Julie Garcia
February 11, 2018

Lab 4
Advanced Genomics and Genetics Analyses

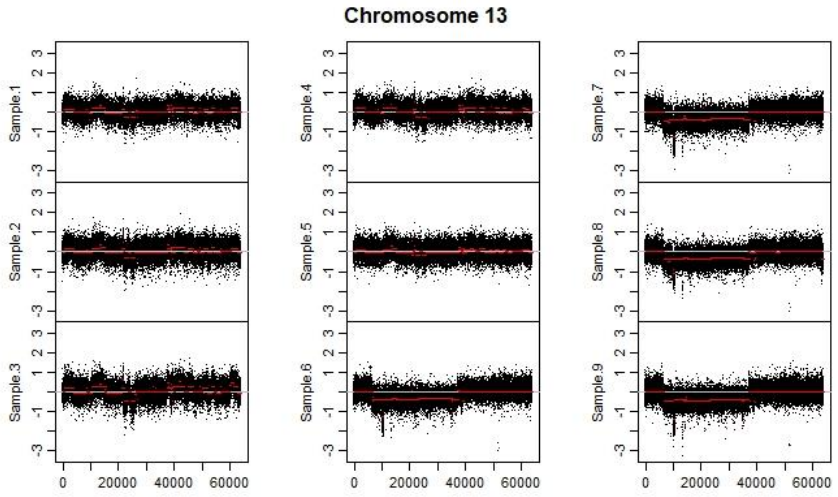
In this lab, we will be analyzing copy number alterations from a set of 9 arrays. These arrays were taken from a prostate cancer cell line study, where each array represents a different cell line. I have already processed the raw CEL files using the APT software and provided you the formatted output data for all 23 chromosomes for these 9 cell lines, to save you time. You will be implementing some tools in various R packages to visualize CN alterations and understand the similarity between these 9 cell lines. Then you will use ABSOLUTE to calculate the ploidy and tumor purity for one of the cell lines in the dataset.

1.) Get the copy number file from the website called CNV_processed_files.zip. This contains both CN state values and $\log_2$ ratio values for the 9 samples

```
cn = read.table("C:\\TEMP\\datasets\\cn_states.txt", header=T)
logratio = read.table("C:\\TEMP\\datasets\\log2ratios.txt", header=T)
dim(cn)
dim(logratio)
```

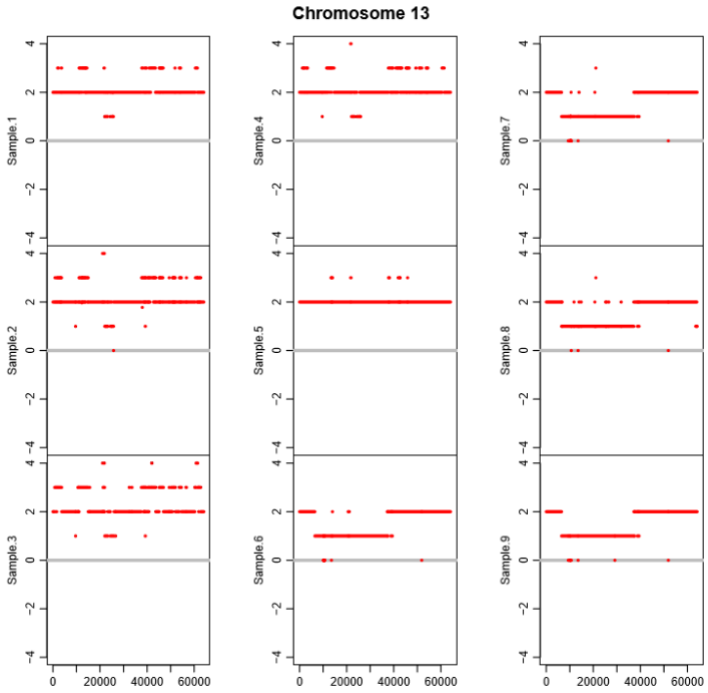
2.) Subset the data by chromosome 13, run smoothing and segmentation with the DNAcopy library. Then plot the curves for all 9 samples for chromosome 13.

```
#2 subset the data by chromosome 13
cn.13 <- cn[cn$Chromosome==13,]
logratio.13 <- logratio[logratio$Chromosome==13,]
dim(cn.13)
dim(logratio.13)

# run smoothing and segmentation
source("https://bioconductor.org/biocLite.R")
biocLite("DNAcopy")
library(DNAcopy)
dx <- logratio.13[,-c(1:3)]
d.logratio <-
CNA(genomdat=as.matrix(dx),chrom=logratio.13$Chromosome,maploc=logratio.
13$Position,data.type=c("logratio"),sampleid=NULL)
d.smoothed <- smooth.CNA(d.logratio)
d.segment <- segment(d.smoothed, verbose=1)

#plot results for chr 1 for all subjects
help(pdf)
pdf("CNV_plot.pdf")
plot(d.segment, plot.type="chrombysample", pt.cex=0.5,lwd=0.5)
dev.off()
```



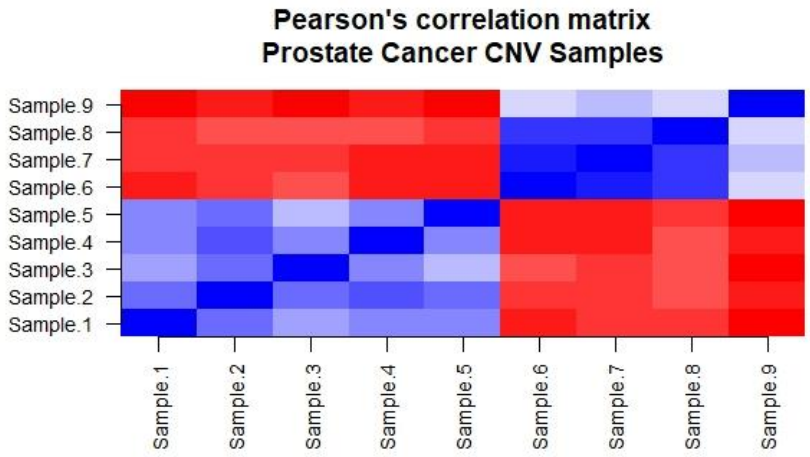
3.) Now do the same type of plots using the plot() function for chromosome 13 using the CN state values for the 9 samples. Make sure to title and label things appropriately.

```
cx <- cn.13[,-c(1:3)]
d.cx <- CNA(genomdat=as.matrix(cx),chrom=cn.13$Chromosome,
            maploc=cn.13$Position,data.type=c("binary"),sampleid=NULL)
d.cx.segment <- segment(d.cx, verbose=1)
pdf("CNV_CNStates_plot.pdf")
plot(d.cx.segment, plot.type="chrombysample", main="CN states\nAll
Chromosomes")
dev.off()
```



4.) Using the CNseg(), getRS(), and rs() functions in the CNTools package on the output from the segment() function, create a data matrix, run correlation between the 9 samples, and plot this correlation heatmap using the image() function. Be sure to label the x and y-axes appropriately and title the plot.

```
source("https://bioconductor.org/biocLite.R")
biocLite("CNTools")
library(CNTools)

#create a data matrix of the combined CN vectors
seg <- CNSeg(d.segment$output)
rs.region <- getRS(seg,by="region", imput=FALSE, XY=FALSE,what =
"mean")
mat <- rs(rs.region)
mat


#plot the correlation matrix
row.names(mat) <- mat$start
mat.data <- mat[,4:12]
mat.data <- data.matrix(mat.data)
mat.cor <- cor(mat.data,use="pairwise.complete.obs")

par(oma=c(3,3,1,1))
colors <- colorRampPalette(c("red", "white", "blue"))(20)
image(mat.cor,main="Pearson's correlation matrix\nProstate Cancer
CNV Samples",axes=F,col=colors)
axis(1,at=seq(0,1,length=ncol(mat.cor)),label=dimnames(mat.cor)[[
2]],cex.axis=0.9,las=2)
axis(2,at=seq(0,1,length=ncol(mat.cor)),label=dimnames(mat.cor)[[
2]],cex.axis=0.9,las=2)
```



5.) Now go back to the $\log_2$ ratio file and extract the first 5 columns, which includes just the first cell line using awk, or some other linux command. Then read this 5 column file into R and load the ABSOLUTE and numDeriv packages. The ABSOLUTE R package is available on the course website.

I used this command to subset the file:

```
awk '{print $1, $2, $3, $4, $5}' log2ratios.txt >
log2ratio5columns.txt
```

In R:

```
#5 read in subsetted file
logratio2 =
read.table("C:\\TEMP\\datasets\\log2ratio5columns.txt", header=T,
row.names=1)
dim(logratio2)

#install absolute & numDeriv
install.packages("numDeriv")
install.packages(pkgs="~//Downloads//ABSOLUTE_1.0.6.tar.gz",
repos=NULL, type="source")
library(numDeriv)
library(ABSOLUTE)
```

6.) Run CNA(), smoothing, and segmentation with the DNAcopy library across the genome for this single cell line.

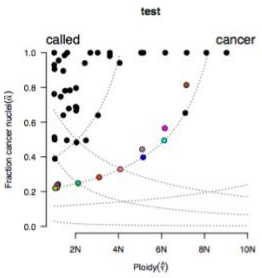
```
dx2 <- logratio2[,-c(1:3)]
d.logratio2 <-
CNA(genomdat=as.matrix(dx2),chrom=logratio2$Chromosome,
maploc=logratio2$Position,data.type=c("logratio"),sampleid=NULL)
d.smoothed2 <- smooth.CNA(d.logratio2)
d.segment2 <- segment(d.smoothed2, verbose=1)
```

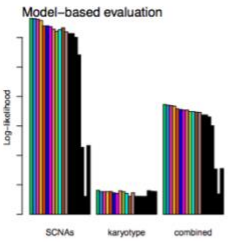
7.) Use the output of the segmentation to run ABSOLUTE with the following parameters (we are coding this as a sequencing technology, though it's really an array technology):
sigma.p <- 0
max.sigma.h <- 0.02
min.ploidy <- 0.95
max.ploidy <- 10
max.as.seg.count <- 1500
max.non.clonal <- 0
max.neg.genome <- 0
genome <- "hg19"
platform <- "Illumina_WES"

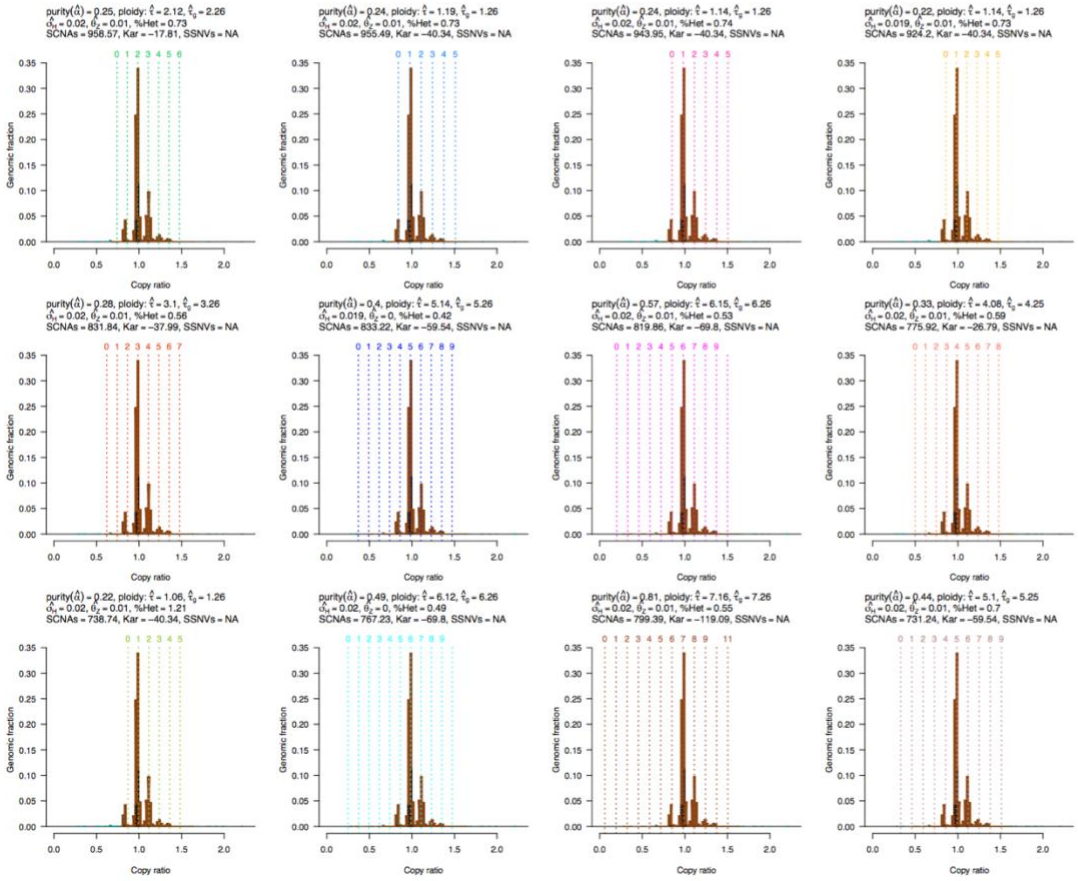
This step will take approximately **a few hours to run**, so plan accordingly. Now looking just at the pdf output generated from the RunAbsolute() function, for the optimal solution determined, what is tumor purity calculated for this cell line as well as approximate ploidy state?

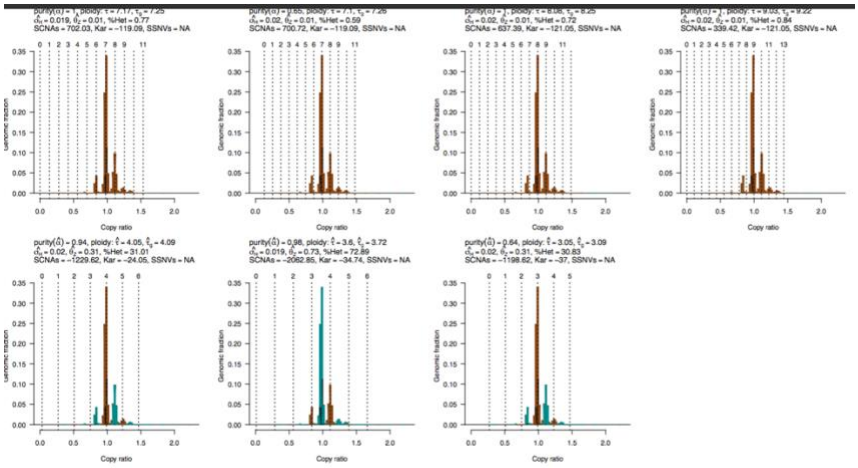
```
dat <- d.segment2$output
dat
names(dat) <- c("ID", "Chromosome", "Start", "End", "Num_Probes",
"Segment_Mean")
dat.fn <- "test.abs.dat"
write.table(file=dat.fn, dat, quote=F, col.names=T, row.names=F,
sep="\t")

primary.disease <- 'cancer'
sample.name <- 'test'
sigma.p <- 0
max.sigma.h <- 0.02
min.ploidy <- 0.95
max.ploidy <- 10
max.as.seg.count <- 1500
max.non.clonal <- 0
max.neg.genome <- 0
genome <- "hg19"
platform <- "Illumina_WES"

RunAbsolute(dat.fn, sigma.p, max.sigma.h, min.ploidy, max.ploidy,
            primary.disease, platform, sample.name, ".",
            max.as.seg.count, max.non.clonal,max.neg.genome,
"total", verbose=TRUE)
```

The tumor purity calculated by RunAbsolute is 0.25 and ploidy is 2.12. I attached the output from the pdf below.


test
called
cancer
Fraction cancer nuclei($\hat{\alpha}$)
Ploidy($\hat{\tau}$)
2N
4N
6N
8N
10N


Model−based evaluation
Log−likelihood
SCNAs
karyotype
combined


purity($\hat{\alpha}$) = 0.25, ploidy: $\hat{\tau}$ = 2.12, $\hat{\tau}_0$ = 2.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.73
SCNAs = 958.57, Kar = −17.81, SSNVs = NA
purity($\hat{\alpha}$) = 0.24, ploidy: $\hat{\tau}$ = 1.19, $\hat{\tau}_0$ = 1.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.73
SCNAs = 955.49, Kar = −40.34, SSNVs = NA
purity($\hat{\alpha}$) = 0.24, ploidy: $\hat{\tau}$ = 1.14, $\hat{\tau}_0$ = 1.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.74
SCNAs = 943.95, Kar = −40.34, SSNVs = NA
purity($\hat{\alpha}$) = 0.22, ploidy: $\hat{\tau}$ = 1.14, $\hat{\tau}_0$ = 1.26
$\hat{\alpha}_0$ = 0.019, $\hat{\theta}_2$ = 0.01, %Het = 0.73
SCNAs = 924.2, Kar = −40.34, SSNVs = NA
purity($\hat{\alpha}$) = 0.28, ploidy: $\hat{\tau}$ = 3.1, $\hat{\tau}_0$ = 3.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.56
SCNAs = 831.84, Kar = −37.99, SSNVs = NA
purity($\hat{\alpha}$) = 0.4, ploidy: $\hat{\tau}$ = 5.14, $\hat{\tau}_0$ = 5.26
$\hat{\alpha}_0$ = 0.019, $\hat{\theta}_2$ = 0, %Het = 0.42
SCNAs = 833.22, Kar = −59.54, SSNVs = NA
purity($\hat{\alpha}$) = 0.57, ploidy: $\hat{\tau}$ = 6.15, $\hat{\tau}_0$ = 6.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.53
SCNAs = 819.86, Kar = −69.8, SSNVs = NA
purity($\hat{\alpha}$) = 0.33, ploidy: $\hat{\tau}$ = 4.08, $\hat{\tau}_0$ = 4.25
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.59
SCNAs = 775.92, Kar = −26.79, SSNVs = NA
purity($\hat{\alpha}$) = 0.22, ploidy: $\hat{\tau}$ = 1.06, $\hat{\tau}_0$ = 1.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 1.21
SCNAs = 738.74, Kar = −40.34, SSNVs = NA
purity($\hat{\alpha}$) = 0.49, ploidy: $\hat{\tau}$ = 6.12, $\hat{\tau}_0$ = 6.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0, %Het = 0.49
SCNAs = 767.23, Kar = −69.8, SSNVs = NA
purity($\hat{\alpha}$) = 0.81, ploidy: $\hat{\tau}$ = 7.16, $\hat{\tau}_0$ = 7.26
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.55
SCNAs = 799.39, Kar = −119.09, SSNVs = NA
purity($\hat{\alpha}$) = 0.44, ploidy: $\hat{\tau}$ = 5.1, $\hat{\tau}_0$ = 5.25
$\hat{\alpha}_0$ = 0.02, $\hat{\theta}_2$ = 0.01, %Het = 0.7
SCNAs = 731.24, Kar = −59.54, SSNVs = NA
Genomic fraction
Copy ratio


SCNAs = 702.03, Kar = −119.09, SSNVs = NA
SCNAs = 700.72, Kar = −119.09, SSNVs = NA
SCNAs = 637.39, Kar = −121.05, SSNVs = NA
SCNAs = 339.42, Kar = −121.05, SSNVs = NA
SCNAs = −1229.62, Kar = −24.05, SSNVs = NA
SCNAs = −2082.85, Kar = −34.74, SSNVs = NA
SCNAs = −1198.62, Kar = −37, SSNVs = NA
Genomic fraction
Copy ratio