MOTIF TUTORIAL

1. Extraction of bound and unbound STAT1 motifs

Background :

Only a fraction of STAT1 motifs that occur in the human genome are actually bound by STAT1 after cytokine stimulation. To investigate the molecular processes that recruit STAT1 to its physiological target sites, it would be useful to have clean lists of bound STAT1 motifs and experimentally verified unbound motifs. We will compile such a list using PWMScan from the PWMTools server. PWMScan scans complete genomes with a PWM and returns a complete list of matches above a user-specified threshold. Next we will extract PWM matches with high tag coverage or zero-tag coverage using the "Enriched Feature Extraction" option of ChIP-Cor. As an application example we will generate single-base resolution conservation plots showing the spacing between doublets of STAT1 motifs. A similar plot based on peak lists generated with FindM is shown in Fig. 22 of the Basic Tutorial.

As in the basic Tutorial, we are going to use the following ChIP-seq data set:

```
Genome:     H. sapiens (Feb 2009 GRCh37/hg19)
Data Type:  ChIP-seq
Series:     Robertson 2007, HeLa S3 cells, Genome-wide STAT1 ...
   Sample:  HeLa S3 Stat1 stim
```

Step-by-step procedure:

1. Go to PWMScan at:

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

   and fill out the input form as follows. On the left side, select genome assembly H. sapiens (Feb 2009 GRCh37/hg19). On the right side, select Motif Library JASPAR CORE 2016 vertebrates, Motif STAT1 MA0137.3, Cut-off P-value 0.0001, Bg base composition (default), Search strand both, Ref. position 6, Non-overlapping matches checked. Submit.

2. The results page reports 806'754 hits. Transfer the hit list to ChIP-Cor with the aid of the direct navigation button. On the left side of the ChIP-Cor input form select Strand oriented, leave the Repeat Masker unchecked for the moment. As Target Feature, select the STAT1 sample indicated above, Strand any, centering 75. Other parameters are not relevant at this stage. Submit.

   Note. PWMScan returns an oriented SGA file, possibly containing matches on the plus and minus strand of the genome. We therefore select strand oriented for the reference feature.

3. On the results page, use the "Feature Extraction Tool" menu with the following inputs. From -100 To 100, Threshold 10, Cut-off 1, Depleted Feature Selection off, Reference Feature Oriented on, Top Enriched/Depleted Features blank. Submit an save the output SGA file to disk under the name:

   stat1_bound.sga

4. To compile a list of high-scoring unbound STAT1 motif matches, repeat the above procedure with the following changes. At Step 1(PWMScan) use a more stringent Cut-off P-value of 0.00001. At Step 3 (Feature Extraction Tool) enter the following inputs. From -100 To 100, Threshold 1, Cut-off 1, Depleted Feature Selection on, Reference Feature Oriented on, Top Enriched/Depleted Features blank. Submit and save the output SGA file to disk under the name:

   stat1_unbound.sga

Note: Specifying Threshold 1 will extract all matches with zero (< 1) tags.

5. Go to ChIP-Cor at :

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

   On the left side of the input form, upload the previously saved file stat1_bound.sga as Reference Feature, select genome assembly H. sapiens (Feb 2009 GRCh37/hg19), and specify: Strand oriented, Repeat Masker off, Beginning -12, End 12, Window width 1, Count Cut-off 10, Normalization count-density. On the right side select as Target Feature:

   ```
   Genome:    H. sapiens (Feb 2009 GRCh37/hg19)
   Data Type: Sequence-derived
   Series:    phyloP base-wise conservation
   Sample:    *PhyloP vertebrate 46way (score >=2)
   ```

   And specify: Strand any, Repeat Masker off. Submit.

6. On the results page, save the text output file under the name:

   stat1_bound_phylop.txt

   or import the results directly 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:

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

7. Repeat the previous step for the unbound list and save the text output file under the name:

   stat1_unbound_phylop.txt

   or import the results directly into an R variable via URL as explained above:

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

8. Make a figure showing the single-base resolution PhyloP profiles for the motif list using the R code shown in Figure 1.2.

Results and Discussion

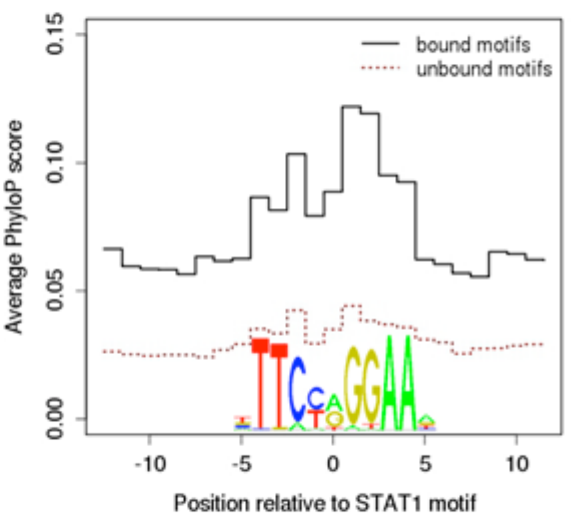
The two lists of bound and unbound STAT1 motif matches contain 14’295 and 43’827 lines, respectively. They are represented in an extended SGA format (Fig. 1.1) with the three additional fields:

Field 6:	PWM score (from PWMScan)
Field 7:	Tag count (from ChIP-Cor, Enriched Feature Selection)
Field 8:	DNA sequence of the match (from PWMScan)

```
NC_000001.10  ChIP_R  877469   +  1  TTTACGGGAAC  1186   11
NC_000001.10  ChIP_R  1069080  -  1  TTTCCAGGAAA  1772   11
NC_000001.10  ChIP_R  1070896  +  1  TTTCTGGGAAA  1722  107
NC_000001.10  ChIP_R  1070917  -  1  CTTCTGGGAAT  1456  116
NC_000001.10  ChIP_R  1175273  +  1  GTTCTGGGAAG  1469   17
NC_000001.10  ChIP_R  1358496  +  1  CTTCCGGGAAT  1497  107
NC_000001.10  ChIP_R  1358517  -  1  TTTCCGGGAAA  1763   98
NC_000001.10  ChIP_R  1368746  -  1  GTTCCAGGAAG  1519   49
NC_000001.10  ChIP_R  1499241  +  1  CTGCTGGGAAA  1097   19
NC_000001.10  ChIP_R  1891748  +  1  CTGCCAGGAAA  1147   10
```

Figure 1.1. Format of extended SGA file containing the list of bound STAT matches.

The single-base resolution PhyloP conservation profiles are shown in Fig. 1.2. Overall, the profiles look very similar to the ones obtained in the basic tutorial.



R Code

```
ylim=c(0,0.15)
xlab=paste("Position relative",
   "to STAT1 motif")
ylab="Average PhyloP score"
plot(bound[,1]-0.5,bound[,2],
  type="s", ylim=ylim, xlab=xlab,
  ylab=ylab, cex.axis=1.3,
  cex.lab=1.3, lwd=2)
par(new=T)
points=plot(unbound[,1]-0.5,
  unbound[,2], t="s", ylim=ylim,
  xlab="", ylab="", xaxt="n", lwd=2,
  yaxt="n", lty=3, col="darkred")
legend("topright", legend=
  c("bound motifs","unbound motifs"),
  col=c("black","darkred"),lwd=2,
  lty=c(1,3), bty="n", cex=1.2)
```

Figure 1.2 Single-base resolution sequence conservation profiles for bound and unbound STAT1 motifs in the human genome.

2. Effects of Repeat-masking

Background:

In the basic tutorial, we sent a repeat-masked list of peak sequences to the MEME-ChIP server for motif discovery. Here we show what happens if we use a non-repeat-masked peak list instead.

Step-by-step instruction.

1. Go to ChIP-Peak.

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

   On the left side of the input form, select the previously used STAT1 sample as input data set with Strand any, Centering 75, Repeat Masker off. On the right side specify Window Width 300, Vicinity Range 300, Peak Threshold / Tag counts 150, Count cut-off 1, Refine Peak Positions checked. Submit.

2. There are 737 peaks detected. On the ChIP-Peak results page use the “Sequence Extraction” menu to extract sequences from -60 to 60. Submit.

3. On the following page, save the sequence file to disk under the name

   stat1_peaks_t150.seq

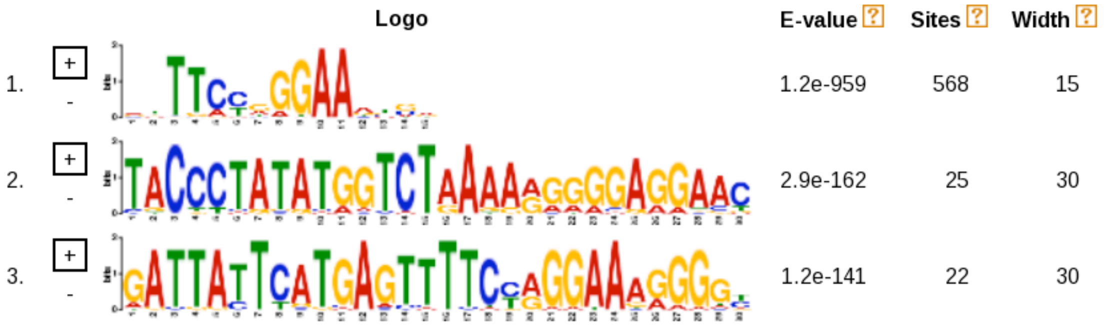
4. Open a new browser window and go the one of the following MEME-ChIP servers.

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

   Select normal discovery mode. You can either input the sequences by uploading the previously saved file or via copy-paste. If you choose copy-paste, you have to open the hyperlink “Sequence File” in the sequence extraction output page and copy the complete content of the page into the Edit buffer of your internet browser. On the MEME-ChIP form, select Type in sequences under “Input the primary sequence” and paste the sequences into the text window. Use defaults for all other options and parameters.

Results and Interpretation

The complete MEME-ChIP output can be found here. The main page provides a summary of the motifs found by different tools, including MEME, DREME and CentriMo. Click on the hyperlink “MEME” to go to the MEME output page. You will see 3 highly significant motifs, *see* Fig. 2.1. The first one clearly resembles the canonical STAT1 motif. The other two motifs do not look like ordinary TF binding motifs in that they are very long, highly conserved but rather rare (only about 25 matches in 737 sequences).



	Logo	E-value ?	Sites ?	Width ?
1.		1.2e-959	568	15
2.		2.9e-162	25	30
3.		1.2e-141	22	30

Figure 2.1. Motifs found by MEME in non-repeat-masked highly occupied STAT1 peaks regions.

The origin of the additional STAT1 motifs is explained in (Schmid & Bucher PLoS One 2010). They are part of a rare repetitive element named MER41B which harbors a pair of STAT1 motifs with a center-to-center spacing of 21 bp. Information about repetitive elements can be found in Repbase at:

www.girinst.org/repbase/

A report on MER41B can be found here. (You need to register to Repbase to see the report, but registration is free). The consensus sequence of the MER41B repeat is shown in Figure 2.2 with the motifs found by MEME indicated by different colors.

```
TGTCAGAGGCGTTTGAACCAGAGCAACTCCATCTTGAATAGGCGCTGGGTAAAATRAGGCTGARACCTAC
TGGGCTGCATTCCCAGACGGTTAAGGCATTCTAAGTCACAGGATGAGATAGGAGGTCGGCACAAGATACA
GGTCATAAAGACCTTGCTGATAAAACAGGTTGCAGTAAAGAAGCCGGCYAAAACCCACCAAAACCAAGAT
GGCCACGAGAGTGACCTCTGGTCGTCCTCACTGCTCATTATATGYTAATTATAATGCATTAGCATGCTAA
AAGACACTCCCACCAGCACCATGACAGTTTACAAATGCCATGGCAACGTCAGGAAGTTACCCTATATGGT
CTAAAAAGGGGAGGAACCCTCAGTTCCGGGAATTGCCCGCCCCTTTCCTKGAAAAYTCATGAATAATCCA
CCCCTTGTTTAGCATATAATCAAGAAATAACCATAAAAATRGGCAACCAGCAGCCCTCGGGGCTGCTCTG
TCTATGGAGTAGCCATTCTTTTATTCCTTTACTTTCTTAATAAACTTGCTTTCACTTTACTCTRTGGACT
CGCCCTGAATTCTTTCTTGCACRAGATCCAAGAACCCTCTCTTGGGGTCTGGATCGGGACCCCTTTCTTG
TAACA1
```

Figure 2.2 MER41B consensus sequence. Occurrences of the 3 motifs found by MEME are marked by different colors (motif 1 red, motif 2 green, motif 3 blue). Motif 3 occurs in reversed orientation and includes a copy of Motif 1.

About 4% of highly occupied STAT1 motifs in the human genome come from this repeat family. The high degree of sequence conservation beyond the STAT1 core motif is not due to functional constraints but simply due to the fact this repeat element has been faithfully replicated in the human genome many times via retro-transposition.

3. Effects of repeats on motif spacing analysis.

We will use the list of highly occupied STAT1 motifs generated in part one 1 for analyzing the spacing between pairs of STAT1 motifs.

Step-by-Step instructions.

1. Go to OProf at

   http://ccg.vital-it.ch/ssa/oprof.php

   and upload the previously saved file

   stat1_bound.sga

   as input data, and select genome assembly H. sapiens (Feb 2009 GRCh37/hg19). Leave the Repeat Masker unchecked, 5' border 0, 3' border 60, window 11, shift 1, search mode forward. On the right side of the input form, select

   ```
   Motif Library: JASPAR CORE 2016 vertebrates
   Motif:         STAT1 MA0137.1 (length=11)
   ```

   Cut-off p-value 0.001, Ref. position 6. Submit.

   Note that this will produce a single-base resolution plot because the window width is identical to the motif length.

2. From the OProf results page, save the text output file under the name

   stat1_spacing_w11.txt

   or transfer the numerical results via URL directly into an R variable named:

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

   Repeat the OProf analysis with Repeat Masker checked and save the output as:

   stat1_spacing_w11_rmsk.txt

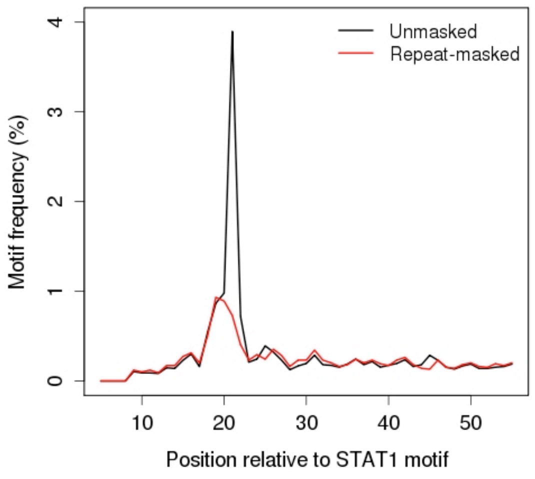
   or transfer the numerical results via URL directly into an R variable named:

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

3. Superpose the high-resolution STAT1 motif spacing plot obtained with and without repeat-masking in one Figure. You may use the R code shown in Figure 3.1 for this purpose.

Results and Discussion

Before repeat masking, we see a high spike (single-base peak) at pos. 21. This spike comes from the MER41B repeats which harbor two STAT1 sites at a center-to-center distance of exactly 21 bp. After repeat-masking we see a lower and somewhat broader peak corresponding to STAT1 motif pairs with a center-to-center distance of 19-23 bp.



R Code

```
masked <- ... ; unmasked <- ...

ylim=c(0,4)
xlab=paste("Position relative",
   "to STAT1 motif")
ylab="Motif frequency (%)"
plot(unmasked, col="black",
  type="l", ylim=ylim, xlab=xlab,
  ylab=ylab, cex.axis=1.3,
  cex.lab=1.3, lwd=2)
par(new=T)
points=plot(masked, col="red", t="l",
  ylim=ylim, xlab="", ylab="",
  xaxt="n", lwd=2, yaxt="n")
legend("topright", legend=
  c("Unmasked","Repeat-masked"),
  col=c("black","red"),lwd=2,
  bty="n", cex=1.2)
```

Figure 3.1 Positional distribution of STAT1 motifs downstream of an *in vivo* bound STAT1 motif with and without repeat-masking. The profiles have been generated at single base resolution (window width = motif length). The R code for generating the Figure is shown on the right side.

What next?

You may look at the distribution of other motifs near *in vivo* bound STAT1 motifs before and after repeat masking. Interesting examples are SRF, Crx, HoxA9, RFX2 and SP1 from the JASPAR CORE collection. Use sequence range from 0 to 100, p-val 0.001 and search mode bidirectional with these examples. Make sure that the window size is identical to the motif length.

4. Correlation between site occupancy and PWM score

Usually only a fraction of predicted sites (PWM matches) are occupied by the corresponding TF *in vivo* .To get a global impression of the relationship between PWM score and site occupancy we propose several graphical plots.

We will use the following ChIP-seq data sets and motifs as examples:

1. STAT1, in interferon-γ stimulated HeLa cells:

```
Genome:     H.sapiens (Feb 2009 GRCh37/hg19)
Data Type:  ChIP-seq
Series:     Robertson 2007, HeLa S3 cells, Genome-wide STAT1 ...
Sample:     HeLa S3 Stat1 stim (75 bp centered)

Motif Library: JASPAR CORE 2016 vertebrates
Motif:         STAT1 MA0137.1 (length=11)
```

2. CTCF in human embryonic stem cell line H1-hESC

```
Genome:     H.sapiens (Feb 2009 GRCh37/hg19)
Data Type:  ENCODE ChIP-seq
Series:     GSE32465, Transcription Factor Binding Sites by ChIP-seq
Sample:     H1-hESC None CTCF (40 bp centered)

Motif Library: JASPAR CORE 2016 vertebrates
Motif:         CTCF MA0139.1 (length=19)
```

Step-by-step instructions:

1. Go the PWMScan at:

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

   On the left side, select H. sapiens (Feb 2009 GRCh37/hg19) as Target Database. On the right side select the above indicated STAT1 matrix, Cut-off P-value 0.0001, Bg base composition default, Search Strand both, Ref. Position 6, Non-overlapping matches checked. Submit.

2. Use the direct navigation button to transfer the match list to ChIP-Cor. On the left side of the input form, specify Strand oriented. Other parameters are not relevant at this stage. On the right side select as Target Feature the STAT1 ChIP-seq data set indicated above with corresponding centering distance. Submit.

3. On the ChIP-Cor output page, fill out the "Feature Selection Tool" menu as follows. From -100 To 100, Threshold 0, Cut-Off 1, Depleted Feature Selection off, Ref. Feature Oriented on, Select Top Enriched/Depleted Ref. Features blank. Submit and save the SGA file posted on the results page to disk under the name:

   stat1_score_counts.sga

4. Repeat Step 1 to 3 with the CTCF matrix and ChIP-seq sample indicated above. Use the same options and parameters as before except: for PMWScan Cut-off P-value 0.00001, Ref. Position 10, and for the Enriched Feature Extraction Option Cut-off 10.

   Explanations: PWMScan takes too long and produces too many matches for the CTCF matrix with P-value 0.0001. The PWMs for STAT1 and CTCF have length 11 and 19, respectively. We place the reference position in the center motifs, *i.e.* pos. 6 and 10. We choose a higher count cut-off for the CTCF ChIP-seq data because this data set has very high average count coverage (≥ 0.01 per bp). Exact duplicates of sequence reads in peak regions are thus expected to occur pulled down sequence fragments and consequently assumed t be real.

5. Generate the plots shown under Results using the R code included in the Figures.

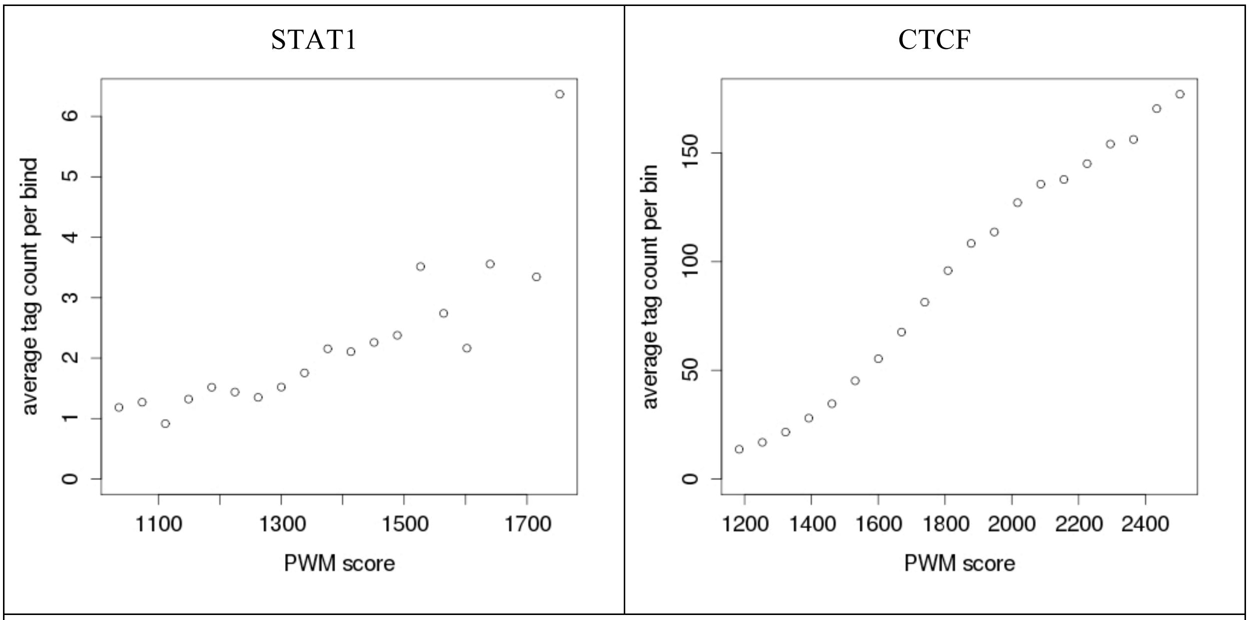
Results and Discussion

A part of the STAT1 peak list generated by the above procedure is shown below:

```
NC_000001.10  ChIP_R  76246251  +  1  TCTCTGGGAAA  1112   0
NC_000001.10  ChIP_R  76248179  +  1  GTGCTAGGAAA  1116   0
NC_000001.10  ChIP_R  76248576  -  1  TTTCTTGTAAA  1017   0
NC_000001.10  ChIP_R  76252361  +  1  CTTCCGGTAAT  1061  13
NC_000001.10  ChIP_R  76253842  -  1  GTTATGGGAAC  1029   2
```

Note that the PWM score and ChIP-seq tag coverage are given in the fields 6 and 7, respectively.

The relationship between PWM score and ChIP-seq tag coverage for STAT1 is visualized by three different plots.



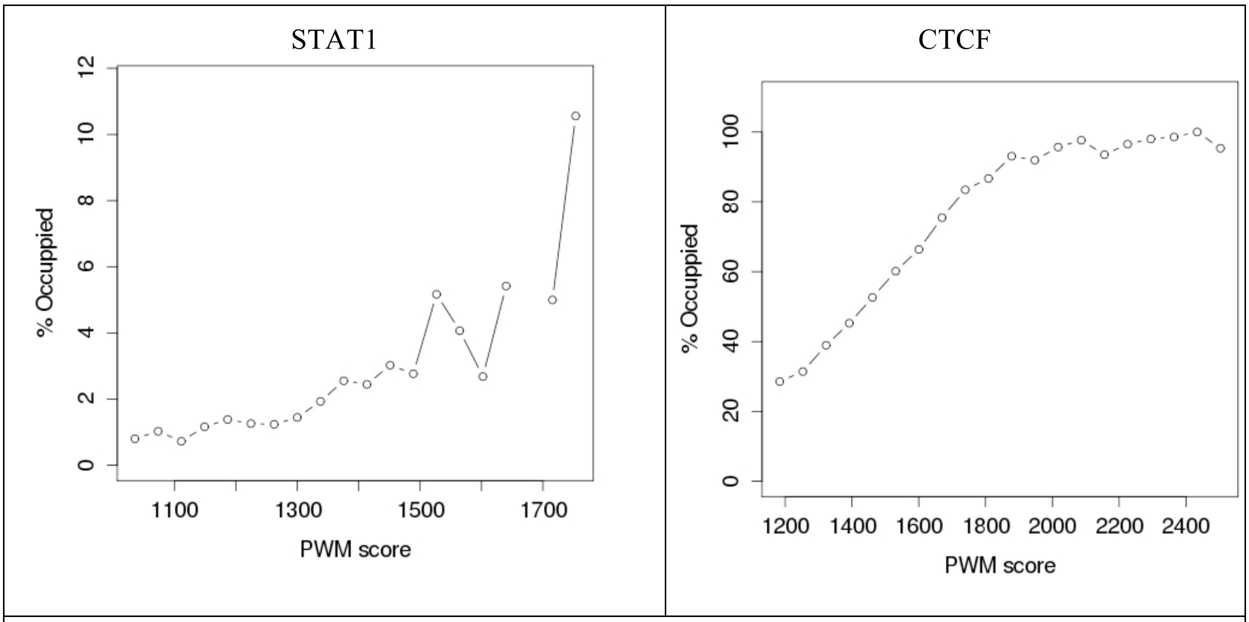
R Code

```
sga=read.table("ctcf_score_counts.sga"),
sga=read.table("stat1_score_counts.sga")

max=max(sga[,7], na.rm=T); min=min(sga[,7], na.rm=T)
r=seq(min,max,(max-min)/20)
L = length(r)-1; x=rep(0,L); y=rep(0,L)
for(i in 1:L) {x[i]=(r[i]+r[i+1])/2;
   y[i]=mean(subset(sga[,8],sga[,7] > r[i] & sga[,7] <= r[i+1]))}
plot(x,y,xlab="PWM score", ylab="average tag count per bin",
   ylim=c(0,max(y, na.rm=T)), cex.lab=1.3, cex.axis=1.3)
```

Figure 4.1 Average ChIP-seq tag counts of motif as a function of the PWM score.

Fig. 4.1.shows the average tag counts for subsets of PWM matches having approximately the same PWM score. The total score range was partitioned into 20 intervals of equal size. We see a generally low coverage for STAT1 motifs starting to significantly increase only in the upper half of the score range. The picture changes completely for CTCF where we see much higher tag coverage in general and an almost linear increase over the entire score range.



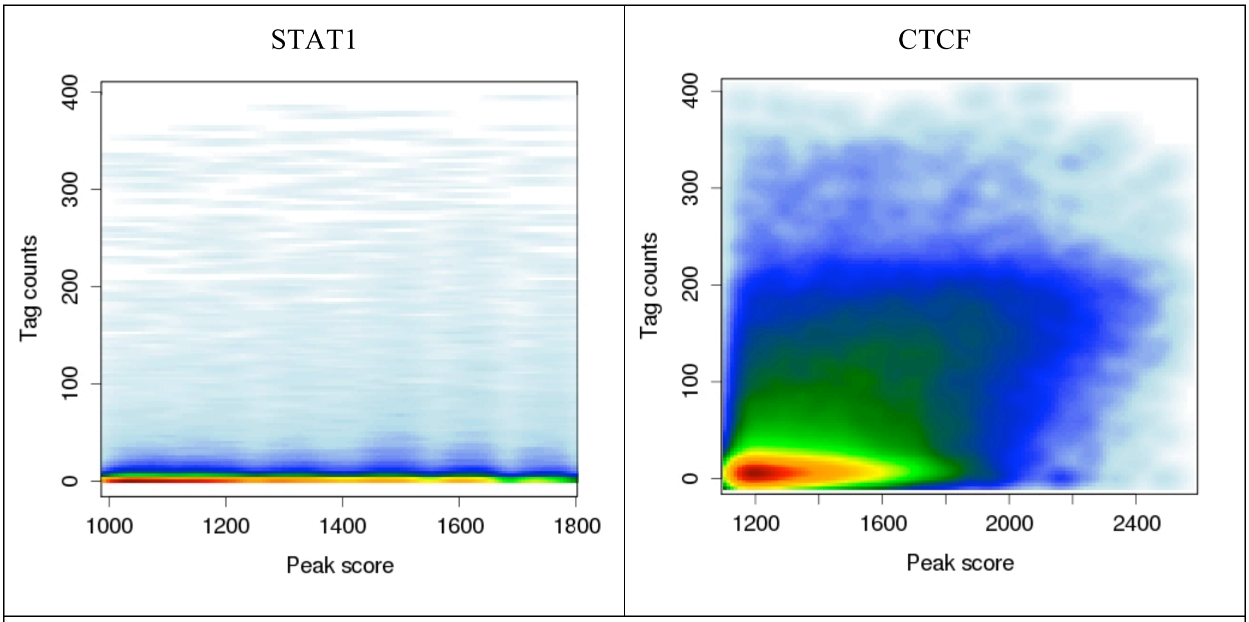
R Code

```
sga=read.table("ctcf_score_counts.sga")
sga=read.table("stat1_score_counts.sga")

occ=as.integer(sga[,8] >= 10)
max=max(sga[,7], na.rm=T); min=min(sga[,7], na.rm=T)
r=seq(min,max,(max-min)/20)
L = length(r)-1; x=rep(0,L); y=rep(0,L)
for(i in 1:L) {x[i]=(r[i]+r[i+1])/2;
   y[i]=100*mean(subset(occ,sga[,7] > r[i] & sga[,7] <= r[i+1]))}
plot(x,y,t="b", xlab="PWM score", ylab="% Occuppied",
   ylim=1.1*c(0,max(y, na.rm=T)), cex.lab=1.3, cex.axis=1.3)
```

Figure 4.2 Percent occupied motifs (>10 tags) as a function the PWM score.

Fig. 4.2 shows the percentage of occupied sites as a function of the score. As for Fig. 4.1, PWM matches were attributed to 20 bins according to their PWM score. Occupancy was defined somewhat arbitrarily as a minimum of 10 ChIP-seq tag counts. The picture emerging from this analysis is consistent with the result shown in Fig. 4.1. For STAT1, less than 10% of PWM matches are occupied even in the highest scoring subclass. For CTCF, the 100% mark is almost reached already in the middle of the PWM score range.



R Code

```
sga=read.table("ctcf_score_counts.sga")
sga=read.table("stat1_score_counts.sga")

colorP=colorRampPalette(c( "white","lightblue","blue", "darkgreen",
  "green","yellow", "orange","darkorange","red", "darkred"))
smoothScatter(sga[,7], sga[,8], nrpoints=0, colramp=colorP,
   xlab="Peak score", ylab="Tag counts", cex.lab=1.3, cex.axis=1.3)
```

Figure 4.3 Smooth scatter plots of tag counts versus PWM score.

Figure 4.3 visualizes the correlation between PWM scores and tag counts with a density scatter plots. For STAT1we notice barely any correlation. The overwhelming majority of PWM matches are concentrated along the base-line of the plot. Nevertheless, we see more white color in the upper left triangle than in the lower right triangle. This is perhaps the only sign of a weak positive correlation between PWM score and tag counts. For CTCF, the correlation is clearly perceptible. We recognize an increase in tag coverage along a diagonal. However, we also see the highest concentration of PWM matches in the lower left corner. This effect is amplified by the generally higher number of PWM matches in the lower scoring range.

Where to go from now?

Make similar plots for different cases. First repeat the analysis with the same PWMs using different ChIP-seq samples from ENCODE. Then look at other TFs (*e.g.* JUND, ESR1, RFX5).

5. Motif discovery with SSA tools

The SSA server also features motif discovery tools. The SList (Signal List) tool finds *k*-mer motifs that are locally over-represented relative to a functional site. It uses an exhaustive word search strategy.

PatOP (Pattern Optimization) optimizes a consensus sequence or weight matrix motif using an iterative alignment algorithm *see* (Bucher, J. Mol. Biol. 1990). It optimizes three components of the model, the weight matrix itself, the cut-off value, and the borders of the preferred region of occurrence, keeping two of these components constant at a time. PatOp also has the capability of extending the matrix to the left and right side if additional consensus is observed, or to drop positions in the opposite case.

This part shows how to generate a PWM for STAT1 with a combination of the two tools. As before, we will use the following data set:

```
Genome:     H.sapiens (Feb 2009 GRCh37/hg19)
Data Type:  ChIP-seq
Series:     Robertson 2007, HeLa S3 cells, Genome-wide STAT1 ...
Sample:     HeLa S3 Stat1 stim (75 bp centered)
```

Step-by-step instructions:

1. We first generate a STAT1 peak list. Go to ChIP-Peak at:

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

   On the left side of the input form, select the above indicated ChIP-seq sample as data input with Strand any, Centering 75, Repeat Masker on. On the right side specify Window Width 300, Vicinity Range 300, Peak Threshold / Enrichment factor 20, Count cut-off 1, Refine Peak Positions checked. Submit.

2. There are 10601 peaks. Because the output page doesn't provide a direct navigation button to SList, we transfer the output FPS file via URL. Right-click on the link named "FPS" and select "Copy Link Location". Then open SList in a new browser window:

   http://ccg.vital-it.ch/ssa/slist.php

   On the left side of the input form activate the checkbox Upload custom Data. Specify format FPS, paste the previously copied URL into the text area provided for this purpose and select Genome H. sapiens (Feb 2009 GRCh37/hg19). Further below select 5'border -499, 3'border 500, Window size 100, shift 25. Under "Selection criteria", select Occurrence frequency over-represented, Calculation mod 2, Selection mode local maxima/minima, St-dev cut-off 10, Sort list by st-dev over/under-representation.

   Note: Calculation mode 1 uses the mean of all word frequencies in the window under consideration as the reference value, whereas Calculation mode 2 uses the mean of the frequencies of the specific word under consideration in all windows except the one under consideration.

   On the right side under the header "Signal Collection" select complete, # of bases 5 and near the bottom of the page, Min. # of matches 5. Submit. (This takes some time.)

3. The results from SList indicate that the most over-represented 5-mer word is TTCCC. We will use this sequence as a seed for optimizing a weight matrix by PatOP. Go to

   http://ccg.vital-it.ch/ssa/patop.php

   and upload the previously generated STAT1 peak list via URL to the PatOp intput form in the same way as you did for SList. Then fill out the remaining parts of the form as follows. Sequence Range: 5'border -499, 3'border 500. Under" Optimization parameters" enter:

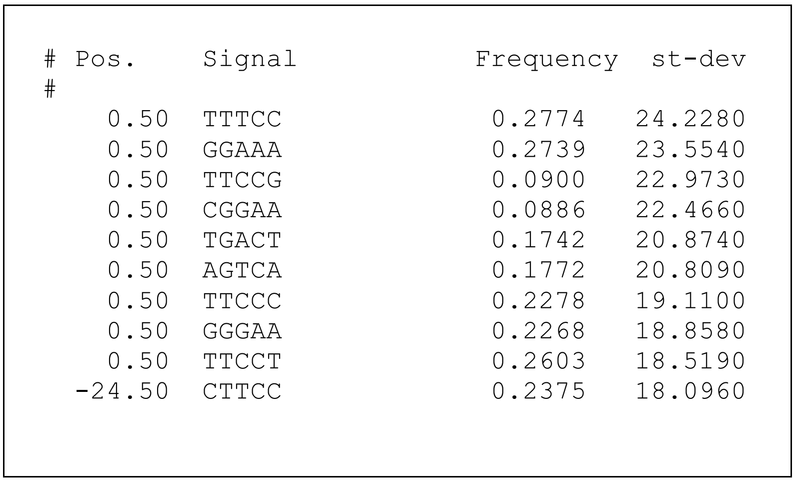
   Window size min. 50 max. 100, increment 10,
   Cut-off % min. 60, max 100, increment 1,
   Search mode bidirectional,
   Selection mode non-overlapping,
   Context range left border -25 right border 25,
   Matrix extension yes, max. gap 1, min chi-sqr 15, Minimal relative entropy 0.1,
   Maximal false-positive rate 20%,
   Normalization mode mononucleotide,
   Initial cut-off-optimization: no
   Smoothing: 2%
   Maximal # cycles: 25

Note: With these parameters setting PatOp will automatically extend the matrix if adjacent positions show a skewed base composition.

On the right side of the form, check Consensus sequence, type TTTCC in the adjacent test area and specify Mismatches 0, Ref-pos. 3. Submit.

Results and Discussion

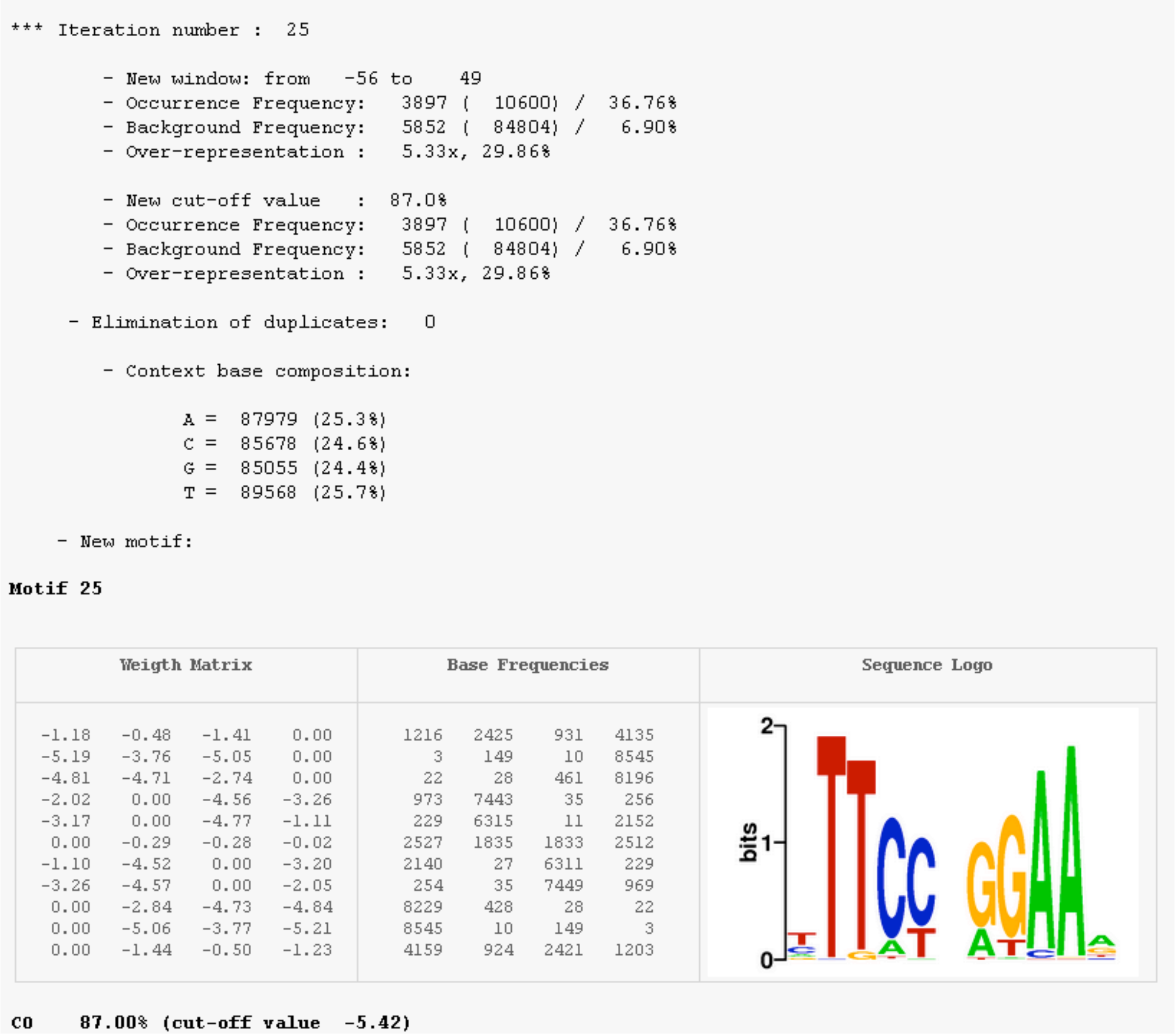
SList returns a list of locally over-represented words *see* Fig. 5.1. We note that TTTCC comes out on top of the list and its preferred region of occurrence is centered right at the center of the peaks (Pos. 0.5).



```
# Pos.     Signal            Frequency  st-dev
#
    0.50   TTTCC               0.2774   24.2280
    0.50   GGAAA               0.2739   23.5540
    0.50   TTCCG               0.0900   22.9730
    0.50   CGGAA               0.0886   22.4660
    0.50   TGACT               0.1742   20.8740
    0.50   AGTCA               0.1772   20.8090
    0.50   TTCCC               0.2278   19.1100
    0.50   GGGAA               0.2268   18.8580
    0.50   TTCCT               0.2603   18.5190
  -24.50   CTTCC               0.2375   18.0960
```

Figure 5.1. Locally over-represented 5-mer words in STAT1 peak regions reported by SList.

The output of PatOp is shown in Fig. 5.2. The sequence logo of the optimized motif resembles those of the JASPAR motif and the motif found by MEME. Note that PatOP reports the final motif both as position weight matrix and base frequency matrix. The weight matrix is scaled such that the highest scoring base at each position always receives a weight of zero. PatOp also reports the optimal cut-off and preferred region of occurrence. In the example presented here, the optimization process comes to an end after 15 iterations.



```
*** Iteration number :  25

        - New window: from    -56 to     49
        - Occurrence Frequency:    3897 (  10600) /  36.76%
        - Background Frequency:    5852 (  84804) /   6.90%
        - Over-representation :    5.33x, 29.86%

        - New cut-off value    :  87.0%
        - Occurrence Frequency:    3897 (  10600) /  36.76%
        - Background Frequency:    5852 (  84804) /   6.90%
        - Over-representation :    5.33x, 29.86%

     - Elimination of duplicates:   0

        - Context base composition:

              A =  87979 (25.3%)
              C =  85678 (24.6%)
              G =  85055 (24.4%)
              T =  89568 (25.7%)

    - New motif:

Motif 25
```

Weigth Matrix				Base Frequencies				Sequence Logo
-1.18	-0.48	-1.41	0.00	1216	2425	931	4135	
-5.19	-3.76	-5.05	0.00	3	149	10	8545	
-4.81	-4.71	-2.74	0.00	22	28	461	8196	
-2.02	0.00	-4.56	-3.26	973	7443	35	256	
-3.17	0.00	-4.77	-1.11	229	6315	11	2152	
0.00	-0.29	-0.28	-0.02	2527	1835	1833	2512	
-1.10	-4.52	0.00	-3.20	2140	27	6311	229	
-3.26	-4.57	0.00	-2.05	254	35	7449	969	
0.00	-2.84	-4.73	-4.84	8229	428	28	22	
0.00	-5.06	-3.77	-5.21	8545	10	149	3	
0.00	-1.44	-0.50	-1.23	4159	924	2421	1203	

```
CO    87.00% (cut-off value  -5.42)
```

Figure 5.2. Optimized STAT1 motif found by PatOp.

What next?

Try to run SList and PatOp with different input parameters. Try to optimize a PWM for another TF, *e.g.* CTCF.