Chromatin Immunoprecipitation-based Sequencing (ChIP-Seq) on the SOLiD™ System

Introduction

Chromatin immunoprecipitation (ChIP) is a technique for identifying and characterizing elements in protein-DNA interactions involved in gene regulation or chromatin organization. Historically, ChIP reactions were analyzed by PCR, sequencing and more recently by microarray technologies also known as ChIP-on-chip. Microarray platforms provide a method for "global" ChIP analysis but direct sequencing of enriched fragments has proven more effective in determining locations of DNA-binding proteins along the genome in an unbiased manner. The massively parallel sequencing capacity, high accuracy and flexibility of the SOLiD™ System make it well suited for ChIP-Seq applications.

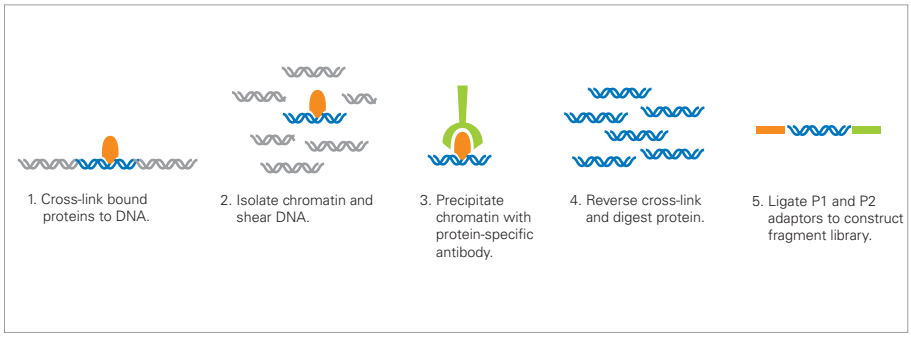

Figure 1: DNA enrichment by chromatin immunoprecipitation(ChIP) and SOLiD™ fragment library construction.

SOLiD System ChIP-Seq Analysis

The SOLiD System's ability to generate over 400 million sequence tags (35 bp sequence reads) in a single run enables whole genome ChIP analysis of complex organisms. Sequence tags are mapped to a reference sequence and counted, to identify specific regions of protein binding. The ultra high throughput of the system provides researchers with the sensitivity and the statistical resolving powers required to map and accurately characterize the protein-DNA interactions of an entire genome. Additionally, the flexible slide format allows researchers to analyze multiple experimental samples and a control sample in a single run.

ChIP analysis with the SOLiD System (Figure 1) begins with a traditional chromatin immunoprecipitation procedure. DNA is cross-linked *in vivo* to DNA-binding proteins with formaldehyde and mechanically sheared using sonication. The DNA-protein complex is then precipitated with an antibody that is specific to the DNA-binding protein. The quality of this antibody is critical to the success of ChIP-Seq protocols, as it determines the level of enrichment over background that is obtained. The DNA is released by reversing the cross-link to the protein and the protein is digested. The size and concentration of the resulting ChIP DNA fragments determines the approach that is taken to process this sample for SOLiD fragment library construction and subsequent sequencing. (Figure 2, page 2)

Typically, DNA derived from the ChIP procedure can range from 100 bp to 2 kb in size and is often limiting in quantity. Therefore, modifications to the standard SOLiD System 2.0 Fragment Library Preparation: Lower Input DNA protocol is used to create the ChIP-Seq library.

Preparation of a negative control consisting of non-immunoprecipitated fragmented DNA of similar size range

or chromatin-immunoprecipitated DNA using unspecific IgG antiserum is suggested to detect differential enrichment. Once these SOLiD ChIP-Seq and negative control libraries are created, the samples are sequenced on the SOLiD™ System. The short sequence reads from the SOLiD System are mapped against genomic sequences, using the SOLiD System alignment tools available through the Applied Biosystems Software Development Community (http://info.appliedbiosystems.com/solidsoftwarecommunity) or third-party tools compatible with color space. Data can then be visualized with a tool such as the University of California, Santa Cruz Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) to identify and quantify the regions of sequence that bind to the protein of interest.

SOLiD System ChIP-Seq Analysis of FOXA3 protein

In collaboration with the laboratory of Dr. Claes Wadelius at Uppsala University, ChIP-Seq analysis was performed using ChIP DNA isolated from hepatic cell lines to identify loci involved in interactions with the FOXA3/HNF3γ protein. This hepatocyte nuclear factor, a member of the forkhead class of DNA-binding proteins, activates transcription for liver-specific genes such as albumin and APOA2 and also interacts with DNA and histones as a pioneer factor that opens chromatin during development.

Materials and Methods:

ChIP DNA was prepared as previously described (Rada-Iglesias et al, 2005) from HepG2 cells using a commercially-available antibody (Santa Cruz Biotechnology; sc-5361) specific to the FOXA3/HNF3γ protein. Immunoprecipitated DNA (0.5 µg) was subjected to size selection and purification using gel electrophoresis to isolate DNA fragments of 3 different size ranges: 150-200 bp, 200-250 bp, and 250-300 bp. Sheared, non-immunoprecipitated hepatic cell line DNA (4 µg), similarly subjected to size selection and purification of 250-300 bp fragments, served as the negative control sample. P1 and P2 adaptor ligation and all subsequent steps were performed on all 4 samples according to SOLiD System 2.0 Fragment Library Preparation: Lower Input DNA protocol. Templated bead generation for each library was performed according to SOLiD System 2.0 User Guide standard protocols. Each sample was deposited on a quadrant of the slide at a target bead density of 60-70 K beads/panel.

TABLE 1. COMPARISON OF CHIP DETECTION PLATFORMS

Feature	SOLiD System (ChIP-Seq)	Microarray (ChIP-chip)
Resolution	> 400 million sequence tags per run	~ 6.5 million oligonucleotides per array
Genome Coverage	UNLIMITED: Entire genome can be sequenced hypothesis-free	LIMITED by probe design
Specificity	No cross-hybridization risks. Identify unique sequence tags	Cross-hybridization risks between closely related elements
Sample Multiplexing	YES	NO

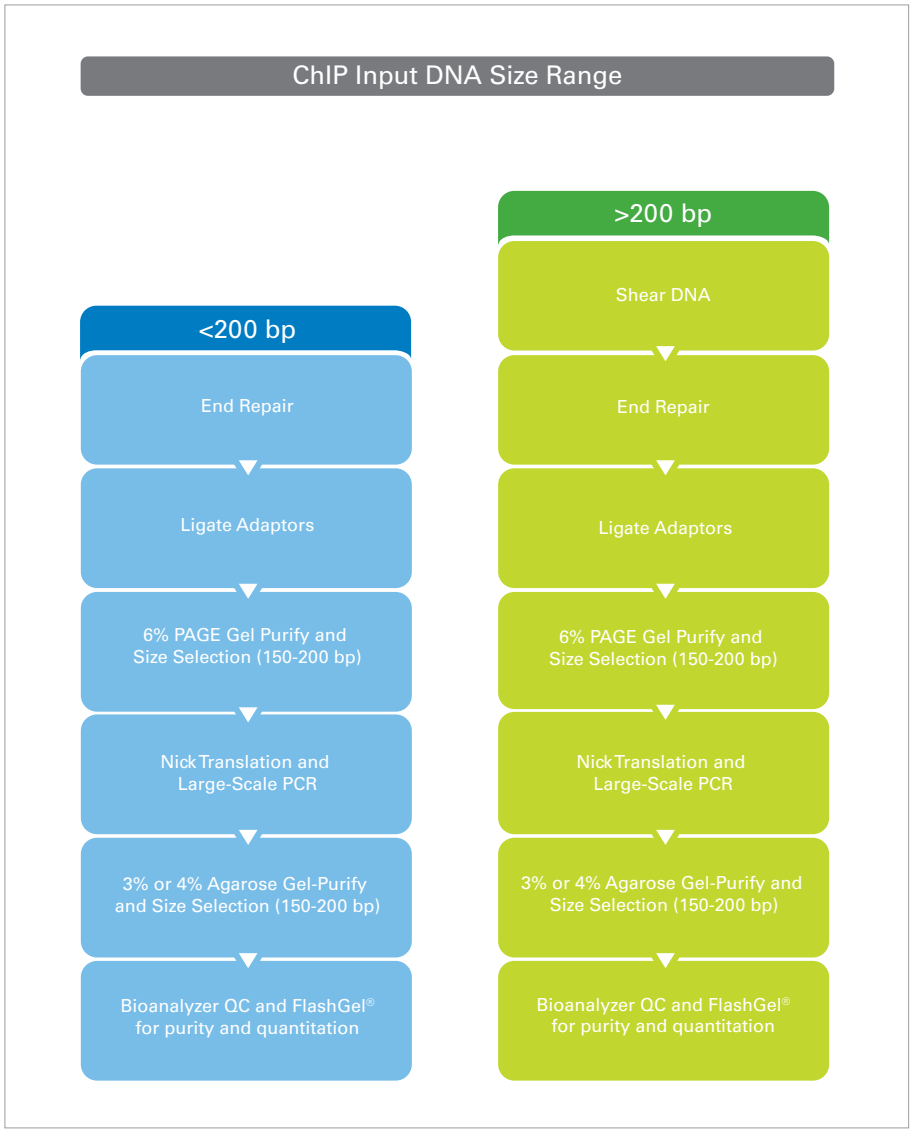

Figure 2. ChIP-Seq SOLiD™ fragment library preparation based on ChIP input DNA size ranges.

A duplicate slide was generated and processed similarly to assess the reproducibility of the system.

High throughput sequencing was performed using the SOLiD™ System and analysis of 35 bp reads was carried out. All reads were filtered for high quality reads (0, 1, 2, or 3 mismatches in color space), as well as for alignment and unique placement in the human reference sequence.

In these experiments, the data from one quadrant generated more than enough reads to map all signals in the genome. We elected to make use of all the data by first verifying a high correlation between the replicates and then merging the reads. The reads were extended *in silico* to represent the selected ChIP fragment sizes (bp). The following process was used to qualify peaks:

1.) A cut-off was calculated for the overlap signal using the binomial distribution as a model under the hypothesis that the reads were randomly placed onto the genome.

2.) FOXA3 peaks also showing significant signal in the control sample were filtered from the analysis.

3.) A peak was required to have a significant signal of forward reads upstream or a reverse signal downstream from its center (an example is shown in Figure 3).

Putative FOXA3 binding regions were visualized using the UCSC Genome Browser (http://genome.ucsc.edu). Detection of putative FOXA3 binding sequences in the enriched peak regions was done using the BCRANK package in R/Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/BCRANK.html).

Results

As expected with an enriched sample, the resulting uniquely mapped reads covered about 10% of the human genome on average (Table 1). The number of reads with a unique starting point averaged approximately 1.07 for all samples, indicating that the sample had unique genomic representation and minimal amplification bias. Based on these data, we estimate that 5-10 million uniquely mapped reads is sufficient to map all protein-binding regions in a complex genome such as human.

All sequence information was used based on the high correlation values between different replicates, ranging from 0.70-0.83. For peak determination, a cut off value of 21 overlapping fragments was used at a Bonferroni corrected p-value <0.01 based on our simulations. Using these stringent

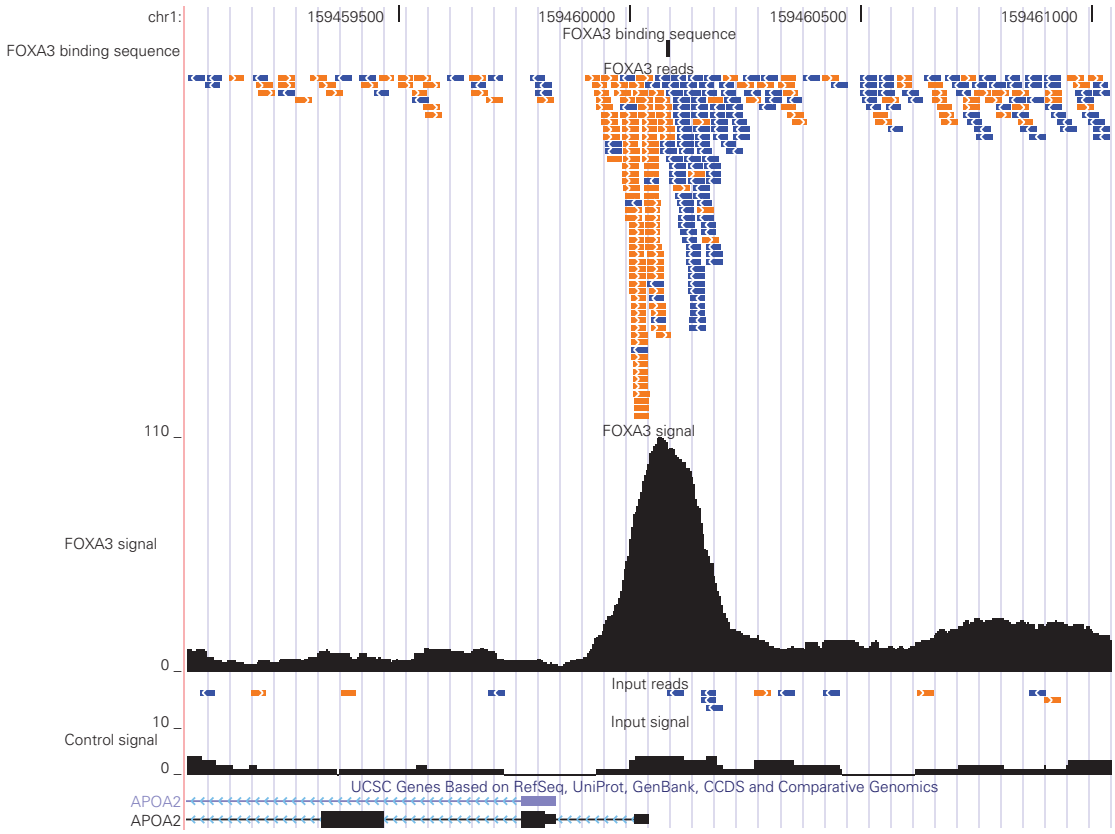

Figure 3. Evidence for FOXA3 protein-DNA interaction at the BCAS1 locus binding to the APOA2 promoter. Graphic representation of alignment of reads derived from size selected SOLiD™ ChIP-Seq libraries (red) and from control input sample (blue) along Chromosome 20 is shown. Illustrated in the UCSC Genome Browser are the mapped forward (red) and reverse (blue) reads and the region of overlap (black). The peak is just upstream of the first exon located and spans the predicted FOXA3 binding motif.

criteria, > 4,000 peaks for FOXA3 were detected either at promoters of protein coding genes or at a distance from a gene that is consistent with FOXA3 interaction with cis-regulatory elements. An example of a FOXA3 peak detected by our peak detection strategy is illustrated in Figure 3 using the UCSC Genome Browser. This peak, located within the APOA2 promoter region which previously has been shown to bind FOXA3, contains 110 overlapping fragments, well above the determined threshold for overlapping fragments. Based on these data, a consensus FOXA3 binding motif was established (Figure 4), which closely resembles the previously characterized FOXA3 binding sequence. Interestingly, this motif was found to be centered at the peak maxima, including that of the APOA2 promoter.

TABLE 2. MAPPING STATISTICS

Sample	Run	Uniquely Mapped Reads
Control 300	1	10.57
IP200	1	8.43
IP250	1	9.39
IP300	1	12.61
Control 300	2	9.85
IP200	2	7.35
IP250	2	9.67
IP300	2	9.02

Table 2. Fraction of uniquely mapped reads (0-3 mismatches) for SOLiD™ ChIP-Seq and control libraries protein. Uniquely mapped reads represent those reads mapping to a single, unique location with less than 3 mismatches in color space.

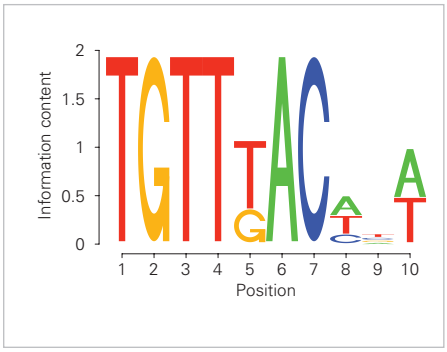

Figure 4. Putative consensus sequence for FOXA3 transcription factor generated using BCRANK. Transcription factor identified by ChIP-Seq on the SOLiD™ System. Information content is plotted as a function of nucleotide position. Sequence logo image was created using an R package called seqLogo (http://bioconductor.org/packages/2.2/bioc/html/seqLogo.html).

Conclusion

The SOLiD™ System provides a level of throughput and sensitivity that cannot be achieved with current hybridization technologies or other next-generation sequencing platforms. The SOLiD System's ability to generate over 400 million sequence tags, to provide a large dynamic range, and to take advantage of multiplexing capabilities, permits multiple hypothesis-neutral ChIP-Seq analyses to be performed in a single run. These system attributes, along with the high degree of accuracy, allow for the determination of regulatory networks in various cellular and pathological states.

Printed in the USA, 11/2008 Publication 139AP14-01

Applied Biosystems

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.327.3002
www.appliedbiosystems.com

International Sales
For our office locations please call the division headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm