1 Introduction

The main purpose of this document is to briefly explain how to reproduce the analyses presented in Figure 2 of the article. This document is not intended to show the exact commands and procedures to reproduce the figure but to give the reader the necessary information to understand the steps that have been made. Here we will show the procedure only for one example for each figure panel, we encourage the user to try to reproduce the other examples.

All the analyses have been done using data stored in the MGA databases in conjunction with its sister analysis tools SSA and ChIP–Seq. These are on–line tools designed to be user–friendly, fast and flexible. For example, they can be used for the analysis of motifs or chromatin states distribution around transcription start sites (TSS).

2 Reproducing Figure 2a

The image in Figure 2a shows the distribution of MNase–seq tags around CTCF sites in GM12878 cells. It has been generated by ChIP–Extract. ChIP–Extract accepts as input two feature called Reference and Target. The reference feature is used as anchor points in the genome (in this case CTCF sites) to check for the distribution of the Target feature (in this case MNase-seq reads). The output is in tabular form with each row representing a Reference feature (e.g. a CTCF site) and each column the number of Target features.
found at a specific distance from the Reference.

The following are the steps that need to be done to study the nucleosome distribution around CTCF sites from ENCODE:

- Go to the **MGA search page**. Select the *H. sapiens* genome assembly hg19 and search for the reference feature (CTCF peaks in GM12878), for examples using the keywords “GM12878 CTCF peaks uniform”.

![MGA Search page with results showing samples for the Uniform TFBS from UCSC series.](image)

Figure 1: MGA Search page with results showing samples for the Uniform TFBS from UCSC series.

- A list of samples matching the criteria will be shown (Figure 1). We used the peak list from Stanford. You can click “Set Reference” next to this sample. A new box will appear at the top of the page showing the selected Reference feature.

- Do the same for Nucleosomes in GM12878, using “nucleosome GM12878” as keywords. Set the Target using sample named “Gm12878|Nucleosome|rep1”.

- Click on ChIP–Extract in the box located at the top of the page. You will be redirected to the ChIP–Extract page loaded with the selected Reference and Target features. Set the other parameters in the page as shown in Figure 1. Note that some parameters have been modified from the default values: Target feature has been centered of 75 bases (half a nucleosome length) and the heatmap has been reordered using the option “Resemblance to overall pattern”. Centering means that reads on the plus strand will be shifted downstream by the provided distance, whereas reads on the minus strand will be shifted upstream by the same distance.

- The output page shows the nucleosome distribution around each CTCF site ordered by similarity to the overall pattern.
Figure 2: ChIP–Extract Input form ready to scan the CTCF bound region from ENCODE in the range -1005 to 1005 bp from the peak center for the presence of MNase-seq reads.

- Alternatively, users can save the TXT file for further analysis. This is a table with MNase–seq tags frequency values for each CTCF site (rows) and each distance range from the peak center (columns) that can be read into R and further analyzed, for example to group promoters according to the MNase distribution.

3 Reproducing Figure 2b

This figure panel shows a comparative analysis of promoter conservation scores from PhyloP in 4 organisms: *H. sapiens*, *M. musculus*, *D. rerio* and *D. melanogaster*. This has been done using the ChIP–Cor tool from ChIP–Seq. Conceptually this tool is similar to ChIP–Extract. It is designed to output the distribution of a target feature (e.g. the conservation scores from PhyloP) around a reference feature (e.g. promoters). Although similar, ChIP–Cor is faster than ChIP–Extract but output only the average profile of the Target feature.

Here we will show the steps to reproduce the chromatin distribution around *H. sapiens* promoters:

- As explained before, go to the **MGA search page**
- Select *H. sapiens* hg19 genome assembly and search for EPDnew promoters using, for example, the keyword “EPDnew”. Set as reference the latest version (005) **4**.
• Search for PhyloP vertebrate 46way and set it as target. Click on ChIP–Cor in the box at the top of the page.

• In the new ChIP–Cor input window, the reference feature is already pre-set to the *H. sapiens* EPDnew database and the Target to PhyloP. Change the other input parameters as shown in figure 3: select “oriented” in the Reference feature additional input options; and increase the count cut–off to 10.

![ChIP–Cor Input form](image)

Figure 3: ChIP–Cor Input form set to scan the EPDnew human promoters in the range -1000 to 1000 bp from the TSS for the presence of PhyloP conservation scores.

• The output page shows the average conservation profile around promoters in graphical form.

• Alternatively, users can save the TXT file for further analysis. This is a table with two columns, the first containing the midpoint of a distance range from the TSS and the second the average conservation score values for the corresponding range. This file can be read into R and plot along the values obtained for other databases.

• Save the TXT file as HsEPDnew_vs_phyloP.out

Results for the other organisms can be obtained selecting the appropriate reference (EPDnew promoter collections) and target (PhyloP conservation scores for each organism) features. Here is the R code that generate the image:
# Read data
hs.phylo <- read.table("HsEPDnew_vs_phyloP.out")
mm.phylo <- read.table("MmEPDnew_vs_phyloP.out")
dr.phylo <- read.table("DrEPDnew_vs_phyloP.out")
dm.phylo <- read.table("DmEPDnew_vs_phyloP.out")

# scale values between 0 and 1 for an easy comparison
hs.phylo.rs <- (hs.phylo[,2] - min(hs.phylo[,2])) / (max(hs.phylo[,2]) - min(hs.phylo[,2]))
mm.phylo.rs <- (mm.phylo[,2] - min(mm.phylo[,2])) / (max(mm.phylo[,2]) - min(mm.phylo[,2]))
dr.phylo.rs <- (dr.phylo[,2] - min(dr.phylo[,2])) / (max(dr.phylo[,2]) - min(dr.phylo[,2]))
dm.phylo.rs <- (dm.phylo[,2] - min(dm.phylo[,2])) / (max(dm.phylo[,2]) - min(dm.phylo[,2]))

col.1 <- c("darkgray","skyblue3","darkgreen","red2")

# Plot data:
plot(x, hs.phylo.rs, type="l", xlab="Distance from TSS (bp)",
ylab="PhyloP scaled average score", frame.plot=F, lwd=2,
col=col.1[1], xlim=c(-500,500))
points(x, mm.phylo.rs, type="l", lwd=2, col=col.1[2])
points(x, dr.phylo.rs, type="l", lwd=2, col=col.1[3])
points(x, dm.phylo.rs, type="l", lwd=2, col=col.1[4])
par(xpd=TRUE)
legend(x=-550, y=1.1, legend=c("H. sapiens","M. musculus",
"D. rerio","D. melanogaster"), lty=1, lwd=1, col=col.1, bty="n")
text(-730, 1.03, labels="B)", font=2, cex=1.5)

4 Reproducing Figure 2c

The image in Figure 2c is a reproduction of the Figure 5A shown in [5] and shows the distribution of MNase–seq tags around YY1 sites that are located near promoters (10kb from a promoter) stratified by the intensity of YY1 binding to DNA (measured as number of YY1 reads in YY1 peaks). Conceptually it has been generated following the following steps:

1. Find YY1 peaks near promoters using data from ENCODE [5] and EPDnew [4]. This
is a two step procedure: first using ChIP-Cor to correlate YY1 peaks and promoters; then using the “Feature Selection Tool” of the ChIP–Cor output page to extract YY1 peaks that fall near promoters. Here are the steps to do:

- In the MGA search page look for YY1 peaks from “Uniform TFBS from UCSC” series under hg19 assembly. Set sample “GM12878|YY1_(SC-281)|None|HudsonAlpha|peaks” as Reference. Search for EPDnew promoters and set the version 005 as target. Go to ChIP-Cor clicking on the button in the box at the top of the page.
- Increase the range from -10000 to 10000 and submit the job
- In the output page set the “Feature Selection Tool” as shown in Figure 4 and submit the job.

![Feature Selection Tool](image)

Figure 4: “Feature Selection Tool” of the ChIP–Cor output page used to select Reference feature that have a certain amount of Target features near them. In this case it is used to select YY1 peaks that have at least 1 promoter (Threshold parameter) in a 20 Kb region around them.

- The output page shows how many peaks have been selected (15571) with the possibility to download them in various formats. A series of “Useful links for Downstream Analysis” is also present and will be used in the next step. Please leave this page open in the browser.

2. Extract YY1 reads near YY1 peaks. This will be used to evaluate the YY1 binding intensity and to stratify YY1 peaks accordingly. As before it is done using ChIP–Extract. Here are the step to be done:

6
• From the previous step output page click on ChIP-Extract button in the “Useful links for Downstream Analysis” section.

• ChIP-Extract is then loaded with the data both on the Reference and Target feature. Change the target feature using the “Select available Data Sets” option. This is another way of accessing MGA samples from ChIP-Seq. Choose assembly hg19, “ENCOD ChIP-seq” as Data Type; “Wang et al. 2012, Transcription Factor Binding Sites from ENCODE/Stanford/Yale/USC/Harvard” as series and “GM12878 - YY1 std - None - Rep1” as sample. Center the Target feature by 90 bp. Submit the job.

• Save the table (TEXT) as YY1peaks YY1reads.mat

3. Extract MNase-seq reads (nucleosome) near YY1 peaks, again using ChIP-Extract. This step is similar to the previous and is perform as follow:

• From the step 1 output page click on ChIP-Extract button in the “Useful links for Downstream Analysis” section.

• ChIP-Extract is then loaded with the data both on the Reference and Target feature. Change the target feature using the “Select available Data Sets” option. Choose assembly hg19, “ENCOD DNase FAIRE, etc.” as Data Type; “GSE35586, Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU” as series and “Gm12878 Nucleosome rep1” as sample. Center the Target feature by 75 bp. Submit the job.

• Save the table (TEXT) as YY1peaks NUCLreads.mat

4. The YY1 peak stratification and MNase distribution is performed in R:

```r
# Read the data:
yy1.read <- as.matrix(read.table("YY1peaks YY1reads.mat", header=T))
yy1.mnase <- as.matrix(read.table("YY1peaks NUCLreads.mat", header=T))

colors <- c("#E41A1C", "#377EB8", "#4DAF4A")
x.lab <- seq(-990, 990, 10)

# find YY1 cover in a region from -200 to 200 around YY1 peaks
cover <- rowSums(yy1.read[,c(80:120)])
cover.q <- quantile(cover, probs=c(0, 0.33, 0.66, 1))

# Divide peaks into 3 groups: low, medium and high cover
```
index <- vector(mode="list", length=3)
for(i in 1:(length(cover.q)-1)){
  limit.lower <- cover.q[i]
  limit.upper <- cover.q[i+1]
  index[[i]] <- which((cover >= limit.lower) &
    (cover < limit.upper))
}

# Plot them:
plot(x.lab, colMeans(yy1.mnase), type='l', lwd=2,
  main="", xlab="Distance from YY1 peak centers (bp)",
  ylab="Mean signal", ylim = c(0.1, 0.75), col=0,
  frame.plot=F)
for(i in 1:length(index)){
  lines(x.lab, colMeans(yy1.mnase[index[[i]]],), lwd=2,
    col=colors[i])
}
legend(x=100, y=0.82, legend=c("33% lower",
  "33% intermediate", "33% upper"), col=colors, lty=1,
  lwd=1, bty='n')

References

[1] Ambrosini, G., Praz, V., Jagannathan, V., and Bucher, P. (2003) Signal search analysis server. *Nucleic acids research*, 31(13), 3618–3620.

[2] Ambrosini, G., Dreos, R., and Bucher, P. (2015) Principles of ChIP-seq data analysis illustrated with examples. *Genomics and Computational Biology*, 1(1), 22.

[3] Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1), 110–121.

[4] Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R., and Bucher, P. (2017) The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research*, 45(D1), D51–D55.

[5] Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9), 1798–1812.