Week 1 assignment of Bioconductor for Genomic Data Science

Yilin

February 21, 2016

Overview:Week 1 lecture covers IRange,GRange and AnnotationHub. After first week learning, I should be able to fetch data from AnnotationHub, find the overlapped region from two different data and calculate the enrichment of peaks A in data1 in peaks B in data2.

Q1: Use the AnnotationHub package to obtain data on "CpG Islands" in the human genome. Question: How many islands exists on the autosomes?

```
library(BiocInstaller)
```

```
## Bioconductor version 3.2 (BiocInstaller 1.20.1), ?biocLite for help
```

```
biocLite("GenomicRanges")
```

```
## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.2 (BiocInstaller 1.20.1), R 3.2.2 (2015-08-14).

## Installing package(s) 'GenomicRanges'

##
## The downloaded binary packages are in
##  /var/folders/pz/ypsrlcks60sfh0hkfwnqysb80000gn/T//Rtmp9W19vc/downloaded_packages

## Old packages: 'curl', 'devtools', 'digest', 'Hmisc', 'httr', 'kernlab',
##   'knitr', 'latticeExtra', 'limma', 'lme4', 'maps', 'memoise', 'mgcv',
##   'munsell', 'nlme', 'nnet', 'quantreg', 'R6', 'Rcpp', 'RcppEigen',
##   'rstudioapi', 'S4Vectors', 'tidyr'
```

```
biocLite("IRanges")
```

```
## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.2 (BiocInstaller 1.20.1), R 3.2.2 (2015-08-14).

## Installing package(s) 'IRanges'

##
## The downloaded binary packages are in
##  /var/folders/pz/ypsrlcks60sfh0hkfwnqysb80000gn/T//Rtmp9W19vc/downloaded_packages

## Old packages: 'curl', 'devtools', 'digest', 'Hmisc', 'httr', 'kernlab',
##   'knitr', 'latticeExtra', 'limma', 'lme4', 'maps', 'memoise', 'mgcv',
##   'munsell', 'nlme', 'nnet', 'quantreg', 'R6', 'Rcpp', 'RcppEigen',
##   'rstudioapi', 'S4Vectors', 'tidyr'
```

```
biocLite("AnnotationHub")
```

```
## BioC_mirror: https://bioconductor.org
```

```
## Using Bioconductor 3.2 (BiocInstaller 1.20.1), R 3.2.2 (2015-08-14).
```

```
## Installing package(s) 'AnnotationHub'
```

```
##
## The downloaded binary packages are in
##  /var/folders/pz/ypsrlcks60sfh0hkfwnqysb80000gn/T//Rtmp9W19vc/downloaded_packages
```

```
## Old packages: 'curl', 'devtools', 'digest', 'Hmisc', 'httr', 'kernlab',
##   'knitr', 'latticeExtra', 'limma', 'lme4', 'maps', 'memoise', 'mgcv',
##   'munsell', 'nlme', 'nnet', 'quantreg', 'R6', 'Rcpp', 'RcppEigen',
##   'rstudioapi', 'S4Vectors', 'tidyr'
```

```
library(GenomicRanges)
```

```
## Warning: package 'GenomicRanges' was built under R version 3.2.3
```

```
## 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, xtabs
```

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

```
## Loading required package: S4Vectors
```

```
## Warning: package 'S4Vectors' was built under R version 3.2.3
```

```
## Loading required package: stats4
```

```
## Loading required package: IRanges
```

```
## Warning: package 'IRanges' was built under R version 3.2.3
```

```
## Loading required package: GenomeInfoDb
```

```
## Warning: package 'GenomeInfoDb' was built under R version 3.2.3
```

```
library(AnnotationHub)
```

```
## Warning: package 'AnnotationHub' was built under R version 3.2.3
```

```
ah=AnnotationHub()
```

```
## snapshotDate(): 2016-01-25
```

```
ah1=subset(ah,species=="Homo sapiens")
qse=query(ah1,"CpG Islands")
grcpg=qse[[1]]
grcpg_autosome=GRanges()
for (i in 1:22){
  grs=subset(grcpg,seqnames(grcpg)==paste0("chr",i))
  grcpg_autosome=append(grcpg_autosome,grs)
}
length(grcpg_autosome)
```

```
## [1] 26641
```

Q2:How many CpG Islands exists on chromosome 4

```
grcpg4=subset(grcpg,seqnames(grcpg)=="chr4")
length(grcpg4)
```

```
## [1] 1031
```

Q3:Obtain the data for the H3K4me3 histone modification for the H1 cell line from Epigenomics Roadmap, using AnnotationHub. Subset these regions to only keep regions mapped to the autosomes (chromosomes 1 to 22) Question: How many bases does these regions cover?

```
library(rtracklayer)
```

```
## Warning: package 'rtracklayer' was built under R version 3.2.3
```

```
qse1=query(ah,c("H3K4me3","H1 cell"))
grH3K4=qse1[[2]]
grH3K4_autosome=GRanges()
for (i in 1:22){
  grs=subset(grH3K4,seqnames(grH3K4)==paste0("chr",i))
  grH3K4_autosome=append(grH3K4_autosome,grs)
}
sum(width(grH3K4_autosome))
```

```
## [1] 41135164
```

Q4:Obtain the data for the H3K27me3 histone modification for the H1 cell line from Epigenomics Roadmap, using the AnnotationHub package. Subset these regions to only keep regions mapped to the autosomes. In the return data, each region has an associated "signalValue" Question: What is the mean signalValue across all regions on the standard chromosomes?

```
qse2=query(ah,c("H3K27me3","H1 cell"))
grH3K27=qse2[[2]]
grH3K27_autosome=GRanges()
for (i in 1:22){
  grs=subset(grH3K27,seqnames(grH3K27)==paste0("chr",i))
  grH3K27_autosome=append(grH3K27_autosome,grs)
}
mean(grH3K27_autosome$signalValue,na.rm=TRUE)
```

```
## [1] 4.770728
```

Q5:Bivalent regions are bound by both H3K4me3 and H3K27me3 Question: Using the regions we have obtained above, how many bases on the standard chromosomes are bivalently marked?

```
sum(width(intersect(grH3K4_autosome,grH3K27_autosome)))
```

```
## [1] 10289096
```

Q6:We will examine the extent to which bivalent regions overlap CpG Islands Question: how big a fraction (expressed as a number between 0 and 1) of the bivalent regions, overlap one or more CpG Islands?

```
grH3K_overlay=intersect(grH3K4_autosome,grH3K27_autosome)
ov=findOverlaps(grH3K_overlay,grcpg_autosome)
length(unique(queryHits(ov)))/length(grH3K_overlay)
```

```
## [1] 0.5383644
```

Q7:How big a fraction (expressed as a number between 0 and 1) of the bases which are part of CpG Islands, are also bivalent marked

```
grcpg_H3K=intersect(grcpg_autosome,grH3K_overlay)
sum(width((grcpg_H3K)))/sum(width((grcpg_autosome)))
```

```
## [1] 0.241688
```

Q8:How many bases are bivalently marked within 10kb of CpG Islands? Tip: consider using the "resize()"" function

```
grcpg_r=resize(grcpg_autosome,width=20000+width(grcpg_autosome),fix="center")
sum(width(intersect(grH3K_overlay,grcpg_r)))
```

```
## [1] 9782086
```

Q9:Question: How big a fraction (expressed as a number between 0 and 1) of the human genome is contained in a CpG Island? Tip 1: the object returned by AnnotationHub contains "seqlengths" Tip 2: you may encounter an integer overflow. As described in the session on R Basic Types, you can address this by converting integers to numeric before summing them, "as.numeric()"

```
seq4=query(ah,"RefSeq")
gr_genome=seq4[[1]]
contain=sum(width(grcpg_autosome))
total=sum(as.numeric(seqlengths(gr_genome)[1:22]))
ratio=contain/total
```

Q10:Compute an odds-ratio for the overlap of bivalent marks with CpG islands

```
table_inout=matrix(rep(0,4),nrow=2,ncol=2)
row.names(table_inout)=c("bivalent_in","bivalent_out")
colnames(table_inout)=c("cpg_in","cpg_out")
table_inout[1,1]=sum(width((grcpg_H3K)))
table_inout[1,2]=sum(width((grcpg_autosome)))-sum(width((grcpg_H3K)))
table_inout[2,1]=sum(width(grH3K_overlay))-sum(width((grcpg_H3K)))
table_inout[2,2]=sum(as.numeric(seqlengths(gr_genome)[1:22]))-sum(width((grcpg_autosome)))-sum(width(gr
odds_ratio=(table_inout[1,1]*table_inout[2,2])/(table_inout[1,2]*table_inout[2,1])
```