Concise Review: ChIP-Seq Bioassay Development Considerations

Abstract

Chromatin immunoprecipitation (ChIP) followed by sequencing is known as a ChIP-Seq assay. They are used for studying transcription factor (TF) binding and histone modifications governing genomic and epigenomic regulation of gene expression. The method has become highly productive with the development of high-throughput sequencing (HTS) platforms; however, there is the possibility of data variability and quality control (QC) issues associated with each individual step of the assay. To meet this challenge a number of technical considerations and working guidelines have been developed by the ENCODE Project and independent labs.

Background

Gene regulation involves complex combinatorial interactions between the genome and epigenome (Mohtat & Susztak, 2010; Reményi et al, 2004). At the genomic level gene expression is an intrinsic function of DNA sequences that either promote or repress expression of the gene product through DNA binding interactions with protein TFs such as TBPs, including protein-protein interactions involving homo- and hetero-oligomerization that also affect the rate of transcription (Reményi et al, 2004). At the epigenomic level, gene expression is further controlled by DNA methylation and histone modifications such as acetylation, methylation, and phosphorylation, among others that affect chromatin structure and access of the TFs and transcriptional machinery to the DNA (Mohtat & Susztak, 2010). H3K27me3 & H3K9me3, associated with regions of repressed transcription, and H3K4me2 & H3K36me3, associated with regions of active transcription, are commonly assayed and often interact in bivalent domains, for example. This complex milieu of TFs and epigenetic modifications are specific for each cell and tissue type and are differentially regulated throughout development and by environmental factors such as mutagens leading to diseases like cancer. Therefore, in order to understand the complex development of organisms and diseases it is essential to understand factors governing gene expression. This insight has been made available through recent developments in HTS methods capable of capturing the precise DNA binding sites of TFs and exact modification of histones at the base level, genome-wide. These methods include ChIP-ChIP using microarrays or a more HT technique known as ChIP-Seq. In general, ChIP-Seq machines provide an advantage over the inherent noise associated with hybridization in microarray ChIP-ChIP, increasing peak detection and resolution (Ho et al, 2011). Though there are cost and performance considerations among different vendors of this technology such as between Applied Biosystems SOLiD versus Illumina's hiSeq systems, this will not be covered in this review. While ENCODE has developed working standards for the most critical aspects of ChIP-Seq such as antibody (Ab) validation, sequencing depth, and data reproducibility & reporting (Landt et al, 2012;

supplementary fig 1), other aspects affecting data quality in a ChIP-Seq assay include cell/tissue preparations, crosslinking, nuclei isolation, fragmentation, and library preparation (Kidder et al, 2011).

ChIP-Seq Assay Lab Protocols

ChIP-Seq lab protocols include the isolation of cells, fixation of chromatin modifying proteins through crosslinking, nuclear isolation, fragmentation, immunoprecipitation (IP), library preparation, and sequencing. The datasets generated are then subjected to QC analysis and computational interpretation. Methods using the HT EZ-Magna ChIP™ HT96 kit (EMD Millipore, 2012) for use with a 96-well microplate followed by sequencing on a hiSeq machine are summarized below.

Cell preparation

Cells or tissues can be used in a ChIP assay. If using a cell culture one should use at least 1x10^5 cells / ChIP rxn at ~%80-90% confluency in growth media (EMD Millipore, 2012). Alternately, one can use fresh tissue containing about 1x10^7 cells for the complete assay. A 10 minute incubation with formaldehyde is the most common fixation method and will induce the formation of covalent crosslinks between the TF and histone proteins and the DNA, capturing the state of the DNA-protein interactions at the time. 10X glycine is used to quench excess formaldehyde. One then removes the medium, homogenizes the sample if using tissue, washes the cells with PBS that also includes protease inhibitors which prevent TF/histone protein degradation, and then pellets them. These cells are then lysed by resuspension in and incubated with a nuclei isolation buffer which also contains protease inhibitors, followed by pelleting and removal of debris. Finally, one needs to shear the DNA by sonication or through MNase treatment if assaying histones (Kidder et al, 2011), optimizing the conditions so that it yields ~150 – 1000 bp fragments. Cell lysates should be kept ice cold to preserve the DNA-protein crosslinks (EMD Millipore, 2012).

IP

IP collects only the DNA-protein binding complexes of interest. This protocol is optimized for use with a robotic workstation and automated liquid handling (EMD Millipore, 2012). This is the most sensitive step and one should make sure to add protease inhibitors to all buffers used. One prepares and places in the workstation an antibody plate, Magna ChIP A/G Magnetic Beads in a tube, and a ChIP plate on top of a magnetic separator. One then adds an empirically determined amount of the selected antibodies that will bind the desired TF or histone proteins, control Ab such as IgG in control wells, and beads to the ChIP plate. Next, one adds the fragmented chromatin to the ChIP plate and this is then left to react overnight at 4 °C. The following morning one then adds specific ChIP and washing buffers, spins down the plate, removes the supernatants, and then transfers the beads containing the attached ChIP DNA to a thermal plate. This thermal plate is then incubated in a Thermomixer at 65 °C for 2 hours in a ChIP elution buffer also containing proteinase K, followed by further incubation at 95 °C for 15 minutes. These heating steps with proteinase K serve to digest the chromatin and reverse the DNA-protein crosslinks. The IP solution is then placed on a magnetic separator and the supernatant containing the IP DNA is collected without the beads. DNA purification is then performed by incubating the recovered DNA with a Agencourt AMPure XP DNA purification beads on a Magna GrIP Rack in order to detect low concentration immuno-precipitates. After washing and eluting one then needs to verify ChIP DNA enrichment using qRT-PCR to compare the IP DNA to a standard IP.

Library Construction

During library construction it is necessary to perform end repair of the DNA by forming phosphorylated blunt ends using T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase in a thermal cycler (Illumina, 2007). One then adds 'A' bases to the 3' blunt end of the DNA using dATPs and the Klenow 3' to 5' exo fragment so that there is one hybridization spot available for the 'T' base overhang on the adapters. One then ligates adapter sequences to both ends. In future sequencing steps, these adapters will in turn hybridize with surface-bound DNA on a flow cell in a clustering machine which captures and amplifies the recovered DNA. For now, this DNA is gel purified in order to remove excess adapters and/or select DNA templates of certain sizes, followed by a round of enrichment of gel-extracted DNA by PCR. Finally, validation of the library should be performed in order to assay the size, purity, and concentration of the DNA. The IP DNA library is now ready for sequencing.

HTS

This step sequences the TFs and/or histone DNA binding sites library previously generated and can be accomplished in a rapid run mode or high-output run mode depending on the desired coverage, read length, sample sizes, and yield desired (Illumina, 2012). If using a hiSeq 2500, for example, one could generate 600 Gb of data in ~10.5 days in high-throughput mode or 120 Gb of data (1 human genome) in ~1 day in rapid-run mode. If using rapid-run mode one would perform the clustering normally done separately on-board the machine and would require SBS kits. Otherwise, one only needs to load the DNA generated during library construction into the template loading station along with the flow cells. All other steps are automated and on-board software controlled. After the sequencing run the base call (bcl) files or FASTQ data can be submitted to the NCBI in SRA format and/or analyzed with off-instrument software such as Model-based Analysis of ChIP-Seq (MACS), PeakSeq, or other algorithms.

BioAssay Development Considerations

ENCODE and modENCODE developed a set of standards to address antibody validation, sequencing parameters, replication, and data quality and reporting in response to the variability in experimental protocols, data quality, and reporting found in ChIP-Seq assays (Landt et al, 2012). Additionally, unofficial standards continue to be developed by vendors and independent labs.

Antibodies

The primary factor in limiting noise and detecting relevant DNA-protein interactions in a ChIP experiment is governed by antibody specificity and quality (Landt et al, 2012). ENCODE has detailed both primary and secondary modes of characterization that must be performed for new antibodies or antibody lots. Primary characterization involves immunoblot or immunoflorescence analyses that show the antibody binds the specified TF or does not bind unmodified or non-histone proteins; however, secondary characterization must be performed using knockdown (siRNA, shRNA, etc) of the target protein or mass spectrometry analyses of the IP TF or histone to verify that the antibody binds only one TF or one type of modified histone specifically (Landt et al, 2012; supplementary fig 2). Commercial suppliers which have performed these characterizations can market 'ChIP grade' Ab; Millipore markets these highly validated Ab as ChIP Ab+, for example (EMD Millipore, 2012). Independent testing of commercially produced Ab is still recommended, however, as research by Egelhofer et al (2011) discovered that over 20% of commercially produced specific Abs failed IP of histone protein QC tests.

Experimental considerations in Ab usage include the empirical determination of the amount of Ab and a dilution series may need to be used if not enough or too much ChIP DNA is recovered (EMD Millipore, 2012). Insufficient recognition of the target protein may be a result of long fixation times or overuse of formaldehyde. Incubation times and conditions can also be variable though temperatures above 65 °C can inactivate proteinase K, resulting in incomplete chromatin digestion and elution. Excessive crosslinking may also irreversibly affect elution. In some cases there may not be an Ab that binds a specific protein and an epitope-tagged protein may need to be used, increasing the potential for perturbation of real expression levels. An IgG Ab that binds non-target, non-nuclear antigens should be used in a control IP (Landt et al, 2012).

Cell Preparations

While 1x10^5 cells may be sufficient for most proteins less abundant proteins may require more cells in order to increase signal to noise (S/N) ratios. One needs to retain enough cells for the two replicates required by ENCODE and controls should be performed for every cell line, developmental stage, culture conditions, and treatments due to possible ploidy or epigenomic alterations that arise, for example (Landt et al, 2012). Sonication conditions and size selection must also be determined empirically as this can vary from lab to lab and can result in non-uniform fragmentation as open chromatin conformations can lead to increased shearing while closed conformations can lead to decreased shearing (Kidder et al, 2011). MNase should be used for ChIP involving histone analyses and overdigestion of chromatin must be avoided to ensure high recovery of the signal. Additionally, while sonication buffers like SDS may improve access of Ab to TFs tightly bound they may also disrupt epigenetic interactions of interest (Landt et al, 2012; Kidder et al, 2011).

Experimental considerations in cell preparation include not enough or too much crosslinking, insufficient cell lysis, not enough washing or aspiration of beads while washing, and improper chromatin shearing such as from sonication but also from protein denaturation - it is important to keep samples on ice during sonication (EMD Millipore, 2012).

Library construction

Relatively small amounts of DNA isolation can result in low complexity leading to reduced site discovery and reproducibility; increasing the read depth may not increase complexity (Landt et al, 2012). Newer technologies such as NuGEN's Mondrian SP digital microfluidics system can increase S/N while reducing sample prep time, pipetting, washing steps, and reagents with automated library construction on a chip, for example (NuGEN Technologies Inc, 2012).

General PCR considerations include variable temperature and amplification conditions, bad primers, or a variable number of cycles that results in too little or overamplification of templates, skewing the results; it is important to use the same number of cycles for the control experiment, also (EMD Millipore, 2012). Illumina recommends ~18 cycles using their procedure, for example (Illumina, 2007). Additionally, when running a gel one will want to purify samples using different gels to reduce the potential for cross-contamination; when visualizing one will want to use a Dark Reader and not UV in order to limit sample degradation.

HTS

The ENCODE minimum recommended number of reads is 20x10^6 and this is usually sufficient to deep sequence most protein targets, though this number can vary depending on the number of genome-wide target sites. Counts tend to increase with more reads, saturating at around 100 million reads for some experiments (Landt et al, 2012). A major concern is that these counts cannot really be ascertained *a priori* and is it therefore critical to obtain the best S/N during IP and deep sequencing. The recommended read length is 36 base pair though this can be increased, increasing the data output but also increasing sequencing time (Roadmap

Epigenomics Project, 2010). Quality scores are also a significant metric of sequencing accuracy; the hiSeq 2500 claims >90% of the reads to contain less than 1 base error in 1000, 99.9% accuracy (Q30), for example (Illumina, 2011; Illumina, 2012).

Data considerations

Data management is a major concern in developing a complete ChIP-Seq experiment as data output for single runs can easily take-up terabytes, increasing storage-associated costs (Park, 2009). Large file transfer and long term public data storage solutions have been met with the NCBI sequence read archive (SRA) which houses raw sequence data and facilitates data sharing, for example. However, the major bottleneck in ChIP-Seq and NGS experiments is data analysis, with large datasets requiring parallel processing and efficient algorithms for genome alignment and peak detection. Software such as MACS can handle analysis of the data efficiently - balancing speed, accuracy, and computational costs (Zhang et al, 2008). For example, ChIP-Seq only determines the 5' and 3' ends of the DNA fragment that the target proteins were bound to; algorithms such as MACS must be used to determine the precise binding sites - capable of numbering > 100,000 genome wide! Moreover, QC analysis of the experiment cannot be performed until the data has been computationally processed. One measure of ChIP enrichment is the fraction of reads in enriched, peak regions (FRiP); ENCODE recommends a FRiP enrichment of 1% or more when using MACS peak detection, for example, and this is generally a good indication of a quality IP (Landt et al, 2012).

Conclusion

ChIP-Seq assays offer great insight into genome-wide, base level TF and epigenetic gene regulation. There are many variables such as cell preparations, Ab quality, library construction, HTS parameters, and data analysis & management methods that can affect the utility of the data, however. Recently, progress has been made by ENCODE and independent labs in standardizing each step of the assay, not only reducing noise but also improving data quality and reproducibility. This review suggests Ab quality still requires improvement and that the number of steps in associated heterogenous assays needs to be reduced and refined in order to further increase the S/N and results integrity.

References

Egelhofer et al (2011) An assessment of histone-modification antibody quality. *Nat Struct Mol Biol*. 18(1): 91–93.

EMD Millipore (2012) *High Throughput Chromatin Immunoprecipitation Kit.* [online] Available at:< http://www.millipore.com/userguides.nsf/a73664f9f981af8c852569b9005b4eee/27646f6bee09bf63852579de007050c1/$FILE/MAGNACHIPHT96MAN.pdf > [Accessed 9 Dec 2012]

Ho et al (2011) ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics*. 12:134

Illumina (2007) *ChIP Sequencing Sample Prep Guide (11257047 A).* [online] Available at:< https://icom.illumina.com/Download/Summary/Yy2liOS-nEWyTUnT5w9UMA > [Accessed 9 Dec 2012]

Illumina (2011) *Quality Scores for NGS*. [online] Available at: < http://www.illumina.com/Documents/%5Cproducts%5Ctechnotes%5Ctechnote_Q-Scores.pdf > [Accessed 9 Dec 2012]

Illumina (2012) *HiSeq 2500 Application Note*. [online] Available at:< http://www.illumina.com/Documents/products/appnotes/appnote_hiseq2500.pdf >[Accessed 9 Dec 2012]

Kidder et al (2011) ChIP-Seq: technical considerations for obtaining high-quality data. *Nat Immunol*. 12(10):918-22.

Landt et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22(9):1813-31.

Mohtat & Susztak (2010) Fine Tuning Gene Expression: The Epigenome. *Semin Nephrol*. 30(5): 468–476.

NuGEN Technologies Inc (2012) *Using Digital Microfluidics and the Mondrian™ SP System for Automated, Cost-Effective and Reproducible NGS Library Preparation.* [online] Available at:< http://www.nugeninc.com/nugen/?LinkServID=EEF626E6-6AAC-4BE7-817CA1D117D263BB > [Accessed 9 Dec 2012]

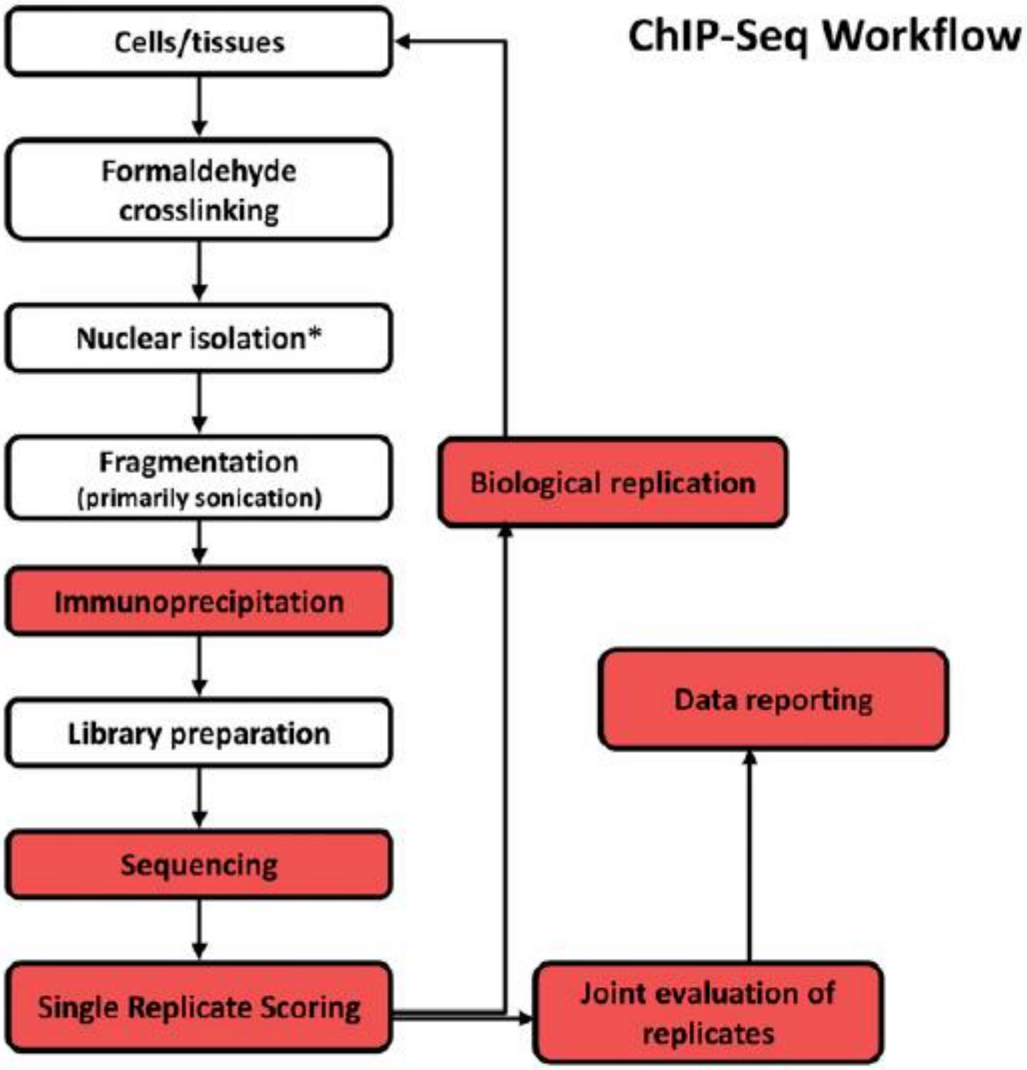
Park (2009) ChIP-Seq: advantages and challenges of a maturing technology. *Nat Rev Genet.* 10(10): 669–680.

Reményi et al (2004) Combinatorial control of gene expression. *Nat Struct Mol Biol*. 11(9):812-5.

Roadmap Epigenomics Project (2010) *REMC Standards and Guidelines for ChIP-seq, V1.0.* [online] Availablet at:< http://www.roadmapepigenomics.org/files/protocols/data/histone-modification/REMC_ChIP-seqStandardsFINAL.pdf >[Accessed 9 Dec 2012]

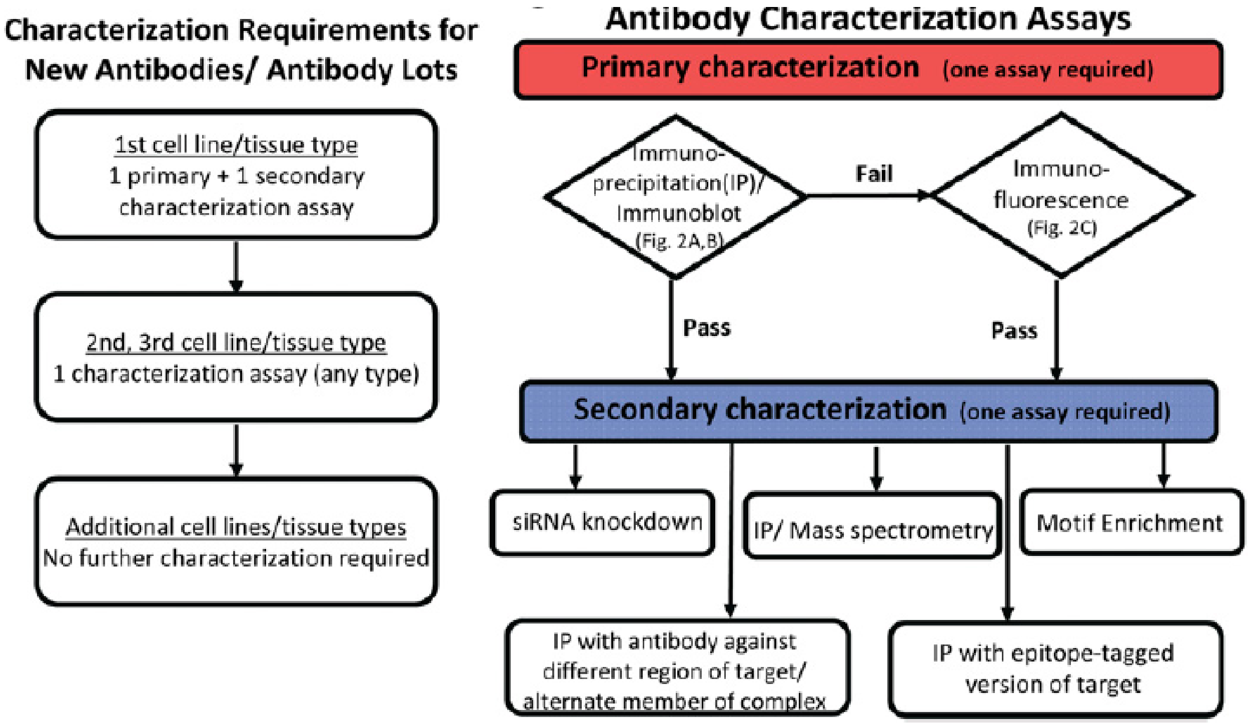
Zhang et al (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*. 9:R137

Supplementary Figure 1 (Landt et al, 2012).



ENCODE has developed standard guidelines in critical ChIP-Seq assay steps (red).

Supplementary Figure 2 (Landt et al, 2012).


Characterization Requirements for New Antibodies/ Antibody Lots
1st cell line/tissue type
1 primary + 1 secondary characterization assay
2nd, 3rd cell line/tissue type
1 characterization assay (any type)
Additional cell lines/tissue types
No further characterization required
Antibody Characterization Assays
Primary characterization (one assay required)
Immuno-precipitation(IP)/ Immunoblot (Fig. 2A,B)
Fail
Immuno-fluorescence (Fig. 2C)
Pass
Pass
Secondary characterization (one assay required)
siRNA knockdown
IP/ Mass spectrometry
Motif Enrichment
IP with antibody against different region of target/ alternate member of complex
IP with epitope-tagged version of target

ENCODE standard guidelines for Ab characterization and associated assays.