ChIP-Seq: The new way to 'seq' genome wide for transcription factor binding patterns

Jaclyn Lim

Biochem218: Computational Molecular Biology

June 6th, 2010

Understanding the transcription regulatory network at the genome level has been a focus of many researchers over the past decade and it remains rightly so. By unlocking how genomic information is translated into gene regulation will allow us to better understand evolution, development and biological processes that have gone astray leading to diseases such as cancer. So far, our understanding of gene regulation has been derived from studies that have focused on detailed characterizations of a specific gene or gene family, but genome-scale analyses are now hinting the re-evaluation of such principles. For example, it is previously thought that a typical RNA polymerase II promoter contains a TATA box that is located 30bp upstream of the transcription start site. However, this may not be entirely true as recent genomic studies have shown that ~50% of human genes have alternative promoters (Kimura et al., 2006). To make matters even more complex, it has been estimated that there are 200-300 transcription factors in humans that bind to core promoters to drive gene transcription (Farnham, 2009). Thus, to fully comprehend the regulatory network of these transcription factors, it is of utmost importance to study the machinery of these proteins at the genomic level.

Fortunately, the technological advancement of chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has allowed investigators to create a global map of specific protein DNA-interactions in a given cell type. ChIP-seq has been widely used to study transcription factor binding, histone modifications and DNA methylation (Aleksic and Russell, 2009). Given the different utilizations of the ChIP-Seq technology, I

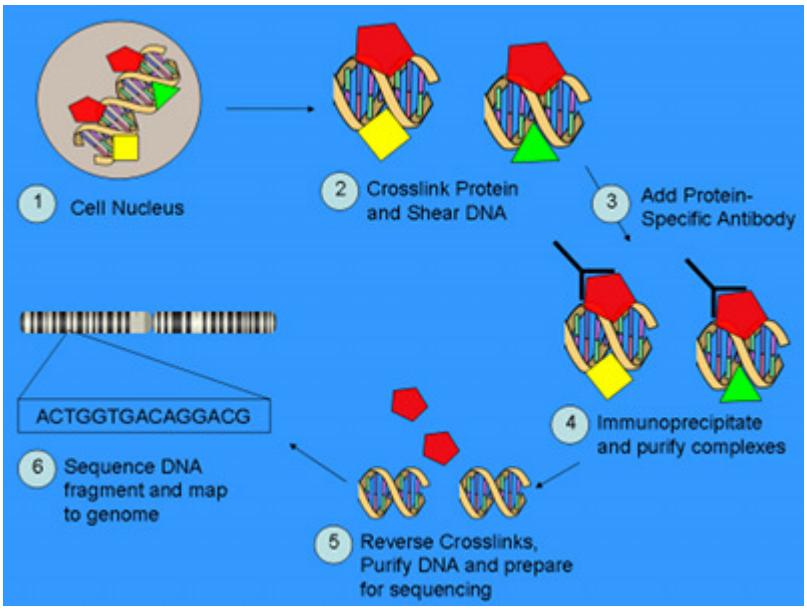

Figure 1: ChIP-Seq is used to analyze protein-DNA interactions (Wikipedia)

will be focusing this review on the application of this technique in identifying transcription factor binding sites (TFBS). In a ChIP-seq experiment (Figure 1), DNA fragments associated with a specific protein are enriched, and then subjected to high-throughput sequencing. The DNA-binding protein is covalently linked to DNA by treating cells with a cross-linking agent, typically formaldehyde. The chromatin is isolated and is sheared by sonication into small fragments, which are generally in the 200–600 bp range. An antibody specific to the protein of interest is used to immunoprecipitate the appropriate DNA–protein complex. The crosslinks are reversed and the released DNA is assayed to determine the sequences bound by the protein. The generated sequences are then aligned back to the genome of interest to identify regions that are enriched with mapped reads, which are often referred to as peaks and mark the location of DNA-protein interaction.

The use of high-throughput sequencing in ChIP-seq experiments offers better advantages over its predecessor, ChIP-chip, where the immunoprecipitated DNA fragments are labeled with fluorescent dyes and hybridized to microarrays. One of the greatest benefits that ChIP-seq offers is its wide coverage that allows for a relatively unbiased genome wide analysis of TFBS. Although arrays can be tiled at a high density in a ChIP-chip experiment,

this requires a large number of probes and poses a significant cost-challenge, especially for large mammalian genomes. For example, a recent genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs used a total of 37 arrays to survey the mouse genome (Barrera et al., 2008). Furthermore, the base pair resolution offered by ChIP-seq is an added bonus of this technique. Probes used in ChIP-chip experiments are typically several hundreds of base pairs in length, making identification of the actual TFBS difficult. Contrary to ChIP-chip, the actual binding site of a factor often lies within 10-30bp of the peak in a ChIP-seq analysis (Kharchenko et al., 2008). The base pair resolution of ChIP-seq can also allow for identification of novel transcription factors binding motif. A recent analysis identifying β-catenin binding regions in colon cancer cells using ChIP-seq clearly illustrate this advantage (Bottomly et al., 2010). Thirdly, ChIP-seq produces fewer artifacts compared to ChIP-chip as it does not suffer from the noise generated by the hybridization step in a ChIP-chip experiment. The GC content, length, concentration and secondary structure of the target and probe sequences often contribute to cross-hybridization between imperfectly matched sequences (Hoffman and Jones, 2009). Fewer DNA amplification cycles are needed for ChIP-seq and this helps in minimizing the number of artifacts that can arise from PCR bias. Lastly, for investigators working with limited samples, ChIP-seq is a preferred tool due to its relatively low input sample requirement (~10ng) (Park, 2009).

Given the repertoire of benefits offered by ChIP-seq, it is still a nascent technology that faces many challenges both experimentally and computationally. In this review, the experimental challenges will be summarized and more attention will be given to address the computational problems associated with ChIP-seq. Like any other ChIP experiments, the accuracy of ChIP-seq depends heavily on the specificity of the antibody used, thereby

necessitating a rigorous validation of the antibody. The ChIP steps in ChIP-seq could give rise to potential artifacts. For example, shearing of the DNA often does not result in uniform fragmentation of the genome. Repetitive sequences can also obscure the validity of ChIP-seq results as those regions might appear to be enriched after the alignment step. Therefore, an appropriate control experiment is necessary to eliminate any sources of artifacts. A well-accepted control in ChIP-seq is comparison with input DNA (DNA prior to precipitation), which corrects for most bias related to the variable solubility of different regions, the shearing of DNA and amplification.

Sequencing errors would result in partial alignment of the short reads with gaps and mismatches. For the widely used sequencing platform, Illumina, the sequencing errors are most prominent at the sequenced 3' tags and notably, the mismatch frequencies towards the 3' termini accounts for 41–75% of all observed mismatches (Kharchenko et al., 2008). To optimize the use of any datasets, caution should be taken to include only partially-aligned tags that will provide biologically important information. Another experimental difficulty in ChIP-seq is determining the depth of sequencing. When a large number of binding sites are present in the genome, one would expect that a larger amount of sequencing is required to obtain a significant fold-enrichment at each bound region. One way to determine the sequencing depth in ChIP-seq would be to figure out the "saturation point" – the number of binding sites identified does not change when more tags are sequenced.

The first computational challenge of ChIP-seq is mapping the reads to the reference genome and this step is one of the most important and computationally intensive of the experiment. A successful ChIP-seq experiment typically generates about 2-20 million mapped reads and it is a daunting task to accurately align these sequences back to the genome. The

large dataset would take conventional alignment algorithms hundreds or thousands of processor hours and thus leads to the development of new generation aligners (Park, 2009). Each aligner is a balance between accuracy, speed, memory and flexibility, and to date, there is no aligner that offers the best of all these aspects. Ideally, the alignment process should be fast and minimizes the number of mismatches due to sequencing errors, SNPs and indels or the difference between the genome of interest and the reference genome. Some of the commonly used aligners are Eland, the default aligner of the Illumina sequencing device which offers efficient and fast alignments of short reads. Another popular aligner that is an excellent SNP detector is Mapping and Assembly with Qualities (MAQ), which utilizes mate-pair information and estimates the error probability of each read alignment using a Bayesian statistical model (Li et al., 2008). Bowtie is an example of an ultrafast aligner that is able to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes (Langmead et al., 2009). Many current analyses do not account for non-unique reads and this would exclude identification of TFBS in repetitive regions. Conversely, including non-unique reads would improve the sensitivity of the analysis, but at the expense of specificity. Therefore, it is imperative to strike a balance between specificity and sensitivity as to optimize the numbers of true positives in the analysis.

The next computational step after alignment is identifying genomic regions that are enriched with sequenced tags. The most basic way to determine enriched domains or "peaks" is by scoring the number of sequenced tags in a window of given size (Pepke et al., 2009). While this method is effective at calling regions with strong ChIP enrichment, it can be highly impaired by the presence of artifacts in the experiment. For example, "open" chromatin states in the genome that are more susceptible to fragmentation, copy number variation and natural polymorphisms will lead to generation of false enrichment regions (Hoffman and Jones, 2009). Additional information will be required to rule out these false-positives and some algorithms have been developed to overcome this challenge. The directionality of sequencing can be adapted to discriminate true binding events from artifacts, as demonstrated by QuEST which uses a kernel density estimation approach to generate peak-calls (Valouev et al., 2008). Because the DNA is sequenced from the 5' end, one would expect that the immunoprecipitated DNA will be sequenced equally from both the Watson and Crick strand,

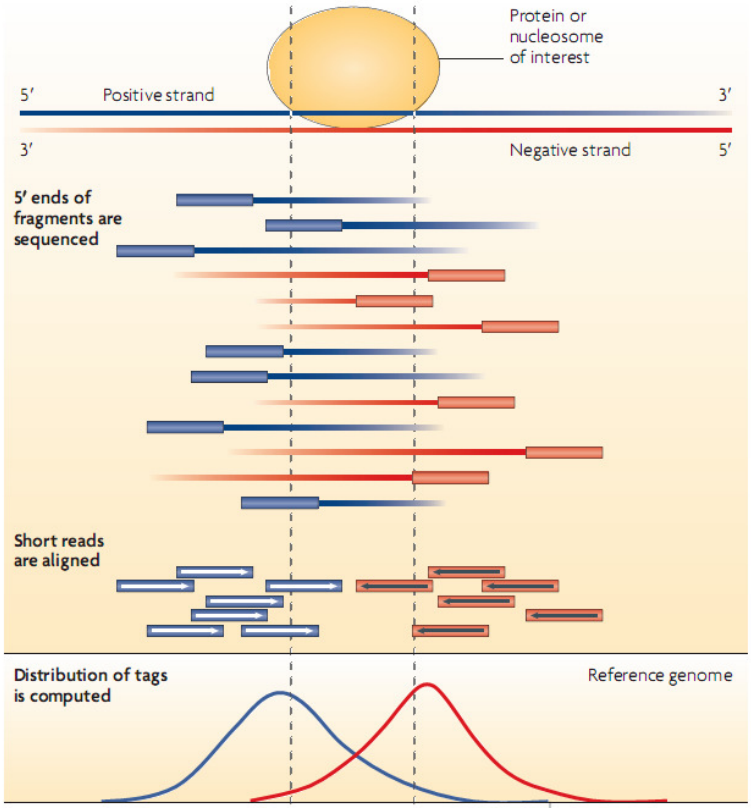

Figure 2: DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5′ end. Therefore, the alignment of these tags to the genome results in two peaks (one on each strand) that flank the binding location of the protein or nucleosome of interest. (Park, 2009)

giving rise to bihorn-peaks that have consistent distance between the two peaks (Figure 2). Peaks that only show enrichment from one direction will indicate presence of artifacts and can be filtered from the analysis. In general, peaks identified in a ChIP-seq experiment can be classified into three groups; sharp peaks that cover a few hundred base pairs or less, localized but broader peaks of up to a few kilobases and broad domains of up to several hundred kilobases. Typically, sharp peaks are associated with protein-DNA binding interactions, as in the case of transcription factors and broader regions are associated with histone modifications of the genomic region (repressive vs. active) (Pepke et al., 2009).

Identification of ChIP-enrichment regions is followed by processing methods that allow for the identification of the original binding site, termed "summits". This is typically done by "smoothening" the profile of each strand (i.e. replacing tag counts at each site with the summed value within the window centered at the site and merging consecutive windows that exceed a threshold value). Both the Watson and Crick strand profiles are combined either by shifting each tag distribution towards the center (Figure 3) or by extending each mapped tags then adding them up together (Figure 4) (Park, 2009; Valouev et al., 2008). In theory, the tag extension method is more precise at identifying the binding site but requires the fragment size to be known and the assumption that all fragment sizes are uniform to be made.

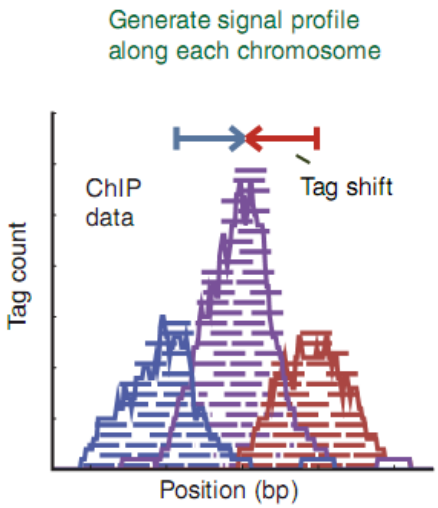

Figure 3: A signal profile is generated by shifting the distribution of the Watson strand (blue) reads and Crick strand (red) reads towards the center (purple). (Pepke et al., 2009)

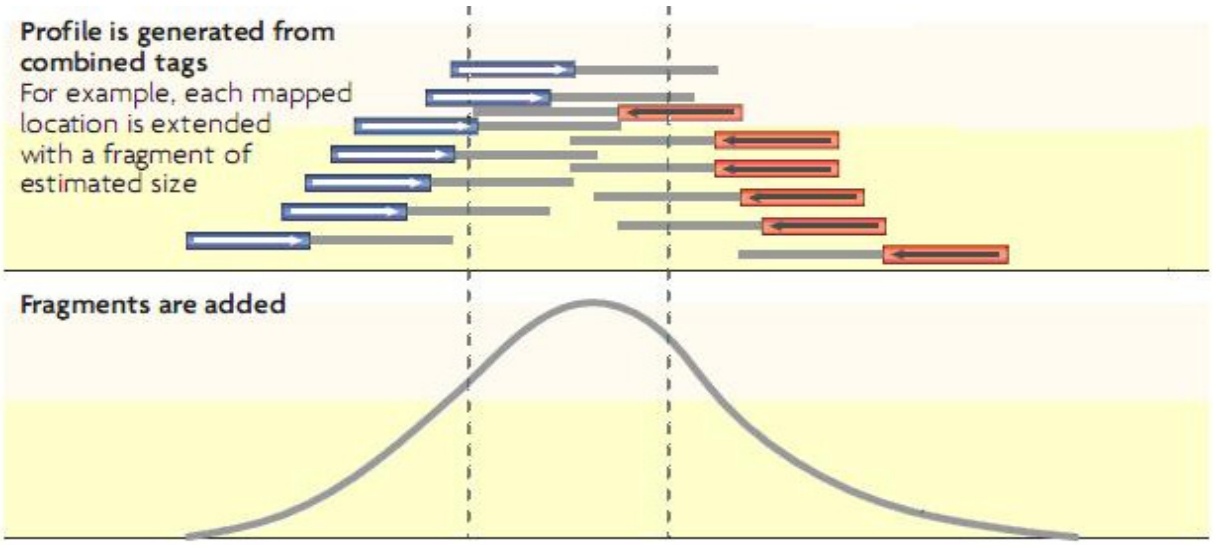

Figure 4: A signal profile is generated by extended each mapped reads by an estimated fragment size and adding them up together. (Park, 2009)

Another important consideration to account for during peak-calling is the background noise level. Measuring the fold ratio of the ChIP sample signal relative to that of a control (e.g. input DNA) is a good indicator of peak validity but it may not be sufficient statistically. For example, a fold ratio of 5 measured from 50 ChIP samples and 10 control samples do not hold the same statistical significance as measured from 500 ChIP samples and 100 control samples (Park, 2009). To bypass this, algorithm developed based on a Poisson distribution model would account for both the number of sequenced tags and fold ratio (Visel et al., 2009). An added advantage of using the Poisson distribution is that it can be modified to account for regional bias in tag density and to model the background tag distribution in the absence of a control (Zhang et al., 2008). Other statistical distributions used to model the background in absence of a control include negative binomial (Ji et al., 2008) and Monte Carlo (Fejes et al., 2008).

Currently, there are many available peak-calling software, each with its pros and cons and are summarized in Table 1. The main criterion in determining the top candidate peaks is either a signal that exceeds a set threshold or a minimum enrichment relative to background or

both. All the software listed in Table 1 utilizes either one or both these criterion and provides a default value. Nevertheless, investigators have the option of adjusting their own parameters when analyzing their data by specifying a false discovery rate that best suits their dataset. Most of the software in Table 1 (e.g. CisGenome) computes a *P*-value and it is often thought that one can compute the FDR based on the given *P*-value for any distribution. However, there is a caveat associated with this method of FDR computation; the distribution assumption made in the *P*-value calculation may not be appropriate and therefore a correct FDR could be very different from the one obtained from the *P*-value threshold. Other programs (e.g. MACS) instead address this problem by calculating the FDR as the ratio of the number of peaks called in the control to the number of peaks called in the ChIP experiment (Pepke et al., 2009).

Once candidate peaks for protein-DNA binding regions are identified, the next step in the analysis is discovery of sequence binding motifs and this can be done through motif-finding algorithms such as MedScan and WebMOTIFS. The sequences for the top candidate peaks are submitted to these algorithms, which will search for potential motifs and return the results with statistical significance associated to every motif identified. In the best case scenario, a single motif will have a much higher statistical significance compared to other matches found but unfortunately, that is not always the case. Sometimes, a series of motif with a gradient of statistical significance is identified and further analysis is needed to examine the possibility of combinatorial interactions between these motifs. The process of motif identification is also not straightforward and any potential binding motif has to be validated before it can be declared bona-fide. As of now, there are no reliable computational methods that can be used to verify a binding motif; instead, experimental methods are used.

Another analysis that can be done after peak-calling is associating these peaks to candidate genes in the genome. The location of the peaks are usually annotated to key features such as the transcriptional start site, exon–intron boundaries and the 3′ ends of genes using correlation analysis and advanced clustering methods. In general, many parameters can be adjusted by researchers when mapping peaks to candidate genes. For example, peaks can be mapped to genes if the peak was within ± 20kb of the gene's transcription start site (TSS) (Johnson et al., 2007) or within -10kb from the TSS to +1kb from the transcriptional termination site (Wederell et al., 2008). Even though there are different methods that can be used to associate peaks to gene, many complications can still arise. This is especially true if the peak falls in a gene-rich region and the gene closest to the peak may not be regulated by that peak. Alternatively, the same peak could be co-regulating all these genes at the same time. The complexity of this issue becomes even more dramatic as one moves from a large mammalian genome to a smaller, more gene-compact genome. These issues could hopefully be solved by better association metrics that are still in development, such as those that uses chromosome conformation capture approaches (Hoffman and Jones, 2009).

The final and most important challenge when analyzing peaks from a ChIP-seq experiment is to determine the functional relevance of the identified binding sites. Many issues have crept up in previous studies that suggest the non-functionality of these binding sites. For example, in the mammalian system, nearly half of the identified binding sites in the mammalian system are associated with inactive genes (Wederell et al., 2008). In addition, the transcription machinery is further complicated by the existence of co-factors that are required for gene transcription. There are a couple of methods that can be used to analyze the functionality of the predicted binding sites, although no large-scale *in vivo* techniques have

been developed to address these concerns. For example, it is known that the TSS of active genes are enriched with histone H3 trimethylated at lysine 4 and enhancers are enriched with histone H3 monomethylated at lysine 4 (Barski et al., 2007; Heintzman et al., 2009). These histone modification landmarks can be accounted for when analyzing transcription factor binding sites identified from a ChIP-seq experiment. Another way to test the functionality of these binding sites is to compare the expression of candidate genes in the presence and absence of the factor of interest. However, the redundant factors that are present in the genome could interfere with the results of this experiment. A more appropriate method to test the functionality of predicted binding sites would be to generate a reporter construct for the binding site of interest. The predicted motif could be hooked up to minimal promoter upstream of a reporter gene and alteration in reporter activity would prove the functionality of the site. The best approach to test the functionality of a binding site is to directly delete or mutate the site *in vivo* and observe if the expression level of the gene is affected. Nevertheless, it is almost impossible, time-wise and cost-wise, to perform this experiment on all the binding sites identified in ChIP-seq.

Even as a relatively nascent technology, ChIP-seq demonstrates promising potentials as the new tool in understanding genome-wide gene regulatory networks. ChIP-seq offers wider coverage of the genome, better resolution and fewer artifacts compared to its predecessor, ChIP-chip. Nevertheless, the technology is not perfect, with many challenges yet to be overcome. The high cost of sequencing, immense data processing and lack of accessible platforms are the biggest barriers for most investigators. Many improvements made in the field these past few years and current work in progress will hopefully solve most of the issues associated with ChIP-seq in the near future.

Table 1 | Publicly available ChIP-seq software packages discussed in this review

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\ control}{\#\ ChIP}$	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\ control}{\#\ ChIP}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ control}{\#\ ChIP}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\#\ control}{\#\ ChIP}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	*P* value	1: Monte Carlo simulation 2: $\frac{\#\ control}{\#\ ChIP}$	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial *P* value	Estimated or user specified	Subtracted before peak calling	*q* value	1, 2: binomial 2: $\frac{\#\ control}{\#\ ChIP}$	Target FDR	No / Yes	20

(Park, 2009)

References:

Aleksic, J., and Russell, S. (2009). ChIPing away at the genome: the new frontier travel guide. Mol Biosyst *5*, 1421-1428.

Barrera, L.O., Li, Z., Smith, A.D., Arden, K.C., Cavenee, W.K., Zhang, M.Q., Green, R.D., and Ren, B. (2008). Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome Res *18*, 46-59.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Bottomly, D., Kyler, S.L., McWeeney, S.K., and Yochum, G.S. (2010). Identification of {beta}-catenin binding regions in colon cancer cells using ChIP-Seq. Nucleic Acids Res.

Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. Nat Rev Genet *10*, 605-616.

Fejes, A.P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S.J. (2008). FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics *24*, 1729-1730.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Hoffman, B.G., and Jones, S.J. (2009). Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. J Endocrinol *201*, 1-13.

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.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol *26*, 1351-1359.

Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H.*, et al.* (2006). Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res *16*, 55-65.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet *10*, 669-680.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat Methods *6*, S22-32.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods *5*, 829-834.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F.*, et al.* (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854-858.

Wederell, E.D., Bilenky, M., Cullum, R., Thiessen, N., Dagpinar, M., Delaney, A., Varhol, R., Zhao, Y., Zeng, T., Bernier, B.*, et al.* (2008). Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Res *36*, 4549-4564.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.