Normalization of ChIP-seq experiments using Background Read Density

Benjamin Leblanc

28.02.2018

Contents

1 Introduction

`Tightrope` is an `R` package proposing a ChIP-seq normalization method, named Background Read Density (BRD)1,2, capable of accurate estimation of normalization factors in conditions involving global variations of chromatin marks.

In these experimental conditions it is known that basic sequencing depth correction can be dramatically misleading, and the fact that such experiments are important for ongoing research on chromatin regulation mechanisms and cancer has increased the need for validated normalization methods in the field of chromatin genomics.

In the recent years, spike-in protocols3-6 were introduced to circumvent ChIP-seq normalization issues with an accessible experimental strategy. Seeking an alternative strategy due to reliability concerns with the spike-in method, we designed the BRD method with a naive and empirical computational approach while we were studying H3K27me3 profiles in a cellular model of DIPG1.

After the initial validation and successful application of the BRD normalization with H3K27me3, we redesigned the underlying algorithm to achieve accurate normalizations with various experimental setups and chromatin marks. Following these improvements, we are currently preparing the manuscript describing the BRD method and presenting comparative analyses between spike-in and BRD normalizations for different datasets2.

The purpose of this document is to provide a practical guide on the use of `R` functions available in `Tightrope` to apply the BRD normalization method. A basic knowledge of the `R` syntax and data structures, and of Bioconductor objects used to represent and manipulate mapped reads and genomic intervals will help to understand and adapt the following source code examples to another ChIP-seq dataset.

1.1 Overview

1.2 Installations

The following softwares must be installed before `Tightrope` can be used.

- R environment version 3.4 or newer.
- To develop and execute `R` scripts we recommend using RStudio.

- CRAN packages `devtools`, `stringr`, `triangle`, `caTools`, `ica`, `FNN`, `igraph`, `mixtools`.
- Bioconductor packages `Rsamtools`, `GenomicAlignments`, `GenomicRanges`, `GenomicFeatures`, `rtracklayer`, `biomaRt`.
- GitHub package Barbouille.

The next subsections provide further indications for these installations.

1.2.1 R and RStudio environments

If `R` and RStudio are not already installed.

- Manually download and install R.
- Manually download and install RStudio.

1.2.2 Required packages

Once `R` and RStudio are installed, open RStudio and run the `R` code below which should install all required packages automatically. While `R` will perfom these tasks, you may be prompted for a confirmation of package installations or updates.

```
# Setting value below to TRUE will reinstall all required packages (optional)
reinstall <- FALSE

# Detect already installed packages
pkg <- ifelse(reinstall, c(), installed.packages()[, "Package"])

# CRAN packages
lst <- c("devtools", "stringr", "triangle", "caTools", "ica", "FNN", "igraph", "mixtools")
lst <- setdiff(lst, pkg)
if(length(lst) > 0) install.packages(lst, dependencies = T)

# Bioconductor packages
source("https://bioconductor.org/biocLite.R")
lst <- c("Rsamtools", "GenomicAlignments", "GenomicRanges", "GenomicFeatures", "rtracklayer", "biomaRt")
lst <- setdiff(lst, pkg)
if(length(lst) > 0) biocLite(lst)

# GitHub packages
library("devtools")
install_github("benja0x40/Barbouille", dependencies = T)
```

1.2.3 Tightrope

Install Tightrope using RStudio as follows:

- Open menu `Tools > Install Packages...`
- Select `Install from: Package Archive File (.tgz; .tar.gz)`
- Browse your computer to select the package archive file (e.g. `Tightrope_0.5.1.tar.gz`)
- Click `Install`

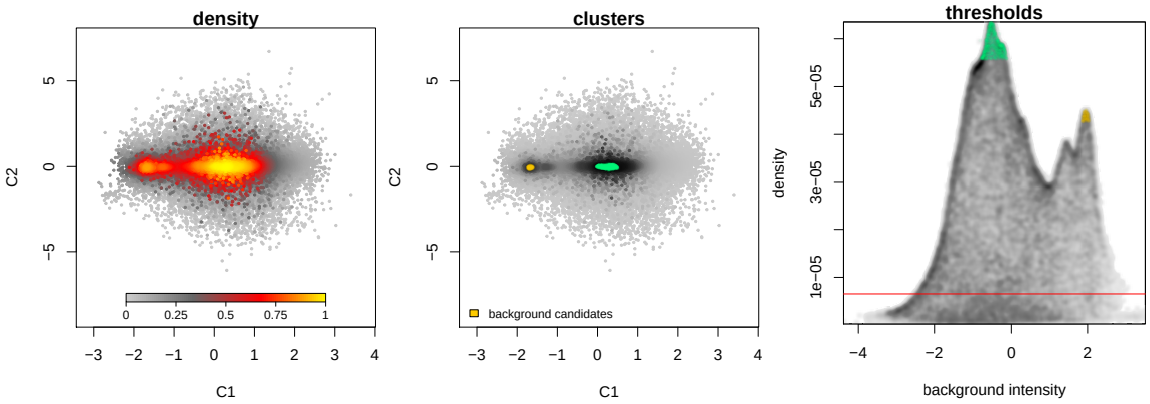
1.3 ChIP-seq experiments

```
library(Tightrope)

META   <- Orlando_METADATA
COUNTS <- Orlando_COUNTS

META   <- ESC_BRD_METADATA
COUNTS <- ESC_BRD_COUNTS
```

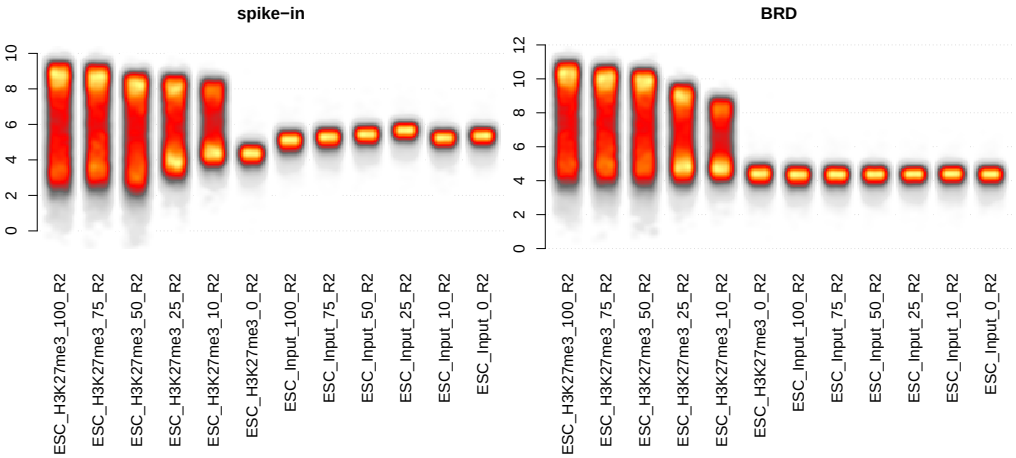
2 Raw read counts

```
CGI <- resize(CGI_hg38, width = 4000, fix = "center", use.names = F)
blacklist <- hg38_blacklist

CGI <- resize(CGI_mm10, width = 4000, fix = "center", use.names = F)
blacklist <- mm10_blacklist

chip <- META[antibody == "H3K79me2"]$sample_id
ctrl <- META[antibody == "Input"]$sample_id

chip <- META[antibody == "H3K27me3" & replicate == "R2"]$sample_id
ctrl <- META[antibody == "Input" & replicate == "R2"]$sample_id

ref <- COUNTS$GNU[, c(chip, ctrl)]
chk <- FiniteValues(log2(ref))
brd <- BRD(ref[chk, ], controls = ctrl, ncl = 2, bdt = c(0.2, 0.03))

chk <- countOverlaps(CGI, blacklist) == 0
cnt <- COUNTS$CGI_UCSC[chk, c(chip, ctrl)]
l2c <- log2(DitherCounts(cnt))

nrm <- t(t(l2c) + brd$normfactors)
```

```
# Adjust graphic options
par(
  pch = 20, mar = c(4.5, 4.5, 1, 1), cex = 1.5 # cex.main = 1.2, cex.lab = 1.2, cex.axis = 1.2
)

# Show the BRD results
PlotBRD(brd)
```




```
# Create a color mapping function to represent read count distributions
cmf <- function(k) colorize(k, mode = "01", col = "Wry")

# Adjust plot margins to show the name of experimental conditions
par(mar = c(12, 4.5, 3, 1), cex.main = 1.2, cex.lab = 1.1, cex.axis = 1.1)

# Show read count distributions after spike-in and BRD normalizations
r <- SideBySideDensity(
  l2c, nx = ncol(cnt) * 25, ny = 150, method = "ash", spacing = 0.4,
  mapper = cmf, las = 2, main = "spike-in"
)
r <- SideBySideDensity(
  nrm, nx = ncol(cnt) * 25, ny = 150, method = "ash", spacing = 0.4,
  mapper = cmf, las = 2, main = "BRD"
)
```



3 References

1. Mohammad et al. 2017 - *EZH2 is a potential therapeutic target for H3K27M-mutant pediatric gliomas.* publisher | pubmed

2. Leblanc B., Mohammad F., Hojfeldt J., Helin K. - *Normalization of ChIP-seq experiments involving global variations of chromatin marks.*
(in preparation)

3. Bonhoure et al. 2014 - *Quantifying ChIP-seq data: a spiking method providing an internal reference for*

sample-to-sample normalization
publisher | pubmed

4. Orlando et al. 2014 - *Quantitative ChIP-Seq normalization reveals global modulation of the epigenome.*
publisher | pubmed

5. Hu et al. 2015 - *Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq.*
publisher | pubmed

6. Egan et al. 2016 - *An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition.*
publisher | pubmed