Principles of ChIP-seq Data Analysis Illustrated with Examples

Giovanna Ambrosini[12], René Dreos[12], and Philipp Bucher[12]

[1] The Swiss Institute for Experimental Cancer Research (ISREC),
Swiss Federal Institute of Technology Lausanne (EPFL), 1015 Lausanne, Switzerland
{giovanna.ambrosini,philipp.bucher}@epfl.ch
[2] Swiss Institute of Bioinformatics, 105 Lausanne, Switzerland
rene.dreos@isb-sib.ch

Abstract. Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) is a powerful method to determine how transcription factors and other chromatin-associated proteins interact with DNA in order to regulate gene transcription. A single ChIP-seq experiment produces large amounts of highly reproducible data. The challenge is to extract knowledge from the data by thoughtful application of appropriate bioinformatics tools. Here we present a concise introduction into ChIP-seq data analysis in the form of a tutorial based on tools developed by our group. We expose biological questions, explain methods and provide guidelines for the interpretation of the results. While this article focuses on ChIP-seq, most of the algorithms and tools we present are applicable to other chromatin profiling assays based on next generation sequencing (NGS) technology as well.

Keywords: ChIP-seq, DNase I hypersensitive sites, transcription factor binding sites, histone marks, bioinformatics analysis

1 Introduction

ChIP-seq is one of several recently introduced high-throughput chromatin profiling assays based on next-generation sequencing (NGS) technology [1]. Others are chromatin accessibility assays using nucleases and DNA methylation profiling based on bisulfite sequencing [2]. The principles for analyzing the data obtained with these techniques are similar, though specialized computer programs have been developed in different application areas. Here we focus on ChIP-seq as a representative chromatin profiling assay while DNAse I hypersensitivity data will be touched briefly.

ChIP-seq is used to locate protein DNA complexes on the genome and works as follows. Protein is first cross-linked to DNA in order to freeze the native chromatin state. Chromatin is then extracted from the cells and cut down to fragments of about 200 bp of DNA, either by sonication or nuclease treatment. Fragments bound by a particular protein are subsequently purified by immunoprecipitation with a specific antibody. After reversing the cross-links, the DNA fragments are extracted and sequenced from the ends with one of the standard NGS platforms. The sequences (25 to 50 bp long) are then mapped to the genome sequence and the coordinates of the matched regions are recorded in a genome feature annotation format such as BED or BAM.

Here we present an introduction into the principles of ChIP-seq data analysis in form of a short tutorial which uses tools from the ChIP-Seq server [5] and the Signal Search Analysis (SSA) server [6], two bioinformatics resources maintained by our group. Both servers provide menu-driven access to large collections of public data sets. In this respect, they represent ideal learning platforms for researchers who would like to make first-hand experiences with ChIP-seq data and familiarize themselves with the corresponding data analysis methods. However, it is not the purpose of this article to provide a comprehensive review of computational methods used in the field, which can be found elsewhere [3] [4].

The reader of this article is invited to carry out the proposed data analysis tasks synchronously on our servers. For each task, we expose the biological motivation, explain the underlying methods, provide step-by-step instructions and provide some guidelines for the interpretation of the results. Nevertheless, due to space limitations we will not be able to explain all the details of the methods. The interested reader is referred to a more comprehensive version of this tutorial posted on the ChIP-Seq server website.

2 ChIP-seq Tutorial

The following tutorial is based on data from an early landmark paper on STAT1 binding sites in γ-interferon stimulated HeLa cells [7]. This data set comprises about 15 million mapped sequence tags and is accessible as a server-resident file from all web input forms of the ChIP-Seq server at:

`http://ccg.vital-it.ch/chipseq/`

Some analysis tasks proposed in this tutorial also rely on programs from the Signal Search Analysis (SSA) server [6] at:

`http://ccg.vital-it.ch/ssa/`

Note further that the figures shown in this paper are not server screenshots. In most cases, they combine results from different program runs and have been generated by downloading the numerical data via links from the server output pages and then re-importing the data into R software.

2.1 5'-3'end Correlation with ChIP-Cor

We start by generating a so-called 5'-3' correlation plot. This analysis serves a dual purpose: quality control and estimation of the average fragment length of the immunoprecipitated fragments. The input data file contains the chromosomal coordinates of the end positions of the mapped sequence reads from the STAT1 ChIP-seq experiment. It is important to know that the reads mapping to the + strand of the genome sequence tend to accumulate upstream of the immunoprecipitated DNA-protein complexes. Likewise, the reads mapping to the -

strand accumulate downstream. A 5'-3' correlation plot reveals the relative shift between + strand (5') tags and - strand (3') tags.

We are going to use the program ChIP-Cor for generating the 5'-3' correlation plot. ChIP-Cor is a very general tool to analyze positional correlations between two genomic features, referred to as 'reference' and 'target' features. It returns a correlation plot showing the average abundance of the target feature at varying distances from the reference feature. The target feature abundance is typically analysed in windows of a certain size and may be expressed as counts per base pair or fold enrichment. The two normalization modes are called 'count density' and 'global' on the ChIP-Cor server page.

For generating a 5'-3' correlation plot, we choose as reference and target features the + strand and - end tags from the same ChIP-seq experiment. If the experiment has worked, we expect a maximum of the - strand tag abundance at a certain distance downstream from position zero, which corresponds to the position of the + strand tags. For the STAT1 data, proceed as follows: Open the ChIP-Cor server input page at:

`http://ccg.vital-it.ch/chipseq/chipcor.php`

and fill out the form as shown in Table 1.

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

Input Data Reference Feature	**Input Data Target Feature**
Select available Data Sets	**Select available Data Sets**
Genome : H.sapiens (March 2006/hg18)	Genome : H.sapiens (March 2006/hg18)
Data type : ChIP-seq	Data type : ChIP-seq
Series : Robertson 2007	Series : Robertson 2007
Sample : Hela S3 STAT1 stim	Sample : Hela S3 STAT1 stim
Additional Input Data Options	**Additional Input Data Options**
Strand : +	Strand : -
Analysis Parameters	
Range : -1000 to 1000	
Histogram Parameters	
Window width : 10	
Count Cut-off : 1	
Normalization : count density	

The resulting plot is shown in Figure 1a. 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. This is an important parameter for 'centering' the data. Centering means shifting the positions of the + strand tags by half the fragment length downstream while shifting the - strand tags by the same distance upstream. Centering increases the power and resolution of the ChIP-seq data analysis as it combines + and - strand tags in an optimal way. In

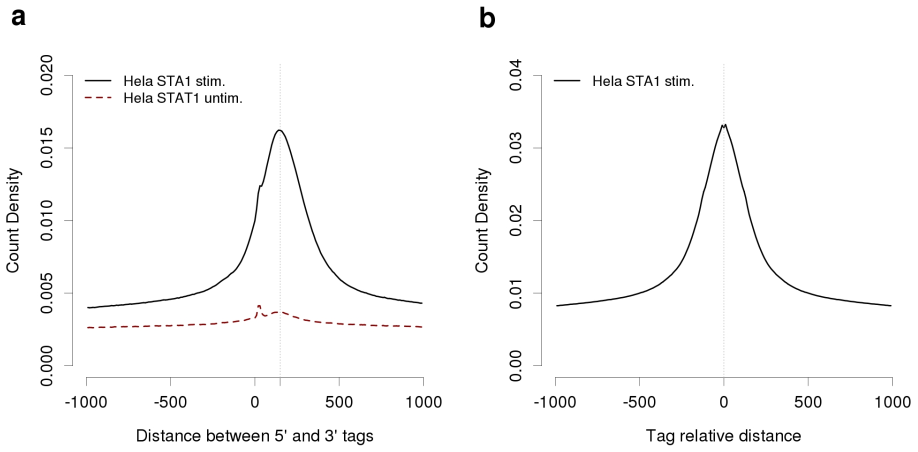


Fig. 1. STAT1 correlation plots. (**a**) 5'-3' end correlation plot for STAT1 ChIP-seq tags *vs* control dataset. (**b**) Autocorrelation plot of 75bp-centered STAT1 ChIP-seq tags.

all subsequent analysis steps, we will center the STAT1 data using a centering distance of 75 bp.

We can generate the same plot for a negative control sample available from the server menu under the name 'Hela S3 STAT1 unstim'. This sample has been generated with chromatin from unstimulated cells where STAT1 is supposed to be localized in the cytoplasm and thus unable to bind DNA. As expected, the corresponding correlation plot is almost flat (Figure 1a), strengthening the conclusion that the peak seen with data from stimulated cells originates from true *in vivo* bound fragments.

Next, we generate a so-called autocorrelation plot for centered STAT1 tags against themselves (same reference and target feature). Use the inputs shown in Table 2 to this end. The result is displayed in Figure 1b.

We see again a Gaussian peak this time with a maximum at position 0. The ChIP-Cor server automatically attempts to fit the correlation plot to a Gaussian curve. If successful, the results are provided via hyperlinks on the output page in graphical (link 'Single Gaussian Fit') and textual form (link 'Parameters'). The text output file (Figure 2) lists the parameters of the fitted function and suggests additional parameters for peak finding.

2.2 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

Table 2. Auto-correlation with ChIP-Cor

Input Data Reference Feature	**Input Data Target Feature**
Select available Data Sets	**Select available Data Sets**
Genome : H.sapiens (March 2006/hg18)	Genome : H.sapiens (March 2006/hg18)
Data type : ChIP-seq	Data type : ChIP-seq
Series : Robertson 2007	Series : Robertson 2007
Sample : Hela S3 STAT1 stim	Sample : Hela S3 STAT1 stim
Additional Input Data Options	**Additional Input Data Options**
Strand : any	Strand : any
Centering : 75	Centering : 75
Analysis Parameters	
Range : -1000 to 1000	
Histogram Parameters	
Window width : 10	
Count Cut-off : 1	
Normalization : count density	

to other programs which report starting and ending position 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 2) suggests to use a window of 286 bp and a peak threshold value of 12 tags (Figure 2). We round the window size to 300.

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

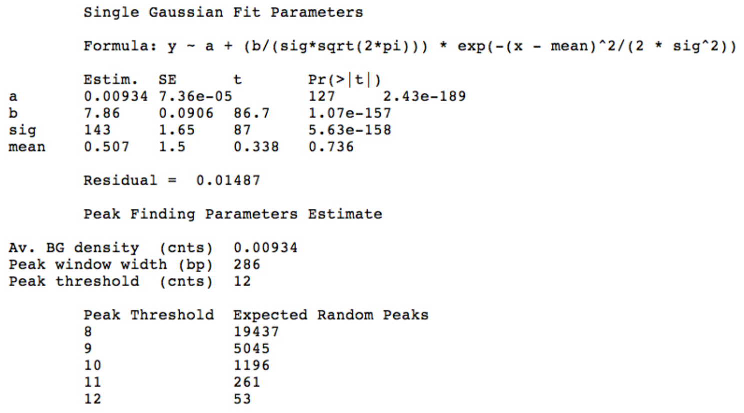
`http://ccg.vital-it.ch/chipseq/chip_peak.php`

and fill it out as shown in Table 3.

Table 3. Peak finding with ChIP-Peak

ChIP-Seq Input Data	**Peak Detection Parameters**
Select available Data Sets	Window Width (bp): 300
Genome : H.sapiens (March 2006/hg18)	Vicinity Range (bp) : 300
Data type : ChIP-seq	Peak Threshold : 100
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 : unchecked (blank)
Centering : 75	Chromosome Region : unchecked
Repeat Masker : checked	

Running then ChIP-Peak with the recommended tag threshold returns 54'473 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



```
         Single Gaussian Fit Parameters

         Formula: y ~ a + (b/(sig*sqrt(2*pi))) * exp(-(x - mean)^2/(2 * sig^2))

         Estim.  SE      t       Pr(>|t|)
a        0.00934 7.36e-05        127     2.43e-189
b        7.86    0.0906  86.7    1.07e-157
sig      143     1.65    87      5.63e-158
mean     0.507   1.5     0.338   0.736

         Residual =  0.01487

         Peak Finding Parameters Estimate

Av. BG density  (cnts)  0.00934
Peak window width (bp)  286
Peak threshold  (cnts)  12

         Peak Threshold  Expected Random Peaks
         8               19437
         9               5045
         10              1196
         11              261
         12              53
```

Fig. 2. Autocorrelation plot: Gaussian fit and parameters for peak finding

BED is a general format understood by many other web-based bioinformatics resources potentially useful for follow-up analysis (*e.g.* gene enrichment analysis). It is therefore recommended to save the peak lists in all three formats. Note further that the output page contains an action button allowing for remapping of the chromosomal coordinates to other genome assemblies. We use this button to remap all peaks from the human genome assembly hg18 to the newer assembly hg19, in order to be able to jointly analyze our peaks with more recent server-resident data. For the following parts of this tutorial, we save the hg19 version of the peak list in FPS format under the name:

stat1_t12_hg19.fps

We have to be aware that 12 is a minimal threshold maximizing sensitivity. For many types of downstream analysis more stringently selected peak lists are preferable. We therefore repeat ChIP-Peak with higher thresholds of 25, 50 and 100 tags and obtain 16'337, 4'445 and 1'522 peaks, respectively. Remap these peak lists to hg19 and save them under the following names.

stat1_t25_hg19.fps, stat1_t50_hg19.fps, stat1_t100_hg19.fps
stat1_t25_hg19.sga

(The 3-letter extension of the file names reflects the format.)

Some of the identified STAT1 peaks fall into repetitive elements of the human genome. These peaks may cause problems for certain types of downstream analysis, for instance cross-species sequence conservation analysis. The input forms of the ChIP-Seq server allow users to filter out tags falling into annotated repeat regions. We will need a repeat-masked peak list later in this tutorial. We therefore rerun ChIP-Peak with tag threshold 25 and the RepeatMasker checkbox activated. Then save the remapped FPS file under the name:

stat1_t25_rmsk_hg19.fps

2.3 Motif Enrichment in Peak Regions

STAT1 is known to bind to a DNA motif resembling 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 of ChIP-seq peak finders [8].

The OProf program of the SSA server can be used for motif enrichment analysis. It returns a graph showing the percentage of sequences containing a motif in a sliding window around a set of reference position. To generate a motif enrichment plot, go to:

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

and fill out the input form as shown in Table 4.

Table 4. Motif Studies with OProf

SSA Input Data	**Signal Description**
Upload FPS file	**Consensus seq**
stat1_t12_hg19.fps	TTCNNNGAA
FPS name : ChIP-Peak	Mismatches : 0
FPS type: STAT1 peaks	
Sequence Range	Name : TTCNNNGAA
Entire sequence range: unchecked	Reference Position : 5
5'border: -499 3'border: 500	
Sliding window parameters	
Window size: 100 Window shift: 5	
Search mode: bidirectional	

The fields 'FS name', 'FPS type' and 'Name' (on the right side) only define text elements displayed on the output page. They do not influence the analysis in any other ways. In general, it is recommended to use the search mode 'bidirectional' because ChIP-seq peaks have no defined +/- strand polarity. As a consequence, the corresponding binding motifs may occur in either orientation. However, this is not really relevant in our case since the STAT1 consensus sequence itself is symmetrical.

We repeat the same type of analysis with the peak lists obtained with tag thresholds 25, 50 and 100. The combined results are shown in Figure 3a. With all peak lists, we see a clear enrichment of STAT1 motifs near position zero (the reported peak center). As expected, we see higher peaks with higher peak thresholds.

The OProf server provides menu-driven access to a large number of published peak lists from ChIP-seq experiments, including two STAT1 peak lists from the ENCODE consortium, one from HeLa and one from K562 cell lines [9]. We can carry out the same analysis as for these peak lists by choosing the following options from the data selection menu:

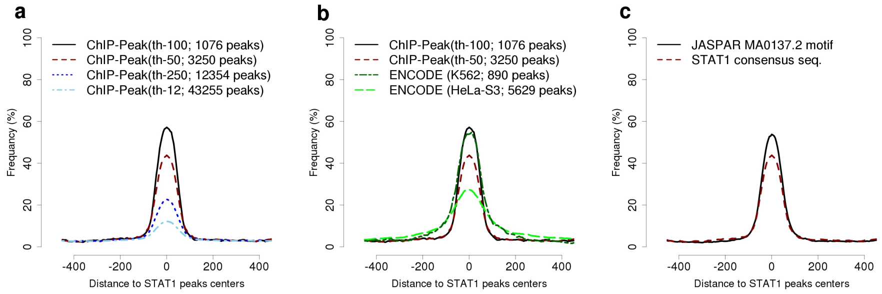


Fig. 3. Peak list evaluation by motif enrichment. (**a**) STAT1 consensus sequence (TTCNNNGAA) enrichment in peak lists obtained at various tag thresholds. (**b**) Peak lists derived in this tutorial versus Peak lists from ENCODE (TTCNNNGAA). (**c**) TTCNNNGAA versus JASPAR weight matrix for tag threshold 50.

Genome : H. sapiens /Feb 2009/hg19)
Data type : ENCODE ChIP-seq-peaks
Series : Wang et al. 2012, Transcription Factor Binding Sites from ENCODE
Sample : Hela-S3 STAT1 std - IFNg30 - peaks

Figure 3b shows consensus sequence enrichment profiles for these peaks lists and the ones generated by ChIP-Peak. Unexpectedly, the peak lists generated from the earlier data by Robertson and coworkers compare favorably to the newer peak lists from ENCODE, 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 more accurately. The OProf server provides menu-driven access to PWMs from several public resources, including a STAT1 matrix from the JASPAR database [10]. To search the JASPAR MA0137.2 STAT1 motif, fill out the OProf input form following the instructions given in Table 5.

Figure 3c shows motif enrichment profiles for the STAT1 consensus sequence, and the JASPAR matrix. Unsurprisingly, the PWM-defined motif shows a higher peak at approximately equal background frequency. Note that for this tutorial, the cut-off P-value in Table 5 was chosen such as to match the background frequency of the consensus sequence motif. This is a necessary condition for fair comparison of the motif enrichment values.

2.4 Exploring the Genomic Context of STAT1 Peaks

ChIP-Cor enables the user to generate aggregation plots 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. Specifically, we are interested in the abundance

Table 5. PWM profile with OProf

SSA Input Data	**Signal Description**
Upload FPS file	**PWMs from Library**
stat1_t50_hg19.fps	Motif Library : MEME-derived JASPAR CORE 2009
FPS name : ChIP-Peak	
FPS type: STAT1 peaks	Motif : MA0137.2 STAT1(length=15)
Sequence Range	
Entire sequence range: unchecked	Cut-off : p-value
5'border: -499 3'border: 500	Value: 0.00011
Sliding window parameters	
Window size: 100 Window shift: 5	MA0137.2 STAT1
Search mode: bidirectional	Reference Position : 8

of an active promoter mark (H3K4me3), an active enhancer mark (H3K27ac) and a repressive chromatin mark (H3K27me3) in the vicinity of STAT1 peaks. Remember in this context that the STAT1 ChIP-seq experiment was carried out in HeLa cells that were stimulated with γ-interferon. On the other hand, the histone modification maps from ENCODE were obtained from unstimulated cells where STAT1 is not supposed to bind to DNA. This analysis thus addresses the questions whether target sites of STAT1 are associated with certain histone marks even at times when they are not bound by STAT1.

To carry out this analysis for H3K4me3, fill out the ChIP-Cor web form as is detailed in Table 6. Rerun ChIP-Cor with the corresponding samples for H3K27ac and H3K27me3.

Table 6. Histone modification profiles with ChIP-Cor

Input Data Reference Feature	**Input Data Target Feature**
Upload Custom Data	**Select available Data Sets**
Format : SGA	H.sapiens (Feb 2009/hg19)
File : *stat1_t25_hg19.sga*	Data type : ENCODE ChIP-seq
Genomes : H. sapiens (Feb 2009/hg19)	Series : GSE29611, Histone Modifications by ChIP-seq
Additional Input Data Options	
Strand : any	Sample : Hela-S3 H3K4me3
Repeat Masker : unchecked	**Additional Input Data Options**
Analysis Parameters	Strand : any
Range : -5000 to 5000	Centering : 70
Histogram Parameters	
Window width : 10	
Count Cut-off : 1	
Normalization : global	

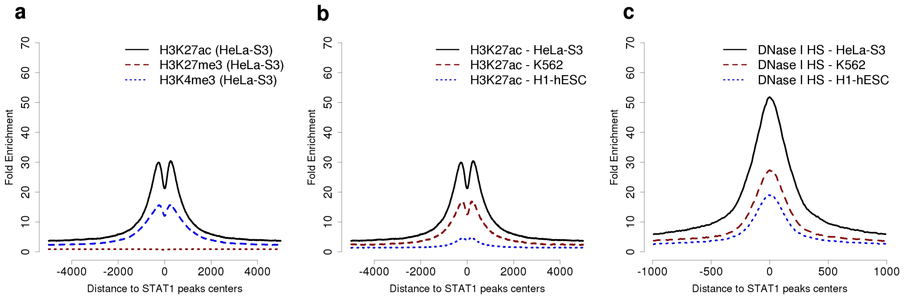


Fig. 4. Histone marks and DNAse I hypersensitivity around STAT1 peaks (**a**) Histone marks around HeLa STAT1 peaks in non-stimulated Hela cells. (**b**) H3K27ac marks in HeLa and other cell types. (**c**) DNAse I hypersensitivity around STAT1 peaks.

Figure 4a shows the histone modification profiles around STAT1 peaks for the three histone marks. We see that STAT1 peaks fall into regions of about 500 base-pairs which are 15-fold enriched in H3K27ac. A 7-fold enrichment is observed with the promoter mark H3K4me3 and no enrichment is seen for H3K27me3. This result suggests that STAT1 primarily binds chromatin domains that are already in an active state before γ-interferon induction. Moreover, STAT1 appears to prefer enhancers over promoters.

We may wonder whether these STAT1 bound enhancers are also active in other cell types. We thus decide to compare H3K27ac marks in HeLa cells along with two other cell types: embryonic stem cells and the cancer-derived K562 cell line. Repeat the step-by-step procedure in Table 6 with the following samples:

'H1-hESC H3K27ac', 'K562 H3K27me3'

The aggregation plots for these cell lines are shown in Figure 4b together with the results obtained for HeLa cells. We see an approximately two-fold higher enrichment in HeLa compared to the other cell types, suggesting a substantial degree of tissue-specificity of STAT1-bound regulatory regions.

Next, we explore DNase I hypersensitivity near STAT1 sites in the same three cell types. To this end, choose the following samples as target features:

Genome : H. sapiens /Feb 2009/hg19)
Data type : ENCODE DNAse FAIRE etc.
Series : Thurman 2012, DNaseI Hypersensitivity by Digital DNaseI ...
Sample : DNaseI HS - Hela-S3 - None - Rep1

and repeat the analysis with the same parameters as shown in Table 6 except:

Range : -1000 to 1000
Centering: (leave blank)

The results are shown in Figure 5c. In summary, STAT1 peaks occur preferentially within DNase hypersensitive regions of about 200 bp.

2.5 High Resolution Aggregation Plots for Bound PWM Matches

According to the motif enrichment analysis (Figure 3), our peak list has a positional precision of +/- 50 bp. Aggregation plots of potentially higher resolution could be obtained by using as anchor points the actual binding sites (defined by the binding motif) rather than the peak centers. The SSA program FindM can be used to compile a list of motifs located in the vicinity of peak center positions. To do so, go to the web form at:

`http://ccg.vital-it.ch/ssa/findm.php`

and fill in the parameters as shown in Table 7. Note that we are using the STAT1 PWM from the JASPAR database as motif definition and that we search corresponding sites between -60 bp and +60 bp relative to the peak center position. To generate a random control set, we also collect an approximately equal number of PWM matches from a region far away from the peak centers. To this end, repeat the search with the following parameter changes:

5'border: 10000 3'border: 12000

Then save the two output files under the following names:

stat1_t25_rmsk_hg19_sites.fps, *stat1_t25_rmsk_hg19_control.fps*

We emphasize that the control set is a random rather than a negative control as we cannot be sure that some of the identified sites lie within peaks. However, since motif matches outnumber peaks by about two orders of magnitude, we can assume that most of these sites are not occupied *in vivo*.

Table 7. Extract occupied motifs with FindM

SSA Input Data	**Signal Description**
Upload FPS file	**PWMs from Library**
stat1_t25_rmsk_hg19.fps	Motif Library : MEME-derived JASPAR CORE 2009
FPS name : Custom FPS	
FPS type: Unknown	Motif : MA0137.2 STAT1(length=15)
Sequence Range	
Entire sequence range: unchecked	Cut-off : p-value
5'border: -60 3'border: 60	Value: 0.0001
Sequence Selection and Search Criteria	
Search mode: forward	MA0137.2 STAT1
Sequence Selection mode: best matches	Reference Position : 8

We are now going to look at cross-species conservation of the *in vivo* bound STAT1 sites and the flanking regions using the PhyloP base-wise conservation scores from UCSC [11], which are also installed at the back-end of the ChIP-Seq server. To this end, fill out the ChIP-Cor input form as shown in Table 8 and subsequently repeat the same procedure for the control set.

Table 8. PhyloP conservation scores with ChIP-Cor

Input Data Reference Feature	**Input Data Target Feature**
Upload Custom Data	**Select available Data Sets**
Format : FPS	H.sapiens (Feb 2009/hg19)
File : *stat1_t25_rmsk_hg19_sites.fps*	Data type : Sequence-derived
Genomes : H. sapiens (Feb 2009/hg19)	Series : phyloP base-wise conservation
Additional Input Data Options	Sample : PhyloP vertebrate 46way (score $\geq$ 2)
Strand : any	
Repeat Masker : unchecked	**Additional Input Data Options**
Analysis Parameters	Strand : any
Range : -1000 to 1000	Repeat Masker : unchecked
Histogram Parameters	
Window width : 10	
Count Cut-off : 10	
Normalization : global	

The results are shown in Figure 5a. 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 which indicates even higher conservation at the actual binding motif. The conservation levels around control sites is much lower.

We can zoom in on the binding motif region by repeating the previous analyses with the following parameter chages:

Range : -12 to 12, Window width : 1

Using these settings, we see increased sequence conservation within the 9 bp region that makes up the STAT1 binding motif (Figure 5b). Note that the sequence logo for the STAT1 matrix has been inserted into the Figure such that the bases in the logo correspond to the positions indicated on the horizontal axis. 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 subect to selective constraints.

References

1. Furey, T.S.: ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet., 13(12), 840-852 (2012)
2. Ku, C.S., Naidoo, N., Wu, M., Soong, R.: Studying the epigenome using next generation sequencing. J Med Genet., 48(11), 721-730 (2011)
3. Rougemont, J., Naef, F.: Computational analysis of protein-DNA interactions from ChIP-seq data Methods Mol Biol., 786, 263-273. (2012)

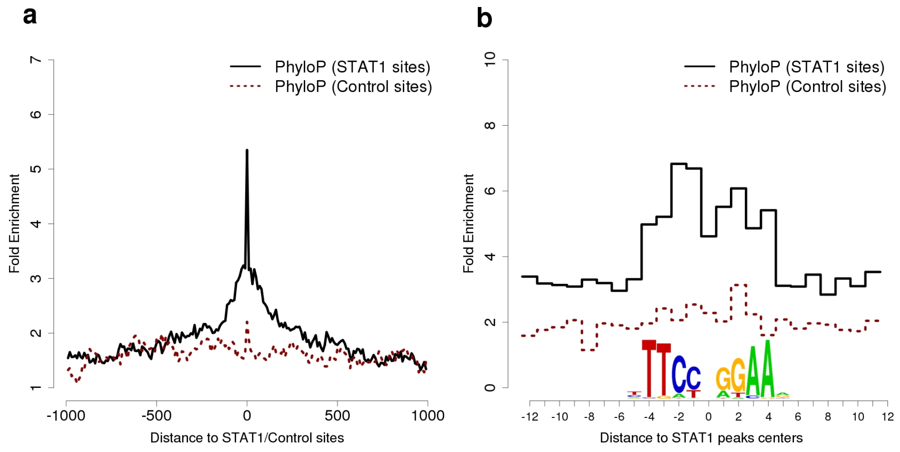


Fig. 5. Sequence conservation around *in vivo* occupied STAT1 motifs. (**a**) Average PhyloP score in 2 kb window around STAT1 motifs evaluated in windows of 10 bp (**b**) Single-base resolution PhyloP profile around STAT1 motif including Sequence Logo.

4. Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., Zhang, J.: Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol., 9(11), e1003326 (2013).
5. The ChIP-Seq web server, `http://ccg.vital-it.ch/chipseq`
6. Ambrosini, G., Praz, V., Jagannathan, V., Bucher, P.: Signal search analysis server. NAR, 31(13), 3618-3628 (2003)
7. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O.L., He, A.: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods, 4(8), 651-7 (2007)
8. Wilbanks E.G., Facciotti M.T.: Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One, 5(7), e11471 (2010)
9. Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., Rando, O.J., Birney, E., Myers, R.M., Noble, W.S., Snyder, M., Weng, Z.: Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res, 22(9), 1798-812 (2012)
10. Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., Wasserman, W.W.: JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res., 42, D142-D147 (2014)
11. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A.: Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res., 20(1), 110-121 (2010)