DNaseR: DNase I footprinting analysis of DNase-seq data

Pedro Madrigal[1,2,*]

February, 2013

[1] Department of Biometry and Bioinformatics, Institute of Plant Genetics, Polish Academy of Sciences, Poznan, Poland
[2] Present address: Wellcome Trust Sanger Institute, Cambridge, United Kingdom

1 Introduction

The combination of DNase I digestion and high-throughput sequencing (DNase-seq) has been used recently to map chromatin accessibility in a given tissue or cell type on a genome-wide scale (Song and Crawford, 2010). In addition to DNase I hypersensitive sites (DHSs), short regions of protected nucleotides known as footprints can be detected using a technique known as "digital genomic footprinting" (DGF). These analyses potentially indicate the location of transcription factor binding ocuppancy events (Neph et al., 2012). However, available software for DGF analysis is still at a very immature state (Madrigal and Krajewski, 2012).

DNaseR is an R package that aims to identify protein binding footprints in DNase I hypersensitive sites sequencing (DNase-seq) data provided in BAM standard aligment format. It relies on the cumulative function of the Skellam distribution (correlation of two Poisson distributions) to detect narrow-depleted regions of read-enrichment formed by the mapped reads in the forward and reversed DNA strands. Study the imbalance of DNase I cuts separately at both DNA strands is of great help in the detection of reliable protein-binding footprints, as proved by the Wellington algorithm (Piper et al., 2013), which uses the binomial cumulative distribution function for that purpose. As in Wellington, DNaseR's main characteristic consists in that consensus DNA sequences (motifs) search is not required *a priori* to detect footprints.

Any BAM file storing aligned reads coming from a DNase-seq experiment is suitable for footprinting analysis, but the ones more deeply sequenced will retrieve a higher number of significant footprints at a fixed *p*-value cutoff.

[*] pm@engineering.com

2 Methodology

DNase I cuts (5' end of the mapped reads) counts are calculated separately for both DNA strands from the alignment files in BAM format using the Bioconductor package Rsamtools. Using the Skellam distribution (Skellam, 1946), DNaseR models at each nucleotide position the discrete signed difference of two Poisson counts at forward and reverse strands, respectively. Then, detecting nearby located significant count differences of opposed sign at both strands (in the direction 5' to 3') allows DNaseR to delimit the flanks of the footprint location at base pair resolution. A one-sided p-value is obtained for each flank using the complementary cumulative Skellam distribution function. To control for multiple testing the p-values delimiting each flank of the footprint (pval.forward and pval.reverse) are corrected using Benjamini-Hochberg procedure (default). A final p-value for each footprint (default cut-off $1e - 9$) is reported as the sum of the two adjusted p-values.

3 Examples

To test DNaseR, we downloaded the DNase-seq data files wgEncodeUwDgfTh1Aln.bam and wgEncodeUwDgfTh1Aln.bam.bai from the ENCODE Project (Neph et al., 2012) [dataType=DnaseDgf; view=Alignments; cell=Th1; origAssembly=hg18; geoSampleAccession=GSM646569; labVersion=Bowtie 0.12.5; type=bam]. We have selected the reads in the first 3000Kb of chrY, and run the digital genomic footprinting analysis in DNaseR by using only one execution of the function `footprints` (see the manual):

```
R> options(width=80)
R> ## hg18. chrY:1 - 3000Kb reads from DNase-seq dataset wgEncodeUwDgfTh1Aln.bam
R> ## from the ENCODE Project.
R> ##
R> ## Downloaded from:
R> ## http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/
R> ## release1/wgEncodeUwDgfTh1Aln.bam
R>
R> owd <- setwd(tempdir())
R> library(DNaseR)
R> bamfile <- "chrY_3Kb_wgEncodeUwDgfTh1Aln.bam"
R> f <- system.file("extdata", bamfile, package="DNaseR",mustWork = TRUE)
R> dgf <- footprints(bam=f, chrN="chrY", chrL=3e6, p=1e-9, width=c(6,40), N=2e6)
R> head(dgf$footprint.events)
```

```
   chr   start     end length pval.forward pval.reverse pval.footprint.event
1 chrY 2709593 2709608     15 8.218011e-22 1.910582e-12         1.910582e-12
2 chrY 2709619 2709637     18 2.197295e-12 9.879344e-19         2.197296e-12
3 chrY 2709800 2709825     25 2.709836e-15 5.125255e-93         2.709836e-15
4 chrY 2709921 2709927      6 2.197295e-12 9.879344e-19         2.197296e-12
5 chrY 2724916 2724924      8 2.197295e-12 1.910582e-12         4.107877e-12
6 chrY 2724949 2724962     13 2.197295e-12 1.910582e-12         4.107877e-12
  log10.pval.footprint.event
1                   11.71883
2                   11.65811
3                   14.56706
4                   11.65811
5                   11.38638
6                   11.38638
```

```
R> nrow(dgf$footprint.events)

[1] 32

R> setwd(owd)
```

32 protein-binding footprints are reported spannig a width range of 6bp-40bp in the first 3000Kb of chrY for this dataset at p-value $\leq 1e-9$.
If we increase the p-value ($\leq 1e-7$) we get a higher number of footprints (40):

```
R> options(width=80)
R> owd <- setwd(tempdir())
R> library(DNaseR)
R> bamfile <- "chrY_3Kb_wgEncodeUwDgfTh1Aln.bam"
R> f <- system.file("extdata", bamfile, package="DNaseR",mustWork = TRUE)
R> dgf <- footprints(bam=f, chrN="chrY", chrL=3e6, p=1e-7, width=c(6,40), N=2e6)
R> head(dgf$footprint.events)

   chr   start     end length pval.forward pval.reverse pval.footprint.event
1 chrY 2657954 2657984     30 3.258105e-09 2.933723e-09         6.191828e-09
2 chrY 2706206 2706212      6 9.484427e-15 2.933723e-09         2.933733e-09
3 chrY 2709818 2709825      7 3.258105e-09 6.406568e-93         3.258105e-09
4 chrY 2709979 2709992     13 3.258105e-09 5.267549e-32         3.258105e-09
5 chrY 2725032 2725039      7 3.258105e-09 1.515236e-46         3.258105e-09
6 chrY 2725106 2725145     39 3.258105e-09 2.933723e-09         6.191828e-09
  log10.pval.footprint.event
1                   8.208181
2                   8.532579
3                   8.487035
4                   8.487035
5                   8.487035
6                   8.208181

R> nrow(dgf$footprint.events)

[1] 40

R> setwd(owd)
```

For several reasons one might be interested only in footprints of a certain size. For example, to report only 15bp width footprints ($p \leq 1e-9$) we can do:

```
R> options(width=80)
R> owd <- setwd(tempdir())
R> library(DNaseR)
R> bamfile <- "chrY_3Kb_wgEncodeUwDgfTh1Aln.bam"
R> f <- system.file("extdata", bamfile, package="DNaseR",mustWork = TRUE)
R> dgf <- footprints(bam=f, chrN="chrY", chrL=3e6, p=1e-9, width=c(15,15), N=2e6)
R> head(dgf$footprint.events)

   chr   start     end length pval.forward pval.reverse pval.footprint.event
1 chrY 2709593 2709608     15 3.081754e-22 1.910582e-12         1.910582e-12
2 chrY 2781394 2781409     15 5.330695e-19 1.185199e-32         5.330695e-19
3 chrY 2803115 2803130     15 2.078565e-25 1.493181e-15         1.493181e-15
  log10.pval.footprint.event
1                   11.71883
2                   18.27322
3                   14.82589

R> nrow(dgf$footprint.events)

[1] 3

R> setwd(owd)
```

However, it is recommended to be flexible in the max. and min. width during footprint search, as transcription factors are not expected to bind forming unique footprint configurations, nor to totally overlap their consensus sequence motifs.

4 References

- Madrigal P, Krajewski P (2012) Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. **Front Genet** 3: 230.

- Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. **Nature** 489: 83-90.

- Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S (2013) Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. **Nucleic Acids Res**, in press.

- Skellam JG (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. **J R Stat Soc Ser A** 109: 296.

- Song L, Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. **Cold Spring Harb Protoc** 2:pdb.prot5384.

5 Details

This document was written using:

```
R> sessionInfo()
```

```
R version 3.0.2 (2013-09-25)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base
```

```
other attached packages:
[1] DNaseR_1.0.0       IRanges_1.20.0     BiocGenerics_0.8.0

loaded via a namespace (and not attached):
[1] Biostrings_2.30.0    GenomicRanges_1.14.0 Rsamtools_1.14.0
[4] XVector_0.2.0        bitops_1.0-6         stats4_3.0.2
[7] tools_3.0.2          zlibbioc_1.8.0
```