

BASIC CHIP-SEQ AND SSA TUTORIAL

In this tutorial we are going to illustrate the capabilities of the ChIP-Seq server using data from an early landmark paper on STAT1 binding sites in γ -interferon stimulated HeLa cells ([Robertson et al., 2007](#)). This data set, which comprises about 15 million mapped sequence tags, is available from the ChIP-Seq server menu.

What follows is a step-by-step description of how the results have been produced.

The list of data and results files that have been used for the analysis can be found at:

<http://ccg.vital-it.ch/chipseq/doc/basic/>

Note that some of the analysis steps described in this tutorial rely on programs from the Signal Search Analysis (SSA) server at:

<http://ccg.vital-it.ch/ssa/>

1. 5'-3' end correlation (ChIP-Cor)

We start by generating a 5'-3' correlation plot using **ChIP-Cor**. We use the 5' (+ strand) tags as reference feature and compute the frequencies of 3' tags as a function of the distance from the reference feature.

To this end, open the **ChIP-Cor** server home page at:

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

Fill out the form as shown in Table 1, and click on the Submit button.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ChIP-seq Series: Robertson 2007 ... Sample: HeLa S3 STAT1 stim	Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ChIP-seq Series: Robertson 2007 ... Sample: HeLa S3 STAT1 stim
Additional Input Data Options Strand: + Analysis Parameters Range: -1000 to 1000	Additional Input Data Options Strand: -
Histogram Parameters Window width: 10 Count Cut-off: 1 Normalization: count density	

Table 1. 5'-3' end correlation with ChIP-Cor.

On the output page you will see the following picture:

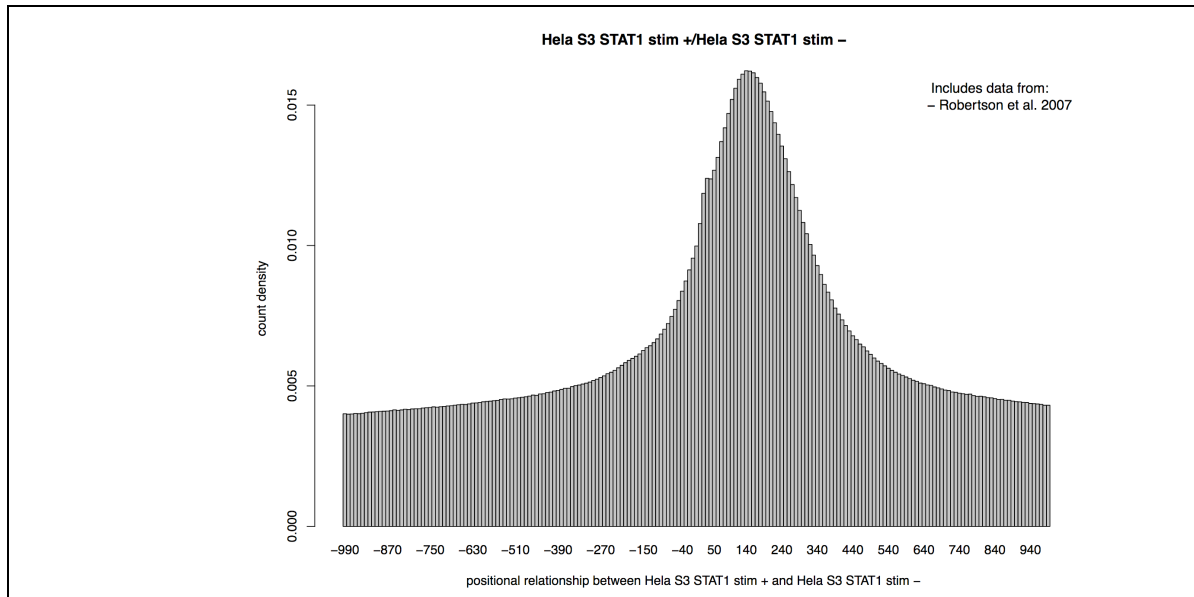


Figure 1. 5'-3'end correlation with ChIP-Cor.

ChIP-Cor offers several options for scaling the abundance of the target feature. Here, we have chosen 'count density', which is defined as the number of target feature tags per base pair. We note a Gaussian peak with a maximum at about position +150, suggesting that the average length of an immunoprecipitated fragment is about 150 bp. In all subsequent analyses, we will therefore use half of this value (75 bp) as centering distance for jointly analyzing 5' and 3' tags. Centering means shifting the positions of tags mapping to the + or - strand of the chromosome by a fixed distance downstream and or upstream, respectively. Centering increases the resolution of the ChIP-Seq data. On the left side, there is a weak shoulder at about position +35 which results from a common artifact seen in almost all ChIP-Seq experiments. We can generate the same plot for the control data set ('HeLa S3 STAT1 unstim' sample) and compare the two distributions. The ChIP-Seq tag distribution for the control data set is similar to the background distribution (Figure 2).

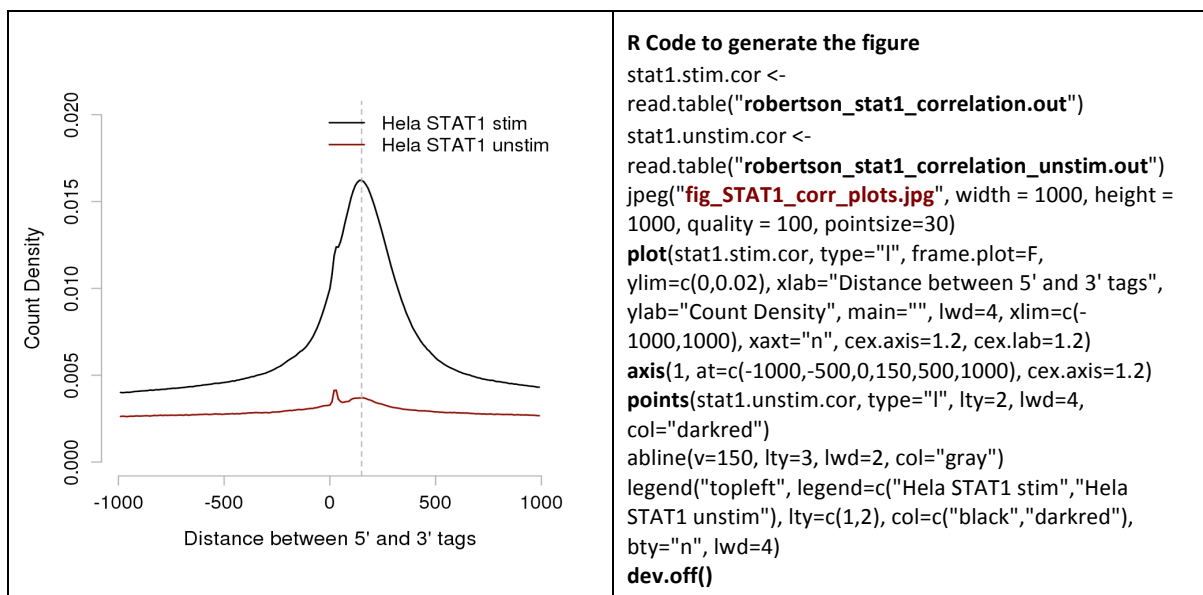


Figure 2. 5'-3'end correlation plot for STAT1 ChIP-seq tags vs control dataset.

The composite Figure 2 has been generated with the R graphics package. For using R, you first need to download the data files from the ChIP-Cor output page.

By clicking on the link ‘TEXT’ save the results as a text file named:

robertson_stat1_correlation.out

For the negative control sample, re-run ChIP-Cor for sample ‘Hela S3 STA1 unstim’, and save the results as:

robertson_stat1_correlation_unstim.out

In Figure 2, as well as in all subsequent composite Figures, the R code is shown on the right.

Next, we generate a so-called autocorrelation plot for centered STAT1 tags against themselves (same reference and target feature).

The **step-by-step procedure** is summarized in Table 2.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
<p>Select available Data Sets</p> <p>Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ChIP-seq Series: Robertson 2007 Sample: HeLa S3 STAT1 stim</p> <p>Additional Input Data Options</p> <p>Strand: any Centering: 75</p> <p>Analysis Parameters</p> <p>Range: -1000 to 1000</p> <p>Histogram Parameters</p> <p>Window width: 10 Count Cut-off: 1 Normalization: count density</p>	<p>Select available Data Sets</p> <p>Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ChIP-seq Series: Robertson 2007 Sample: HeLa S3 STAT1 stim</p> <p>Additional Input Data Options</p> <p>Strand: any Centering: 75</p>

Table 2. Autocorrelation with ChIP-Cor.

The results are shown below:

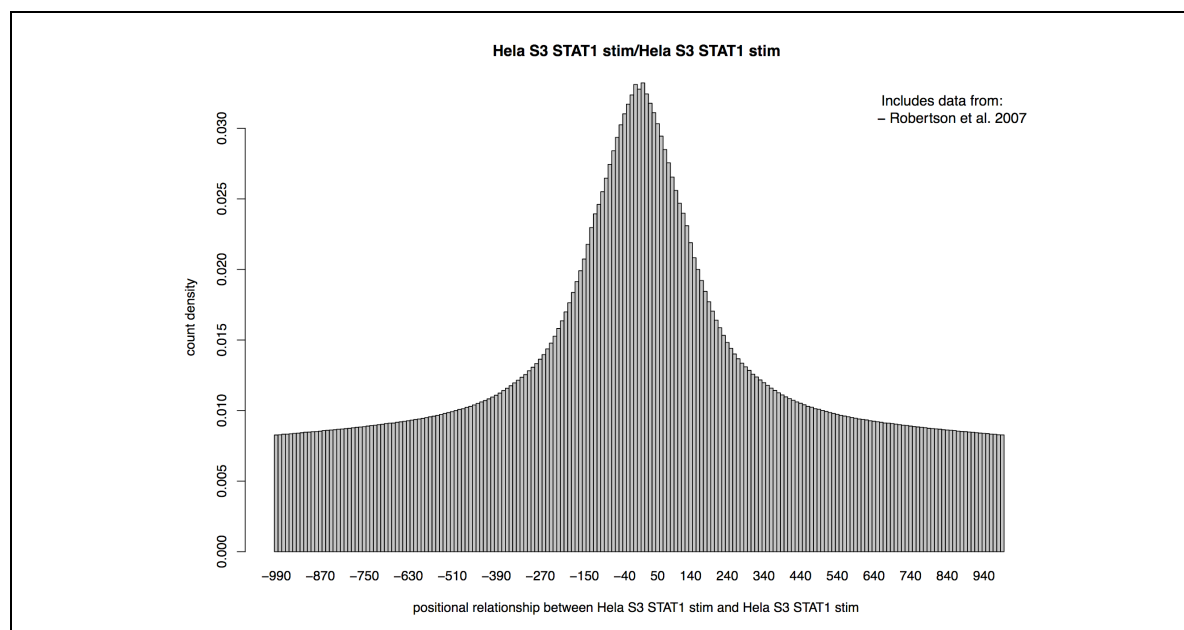


Figure 3. Autocorrelation with ChIP-Cor.

We see again a Gaussian peak this time with a maximum at 0. The **ChIP-Cor** server automatically attempts to fit the correlation histogram to a Gaussian curve. If successful, the results of the fit can be accessed via hyperlinks on the output page. Results are provided in both graphical and textual forms. The link 'Single Gaussian Fit' takes you to a figure showing an optimal Gaussian fit and the corresponding parameters (the parameter μ which is the mean of the distribution and corresponds to the peak center position, here $\mu = 0$).

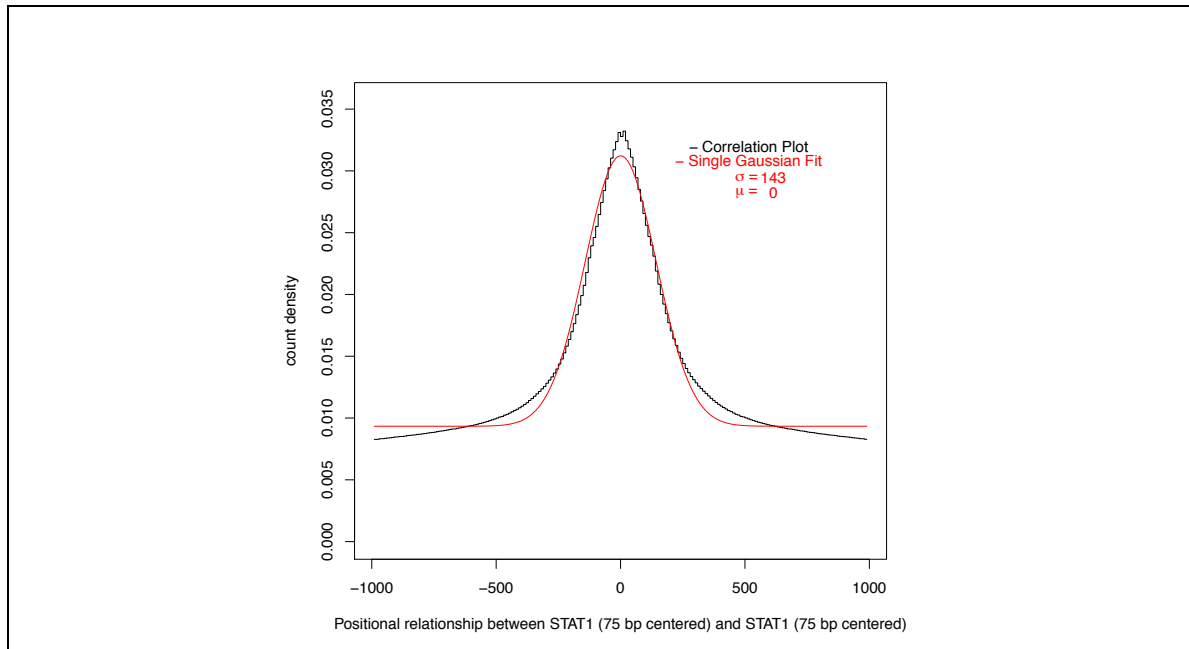


Figure 4. Gaussian fit of autocorrelation histogram.

2. Peak detection

We will use the ChIP-Peak program to identify peaks in the STAT1 data set. ChIP-Peak implements a simple window scanning algorithm. In essence, windows which contain more than a threshold number of tags and in addition constitute a local maximum within a certain distance range are reported as peaks. In contrast to other programs which report starting and ending positions of a peak region, ChIP-Peak returns single positions corresponding to peak centers. The Gaussian fit to the auto-correlation plot of the STAT1 data (Figure 4) suggests to use a window of 300 bp (approximately twice the standard deviation σ). The ChIP-Peak program on the Web allows users to set the peak threshold as either a tag count threshold or an enrichment factor relative to a mean background distribution of fragments across the whole genome. The mean background distribution probably represents a low estimate, as there are regions of the genome that are not mappable. A relative fold enrichment of 10 is a good first guess in most cases.

To generate a peak list, go to the **ChIP-Peak** input form at:

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

Run the default example, by clicking on the Example and subsequently on the Submit buttons. **ChIP-Peak** returns **14'113** peaks.

The peak lists are posted in three formats: SGA, FPS, and BED. SGA is the native format of the ChIP-Seq server, FPS is used by the SSA server and BED is a general format understood by many other web-based bioinformatics resources potentially useful for follow-up analysis (e.g. gene enrichment analysis).

If the input relates to a supported genome assembly, like in this case, a number of additional action buttons will be displayed on the **ChIP-Peak** output page (Figure 5). Hyperlinks are provided for sending the peak list directly to external servers for peak annotation, and for viewing the results in a UCSC genome browser window. The ‘Sequence Extraction Option’ enables users to extract sequences around the peak centers in Fasta format. Direct navigation buttons enable downstream analysis with other tools from the ChIP-Seq and SSA servers.

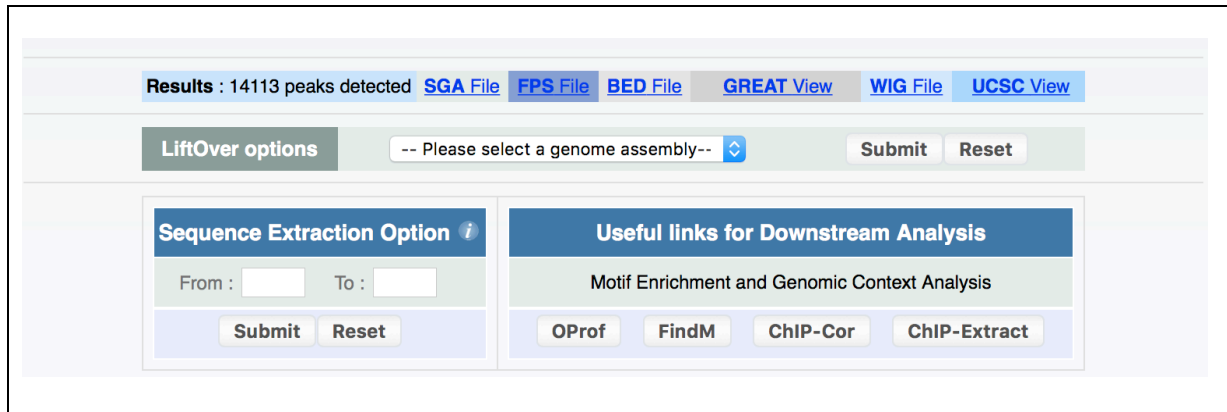


Figure 5. ChIP-Peak output page.

As we will see in the following Paragraphs, the direct navigation buttons which connect web server output pages to input forms allow users to carry out complex analysis tasks with a minimal number of mouse clicks.

3. Peak analysis with external tools

3.1 Viewing STAT1 peaks in the UCSC genome browser

The link to the UCSC browser enables the user to view individual STAT1 peaks in the context of other genomic features.

If you are interested in a particular gene, you can explore the genomic context of nearby peaks by using the direct hyperlink to the UCSC genome browser provided by the ChIP-Peak results page. This link will automatically load a custom track file in [WIG](#) format.

Step-by-step procedure

- Go to the bookmarked ChIP-Peak output page or simply rerun ChIP-Peak with the example provided on the input form.
- Navigate to the UCSC browser using the hyperlink ‘UCSC view’.
- Enter ICAM1 into the text area near the top of the browser page. The next page displays a list of hyperlinks to genes matching to the query. Click on any of these links that seems appropriate according to the gene description.
- On the next page, you will see a green bar near the 5' end of the gene. Zoom in on this region by moving the pointer to the top of the image slightly to the left of the STAT1 peak. Now you select the genomic region you want to enlarge with the left mouse button. Then zoom out again to view a region of about 2kb using the “zoom out” button on the right side near the top of the page. To precisely reproduce the image shown in Figure 6, type “chr19:10,381,000-10,383,000” into the text area just left to the “go” button.

- To provide a detailed view of ChIP-seq peaks identified by the ENCODE consortium, right-click on the track named "Txn factor ChIP" on the left side of the image and select "pack". This will change the display mode of the track. If you don't find the "Txn factor ChIP" track in the image, look in the track menu below the image for the "ENCODE Regulation ..." super-track under the heading "Regulation" and click on the corresponding hyperlink. A menu will appear where you can set the display mode of the "Txn factor ChIP" track to "pack".

Once you have gone through the above steps, you should see something like the following picture (Figure 6) in your UCSC browser window:

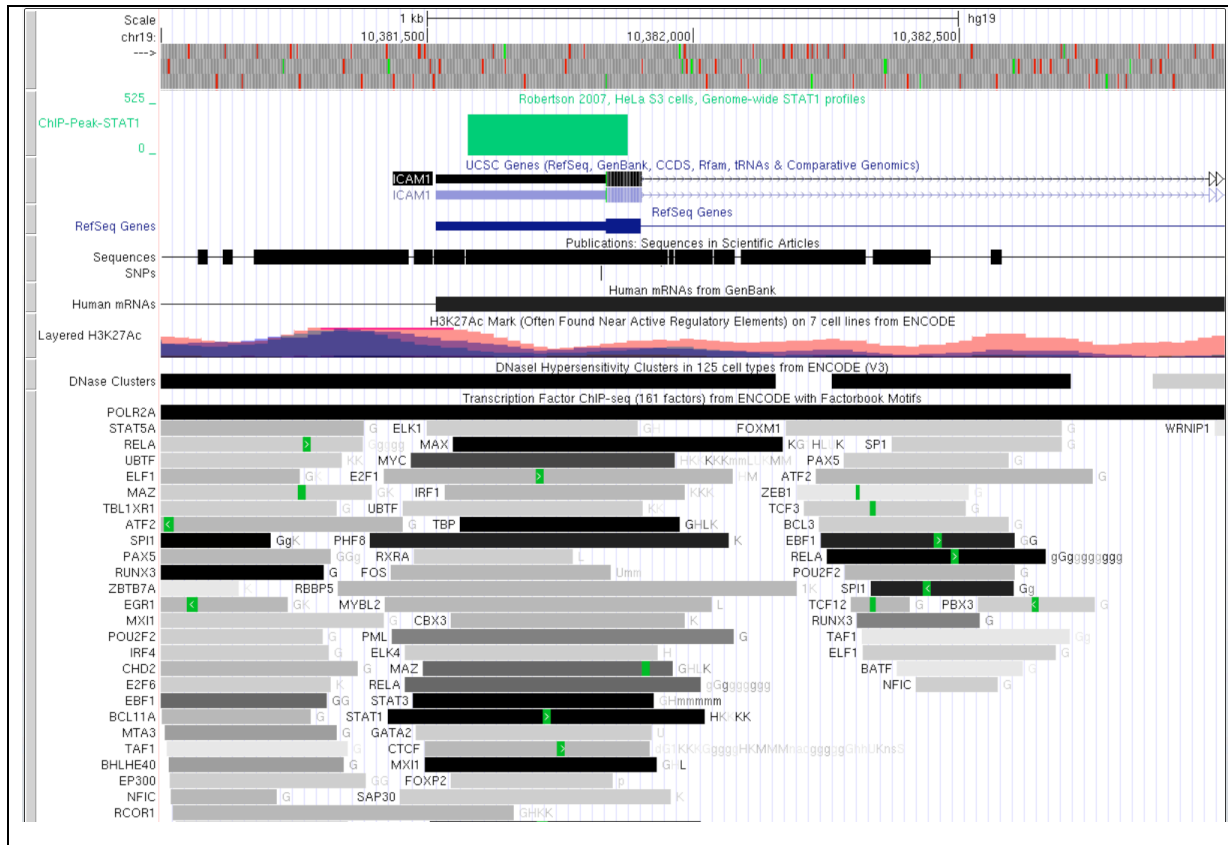


Figure 6. ICAM1 promoter region at UCSC.

Note the STAT1 peak near the bottom of the picture. The grey-scale color of the bar reflects the peak strength, the small green area at the center of the bar represents a STAT1 motif match. Clicking on the peak icon will open a new window with more information in STAT1 ChIP-seq assays carried out by the ENCODE consortium.

You may be surprised to see so many other peaks in the ICAM1 promoter region. Not all of them may be biologically relevant. Keep also in mind that they come from ChIP-seq assays carried out in many different cell types. Nevertheless, look at some of them that appear interesting. You may also look more closely at other tracks, for instance the conservation track ("Conservation" under "Comparative Genomics) or SNP track ("Common SNPs(147)" under "Variation"). Are there common SNPs near the peak center? Is the STAT1 motif indicated in the ENCODE track conserved across vertebrates?

3.2 GO term enrichment analysis with GREAT

The **ChIP-Peak** output page may also contain a hyperlink to the **GREAT** server ([McLean, et al., 2010](#)), which performs GO enrichment term analysis of the genes in the neighborhood of the peaks. A mouse-click on this link returns the Annotation terms that are significantly associated with the set of input STAT1 peaks. This table is sorted in order of decreasing Binomial p-value and can be visualized as a bar chart of sorted p-values as shown here below:

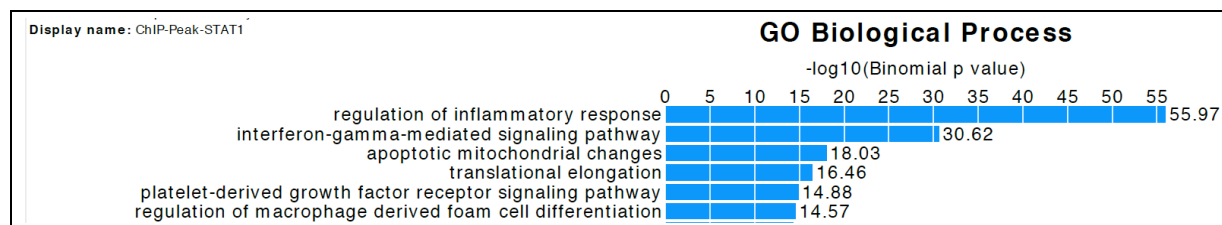


Figure 7. GO term enrichment analysis with GREAT.

We note that the majority of terms relate to interferon gamma-mediated signaling consistent with the reported biological function of STAT1. Note that due to file size restrictions at the GREAT site, the link to GREAT cannot always be provided on the ChIP-Peak output page. You can anyway download the BED file and upload it to the GREAT server at <http://bejerano.stanford.edu/great/public/html/>.

3.3 Genomic annotation of ChIP-seq peaks

Another topic of interest is the location of TF binding peaks relative to protein coding genes. One web-based resource performing such an analysis is **Nebula** ([Boeva, et al., 2012](#)). It returns graphics showing the abundance of peaks within promoter regions, gene bodies, intergenic regions, and components of genes (intron, exons, etc.). Note that the **Nebula** server might not be always working properly in which case you may try later.

Step-by-step procedure

We first run ChIP-Peak with threshold 50 (tag counts) and Repeat Masker on. The peak lists are posted in three formats, SGA, FPS, and BED. For Nebula, we copy the link location of the peak list in BED format (by right-clicking on the link ‘BED File’).

Go to [Nebula](#) and follow the instructions given below:

- Import the BED-formatted peak file generated with ChIP-Peak by selecting ‘Upload File’ under ‘Get Data’ (on the left side of the window), paste the BED file link location in the designated URL/Text input text area, and select bed as file format and Human Feb. 2009 (GRCh37/hg19) as genome version. The uploaded file will appear on the right side of the page.
- On the left side, select ‘NGS: Peak Annotation’ under ‘NGS TOOLBOX’. A number of tools will appear. Choose ‘Genomic annotation of Chip-Seq peaks’. A new menu will appear. Use default parameters and click on ‘Execute’. The results files (text and image) will appear on the right side.

Results are shown in Figure 8.

Nebula also returns a peak annotation table, indicating for each peak the nearest gene and its relative location to that gene. This dataset can be downloaded by clicking on the download icon under ‘Annotated Peaks’. Figure 9 shows a snapshot of the tabular output.

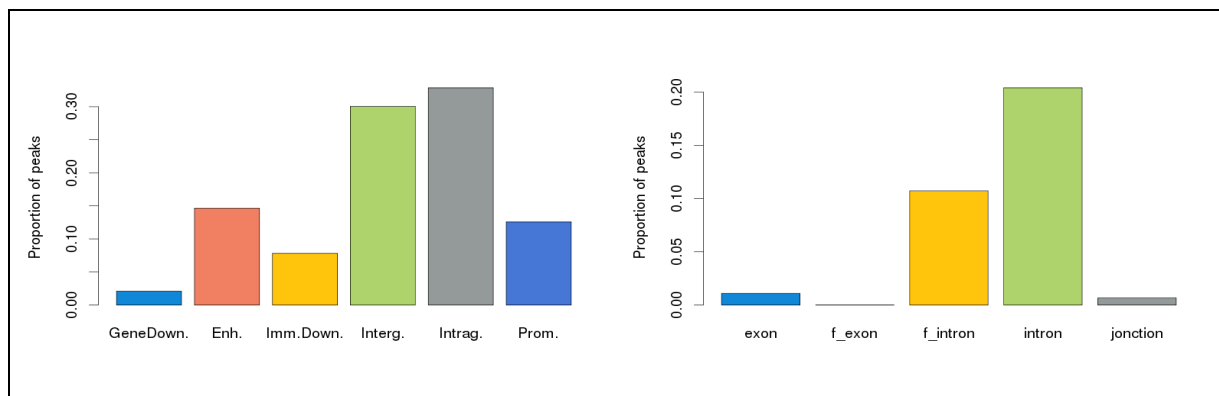


Figure 8. Peak location statistics with Nebula.

Chromosome	Start	End	Max Score	DistTSS	Type	TypeIntra	Reg	FoldChange	DistTE	GeneName	GeneCoordinates	
chr1	566839	566840	566839	59	55195	intergenic	NA NA	1	54256	OR4F3	chr1:621095-622034;621095,-622034,	
chr1	1310757	1310758	1310757	51	-195	promoter	NA NA	1	-1647	AURKAIP1	chr1:1309110-1310562;1309110,1309379	
chr1	1368774	1368775	1368774	58	-26081	enhancer	NA NA	1	-31499	MRPL20	chr1:1337275-1342693;1337275,1341188,134	
chr1	2321977	2321978	2321977	160	-1236	promoter	NA NA	1	-14908	RER1	chr1:2323213-2336885;2323213,2327222,232	
chr1	2460650	2460651	2460650	95	1034	immediateDownstream	exon	NA	1	-467	HES5	chr1:2460183-2461684;2460183
chr1	3624812	3624813	3624812	54	10203	intragenic	intron	NA	1	-27953	TP73	chr1:3614609-3652765;3614609,3624112
chr1	6221659	6221660	6221659	108	18535	intragenic	intron	NA	1	-59813	CHD5	chr1:6161846-6240194;6161846,6166293
chr1	6294685	6294686	6294685	227	1359	immediateDownstream	intron	NA	1	-13433	ICMT	chr1:6281252-6296044;6281252
chr1	6464907	6464908	6464907	106	-19940	enhancer	NA NA	1	-56097	ESPN	chr1:6484847-6521004;6484847,6488285,650	
chr1	7727870	7727871	7727870	97	882487	intragenic	intron	NA	1	-101896	CAMTA1	chr1:6845383-7829766;6845383,6880240

Figure 9. Peak list annotated with nearest genes (Nebula).

4. Motif studies in peak regions

STAT1 is known to bind to a DNA motif approximately described by the consensus sequence TTCNNNGAA. If the peaks found by **ChIP-Peak** were real binding sites, one would expect this motif to be over-represented near the peak center positions. In fact, motif enrichment analysis is commonly used for benchmarking the performance ChIP-Seq peak finders ([Wilbanks and Facciotti, 2010](#)). The **OProf** program of the SSA server can be used for this purpose. It returns a graph showing the percentage of sequences containing a motif in a sliding window around a reference position.

Step-by-step procedure

Go to [ChIP-Peak](#), click on Example, set the 'Peak Threshold' enrichment factor to 10 and, subsequently, Submit. The **ChIP-Peak** output page provides a direct link to the [OProf](#) program via an action button located in the lower right corner under 'Useful links for Downstream Analysis'. Clicking on the button will take you to the **OProf** input form with the list of peaks already loaded as a server-resident FPS file. Complete the **OProf** input form as shown in Table 3.

SSA Input Data	Signal Description
Sequence input via server-resident FPS Files	Consensus seq
Name(s): <i>hourly/chippeak_*.fps</i>	TTCNNNGAA
Sequence Range	Mismatches: 0
Entire sequence range: unchecked	Name: TTCNNNGAA
5'border: -499 3'border: 500	Reference Position: 5
Sliding window parameters	
Window size: 100	
Window shift: 5	
Search mode: bidirectional	

Table 3. Motif studies with OProf.

We use the search mode ‘bidirectional’ because ChIP-Seq peaks are un-oriented. In the output, you can see a clear peak centered at position 0.

Results from OProf can be saved in textual format under the name:

stat1_tr-10e.dat

To select different peak thresholds, you can go back to the **ChIP-Peak** output page and click on the ‘ChIP-Cor’ action button located in the lower right corner under ‘Useful links for Downstream Analysis’. Similarly to what happened with OProf, clicking on the button will take you to the **ChIP-Cor** input form with the list of peaks loaded as both the reference and the target features.

We first look at ChIP-seq tag coverage around peak regions using the server-resident dataset from ([Robertson et al., 2007](#)). To this end, complete the ChIP- Cor input form as shown in Table 4.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
<p>Server-resident SGA Files by Filename</p> <p>Filename: <i>hourly/chippeak_*.sga</i></p> <p>Experiment: Custom SGA</p> <p>Feature: (blank)</p> <p>Additional Input Data Options</p> <p>Strand: any</p> <p>Analysis Parameters</p> <p>Range: -1000 to 1000</p> <p>Histogram Parameters</p> <p>Window width: 10</p> <p>Count Cut-off: 1</p> <p>Normalization: count density</p>	<p>Select available Data Sets</p> <p>Genome: H. sapiens (Feb 2009 GRCh37/hg19)</p> <p>Data type: ChIP-seq</p> <p>Series: Robertson 2007</p> <p>Sample: HeLa S3 STAT1 stim</p> <p>Additional Input Data Options</p> <p>Strand: any</p> <p>Centering: 75</p>

Table 4. Peak threshold selection with ChIP-Cor.

On the **ChIP-Cor** output page we see, as expected, a high peak centered at 0. At the bottom of the page appears a new menu under the title ‘**Feature Selection Tool**’. It is the input form to a program (*chipscore*) that extracts reference feature coordinates that are enriched in target feature tags according to a given tag count threshold or score. We will therefore use the ‘**Feature Selection Tool**’ to extract STAT1 peaks with different tag count thresholds.

To this end, fill the ‘**Feature Selection Tool**’ menu as shown in table 5 and click on Submit.

Feature Selection Tool	
From : -150	To : 150
Threshold:	15
Cut-Off:	1
Switch to Depleted Feature Selection	off
Ref Feature oriented	off
Select Top/Enriched/Depleted Ref Feature	(blank)
Genomes	H. sapiens (Feb 2009 GRCh37/hg19)

Table 5. Select different peak thresholds using ‘Feature Selection Tool’.

This returns a peak list of **38’347** sites. Note that, if the ‘peak refinement’ option is set, the number of peaks returned by running **ChIP-Peak** with a tag count threshold of 15 does not correspond to the number of sites returned by re-scoring the peak list with ‘**Feature Selection**

Tool. If peak refinement is selected, the peak finding program re-computes the position of each peak by taking the center of gravity of the counts within the peak region defined by the ‘window’ parameter. Thus, the refined peak regions might have a different tag-count score than the original one.

Similarly to **ChIP-Peak**, the ‘**Feature Selection Tool**’ output page provides a direct link to the OProf program via an action button located in the lower right corner under ‘Useful links for Downstream Analysis’. Clicking on the button will take you to the OProf input form with the list of peaks already loaded as a server-resident FPS file. Complete the OProf input form as shown in Table 3 to compute the STAT1 motif occurrence profile.

Repeat the same type of analysis with the peak lists obtained with enrichment factor values 20, 35 and 75. The combined results are shown in Figure 9. Results from OProf can be saved in textual format under the names:

stat1_tr-10e.dat, stat1_tr-20e.dat, stat1_tr-35e.dat, stat1_tr-75e.dat

Figure 10 shows the motif occurrence profiles for TTCNNNGAA for the five different STAT1 peak lists obtained with different peak thresholds (10e, 20e, 35e, 75e). With all peak lists, we see a clear enrichment of STAT1 motifs near position zero (the reported peak center). As expected, the peak height is inversely correlated to the number of the peaks.

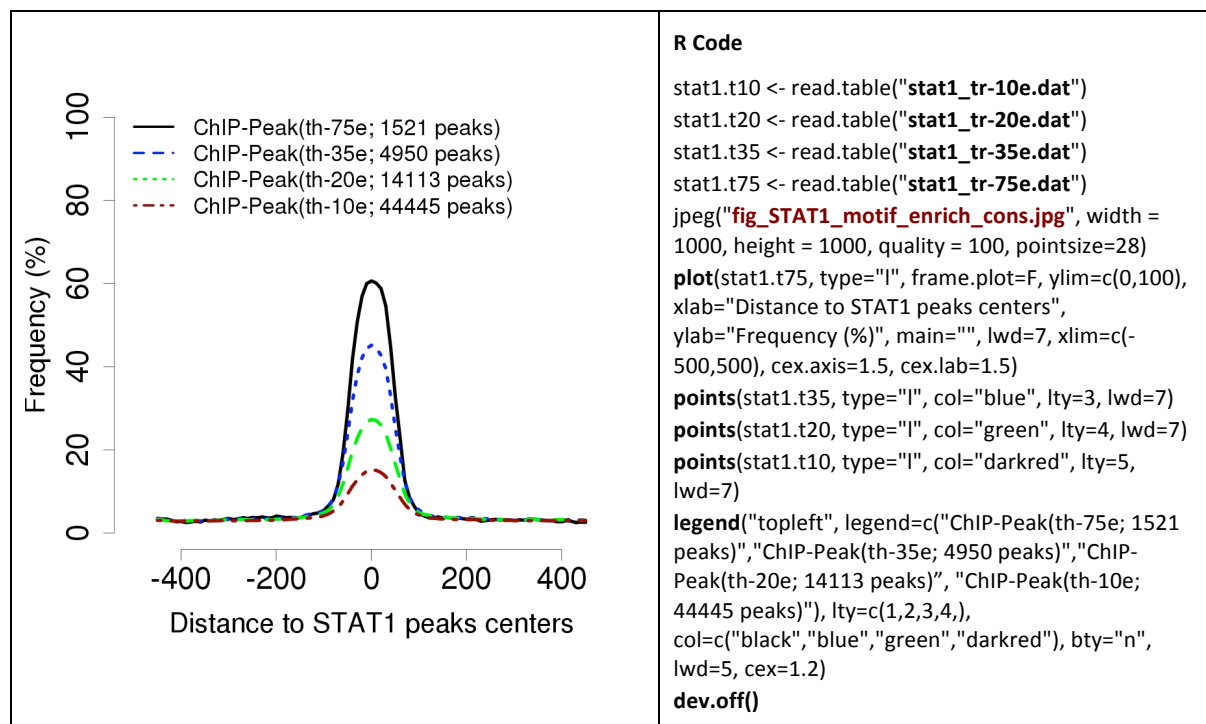


Figure 10. Peak list evaluation by motif enrichment. STAT1 consensus sequence (TTCNNNGAA) enrichment in peak lists obtained at various tag thresholds.

The **OProf** server provides access to a large number of functional sites collections, including two STAT1 peak lists by the ENCODE consortium, one from HeLa and the other from K562 cell lines ([Wang et al., 2012](#)).

Step-by-step procedure

Go to:

- <http://cgg.vital-it.ch/ssa/oprof.php>

Then fill out the input form as shown in Table 5.

SSA Input Data	Signal Description
<p>Select available Data Sets</p> <p>Genome : H.sapiens (Feb2009/hg19) Data type : ENCODE ChIP-seq-peak Series : Wang et al. 2012, Transcription Factor Binding Sites from ENCODE Stanford/Yale/USC/Harvard Sample : HeLa-S3 STAT1 std - IFNg30 - peaks</p> <p>Sequence Range</p> <p>Entire sequence range: unchecked 5'border: -499 3'border: 500</p> <p>Sliding window parameters</p> <p>Window size: 100 Window shift: 5 Search mode: bidirectional</p>	<p>Consensus seq</p> <p>TTCNNNGAA Mismatches: 0</p> <p>Name: TTCNNNGAA Reference Position: 5</p>

Table 5. Motif studies with OProf (use ENCODE data).

Carry out the same analysis on sample 'K562 - STAT1 std - IFNg30 – peaks'.

Save the results as a text file named respectively: **Hela-Stat1_encode.dat**, and **K562-Stat1_encode.dat**. Figure 11 shows consensus sequence enrichment profiles for these peaks lists and the ones generated by ChIP-Peak.

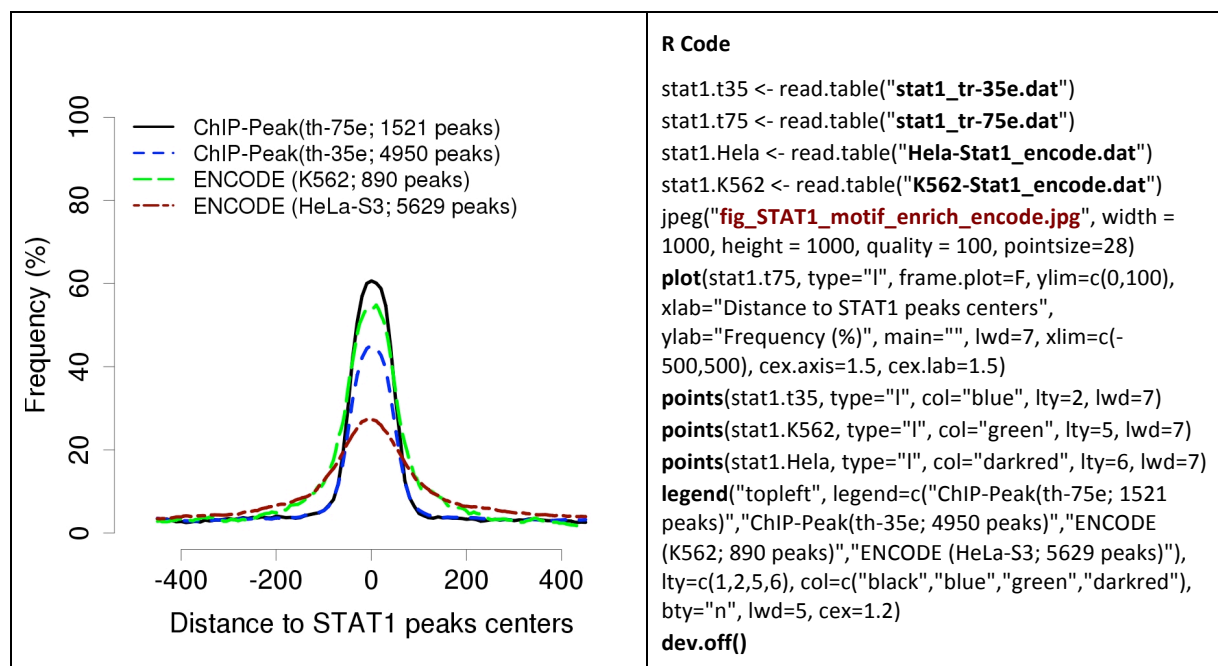


Figure 11. Peak list evaluation by motif enrichment. Peak lists derived in this tutorial versus Peak lists from ENCODE (TTCNNNGAA).

We note that our lists generated from earlier data compares favorably to the ENCODE peak lists, both in terms of enrichment (peak volume) and positional resolution (peak width).

For most TFs, consensus sequences can only provide approximations of the true binding motifs. Position weight matrices (PWMs) are widely used to describe the binding specificity of TFs. The **OProf** server provides menu-driven access to PWMs from several public resources, including a STAT1 matrix from the JASPAR database.

To search the JASPAR MA0137.3 STAT1 motif (Figure 12), access the **OProf** program by first running ChIP-Peak on the Robertson STAT1 data with relative enrichment factor threshold of 35. On the results page, use the ‘OProf’ action button to directly go to the **OProf** input form and have the peak list already loaded. Complete the **OProf** input form following the instructions given in Table 6, and click Submit.

SSA Input Data	Signal Description
Sequence input via server-resident FPS Files Name(s): <i>hourly/chippeak_*.fps</i> Sequence Range Entire sequence range: unchecked 5'border: -499 3'border: 500 Sliding window parameters Window size: 100 Window shift: 5 Search mode: bidirectional	PWMs from Library Motif Library: JASPAR CORE 2016 vertebrates Motif: STAT1 MA0137.3 (length=11) Cut-off: p-value Value: 0.0001 Name: MA0137.3 STAT1 Reference Position: 6

Table 6. Motif studies with OProf (use the JASPAR STAT1 matrix).

Save the results as: **stat1_jaspar2016_tr-35e.dat**

5. De novo motif discovery with MEME-ChIP

To search the MEME-derived motif (Figure 12), we use the **MEME-ChIP** server . The sequence file provided as input to the MEME-ChIP server can be generated using the sequence extraction tool on the ChIP-Peak output page.

Step-by-step procedure

Go to the **ChIP-Peak** page at:

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

Fill it out as shown in Table 7.

ChIP-Seq Input Data	Peak Detection Parameters
Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ChIP-seq Series: Robertson 2007 Sample: HeLa S3 STAT1 stim Additional Input Data Options Strand: any Centering: 75 Repeat Masker : checked	Window Width (bp): 300 Vicinity Range (bp) : 300 Peak Threshold : relative enrichment factor (checked) 75 Count Cut-off : 1 Refine Peak Position : checked Genome Viewing Parameters : Wig Track name : blank Chromosome Region : blank

Table 7. Extract peak lists with ChIP-Peak (motif discovery with MEME-ChIP).

We use the repeat-masked peak list obtained with enrichment factor threshold 75. Extract sequences from position -60 to +60 relative to the peaks center, using the menu that appears at the bottom of the **ChIP-Peak** result page:

Sequence Extraction Option From : -60 To: +60

This returns a list of sequences around the peaks in FASTA format. A new link ‘Sequence File’ will appear. Open this link.

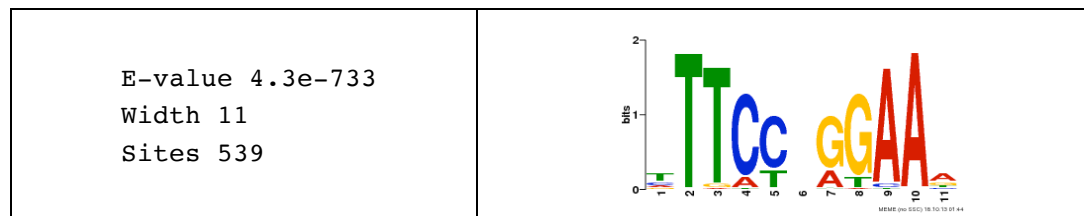
Now, open the home page of the **MEME** Suite in another browser window:

- <http://meme.sdsc.edu>

Click on the **MEME-ChIP** link near the lower right corner of the page. Enter your email address, then copy&paste the sequences from the open **ChIP-Peak** window into the text area of the **MEME-ChIP** page that serves for sequence input.

Since we expect the STAT1 binding motif to be palindromic, we restrict the **MEME-ChIP** search to palindromic motifs. Keep all the default parameters except the maximum motif width (=15), and check the box near ‘look for palindromes only’.

On the **MEME-ChIP** result page, click on the link ‘MEME-ChIP html output’ to get the result page. Here is the main motif found by MEME:



You can expand all clusters to show all motifs. Besides the graphical display of motif 1, you can click on the link “MEME” to view the MEME output. At the “Discovered Motifs” section, beside the logo of motif 1, you click on the arrow “Submit/Download”. A pop-up window will appear allowing you to download the motif. There, click on "Download Motif" and select the “Minimal MEME” format. Both probability and log-odds matrices will be displayed. The ‘log-odds’ scoring matrix is the following:

-74	-1	-116	97
-808	-448	-506	196
-462	-606	-261	191
-82	174	-606	-350
-450	158	-342	-17
3	-3	-3	3
-17	-342	158	-450
-350	-606	174	-82
191	-261	-606	-462
196	-506	-448	-808
97	-116	-1	-74

Keep this window open and access the **OProf** program by first running ChIP-Peak on the Robertson STAT1 data with relative enrichment factor threshold of 35. On the results page use the ‘OProf’ action button to directly go to the **OProf** input form and have the peak list already loaded.

Complete the **OProf** input form as shown in Table 8 here below.

SSA Input Data	Signal Description
<p>Sequence input via server-resident FPS Files</p> <p>Name(s): <i>hourly/chippeak_*.fps</i></p> <p>Sequence Range</p> <p>Entire sequence range: unchecked 5'border: -499 3'border: 500</p> <p>Sliding window parameters</p> <p>Window size: 100 Window shift: 5 Search mode: bidirectional</p>	<p>Custom Weight Matrix</p> <p>Format: Integer PWM PWM text: cut&paste matrix from MEME output into text area</p> <p>Cut-off: p-value Value: 0.00015</p> <p>Name: MEME-ChIP motif Reference Position: 6</p>

Table 8. PWM profile with OProf.

The p-value thresholds for all motifs have been chosen so as to get an equal background distribution for all profiles.

Save the results as: **stat1_MEME-ChIP_tr-35e.dat**

Figure 12 shows motif enrichment profiles for the STAT1 consensus sequence, the JASPAR matrix, and the *de novo* generated matrix. We note that PWMs show higher enrichment than the consensus sequence. However, there isn't an obvious difference between the matrix generated by MEME and the JASPAR matrix.

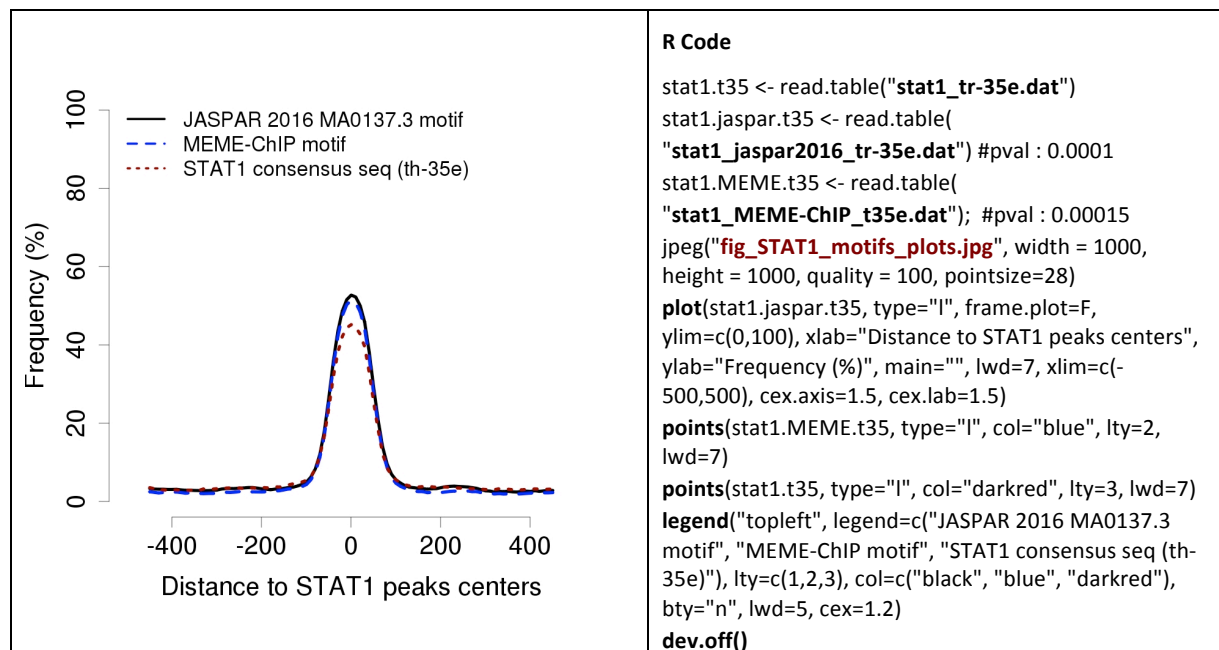
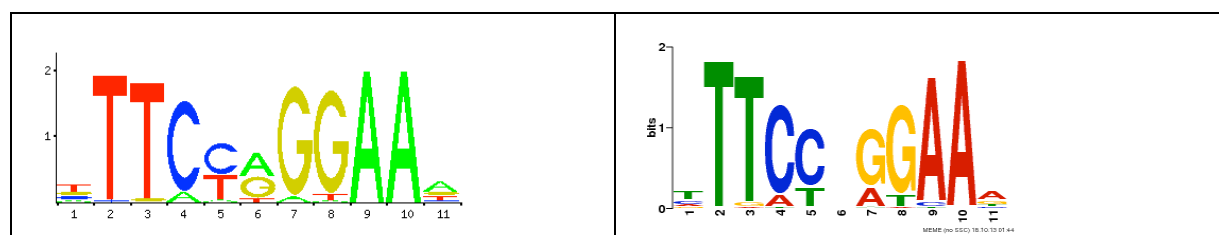


Figure 12. Peak list evaluation by motif enrichment. TTCNNGAA versus JASPAR and MEME-ChIP weight matrices for relative enrichment factor threshold 35.

The sequence logos for both the JASPAR and the MEME-ChIP matrices are shown here below:



JASPAR and MEME-ChIP PWM logos.

6. Searching peak regions for known motifs with CentriMo

TFs may cooperate with each other when bound to nearby target sites. Strong co-localization of diverse TF binding sites has been documented by many ChIP-seq studies. We therefore expect enrichment of other TF binding sites in the STAT1 peak regions. The program **CentriMo** from the MEME suite is well suited to discover such motifs.

CentriMo identifies known motifs from databases that show significant central enrichment in a set of uploaded sequences of equal length. **CentriMo** is much faster than MEME-ChIP. Therefore larger sequence sets can be submitted to the server. Note further that **CentriMo** requires longer input sequences than MEME-ChIP for effective assessment of central enrichment.

Step-by-step procedure

Run **ChIP-Peak** on the Robertson STAT1 data with the parameters shown in Table 9.

ChIP-Seq Input Data	Peak Detection Parameters
Select available Data Sets	Window Width (bp): 300
Genome: H. sapiens (Feb 2009 GRCh37/hg19)	Vicinity Range (bp) : 300
Data type: ChIP-seq	Peak Threshold : relative enrichment factor (checked) 75
Series: Robertson 2007	Count Cut-off : 1
Sample: HeLa S3 STAT1 stim	Refine Peak Position : checked
Additional Input Data Options	Genome Viewing Parameters :
Strand: any	Wig Track name : blank
Centering: 75	Chromosome Region : blank
Repeat Masker : checked	

Table 9. Extract peak list for CentriMo analysis.

On the results page, use the 'Sequence Extraction Option' menu to extract sequences from -300 to +300. Then, on the next page, open or save the sequences via the hyperlink labeled "Sequence file". Name the saved file: **stat1_centrimo_input.seq**.

Go to the **CentriMo** input form at:

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

You can upload the sequences from the STAT1 peak regions either via *copy&paste* or from disk ('Upload sequences'). If you choose *copy&paste*, then select 'Type in sequences' under the menu heading 'Input the primary sequences', otherwise select 'Upload sequences'.

The full results from the above analysis can be found [here](#). **CentriMo** returns a graphic showing the positional distribution of three motifs plus a list of all centrally enriched motifs ranked by P-value. The top of this list is shown in Figure 13 here below.

<input checked="" type="checkbox"/>	ID [?]	Name [?]	E-value [?]	Region Width [?]	Region Matches [?]
<input type="checkbox"/>	MA0518.1	Stat4	4.4e-423	82	782
<input checked="" type="checkbox"/>	MA0137.3	STAT1	5.7e-411	79	746
<input type="checkbox"/>	MA0144.2	STAT3	2.6e-345	77	703
<input type="checkbox"/>	MA0519.1	Stat5a::Stat5b	1.9e-271	85	637
<input type="checkbox"/>	UP00043_1	Bcl6b_primary	1.2e-133	54	394
<input type="checkbox"/>	MA0520.1	Stat6	9.2e-98	55	302
<input type="checkbox"/>	UP00049_1	Sp100_primary	3.0e-88	60	310
<input type="checkbox"/>	MA0463.1	Bcl6	1.0e-80	86	393
<input type="checkbox"/>	UP00090_1	Elf3_primary	1.6e-64	101	429
<input type="checkbox"/>	UP00015_1	Ehf_primary	6.6e-64	51	285
<input type="checkbox"/>	EHF_full		5.1e-62	88	320
<input checked="" type="checkbox"/>	MA0062.2	GABPA	1.6e-60	123	388
<input type="checkbox"/>	ELF3_DBD		8.0e-59	52	192
<input type="checkbox"/>	ELF5_full		2.9e-55	117	447
<input type="checkbox"/>	ELF4_full		6.2e-55	52	192
<input type="checkbox"/>	MA0076.2	ELK4	1.4e-54	109	367
<input type="checkbox"/>	Elf5_DBD		6.0e-53	103	410
<input type="checkbox"/>	ELF5_DBD		1.8e-52	101	395
<input type="checkbox"/>	ELF1_full		2.2e-49	64	173
<input type="checkbox"/>	ETS1_full_1		7.7e-48	114	391
<input type="checkbox"/>	ELF1_DBD		2.6e-47	66	191
<input checked="" type="checkbox"/>	MA0491.1	JUND	9.4e-47	271	366

Figure 13. List of top-ranked Motifs from CentriMO.

Unsurprisingly, the STAT1 motif appears among the top ranked motifs. Further below we find motifs from the EHF (GABPA) and AP1 (JUND) family of transcription factors.

CentriMo returns an interactive HTML page allowing the user to select the motifs to be displayed in the graphic. Figure 14 shows the positional distributions of the top-scoring motifs from the three transcription factor families STAT, ETS, and AP1.

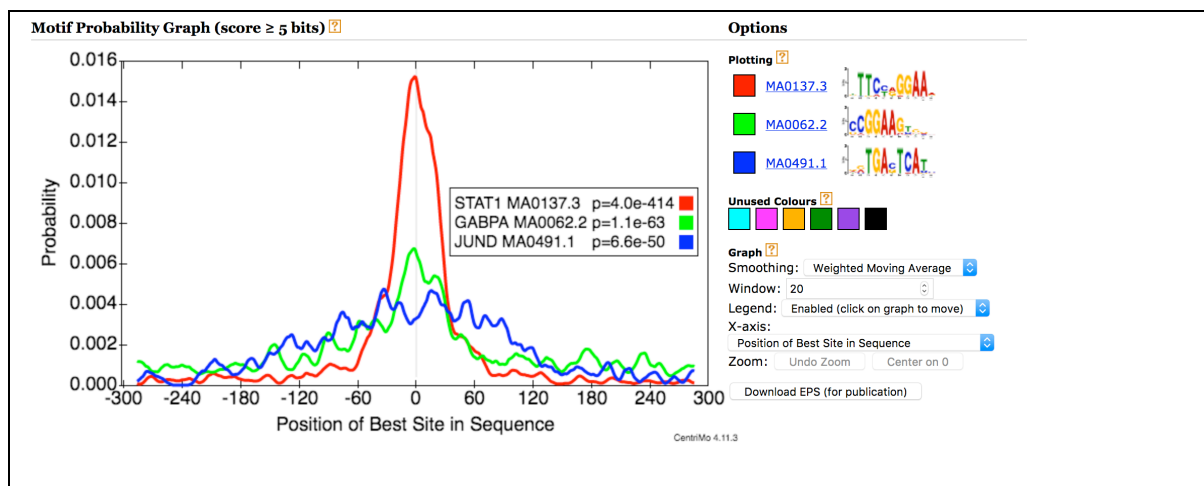


Figure 14. Motif probability graph (CentriMO).

We note that the distributions of the STAT1 motif (MA0137.3) and the ETS family EHF motif (GABPA) are similarly narrow whereas the AP1 family JUND motif (MA0491.1) has a broader distribution. Note further that the EHF motif resembles the STAT1 motif whereas the JUND motif is completely different. Taken together, this suggests that the central enrichment of the ETS-like motifs is the consequence of their similarity to the STAT1 motif and mostly likely results from direct overlaps with STAT1 motifs. On the other hand, since AP1 motifs are not expected to overlap STAT1 motifs, their enrichment near STAT1 binding sites may reflect true biological cooperativity.

7. Exploring the genomic context of STAT1 peaks

ChIP-Cor enables the user to generate aggregation plots (APs) for a great variety of target features from peak lists. We first investigate whether the STAT1 binding sites are associated with active or repressive histone marks. Since the STAT1 binding experiment was carried out in HeLa cells, we choose histone modification data from the same cell type generated by the ENCODE consortium. Specifically, we would like to test an active promoter mark (H3K4me3), an active enhancer mark (H3K27ac) and a repressive chromatin mark (H3K27me3). Remember in this context, that STAT1 peaks were discovered in HeLa cells that were stimulated with gamma-interferon. On the other hand, the histone modification maps from ENCODE were obtained from non-stimulated cells, in which STAT1 is not supposed to be bound to genomic target sites.

Step-by-step procedure

Run **ChIP-Peak** on the Robertson STAT1 data with the parameters shown in Table 10.

ChIP-Seq Input Data	Peak Detection Parameters
Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ChIP-seq Series: Robertson 2007 Sample: HeLa S3 STAT1 stim	Window Width (bp): 300 Vicinity Range (bp) : 300 Peak Threshold : relative enrichment factor (checked) 20 Count Cut-off : 1
Additional Input Data Options Strand: any Centering: 75	Refine Peak Position : checked Genome Viewing Parameters : Wig Track name : (blank) Chromosome Region : (blank)

Table 10. Extract peak list for genomic context analysis.

On the results page, use the action button ‘ChIP-Cor’ under ‘Useful links for Downstream Analysis’ to directly go to the **ChIP-Cor** input form with the peak list already loaded. For the following analysis it is convenient to keep this page open in a separate browser window for reference.

Complete the **ChIP-Cor** input form as shown in Table 11 here below.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
Server-resident SGA Files by Filename Filename: <i>hourly/chippeak_*.sga</i> Experiment: Custom SGA Feature: (blank)	Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ENCODE ChIP-seq Series: GSE29611, Histone Modifications by ChIP-seq Sample: Hela-S3 H3K4me3

Additional Input Data Options

Strand: any

Analysis Parameters

Range: -5000 to 5000

Histogram Parameters

Window width: 10

Count Cut-off: 1

Normalization: global

Additional Input Data Options

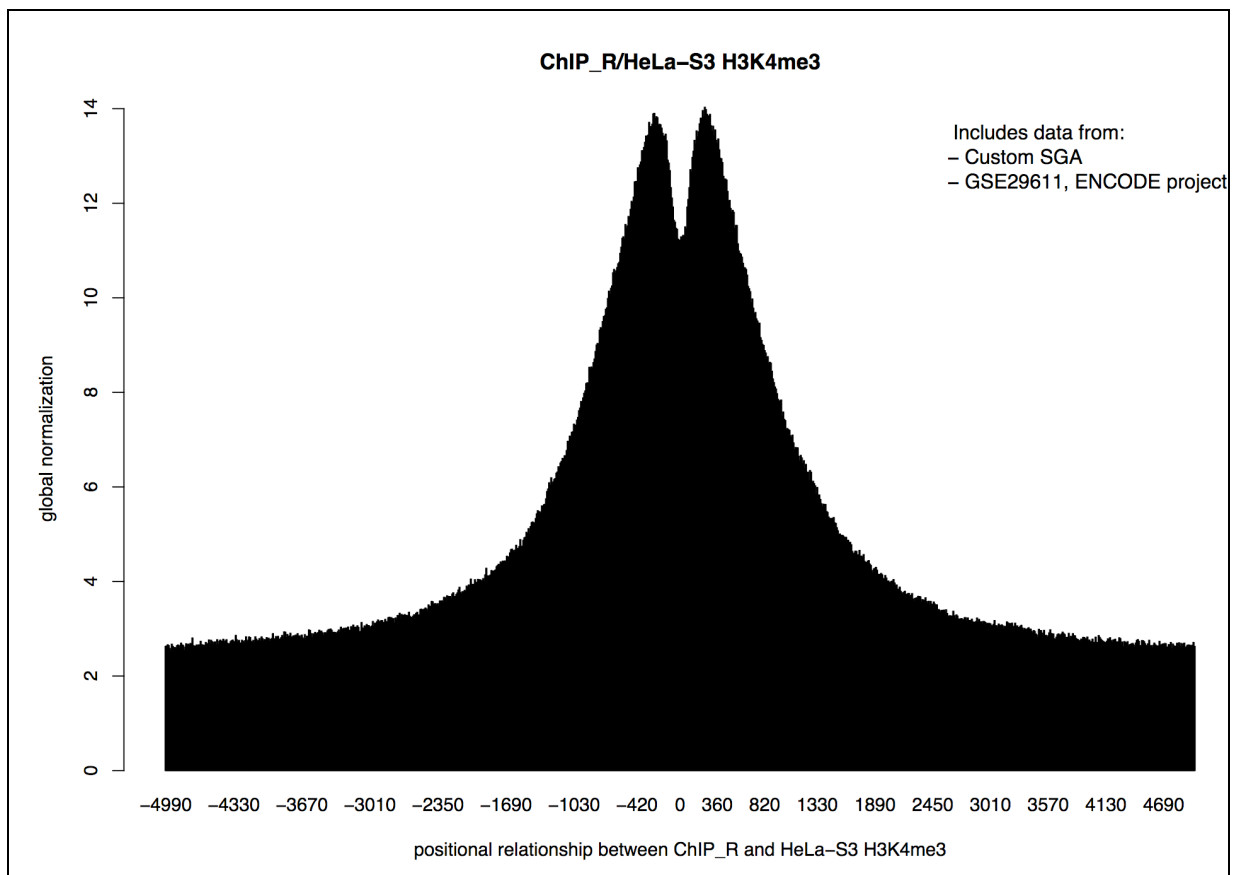
Strand: any

Centering: 70

Table 11. Histone modification profiles with CHIP-Cor.

Use 70 bp as the centering distance, as it would correspond to the average nucleosome DNA length (of about 140 bp). An aggregation plot shows the distribution of a particular genomic feature (e.g. a ChIP-seq signal) relative to a specified anchor point (e.g. a transcription start site) within a set of genomic regions.

The results are shown in Figure 15 here below:

**Figure 15. CHIP-cor output page. H3K4me3 profile around STAT1 peaks.**

Repeat the same procedure for samples 'HeLa-S3 H3K27ac' and 'HeLa-S3 H3K27me3'.

Save the results from the three runs under the following file names:

**robertson_stat1_H3K4me3.out, robertson_stat1_H3K27ac.out,
robertson_stat1_H3K27me3.out**

Figure 16 shows the results for the three histone marks.

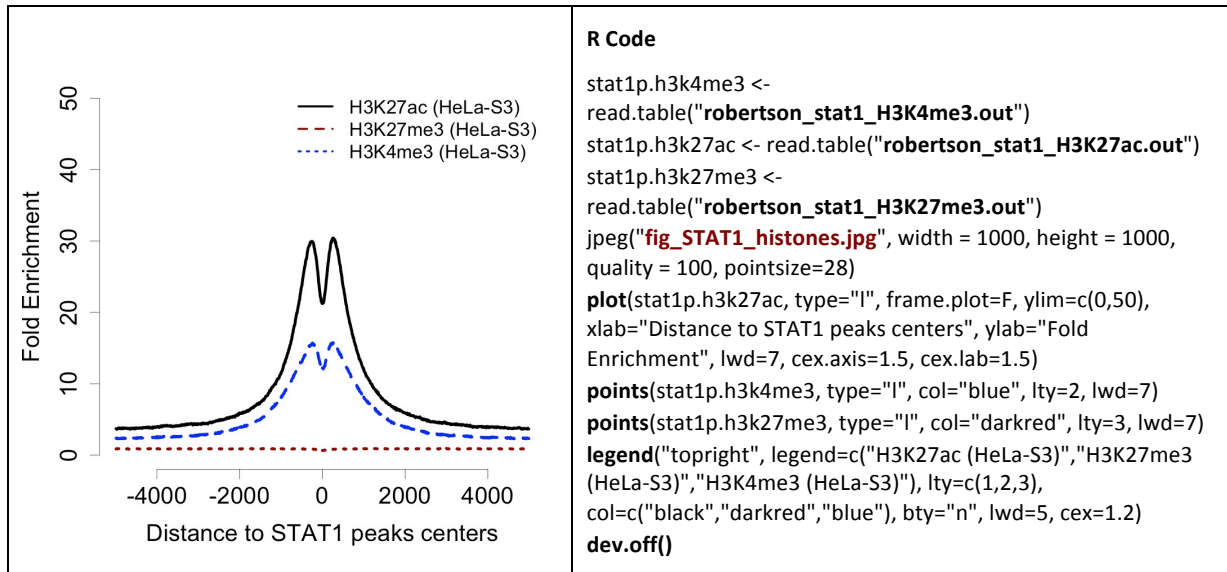


Figure 16. Histone marks around STAT1 peaks. Histone marks around HeLa STAT1 peaks in non-stimulated HeLa cells.

We see that STAT1 peaks fall into regions of about 500 base-pairs which are 15-fold enriched in H3K27ac and 7-fold in H3K4me3 compared to the background level. Conversely, no enrichment is seen for H3K27me3 in the vicinity of STAT1 peaks. These results suggest that STAT1 binds primarily to regions that are already in an active chromatin state before interferon induction. Note further the bimodal distribution of the active histone marks with maxima symmetrically positioned around the peak center. This may indicate that STAT1 preferentially binds to target sites that are nucleosome-free in un-stimulated cells.

We may wonder whether these STAT1 bound enhancers are also in an active state and nucleosome-free in other cell types. To answer this question, we generate APs for H3K27ac in the embryonic stem cell line H1-hESC and the leukemia-derived cell line K562.

Repeat the **step-by-step procedure** as described in Table 11 for samples: ‘HeLa-S3 H3K27ac’, ‘H1-hESC H3K27ac’, and ‘K562 H3K27ac’.

The aggregation plots for these cell lines are shown in Figure 17 together with the results obtained for HeLa cells.

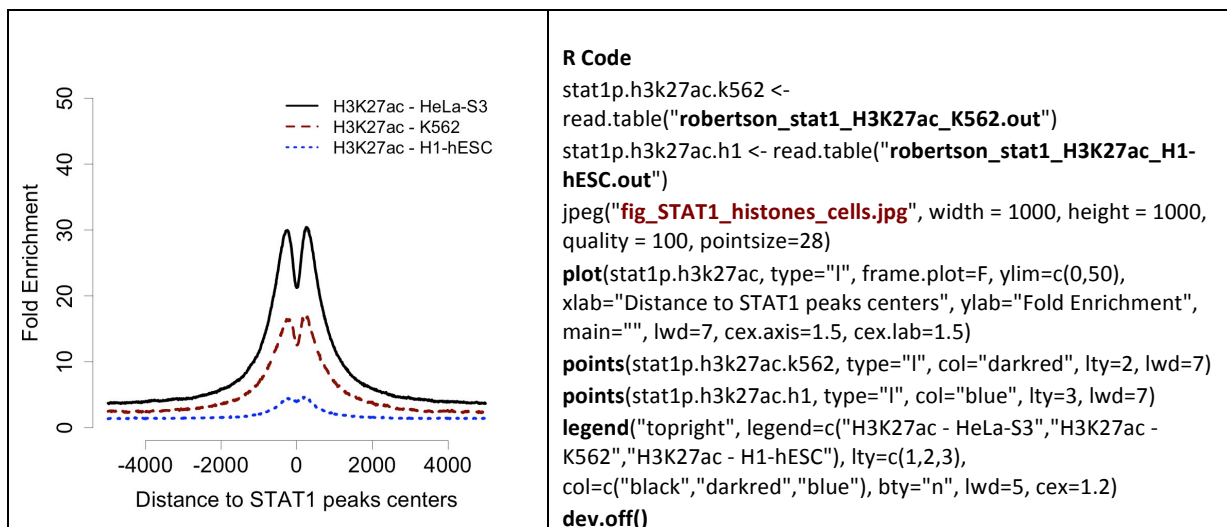


Figure 17. Histone marks around STAT1 peaks. H3K27ac marks in HeLa and other cell types.

We see an approximately two-fold higher enrichment in HeLa cells over K562 and an almost flat H2K27ac profile in H-hESC, suggesting a substantial degree of tissue-specificity of the regulatory regions that are bona fide accessible to STAT1 by virtue of their chromatin state.

In addition, we can explore DNase I hypersensitivity, sequence conservation, and population variation data near STAT1 sites.

To this end, we proceed as follows.

Step-by-step procedure

If necessary, repeat the step-by-step procedure described in Table 10 for extracting a STAT1 peak list with relative enrichment factor threshold of 20, otherwise go directly to the ChIP-Cor input form (via the action button on the ChIP-Peak output page) and complete the form as indicated in Table 12 here below.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
<p>Server-resident SGA Files by Filename</p> <p>Filename: <i>hourly/chippeak_*.sga</i></p> <p>Experiment: Custom SGA</p> <p>Feature: (blank)</p> <p>Additional Input Data Options</p> <p>Strand: any</p> <p>Analysis Parameters</p> <p>Range: -1000 to 1000</p> <p>Histogram Parameters</p> <p>Window width: 10</p> <p>Count Cut-off: 1</p> <p>Normalization: global</p>	<p>Select available Data Sets</p> <p>Genome: H.sapiens (Feb 2009/hg19)</p> <p>Data type: ENCODE DNase FAIRE etc.</p> <p>Series: Thurman 2012, DNaseI Hypersensitivity by Digital DNaseI from ENCODE/OpenChrom</p> <p>Sample: DNaseI HS – H1-hESC – None – Rep1</p> <p>Additional Input Data Options</p> <p>Strand: any</p> <p>Centering: (blank)</p>

Table 12. DNase I hypersensitivity with ChIP-Cor.

Do the same for samples: ‘DNaseI HS – HeLa-S3 – None- Rep1’, and ‘DNaseI HS – K562 – None- Rep1’. Note that DNase data do not require centering. We don’t center DNaseI tags because there is no shift according to 5’-3’ correlation plots (neither is there a biological reason for such a shift), given that DNase tags represent sites sensitive to cleavage by DNase I. Save the results under the following names: **robertson_stat1_DNaseI_HeLa.out**, **robertson_stat1_DNaseI_K562.out**, **robertson_stat1_DNaseI_H1-hESC.out**

The aggregation plots for DNaseI hypersensitivity are shown in Figure 18.

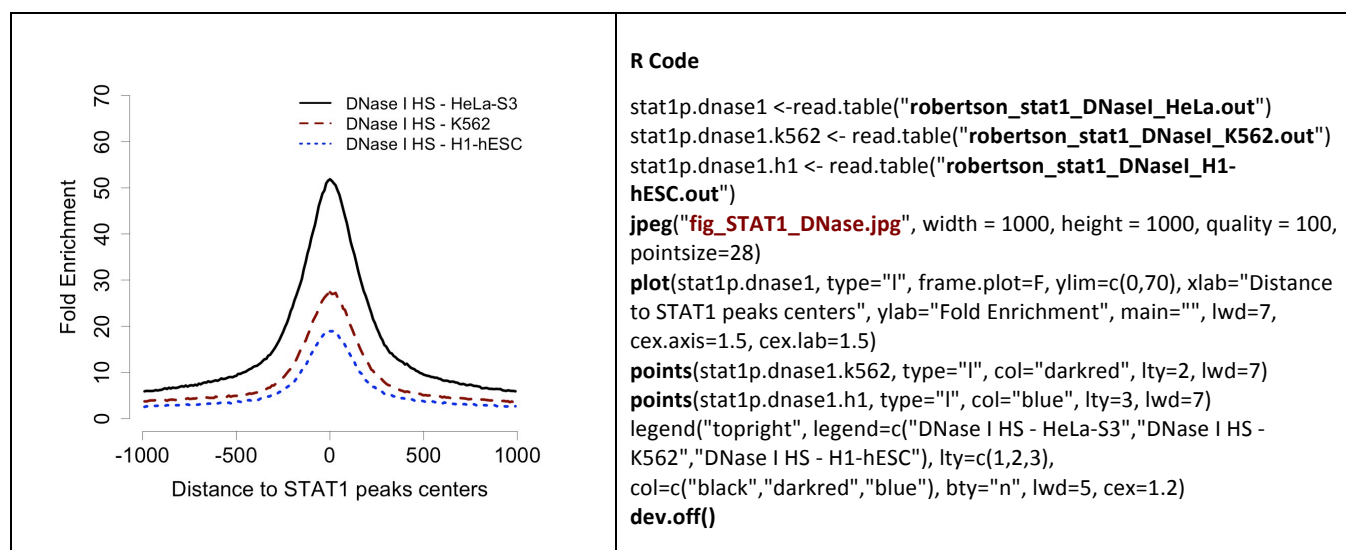


Figure 18. DNase I hypersensitivity around STAT1 peaks.

In summary, STAT1 peaks occur preferentially within DNaseI hypersensitive regions of about 500 bp.

For sequence conservation and population variation studies, you can proceed as follows.

Step-by-step procedure

Complete the ChIP-Cor input form as shown in Table 13 here below.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
<p>Server-resident SGA Files by Filename</p> <p>Filename: <i>hourly/chippeak_*.sga</i></p> <p>Experiment: Custom SGA</p> <p>Feature: (blank)</p> <p>Additional Input Data Options</p> <p>Strand: any</p> <p>Analysis Parameters</p> <p>Range: -1000 to 1000</p> <p>Histogram Parameters</p> <p>Window width: 10</p> <p>Count Cut-off: 10</p> <p>Normalization: global</p>	<p>Select available Data Sets</p> <p>Genome: H. sapiens (Feb 2009 GRCh37/hg19)</p> <p>Data type: Sequence-derived</p> <p>Series: Vertebrate conservation (phastCons46way)</p> <p>Sample: PHASTCONS VERT46</p> <p>Additional Input Data Options</p> <p>Strand: any</p> <p>Repeat Masker: unchecked</p>

Table 13. PhastCons conservation scores with ChIP-Cor.

Note that we chose a count cut-off of 10 because the PhastCons dataset is a special low-resolution representation of the original UCSC track in which all lines have counts = 10.

For population studies, select the following data sets as target features:

- Series: SNP collection from 1000 genome project
- Sample: Common SNPs
- Sample: All indels

Repeat the analysis with the same parameters as shown in Table 12.

Save the results under the following file names:

stat1_t20e_hg19_PhastCons_foldenrich.out, stat1_t20e_indels_1000GP_foldenrich.out, stat1_t20e_hg19_commonSNPs_1000GP_foldenrich.out

The results are shown in Figure 19 here below.

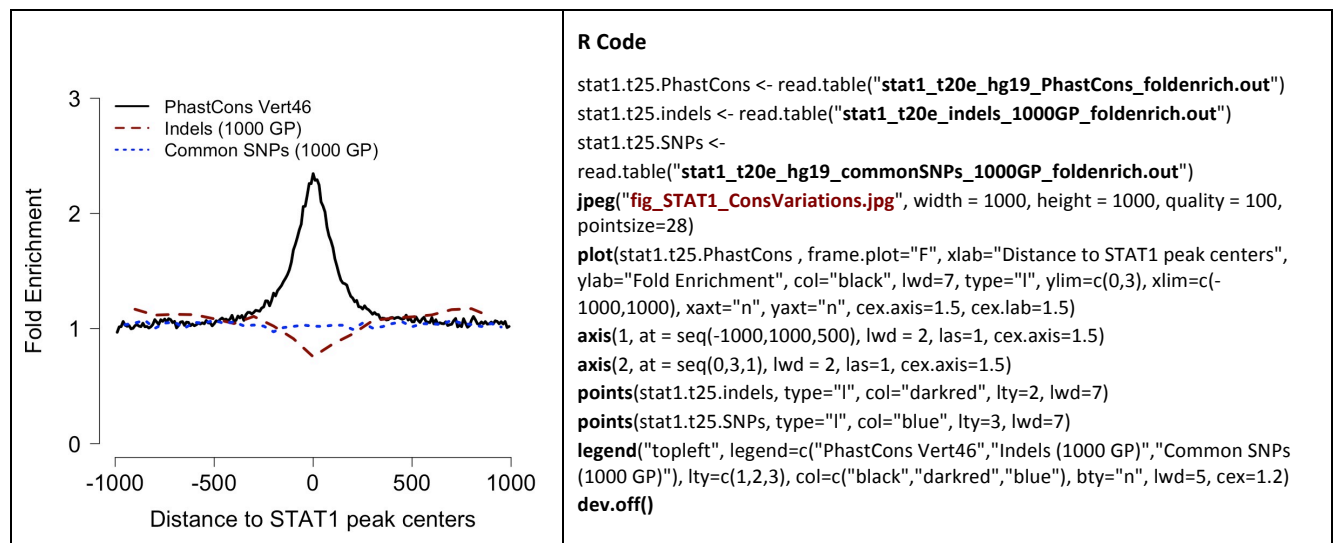


Figure 19. PhastCons score and population variation around STAT1 peaks.

Increased cross-species conservation is observed in a slightly narrower region of about 300 bp. Consistent with this finding, we see depletion of indel variation in the same region. However, contrary to expectation, there appears to be no depletion of common SNPs.

7.1 Histone modification studies with ChIP-Extract

Similarly to ChIP-Cor, the program **ChIP-Extract** correlates the genomic tag count distributions of two features, reference and target feature respectively, and returns the results in a tabular format. **ChIP-Extract** extracts target feature counts in binned genomic regions for each reference feature individually. The output is an integer matrix that can be visualized as a heat map. The output of **ChIP-Extract** can be easily imported into any statistical program, such as the R package, for further analysis.

We use **ChIP-Extract** to study H3K4me3 and H3K27me3 around STAT1 peaks.

Step-by-step procedure

If necessary, repeat the step-by-step procedure described in Table 10 for extracting a STAT1 peak list with relative enrichment threshold 20, otherwise go directly to the **ChIP-Extract** input form (via the action button named '**ChIP-Extract**' on the ChIP-Peak output page) and complete the form as indicated in Table 14 here below.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
<p>Server-resident SGA Files by Filename Filename: <i>hourly/chippeak_*.sga</i> Experiment: Custom SGA Feature: (blank)</p> <p>Additional Input Data Options Strand: any</p> <p>Analysis Parameters Range: -5000 to 5000</p> <p>Histogram Parameters Window width: 50 Count Cut-off: 1</p>	<p>Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: ENCODE ChIP-seq Series: GSE29611, Histone Modifications by ChIP-seq Sample: HeLa-S3 H3K4me3</p> <p>Additional Input Data Options Strand: any Centering: 70</p> <p>Heatmap with ordering : checked</p>

Table 14. Histone modification profiles with ChIP-Cor.

Note that the 'Heatmap with ordering' option is checked. In this case, the order of the rows of the heat map is re-calculated by by computing the correlation with the mean profile. Thus, the loci that mostly contribute to the aggregate profile are ranked on top of the observed heat map. Heat map ordering does not change the output matrix.

Repeat the same procedure for samples 'HeLa-S3 H3K27ac'.

The results are shown in Figure 20. Note that the x-axis labels have been re-edited with R and a text indicating the histone mark has been added to the aggregation plots.

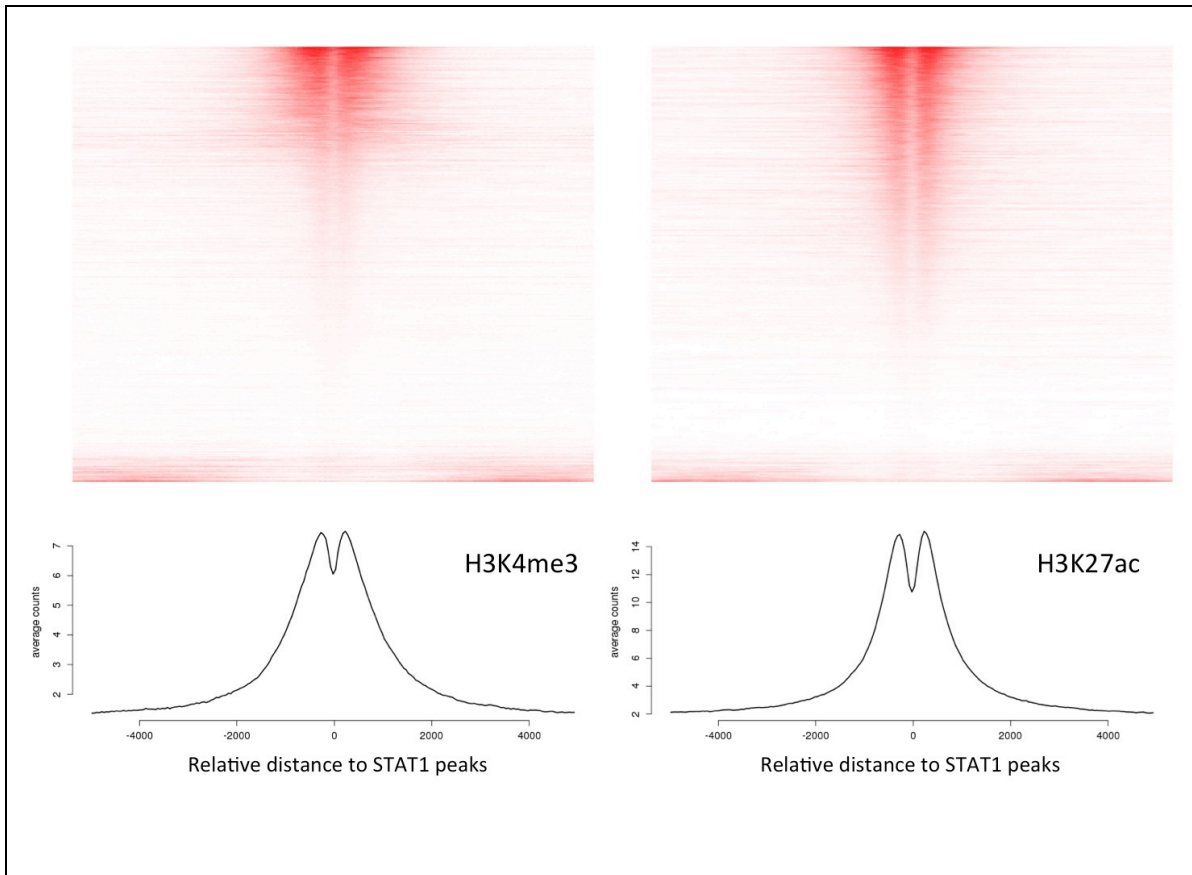


Figure 20. Histone profiles (H3K4me3 and H3K27ac) around STAT1 peaks with ChIP-Extract.

As we can see from both heat maps in Figure 19, a larger number of sites contribute to the H3K27ac profile compared to H3K4me3.

The last section (**Section 8**) of this tutorial is more advanced and it is meant to show how to obtain higher resolution aggregation plots of STAT1 binding sites by using binding site motif matches (or PWM matches) instead of ChIP-seq peak center coordinates to identify with higher accuracy the protein-DNA interaction loci. The idea is to extract the motif that STAT1 recognizes from the *in vivo* occupied binding sites (that is the STAT1 peaks) and to use the motif matches as new anchor points for generating aggregation plots.

In particular, we will investigate sequence conservation at a single base level within *in vivo* occupied STAT1 motif regions as well as interactions between motifs.

8. High resolution aggregation plots for bound PWM matches.

According to the motif occurrence analysis (Figure 10), our peak list has a positional precision of +/- 50 bp as indicated by the average peak width. Aggregation plots of potentially higher resolutions could be obtained by using the *in vivo* occupied binding motif locations as anchor point rather than the peak center positions inferred from ChIP-Seq data. The SSA program **FindM** (<http://ccg.vital-it.ch/ssa/findm.php>) allows for generating such a list of *in vivo* occupied motifs. FindM is a program that retrieves nucleotide sequences around functional genomic sites that may or may not contain given motifs defined by a consensus sequence or a weight matrix. We upload the peak list obtained with relative enrichment factor 20 to the **FindM** web form by rerunning ChIP-Peak once more with the RepeatMasker checkbox activated, and then using the 'FindM' action button on the ChIP-Peak results page. We then select the STAT1 PWM from JASPAR in the menu area reserved for this purpose and search for motifs matches within peak regions from -60 bp to +60 bp relative to the peak center. To generate a random control set, we also collect an approximately equal number of PWM matches from far outside the peak region (+10000 to +12000). Some of the identified STAT1 peaks fall into repetitive elements of the human genome. If these sites are not removed, they introduce a bias in motif-driven down-stream analysis that may lead to wrong conclusions. Repeat-masking allows users to filter out tags falling into repeat regions.

Step-by-step procedure

Fill out the **FindM** input form as shown in Table 15.

SSA Input Data	Signal Description
Sequence input via server-resident FPS Files Name(s): <i>hourly/chippeak_*.fps</i> Sequence Range Entire sequence range: unchecked 5'border: -60 3'border: 60 Sliding window parameters Search mode: bidirectional Sequence selection mode: best matches	PWMs from Library Motif Library: JASPAR CORE 2014 vertebrates Motif: STAT1 MA0137.3 (length=11) Cut-off: p-value Value: 0.0001 Name: MA0137.3 STAT1 Reference Position: 6

Table 15. Extract occupied sites with FindM.

After submission, go to the **ChIP-Cor** program via the action button located in the lower right corner of the **FindM** output page.

To study sequence conservation using PhyloP base-wise conservation scores across STAT1 *in-vivo* occupied binding sites, complete the ChIP-Cor input form as shown in Table 16.

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
Server-resident SGA Files by Filename Filename: <i>hourly/findm_*.sga</i> Experiment: Custom SGA Additional Input Data Options Strand: oriented Analysis Parameters Range: -1000 to 1000 Histogram Parameters Window width: 10 Count Cut-off: 10 Normalization: global	Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: Sequence-derived Series: PhyloP base-wise conservation Sample: PhyloP vertebrate 46way (score ≥ 2) Additional Input Data Options Strand: any Repeat Masker: unchecked

Table 16. PhyloP conservation scores with ChIP-Cor.

Note that we chose the ‘global’ normalization option for the histogram. In such case, we display the target feature abundance as a fold-change relative to the genome average.

Repeat the above procedure (as described in both Tables 15 and 16) for the control list by changing the Sequence Range to 5’border: +10000, 3’border: +12000 (Table 15).

Save the results in TEXT format under the names: **stat1_phyloP_score.out**, and **stat1_phyloP_control.out**

Results are shown in Figure 21 here below.

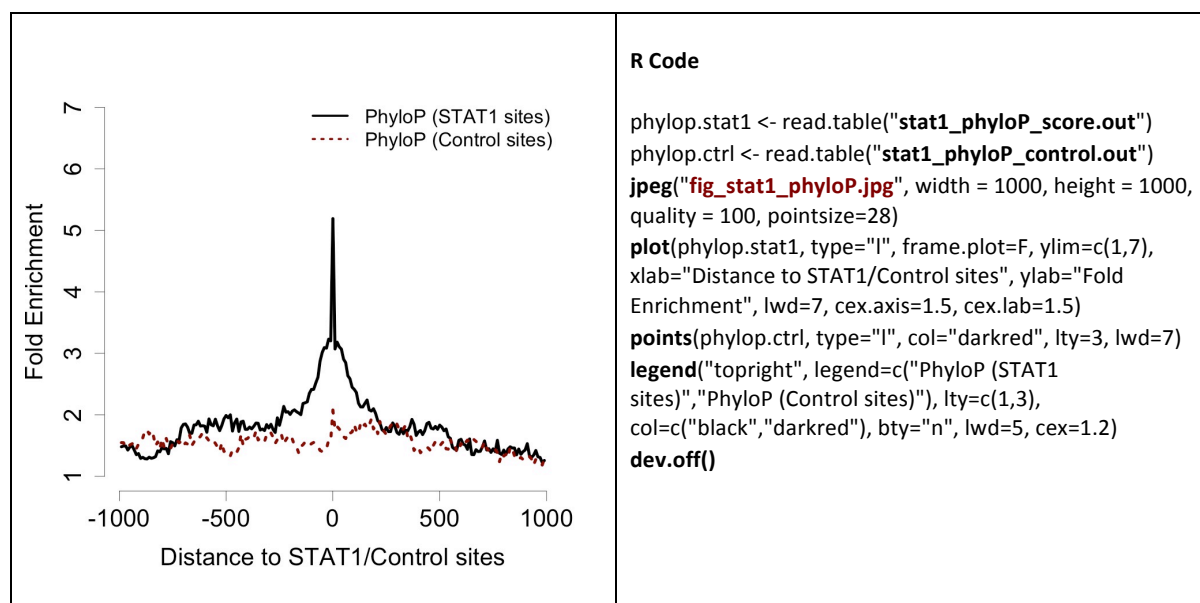


Figure 21. Sequence conservation around *in vivo* occupied STAT1 motifs. Average PhyloP score in 2 kb window around STAT1 motifs evaluated in windows of 10 bp.

Figure 21 shows single base resolution plots for sequence conservation (PhyloP scores). We see that STAT1 sites are surrounded by a region of increased sequence conservation of at least 200 bp. At the center of the plot we notice a spike that indicates even higher conservation at the actual binding motif. The conservation levels around control sites are much lower.

We can zoom in on the binding motif region to better characterize single base conservation within motif regions by repeating the previous analyses with the following parameter changes: Range: -12 to 12, Window width: 1 (Table 17).

ChIP-Seq Input Data Reference Feature	ChIP-Seq Input Data Target Feature
<p>Server-resident SGA Files by Filename Filename: <i>hourly/findm_*.sga</i> Experiment: Custom SGA</p> <p>Additional Input Data Options Strand: oriented</p> <p>Analysis Parameters Range: -12 to 12</p> <p>Histogram Parameters Window width: 1 Count Cut-off: 10 Normalization: global</p>	<p>Select available Data Sets Genome: H. sapiens (Feb 2009 GRCh37/hg19) Data type: Sequence-derived Series: PhyloP base-wise conservation Sample: PhyloP vertebrate 46way (score ≥ 2)</p> <p>Additional Input Data Options Strand: any Repeat Masker: unchecked</p>

Table 17. PhyloP conservation scores with ChIP-Cor.

Repeat the same procedure for the control list, and save the results as:
stat1_phyloP_score_zoom.out, stat1_phyloP_control_zoom.out

Results are shown in Figure 22. The motif sequence logo has been superimposed using GIMP such that the bases in the logo correspond to the positions indicated on the horizontal axis.

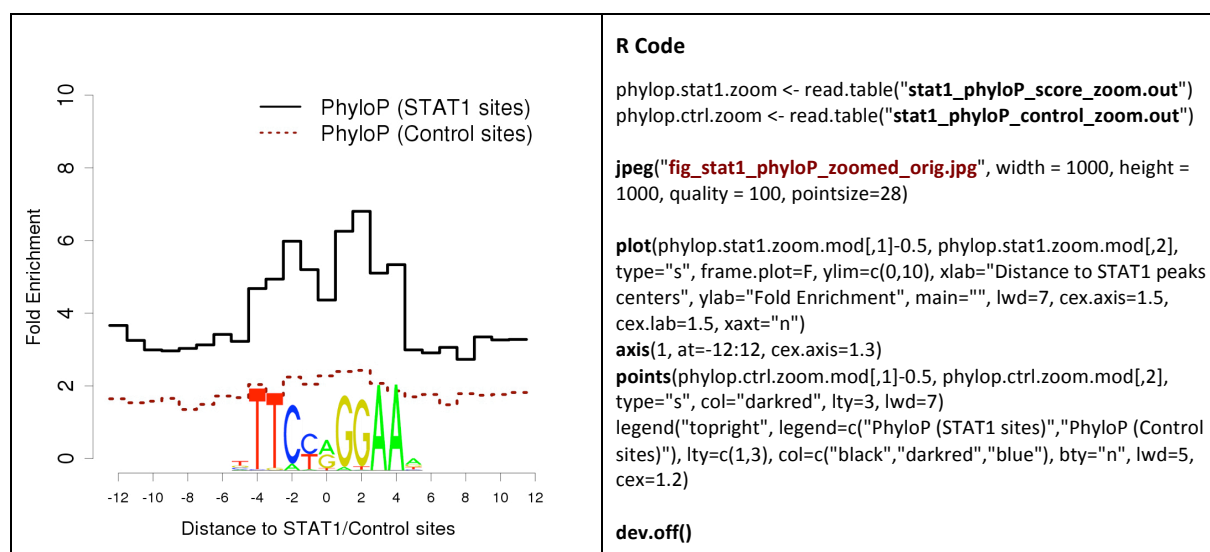


Figure 22. Sequence conservation around *in vivo* occupied STAT1 motifs. Single-base resolution PhyloP profile around STAT1 motif including Sequence Logo.

We see increased sequence conservation within the 9 bp regions that make up the STAT1 binding motif.

As expected, the center position (which is essentially unconstrained according to motif logo) is not more conserved than the flanking regions. The degree of sequence conservation of the control sites is essentially at background levels. In summary, the binding site conservation analysis suggests that *in vivo* bound STAT1 motifs are functionally important whereas unbound motifs are not subject to selective constraints.

Lists of occupied motifs rather than peak centers are also useful to investigate interactions between sequence motifs. This can be done using **OPProf** to generate a motif occurrence profile of STAT1 PWM matches downstream of *in vivo* occupied STAT1 sites.

Step-by-step procedure

Go to **OPProf** via the link on the FindM output page (see procedure described in Table 15).

Complete the **OPProf** input form as shown in Table 18.

SSA Input Data	Signal Description
<p>Sequence input via server-resident FPS Files</p> <p>Name(s): <i>hourly/findm_*.fps</i></p> <p>Sequence Range</p> <p>Entire sequence range: unchecked 5'border: 0 3'border: 100</p> <p>Sliding window parameters</p> <p>Window size: 13 Window shift: 1 Search mode: bidirectional</p>	<p>PWMs from Library</p> <p>Motif Library: JASPAR CORE 2014 vertebrates Motif: STAT1 MA0137.3 (length=11)</p> <p>Cut-off: p-value Value: 0.001</p> <p>Name: MA0137.3 STAT1 Reference Position: 6</p>

Table 18. JASPAR motif occurrence profile using OProf.

Repeat the same analysis for the control set, and save the results in Text format as:

JASPAR_motif_stat1_t20e_rmsk_sites.dat,
JASPAR_motif_stat1_t20e_rmsk_control.dat

Results are shown in Figure 23.

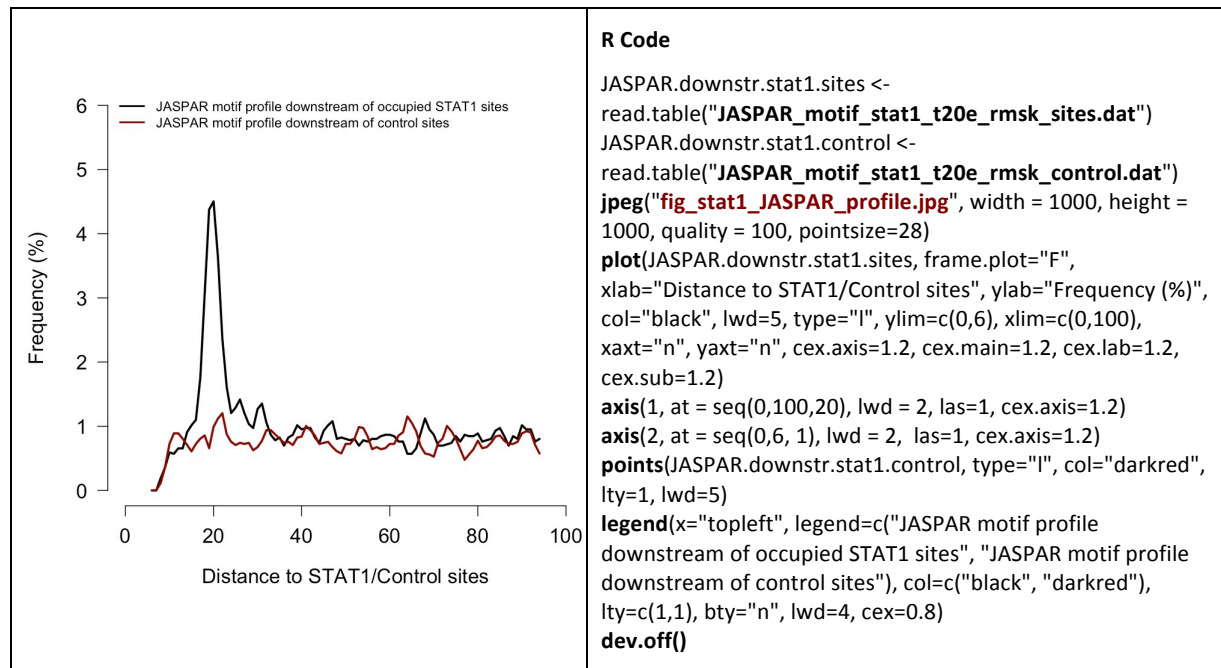


Figure 23. Occurrence profile of JASPAR motif downstream of occupied/non occupied STAT1 sites.

We see a narrow peak (± 2 bp) centered 21 bp downstream of the reference site that is absent in the plot generated with the control set. This previously observed preferential occurrence of two *in vivo* occupied STAT1 sites at a center-to-center distance of two helical turns may be related to a tetrameric binding mode documented for some members of the STAT family ([Schmid and Bucher, 2010](#)).