3. dat.cor <- cor(data,use="pairwise.complete.obs")
   layout(matrix(c(1,1,1,1,1,1,1,1,2,2), 5, 2, byrow = TRUE))
   par(oma=c(5,7,1,1))
    cx <- rev(colorpanel(25,"red","white","blue"))
   leg <- seq(min(dat.cor,na.rm=T),max(dat.cor,na.rm=T),length=10)
    image(dat.cor,main="Pearson's correlation matrix\n cRCC tumors",axes=F,col=cx)
   axis(1,at=seq(0,1,length=ncol(dat.cor)),label=dimnames(dat.cor)[[2]],cex.axis=0.9,las=2)
   axis(2,at=seq(0,1,length=ncol(dat.cor)),label=dimnames(dat.cor)[[2]],cex.axis=0.9,las=2)
   image(as.matrix(leg),col=cx,axes=F)
   tmp <- round(leg,2)
    axis(1,at=seq(0,1,length=length(leg)),labels=tmp,cex.axis=1)



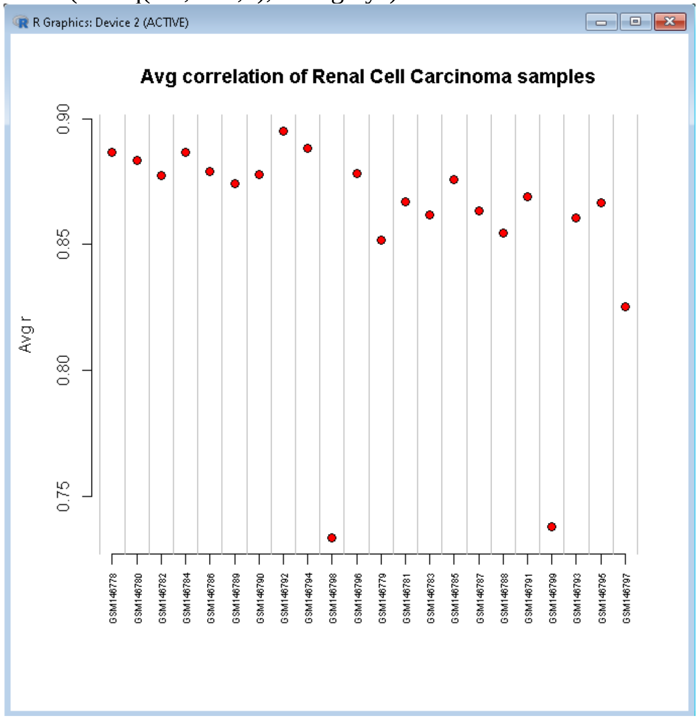
```
> dist <- dist(dat.cor , diag=TRUE)
> hc <- hclust(dist)
> plot(hc)
```



```
> dat <- as.data.frame(data)
> dat.mean <- apply(log2(dat),2,mean)
> dat.sd <- sqrt(apply(log2(dat),2,var))
> dat.cv <- dat.sd/dat.mean
> plot(dat.mean,dat.cv,main="Renal Cell carcinoma Dataset\nSample CV vs.
Mean",xlab="Mean",ylab="CV",col='blue',cex=1.5,type="n")
> points(dat.mean,dat.cv,bg="lightblue",col=1,pch=21)
> text(dat.mean,dat.cv,label=dimnames(dat)[[2]],pos=1,cex=0.5)
```



```
dat.avg <- apply(dat.cor,1,mean)
par(oma=c(3,0.1,0.1,0.1))
plot(c(1,length(dat.avg)),range(dat.avg),type="n",xlab="",ylab="Avg r",main="Avg correlation of Renal
Cell Carcinoma samples",axes=F)
points(dat.avg,bg="red",col=1,pch=21,cex=1.25)
axis(1,at=c(1:length(dat.avg)),labels=dimnames(dat)[[2]],las=2,cex.lab=0.4,cex.axis=0.6)
axis(2)
abline(v=seq(0.5,62.5,1),col="grey")
```

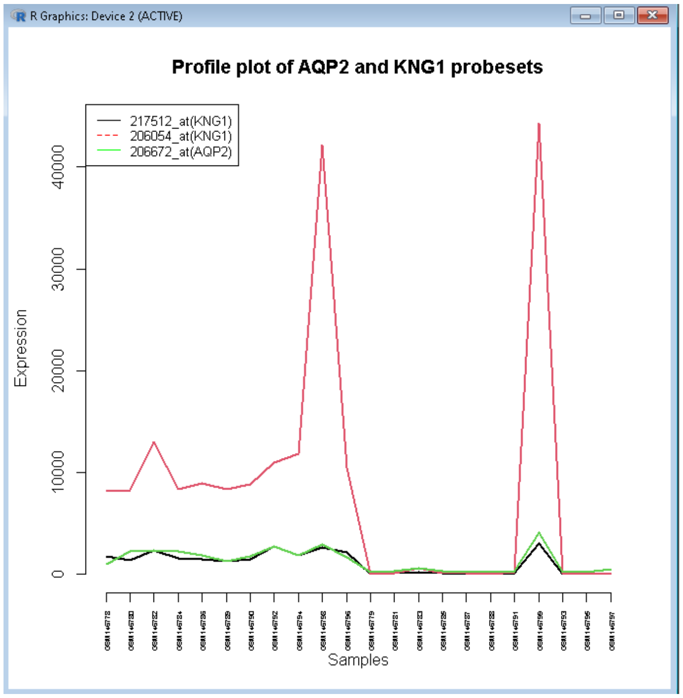


From looking through all of the data sets that are identified in these graphs the outliers are GSM146798 and GSM146799.

```
6. library(impute)
7. drop <- c("GSM146798","GSM146799")
df = dat[,!(names(dat) %in% drop)]
```

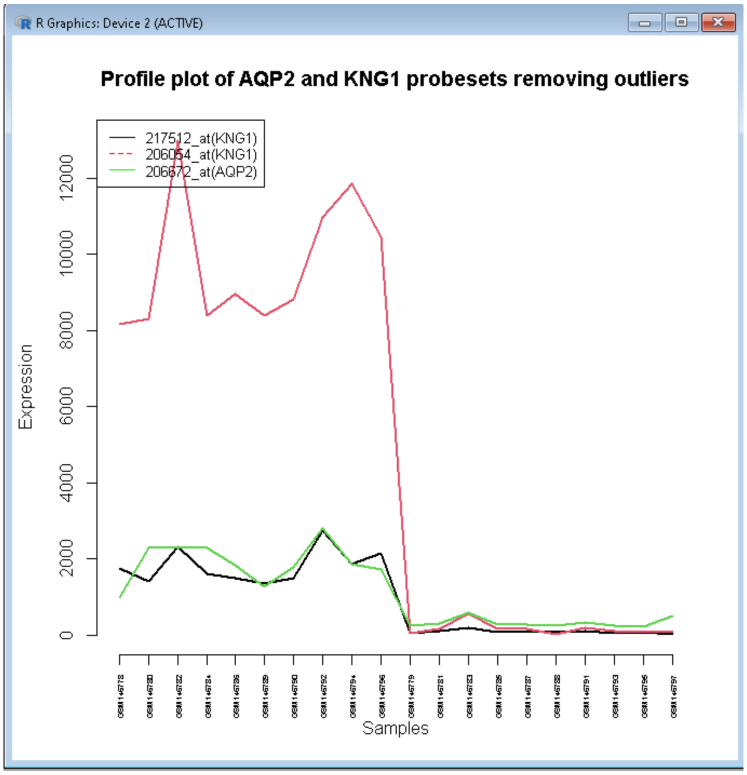
8.Plotted both data sets with outliers present and outliers removed was unsure which one was wanted.
probesets for KNG1 = 217512_at, and 206054_at
AQP2 = 206672_at

```
probesets <- dat[c("206054_at","217512_at","206672_at"),]
plot(c(1,ncol(probesets)),range(probesets[,]),type='n',main="Profile plot of AQP2 and KNG1 probesets",xlab="Samples",ylab="Expression",axes=F)
axis(side=1,at=c(1:22),labels=dimnames(probesets)[[2]],cex.axis=0.4,las=2)
axis(side=2)
for(i in 1:length(dimnames(probesets)[[1]])){
dat.y <- as.numeric(probesets[p[i],])
lines(c(1:ncol(probesets)),dat.y,col=i,lwd=2)
}
legend("topleft", legend = c("217512_at(KNG1)", "206054_at(KNG1)", "206672_at(AQP2)"), col=c("black", "red", "green"), lty=1:2, cex=0.8)
```



The data from this plot does not seem to indicate normal renal function for the probeset of 206054 which is depicted as red line from the legend we are seeing huge spikes in expression in comparison to the two other probesets.

```
plot(c(1,ncol(probesets2)),range(probesets2[,]),type='n',main="Profile plot of AQP2 and KNG1 probesets removing outliers",xlab="Samples",ylab="Expression",axes=F)
axis(side=1,at=c(1:20),labels=dimnames(probesets)[[2]],cex.axis=0.4,las=2)
axis(side=2)
for(i in 1:length(dimnames(probesets2)[[1]])){
dat.y <- as.numeric(probesets2[p[i],])
lines(c(1:ncol(probesets2)),dat.y,col=i,lwd=2)
}
legend("topleft", legend = c("217512_at(KNG1)", "206054_at(KNG1)", "206672_at(AQP2)"), col=c("black", "red", "green"), lty=1:2, cex=0.8)
```



9. repalced2<- replace(data,data==8385.3,NA)
datam<- data.matrix(repalced2)
Without outliers:
repalced3<- replace(df,df==8385.3,NA)
dataf<- data.matrix(repalced3)

10.
new<-impute.knn(datam, k = 6)
showing the value produced:
imputematrix <- as.matrix(new)
imputeddata <- data.frame(imputematrix[1,])
imputeddata["206054_at","data.GSM146784"]

11.
With out outliers removed:
8385.3
7632.35
Relative error = |(8385.3-7632.35)|/ 8385.3= 0.0898= 9%
Removing outliers:
8385.3
7559.533
|(8385.3-7559.533)|/ 8385.3 = 0.0985 = 10%

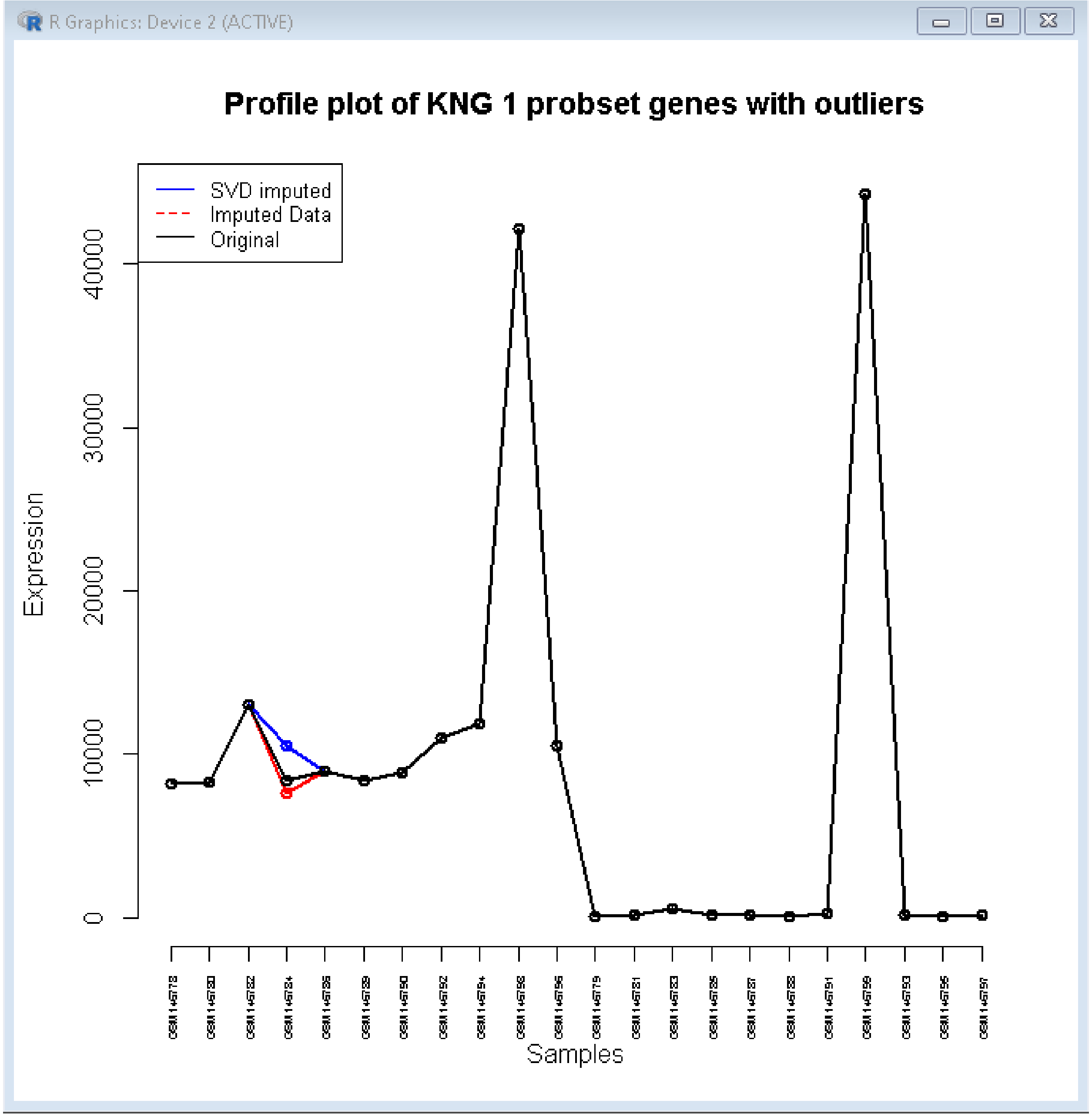
12. pc<-pca(datam, nPcs = 9, method="svdImpute")
imputedSVD <- completeObs(pc)
imputedSVD["206054_at", "GSM146784"]
value with outliers = 10541.34
For the Values without outliers:
pc2<-pca(datam, nPcs = 9, method="svdImpute")

```
    imputedSVD2 <- completeObs(pc2)
```

value produced from imputation without out liers = 10418

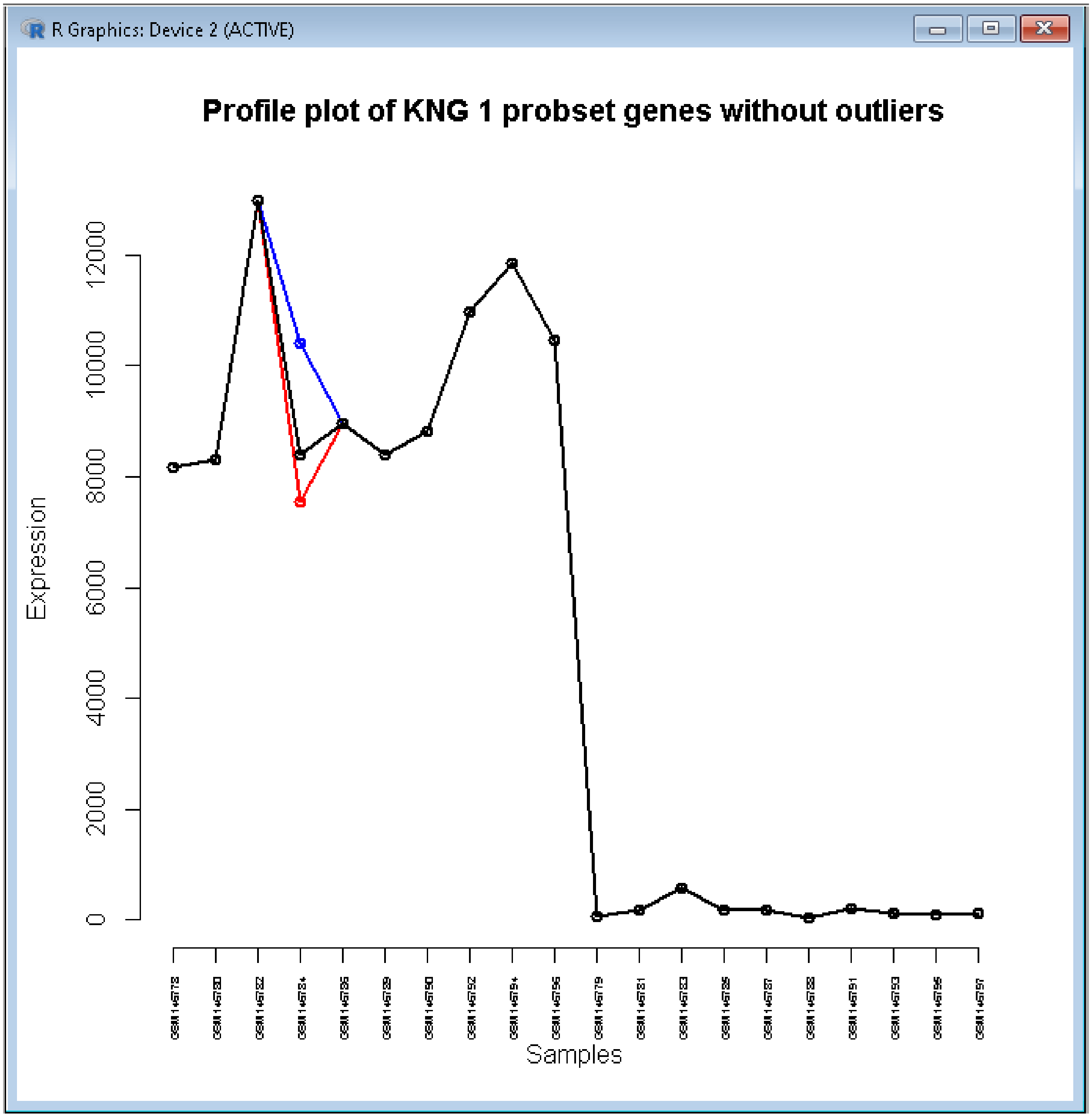
13.

```
 colnames(imputdat) <- c("GSM146778",
"GSM146780","GSM146782","GSM146784","GSM146786","GSM146789","GSM146790","GSM146792",
    +
    "GSM146794","GSM146798","GSM146796","GSM146779","GSM146781","GSM146783","GSM146785
    ","GSM146787","GSM146788","GSM146791","GSM146799","GSM146793","GSM146795","GSM14679
    7")
imputedSVD2<- data.frame(imputedSVD)
SVD <- imputedSVD2["206054_at",]
Imputdat <- imputeddata[“206054_at”,]
rand.genes <- sample(dimnames(data)[[1]])
plot(c(1,ncol(dat)),range(imputdat[,]),type='n',main="Profile plot of KNG 1 probset genes with
outliers",xlab="Samples",ylab="Expression",axes=F)
axis(side=1,at=c(1:22),labels=dimnames(data)[[2]],cex.axis=0.4,las=2)
axis(side=2)
for(i in 1:length(rand.genes)) {
dat.n <- as.numeric(SVD[rand.genes[i],])
dat.y <- as.numeric(imputdat[rand.genes[i],])
dat.z<- as.numeric(dat[rand.genes[i],])
lines(c(1:ncol(dat)),dat.n,col="blue",lwd=2)
lines(c(1:ncol(dat)),dat.y,col="red",lwd=2)
lines(c(1:ncol(dat)),dat.z,col="black",lwd=2)
points(c(1:ncol(dat)),dat.n,col="blue",lwd=2)
points(c(1:ncol(dat)),dat.y,col="red",lwd=2)
points(c(1:ncol(dat)),dat.z,col="black",lwd=2)
}
legend("topleft",legend=c("SVD imputed", "Imputed Data", "Original"),col=c("blue","red","black"),lty=1:2,
cex=0.8)
```



Without the outliers:

```
df2<- df["206054_at",]
imputdat2 <- imputeddata2["206054_at",]
imputedSVD3 <- data.frame(imputedSVD2)
SVD2 <- imputedSVD3["206054_at",]
rand.genes2 <- sample(dimnames(df)[[1]])
plot(c(1,ncol(df)),range(imputdat2[,]),type='n',main="Profile plot of KNG 1 probset genes without outliers",xlab="Samples",ylab="Expression",axes=F)
axis(side=1,at=c(1:20),labels=dimnames(df)[[2]],cex.axis=0.4,las=2)
axis(side=2)
for(i in 1:length(rand.genes)) {
dat.n <- as.numeric(SVD2[rand.genes2[i],])
dat.y <- as.numeric(imputdat2[rand.genes2[i],])
dat.z<- as.numeric(df2[rand.genes2[i],])
lines(c(1:ncol(df)),dat.n,col="blue",lwd=2)
lines(c(1:ncol(df)),dat.y,col="red",lwd=2)
lines(c(1:ncol(df)),dat.z,col="black",lwd=2)
points(c(1:ncol(df)),dat.n,col="blue",lwd=2)
points(c(1:ncol(df)),dat.y,col="red",lwd=2)
points(c(1:ncol(df)),dat.z,col="black",lwd=2)
}
legend("topleft",legend=c("SVD imputed", "Imputed Data", "Original"),col=c("blue","red","black"),lty=1:2,
cex=0.8)
```


R Graphics: Device 2 (ACTIVE)
Profile plot of KNG 1 probset genes without outliers
Expression
0
2000
4000
6000
8000
10000
12000
Samples
GSM146778
GSM146780
GSM146782
GSM146784
GSM146786
GSM146789
GSM146790
GSM146792
GSM146794
GSM146796
GSM146779
GSM146781
GSM146783
GSM146785
GSM146787
GSM146788
GSM146791
GSM146793
GSM146795
GSM146797