DNaseR: DNase I footprinting analysis of DNase-seq data

Pedro Madrigal

Created: February, 2013. Last modified: January, 2014. Compiled: October 13, 2014

[1] Department of Biometry and Bioinformatics, Institute of Plant Genetics, Polish Academy of Sciences, Poznan, Poland
[2] Present address: Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[3] Present address: Department of Surgery, University of Cambridge, UK

Contents

1 Introduction to Digital Genomic Footprinting **1**

2 Methodology **2**
2.1 Brief summary 2
2.2 Difference of two Poisson variables: The Skellam distribution 2

3 Examples **2**

4 FAQ **5**

5 References **5**

6 Details **5**

1 Introduction to Digital Genomic Footprinting

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 methodology can potentially indicate the location of transcription factor binding ocuppancy events at a nucleotide resolution (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 'double-hit' 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. Studying the imbalance of DNase I cuts separately at both DNA strands is of great help in the detection of reliable protein-binding footprints, as demostrated 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) are 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.

2 Methodology

2.1 Brief summary

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 a 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.

2.2 Difference of two Poisson variables: The Skellam distribution

Footprints from a DGF DNase-seq assay are retrieved using the probability density function of the Skellam distribution (Skellam, 1948), on the DNase I cuts mapped at forward and reverse strands. It has been shown that at the flanking edges of a protein-DNA footprint in DNase-seq data, the difference between DNase I cuts (read-start sites) is higher than on loci not occupied by a TF (Neph et al., 2012; Piper et al., 2013). *DNaseR* uses the Skellam distribution to model DNase I cuts differences around the footprints as a cross-correlation of two Poisson distributions. We model the count difference $n_1 - n_2$ of two statistically independent random variables N_1 (DNase I cuts in $+$ strand) and N_2 (DNase I cuts in - strand), each one having Poisson distribution with different (but almost equal in practice, for big samples) expected values μ_1 and μ_2. This is done under the assumption that DNase I enzyme cleaves each strand of DNA independently, and cleavage sites are random. The probability mass function for the Skellam distribution for a count difference $k = n_1 - n_2$ from two Poisson distributed variables with means μ_1 and μ_2 is given by:

$$f(k;\mu_1,\mu_2) = e^{-(\mu_1+\mu_2)}\left(\frac{\mu_1}{\mu_2}\right)I_k(2\sqrt{\mu_1\mu_2}) \tag{1}$$

Where $I_k(z)$ is the modified Bessel function of the first kind,

$$I_k(z) = \left(\frac{z}{2}\right)^k \sum_{j=0}^{\infty} \frac{\left(\frac{z^2}{4}\right)^j}{j!\Gamma(k+j+1)} \tag{2}$$

Where $\Gamma(a)$ is the gamma function. Then, a footprint ranging between a minimum η_{min} (bp) and a maximum η_{max} (bp.) width is reported when two consecutive statistically significant events are encountered, and only in the cases where the count difference is significantly positive in the direction 5' to 3' ($n_1 - n_2 > 0$), and significantly negative from 3' to 5' ($n_1 - n_2 < 0$). One side p-values, one for each strand, are calculated for each footprint flank using the complementary cumulative Skellam distribution. Benjamini-Hochberg correction for multiple testing was used to adjust the p-values (Benjamini and Hochberg, 1995). As a conservative estimate, *DNaseR* sums the p-values calculated at each strand for each footprint event, but other ways to combine p-values, as the Fisher's method are under study.

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 DGF analysis in by using only one execution of the function `footprints` (see the manual of *DNaseR*):

```
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 DNaseIcuts.start.forward DNaseIcuts.start.reverse
1 chrY 2709593 2709608     15                       6                       0
2 chrY 2709619 2709637     18                       4                       1
3 chrY 2709800 2709825     25                       4                       0
4 chrY 2709921 2709927      6                       3                       0
5 chrY 2724916 2724924      8                       3                       0
6 chrY 2724949 2724962     13                       3                       0
  pval.forward DNaseIcuts.end.forward DNaseIcuts.end.reverse pval.reverse
1 8.218011e-22                     0                     3 1.910582e-12
2 2.197295e-12                     0                     5 9.879344e-19
3 2.709836e-15                     0                    25 5.125255e-93
4 2.197295e-12                     0                     5 9.879344e-19
5 2.197295e-12                     0                     3 1.910582e-12
6 2.197295e-12                     0                     3 1.910582e-12
  pval.footprint.event log10.pval.footprint.event
1         1.910582e-12                   11.71883
2         2.197296e-12                   11.65811
3         2.709836e-15                   14.56706
4         2.197296e-12                   11.65811
5         4.107877e-12                   11.38638
6         4.107877e-12                   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 DNaseIcuts.start.forward DNaseIcuts.start.reverse
1 chrY 2657954 2657984     30                       2                       0
```

```
2 chrY 2706206 2706212      6                       4                       0
3 chrY 2709818 2709825      7                       2                       0
4 chrY 2709979 2709992     13                       2                       0
5 chrY 2725032 2725039      7                       2                       0
6 chrY 2725106 2725145     39                       2                       0
  pval.forward DNaseIcuts.end.forward DNaseIcuts.end.reverse pval.reverse
1 3.258105e-09                     0                       2 2.933723e-09
2 9.484427e-15                     0                       2 2.933723e-09
3 3.258105e-09                     0                      25 6.406568e-93
4 3.258105e-09                     0                       9 5.267549e-32
5 3.258105e-09                     0                      13 1.515236e-46
6 3.258105e-09                     0                       2 2.933723e-09
  pval.footprint.event log10.pval.footprint.event
1         6.191828e-09                   8.208181
2         2.933733e-09                   8.532579
3         3.258105e-09                   8.487035
4         3.258105e-09                   8.487035
5         3.258105e-09                   8.487035
6         6.191828e-09                   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 DNaseIcuts.start.forward DNaseIcuts.start.reverse
1 chrY 2709593 2709608     15                       6                       0
2 chrY 2781394 2781409     15                       6                       1
3 chrY 2803115 2803130     15                      12                       5
  pval.forward DNaseIcuts.end.forward DNaseIcuts.end.reverse pval.reverse
1 3.081754e-22                     0                       3 1.910582e-12
2 5.330695e-19                     0                       9 1.185199e-32
3 2.078565e-25                     0                       4 1.493181e-15
  pval.footprint.event log10.pval.footprint.event
1         1.910582e-12                   11.71883
2         5.330695e-19                   18.27322
3         1.493181e-15                   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 highly-scored matches of their consensus sequence motifs.

4 FAQ

- **How replicates are treated in DNaseR?** The package analyses just one sample. If you have several biological replicates you might like to consider as a result only the consensus list of footprints, or submitting footprint scores (`log10.pval.footprint.event`) for Irreproducible Discovery Rate analysis (Li et al., 2011), which is an accepted practice in RNA-seq and ChIP-seq (Bailey et al., 2013).
- **Can I get False Discovery Rate for the list of footprints reported?** See the Bioconductor package *qvalue*.

5 References

- Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J (2013) Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. **PLoS Comput Biol** 9(11): e1003326.
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. **J R Stat Soc Ser B** 57: 289-300.
- Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. **Ann Appl Stat** 5(3): 1699-2264.
- 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.

6 Details

This document was written using:

```
R> sessionInfo()

R version 3.1.1 Patched (2014-09-25 r66681)
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] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
[1] DNaseR_1.4.0        IRanges_2.0.0       S4Vectors_0.4.0
```

```
[4] BiocGenerics_0.12.0

loaded via a namespace (and not attached):
[1] BiocStyle_1.4.0      Biostrings_2.34.0    GenomeInfoDb_1.2.0
[4] GenomicRanges_1.18.0 Rsamtools_1.18.0     XVector_0.6.0
[7] bitops_1.0-6         tools_3.1.1          zlibbioc_1.12.0
```