DBChIP: Detecting differential binding of transcription factors with ChIP-seq

Kun Liang[1,2] and Sündüz Keleş[1,2]
[1]Department of Statistics, University of Wisconsin
Madison, WI 53706.
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin
Madison, WI 53706.

April 16, 2015

Contents

1 Introduction

This document provides an introduction to the differential binding analysis of ChIP-seq data with the `DBChIP` package [1]. We focus on the binding sites that have similar read profiles and well-defined centers throughout the genome. These binding sites tend to have read profiles that look like sharp peaks. The examples are transcription factor binding and some histone modifications measured by ChIP-seq.

An increasing number of ChIP experiments are investigating the same type of binding event (protein-DNA binding or histone modification) under different conditions (treatments, time points, cell lines, etc.). A natural question is which binding sites behave differently across the conditions, and we set out to address this question.

2 Overview

Here are the major steps `DBChIP` performs:

1. Consensus site: Binding site predictions from multiple conditions are merged into consensus binding sites. This step is necessary because the predictions for the same binding site across different conditions are unlikely to be exactly the same.

2. Counting reads: The number of reads contributing to the binding are counted for each consensus site from the aligned sequence reads data.

3. Detecting differential binding: We formally test a null hypothesis of non-differential binding at each consensus site. The tests are generally carried out through a generalized linear model with Negative Binomial distribution to account for the over-dispersion among replicates. We report a p-value and fold change estimates between conditions for each site. `DBChIP` can then report significantly differentially bound sites according to a pre-specified false discovery rate (FDR) threshold.

3 Example

The example dataset cames from a study of the transcription factor PHA-4 in *C.elegans* under two developmental conditions: embryonic and the first stage of larval development (L1) under starvation conditions [2]. There are two replicates in each condition. In `DBChIP`, replicates stand for biological replicates; technical replicates usually can be merged after their consistency is established. To keep the data size small, only alignment data (ChIP and control) and identified binding sites in chromosome I with position < 0.9M bp are included.

First, we load the `DBChIP` library and the `PHA4` dataset.

```
> library(DBChIP)
> data("PHA4")
```

Here we specify the experiment condition of each ChIP replicate.

```
> conds <- factor(c("emb","emb","L1", "L1"), levels=c("emb", "L1"))
```

`DBChIP` requires a set of binding site predictions from each experiment condition. The binding site predictions should contain the following fields: `chr`, chromosome; `pos`, the binding position; (optional) `weight`, a measure of strength of the binding (for example, the number of reads in the peak). Here we read the predictions into `binding.site.list` from local files.

```
> path <- system.file("ext", package="DBChIP")
> binding.site.list <- list()
> binding.site.list[["emb"]] <- read.table(paste(path, "/emb.binding.txt", sep=""),
+ header=TRUE)
> head(binding.site.list[["emb"]])
```

```
  chr    pos weight
1   I 397898  152.6
2   I 536855  147.7
3   I 315229  122.9
4   I 382525  130.3
5   I 535441  135.6
6   I 882789   27.4
```

```
> binding.site.list[["L1"]] <- read.table(paste(path, "/L1.binding.txt", sep=""),
+ header=TRUE)
> bs.list <- read.binding.site.list(binding.site.list)
```

Then the binding sites from different conditions are merged into consensus sites.

```
> consensus.site <- site.merge(bs.list)

merging sites from different conditions to consensus sites.done
```

PHA4 data contain the raw ChIP (`chip.data.list`) and control/input (`input.data.list`) data. ChIP data are organized as biological replicates.

```
> names(chip.data.list)

[1] "emb_rep1" "emb_rep2" "L1_rep1"  "L1_rep2"

> head(chip.data.list[["emb_rep1"]])

   chr strand    pos
50   I      + 546758
53   I      + 184288
62   I      + 255673
85   I      + 180366
87   I      + 492186
89   I      + 881468
```

Here the alignment data are in Minimum ChIP-Seq (MCS) format, which is a data.frame with following fields: `chr` (factor), `pos` (integer) and `strand` (factor, "+" and "-"). Note that the `pos` is the $5'$ position of the read. Most commonly used alignment formats are supported in `DBChIP`, see Section 4.1 for more details. On the other hand, input data are usually organized per condition with replicates within each condition merged. This is because the focus in differential binding analysis is on the biological variation among ChIP replicates, while the input data are mainly used to provide estimates of the background noise level.

```
> names(input.data.list)

[1] "emb" "L1"
```

To facilitate the data loading, we use `load.data` function:

```
> dat <- load.data(chip.data.list=chip.data.list, conds=conds, consensus.site=
+ consensus.site, input.data.list=input.data.list, data.type="MCS")

reading data...done
computing normalization factor between ChIP and control samples.done
```

Then we count the reads around each consensus binding site.

```
> dat <- get.site.count(dat)
```

```
count ChIP reads around each binding site.done
```

Differential binding detection

```
> dat <- test.diff.binding(dat)

Common dispersion: 0.05915362

> rept <- report.peak(dat)
> rept
```

	chr	pos	nsig	origin	ori.pos	FC.L1	pval	FDR
1	I	260346	1	emb	260346	0.1129558	1.057645e-11	5.076698e-10
2	I	673122	1	emb	673122	0.1168590	1.798443e-10	4.316264e-09
3	I	757094	1	emb	757094	0.1717831	4.012809e-09	6.420495e-08
4	I	454904	1	emb	454904	0.1632294	1.349896e-08	1.619875e-07
5	I	547611	1	L1	547611	5.6827777	2.104486e-08	2.020307e-07
6	I	41410	1	L1	41410	4.2984821	4.548449e-07	3.638759e-06
7	I	546710	1	L1	546710	4.3025538	7.621600e-06	5.226240e-05
8	I	43116	1	emb	43116	0.2581921	6.466525e-05	3.879915e-04
9	I	159188	1	L1	159188	2.9888809	3.449015e-04	1.839475e-03
10	I	3907	1	emb	3907	0.3718658	7.035650e-04	3.377112e-03

The column `FC.L1` contains the fold change of the L1 condition with respect to the embryonic condition (`emb` is the first condition and is used as the baseline). By default, `report.peak` returns top 10 most differentially bound sites. The number of sites to return can be specified through parameter `n`. We can also specify a FDR level to return only the sites deemed significant enough. Finally, we can inspect our results by looking at their coverage plots.

In Figure 1, each read is extended by the average fragment length (default 200 bp) from its $5'$ end towards its $3'$ end. The coverage at each nucleotide is defined as the number of extended reads covering the position and is computed separately for ChIP sample forward strand (blue), ChIP sample reverse strand (red), control sample forward strand (green), control sample reverse strand (orange).

4 Technical details

4.1 Alignment data

Besides the Minimum ChIP-Seq (MCS) format used in our example, most commonly used alignment formats (Eland, MAQ, Bowtie, SOAP, BAM, etc.) are supported through the `AlignedRead` object from `Bioconductor ShortRead` package. For example, a BAM file can be read into an `AlignedRead` object as follows:

```
> library(ShortRead)
> aln <- readAligned("./", pattern="emb.bam", type="BAM")
> chip.data.list[["emb"]] <- aln
```

The last option is through BED files, where we require at least the first 6 fields (chrom, start, end, name, score and strand). Here we provide the list of BED file names.

```
> chip.data.list <- list()
> chip.data.list[["emb"]] <- "/path/emb.bed.file"
> chip.data.list[["L1"]] <- "/path/L1.bed.file"
```

Then we can simply specify `data.type="BED"` in the `DBChIP` or `load.data` function.

4.2 Consensus site

Here we provide more operational details about obtaining consensus sites through the `site.merge` function. Because the predictions for the same binding site across multiple conditions tend to cluster together, we employ agglomerative (bottom-up) hierarchical clustering with centroid linkage to group predicted locations into different clusters. The centroid is computed as the average of the locations within each cluster. If the distance between centroids of two clusters are smaller than `in.distance` (default 100 bp), the clusters are considered as coming from the same binding site and are merged into one cluster. On the other hand, two clusters are considered as coming from separate sites if the distance between two respective centroids are larger than `out.distance` (default 250 bp). If the distance between the centroids of two clusters is between `in.distance` and `out.distance`, the cluster with higher weight will be kept. Finally, the consensus position within each cluster is an (weighted) average of original positions.

5 Session Info

```
> sessionInfo()

R version 3.2.0 RC (2015-04-08 r68161)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server 2008 R2 x64 (build 7601) Service Pack 1

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

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

other attached packages:
[1] DBChIP_1.12.0       DESeq_1.20.0        lattice_0.20-31
[4] locfit_1.5-9.1      Biobase_2.28.0      BiocGenerics_0.14.0
[7] edgeR_3.10.0        limma_3.24.0

loaded via a namespace (and not attached):
 [1] IRanges_2.2.0         XML_3.98-1.1          GenomeInfoDb_1.4.0
 [4] grid_3.2.0            xtable_1.7-4          DBI_0.3.1
 [7] stats4_3.2.0          RSQLite_1.0.0         genefilter_1.50.0
```

```
[10] annotate_1.46.0      S4Vectors_0.6.0      splines_3.2.0
[13] RColorBrewer_1.1-2   tools_3.2.0          geneplotter_1.46.0
[16] survival_2.38-1      AnnotationDbi_1.30.0
```

References

[1] K. Liang and S. Keleş. Detecting differential binding of transcription factors with chip-seq. *Bioinformatics*, 28(1):121–122, 2012.

[2] M. Zhong, W. Niu, Z.J. Lu, M. Sarov, J.I. Murray, J. Janette, D. Raha, K.L. Sheaffer, H.Y.K. Lam, E. Preston, et al. Genome-wide identification of binding sites defines distinct functions for caenorhabditis elegans pha-4/foxa in development and environmental response. *PLoS Genet*, 6(2):e1000848, 2010.

```
> plotPeak(rept[1,], dat)
```

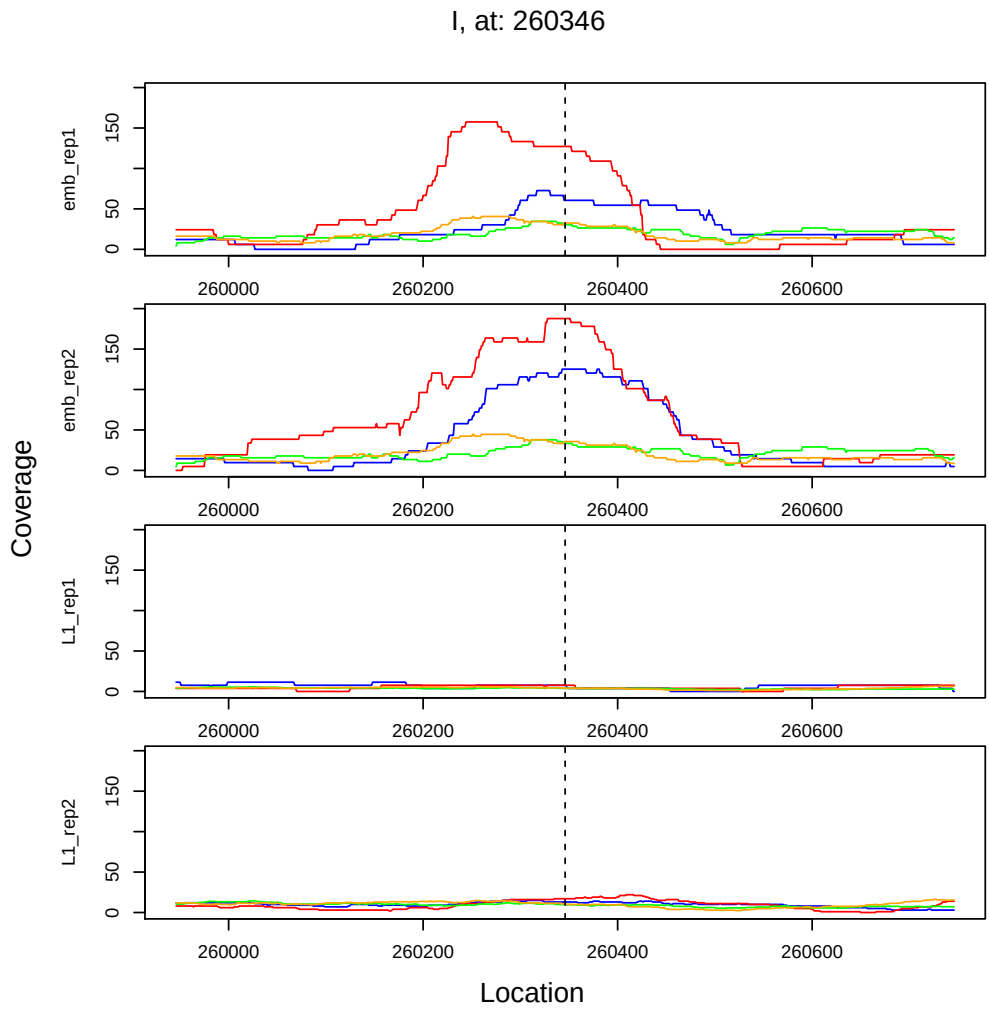

Figure 1: Coverage plot for the most significantly differentially bound site (chromosome I location 260346). Color index: ChIP sample forward strand (blue), ChIP sample reverse strand (red), control sample forward strand (green), control sample reverse strand (orange).