## Prepare environment and requirement

install.packages("BiocManager", repos="https://cran.ms.unimelb.edu.au/")

## Installing package into 'C:/Users/iefad/Documents/R/win-library/4.0'  
## (as 'lib' is unspecified)

## package 'BiocManager' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\iefad\AppData\Local\Temp\RtmpofNldM\downloaded\_packages

BiocManager::install()

## Bioconductor version 3.11 (BiocManager 1.30.10), R 4.0.2 (2020-06-22)

## Installation path not writeable, unable to update packages: MASS, mgcv, nlme,  
## survival

## Old packages: 'digest'

BiocManager::install(c("copynumber", "AnnotationDbi"))

## Bioconductor version 3.11 (BiocManager 1.30.10), R 4.0.2 (2020-06-22)

## Installing package(s) 'copynumber', 'AnnotationDbi'

## package 'copynumber' successfully unpacked and MD5 sums checked  
## package 'AnnotationDbi' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\iefad\AppData\Local\Temp\RtmpofNldM\downloaded\_packages

## Installation path not writeable, unable to update packages: MASS, mgcv, nlme,  
## survival

## Old packages: 'digest'

if(!require(squash)) {  
 install.packages("squash", repos="https://cran.ms.unimelb.edu.au/")  
}

## Loading required package: squash

if(!require(sequenza)) {  
 install.packages("sequenza", repos="https://cran.ms.unimelb.edu.au/")  
}

## Loading required package: sequenza

if(!require(stringr)) {  
 install.packages("stringr", repos="https://cran.ms.unimelb.edu.au/")  
}

## Loading required package: stringr

##Set up work directory

setwd("C:/Users/iefad/Desktop/BINF90004")  
input.files<-list.files(path="./data",full.names=TRUE)  
input.files

## [1] "./data/CA0023-61-2.chr3\_bin50.seqz.gz"  
## [2] "./data/CA0023-61-3.chr3\_bin50.seqz.gz"  
## [3] "./data/CA0023-65.chr3\_bin50.seqz.gz"   
## [4] "./data/CA0023-66.chr3\_bin50.seqz.gz"

cur.path <- getwd()

##load library

library(copynumber)

## Loading required package: BiocGenerics

## Loading required package: parallel

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':  
##   
## clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,  
## clusterExport, clusterMap, parApply, parCapply, parLapply,  
## parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, append, as.data.frame, basename, cbind, colnames,  
## dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,  
## grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,  
## order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,  
## rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,  
## union, unique, unsplit, which, which.max, which.min

library(stringr)  
library(stringr)

## Pre-explore the data

Before formally run Sequenza, the data is firstly pre-tested.

### Data loading and specific the position

dat.1 <- read.delim(gzfile(input.files[1]), sep="\t")  
dat.1$x = dat.1$position/1e6  
head(dat.1)

## chromosome position base.ref depth.normal depth.tumor depth.ratio Af Bf  
## 1 3 239221 N 25 10 0.391 1.000 0.0  
## 2 3 239245 C 25 10 0.400 0.800 0.2  
## 3 3 239272 N 40 29 0.718 1.000 0.0  
## 4 3 239323 N 77 66 0.872 1.000 0.0  
## 5 3 239374 N 113 100 0.888 1.000 0.0  
## 6 3 239382 G 102 92 0.902 0.989 0.0  
## zygosity.normal GC.percent good.reads AB.normal AB.tumor tumor.strand  
## 1 hom 71 51 N . 0  
## 2 het 71 10 CG . 0  
## 3 hom 68 51 N . 0  
## 4 hom 65 51 N . 0  
## 5 hom 68 51 N . 0  
## 6 hom 68 92 G A0.011 A0.0  
## x  
## 1 0.239221  
## 2 0.239245  
## 3 0.239272  
## 4 0.239323  
## 5 0.239374  
## 6 0.239382

dat.2 <- read.delim(gzfile(input.files[2]), sep="\t")  
dat.2$x = dat.2$position/1e6  
head(dat.2)

## chromosome position base.ref depth.normal depth.tumor depth.ratio Af Bf  
## 1 3 239203 N 18 11 0.655 1.000 0.000  
## 2 3 239245 C 25 14 0.560 0.857 0.143  
## 3 3 239254 N 36 30 0.801 1.000 0.000  
## 4 3 239304 A 48 48 1.000 0.979 0.000  
## 5 3 239305 N 58 65 1.091 1.000 0.000  
## 6 3 239341 C 74 85 1.149 0.988 0.000  
## zygosity.normal GC.percent good.reads AB.normal AB.tumor tumor.strand  
## 1 hom 74 51 N . 0  
## 2 het 74 14 CG . 0  
## 3 hom 66 51 N . 0  
## 4 hom 66 48 A C0.021 C1.0  
## 5 hom 69 51 N . 0  
## 6 hom 69 85 C A0.012 A1.0  
## x  
## 1 0.239203  
## 2 0.239245  
## 3 0.239254  
## 4 0.239304  
## 5 0.239305  
## 6 0.239341

dat.3 <- read.delim(gzfile(input.files[3]), sep="\t")  
dat.3$x = dat.3$position/1e6  
head(dat.3)

## chromosome position base.ref depth.normal depth.tumor depth.ratio Af Bf  
## 1 3 239221 N 25 3 0.131 1.000 0.000  
## 2 3 239245 C 25 3 0.120 0.667 0.333  
## 3 3 239272 N 40 7 0.170 1.000 0.000  
## 4 3 239323 N 77 21 0.282 1.000 0.000  
## 5 3 239374 N 113 32 0.285 1.000 0.000  
## 6 3 239407 T 120 35 0.292 0.971 0.000  
## zygosity.normal GC.percent good.reads AB.normal AB.tumor tumor.strand  
## 1 hom 71 51 N . 0  
## 2 het 71 3 GC . 0  
## 3 hom 68 51 N . 0  
## 4 hom 65 51 N . 0  
## 5 hom 68 51 N . 0  
## 6 hom 68 35 T C0.029 C1.0  
## x  
## 1 0.239221  
## 2 0.239245  
## 3 0.239272  
## 4 0.239323  
## 5 0.239374  
## 6 0.239407

dat.4 <- read.delim(gzfile(input.files[4]), sep="\t")  
dat.4$x = dat.4$position/1e6  
head(dat.4)

## chromosome position base.ref depth.normal depth.tumor depth.ratio Af Bf  
## 1 3 239223 N 26 5 0.166 1.000 0  
## 2 3 239245 C 25 4 0.160 1.000 0  
## 3 3 239274 N 41 13 0.309 1.000 0  
## 4 3 239285 T 35 12 0.343 0.917 0  
## 5 3 239308 C 49 16 0.327 0.938 0  
## 6 3 239325 N 79 40 0.521 1.000 0  
## zygosity.normal GC.percent good.reads AB.normal AB.tumor tumor.strand  
## 1 hom 70 49 N . 0  
## 2 het 70 4 CG . 0  
## 3 hom 68 51 N . 0  
## 4 hom 68 12 T A0.083 A1.0  
## 5 hom 68 16 C T0.063 T1.0  
## 6 hom 65 51 N . 0  
## x  
## 1 0.239223  
## 2 0.239245  
## 3 0.239274  
## 4 0.239285  
## 5 0.239308  
## 6 0.239325

### Summary statistics of read depth for normal and tumor samples

data1

summary(dat.1$depth.normal)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 37.00 76.00 87.86 126.00 510.00

summary(dat.1$depth.tumor)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 37.0 85.0 106.2 152.0 814.0

data2

summary(dat.2$depth.normal)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 35.00 74.00 85.77 124.00 503.00

summary(dat.2$depth.tumor)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 37.00 80.00 95.23 136.00 830.00

data3

summary(dat.3$depth.normal)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.0 39.0 76.0 87.8 123.0 506.0

summary(dat.3$depth.tumor)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 16.00 43.00 64.18 91.00 589.00

data4

summary(dat.4$depth.normal)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.00 39.00 78.00 89.23 127.00 506.00

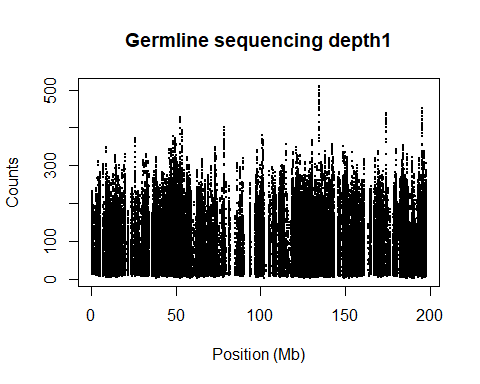
summary(dat.4$depth.tumor)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 36.0 90.0 117.3 170.0 803.0

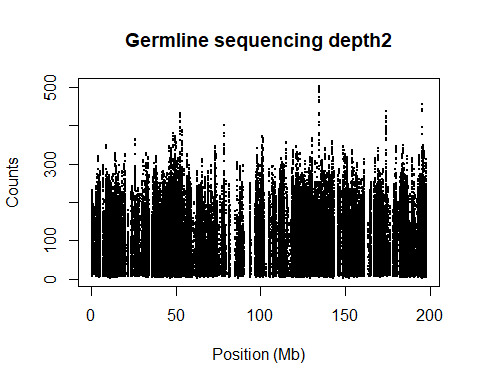
### Plot the read depth counts in normal samples:

data1

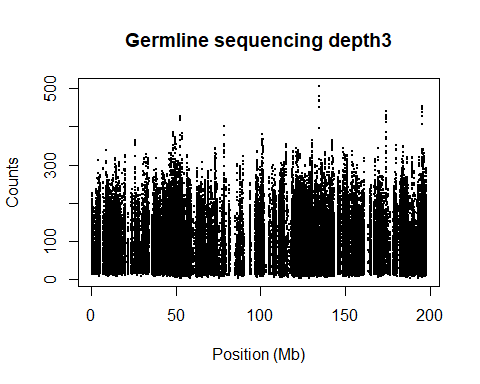
plot(depth.normal~x, dat.1, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Germline sequencing depth1")

 data2

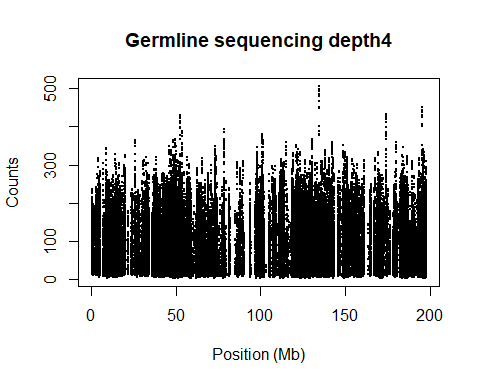
plot(depth.normal~x, dat.2, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Germline sequencing depth2")

 data3

plot(depth.normal~x, dat.3, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Germline sequencing depth3")

 data4

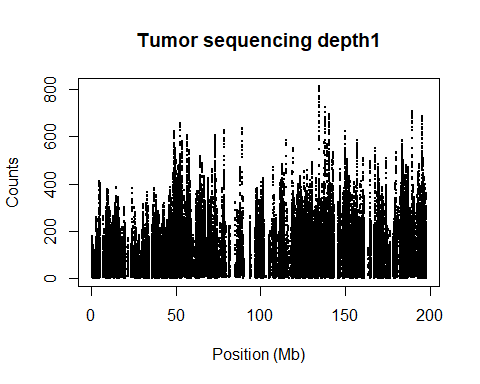
plot(depth.normal~x, dat.4, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Germline sequencing depth4")



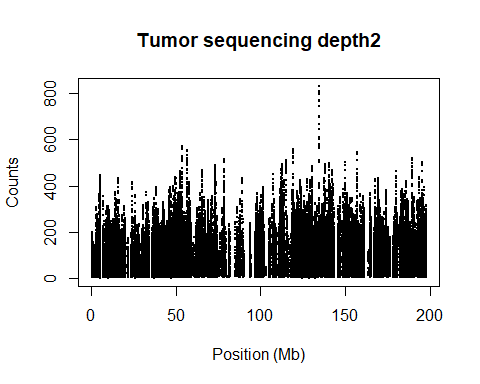
### Plot the read depth counts in tumour samples:

data1

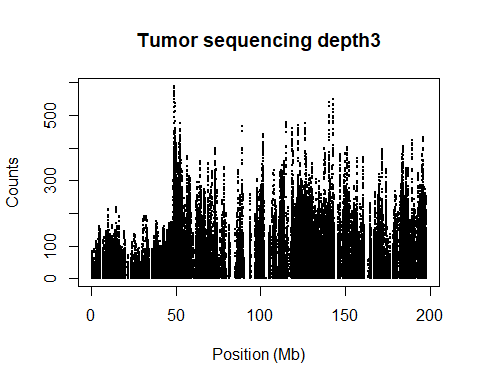
plot(depth.tumor~x, dat.1, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Tumor sequencing depth1")

 data2

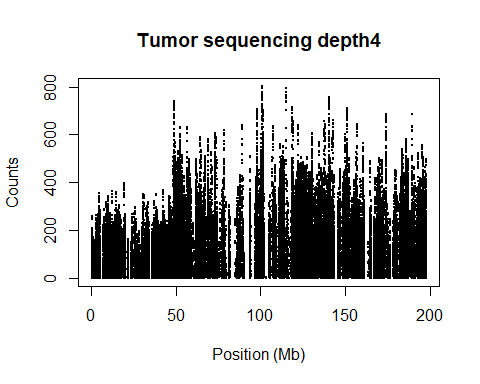
plot(depth.tumor~x, dat.2, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Tumor sequencing depth2")

 data3

plot(depth.tumor~x, dat.3, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Tumor sequencing depth3")

 data4

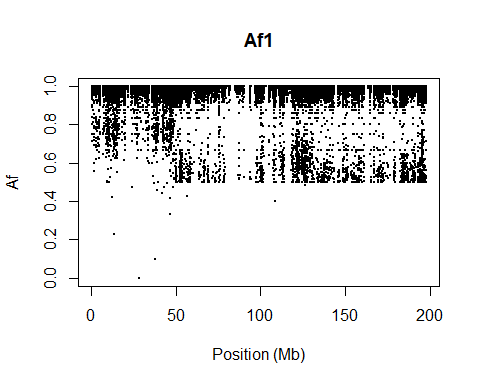
plot(depth.tumor~x, dat.4, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Counts"  
 , main="Tumor sequencing depth4")



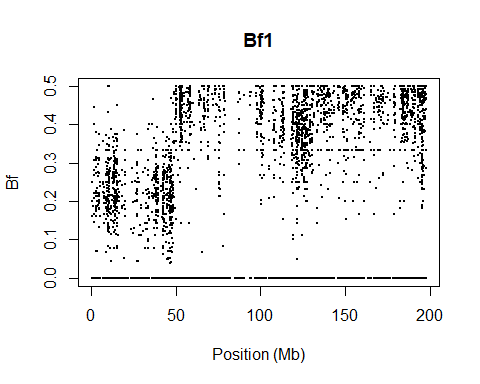
### The allele frequencies are captured in the raw data.

data1

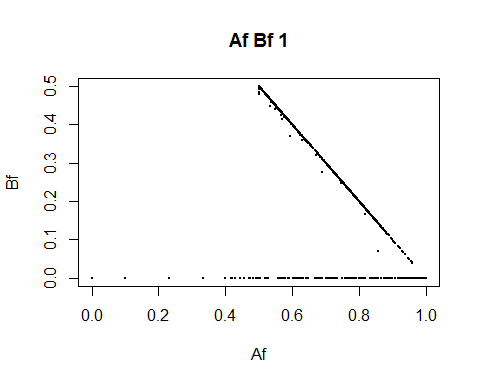
plot(Af~x, dat.1, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Af"  
 ,main="Af1")



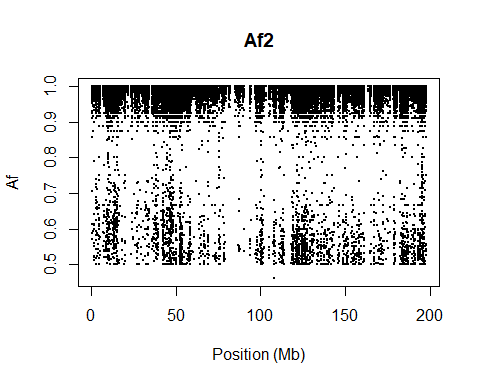
plot(Bf~x, dat.1, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Bf"  
 ,main="Bf1")



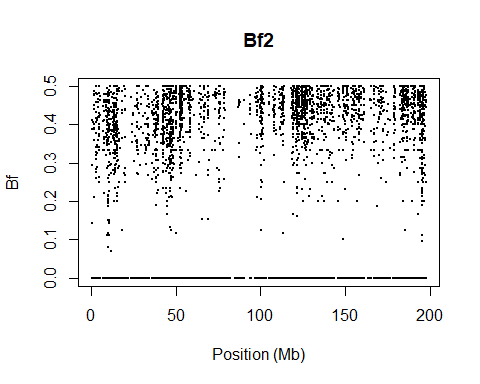
plot(Bf~Af, dat.1, pch=20, cex=0.3, xlab="Af", ylab="Bf", main="Af Bf 1")

 data2

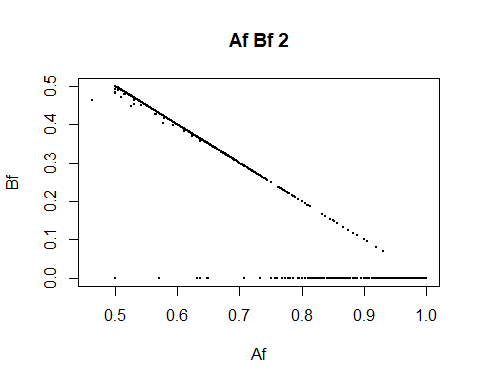
plot(Af~x, dat.2, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Af"  
 ,main="Af2")



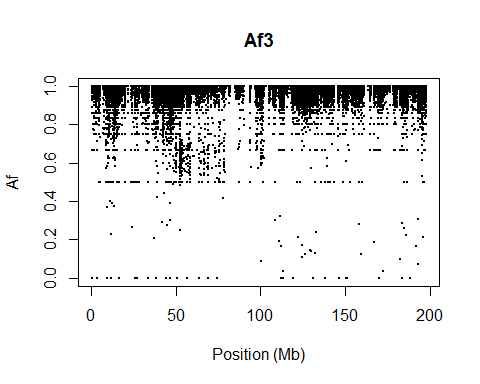
plot(Bf~x, dat.2, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Bf"  
 ,main="Bf2")



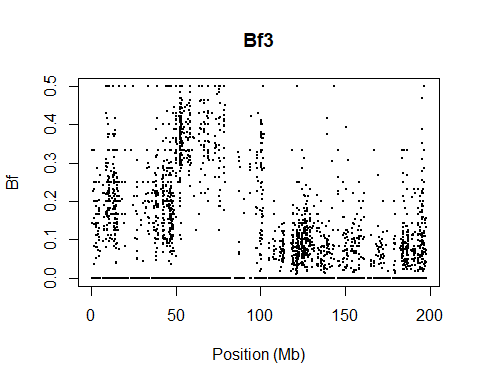
plot(Bf~Af, dat.2, pch=20, cex=0.3, xlab="Af", ylab="Bf", main="Af Bf 2")

 data3

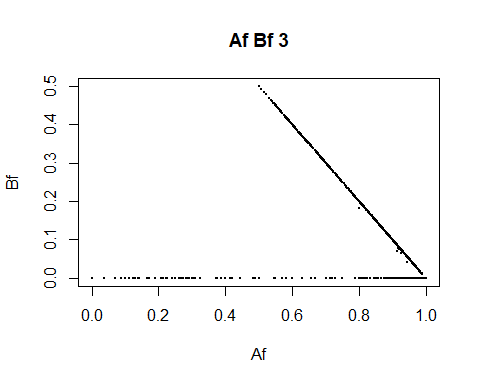
plot(Af~x, dat.3, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Af"  
 ,main="Af3")



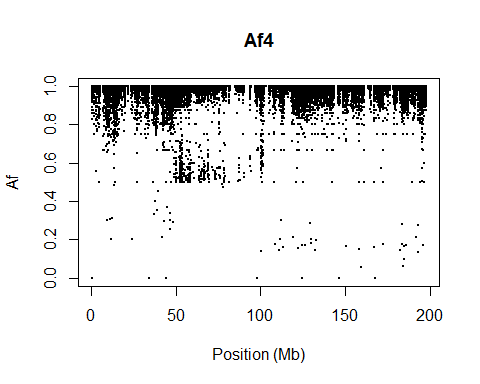
plot(Bf~x, dat.3, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Bf"  
 ,main="Bf3")



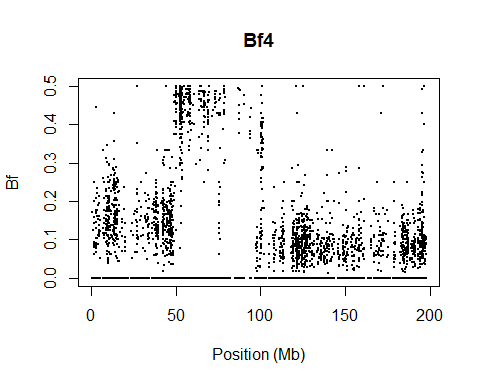
plot(Bf~Af, dat.3, pch=20, cex=0.3, xlab="Af", ylab="Bf", main="Af Bf 3")

 data4

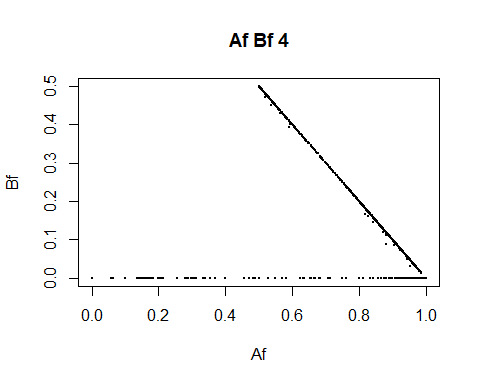
plot(Af~x, dat.4, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Af"  
 ,main="Af4")



plot(Bf~x, dat.4, pch=20, cex=0.3, xlab="Position (Mb)", ylab="Bf"  
 ,main="Bf4")



plot(Bf~Af, dat.4, pch=20, cex=0.3, xlab="Af", ylab="Bf", main="Af Bf 4")

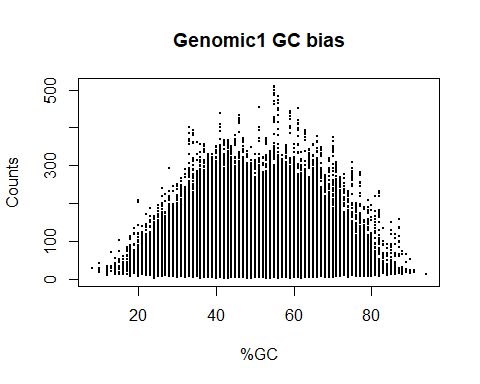


### Compare the sequencing depths with the genomic GC to plot the GC bias.

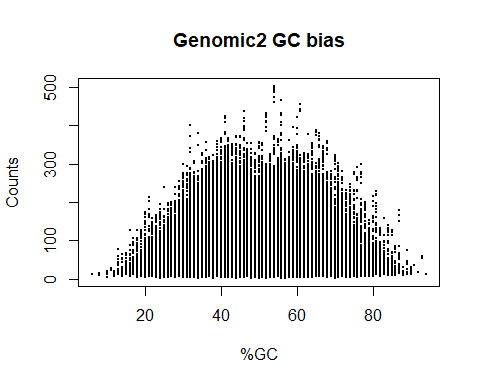
### Strong GC bias is presented.

data1

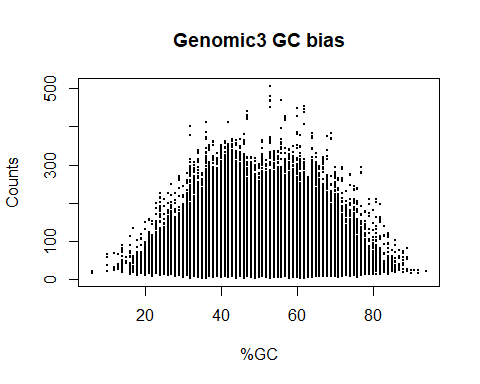
plot(depth.normal~GC.percent, dat.1, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Genomic1 GC bias")

 data2

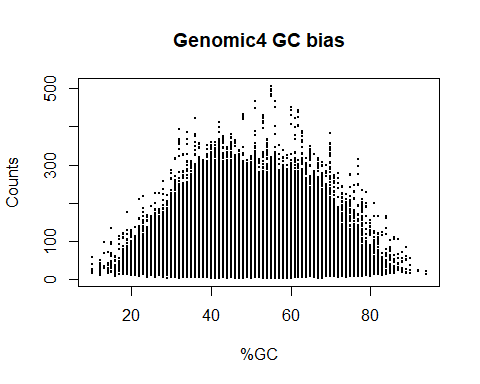
plot(depth.normal~GC.percent, dat.2, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Genomic2 GC bias")

 data3

plot(depth.normal~GC.percent, dat.3, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Genomic3 GC bias")

 data4

plot(depth.normal~GC.percent, dat.4, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Genomic4 GC bias")

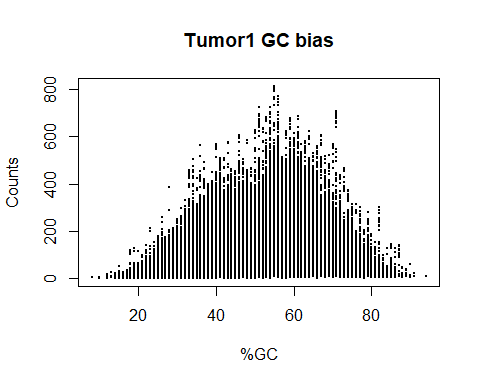


### Compare the sequencing depths with the tumor GC to plot the GC bias.

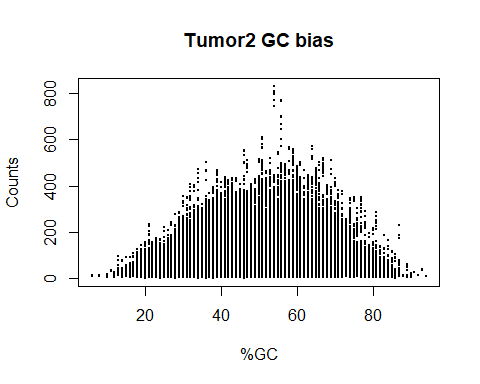
### Strong GC bias is presented as well.

data1

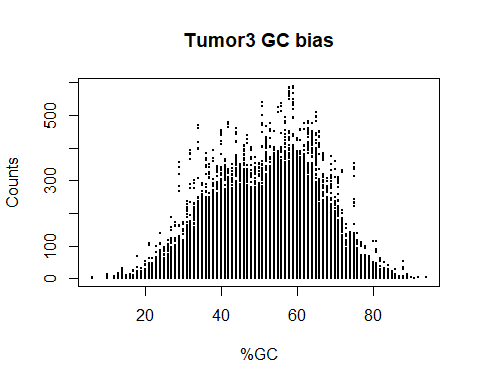
plot(depth.tumor~GC.percent, dat.1, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Tumor1 GC bias")

 data2

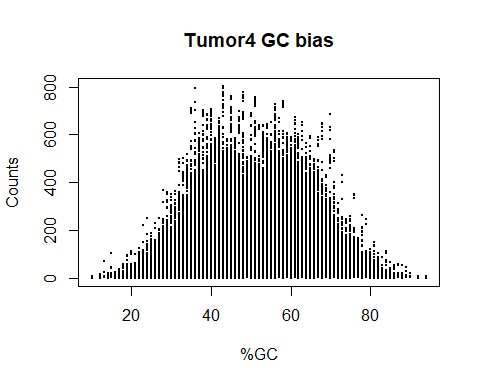
plot(depth.tumor~GC.percent, dat.2, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Tumor2 GC bias")

 data3

plot(depth.tumor~GC.percent, dat.3, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Tumor3 GC bias")

 data4

plot(depth.tumor~GC.percent, dat.4, pch=20, cex=0.3, xlab="%GC", ylab="Counts"  
 , main="Tumor4 GC bias")

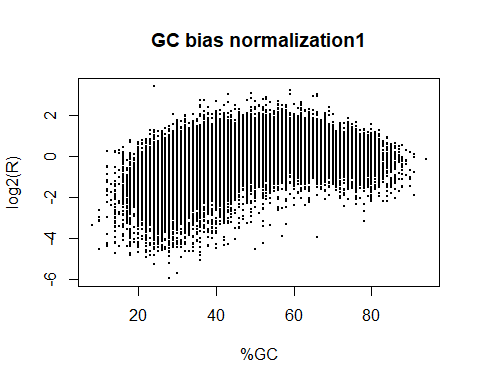


### Plot depth ratio vs GC-content values. GC bias still exists

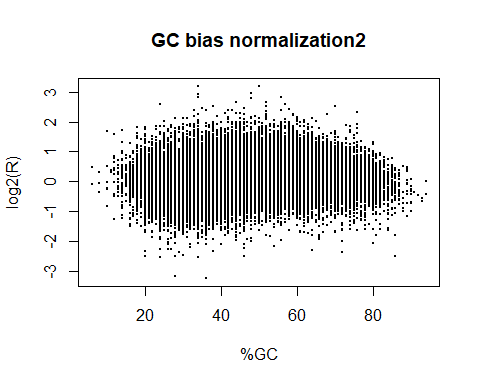
### after normalizaition of simply division and logarithmization

data1

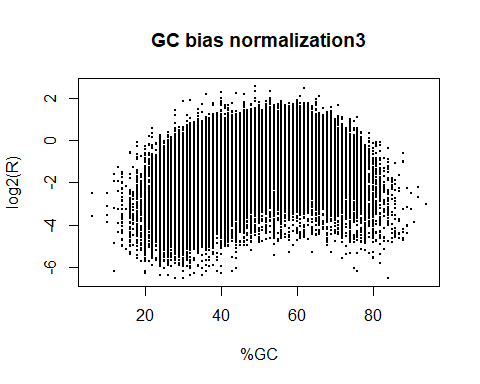
plot(log2(depth.ratio)~GC.percent, dat.1, pch=20, cex=0.3, xlab="%GC",   
 ylab="log2(R)", main="GC bias normalization1")

 data2

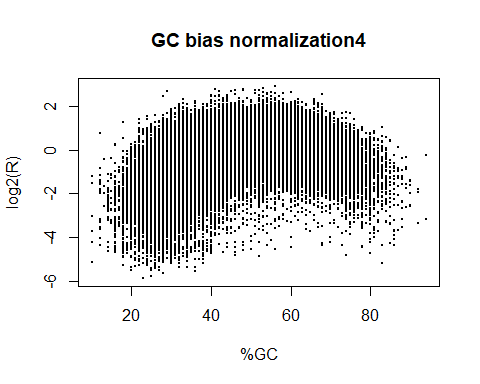
plot(log2(depth.ratio)~GC.percent, dat.2, pch=20, cex=0.3, xlab="%GC",   
 ylab="log2(R)", main="GC bias normalization2")

 data3

plot(log2(depth.ratio)~GC.percent, dat.3, pch=20, cex=0.3, xlab="%GC",   
 ylab="log2(R)", main="GC bias normalization3")

 data4

plot(log2(depth.ratio)~GC.percent, dat.4, pch=20, cex=0.3, xlab="%GC",   
 ylab="log2(R)", main="GC bias normalization4")

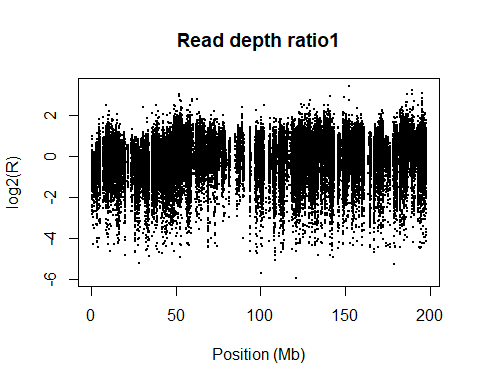


### Plot read depth ratio after divison.

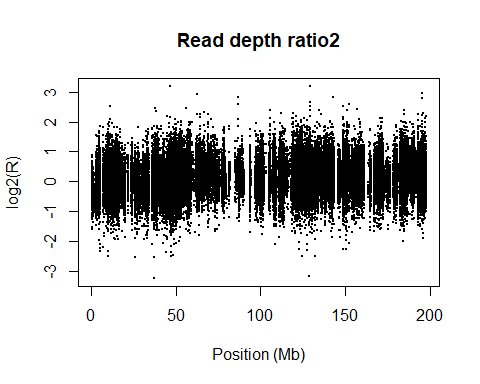
### It appears clearer.

data1

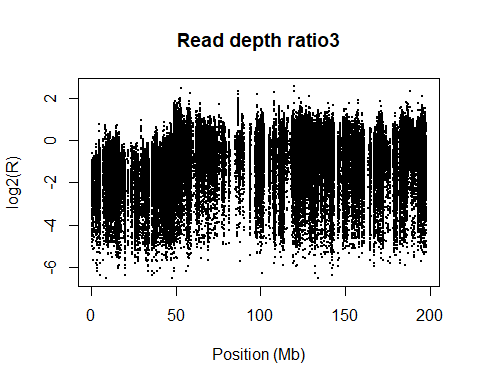
plot(log2(depth.ratio)~x, dat.1, pch=20, cex=0.3, xlab="Position (Mb)",  
 ylab="log2(R)", main="Read depth ratio1")

 data2

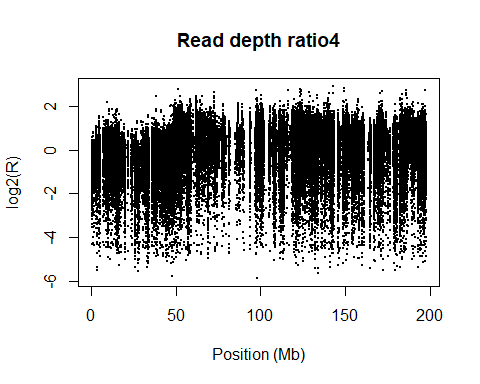
plot(log2(depth.ratio)~x, dat.2, pch=20, cex=0.3, xlab="Position (Mb)",   
 ylab="log2(R)", main="Read depth ratio2")

 data3

plot(log2(depth.ratio)~x, dat.3, pch=20, cex=0.3, xlab="Position (Mb)",   
 ylab="log2(R)", main="Read depth ratio3")

 data4

plot(log2(depth.ratio)~x, dat.4, pch=20, cex=0.3, xlab="Position (Mb)",   
 ylab="log2(R)", main="Read depth ratio4")



## Finally Run Sequenza

### Output files can be found in ‘output’ directory.

for(i in input.files) {  
 print(paste("Processing input file ", i, sep=""))  
 sample.name <- str\_replace(i, "\_bin50.seqz.gz", "")  
 sample.name <- str\_replace(sample.name, "./data/", "")  
   
 seq.ext <- sequenza.extract(i, normalization.method="median")  
 seq.fit <- sequenza.fit(sequenza.extract=seq.ext)  
 out.dir <- paste("./output/", sample.name, "\_OUTPUT", sep="")  
 sequenza.results(sequenza.extract=seq.ext, cp.table=seq.fit, sample.id=sample.name, out.dir=out.dir)  
}

## [1] "Processing input file ./data/CA0023-61-2.chr3\_bin50.seqz.gz"

## Collecting GC information . done  
##   
## Processing 3:  
## 57 variant calls.  
## 4 copy-number segments.  
## 2520 heterozygous positions.  
## 192026 homozygous positions.

## [1] "Processing input file ./data/CA0023-61-3.chr3\_bin50.seqz.gz"

## Collecting GC information . done  
##   
## Processing 3:  
## 44 variant calls.  
## 6 copy-number segments.  
## 2599 heterozygous positions.  
## 197274 homozygous positions.

## [1] "Processing input file ./data/CA0023-65.chr3\_bin50.seqz.gz"

## Collecting GC information . done  
##   
## Processing 3:  
## 25 variant calls.  
## 6 copy-number segments.  
## 2261 heterozygous positions.  
## 141981 homozygous positions.

## [1] "Processing input file ./data/CA0023-66.chr3\_bin50.seqz.gz"

## Collecting GC information . done  
##   
## Processing 3:  
## 42 variant calls.  
## 8 copy-number segments.  
## 2437 heterozygous positions.  
## 181031 homozygous positions.