ChIP-seq, SSA and PWMTools: Advanced Tutorial

1. Applications to ChIP-seq data for transcription factors

1.1. Quality assessment of ChIP-seq data with 5'-3' correlation analysis

Background:

5'-3' correlation analysis introduced in part 1 of the Basic ChIP-Seq Tutorial. This type of analysis is not only useful for estimating the average fragment length of a ChIP-seq experiment. It is also a powerful graphical method to evaluate the quality of a ChIP-seq experiment as pointed out by ENCODE (Landt et al. 2012, Genome Res.).

The use of strand correlation analysis will be illustrated with data from ChIP-seq experiments targeted at 15 transcription factors in mouse embryonic stem cells.

```
Genome:     M. musculus (July 2007 NCBI37/mm9)
Data Type:  ChIP-seq
Series:     Chen 2008, ES cells, 16 transcription factors,...
Samples:    ES Nanog
            ES Smad1
            ES CTCF
```

The study is described in (Chen et al. Cell 2008) and data have been deposited in GEO entry GSE11431.

Step-by-step procedure:

1. Let's start with Nanog. Open the ChIP-Cor input form at

   http://ccg.vital-it.ch/chipseq/chip_cor.php

   and fill it out as shown below:

```
Input Data Reference Feature:        Input Data Target Feature:

x Select available Data Sets         x Select available Data Sets

  Genome: M. musculus (July 2007...    Genome: M. musculus (July 2007...
  Data Type: ChIP-Seq                  Data Type: ChIP-Seq
  Series: Chen 2008, ES cells ...      Series: Chen 2008, ES cells ...
  Sample: ES Nanog                     Sample: ES Nanog

Additional Input Data Options       Additional Input Data Options

  Strand: +                            Strand: -
  Centering: (blank)                   Centering: (blank)
  Repeat Masker: unchecked             Repeat Masker: unchecked

Analysis Parameters

  Input Range : -1000 to 1000

Histogram Parameters

  Window width: 10
  Count Cut-off value: 1
  Normalization: count density
```

2. Repeat the same analysis for Smad1 and CTCF and for the human breast cancer samples indicated above.

Results and Discussion.

The 5'-3' correlation plots for the three TFs assayed in mouse ES cells are shown in Figure 1.1.1.

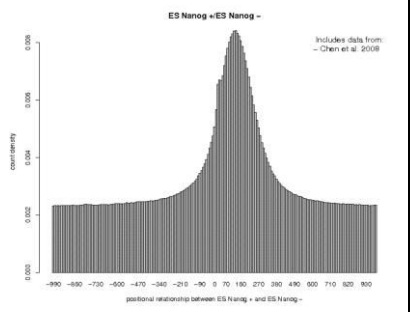
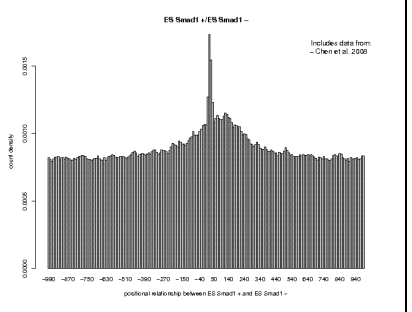
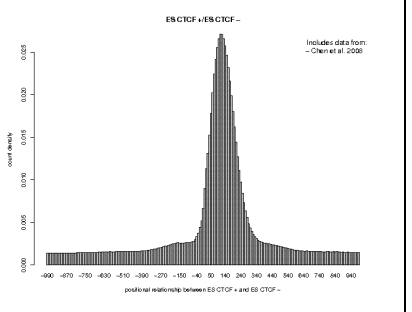
Nanog	Smad1	CTCF
		

Figure 1.1.1 5'-3 correlation plots for three ChIP-seq experiments targeted at Nanog, Smad1, and CTCF in mouse embryonic stem cells.

The main quality indicator is the signal-to-noise ratio as reflected by the peak height relative to the background signal observed in the peripheral parts of the plots. According to his criterion the quality ranking is CTCF > Nanog > Smad1. Poor quality data such as the Smad1 data often show an additional spike (thin peak) closer to position zero. This is an artifact. The spike position corresponds to the read length that has been used for mapping the ChIP-seq fragments to the genome. For Smad1, the spike is considerably higher than the true ChIP-seq peak. For Nanog, it is barely visible as a shoulder on the 5' side of the main peak. For CTCF it is completely hidden.

To look at a second example, we propose to repeat the above procedure with ChIP-seq data for estrogen receptor α (gene symbol ESR1) in ER+ breast cancer tissues, cell lines and needle biopsies:

```
Genome:     H.sapiens (Feb 2009/hg18)
Data Type:  ChIP-seq
Series:     Ross-Inness 2012, breast cancer, ER and FOXA1
Samples:    tumor CC135745(poor)ER
            tumor NT2433(good)ER
            cell-line MCF7 ER rep1

Genome:     H.sapiens (Feb 2009/hg19)
Data Type:  ChIP-seq
Series:     Zwart 2013, ERalpha ChIP-seq from breast cancer ...
Samples:    breast_cancer BIOPSY7367 ERalpha None
            breast_cancer BIOPSY7367 ERalpha mRNA+histone
            MCF7 ERalpha None_1
```

The first two samples were obtained from surgically removed tumor samples. MCF-7 is an ER+ breast cancer cell line. The samples from the second studies are taken from needle biopsies. The last sample in the list comprises a small number (10'000) of cultured MCF7 cells. mRNA and histone proteins are used as "carrier" to enhance the efficacy of ChIP-seq with small biosamples. The quality of the ChIP-seq data from these samples may be compromised by the cell type heterogeneity in the case of of breast tumors and by the small number of cells in the case of needle biopsies.

The ChIP-seq experiments from these studies are described in (Ross-Innes et al. Nature 2012) and (Zwart et al. BMC Genomics 2013). The source data have been deposited in GEO entry GSE322222 and ArrayExpress E-MTAB-1534. The results obtained from these data are shown in Figure 1.1.2.

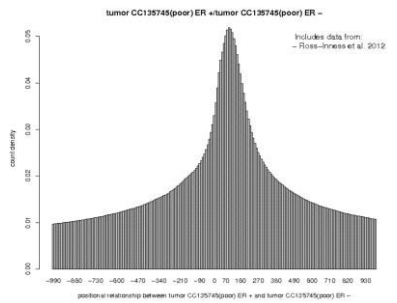
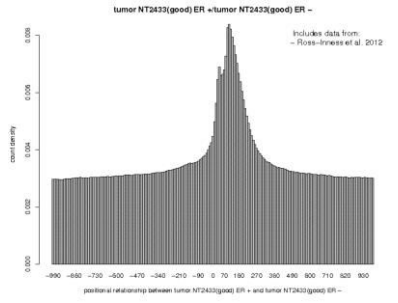
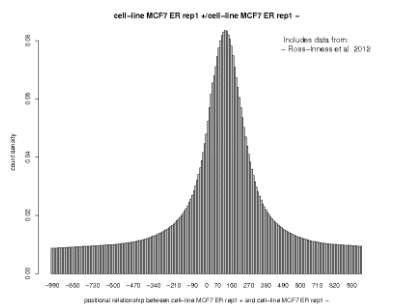
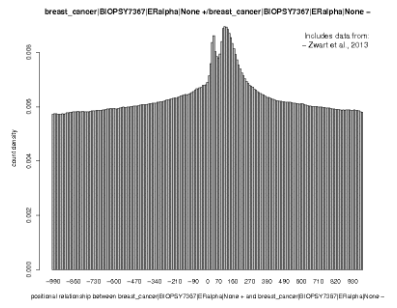
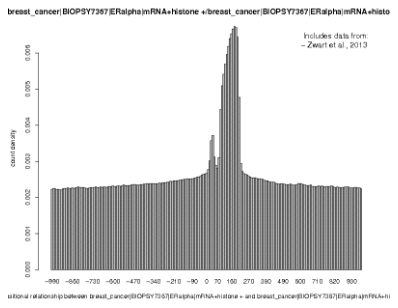
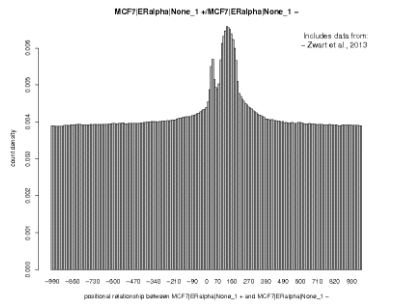
Breast tumor CC135745	Breast tumor NT2433	MCF7 cell line
		
Biopsy 7367 no carrier	Biopsy 7367 (mRNA+histone)	MCF7 (only 10'000 cells)
		

Figure 1.1.2 5'-3 correlation plots for ChIP-seq data for ER in breast cancer tissues and cell lines.

We notice a surprisingly large variation in quality for ChIP-seq data derived from surgically removed tumor tissue. For the needle biopsies, the carrier composed of mRNA and histone appears to be effective even though the shape of the main peak is atypical and possible indicative of some artifact.

1.2. Quality assessment of ChIP-seq peak lists via motif enrichment

Background:

Rather than assessing the quality of the primary read-mapping data from a ChIP-seq experiment, one could assess the quality of the derived peak lists by a motif enrichment test. For fair comparison, it is important that the same number of top-scoring peaks is analyzed, since larger peak lists tend to be less enriched in binding site motifs for the targeted TF:

To illustrate this procedure, we are going to use the same data as in part 1.1.1 of this Tutorial:

```
Genome:      M. musculus (July 2007 NCBI37/mm9)
Data Type:   ChIP-seq
Series:      Chen 2008, ES cells, 16 transcription factors,...
Samples:     ES Nanog
             ES Smad1
             ES CTCF

Genome:      H.sapiens (March 2006 NCBI36/hg18)
Data Type:   ChIP-seq
Series:      Ross-Inness 2012, breast cancer, ER and FOXA1
```

```
Samples:    tumor CC135745(poor)ER
            tumor NT2433(good)ER
            cell-line MCF7 ER rep1

Genome:     H.sapiens (Feb 2009 GRCh37/hg19)
Data Type:  ChIP-seq
Series:     Zwart 2013, ERalpha ChIP-seq from breast cancer ...
Samples:    breast_cancer|BIOPSY7367|ERalpha|None
            breast_cancer|BIOPSY7367|ERalpha|mRNA+histone
            MCF7|ERalpha|None_1
```

As TF binding site motifs, we recommend the following position weight matrices:

```
Motif Library: HOCOMOCO v10 Mouse TF Collection
Motifs:        NANOG_MOUSE.H10MO.A (length=17)
               SMAD1_MOUSE.H10MO.D (length=12)

Motif Library: JASPAR CORE 2014 vertbrates
Motifs:        CTCF MA0139.1(length=19)
               ESR1 MA0112.2 (Length=20)
```

Step-by-step procedure

Example mouse ES Nanog:

1. Go to the ChIP-Peak input form at:

   http://ccg.vital-it.ch/chipseq/chip_peak.php

   As input data, select the Nanog sample indicated above with centering distance 60. Use defaults for all the other parameters: Repeat Masker off, window 200, Vicinity range 200, Peak Threshold enrichment factor 10, Count Cut-off 1, Refine Peak Position checked. Submit.

2. Check the number of peaks on the results page. For Nanog you will get 28'076 peaks. Should get less than 1000 peaks with other data sets, repeat the analysis with a lower threshold until you get at least 1000 peaks. Then transfer the peak list to ChIP-Cor using the direct navigation button provided for this purpose.

3. On the ChIP-Cor input form, select the sample that was used for generating the peak list as Target Feature with the same centering distance. For the Reference Feature, make sure that the following parameter settings are specified: Strand any, Centering blank, Count Cut-off 1. Other parameters are not important at this stage. Submit.

4. On the ChIP-Cor results page, use the Enriched Feature Selection menu with the following inputs. From -100 To 100, Threshold 0, Cut-Off: 1, Depleted Feature Selection off, Ref Feature Oriented off, Top enriched Features 1000. Submit.

5. From the Feature Selection results page, transfer the list of the top 1000 Nanog peaks to OProf using the direct navigation button provided for this purpose. On the left site of the OProf input form, select the following parameters: 5'border -499, 3'border 500, window 100, shift 10, Search mode bidirectional. On the right side, select the Nanog PWM recommended above with Cut-off p-value 0.0001, Ref. position 9.

6. Repeat this analysis with the other samples and corresponding PWMs.

Results and Discussion:

Selected results are shown in Fig. 1.2.1.

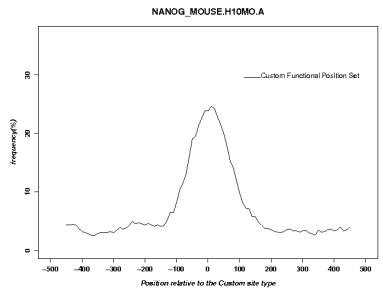
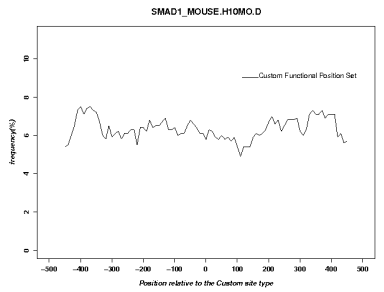
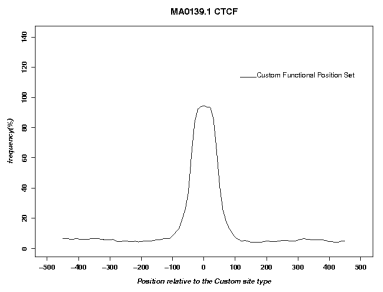
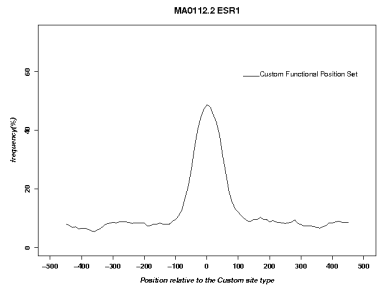
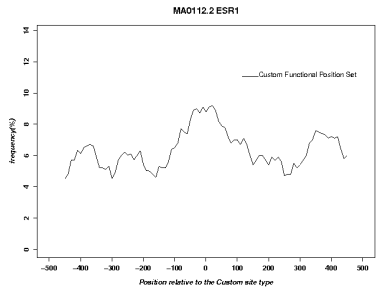
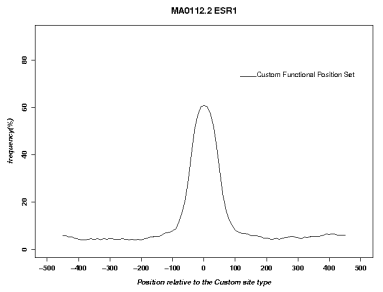
Mouse ES: Nanog	Mouse ES: Smad1	Mouse ES: CTCF
		
Breast tumor CC135745: ERα	Biopsy 7367 +carrier: ERα	MCF7 cell line
		

Figure 1.2.1. Peak list evaluation with motif occurrence profiles.

We see virtually no enrichment for the low quality ChIP-seq for Smad1. However, of you search this peak list with the Nanog motif, you will see a weak enrichment (Try it). The interpretation of this puzzling observation is not clear. Smad1 may indirectly bind to the chromosomal regions via Nanog. The example may serve as illustration that the biological processes that recruit TFs to their *in vivo* target sites may in some cases be complex, manifold and difficult to elucidate.

Also disappointing is the result for the needle biopsy treated with carrier. This confirms out suspicion that the unusual shape of main peak seen in a 5'-3' correlation plot reflects an artifact.

1.3. Characterizing the diversity of TF binding sites with aggregations plots (APs)

Background: *in vivo* binding sites of different transcription occur in different genomic contexts. Some tend to be associated with active chromatin marks, others with repressive ones. Moreover, they may differ in other respects, for instance with regard to their tendency to be conserved through evolution. In this exercise, we are going to generate feature correlation plots for ChIP-seq peak lists for 15 transcription factors in mouse embryonic (ES) cells. The peak lists have been publised by Chen et al. 2008 and are available from the ChIP-seq server menu under the followig headings :

```
Genome:      M. musculus (July 2007 NCBI37/mm9)
Data Type:   ChIP-seq-peak
Series:      Chen 2008,ES cells, 16 transcription...
Samples:     ES Nanog peaks (10343)
```

In addition we will use data from Mikkelsen et al. 2007:

```
Genome:      M. musculus (July 2007 NCBI37/mm9)
Data Type:   ChIP-seq
Series:      Mikkelsen 2007, histone modifications in mouse embryo...
Samples:     ES H3K4me3
             ES H3K27me3
             ...
```

Fig. 1.3.1 shows H3K4me3 profiles around 8 selected transcription factors from this data series. Note the big differences in signal intensities : c-Myc, Zfx, Klf4, and Suz12 have high peaks whereas classical pluripotend stem cell transcription factors including Sox2, Nanog and Smad1, as well as Ctcf, show almost flat curves at the bottom of theFigure.

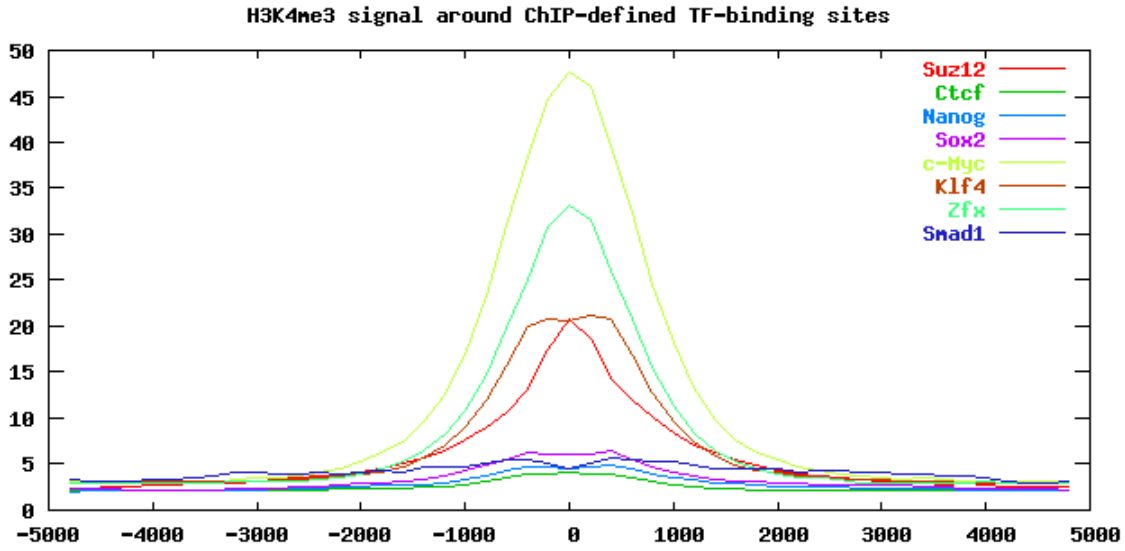


Figure 1.3. H3K4me3 levels around TF binding peaks in mouse ES cells.

Step-by-step procedure

Go to the ChIP-Cor input form at:

http://ccg.vital-it.ch/chipseq/chip_cor.php

select as reference feature the Nanog peak list from the above indicated data series and as reference feature the ES H3K4me3 sample Mikkelsen et al. 2007.

Additional parameters : for reference feature strand any and centering (blank), for target feature strand any and centering 100, Repeat Masker off on both sides, beginning -5000, end 5000, window 100, normalization global. Then save the output file under then name

nanog_h3k4me3.txt

repeat the same procedure for the other transcription factors appearing in Fig. 3.1.1. and save the output files under the names :

sox2_h3k4me3.txt
smad1_h3k4me3.txt
ctcf_h3k4me3.txt
zfx_h3k4me3.txt
klf4_h3k4me3.txt
c-myc_h3k4me3.tx
suz12_h3k4me3.txt

You can now regenerate Figure 1.3.1 using the R code below:

```
x=read.table("suz12_h3k4me3.txt")[,1]

data=matrix(0,nrow=8,ncol=length(x)); names=rep("",8)
data[1,]=read.table("suz12_h3k4me3.txt")[,2]; names[1]="Suz12"
data[2,]=read.table("ctcf_h3k4me3.txt")[,2] ; names[2]="Ctcf"
data[3,]=read.table("nanog_h3k4me3.txt")[,2]; names[3]="Nanog"
data[4,]=read.table("sox2_h3k4me3.txt")[,2] ; names[4]="Sox2"
```

```
data[5,]=read.table("c-myc_h3k4me3.txt")[,2]; names[5]="c-Myc"
data[6,]=read.table("klf4_h3k4me3.txt")[,2] ; names[6]="Klf4"
data[7,]=read.table("zfx_h3k4me3.txt")[,2]  ; names[7]="Zfx"
data[8,]=read.table("smad1_h3k4me3.txt")[,2]; names[8]="Smad1"

plot(x,data[1,], type="l", ylim=c(0,50),
xlab="Distance to peak center", ylab="Fold enrichment",
main="", lwd=2, col=1, xaxt="n", cex.axis=1.2, cex.lab=1.2)
for(i in 2:8) {points(x,data[i,], lwd=2, type="l", col=i)}

axis(1, at=seq(-5000,5000,2000), cex.axis=1.2)

legend("topright",legend=names,col=1:8, lwd=rep(2,8), lty=rep(1,8))
```

What next:

- Make similar plots for other histone marks
- Make similar plots for conservation scores
- Analyze TF binding peaks in other cell types

2. Applications to histone modifications

2.1. Strand correlation and autocorrelation analysis

(in preparation)

2.2. Using ChIP-Part for defining enriched regions of variable size

The application ChIP-Part finds ChIP-seq signal enriched regions of variable size. Similar programs are sometimes referred as "broad peak" finder in the bioinformatics literature. Using ChIP-Part rather than ChIP-Peak is indicated if there is reason to believe that the enriched regions are relatively large and of variable. This is usually the case for histone variants and histone modifications.

ChIP-Part uses a segmentation algorithm which splits the genome into an alternating series of signal-enriched and signal-depleted regions. The process is guided by two parameters:

1. count density ***threshold***: regions with count density > ***threshold*** will be considered "enriched", others will be considered depleted.
2. Transition ***penalty***: a negative weight that is applied whenever a transition between the two types of domains occurs. The function of this parameter is to prevent excessive fragmentation of the genome into very small regions,

ChIP-Part generates the optimal genome segmentation by maximizing the following score:

*total-enriched-region-length x (average-enriched-region-count-density – **threshold**)*
*+ total-depleted-region-length x (**threshold** – average-depleted-region-count-density)*
*– number-of-transitions x **penalty***

We will use the ChIP-Part tool to compare H3K4me3 chromatin domains in four different mouse cell types: embryonic stem cells and cell lines (ES and ESHyb), embryonic fibroblasts (MEF), and neural progenitors (NP), using the following data sets:

```
Genome:     M. musculus (July 2007 NCBI37/mm9)
Data Type:  ChIP-seq
Series:     Mikkelsen 2007, histone modifications in mouse ...
Samples:    ES H3K4me3
            EShyb H3K4me3
            MEF H3K4me3
            NP H3K4me3
```

Step-by-step instructions (`ES H3K4me3`):

1. Go to ChIP-part at:

   http://ccg.vital-it.ch/chipseq/chip_part.php

   On the left side of the input form under "ChIP-seq Input Data" select the sample ES H3K4me3 from the above indicated series. Further below under "Additional Input Data Options" select strand any, centering 100, Repeat Masker off. On the right site under "Partitioning Parameters" enter Threshold / relative enrichment factor 2, count cut-off 1, Breaking Cost counts checked 10, DNA length equivalent checked 1000. Further below under "Genome Viewing Parameters" enter BED Track Name checked ES.H3K4me3. Submit.

   Note: The ChIP-Part web application allows you to specify the partitioning parameters in absolute (count density, counts) or relative terms (relative enrichment factor, DNA length equivalent). The DNA length equivalent is the minimal length of a count-free region that breaks an enriched domain into two parts. The server automatically converts the corresponding value into an absolute penalty. If both types of breaking costs are checked then ChIP-part will take the higher one of the two penalty values.

2. On the ChIP-Part output page, click on the link "Partitioning Statistics". This will open a page showing various statistics of the results. You will see that there are 28031 enriched fragments with an average length of 2883 bp. Now, click on the link "USCS View" to navigate to the UCSC genome browser. Enter genome positions chr6:122,650,000-122,800,000 into the text area provided for this purpose. Keep this browser window open during the next steps.

3. On the ChIP-Seq Web server, we provide custom tracks for viewing selected ChIP seq samples in the UCSC genome browser. Go to

   http://ccg.vital-it.ch/chipseq/chip_seq_wig.html

   and copy the URL corresponding to the ES H3K4me3 sample into the edit buffer of your browser window. Then return to the open UCSC genome browser window and click on the action button "manage custom tracks" under the graphical display. A new window will open. Click on "add custom track" and then paste the URL of the ES H3K4me3 custom track into the text area provided for this purpose. Submit. On the next page, use the "go" button to return to standard genome browser view. You will now see the enriched domain defined by ChIP-Part together with a track showing the corresponding ChIP-seq tag density profile.

4. Repeat step 1-3 for the other three ChIP-seq samples indicated above. Upload all custom track files produced by ChIP_Part to the UCSC genome browser along with corresponding wiggle files for the input data provided by the ChIP-seq server.

Results and Discussion:

Fig. 2.2.1 shows the H3K4me3-enriched domains found by ChIP-Part for different mouse cell types along with the corresponding input data (ChIP-seq tag density profiles).

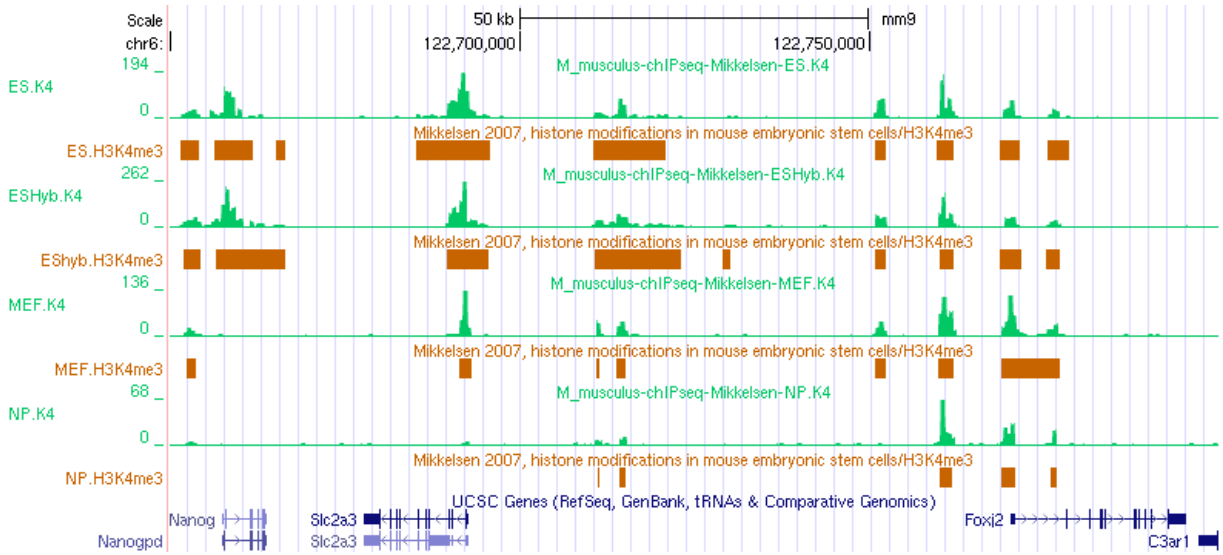


Figure 2.2.1. H3K4me3 enriched domains found by ChIP-Part for 4 different cell types (brown) together with the corresponding input data (green) in a genomic region including the Nanog gene.

We first note that ChIP-part partitions the gnome in a way that is consistent with human intuition. The genome snapshot is taken from a genomic region which includes Nanog, a hallmark gene of embryonic stem cells. We note that the promoter region of this gene is marked by H3K4me3 only in the two ES samples. An enriched region further downstream disappears in neural progenitor cells but not in mouse embryonic fibroblasts. Other enriched domains persist in all analyzed cell types.

What next ?

Repeat this analysis with other histone marks from the same series. H3K36me3 is an interesting example as this histone mark reportedly associates with transcribed region (gene bodies).

2.3. Analyzing chromatin domain boundaries by ChIP-part

The ChIP-Part tool can be used to define chromatin domain boundaries and to investigate correlated features. The zinc finger protein CTCF has been reported to bind to such domain boundaries and to prevent propagation of some histone modifications across them. To test this hypothesis, we will use ChIP-Part to delineate chromatin domains enriched in the histone variant H2AZ and then look at the distribution of CTCF ChIP-seq tags and peaks around the 5' boundaries of such domains. We are going to use the following data sets from Barski et al. 2007.

```
Genome:      H.sapiens (Feb 2009/hg19)
Data Type:   ChIP-seq
Series:      Barski 2007, CD4+ cells, histone marks ...
   Sample:   CD4+ H2AZ
   Sample:   CD4+ CTCF
```

Step-by-step instructions:

- Open an R terminal window.
- Go to ChIP-Part at:

  http://ccg.vital-it.ch/chipseq/chip_part.php

  and fill out the input form as follow. As Reference Feature select the H2AZ sample indicated above, Strand any, Centering 60, Repeat Masker off, Threshold: relative

enrichment factor 2, Breaking Cost: counts checked 10, Breaking Cost: DNA length equivalent (bb) checked 1000, BED Track Name checked "barsk07_h2az_domains" Chromosomal regions: unchecked.

- On the ChIP-part output page, click on "Partitioning Statistics" and have a look at the numbers. Then right-click on the "ChIP-Cor" navigation button in the center of the page. This will send an oriented SGA file to ChIP-cor, in which genome positions with strand indication "+" and "–" mark the beginning and end of enriched domains, respectively. The output file resulting from this step is accessible via the link below.

  h2az_domains.sga

- We first look at the distribution of H2AZ near the beginning of H2AZ-enriched domains. Fill out the ChIP-cor input form as follows: On the left side, leave the Reference Feature Input Section as is, Strand oriented, Centering blank, Repeat Masker off, Range from -1000, to 1000, Window Width 10, Count Cut-off 1, Normalization global. On the right side, select the H2AZ sample indicated above as Target Feature, Strand any, Centering 40, Repeat Masker off.

- In the figure displayed on the output page you will see a very high spike at position zero. This is normal, since ChIP-part always puts the beginning of enriched domains at positions occupied by at least one ChIP-seq tag. Otherwise, the distribution looks as expected, with low signal in the upstream region and a more or less constant high signal in the downstream part. Import the results from this analysis into R by right-clicking on the hyperlink labelled "TEXT" and using the "Copy Link Location" mechanism to paste the URL into the R command line:

  ```
  h2hz=read.table("http://ccg.vital-it.ch/...")
  ```

  or save the file under the name:

  h2az_at_h2az_boundary.txt

- Go back to the ChIP-Cor input page und repeat the analysis with CTCF as target feature (centering 40). In the figure displayed on the output page you will see a major CTCF peak at approximately +250 flanked by minor peaks on both sides. Import the results into R using the same mechanism as before

  ```
  ctcf=read.table("http://ccg.vital-it.ch/...")
  ```

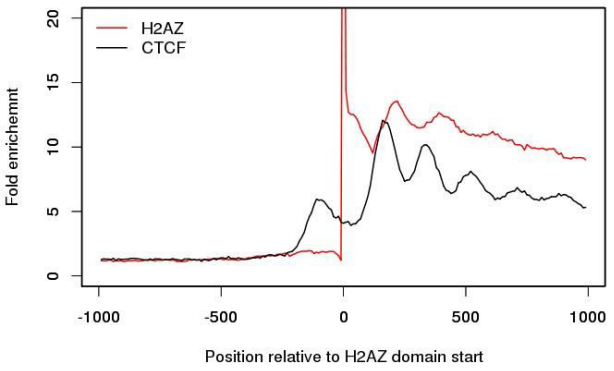
  or save the file under the name:

  ctcf_at_h2az_boundary.txt

- Make a Figure showing the distribution of the H2AZ and CTCF signal near the beginning of H2AZ-enriched domains. You may use the R code given next to Figure 2.3.1.

Results and Discussion

There are 32837 H2AZ-enriched domains of an average length of 2965 bp. The distribution of the H2AZ and CTCF signal around H2AZ domain boundaries is shown in Fig. 2.3.1.




```
ylim=c(0,20)
xlab="Position relative to H2AZ
domain start"
ylab="Fold enrichment"

plot(h2az,t="l", col="red",
ylim=ylim, xlab=xlab, ylab=ylab)
par(new=T)

plot(ctcf,t="l", col="black",
ylim=ylim, xlab="", ylab="")
legend(x="topleft",
legend=c("H2AZ","CTCF"),
col=c("red","black"),lty=c(1,1),
bty="n")
```

Figure 2.3.1 Distribution of the H2AZ and CTCF signal relative to the 5'end of H2AZ-enriched chromatin domains. The R code for generating the Figure is given on the right side,

We note a periodic distribution of both H2AZ and CTCF, presumably reflecting positioned nucleosomes. The start sites of H2AZ-enriched domains defined by ChIP-Part appear to be preferentially located between the 5' end and the center of a nucleosome. The CTCF signal shows a maximum at about 200 bp downstream of the 5' boundary of H2AZ enriched domains. The results should be interpreted with caution as the chromatin boundaries may be intrinsically fuzzy (variable between individual cells) or statistically unstable due to low tag count numbers.

What next?

You may repeat this analysis with other histone modifications from the same series, *e.g.* H3K36me3 or H3K27me3, or with human cell line data from ENCODE, *e.g.*

```
Genome:     H.sapiens (Feb 2009/hg19)
Data Type:  ChIP-seq
Series:     GSE29611, Histone Modifications by ChIP-seq
   Sample:  GM12878 H2A.Z
   Sample:  GM12878 CTCF
```

2.4. Quality assessment of histone modification data by gene set enrichment analysis

(in preparation)

2.5. Identifying differentially modified chromatin regions

(in preparation)

3. Data viualization

3.1. Exploring data heterogeneity with ordered and clustered heatmaps

Aggregation plots such as produced by ChIP-cor often hide data heterogeneity. In fact, no single genomic region may resemble the consensus pattern. If we see two peaks in an aggregation plot, each peak my come from a distinct sub-population.In this exercise you will learn how to uncover and characterize heterogeneneity of genomic regions by means of ordered and clustered heatmaps.

Step-by-step instructions :

As input data you will use H3K4me3 and nucleosome profiles of human promoters from CD4+ T cells. The count tables files which you will import into R can be found here :

hs_epd04_h3k4me3.txt, hs_epd04_nucleosomes.txt

Download these files to your computer or generate them yourself with

ChIP-Extract

Select as reference feature :

```
Genome
Data type: Genome ann
Series :   EPDnew, the Human Curated Promoter Database
Sample :   TSS from hg18 EPDnew rel 004
```

Additional input parameters : Strand oriented, centering (blank), Repeat Masker off, Beginning -1000, End 1000, window width 20, count cut-off 10.

Select as target feature :

```
Genome:    H. sapiens (March 2006 NCBI36/hg18)
Data type: ChIP-seq
Series:    Barski 2007, CD4+ cells, histone marks
Sample:    CD4+ T H3K4me3 (MNase)
```

with options strand any, centering 70, Repeat Masker off, Heatmap ordering off .

Save the table output file under the name hs_epd04_h3k4me3.txt.

Then repeat the same procedure with the following target feature :

```
Genome     H. sapiens (March 2006 NCBI36/hg18)
Data type: DNase FAIRE etc.
Series:    Schones 2008, CD4+ cells, nucleosome profiling, GSE10437
Sample:    CD4+ Resting Nucleosomes
```

And save the table output file under the name hs_epd04_nucleosomes.txt.

The reaming part of the exercies is based on R. Open R in a terminal window, then read in the H3K4me3 table file :

```
fname="hs_epd04_h3k4me3.txt"
data <- as.matrix(read.table(fname, check.names=F))
N=dim(data)[1]; x=as.numeric(colnames(data));
```

a) Plotting the heatmap with the original row ordering

```
order=1:dim(data)[1]
main="Original ordering"
```

To speed up plotting we define a function to condense the matrix :

```
condense.matrix <- function(mat, rows) {
   groups <- sort(rep(1:ceiling(dim(mat)[1]/rows), rows))[1:dim(mat)[1]]
   c.mat <- matrix(ncol=dim(mat)[2], nrow=max(groups))
   for (I in 1:dim(mat)[2]){
      c.mat[,I] <- tapply(mat[,I], groups, mean, na.rm=T)
      }
   return(c.mat)
   }
```

We condense the matrix to a maximum of 1000 rows :

```
w=ceiling(dim(data)[1]/1000)
z <- condense.matrix(data[order,], w)
```

To increase color intensity for weak signals, we reduce the very high matrix elements to the 0.99 quantile value :

```
q=0.99; max=sort(z)[q*dim(z)[1]*dim(z)[2]]
for(i in 1:dim(z)[1]) {for (j in 1:dim(z)[2]) {z[i,j]=min(z[i,j],max)}}
```

We are ready to plot :

```
color  <- colorRampPalette(c("white", "red"), space = "rgb")(100)
image(x,z=t(z), xaxt="n", col=color, yaxt="n", bty="n",
xlab="Distance from TSS"); title(main=main); axis(1)
```

b) Plotting the heatmap with different row orderings

Let's first define a function for plotting ordered, condensed heatmaps :

```
plot.matrix <- function(x, mat, order, rows, q, color, xlab, main) {
   w=ceiling(dim(mat)[1]/rows)
   z <- condense.matrix(mat[order,], w)
   max=ceiling(sort(z)[q*dim(z)[1]*dim(z)[2]])
   for(i in 1:dim(z)[1]) {for (j in 1:dim(z)[2]) {z[i,j]=min(z[i,j],max)}}
   image(x,z=t(z), xaxt="n", col=color, yaxt="n", bty="n", xlab=xlab)
   title(main=main); axis(1)
   }
```

Order the rows in different ways

```
main="By consensus pattern"
order=order(cov(t(data),colMeans(data)))
plot.matrix(x, data, order, 1000, 0.99, color, "Distance from TSS", main)

main="By overall intensity"
order=order(rowSums(data))
plot.matrix(x, data, order, 1000, 0.99, color, "Distance from TSS", main)

main="Left to right"
order=order(((dim(data)[2]:1)-dim(data)[2]/2-0.5) %*% t(data))
plot.matrix(x, data, order, 1000, 0.99, color, "Distance from TSS", main)

# Find promoter subclasses with K-means
```

c) Finding sublcasses of promoters with K-means clustering

We try 10 subclasses :

```
main="K-means clustering";
k=10; km=kmeans(data,k); order=order(km$cluster)
```

To display class boundaries, we define a new plotting function :

```
plot.matrix.2 <- function(x, mat, order, k, rows, q, color, xlab, main) {
   w=ceiling(dim(mat)[1]/rows)
   z <- condense.matrix(mat[order,], w)
   max=ceiling(sort(z)[q*dim(z)[1]*dim(z)[2]])
   for(i in 1:dim(z)[1]) {for (j in 1:dim(z)[2]) {z[i,j]=min(z[i,j],max)}}
   image(x,z=t(z), xaxt="n", col=color, yaxt="n", bty="n", xlab=xlab)
   title(main=main); axis(1)
   if(k == 1) {h=1} else
      {h=km$size[1]; for (i in 2:k) {h=c(h,h[i-1]+km$size[i])}}
   abline(h=c(0,h/h[k]), v=c(max(x),min(x)), col="black")
   }
```

Now let's plot the heatmap ordered by sub-classes :

```
plot.matrix.2(x, data, order, k, 1000, 0.99, color, "Distance from TSS",
main)
```

K-means is a stochastic algorithm producing slightly different results each time you run it. To test this behavior, run it again:

```
k=10; km=kmeans(data,k); order=order(km$cluster)
dev.new() # to open a new graphics window
plot.matrix.2(x, data, order, k, 1000, 0.99, color, "Distance from TSS",
main)
```

You will see that the classes found are more or less the same, but the order in which they appear has changed.

d) Correlation of H3K4me3 signal with nucleosome density

How does the H3K4me3 signal correlate with other genomic features, e.g. nucleosome organisation. Do the promoters which have low levels of H3K4me3, nevertheless have organized nucleosomes?

We first import the nucleosome data :

```
fname="hs_epd04_nucleosomes.txt"
data2 <- as.matrix(read.table(fname, check.names=F))
N=dim(data)[1]; x2=as.numeric(colnames(data))
```

Then we define a new color palette

```
color2 <- colorRampPalette(c("white", "darkgreen"), space = "rgb")(100)
```

… adnd a new plotting function for plotting two heatmaps side by side

```
plot.2matrices <- function(
   x, mat, order, k, rows, q, color, xlab, main,
   x2, mat2, color2, main2) {

   layout(matrix(c(1,2), nrow=1, ncol=2), widths=c(2.5,2.5), heights=c(5.0))
   par(mar=c(3,4,4,1), oma=c(0,0,2,0))

   w=ceiling(dim(mat)[1]/rows)
   z <- condense.matrix(data[order,], w)
   max=ceiling(sort(z)[q*dim(z)[1]*dim(z)[2]])
   for(i in 1:dim(z)[1]) {for (j in 1:dim(z)[2]) {z[i,j]=min(z[i,j],max)}}
   image(x,z=t(z), xaxt="n", col=color, yaxt="n", bty="n", xlab=xlab)
   title(main=main); axis(1)
   if(k == 1) {h=1} else
      {h=km$size[1]; for (i in 2:k) {h=c(h,h[i-1]+km$size[i])}}
   abline(h=c(0,h/h[k]), v=c(min(x),max(x)), col="black")

   z <- condense.matrix(mat2[order,], w)
   max=ceiling(sort(z)[q*dim(z)[1]*dim(z)[2]])
   for(i in 1:dim(z)[1]) {for (j in 1:dim(z)[2]) {z[i,j]=min(z[i,j],max)}}
   image(x2,z=t(z), xaxt="n", col=color2, yaxt="n", bty="n", xlab=xlab)
   title(main=main2); axis(1)
   if(k == 1) {h=1} else
      {h=km$size[1]; for (i in 2:k) {h=c(h,h[i-1]+km$size[i])}}
   abline(h=c(0,h/h[k]), v=c(min(x2),max(x2)), col="black")
   }
```

First, let's order the rows by overall intensity

```
k=1; order=order(rowSums(data))

main="H3K4me3"; main2="Nucleosomes"
plot.2matrices(
   x, data, order, k, 1000, 0.99, color, "Distance from TSS", main,
   x2, data2, color2, main2)
```

You will see that promoters with low H3K4me3 signal have a more diffuse nucleosome organization.

Now let's do clustering wihthK-means :

```
k=10; km=kmeans(data,k); order=order(km$cluster)
plot.2matrices(
   x, data, order, k, 1000, 0.99, color, "Distance from TSS", main,
   x2, data2, color2, main2)
```

What next :

- Correlate the H3K4me3 signal with other histone marks from the Barski 2007 series.
- Using the CTCF data from the same series, find peaks and extract ChIP-seq data for H2AZ in a window of 2000 bp around the peak centers. Then explore the heterogeneity of genomic regions containing CTCF sites in the same way as exemplified above
- Apply the same approach to the heatmaps generated under exercise 7.1 of the Basic ChIP-Seq and SSA Tutorial.

4. Motif-oriented applications

4.1. High-resolution DNase I footprints of occupied TF binding sites

(in preparation)

4.2. Motif association and spacing analysis

Background:

In *in vivo* occupied STAT1 sites often occur as pairs with a center-to-center spacing of about 20-24 base pars, see (Schmid & Bucher 2010) and part 3 of our Motif tutorial. A search for centrally enriched motifs using CentriMo from the MEME suite further indicated that STAT1 peaks often are enriched in binding site motifs of the ETS and AP1 families of transcription factors. The enrichment of the ETS-like motifs may simply be a consequence of their close resemblance to the STAT1 motif.

In this part of the tutorial, we analyze the co-localization and spacing preferences of STAT1 and AP1 binding motifs in STAT1 peak regions. As in previous examples, we use the following input data:

```
Genome:      H.sapiens (Feb 2009/hg19)
Data Type:   ENCODE ChIP-seq
Series:      Robertson 2007, HeLa S3 cells, Genome-wide STAT1 ...
   Sample:   HeLa S3 Stat1 stim
```

We will first compile a list of 5000 highly occupied STAT1 motifs by first scanning the entire human genome with the JASPAR STAT1 matrix at a permissive P-value threshold and then selecting those motif matches with high STAT1 ChIP-seq tag coverage. To avoid biases by repetitive elements, we will work with repeat masked peak lists.

Step-by-step instructions:

1. Open the PWMScan input form at:

   http://ccg.vital-it.ch/pwmtools/pwmscan.php

   On the left side under “Genome Assemblies” select Home sapiens GRCh37 (hg19). On the right side under “PWMs from Library” select Motif Library JASPAR CORE 2014 vertebrates, Motif MA0137.3 STAT1, Cut-off P-value 0.0001, background base composition 0.29,0.21,0.21,0.29, Search strand both, Reference Position 6, Non-overlapping matches checked. There are 839’152 hits.

2. Transfer the PWMScan match list to ChIP-Cor via the direct navigation button. Use the STAT1 motif match list as reference feature. Select strand option oriented and Repeat Masker

checked. As target feature, use the STAT1 ChIP-seq data indicated above with centering 75 and count cut-off 1. Other parameters are not important at this stage. Run ChIP-Cor.

3. On the ChIP-Cor results page, use the "Enriched Feature Selection Menu" with the following settings: From -100 To 100. Threshold 10, Cut-Off 1, Depleted Feature Selection off, Ref Feature Oriented on, Select Top Enriched/Depleted Features 5000. Submit.

4. Next, analyze the peak sequences with CentriMo. To this end, we first have to extract sequences adjacent to the STAT1 motifs from the list generated by step 3. On the Feature Selection results page, use the "Sequence Extraction" menu to extract sequences from 10 to 250. Submit. On the following page, save the sequence file to disk under the name:

   stat1_sites_downstream.seq

   Incidentally, you may also save the corresponding SGA file as:

   stat1_sites_5k.sga

   The purpose of selecting downstream regions is to exclude the STAT1 motif matches at position 0. In doing so, we prevent CentriMo from reporting the STAT1 motifs at positon zero as centrally enriched plus similar motifs overlapping with the STAT1 motif. (We could of course also select an upstream region instead of a downstream region.)

5. Now go to the CentriMo input form at one of the following locations.

   http://meme-suite.org/tools/centrimo

   http://alternate.meme-suite.org/tools/centrimo

   Near the top of the page, choose local enrichment Anywhere rather than Central. This is important because we are going to analyze sequences immediately downstream of a STAT1 site, and consequently expect associated motifs to be enriched near the 5'end rather than the center of the input sequences. Upload the previously saved sequence file and press the "Start Search" button.

6. A screenshot of the CentriMo output page is shown below:

☒	ID	Name	*E*-value	Region Center	Region Width
■	MA0137.3	STAT1	2.9e-86	-110.5	4
□	MA0518.1	Stat4	3.9e-73	-110.5	7
□	MA0144.2	STAT3	4.3e-64	-110	5
□	MA0519.1	Stat5a::Stat5b	9.8e-41	-110.5	6
■	MA0490.1	JUNB	4.4e-35	-76.5	78
□	MA0477.1	FOSL1	6.7e-33	-77	77

   The full CentriMo output page is pasted here:

   http://ccg.vital-it.ch/chipseq/new/adv/stat1_centrimo_downstream.html

   We notice that STAT1 and AP1-like motifs (JUNB, FOSL1) appear at the top of the list. The discovery of STAT1 as an enriched motif is all but trivial and constitutes an independent confirmation that *in vivo* occupied STAT1 sites often occur as pairs.

7. Next we will investigate the positional distribution of AP1 –like motifs downstream of occupied STAT1 motifs in more detail. To this end, use the backwards button of your internet browser to go to the Enriched Feature Selection output page generated by step 3. From there, send the list of the 5000 most occupied STAT1 motifs to OProf using the directive navigation button provided for this purpose. Fill out the OProf input form as follows. SSA Input Data: leave as is except Feature STAT1, Additional Input Data Options: Sequence range 5'border -500 3'border 500, window 25, shift 5, search mode bidirectional. Signal Description: Motif

library JASPAR CORE 2014 vertebrates, motif JUNB MA0490.1, cut-off P-value 0.0002, ref. pos. 7. The resulting motif occurrence profile is shown in Figure 3.2.1.

8. To analyze the spacing between STAT1 and JUNB at higher resolution, return to the OProf input form and change the following parameters: 5'border 0, 3'border 100, window 15, shift 1, search mode forward. Open now an R in a terminal window and import the numbers underlying the motif occurrence profile into R by right-clicking on the link "Text 1" and selecting "Copy Link Location" to copy-paste the URL into the following R command:

```
junb=read.table("http://ccg.vital-it.ch/...")
```

Alternatively, you may save the numerical data file to disk under the name:

spacing_stat1_junb.txt

Two compare the spacing between STAT1 and JUNB with the spacing between pairs of STAT1 sites, repeat the same analysis with motif STAT1 MA0137.3 from JASPAR Core. Leave all other parameters as is except ref. pos. 6. The reference positions are chosen such that they correspond to the symmetry center of the respective motifs, which are both near-palindromic. From the OProf results page, import the numerical data via URL into R:

```
stat1=read.table("http://ccg.vital-it.ch/...")
```

Alternatively, you may save the numerical data file to disk under the name:

spacing_stat1_stat1.txt

Results and Discussion

In part 6 of the Basic ChIP-seq tutorial, CentriMo is used to search for centrally enriched motifs in 500 bp long sequences symmetrically positioned around STAT1 ChIP-seq peak centers . There, many more motifs were found including a large number of ETS-like motifs. The absence of ETS-like motif confirms the hypothesis that the enrichment of these motifs is due to their resemblance to the STAT1 motif and results from exact co-localization.

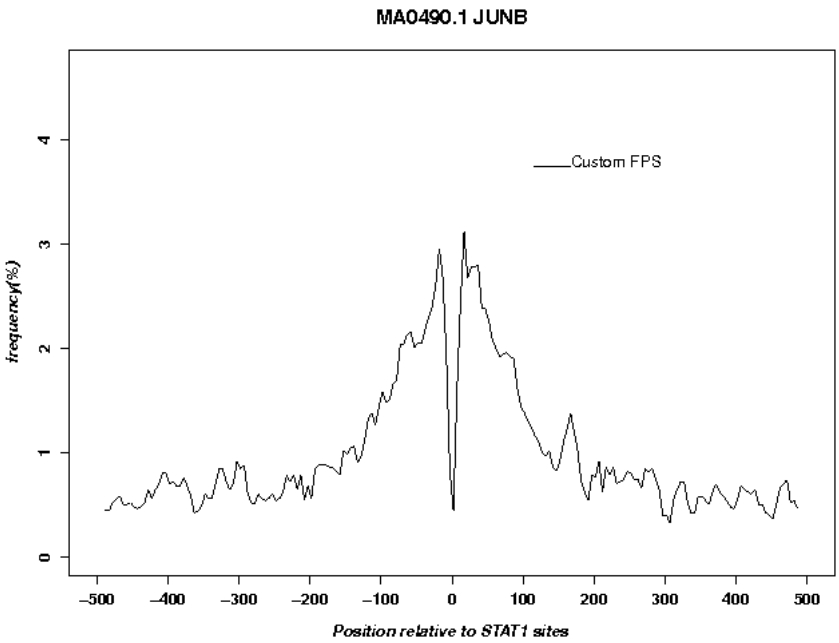
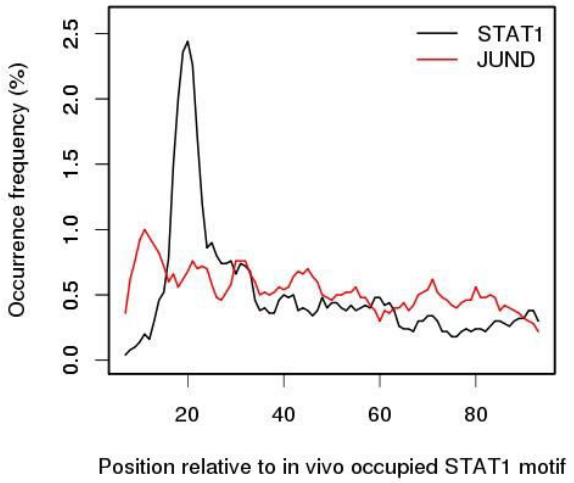


Figure 4.2.1 Positional distribution of JUNB motifs in the vicinity of *in vivo* bound STAT1 motifs.

In the analysis proposed here, we used STAT1 motif matches with high ChIP-seq tag coverage instead of fuzzily defined ChIP-seq peak centers as reference points. Moreover we submitted sequences immediately adjacent but not including the motif matches themselves to CentriMo. Furthermore, we excluded sequences from annotated repeat regions. In this new analysis, only motifs corresponding to

the STAT1 and AP1 families were found. A medium resolution plot of JUNB motifs around STAT1 motifs (Fig. 4.2.1.) shows over-representation of JUNB motifs within about 100 bp around the bound STAT1 motif. Note the deep valley at position zero, indicating that there are virtually no JUNB motifs directly overlapping with STAT1 motifs.




```
stat1=read.table(...)
junb =read.table(...)

par(lwd=2)
ylim=c(0,2.6)
xlab=paste("Position relative to",
   "in vivo occupied STAT1 motif")
ylab="Occurrence frequency (%)"
plot(stat1,t="l", col="black",
   ylim=ylim, xlab=xlab, ylab=ylab)
par(new=T)
plot(junb,t="l", col="red",
ylim=ylim,
   xlab="", ylab="",xaxt="n")
legend(x="topright",
   legend=c("STAT1","JUND"),
bty="n",
   col=c("black","red"),lty=c(1,1))
```

Figure 4.2.2 High-resolution positional distribution of STAT1 and JUNB motifs downstream of *in vivo* occupied STAT1 motifs. The R code for making this Figure is given on the left side.

The results from the motif spacing analysis at higher resolution (step 8) are shown in Fig, 4.2.2. We note a sharp peak at about position 21 for STAT1. The distribution of JUNB motifs is much broader and shows weak 10 bp periodicity in the STAT1 proximal region, the first maximum being located at pos. 12.

What next?

Interesting examples for motif association and motif spacing analysis are:

- Ptf1a peaks in amouse pancreatic cell line

```
Genome:     M. musculus (July 2007 NCBI37/mm9)
Data Type:  ChIP-seq
Series:     Thompson 2012, Ptf1a in 266-6 pancreatic cell line, ...
```

- RUNX1 peaks in AML

```
Genome:     H.sapiens (March 2006 NCBI36/hg18)
Data Type:  ChIP-seq
Series:     Ben-Ami 2013, AML1 (Kazumi -1) cells, ChIP Seq for …
```

- ER peaks in breast tumors

```
Genome:     H.sapiens (March 2006 NCBI36/hg18)
Data Type:  ChIP-seq
Series:     Ross-Inness 2012, breast cancer, ER and FOXA1
```

4.3 Performance evaluation of PWMs with ChIP-seq data

Background:

For many transcription factors, there are several PWMs available on our server. In some cases, these matrices correspond to primary and secondary motifs extracted by a motif discovery programs from high-throughput data.

PWMEval from the PWMTools server is a tool for evaluating the performance of a PWM in terms of predicting ChIP-seq peaks. The server uses a method introduced by (Orenstein and Shamir, Nucleic Acids Res. 2014). Input to the procedure is:

1. a scored (ranked) ChIP-Seq peak list and
2. a PWM represented as base probability matrix

The top *N* peaks are selected as positive examples. Genomic regions of the same size but located at some user-specified distance downstream or upstream of the peak centers are selected as negative examples. All sequences are then scored with the base probability matrix using the following formula:

$$Score(s,\Theta) = \sum_{t=0}^{|s|-k} \left[\prod_{i=1}^{k} \frac{\Theta_i(s_{t+i})}{q(s_{t+i})} + \prod_{i=1}^{k}{}_i \frac{\overline{\Theta}_i(s_{t+i})}{q(s_{t+1})} \right]$$

Here, s denotes the sequence, q the background base composition and Θ the base probability matrix ($\overline{\Theta}$ is the reverse complementary matrix). The performance of this score in distinguishing positive from negative examples is then assessed by a ROC curve and the area under the curve (AUC) is reported as performance measure.

To illustrate this tool, we propose the following examples:

Factor	Peak files from: ENCODE ChIP-seq-peak Uniform TFBS from UCSC	Position weight matrices from:
RXRA	GM12878 RXRA None HudsonAlpha peaks H1-hESC RXRA None HudsonAlpha peaks HepG2 RXRA None HudsonAlpha peaks	JASPAR CORE 2014 vertebrates MA0074.1 RXRA::VDR MA0512.1 Rxra: Jolma2013 Rxra_nuclearreceptor_DBD dimeric Mouse UniPROBE: UP00053_1 Rxra_primary UP00053_2 Rxra_secondary HOCOMOCO v9: RXRA_f1
JUND	K562 eGFP-JunD None UChicago peaks GM12878 JunD None Yale peaks H1-hESC JunD None HudsonAlpha peaks H1-hESC JunD None Stanford peaks HeLa-S3 JunD None Stanford peaks HepG2 JunD None HudsonAlpha peaks HepG2 JunD None Stanford peaks K562 JunD None Stanford peaks	JASPAR CORE 2014 vertebrates MA0491.1 JUND MA0492.1 JUND Mouse UniPROBE: UP00103_1 Jundm2_primary UP00103_2 Jundm2_secondary HOCOMOCO v9 JUND_f1

Step-by-step instructions: (for one example)

1. Go to PWMEval at:

   http://ccg.vital-it.ch/pwmtools/pwmeval.php

   On the left side of the input form, select:

   ```
   Genome:     H.sapiens
   Data type:  ranked ENCODE ChIP-seq-peak
   Series:     Uniform TFBS from UCSC
   Sample:     RXRA Gm12878 None HudsonAlpha peaks
   ```

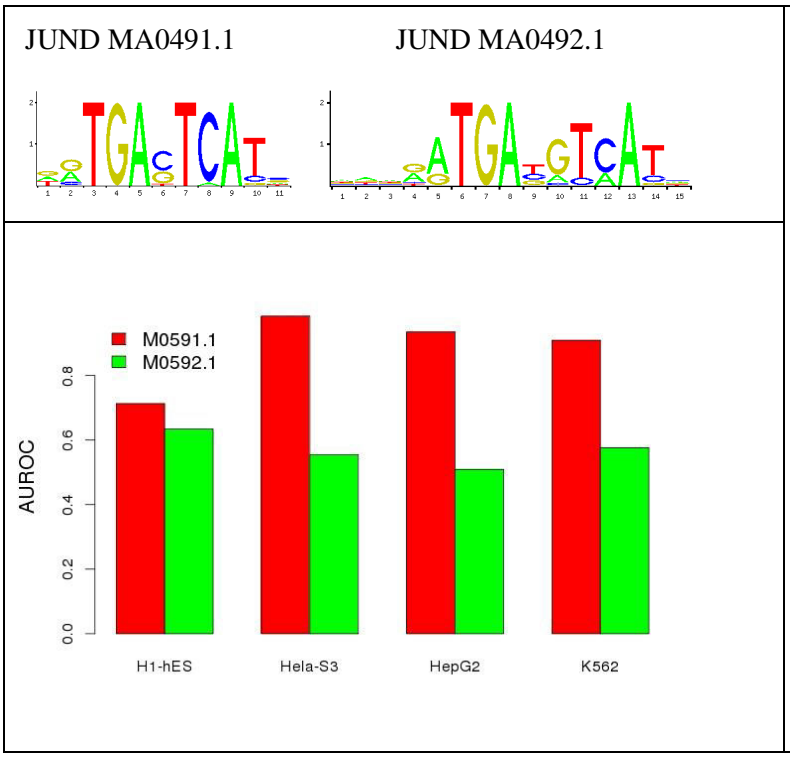
   Leave Repeat Masker off. On the right side, select:

   ```
   Motif Library:  JASPAR CORE 2014 vertebrates
   Motif:          RXRA::VDE MA0074.1
   ```

   Under "Options for Binding Prediction Method" specify: ChIP-seq peak ranking From 1 To 500, Sequence Length 250, Downstream Shift of Negative Set 300, Reverse peak ranking off, Odds-ration mode Seq Library bg. Submit.

Result and Discussion

The AUROC values for two JUND matrices evaluated with ChIP-seq data from Stanford are shown in Figure 4.3.1. We note a generally better performance of the matrix with the shorter spacing between the two reverse-complementary half sites (consensus sequence TGANTCA). However, there seem to be cell-type specific differences. In the embryonic stem cell line H1-hESC, the two matrices perform almost equally well whereas in HepG2, matrix MA0492.1 has an AUROC value very close to 0.5, the expected value for random predictions. These differences may be due to different AP1 heterodimers active in different cell types. For instance JUND homodimers may have different spacing preferences than JUND-FOS heterodimers. Both complexes would however be detected by antibodies against JUND.




```
cells=c("H1-hES","Hela-S3",
 "HepG2","K562")
motifs=c("M0591.1","M0592.1")
data=matrix(0,nrow=2,ncol=4)
data[,1] = c(0.713,0.634)
data[,2] = c(0.984,0.554)
data[,3] = c(0.935,0.509)
data[,4] = c(0.909,0.576)
colors=c("red","green")
barplot(data, main="",
  ylab ="AUROC", cex.lab=1.3,
  cex.main = 1.3, beside=TRUE,
  col=colors, names.arg=cells)
legend("topleft", motifs,
cex=1.2,
  bty="n", fill=colors)
```

Figure 4.3.1 Evaluation of two JUND matrices from JASPAR with ENCODE ChIP-seq data from Stanford. The R code used for producing the Figure is given on the right side.