chipseeker: an R package for ChIP peak Annotation, Visualization and Comparison

Guangchuang Yu
The University of Hong Kong

April 25, 2014

Contents

1 Introduction 1

2 ChIP profiling 2
2.1 ChIP peaks over Chromosomes 3
2.2 Heatmap of ChIP binding to TSS regions 3
2.3 Average Profile of ChIP peaks binding to TSS region 3

3 Peak Annotation 3

4 Visualize Genomic Annotation 7

5 Visualize distribution of TF-binding loci relative to TSS 8

6 Compare among several ChIPseq data 9

7 Overlap of peaks and annotated genes 11

8 Functional enrichment analysis 12

9 Session Information 13

1 Introduction

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has become standard technologies for genome wide identification of DNA-binding protein target sites. After read mappings and peak callings, the peak should be annotated to answer the biological questions. I developed an R package called *chipseeker* for annotating nearest genes and genomic features to peaks. Several plot functions are implemented to summarize peaks and peak

annotation. Functional enrichment analysis of the peaks can be performed by my Bioconductor packages *DOSE* , *ReactomePA*, *clusterProfiler* [1] .

```
## loading packages
require(ChIPseeker)
require(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
require(clusterProfiler)
```

2 ChIP profiling

The datasets in this vignettes were downloaded from GEO (GSE40740) [2] . *ChIPseeker* provides `readPeakFile` to load the peak and store in `GRanges` object. Most of the functions in *ChIPseeker* can accept input in peak file (bed format) or `GRanges` object.

```
files <- getSampleFiles()
print(files)

## $CBX6
## [1] "/usr/local/Cellar/r/3.1.0/R.framework/Versions/3.1/Resources/library/ChIPse
##
## $CBX7
## [1] "/usr/local/Cellar/r/3.1.0/R.framework/Versions/3.1/Resources/library/ChIPse
##
## $CBX8
## [1] "/usr/local/Cellar/r/3.1.0/R.framework/Versions/3.1/Resources/library/ChIPse
##
## $RING1
## [1] "/usr/local/Cellar/r/3.1.0/R.framework/Versions/3.1/Resources/library/ChIPse
##
## $RING2
## [1] "/usr/local/Cellar/r/3.1.0/R.framework/Versions/3.1/Resources/library/ChIPse

peak <- readPeakFile(files[[1]])
peak

## GRanges with 1331 ranges and 2 metadata columns:
##          seqnames                 ranges strand   |             V4        V5
##             <Rle>              <IRanges>  <Rle>   |       <factor> <numeric>
##      [1]     chr1     [ 815092,  817883]      *   |    MACS_peak_1     295.8
##      [2]     chr1     [1243287, 1244338]      *   |    MACS_peak_2      63.2
##      [3]     chr1     [2979976, 2981228]      *   |    MACS_peak_3     100.2
##      [4]     chr1     [3566181, 3567876]      *   |    MACS_peak_4     558.9
```

```
##      [5]     chr1     [3816545, 3818111]      *   |    MACS_peak_5      57.6
##      ...      ...                    ...    ... ...            ...       ...
##   [1327]     chrX [135244782, 135245821]      *   | MACS_peak_1327      55.5
##   [1328]     chrX [139171963, 139173506]      *   | MACS_peak_1328     270.2
##   [1329]     chrX [139583953, 139586126]      *   | MACS_peak_1329     918.7
##   [1330]     chrX [139592001, 139593238]      *   | MACS_peak_1330     210.9
##   [1331]     chrY [ 13845133,  13845777]      *   | MACS_peak_1331      58.4
##   ---
##   seqlengths:
##     chr1 chr10 chr11 chr12 chr13 chr14 ...  chr6  chr7  chr8  chr9  chrX  chrY
##       NA    NA    NA    NA    NA    NA ...    NA    NA    NA    NA    NA    NA
```

2.1 ChIP peaks over Chromosomes

After peak calling, we would like to know the peak locations over the whole genome, `plotChrCov` function calculates the coverage of peak regions over chromosomes and generate a figure to visualize.

```
plotChrCov(peak, weightCol = "V5")
```

2.2 Heatmap of ChIP binding to TSS regions

```
plotPeakHeatmap(files[[1]], weightCol = "V5", TranscriptDb = txdb,
    upstream = 3000, downstream = 3000, color = "red")
```

2.3 Average Profile of ChIP peaks binding to TSS region

```
plotPeakProf(files[[1]], TranscriptDb = txdb, upstream = 3000,
    downstream = 3000)
```

```
plotPeakProf(files, TranscriptDb = txdb, upstream = 3000,
    downstream = 3000)
```

3 Peak Annotation

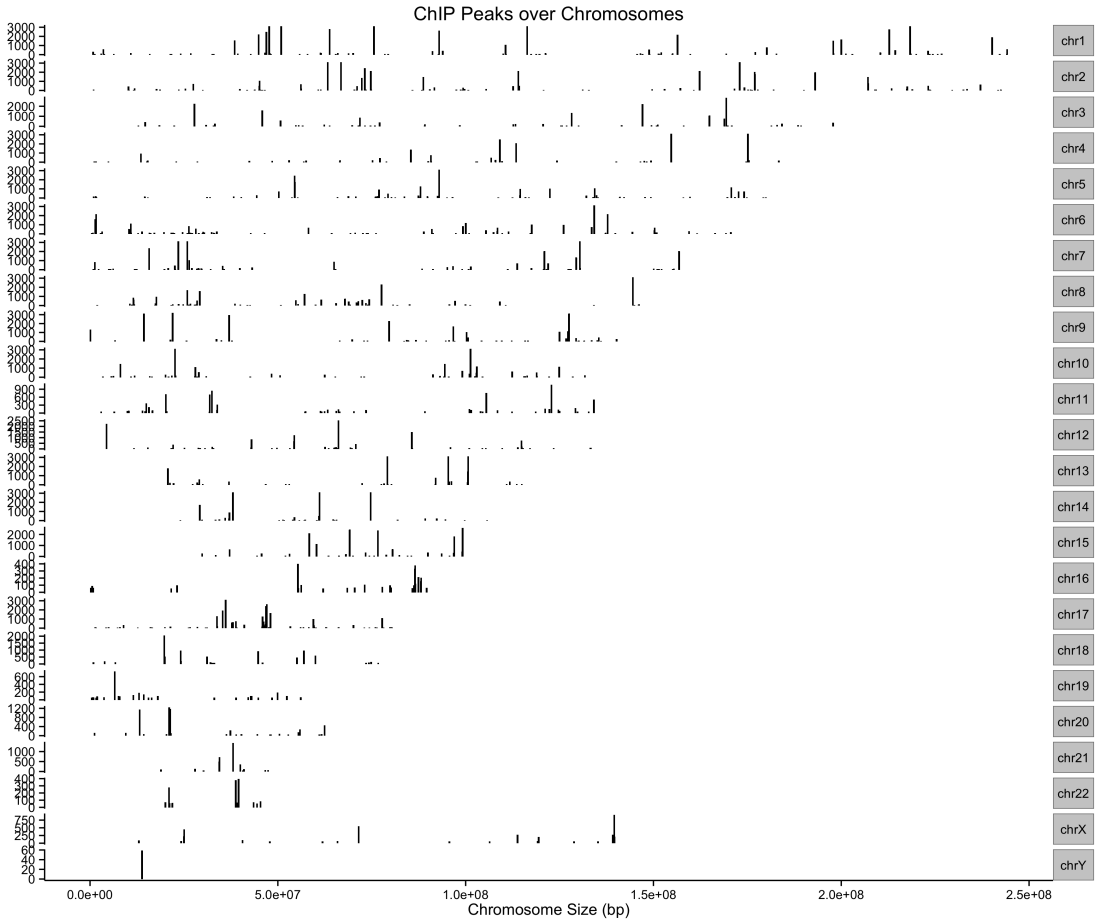

Figure 1: ChIP peaks over Chromosomes

```
peakAnno <- annotatePeak(files[[1]], tssRegion = c(-3000,
    3000), as = "GRanges", TranscriptDb = txdb, annoDb = "org.Hs.eg.db")

## >> loading peak file...  2014-04-25 21:49:12
## >> preparing features information...  2014-04-25 21:49:12
## >> identifying nearest features...  2014-04-25 21:49:22
## >> calculating distance from peak to TSS...  2014-04-25 21:49:23
## >> assigning genomic annotation...  2014-04-25 21:49:23
## >> adding gene annotation...  2014-04-25 21:50:09
## >> assigning chromosome lengths  2014-04-25 21:50:10
## >> done...  2014-04-25 21:50:10

head(peakAnno)

## GRanges with 6 ranges and 13 metadata columns:
##       seqnames             ranges strand |          V4        V5
##          <Rle>          <IRanges>  <Rle> |    <factor> <numeric>
##   [1]     chr1 [ 815092,  817883]      * | MACS_peak_1     295.8
##   [2]     chr1 [1243287, 1244338]      * | MACS_peak_2      63.2
```

Figure 2: Heatmap of ChIP peaks binding to TSS regions

```
##   [3]     chr1 [2979976, 2981228]      * | MACS_peak_3     100.2
##   [4]     chr1 [3566181, 3567876]      * | MACS_peak_4     558.9
##   [5]     chr1 [3816545, 3818111]      * | MACS_peak_5      57.6
##   [6]     chr1 [6304864, 6305704]      * | MACS_peak_6      54.6
##                    annotation  geneChr geneStart   geneEnd geneLength
##                   <character> <factor> <integer> <integer>  <integer>
##   [1]              Intergenic     chr1    803451    812182       8732
##   [2]                Promoter     chr1   1227764   1244989      17226
##   [3] Exon (4267 exon 1 of 6)     chr1   2976181   2984289       8109
##   [4]                Promoter     chr1   3569129   3652765      83637
##   [5]  Exon (197 exon 1 of 4)     chr1   3773845   3801993      28149
##   [6]                Promoter     chr1   6304252   6305638       1387
##       geneStrand      geneId distanceToTSS         ENSEMBL      SYMBOL
##         <factor> <character>     <integer>     <character> <character>
##   [1]          -      284593          5701 ENSG00000230368      FAM41C
##   [2]          -      116983           651 ENSG00000131584       ACAP3
##   [3]          -      440556          3061 ENSG00000177133   LINC00982
```

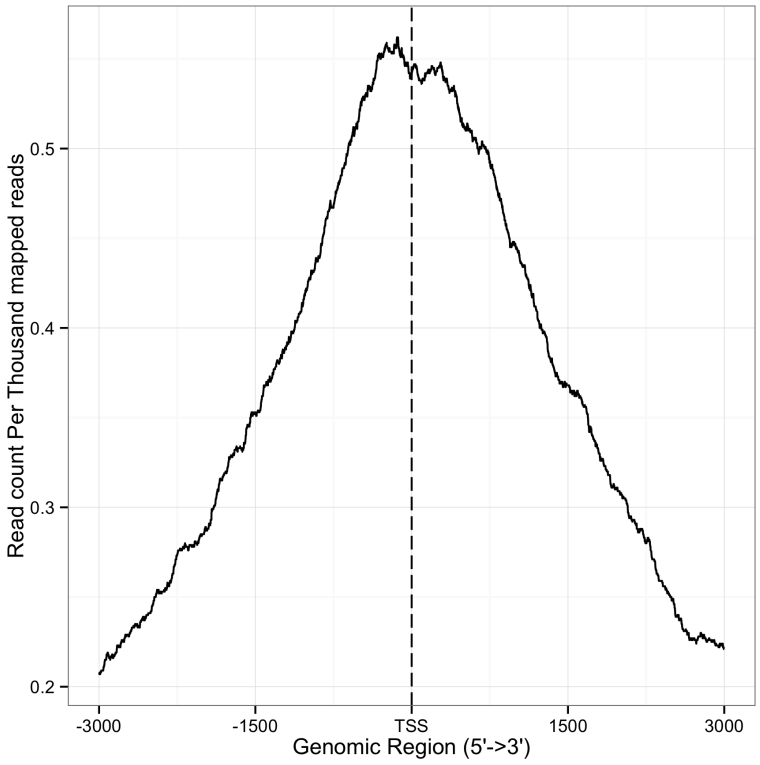

Figure 3: Average Profile of ChIP peaks binding to TSS region

```
##   [4]           +        7161         -2948 ENSG00000078900        TP73
##   [5]           +        1677       -42700 ENSG00000169598        DFFB
##   [6]           +      390992         -612 ENSG00000173673        HES3
##                                                                      GENENAM
##                                                                    <character
##   [1]                                  family with sequence similarity 41, member
##   [2]                    ArfGAP with coiled-coil, ankyrin repeat and PH domains
##   [3]                                    long intergenic non-protein coding RNA 98
##   [4]                                                          tumor protein p7
##   [5] DNA fragmentation factor, 40kDa, beta polypeptide (caspase-activated DNase
##   [6]                                       hes family bHLH transcription factor
##   ---
##   seqlengths:
##         chr1     chr10     chr11     chr12 ...      chr9      chrX      chrY
##    249250621 135534747 135006516 133851895 ... 141213431 155270560  59373566
```

Peak Annotation is performed by `annotatePeak`. User can define TSS (transcription start site) region, by default TSS is defined from -3kb to +100bp. The argument *as* can be one of "GRanges", "data.fram" and "txt" to specify the output format return by `annotatePeak`. If *as* is set to "txt", the output will save to a TXT file with name suffix by anno.txt.

TranscriptDb object should be passed for annotation. annoDb is optional, if provided, extra columns such as SYMBOL, GENENAME will be added.

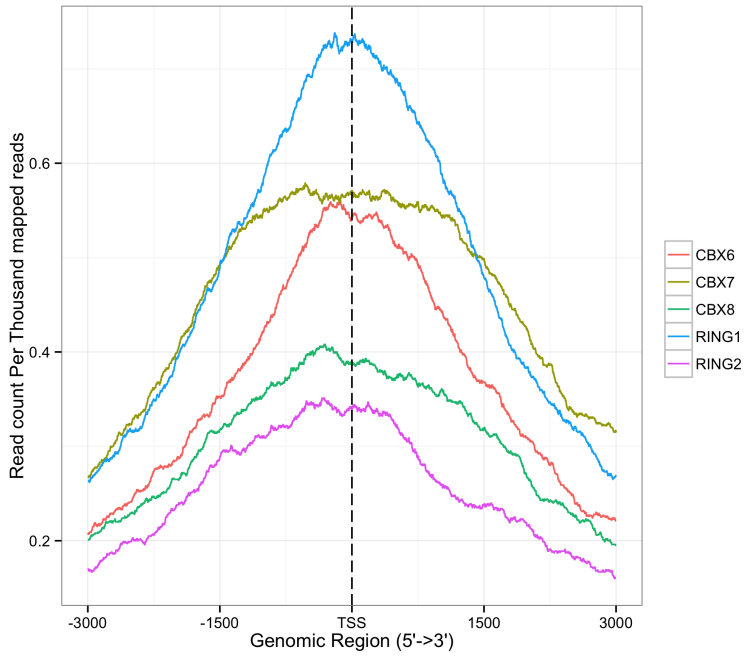

Figure 4: Average Profiles of ChIP peaks among different experiments

All the peak information contained in peakfile will be retained in the output of `annotatePeak`. The position and strand information of nearest genes are reported. The distance from peak to the TSS of its nearest gene is also reported. The genomic region of the peak is reported in annotation column. `annotatePeak` report detail information, for instance "Exon (38885 exon 3 of 11)", means that the peak is overlap with an Exon of gene 38885 (EntrezID), and this overlaped exon is the 3rd exon of the 11 exons that gene 38885 prossess.

4 Visualize Genomic Annotation

To annotate the location of a given peak in terms of genomic features, `annotatePeak` assigns peaks to genomic annotation in "annotation" column of the output, which includes whether a peak is in the TSS, Exon, 5' UTR, 3' UTR, Intronic or Intergenic. Many researchers are very interesting in these annotations. TSS region can be defined by user and `annotatePeak` output in details of which exon/intron of which genes as illustrated in previous section.

Pie and Bar plot are supported to visualize the genomic annotation.

```
plotAnnoPie(peakAnno)
```

```
plotAnnoBar(peakAnno)
```

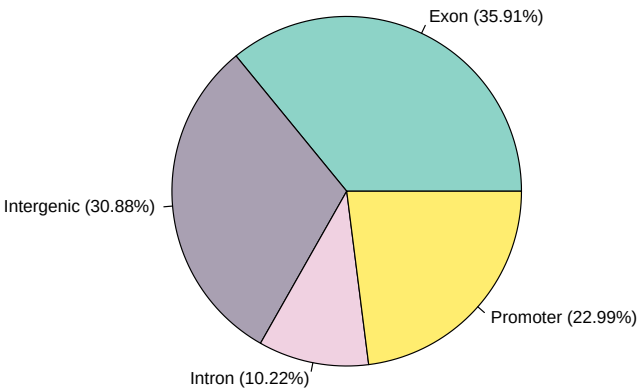

Figure 5: Genomic Annotation by pieplot

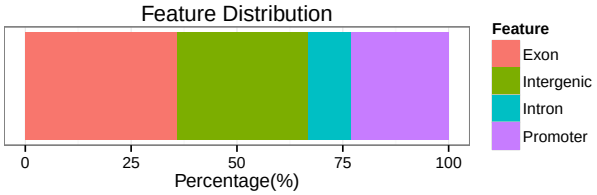

Figure 6: Genomic Annotation by barplot

5 Visualize distribution of TF-binding loci relative to TSS

The distance from the peak (binding site) to the TSS of the nearest gene is calculated by `annotatePeak` and reported in the output. We provide `plotDistToTSS` to calculate the percentage of binding sites upstream and downstream from the TSS of the nearest genes, and visualize the distribution.

```
plotDistToTSS(peakAnno, title = "Distribution of transcription factor-binding loci
```

```
## Warning:  Stacking not well defined when ymin != 0
```

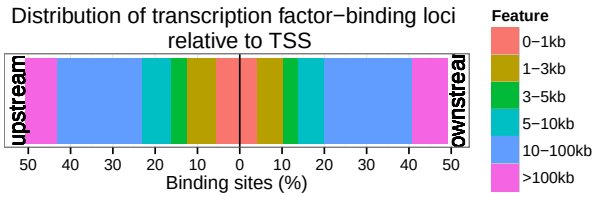

Figure 7: Distribution of Binding Sites

6 Compare among several ChIPseq data

The `plotAnnoBar` and `plotDistToTSS` can also accept input of a named list of annotated peaks (output of `annotatePeak`).

For illustration, here I create a named list from peakAnno object:

```
set.seed(123)
peakAnnoList <- lapply(1:3, function(i) peakAnno[sample(1:length(peakAnno),
    100), ])
names(peakAnnoList) <- paste("peak", 1:3, sep = "_")
lapply(peakAnnoList, head, n = 2)

## $peak_1
## GRanges with 2 ranges and 13 metadata columns:
##       seqnames                ranges strand |             V4        V5
##          <Rle>             <IRanges>  <Rle> |       <factor> <numeric>
##   [1]    chr14 [ 99737653,  99738328]      * |  MACS_peak_383      59.5
##   [2]     chr6 [105403632, 105406129]      * | MACS_peak_1049     389.9
##                    annotation  geneChr geneStart   geneEnd geneLength
##                   <character> <factor> <integer> <integer>  <integer>
##   [1] Exon (53406 exon 1 of 4)    chr14  99864083  99947226      83144
##   [2] Exon (25132 exon 1 of 4)     chr6 105384169 105388402       4234
##       geneStrand      geneId distanceToTSS         ENSEMBL      SYMBOL
##         <factor> <character>     <integer>     <character> <character>
##   [1]          -       84193        208898 ENSG00000183576       SETD3
##   [2]          -   100113403         17727 ENSG00000203809   LINC00577
##                                     GENENAME
##                                    <character>
##   [1]                   SET domain containing 3
##   [2] long intergenic non-protein coding RNA 577
##   ---
##   seqlengths:
##         chr1     chr10     chr11     chr12 ...      chr9      chrX      chrY
##    249250621 135534747 135006516 133851895 ... 141213431 155270560  59373566
##
## $peak_2
## GRanges with 2 ranges and 13 metadata columns:
##       seqnames                ranges strand |            V4        V5
```

```
##          <Rle>            <IRanges>  <Rle> |      <factor> <numeric>
##   [1]     chr3 [73196998, 73198722]      * | MACS_peak_799      52.8
##   [2]    chr16 [70472724, 70474054]      * | MACS_peak_443      63.5
##                     annotation  geneChr geneStart   geneEnd geneLength
##                    <character> <factor> <integer> <integer>  <integer>
##   [1]               Intergenic     chr3  73110810  73112471       1662
##   [2] Exon (59794 exon 1 of 7)    chr16  70488498  70514177      25680
##       geneStrand      geneId distanceToTSS         ENSEMBL      SYMBOL
##         <factor> <character>     <integer>     <character> <character>
##   [1]          +       55096        -86188 ENSG00000255423       EBLN2
##   [2]          +      197258        -15774 ENSG00000157353         FUK
##                                       GENENAME
##                                    <character>
##   [1] endogenous Bornavirus-like nucleoprotein 2
##   [2]                                 fucokinase
##   ---
##   seqlengths:
##         chr1     chr10     chr11     chr12 ...      chr9      chrX      chrY
##    249250621 135534747 135006516 133851895 ... 141213431 155270560  59373566
##
## $peak_3
## GRanges with 2 ranges and 13 metadata columns:
##       seqnames                ranges strand |             V4        V5
##          <Rle>             <IRanges>  <Rle> |       <factor> <numeric>
##   [1]    chr13 [ 79165492,  79166384]      * |  MACS_peak_318      70.7
##   [2]     chr9 [100617729, 100619111]      * | MACS_peak_1280      98.6
##                         annotation  geneChr geneStart   geneEnd geneLength
##                        <character> <factor> <integer> <integer>  <integer>
##   [1] Intron (49787 intron 3 of 5)    chr13  79173230  79177695       4466
##   [2]                     Promoter     chr9 100615537 100618997       3461
##       geneStrand      geneId distanceToTSS         ENSEMBL      SYMBOL
##         <factor> <character>     <integer>     <character> <character>
##   [1]          -        5457         11311 ENSG00000152192      POU4F1
##   [2]          +        2304         -2192 ENSG00000178919       FOXE1
##                                             GENENAME
##                                          <character>
##   [1]                           POU class 4 homeobox 1
##   [2] forkhead box E1 (thyroid transcription factor 2)
##   ---
##   seqlengths:
##         chr1     chr10     chr11     chr12 ...      chr9      chrX      chrY
##    249250621 135534747 135006516 133851895 ... 141213431 155270560  59373566
```

We can use `plotAnnoBar` to comparing their genomic annotation.

```
plotAnnoBar(peakAnnoList)
```

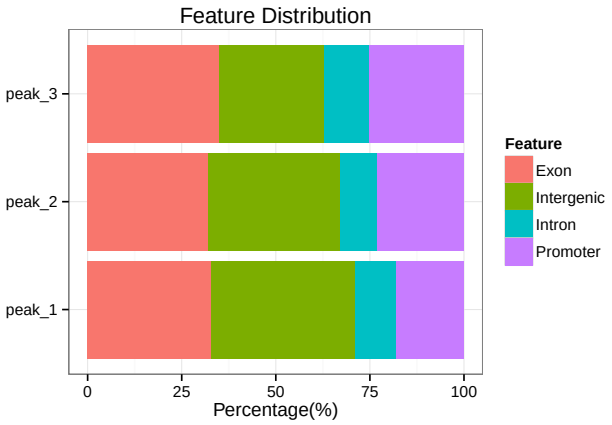

Figure 8: Genomic Annotation among different ChIPseq data

R function `plotDistToTSS` can use to comparing distance to TSS profiles among ChIPseq data.

```
plotDistToTSS(peakAnnoList)

## Warning:  Stacking not well defined when ymin != 0
```

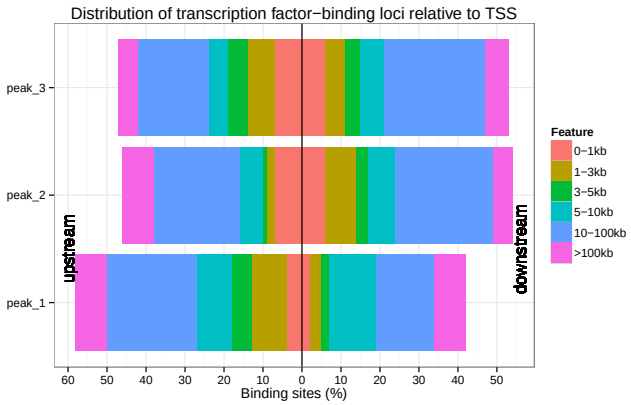

Figure 9: Distribution of Binding Sites among different ChIPseq data

7 Overlap of peaks and annotated genes

User may want to compare the overlap peaks of replicate experiments or from different experiments. *ChIPseeker* provides `peak2GRanges` that can read peak file and stored in GRanges object. Several files can be read simultaneously using lapply, and then passed to `vennplot` to calculate their overlap and draw venn plot.

`vennplot` accept a list of object, can be a list of GRanges or a list of vector. Here, I will demonstrate using `vennplot` to visualize the overlap of the nearest genes stored in peakAnnoList.

```
genes = lapply(peakAnnoList, function(i) unlist(i$geneId))
vennplot(genes)
```

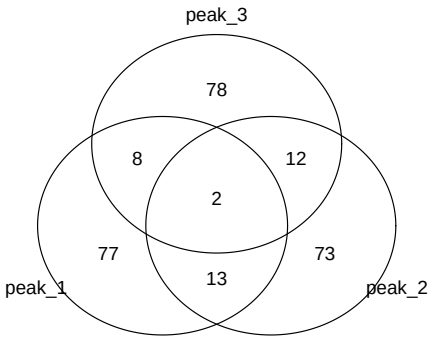

Figure 10: Overlap of annotated genes

8 Functional enrichment analysis

Once we have obtained the annotated nearest genes, we can perform functional enrichment analysis to identify predominant biological themes among these genes by incorporating biological knowledge provided by biological ontologies. For instance, Gene Ontology (GO) [3] annotates genes to biological processes, molecular functions, and cellular components in a directed acyclic graph structure, Kyoto Encyclopedia of Genes and Genomes (KEGG) [4] annotates genes to pathways, Disease Ontology (DO) [5] annotates genes with human disease association, and Reactome [6] annotates gene to pathways and reactions.

Enrichment analysis is a widely used approach to identify biological themes. I have developed several Bioconductor packages for investigating whether the number of selected genes associated with a particular biological term is larger than expected, including *DOSE* for Disease Ontology, *ReactomePA* for reactome pathway, *clusterProfiler* [1] for Gene Ontology and KEGG enrichment analysis.

```
require(clusterProfiler)
bp <- enrichGO(unlist(peakAnno$geneId), ont = "BP",
    readable = TRUE)

## Loading required package:  GO.db
```

```
head(summary(bp))
```

```
##                    ID                          Description GeneRatio
## GO:0008150 GO:0008150                   biological_process   734/734
## GO:0007275 GO:0007275  multicellular organismal development   366/734
## GO:0044767 GO:0044767 single-organism developmental process   393/734
## GO:0032502 GO:0032502                developmental process   395/734
## GO:0048731 GO:0048731                   system development   321/734
## GO:0048513 GO:0048513                    organ development   259/734
##                BgRatio   pvalue p.adjust   qvalue
## GO:0008150 15034/18207 3.65e-63 5.70e-60 2.71e-60
## GO:0007275  4274/18207 6.95e-57 5.43e-54 2.59e-54
## GO:0044767  4848/18207 3.86e-56 2.01e-53 9.58e-54
## GO:0032502  4899/18207 6.76e-56 2.64e-53 1.26e-53
## GO:0048731  3530/18207 1.42e-53 4.43e-51 2.11e-51
## GO:0048513  2489/18207 5.42e-52 1.41e-49 6.72e-50
##
## GO:0008150 TCF24/FRAT1/CDH6/LOC100506422/CASP12/CDH8/EDIL3/AASS/OLIG2/SLC17A2/SP
## GO:0007275
## GO:0044767
## GO:0032502
## GO:0048731
## GO:0048513
##            Count
## GO:0008150   734
## GO:0007275   366
## GO:0044767   393
## GO:0032502   395
## GO:0048731   321
## GO:0048513   259
```

More information can be found in the vignettes of Bioconductor packages *DOSE* , *ReactomePA*, *clusterProfiler* [1], which also provide several methods to visualize enrichment results. The *clusterProfiler* package is designed for comparing and visualizing functional profiles among gene clusters, and can directly applied to compare biological themes at GO, DO, KEGG, Reactome perspective.

9 Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 3.1.0 (2014-04-10), `x86_64-apple-darwin13.1.0`
- Locale: `C`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils

- Other packages: AnnotationDbi 1.26.0, Biobase 2.24.0, BiocGenerics 0.10.0, ChIPseeker 1.1.2, DBI 0.2-7, GO.db 2.14.0, GenomeInfoDb 1.0.2, GenomicFeatures 1.16.0, GenomicRanges 1.16.2, IRanges 1.22.3, RSQLite 0.11.4, TxDb.Hsapiens.UCSC.hg19.knownGene 2.14.0, XVector 0.4.0, clusterProfiler 1.12.0, ggplot2 0.9.3.1, knitr 1.5, org.Hs.eg.db 2.14.0

- Loaded via a namespace (and not attached): BBmisc 1.5, BSgenome 1.32.0, BatchJobs 1.2, BiocParallel 0.6.0, Biostrings 2.32.0, DO.db 2.8.0, DOSE 2.2.0, GOSemSim 1.22.0, GenomicAlignments 1.0.0, KEGG.db 2.14.0, KernSmooth 2.23-12, MASS 7.3-31, Matrix 1.1-3, RColorBrewer 1.0-5, RCurl 1.95-4.1, Rcpp 0.11.1, Rsamtools 1.16.0, XML 3.98-1.1, biomaRt 2.20.0, bitops 1.0-6, brew 1.0-6, caTools 1.17, codetools 0.2-8, colorspace 1.2-4, digest 0.6.4, evaluate 0.5.3, fail 1.2, foreach 1.4.2, formatR 0.10, gdata 2.13.3, gplots 2.13.0, grid 3.1.0, gtable 0.1.2, gtools 3.4.0, highr 0.3, igraph 0.7.1, iterators 1.0.7, labeling 0.2, lattice 0.20-29, munsell 0.4.2, pheatmap 0.7.7, plyr 1.8.1, proto 0.3-10, qvalue 1.38.0, reshape2 1.2.2, rtracklayer 1.24.0, scales 0.2.4, sendmailR 1.1-2, stats4 3.1.0, stringr 0.6.2, tcltk 3.1.0, tools 3.1.0, zlibbioc 1.10.0

References

[1] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, May 2012.

[2] Helen Pemberton, Emma Anderton, Harshil Patel, Sharon Brookes, Hollie Chandler, Richard Palermo, Julie Stock, Marc Rodriguez-Niedenfhr, Tomas Racek, Lucas de Breed, Aengus Stewart, Nik Matthews, and Gordon Peters. Genome-wide co-localization of polycomb orthologs and their effects on gene expression in human fibroblasts. 15(2):R23. PMID: 24485159.

[3] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. 25:25–29.

[4] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. 32:D277–D280. PMID: 14681412.

[5] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease ontology: a backbone for disease semantic integration. 40:D940–D946.

[6] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The reactome pathway knowledgebase. 42:D472–D477.