Optimize DNA Library Preparation of ChIP-DNA for Next-Generation Sequencing

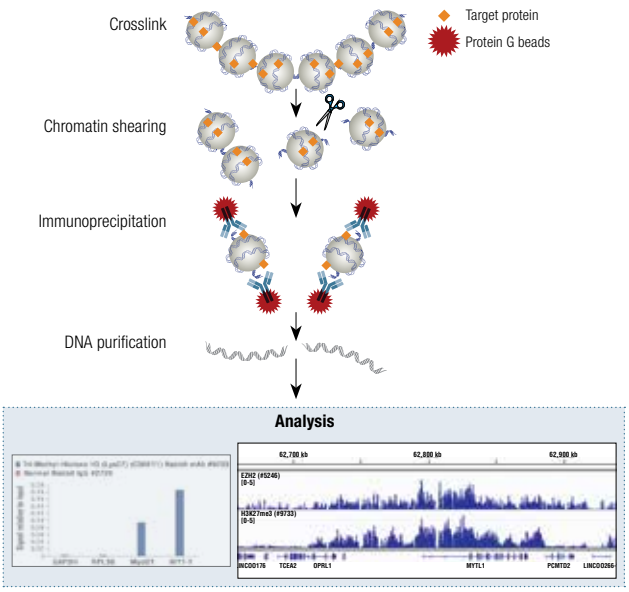
Introduction

ChIP-sequencing (ChIP-seq) is a powerful technique that examines protein-DNA interactions across a genome (1). It couples chromatin immunoprecipitation (ChIP) with next-generation sequencing (NGS) to identify how histones, transcription factors, and cofactors interact with DNA (2,3,4). ChIP-seq experiments typically begin with the formaldehyde cross-linking of protein-DNA complexes in cells or tissue. The chromatin is then extracted and fragmented, either through enzymatic digestion or sonication, and DNA-protein fragments are immunoprecipitated with target-specific antibodies.

Generating reliable ChIP-seq data depends on using antibodies that have been validated for target specificity and acceptable signal-to-noise ratios to perform the ChIP experiment. Cell Signaling Technology® (CST®) has an extensive list of ChIP-seq validated antibodies as well as ChIP-specific protocols and resources to help optimize your ChIP experiments. The amount of DNA in each ChIP sample can be quickly determined using a PicoGreen™-based double-stranded DNA detection method.

Following the ChIP experiment, qPCR is performed to confirm the immuno-enrichment of a few known target genes and to validate the quality of the ChIP-DNA. Once the enrichment and quality of the ChIP DNA have been confirmed, a DNA library is constructed using the ChIP-DNA. Individual DNA libraries are then pooled into a mixture of libraries to be sequenced together on one lane of an NGS platform to identify protein-DNA binding sites. In this application note, we'll walk you through some of the parameters to consider when preparing your DNA library, evaluating library quality, and preparing samples for NGS. We'll also review how to quickly assess ChIP-seq data to increase your confidence in the data before proceeding with more in-depth, extensive analysis.

CST SimpleChIP® Chromatin IP Protocol



Determining the Optimal Amount of DNA to Use in Your Library Preparation

The amount of ChIP-DNA to use when creating a DNA library is influenced by factors like the amount of DNA obtained from the actual ChIP, the desired library yield, and the limits on PCR amplification required to minimize duplicate sequencing reads. In general, using more ChIP-DNA for library construction is preferred, as it will require fewer PCR amplification cycles and will improve sequencing library diversity.

A typical histone ChIP experiment using 10 µg of input chromatin DNA per immunoprecipitation yields approximately 100 to 1000 ng of ChIP-DNA. In comparison, a transcription factor or cofactor ChIP experiment yields approximately 5 to 25 ng of ChIP-DNA. CST® recommends using 50 ng of histone ChIP-DNA and 5 ng of transcription factor and cofactor ChIP-DNA for library construction.

PCR protocols should be adjusted based on the amount of DNA used for the sequencing library preparation. We recommend using 6 PCR amplification cycles when starting with 50 ng of ChIP-DNA and 10 amplification cycles when starting with 5 ng of ChIP-DNA (**Table 1** and **Figure 1**). The DNA library yield drops when using less than 1 ng of ChIP-DNA. However, in situations where ChIP-DNA is limited, library DNA can be generated with as little as 0.5 ng of ChIP-DNA and 14 cycles of PCR amplification. The number of identified peaks appears to be independent of the amount of starting material (**Table 1** and **Figure 1**). Just remember that more PCR amplification cycles can lead to lower library diversity and a higher duplication rate of reads (**Table 1**).

Each DNA library is generated with a unique barcode and individual library samples are combined in a mixture of libraries to be sequenced together. The barcode is used to de-convolute data from the pooled sample in order to obtain sequences for each library. The number of sequence reads for a given sample DNA library depends on the amount of the given sample library DNA added to the pooled mixture of DNA libraries that is sent to NGS. If the concentration of a given sample DNA library is lower than others, we suggest adding a larger volume of that particular sample DNA library to the pooled mixture in order to obtain similar reads as other samples in the same sequencing library pool.

Optimize DNA Library Preparation of ChIP-DNA for Next-Generation Sequencing

Performance of DNA Library Kit and Oligos on Different Starting Amounts of ChIP-DNA (Re1)

Antibody	Library yield (ng)	Total Reads	Reads mapped to genome (%)	Duplication Rate (%)	Identified peaks
TCF4 #2569 (50 ng)	140.8	39.82	97.80	5.93	9425
TCF4 #2569 (5 ng)	187.8	34.37	97.77	6.16	10506
TCF4 #2569 (0.5 ng)	67.3	33.71	96.02	20.55	9416
H3K4me3 #9733 (50 ng)	109.5	24.40	98.46	10.36	26903
H3K4me3 #9733 (5 ng)	154.2	26.90	98.30	11.86	26365
H3K4me3 #9733 (0.5 ng)	50.1	26.16	94.65	26.55	36254
Input (50 ng)	112.2	26.78	97.94	4.73	n/a

Table 1: Comparison of DNA sequencing libraries prepared from different starting amounts of ChIP-DNA generated using the SimpleChIP® Plus Enzymatic Chromatin IP Kit (Magnetic Beads) #9005. ChIP was performed using 4 x 10^6 HCT 116 cells and either TCF4/TCF7L2 (C48H11) Rabbit mAb #2569 or Tri-Methyl-Histone H3 (Lys4) (C42D8) Rabbit mAb #9751. Libraries were generated using the SimpleChIP® ChIP-seq DNA Library Prep Kit for Illumina® #56795 and the dual index primers provided in SimpleChIP® ChIP-seq Multiplex Oligos for Illumina® (Dual Index Primers) #47538, pooled into one sample, and sequenced on an Illumina® Next-Seq platform. **As shown, the number of identified peaks is independent of the starting amounts of ChIP-DNA, but yield was lower and duplication rate was higher for libraries generated using only 0.5 ng of starting material.**

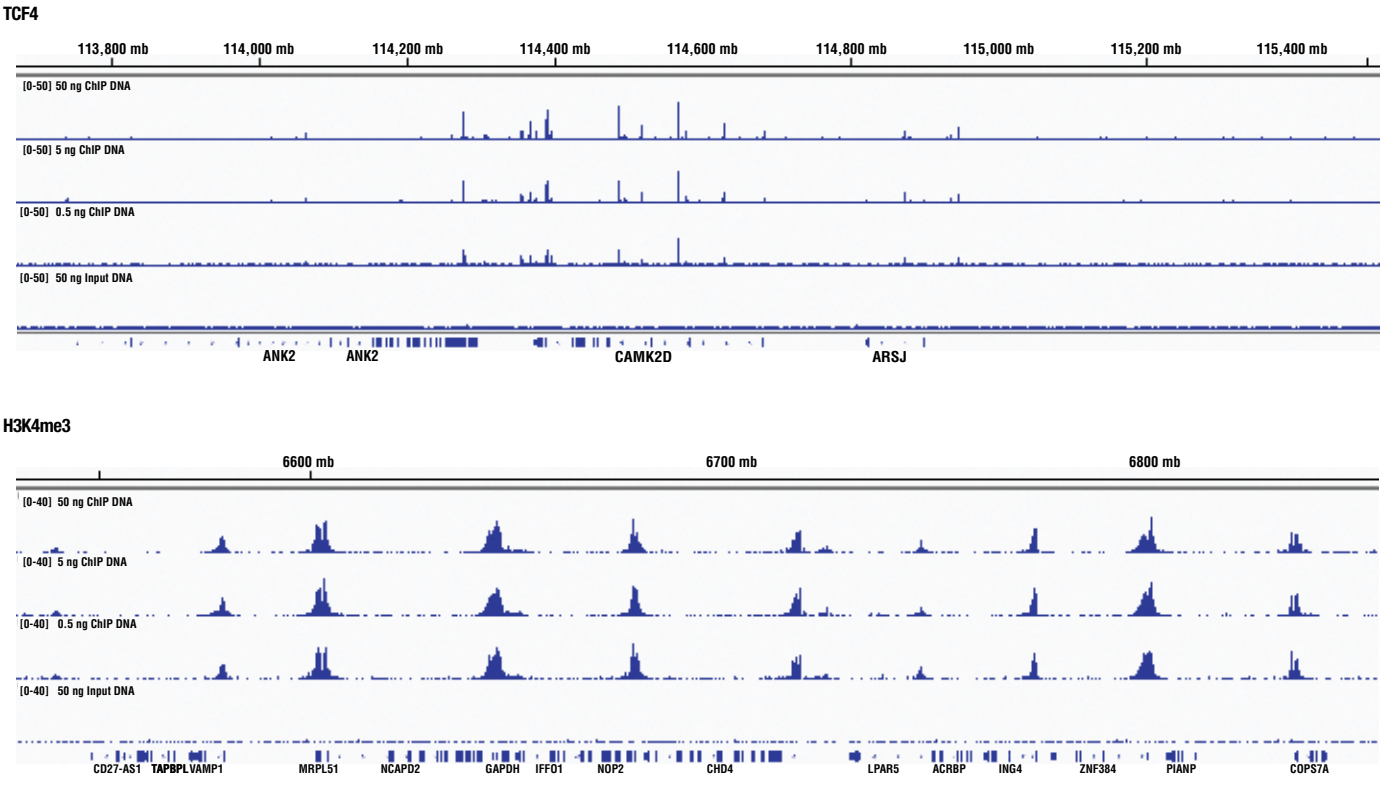


Figure 1: ChIP-seq data generated using TCF4/TCF7L2 (C48H11) Rabbit mAb #2569 (upper) and Tri-Methyl-Histone H3 (Lys 4) (C42D8) Rabbit mAb #9751 (lower) using different starting amounts of ChIP-DNA, as described in **Table 1**. The figure shows binding with CaMK2D and GAPDH, known targets of TCF4/TCF7L2 and H3K4me3, respectively. **As shown, the peak localizations and heights are similar for all three starting amounts of ChIP-DNA, indicating ChIP-seq data are independent of the amount of starting material as long as the quality of starting ChIP-DNA is good and PCR conditions used to construct the DNA library are optimized.**

Optimize DNA Library Preparation of ChIP-DNA for Next-Generation Sequencing

Confirming DNA Library Quality

You'll want to confirm DNA library quality by assessing the library yield and fragment size before sequencing the samples. The DNA library yield can help you evaluate whether PCR conditions used to generate the library are optimal. The PCR recommendations described above using 50 or 5 ng of ChIP-DNA typically generate approximately 30 ng/µL of library DNA (900 ng total DNA). Starting with less than 1 ng of ChIP-DNA typically generates approximately 10 ng/µL of DNA (300 ng total DNA). If your PCR protocol generates a library with significantly higher yields, it is an indication that you may want to consider re-making the DNA library with fewer PCR amplification cycles in order to avoid amplification preferences and poor library diversity.

To assess fragment size, use a system like Agilent's Bioanalyzer to evaluate size profiles for enzymatic and sonication ChIP-DNA library preparations (**Figure 2A**). You should see several peaks between 290 and 740 bp when you run an enzymatic ChIP-DNA library on the Bioanalyzer. This ladder of bands is consistent with the nucleosomal DNA fragments generated with enzymatically digested chromatin fragmentation. Sonicated ChIP-DNA libraries will appear as a single broad peak on the Bioanalyzer, since sonication shears the DNA between and within the nucleosome (**Figure 2B**).

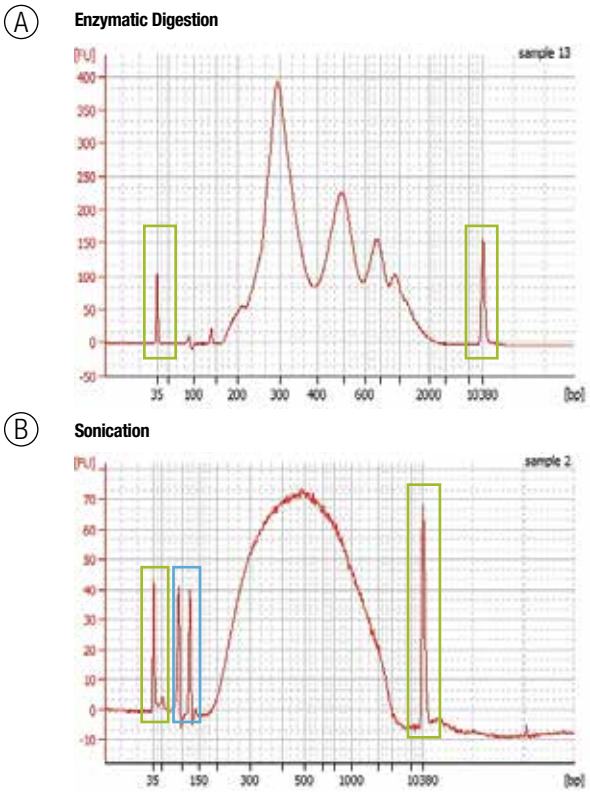


Figure 2: Bioanalyzer data for enzymatic digestion (A) and sonication (B) ChIP-seq library DNA data displayed as electropherograms. Enzymatic ChIP-DNA libraries will appear as a ladder of bands between 290 and 740 bp, while sonicated ChIP-DNA libraries will appear as a single broad peak. Two sharp peaks at 70 and/or 140 bp in the sonication DNA data (blue circle) indicates an excess of adaptors and/or adaptor dimers that can be removed by repeating library cleanup. Molecular weight markers (green circles) should bracket the ends of the electropherogram so as to not interfere with the DNA fragment analysis.

An excess of adaptors during the library preparation process can result in adaptor contamination in your DNA library. Adaptor and/or adaptor dimer contamination in your library can be detected through the presence of two sharp peaks around 70 and 140 bp (Figure 2B). We recommend repeating the library cleanup procedures (SimpleChIP® ChIP-seq DNA Library Prep Kit for Illumina® #56795 protocol section V) to remove the adaptor and adaptor dimers. Otherwise, you may observe decreased sequencing depth in your NGS data, since the adaptors may occupy too many reads due to a sequencing bias for smaller fragments.

Some Sequencing Service Centers like to perform size selection on the DNA library to remove DNA fragments larger than 1 kb that NGS cannot sequence. However, this is not necessary. We suggest not performing size selection because the presence of large fragments will not interfere with the sequencing of the small fragments in the sample, and the size-selection process can lead to additional loss of the smaller fragments, resulting in decreased diversity of the sequencing library. If your sample contains a lot of large fragments, we suggest normalizing the library concentration based on the percentage of small fragments in your sample in order to obtain optimal cluster density and sequencing depth for your sample.

Preparing Samples for Sequencing Reactions

DNA libraries are typically pooled into one mixture that is applied to a single sequencing lane on an NGS platform. Sample libraries should be pooled in a way that will ensure each library will obtain similar reads. To do so, dilute each library DNA sample to the same concentration, around 2 to 10 nM, and then pool libraries together as one sample for one NGS reaction. Consult your Sequencing Service Center for required concentration and volumes for pooled library samples.

After each library sample is normalized to the same concentration as described above, adjust the volume of each sample library in the mixture to get the desired sequencing depth for each sample library. For example, if you want 45 million reads for ChIP sample 1 and 15 million reads for ChIP sample 2, combine 3 volumes of sample 1 library DNA with 1 volume of sample 2 library DNA in the pooled mixture.

ChIP-seq data generate two different binding patterns: narrow peaks covering a small region of the gene and broad peaks covering several genes. Transcription factors typically generate narrow binding patterns, while binding patterns for cofactors vary between narrow binding and broad binding. Histone modifications also vary, as some, like H3K4me3, have narrow binding patterns and others, like H2K27me3, have broad binding patterns. The ENCODE (Encyclopedia of DNA Elements) consortia guidelines recommend generating 40 to 50 million reads for histone modification ChIP samples with broad binding patterns and 15 to 20 million reads for transcription factors, cofactor ChIP samples, and histone modification ChIP samples with narrow binding patterns.

Optimize DNA Library Preparation of ChIP-DNA for Next-Generation Sequencing

The number of libraries you combine will depend on the number of library samples you have, the desired number of reads for each library sample, and the total capacity of the sequencing platform that will be used. For example, 3 histone modification samples (40 million reads per sample) and 7 transcription or cofactor samples (15 million reads per sample) will require 225 million reads. Therefore, you'll want to sequence the samples on, for example, an Illumina HiSeq 2500 system that can provide 250 million reads per lane compared to an Illumina Next-Seq 500 system that provides 400 million reads per lane. This covers the 225 million theoretical reads required and leaves 25 million reads to cover any unexpected reads. Combining too many samples will cause each sample to have less sequencing depth, while combining too few samples may possibly waste your sequencing resources.

You'll also want to prepare a negative control for your ChIP-seq experiment. We recommend using input ChIP-DNA, since it's the exact same chromatin that you used for the immunoprecipitation and, for this reason, is the most frequently used negative control for ChIP-seq experiments. Input ChIP-DNA will produce a complex sequencing library, and peak calling algorithms are typically written with the assumption that input ChIP-DNA will be used as the negative control. Using IgG ChIP-DNA as a negative control is not as ideal as it is more labor-intensive to produce, it typically doesn't generate enough DNA for library preparation, and the DNA purified is biased due to the limited genomic regions recognized by the IgG. This bias will be further amplified by PCR during library construction.

Evaluating ChIP-seq Data

Raw NGS data are typically reported as FASTQ files. Once you have received this file, you'll want to quickly check the quality and quantity of reads for each sample to confirm confidence in your data before proceeding with more in-depth analysis. You will also need to map good-quality reads back to the destination genome using software like the Bowtie2 program, since the mapping rate will indicate whether there was contamination from other species in your DNA library.

In order to visualize the data, you will want to use software like Integrative Genomics Viewer (IGV) or upload your data to the UCSC Genome Browser to see binding peaks at any genomic locus. As described above, transcription factors typically generate narrow binding patterns, while binding patterns for cofactors and different histone modifications can vary between narrow binding and broad binding.

When evaluating narrow binding patterns, you'll want to determine the number of identified binding peaks as well as the signal-to-noise ratio across the whole genome. A sequencing depth of 15 million reads should yield roughly 1000 peaks for transcription factors and 5000 peaks for the histone modifications known to generate narrow peaks. The average signal to noise across the genome should be at least 3- to 4-fold. For transcription factors, you also may want to use the MEME online tool in order to analyze whether its known binding motif sequence is enriched in the ChIP-DNA.

Broad binding pattern proteins should be evaluated based on the genomic signal-to-noise ratio, since peak number is not a reliable way to assess data quality. The average signal to noise across the genome should be around 6- to 8-fold.

If possible, it is ideal to validate your ChIP-seq data using a knock-out cell line or tissue to show that enriched peaks depend on the protein of interest. However, it is not always possible to include a knock-out negative control. Instead, you can utilize antigenically distinct antibodies to your target protein or antibodies against different subunits of your target complex and compare the ChIP-seq data for each. Antibodies against the same target protein or different subunits of a protein complex should generate overlapping genome enrichments. Finally, you can further validate your ChIP-seq data by ChIP-qPCR using primers specific to your genomic regions of interest.

Conclusion

You need to optimize your ChIP experiment, DNA library construction, and NGS sample preparation in order to generate high-quality ChIP-seq data. CST provides multiple resources and many ChIP-seq validated antibodies to help you achieve ChIP-seq success! This document discusses important experimental parameters that help you to maximize the diversity and number of reads in your DNA library preparation and assess the quality of your ChIP-seq data. With this knowledge, you'll be able to experience all that ChIP-seq has to offer and determine how various proteins and histone modifications act to regulate gene expression and impact a variety of biological processes and diseased states.

References

1. Pellegrini, M. and Ferrari, R. (2012) *Methods Mol Biol* 802:377-387.
2. Jiang, S. and Mortazavi, A. (2018) *Brief Func Genomics* 17(2):104–115.
3. Orlando, V. (2000) *Trends Biochem Sci* 25(3):99-104.
4. Kuo, M.H. and Allis, C.D. (1999) *Methods* 19:425-433.

Technical Support

At CST, providing exceptional customer service and technical support are top priorities. Our scientists work at the bench daily to produce and validate our antibodies, so they have hands-on experience and in-depth knowledge of each antibody's performance. In the process, these same scientists generate valuable reference information that they use to answer your questions and help troubleshoot your experiment by phone or email.

For questions about how to customize your protocol, please contact technical support by emailing support@cellsignal.com, visiting **www.cellsignal.com/support**, or calling 1-877-678-8324.

Ordering Information

www.cellsignal.com/orderinfo

For a complete list of CST offices and distributors, please visit **www.cellsignal.com/contactus**