The ChIP-Seq web server, an online resource for analyzing ChIP-Seq and other types of mass genome annotation data

Giovanna Ambrosini[1,2], Christoph D. Schmid[1,2,*], Philipp Bucher[1,2,†]

[1] Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, 1015 Lausanne
* Current address: Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel.

ABSTRACT

The ChIP-Seq web server provides access to a set of useful tools performing common ChIP-Seq data analysis tasks, including positional correlation analysis, peak detection, and genome partitioning into signal-rich and signal-poor regions. It is an open system designed to allow interoperability with other resources, in particular the motif discovery programs from the Signal Search Analysis (SSA) server. Even though initially developed for ChIP-Seq data, the computational methods are generic enough to be applicable to other mass genome annotation data including CAGE, or even non-experimental annotations such as cross-genome conservation scores. Since the Chip-Seq server uses speed-optimized algorithms and programs, response times are short. Input data can be uploaded in various formats, including BED and GFF. The server also provides access to a large collection of server resident data from landmark papers for comparison and demonstration purposes. Results are provided in graphical form as well as in formats that can be used as input to another web form. Complex analysis tasks can thus be carried out by running several services in a pipeline. Documentation is provided by a comprehensive tutorial featuring many biologically interesting examples. The source code of the back-end programs is publicly available at http://sourceforge.net/projects/chip-seq. The ChIP-Seq web server is freely accessible at http://ccg.vital-it.ch/chipseq.

.

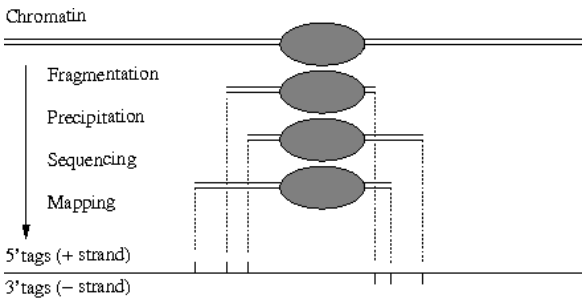
INTRODUCTION

The ChIP-Seq technique, recently reviewed in (1), allows for genome-wide mapping of *in vivo* DNA-protein interactions at high resolution and low cost. It constitutes a breakthrough in the study of gene regulation for the following reason. Before the advent of chromatin-immunoprecipitation (ChIP), *in vivo* protein-DNA interactions were essentially non-observable events. With ChIP-Seq (a combination of ChIP with massively parallel sequencing) these interactions have at once become visible on a genome-wide scale. As a consequence, ChIP-Seq has become a widely used standard technique in less than five years.

A ChIP-Seq experiment produces huge amounts of reproducible and potentially informative data. It is broadly recognized that data analysis is the major bottleneck in the application of this technique. In order to explain the computational methods involved in ChIP-Seq data analysis, it is useful to recapitulate how the technique works (Figure 1a). Chromatin, which consists of DNA associated with proteins, is first purified and cross-linked. The cross-linking ensures that a protein bound at a specific location on a chromosome will stay at the same place during subsequent manipulations. After cross-linking, the chromatin is cut down to short fragments, usually by sonication or nuclease digestion. The DNA fragments attached to a particular protein are then purified with the aid of a specific antibody. After that, the DNA is released from the protein by reversal of the cross-linking reaction. Subsequently, the ends of the purified DNA fragments are sequenced by a massively parallel sequencing technology.

The procedure described above typically results in millions, or tens of millions of short sequence reads of 25 to 75 base pairs. In order to localize the corresponding protein-binding sites, these reads need to be mapped to the genome. The result is a new data structure representing the distribution of mapped sequence tags along the genome. For concise representation, only the base positions corresponding to the 5'ends of the sequence reads need to be recorded. If the experiment has worked properly, we will observe a cluster of tags matching the +strand of the chromosome upstream of the protein binding region, and a second cluster of tags matching the −strand downstream of this region (Figure 1a). The characteristic shift between the peaks corresponds to the average length of the pulled-down fragments and is an important parameter for downstream analysis methods.

A



B

Sequence ID	Feature	position	strand	counts
NC_000001.9	STAT1	559786	+	3
NC_000001.9	STAT1	559787	+	1
NC_000001.9	STAT1	559789	+	2
NC_000001.9	STAT1	559794	-	3
NC_000001.9	STAT1	559801	-	1
NC_000001.9	STAT1	559801	-	1

Figure 1: ChIP-Seq technique and data representation: (**A)** Schematic representation of the process leading to the input data for the ChIP-Seq server. Multiple fragments bound by the same protein and originating from the same chromosomal location are sequenced from one or both ends. The short sequence reads are mapped to the genome. This leads to a kind of tag count histograms for both strands of the chromosomes. Note that tags mapped to the +and −strand will accumulate upstream and downstream of the protein-bound region, respectively. (**B**) Representation of mapped sequence tags in SGA format, the internal working format of the ChIP-Seq server. SGA is tab-delimited text file format with 5 obligatory fields, sorted by sequence ID, position and strand. The feature field refers to a particular ChIP-Seq experiment. Note that chromosomes are identified by unique sequence IDs from NCBI's RefSeq database. This prevents mix-ups between different species and genome assemblies.

The reaction of the bioinformatics community to ChIP-Seq was quick. Many software resources have been developed over the last few years, reviewed for instance recently in (2). The computational analysis of ChIP-Seq data concerns three levels:

1. Quality filtering and tag mapping
2. Identification of signal enriched regions (peak finding)
3. Follow-up analysis: motif discovery, mapping peaks to genes, etc.

Level 1 and 2 involve standard tasks that are usually accomplished by stand-alone programs with a command line interface, or by R packages such as Rolexa (3). Level 3 analysis, which is downstream of peak detection, is much more diverse. Some of the more common tasks can be carried with general purpose genome analysis platforms such as Galaxy (4). Note in this context that most peak-finders return results in formats that can be uploaded to the UCSC genome browser or Galaxy. In addition to general purpose platforms, a number of specialized software solutions for ChIP-seq data have been developed. These programs typically offer a graphical user interface and combine peak finding with third level functions. Two prominent examples are CisGenome (5) and seqMiner (6). As these software platforms are not web-based, they need to be installed on a local computer.

The high volume of data that need to be transferred and processed in ChIP-Seq data analysis makes it difficult to develop a web-based solution with acceptable response time. This may explain why such web servers are rare. We know of only two resources described in journal articles, W-ChIPeaks (7) and Sole-Search (8). In this paper, we present a third resource, the ChIP-Seq web server, which offers a variety of second and third level analysis services at acceptable response times.

OVERVIEW AND DESIGN PRINCIPLES

The ChIP-Seq server offers three main programs: ChIP-Peak, ChIP-Part, and ChIP-Cor. The first two serve to detect signal-enriched regions. The main difference between ChIP-Peak and ChIP-Part lies in the output formats. ChIP-Peak returns peak center positions and is typically used for detecting transcription factor binding sites. ChIP-Part returns a list of signal-enriched regions defined by start and end positions. It is more useful for analyzing the genomic distribution of epigenetic marks such as histone modification marks, especially those that spread over large regions (*e.g.* H3K36me3). The most versatile tool is ChIP-Cor, which produces a positional correlation diagram for two genomic features. Input features may be ChIP-Seq tag positions, peaks found by ChIP-Peak, or any type of genome annotation that can be mapped to a single base on a chromosome. The power of this method will be illustrated by several examples in the next Section.

The ChIP-Seq server is a front-end to the ChIP-Seq tools, a collection of C programs and Perl scripts which can be downloaded from SourceForge.net. The C programs are primarily optimized for speed. For this reason, they use their own compact format for ChIP-Seq data representation called SGA (Figure 1b). The SGA (Simple Genome Annotation) format differs in one very important aspect from similar formats such as BED or GFF. It is required to be sorted by genome positions. The fact that ChIP-Seq tools programs operate on sorted input files enables them to generate results in one sweep through the genome.

A potentially vulnerable component of ChIP-Seq data analysis resources is data input. We support data upload in SGA, FPS, BED and GFF formats. FPS is the specific format used by the Signal Search Analysis server (9), a motif discovery platform developed by our group. Users have to pay attention that chromosomes are uniquely identified. Both the ChIP-Seq and Signal Search Analysis servers use RefSeq IDs including version numbers for internal representation (Figure 1b). The ChIP-Seq server also accepts chromosome names as spelled by the UCSC genome browser on input. If chromosome names are used, the corresponding species and gene assembly should be selected from a menu button on the web form. The basic algorithms will still work with non-standard sequence identifiers, but many accessory functions will not be available, including sequence extraction and links for viewing the results in the UCSC genome browser.

The software at the back-end of the ChIP-Seq server has a modular design. Each program performs an elementary data processing step in an efficient manner. The same format, SGA, is used for input and output,

making it possible to carry out complex analysis tasks by running several programs in a pipeline. The web server mirrors this modular design to a large extent. Generally, there is a one-to-one relationship between web forms and back-end programs. The output of one service can be used as input to another service. This interoperability extends to programs of the SSA server, as both resources share common input and output formats. By combining different programs in an innovative fashion, creative users can run novel types of analyses not even anticipated by the developers.

The ChIP-Seq server offers a rich collection of server-resident public data sets from landmark papers, which have been widely used by the bioinformatics community for benchmarking and testing new algorithms. New users can familiarize themselves with our methods by running the same program with different parameter settings on server-resident public data. The server also features a collection of so-called genome annotation files, including transcription start site collections, which are primarily used as input to the ChIP-Cor tool. Documentation is provided via a tutorial illustrating the capabilities of the methods with many biologically interesting examples. In addition, we offer a technical document, describing data formats and analysis methods in more detail. The back-end programs are documented by UNIX-style man pages included in the source code distribution.

EXAMPLES

In this section we are going to illustrate the capabilities of the ChIP-Seq server with typical application examples. To this end, we will use a well known data set from an experiment aimed at mapping STAT1 binding sites in γ-interferon stimulated HeLa cells (10). Note that this experiment produced about 15 million tags that could be mapped to the human genome sequence. This data set can be accessed as a server-resident file. Precise instructions on how to carry out the following analyses can be found in the tutorial posted on the ChIP-Seq server home page.

5'-3' end correlation

This type of analysis is usually carried out first. It serves as a quality control step and helps to choose optimal parameters for subsequent peak detection. In particular, it reveals the characteristic shift between tags mapping to the + and −strand of the reference genome. The results shown here were obtained with the program ChIP-Cor. The output is a tag correlation diagram (Figure 2a) showing the abundance of the "target" feature (STAT1 3' tags) as a function of the distance from the "reference" feature (STAT1 5'tags). The server offers several options for scaling the abundance of the target feature. Here we have chosen "count density", which is defined as counts per base pair. The correlation diagram shows a peak centered at about position +140. The absence of such a peak would indicate that the experiment has not worked. We further note that the background frequency of −strand tags is about 0.005 tags per bp.

Peak detection

The ChIP-Peak tool is used for detecting peaks in ChIP-Seq data targeted at transcription factors. It implements a very simple method which works as follows. The number of tags is counted in a sliding window of fixed width. Speed is gained by considering only those windows which have at least one tag at the center position. At the end, all windows which have tag numbers greater or equal to a threshold value, and in addition are locally maximal within a so-called "vicinity range", are reported as peaks. Note that only the peak center positions are included in the output. In spite of its simplicity, ChIP-Peak appears to perform equally well or better that other peak finders based on more advanced statistical models (11). The ChIP-Peak web service offers some data-preprocessing options, notably tag centering, which has the effect that tags mapped to the +strand are shifted by a user-specified distance downstream the chromosome while tags mapped to the −strand are shifted by the same distance in the upstream direction. Based on the results from the 5'-3' correlation analysis, we choose 70 as the centering distance and 200 as the window width. The choice of an optimal peak threshold is more a matter of taste, as significant peaks vary in tag coverage over a large dynamic range. Peaks with low tag coverage may be significant from a statistical viewpoint but functionally irrelevant. In any case the threshold should be significantly higher than the expected number of + and −strand tags in a window of the specified width. Based on Figure 2a, this number is in the order of two

tags for a window of 200 bp. With a threshold of 15 tag per windows, ChIP-Peak finds 23720 peaks. ChIP-Peak returns results in various formats. Figure 2b shows a portion of the SGA-formatted output file, which can be used as input to other ChIP-Seq tools, for instance feature correlation analysis. The same information is also provided in the largely equivalent FPS format, which is used by the Signal Search Analysis Server (SSA). The output is further provided as a BED file, which can be uploaded to the UCSC genome browser via a direct link from the ChIP-Peak results page. Figure 2c shows the peak highlighted in Figure 2b in a UCSC browser window. Note that this peak co-localizes with a known STAT1 binding site recorded in the ORegAnno database (12).

Motif enrichment in peak regions

STAT1 is known to bind to the consensus sequence TTCNNNGAA. If the peaks found by ChIP-Peak indeed correspond to real binding sites, one would expect to find this motif in the vicinity of the returned peak center positions. In fact, this type of motif enrichment test is commonly used for benchmarking the performance of peak finders (13). As we will search for exact matches to the motif, we use ChIP-Peak with a higher threshold of 30 counts. The output is saved in FPS format. The program OProf (Occurrence Profile) from the SSA server (http://ccg.vital-it.ch/ssa/) is used to determine the frequency of STAT1 motifs at relative distances from the STAT1 peak center in a sliding window of 100 bp (Figure 2d).

Motif discovery in peak regions

Consensus sequences can only approximately describe the binding specificity of a transcription factor. Position weight matrices are potentially far more accurate in predicting binding sites. Moreover, the DNA-binding specificity of a protein under investigation may not be known in advance. These are some of the reasons why a user may want to apply an *ab initio* motif discovery program such as MEME (14) to the peak regions found by ChIP-Peak. In order to do so, we proceed as follows. Since MEME can only process a limited amount of DNA sequence data in reasonable time, we repeat the peak finding step with a high count threshold of 200 counts per window. We also activate the RepeatMasker checkbox on the web form, which excludes peaks falling into annotated repeat regions from the output. This is important, because parts of dispersed repeats such as SINEs and LINEs may be picked up as motifs by programs like MEME, if present in the input sequence set. The results page of ChIP-Peak includes a web form, which enables users to extract DNA sequences within an adjustable range relative to the peak center positions. Figure 2e shows the sequence around the STAT1 peak shown in Figure 2c as extracted by this tool. After uploading the complete Fasta-formatted sequence file to the MEME server, one obtains the motif displayed in Figure 2f as a sequence logo.

Cross-genome conservation of STAT1 sites

An *in vivo* occupied STAT1 site may still not be functional. One way to address the question of biological function is by measuring the degree of cross-genome sequence conservation across related species. To allow for this type of analysis, the ChIP-Seq server offers as server-resident file, a compact SGA-formatted version the PhastCons genome conservation track from the UCSC genome browser database. Figure 2g shows the degree of cross-genome conservation around consensus STAT1 binding sites found in peak regions. This figure has been generated by successively using three tools from ChIP-Seq and SSA servers in a pipeline. ChIP-Peak was first used to generate a peak list. This list was then uploaded to the FINDM program of the SSA server in order to generate a new list containing the genomic coordinates of STAT1 consensus sequences occurring within 100 bp from a peak center position. This new list was subsequently uploaded as reference feature to the ChIP-Cor server in order to generate the sequence conservation profile shown in Fig 2g. We observe a sharp peak centered at position zero surrounded by a larger region of increased sequence conservation, suggesting that *in vivo* bound STAT1 sites are indeed conserved across species, and in addition tend to occur as parts of larger conserved regulatory regions of up to 300 bp.

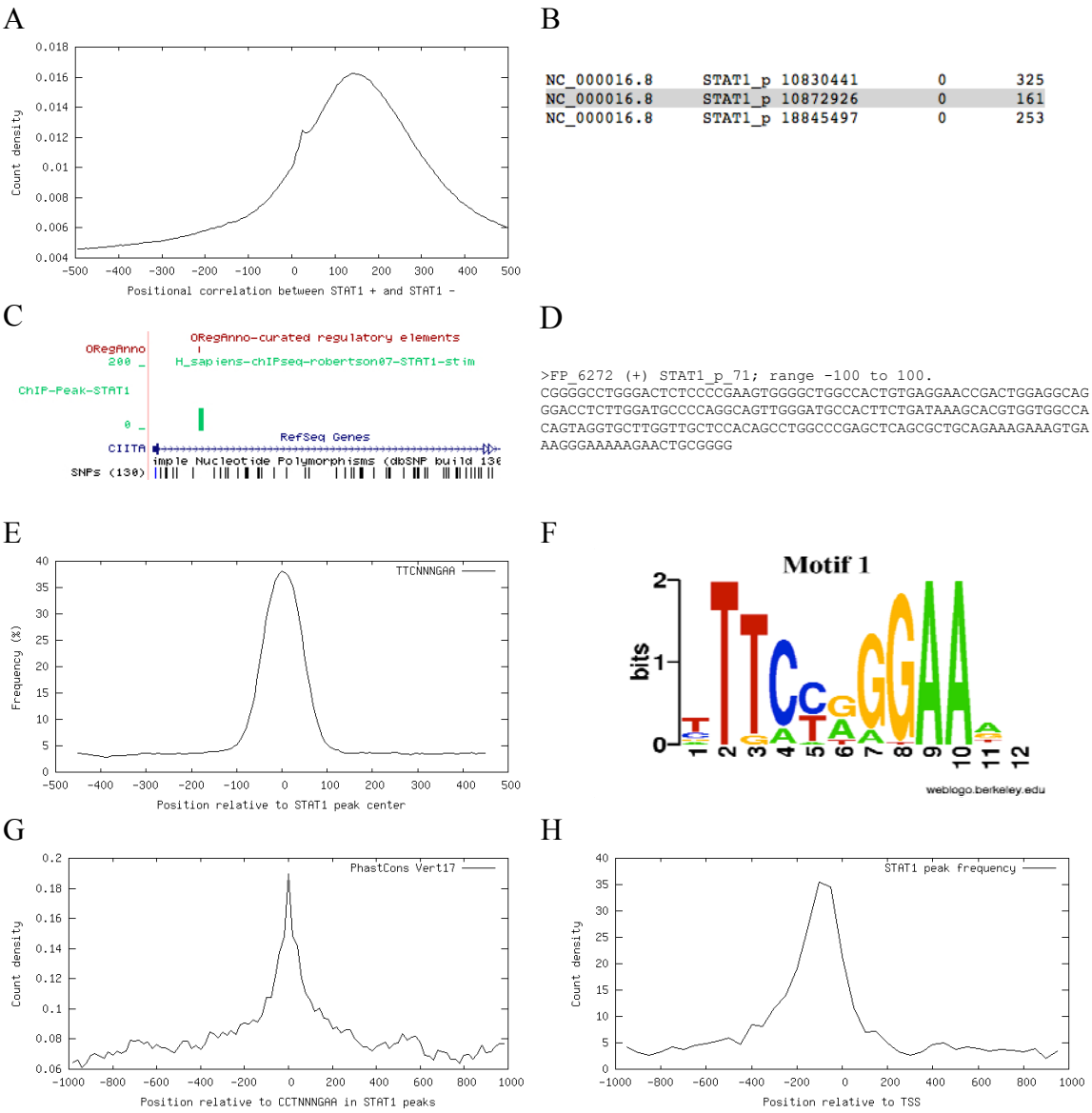


Figure 2: Examples of results produced with the ChIP-Seq server. (**A**) 5'-3' tag correlation analysis generated with ChIP-Cor. (**B**) List of STAT1 peaks in SGA format generated by ChIP-Peak. The zero value in the strand field indicates that this feature has no intrinsic orientation. The fourth field contains the peak center position, and the fifth field indicates the number of tags occurring in a surrounding window of 200 bp. (**C**) ChIP-Peak output viewed in the UCSC genome browser. The vertical bar in the STAT1 track corresponds to the line highlighted in B. Also shown is a corresponding STAT1 site from the ORegAnno track provided by the genome browser. (**D**) Sequence in Fasta format extracted from the ChIP-Peak results page. The sequence corresponds to the peak shown in C. (**E**) Consensus sequence enrichment in the vicinity of STAT1 peaks. This graph was produced with the OProf (Occurrence profile) program from the SSA (Signal Search Analysis) server. (**F**) Sequence logo of the binding site motif discovered in STAT1 peak regions by MEME software. (**G**) Cross-genome conservation profile around STAT1 consensus sequences occurring near peak center positions. This program was generated by successively running the programs ChIP-Peak, FINDM, and ChIP-Cor via the web interfaces of the ChIP-Seq and SSA servers, illustrating the interoperability of the two resources. (**H**) Distribution of STAT1 peaks relative to transcription start sites. This figure has been generated by ChIP-Cor using the server-resident transcription start site collection from the Ensembl database as reference feature.

Enrichement of STAT1 peaks in promoter regions

Another question of interest to biologists is whether *in vivo* bound STAT1 binding sites preferentially occur in promoter regions. The ChIP-Seq server offers several transcription start site collections as server-resident files. In order to generate the STAT1 peak binding site occurrence profile shown in Figure 2h, we upload the SGA-formatted output from ChIP-Peak as target feature to the ChIP-Cor server and select the Ensembl TSS collections as reference feature. Here we selected the "global" normalization mode, which displays the target feature abundance as fold change relative to the genome average. The result of this analysis shows that *in vivo* bound STAT1 binding sites are indeed strongly over-represented in promoter regions.

Other applications

The tools offered by the ChIP-Seq server are also well suited for the analysis of genome-wide maps of epigenetic marks such as histone modification, another major application area of the ChIP-Seq technology. In fact, data from many other high-throughput functional genomics techniques can be analyzed as well. The only condition is that the resulting data can be represented by an SGA file in a meaningful manner. For instance, the ChIP-server offers CAGE data from the Fantom4 projects as server resident files. CAGE is a technology for experimentally mapping transcription start sites in eukaryotic genomes. The interested reader is referred to the ChIP-Seq tutorial, which presents many examples from these other application areas as well.

DISCUSSION

We will by briefly compare our web server to the other two web-based resources mentioned in the introduction. All three web servers offer peak detection tools. However, our peak detection program is much faster. ChIP-Peak often responds after less than one minute whereas the other two peak finders take at least an hour for the same task according to our experience. The ChIP-seq server is the only resource which offers a tool (ChIP-Part) specifically designed to find very large signal-enriched regions, occurring for instance in histone modification maps. W-ChIPeaks does not support any third level analysis. Sole-Search offers some support for third level analysis, including sequence extraction for subsequent motif analysis, and a service for mapping peaks to the closest gene in the genome. The peak-to-gene mapping function is missing at our server. On the other hand, we offer a rich collection of server-resident data files for learning and demonstration purposes. Automatic repeat masking is another unique feature of the ChIP-Seq server, particularly useful in combination with DNA motif analysis. The modular design and the interoperability with the motif analysis platform SSA are perhaps the most distinctive features of our resource. Together, they enable users to carry out complex analysis tasks by using individual tools in a pipeline, thereby expanding the range of potential application in a combinatorial fashion.

The development of the ChIP-Seq server is still an ongoing process. There are many ways how the current server could be improved with relatively little work investment. We will mention only a few of them. We will add more genomes to the seven assemblies already supported now. We are continuously adding more experimental data sets and genome annotation tracks to the database of server-resident SGA files. We further plan to support more input data formats in the near future, and will try to make data-upload more robust by improving the format checking utilities. Finally, we plan to develop the educational aspects of the server by posting ready-to-use teaching material including web-based exercises for hands-on courses in ChIP-Seq data analysis. We hope that such improvements will make the ChIP-Seq server even more useful to a larger user community in the near future.

ACKNOWLEDGEMENTS

The ChIP-Seq server is funded by a grant from the Swiss Federal government. We would like to thank the Vital-IT team from the Swiss Institute of Bionformatics for maintenance of the computer hardware behind the server.

REFERENCES

1. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 10, 669-680.
2. Leleu, M., Lefebvre, G. and Rougemont, J. (2010) Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. Brief Funct Genomics, 9, 466-476.
3. Rougemont, J., Amzallag, A., Iseli, C., Farinelli, L., Xenarios, I. and Naef, F. (2008) Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics, 9, 431.
4. Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol, 11, R86.
5. Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M. and Wong, W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol, 26, 1293-1300.
6. Ye, T., Krebs, A.R., Choukrallah, M.A., Keime, C., Plewniak, F., Davidson, I. and Tora, L. (2010) seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res.
7. Lan, X., Bonneville, R., Apostolos, J., Wu, W. and Jin, V.X. (2011) W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Bioinformatics, 27, 428-430.
8. Blahnik, K.R., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A.R., Iyengar, S., Nicolet, C.M., Ludascher, B., Korf, I. et al. (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. Nucleic Acids Res, 38, e13.
9. Ambrosini, G., Praz, V., Jagannathan, V. and Bucher, P. (2003) Signal search analysis server. Nucleic Acids Res, 31, 3618-3620.
10. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y.J., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods, 4, 651-657.
11. Schmid, C.D. and Bucher, P. (2010) MER41 repeat sequences contain inducible STAT1 binding sites. PLoS One, 5, e11425.
12. Montgomery, S.B., Griffith, O.L., Sleumer, M.C., Bergman, C.M., Bilenky, M., Pleasance, E.D., Prychyna, Y., Zhang, X. and Jones, S.J. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics, 22, 637-640.
13. Wilbanks, E.G. and Facciotti, M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One, 5, e11471.
14. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res, 37, W202-208.