Command line analysis of ChIP-seq results

Endre Barta[1,2]

[1]Department of Biochemistry and Molecular Biology and Apoptosis and Genomics Research Group of the Hungarian Academy of Sciences, Hungary;
[2]University of Debrecen, Medical and Health Science Center, Research Center for Molecular Medicine, Egyetem ter 1. Debrecen, H-4010, Hungary

Abstract

Among the emerging next-generation sequencing technologies, ChIP-seq provides a very important tool for functional genomics studies. From the bioinformatics point of view, ChIP-seq analysis involves more than simply aligning the short reads to the reference genome. It also completes several other downstream steps, like determination of peaks, motif finding and gene ontology enrichment calculation. For these, several programs, applications and packages are available, both free and commercial. In this article I describe the usage of two free ChIP-seq analysis packages, the HOMER and ChIPseeqer along with the MACS and MEME programs. I also provide a customisable script suitable for the complete analysis of raw ChIP-seq sequencing data either from a sequence read repository or directly from sequencing.

Introduction

In the post-genomic era, ChIP-seq (Chromatin immunoprecipitation followed by next-generation sequencing) [1] soon became one of the most exciting technologies for functional genomic studies. Using ChiP-seq one can nearly determine the exact transcription factor binding sites (TFBSs) and histone modifications genome-wide. The experimental part of ChIP-seq consists of crosslinking proteins sitting on the chromosomes into the DNA, followed by a fragmentation step, which ideally yields 100-200 basepairs DNA-protein pieces. The next step is the immunoprecipitation of the specific fragments with a corresponding antibody. After some purification steps and library preparation, a short tag from either of the ends of the precipitated fragments will be sequenced using a next-generation sequencing method. The bioinformatic part of this analysis consists of the following three main steps:

- alignment of the short reads to the reference genome;
- finding significant peaks;
- downstream analysis, including de novo and known motif finding or analysis of the gene lists associated with the peaks.

In a typical ChIP-seq experiment it may be enough as less as 10 million reads for the saturation in an analysis. Hence, the main challenge today is neither the alignment of the reads to the reference genome, nor the peak finding, but rather the downstream bioinformatic analysis and the overall handling and maintaining of the different types of data generated by the analysis.

We would need to carry out a complex ChIP-seq analysis for the following reasons:

- to process our own experimental ChIP-seq data;
- to re-process already published ChIP-seq data to compare more detailed results with what the authors provided in the original articles;
- to make a meta-analysis of similarly processed ChIP-seq experiments from different sources.

In this article I describe an approach how it is possible to process ChIP-seq data from different experiments automatically, starting either from the SRA format files from NCBI [2], or FASTQ format files, or BAM format files which contain aligned reads. Among the many different available genome alignment tools I use the BWA. After aligning the short reads to the reference genome, the resulted SAM format files are converted on fly into the binary BAM format using the SAMtools program [3]. From the BAM format alignment files I use three different methods to find and analyse peaks. The first is the MACS (Model-based Analysis for ChIP-Seq) program [4] followed by the de novo motif finding using the MEME (Motif-based sequence analysis tools) program [5]. The second is the HOMER software [6] for motif discovery and ChIP-Seq analysis. The third one is the ChIPseeqer [7], which is a comprehensive framework for the analysis of ChIP-seq data. Using the three different approaches in parallel ensures that the analysis will be comprehensive and will cover every aspect of the possible questions.

Installation of the programs

Hardware requirements

To carry out such a comprehensive analysis we need a UNIX based computer. The operating system in principle can be any UNIX distribution. I have tested several LINUX and MacOSX Snow leopard based machines. Many steps in the analysis do not require any extra computing power, but the short read alignment, the peak and the *de novo* motif finding can run for a long time on slow processors. Furthermore, these steps need more memory and of course, the more memory we have the faster the finishing of the processes is. The storage capacity is an important factor as well. We need to have reference genome sequences and indices in place for the mapping. The raw sequencing reads and other files created during the analysis also need a lot of disk space. However, ChIP-seq analysis needs much less computing power, memory and storage capacity than other NGS tools. In summary, a PC with 1-2 terabytes (TB) of disk, minimum of 8 gigabytes (GB) of RAM and one recent processor with at least two cores can be enough, although in this case the alignment and the de novo motif finding steps can take days or even more than a week. I tested the programs on a mid 2010 Apple iMac computer (2.93 GHz Intel core i7, 16GB RAM, 2 TB disk), and it worked smoothly. Ideally, we need a machine with at least 16 GB of RAM, two processors with several cores, and at least 4 TB of raid storage.

Software environment

For the analysis several UNIX based (C, C++, PERL, PYTHON etc.) programs and scripts need to be installed. To run and compile these tools, beside the standard UNIX packages (like PERL, PYTHON or the wget), we will also need the developmental packages (Xcode for MacOSX or dev packages for LINUX distributions) for compiling programs from source code. For running MEME in parallel we need one of the MPI (Message Passing Interface) implementation and PBS (Product Breakdown Structure), for example if we want to run it on a supercomputing environment.

In general, installing the main C, C++ developmental packages, with their dependencies on a LINUX based machine, or the XCode package on a MacOSX based machine, can be enough. Otherwise, during the installation we can learn from the logs and error messages which package is missing or needs to be upgraded.

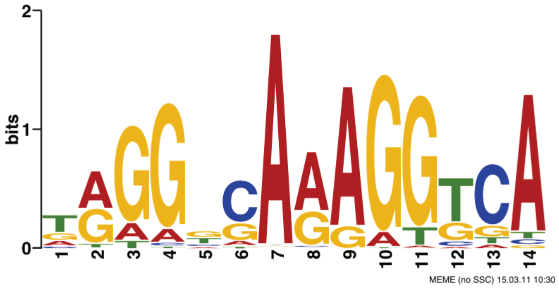


Figure 1. Motif finding result using parallel MEME on a macrophage PPARg ChIP result [14].

Specific programs/packages for the analysis

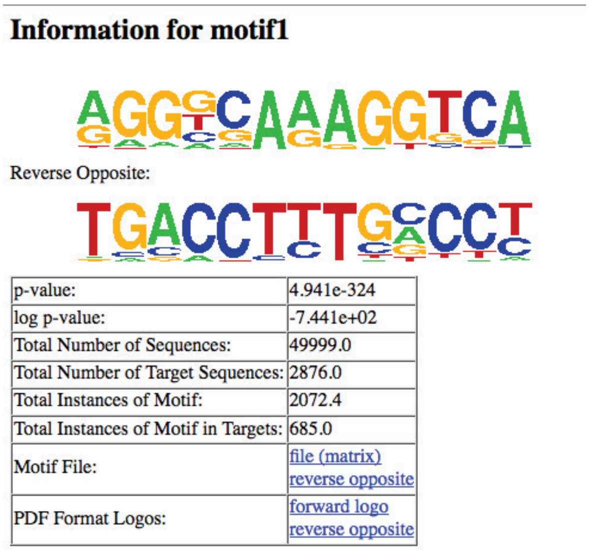
SRA toolkit is needed if we would like to download and process published raw ChIP-seq sequencing data from the NCBI SRA database. We can use both SRA and SRA-lite format data. Compiled binaries for several operating systems can be downloaded. After unzipping and untarring the files we only need to put the directory into the $PATH variable in the SHELL. The advantage of using the NCBI's SRA approach is that we need to transfer smaller files and we do not need to take care of the sequencing methods used.

BWA (Burrows-Wheeler Aligner) [8] is used for the alignment of the short reads to the reference genome. We need to download and compile the latest source code and to put it in the $PATH (I am using /usr/local/molbio/bin). For the alignment we also need to download and index the reference genomes. For this, I am using the human hg18 (because this is available for ChIPseeqer) and the mouse mm9 genome sequences. A script for making the indexing is available as well.

MACS is the most wildly used peak finding program. It is written in PYTHON, so we need to have it installed before installing MACS.

MEME is a de novo motif finding program, suitable for scan ChIP-seq peaks, or peak-centred regions for possible motifs (binding sites for the transcription factor used in the experiment). The main advantage of the MEME program is that it is still actively developed and it can be used under supercomputing environment. Compiling MEME on a grid computer needs properly installed MPI environment.

HOMER is a software package for motif discovery and ChIP-Seq analysis. Installing HOMER is very simple, we only need to have the proper



Information for motif1

Reverse Opposite:

p-value:	4.941e-324
log p-value:	-7.441e+02
Total Number of Sequences:	49999.0
Total Number of Target Sequences:	2876.0
Total Instances of Motif:	2072.4
Total Instances of Motif in Targets:	685.0
Motif File:	file (matrix) reverse opposite
PDF Format Logos:	forward logo reverse opposite

Figure 2. Motif finding result from the HOMER analysis on a macrophage PPARg ChIP result [14].

software environment and to download and run the *configureHomer.pl* script. There are also some third-party software needed for proper working of HOMER programs and scripts. The *configureHomer.pl* script is also suitable to download the different genomes for the analysis. In the HOMER web page there is a detailed instruction how to install and use the package.

ChIPseeqer [7] is an integrative and comprehensive framework that allows the users to perform in-depth analysis of ChIP-seq datasets through easily customised workflows. Besides the command-line tools, the newest ChIPseeqer also contains a GUI (Graphical User Interface) suitable for detailed analysis of ChIP-seq results. The ChIPseeqer package relies on some other programs (like FIRE, PAGE etc.) developed on the same laboratory [9]. In the latest version these programs are compiled together with the main programs and scripts, but we still need to manually set some variables (CHIPSEEQERDIR, FIREDIR, PAGEDIR, MYSCANACEDIR) and the PATHs for programs and PERL libraries.

For this analysis I have used the BWA version 0.5.9, MACS version 1.4.0rc2, which is working with PYTHON 2.6 to 2.7 (2.6.5 recommended). MEME version was 4.5 (4.6 available now with specific option for ChIP-seq region analysis). The HOMER version was 2.6 and the ChIPseeqer version was 1.0.

For converting the SAM files to BAM format I have used the SAMtools program, while for converting the BAM files to BED format for HOMER, I have used the BEDTools utility [10]. For tag statistics I have used the BAMTools utility [15].

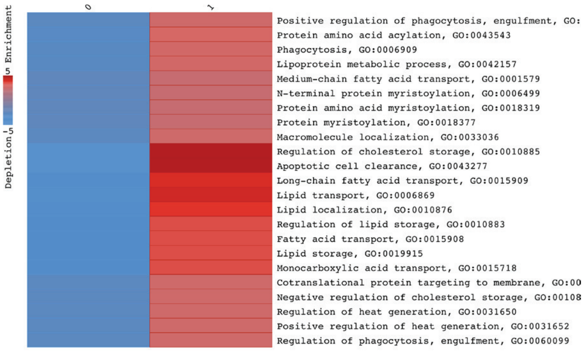


Figure 3. ChIPseeqer gene ontology analysis result on a macrophage PPARg ChIP result [14].

Implementation

The main goal for using these tools together is to generate data for comparison of the ChIP-seq results. Furthermore, this approach is suitable to process data from different sources. To make the analysis I had written two BASH SHELL scripts. The first is to run the programs and to carry out the analysis, while the second is to extract some statistics from the result. In the current implementation (version 1.1), the analysis script accepts either SRA or FASTQ (not colour spaced) format raw sequencing data, or BAM format alignments. If the BAM format alignment files are present in the proper location, the script omits the alignment steps. The script basically needs two parameters, the name of the file where the experiments to be analysed are listed, and the location of the basic directory for the analysis. It is also possible to provide an additional directory name, where the big files (SRA, FASTQ, SAI) will be stored and can be removed later safely. In the list file, there can be two fields separated by a space. The first is the name of the experiment starting with mm (mouse) or hs (human). I am using the format *hs _ celltype _ antibody _ condition*. The second field can be the FTP locations of SRA or SRA-LITE format raw sequencing data at the NCBI, or the FTP locations of fastq.gz format raw sequence data at the EBI SRA FTP site. If either the FASTQ or the BAM format file for the given experiment is available at the proper location with the proper name, e.g. *base _ directory/experiment _ name/bam/experiment _ name.bam*, this field can be

Table 1. Example of the statistics extracted from the log and result files of the analysis using the get _ tag _ statistics-v1 _ 0.sh script. The raw sequence data for the basic analysis with the *ChIP-seq _ anal-v1 _ 0.sh* script were obtained from different sources [11-14].

Article	experiment	FASTQ (raw reads)		BAM (BWA mapping)		HOMER analysis								MACS analysis					ChIPseeqer analysis		
		No of reads	% unique reads	No of total reads	% reads mapped	total tags after filtering	Average tags per peaks	Peaks width	Frag-ment length	No of peaks	IP effi-ciency	% peaks filtered by local signal	% clonal peaks filtered	tag-size	Total tags	% tags filtered out	Frag-ment length	No of peaks	No of peaks	Avg peak height	Avg peak size
Mikkelsen	mm_L1_PPARg1_d7	17,885,244	69.8%	17,996,284	46.7%	7,641,053	1.0775	285	53	11,160	3.12	15.1%	0.01%	61	8,396,770	15.6%	243	8,561	1,558	27	572
	mm_L1_PPARg2_d7	11,938,762	63.8%	12,212,356	52.6%	5,695,805	1.0868	96	53	4,375	1.35	26.0%	0.00%	71	6,426,779	18.5%	301	3,793	818	26	669
	mm_L1_CTCF_d7	24,082,065	56.6%	24,527,548	62.8%	9,638,738	1.0505	240	111	54,437	31.17	8.3%	0.01%	36	15,405,658	19.3%	96	59,551	26,093	48	434
	hs_ASC_CTCF_d9	21,972,388	63.7%	22,315,270	47.3%	8,979,919	1.0086	280	173	48,311	40.04	5.2%	0.00%	76	10,546,650	15.6%	154	47,179	30,515	53	429
	Hs_ASC_PPARg1_d9	34,635,364	66.1%	34,775,194	30.3%	9,315,493	1.0043	255	136	48,327	11.54	10.3%	0.00%	76	10,534,909	11.9%	156	39,902	9,930	27	325
	Hs_ASC_PPARg2_d9	42,987,395	65.7%	43,123,317	16.4%	6,316,958	1.0031	244	133	31,581	8.89	5.7%	0.00%	76	7,062,963	10.8%	154	27,245	5,749	23	301
O'Geen	hs_GM_TR4_2	28,598,265	83.9%	28,177,320	79.5%	21,896,016	1.0056	264	41	10,387	2	30.5%	0.13%	32	22,400,007	1.3%	33	7,968	1,616	36	402
	hs_HeLa_ELK1_1	69,333,573	36.3%	43,345,862	54.2%	21,241,382	1.0066	309	41	6,551	1.69	41.3%	0.12%	32	23,474,148	2.2%	33	6,581	1,587	30	332
	hs_HeLa_ELK4_2	51,888,861	23.0%	28,402,913	59.8%	14,107,350	1.0078	331	41	9,831	2.6	23.7%	0.09%	32	16,982,796	3.9%	34	11,898	2,170	25	315
	hs_HeLa_TR4_2	25,624,681	88.8%	25,129,235	52.6%	12,892,737	1.0033	155	106	9,926	2.38	18.3%	0.06%	32	13,208,389	0.9%	47	6,349	2,011	47	406
	hs_HepG2_TR4_2	16,347,187	85.7%	16,276,605	79.4%	12,509,867	1.0042	275	41	4,904	2	31.3%	0.17%	32	12,917,727	1.0%	35	4,671	1,307	39	430
	hs_K562_TR4_2	14,851,399	88.0%	14,855,527	72.1%	10,490,292	1.0057	192	41	2,059	1.28	60.1%	0.14%	32	10,705,127	2.6%	33	1,559	679	28	394
Nielsen	mm_L1_PPARg_d0	12,025,045	74.9%	12,025,037	74.1%	7,872,656	1.0375	52	85	6,502	1.38	21.5%	0.19%	32	8,911,122	14.9%	39	4,036	635	25	421
	mm_L1_PPARg_d1	13,588,664	63.1%	13,588,660	74.4%	8,325,879	1.0363	58	90	7,639	1.48	19.9%	0.43%	32	10,104,943	20.6%	40	4,226	731	26	463
	mm_L1_PPARg_d2	18,154,176	61.3%	18,154,159	40.4%	5,990,659	1.0482	96	41	15,961	3.8	7.6%	3.40%	32	7,339,281	22.1%	42	11,075	918	23	302
	mm_L1_PPARg_d3	14,392,918	71.8%	14,392,892	76.6%	9,615,931	1.0393	67	87	10,816	1.73	16.8%	0.19%	32	11,023,487	16.2%	40	5,191	721	29	632
	mm_L1_PPARg_d4	15,034,953	57.7%	15,034,953	60.2%	6,938,366	1.0419	70	80	19,773	3.34	8.4%	0.26%	32	9,048,496	10.4%	50	13,936	1,512	23	328
	mm_L1_PPARg_d6	15,140,624	62.1%	15,140,621	50.0%	6,094,142	1.0392	94	88	36,277	8.67	5.8%	0.32%	32	7,573,685	22.6%	69	34,417	6,340	27	330
	mm_L1_RNAPII_d0	8,573,748	82.0%	8,573,748	89.9%	7,204,654	1.0208	100	106	22,922	5.45	25.2%	0.00%	32	7,707,149	8.4%	111	15,948	4,665	23	368
	mm_L1_RNAPII_d1	9,715,568	80.9%	9,715,568	83.7%	7,500,956	1.0181	89	78	32,578	8.07	39.8%	0.01%	32	8,127,267	9.3%	70	19,524	7,137	25	422
	mm_L1_RNAPII_d2	10,318,412	73.4%	10,318,404	80.7%	7,308,686	1.0173	90	88	32,151	8.06	42.0%	0.01%	32	8,324,284	13.7%	85	18,056	6,803	24	430
	mm_L1_RNAPII_d3	8,905,079	81.4%	8,905,071	83.7%	6,890,426	1.0155	84	86	30,990	7.4	36.3%	0.01%	32	7,453,351	9.0%	86	18,140	5,588	23	428
	mm_L1_RNAPII_d6	8,979,276	65.1%	8,979,275	87.4%	6,502,922	1.048	48	79	12,327	2.44	23.1%	0.01%	32	7,850,700	21.0%	70	10,704	1,293	21	381
	mm_L1_RXR_d0	7,767,412	73.0%	7,767,407	67.7%	4,634,488	1.0153	61	100	17,134	3.78	7.0%	0.30%	32	5,259,011	13.2%	49	10,298	928	20	285
	mm_L1_RXR_d1	8,549,694	67.8%	8,549,651	81.3%	5,923,120	1.0182	82	96	32,111	5.74	6.6%	0.05%	32	6,953,474	16.3%	93	22,644	2,844	21	302
	mm_L1_RXR_d3	9,503,443	68.6%	9,503,443	63.4%	5,093,136	1.0317	71	78	38,857	6.22	4.5%	0.10%	32	6,021,891	18.0%	56	13,198	1,619	21	294
	mm_L1_RXR_d4	14,468,617	44.9%	14,468,610	40.8%	3,948,339	1.0334	73	85	36,722	8.49	4.3%	0.47%	32	5,902,838	35.3%	52	17,633	2,302	20	286
	mm_L1_RXR_d6	9,395,192	68.0%	9,395,188	66.1%	5,191,748	1.0273	93	92	39,302	9.95	5.5%	0.11%	32	6,207,267	18.6%	87	32,595	6,290	25	321
Lefte-rova	mm_L1_PPARg	7,502,114	66.0%	7,502,114	88.9%	5,660,584	1.0483	102	41	5,540	1.82	19.3%	0.06%	36	6,668,116	19.0%	79	6,866	618	27	863
	mm_mac_CEBP	7,898,957	42.6%	7,898,957	84.7%	4,665,715	1.0335	98	81	38,024	13.44	3.8%	0.01%	36	6,689,884	32.5%	80	34,765	10,310	26	307
	mm_mac_PPARg	17,389,919	66.9%	17,323,524	75.1%	11,348,348	1.0338	121	41	3,769	0.81	20.9%	0.29%	36	13,002,396	6.9%	39	2,635	500	30	386
	mm_mac_PU1	9,438,501	65.9%	9,382,873	77.0%	6,278,059	1.0155	105	82	60,996	17.75	4.0%	0.00%	36	7,220,167	14.4%	76	48,032	15,728	26	300

empty. The script will create the directory structure for the experiments and run the programs. All the logs and standard error messages related to program running will be put to the *base _ directory/experiment _ name/logs* directory, while the results will be put to the *bam, macs, homer and chipseeqer* directories.

The scripts performs the most basic ChIP-seq related analyses including the alignment to the reference genome, peak finding using MACS, HOMER and ChIPseeqer, quality control calculations using HOMER and ChIPseeqer, *de novo* motif finding using MEME (Fig. 1), HOMER using FindPeaks (Fig. 2) and ChIPseeqer. HOMER and ChIPseeqer also carry out a detailed annotation of the peak regions as well as a Gene Ontology analysis (Fig. 3). They also have some other programs and scripts suitable for many different tasks connected to the analysis. These include, among others, the comparing of the ChIP regions (peaks) and combining them into common sets. The results are mostly available as BED format files for peak regions that are suitable for visualisation in any of the available genome browsers. The results of de novo motif finding programs are available as local HTML pages (HOMER, MEME), or in EPS and PDF format files (ChIPseeqer). The annotation files generally can be found as tab delimited text files available for direct import into spreadsheet programs. The summary of statistics about the experiment is generated as comma separated values (CSV).

Conclusions

It is now getting easier and easier to carry out parallel ChIP-seq experiments using multiplexed next-generation sequencing. The SRA at the NCBI, and the ENA (European Nucleotide Archive) at the EBI, host more and more raw and processed ChIP-seq sequencing result. As a consequence, this growing data are available for detailed analysis and for comparing to our own results. To deal with this ever-increasing amount of next-generation sequencing data, we need bioinformaticians capable to work on a UNIX environment. For them, using command line tools for processing and comparing ChIP-seq data can be a good alternative to commercially available GUI version program packages. Here, I provide a layout with example scripts to conveniently analyse, and thus easily compare, ChIP-seq experiments from different sources. The method can be very useful not only for processing our own data but also to compare our data to other's, or simply to make a ChIP-seq meta-analysis using the available ChIP-seq raw sequence data at the sequence read repositories. The main advantages of this approach are that: i) it can be run on a minimal, low-budget hardware; ii) provides comparable data for every aspect of the analysis; iii) is easy to customise for any personal needs.

Availability

The scripts are available upon E-mail request or from the Facebook Page "Command line ChIP-seq analysis". The Facebook Page is also meant to be a place for providing further details about installing and using these programs and scripts, for announcing improvements or new versions of programs and scripts, and also to exchange experiences about using command line tools for ChIP-seq analysis.

Acknowledgments

I would like to thank Prof. László Nagy and Bálint L. Bálint for inspiring and helping my work, and for the support of the Hungarian Academy of Sciences Bolyai Scholarships. This work was supported by grants from the Hungarian Science Research Fund OTKA NK72730, and the NKTH-ANR TéT BOV-rSNP. Some of the programs were running on the GenaGrid supercomputer, kindly provided by the GenaGrid consortium.

Competing interest statement

None declared

References

1. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669-680.
2. Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. Nucleic Acids Res 39: D19-21.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
4. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.
5. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.
6. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38: 576-589.
7. Giannopoulo E, Elemento O (2011) Characterizing ChIP-seq peaks using customizable analysis workflows (submitted).
8. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
9. Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28: 337-350.
10. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
11. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, et al. (2010) Comparative epigenomic analysis of murine and human adipogenesis. Cell 143: 156-169.
12. O'Geen H, Lin YH, Xu X, Echipare L, Komashko VM, et al. (2010) Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. BMC Genomics 11: 689.
13. Nielsen R, Pedersen TA, Hagenbeek D, Moulos P, Siersbaek R, et al. (2008) Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. Genes Dev 22: 2953-2967.
14. Lefterova MI, Steger DJ, Zhuo D, Qatanani M, Mullican SE, et al. (2010) Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. Mol Cell Biol 30: 2078-2089.
15. github SOCIAL CODING: [http://github.com/pezmaster31/bamtools]