Moffitt Cancer Center
Biostatistics & Bioinformatics

Morsani College of Medicine
Oncologic Sciences

Practice: ChIP-seq data analysis

USF Master Program in Bioinformatics

Mingxiang Teng

March 23, 2023

Steps

- **Setting up R/Rstudio**
- **Data preparation**
- **Exploratory data analysis (QC)**
- **Peak detection**

Install R

r-project.org

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Get Involved: Contributing

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 4.3.0 (Already Tomorrow) prerelease versions** will appear starting Tuesday 2023-03-21. Final release is scheduled for Friday 2023-04-21.
- **R version 4.2.3 (Shortstop Beagle)** has been released on 2023-03-15.
- **R version 4.1.3 (One Push-Up)** was released on 2022-03-10.

Install Rstudio

posit.co/downloads/

posit

PRODUCTS SOLUTIONS LEARN & SUPPORT EXPLORE MORE PRICING

DOWNLOAD RSTUDIO

RStudio Desktop

Find out more about RStudio Desktop and RStudio Desktop Pro below.

DOWNLOAD RSTUDIO

Biostatistics & Bioinformatics

Download course material

github.com/tengmx/gms7930/blob/master/gms7930-rgs.Rmd

Product Solutions Open Source Pricing Search / Sign in Sign up

tengmx / gms7930 Public

Notifications Fork 1 Star 1

Code Issues Pull requests Actions Projects Security Insights

master gms7930 / gms7930-rgs.Rmd Go to file ···

tengmx 2023 Spring Latest commit 92f7acd yesterday History

1 contributor

271 lines (231 sloc) 13.8 KB

Raw

Open Link in New Tab
Open Link in New Window
Open Link in Incognito Window
Save Link As...
Copy Link Address
Inspect

```
---
title: "ChIP-seq Data Analysis"
author: "Mingxiang Teng"
date: "3/23/2023"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE,warning=FALSE, message=FALSE)
```

```

Biostatistics & Bioinformatics

# Open .Rmd file in Rstudio

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

gms7930-rgs.Rmd*

Packages BiocManager, knitr, and RColorBrewer required but are not installe. Install | Don't Show Again

```
---
author: "Mingxiang Teng"
date: "1/26/2021"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE,warning=FALSE, message=FALSE)
```

Install R packages

In order to perform ChIP-seq data analysis with R, we have to first make sure necessary R packages
installed before any processing. Bioconductor <http://www.bioconductor.org> is a large community
which documents thousands of R packages dedicated to biomedical data analysis.
We will pick several of those packages from Bioconductor for ChIP-seq data analysis. Note that
some widely-used tools for ChIP-seq data analysis were not developed under R environment.
Thus, R packages chosen here are for demostration purpose, and might not be well adopted by community users.

```{r install, echo=TRUE, results='hide',message=FALSE}
## install necessary Bioconductor packages
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install(version = "3.10")
BiocManager::install("knitr") ## dealing with rmarkdown
BiocManager::install("GenomicRanges") ## dealing with genomic intervals
BiocManager::install("normr") ## dealing with peak calling
BiocManager::install("ChIPseeker") ## dealing with peak annotation
BiocManager::install("TxDb.Hsapiens.UCSC.hg19.knownGene") ## human genome and transcriptome
BiocManager::install("org.Hs.eg.db") ## gene name ontology
BiocManager::install("annotate") ## genome annotation
```

4:12 | ChIP-seq Data Analysis | R Markdown

Console | Terminal | Jobs

```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Environment | History | C

Import Dataset

R | Global Environment

Files | Plots | Packages

New Folder | Delete

Home

Name

desktop.ini

My Music

My Pictures

My Videos

R
```

github.com/tengmx/gms7930/blob/master/gms7930-rgs.Rmd

```
---
title: "ChIP-seq Data Analysis"
author: "Mingxiang Teng"
date: "3/23/2023"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE,warning=FALSE, message=FALSE)
```

Install R packages

In order to perform ChIP-seq data analysis with R, we have to first make sure necessary R packages
installed before any processing. Bioconductor <http://www.bioconductor.org> is a large community
which documents thousands of R packages dedicated to biomedical data analysis.
We will pick several of those packages from Bioconductor for ChIP-seq data analysis. Note that
some widely-used tools for ChIP-seq data analysis were not developed under R environment.
Thus, R packages chosen here are for demonstration purpose, and might not be well adopted by community users.

```{r install, echo=TRUE, results='hide',message=FALSE}
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org" #https://cran.rstudio.com/
options(repos = r)
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install()
install.packages('rmarkdown')
BiocManager::install("knitr") ## dealing with rmarkdown
BiocManager::install("GenomicRanges") ## dealing with genomic intervals
BiocManager::install("normr") ## dealing with peak calling
BiocManager::install("ChIPseeker") ## dealing with peak annotation
BiocManager::install("TxDb.Hsapiens.UCSC.hg38.knownGene") ## human genome and transcriptome
BiocManager::install("GenomeInfoDb")
BiocManager::install("BSgenome.Hsapiens.UCSC.hg38")
BiocManager::install("org.Hs.eg.db") ## gene name ontology
BiocManager::install("annotate") ## genome annotation
BiocManager::install("Rsamtools") ## dealing with BAM files
BiocManager::install("GenomicAlignments") ## dealing with sequencing reads
BiocManager::install("rtracklayer") ## genomic sequencing tracks
BiocManager::install("RColorBrewer") ## color code
```

After all required packages installed, we need to load these packages.

```{r load, echo=TRUE, results='hide',message=FALSE}
library(Rsamtools)
library(GenomicAlignments)
library(GenomicRanges)
library(normr)
library(TxDb.Hsapiens.UCSC.hg38.knownGene)
library(BSgenome.Hsapiens.UCSC.hg38)
library(org.Hs.eg.db)
library(ChIPseeker)
library(annotate)
library(rtracklayer)
library(RColorBrewer)
```

Non-code text

Install R packages

Load R packages

Biostatistic
Bioinforma
```

How to run R code in .Rmd file

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

gms7930-rgs.Rmd

Knit Run

```
under R environment.
Thus, R packages chosen h                                  and might
not be well adopted by co
```{r install, echo=TRUE,
## install necessary Bioc
if (!requireNamespace("Bi
install.packages("BiocMan
BiocManager::install()
BiocManager::install("kni
BiocManager::install("Gen                                  omic
intervals
BiocManager::install("nor
BiocManager::install("ChI                                  nnotation
BiocManager::install("TxD                                  ## human
genome and transcriptome
BiocManager::install("org
BiocManager::install("ann
BiocManager::install("Rsa                                  s
BiocManager::install("GenomicAlignments") ## dealing with sequencing
reads
BiocManager::install("rtracklayer") ## genomic sequencing tracks
BiocManager::install("RColorBrewer") ## color code
```
```

Run Selected Line(s) Ctrl+Ente
Run Current Chunk Ctrl+Shift+Ente
Run Next Chunk
Run the current code chunk
Run Setup Chunk
Run Setup Chunk Automatically
Run All Chunks Above Ctrl+Alt+Shift+P
Run All Chunks Below
Restart R and Run All Chunks
Restart R and Clear Output
Run All Ctrl+Alt+R

28:1 Chunk 2: install R Markdown

Console Terminal Jobs

Setting up R analysis environment

```r
## install necessary Bioconductor packages
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install(version = "3.10")
BiocManager::install("knitr") ## dealing with rmarkdown
BiocManager::install("GenomicRanges") ## dealing with genomic intervals
BiocManager::install("normr") ## dealing with peak calling
BiocManager::install("ChIPseeker") ## dealing with peak annotation
BiocManager::install("TxDb.Hsapiens.UCSC.hg19.knownGene") ## human genome and transcriptome
BiocManager::install("org.Hs.eg.db") ## gene name ontology
BiocManager::install("annotate") ## genome annotation
BiocManager::install("Rsamtools") ## dealing with BAM files
BiocManager::install("GenomicAlignments") ## dealing with sequencing reads
BiocManager::install("rtracklayer") ## genomic sequencing tracks
BiocManager::install("RColorBrewer") ## color code
```

Make sure to install these R packages before data analysis

Questions you need to answer if installing packages one by one

```
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
  converting help for package 'TxDb.Hsapiens.UCSC.hg19.knownGene'
    finding HTML links ... done
    package                                 html
** building package indices
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation pat
h
* DONE (TxDb.Hsapiens.UCSC.hg19.knownGene)

The downloaded source packages are in
        'C:\Users\tengm\AppData\Local\Temp\Rtmpw5Oztw\downloaded_packages'
Update all/some/none? [a/s/n]:
n
```

Data to be analyzed

ChIP: Gm12878Ctcf.bam
INPUT: Gm12878Control.bam

```
workdir = '~/' ##
download.file('https://github.com/tengmx/gms7930/raw/master/data/Gm12878Control.bam',paste0(workdir,'Gm12878Control.bam'))
download.file('https://github.com/tengmx/gms7930/raw/master/data/Gm12878Control.bam.bai',paste0(workdir,'Gm12878Control.bam.bai'))
download.file('https://github.com/tengmx/gms7930/raw/master/data/Gm12878Ctcf.bam',paste0(workdir,'Gm12878Ctcf.bam'))
download.file('https://github.com/tengmx/gms7930/raw/master/data/Gm12878Ctcf.bam.bai',paste0(workdir,'Gm12878Ctcf.bam.bai'))
```

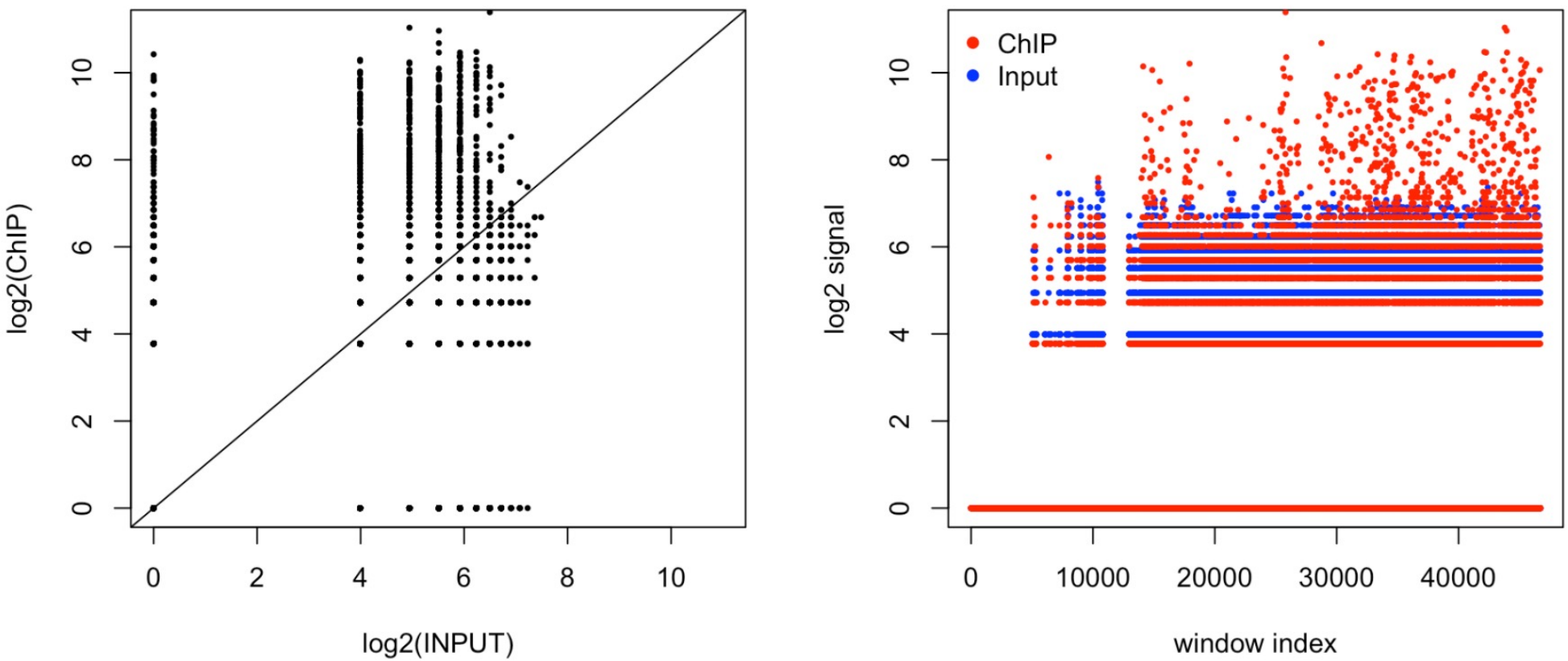
Data table generation

```{r bam, echo=TRUE}
## sequencing files that only contain reads from chromosome 21
chip = file.path(workdir,'Gm12878Ctcf.bam')
input = file.path(workdir,'Gm12878Control.bam')
## chromosome length info from human genome build hg19
hg_chrs = getBSgenome("hg38")
seqlen_chr21 = seqlengths(hg_chrs)['chr21']
## create genomic 1000bp windows and store them using GenomicRanges
window_gr = unlist(tileGenome(seqlen_chr21, tilewidth=1000))
## count reads from both files for all windows
rc = summarizeOverlaps(window_gr, c(chip, input))
rc = assays(rc)[[1]]
## simple normalization based on library size
cpm_chr21 = t(t(rc)*(1000000/colSums(rc)))
head(cpm_chr21)
```

```
     Gm12878Ctcf.bam Gm12878Control.bam
[1,]               0                  0
[2,]               0                  0
[3,]               0                  0
[4,]               0                  0
[5,]               0                  0
[6,]               0                  0
```

Exploratory comparing ChIP to Input

```
par(mfrow=c(1,2))
## read counts in all windows
plot(log2(cpm_chr21[,2]+1),log2(cpm_chr21[,1]+1),pch=16,cex=0.5,xlab='log2(INPUT)',ylab='log2(ChIP)',ylim=c(0,11),xlim=c(0,11))
abline(a=0,b=1)
## read counts along the chromosome
plot(log2(cpm_chr21[,2]+1),col='blue',pch=16,cex=0.5,xlab='window index',ylab='log2 signal',ylim=c(0,11))
points(log2(cpm_chr21[,1]+1),col='red',pch=16,cex=0.5)
legend('topleft',c('ChIP','Input'),pch=16,col=c('red','blue'),bty='n')
```



Peak detection at bin level

```{r range, echo=TRUE,message=FALSE}
## formalizing peak info with GRanges container
peaks = getRanges(peakfit)
peaks$sig = getEnrichment(peakfit) ## add enrichmend signal
peaks$pvalue = getPvalues(peakfit) ## add enrichment significance
peaks
```

```
GRanges object with 12353090 ranges and 3 metadata columns:
             seqnames            ranges strand | component          sig    pvalue
                <Rle>         <IRanges>  <Rle> | <integer>    <numeric> <numeric>
         [1]     chr1             1-250      * |      <NA> -8.27017e-18         1
         [2]     chr1           251-500      * |      <NA> -8.27017e-18         1
         [3]     chr1           501-750      * |      <NA> -8.27017e-18         1
         [4]     chr1          751-1000      * |      <NA> -8.27017e-18         1
         [5]     chr1         1001-1250      * |      <NA> -8.27017e-18         1
         ...      ...               ...    ... .       ...          ...       ...
  [12353086]     chrY 57226251-57226500      * |      <NA> -8.27017e-18         1
  [12353087]     chrY 57226501-57226750      * |      <NA> -8.27017e-18         1
  [12353088]     chrY 57226751-57227000      * |      <NA> -8.27017e-18         1
  [12353089]     chrY 57227001-57227250      * |      <NA> -8.27017e-18         1
  [12353090]     chrY 57227251-57227415      * |      <NA> -8.27017e-18         1
  -------
  seqinfo: 24 sequences from an unspecified genome
```

Select significant peaks

```{r peaksig, echo=TRUE, message=TRUE}
## only consider a peak as significantly enriched if its q-value is less than 0.01
peakssig = peaks[which(peaks$pvalue<0.01)]
peakssig = peakssig[order(peakssig$pvalue)]
peakssig
```

```
GRanges object with 706 ranges and 3 metadata columns:
        seqnames            ranges strand | component       sig      pvalue
           <Rle>         <IRanges>  <Rle> | <integer> <numeric>   <numeric>
     [1]    chr21 25801501-25801750      * |         1   1.24785 9.73874e-35
     [2]    chr21 25801751-25802000      * |         1   1.18900 3.17057e-28
     [3]    chr21 43789251-43789500      * |         1   1.17497 7.37378e-27
     [4]    chr21 43939251-43939500      * |         1   2.86317 9.60544e-27
     [5]    chr21 43975501-43975750      * |         1   2.85912 2.15827e-26
     ...      ...                ...    ... .       ...       ...          ...
   [702]    chr21 37582501-37582750      * |         1  0.683697  0.00837267
   [703]    chr21 42310001-42310250      * |         1  0.683697  0.00837267
   [704]    chr21 44298001-44298250      * |         1  0.683697  0.00837267
   [705]    chr21 45308751-45309000      * |         1  0.683697  0.00837267
   [706]    chr21 45477751-45478000      * |         1  0.683697  0.00837267
   -------
   seqinfo: 24 sequences from an unspecified genome
```