ChIP-Seq Profiling of Estrogen Receptor Alpha Binding Sites Using the Illumina Genome Analyzer

Contributed by Willem-Jan Welboren, MSc and Prof. Dr. Henk Stunnenberg
Department of Molecular Biology, Radboud University, Nijmegen, The Netherlands

Introduction

Regulation of gene expression is a key event during development, differentiation, signaling, and adaptation to environmental cues. Gene expression is regulated at many levels. Histone modifications and binding of transcription factors to their cognate target sites in promoters and enhancers lead to the recruitment of activating or repressing cofactors and ultimately in alteration of mRNA levels. To decipher the network of target genes governed by transcription factors, it is essential to identify their genomic binding site repertoire.

The presence or absence of a protein (or histone modification) at a specific genomic location is typically determined using chromatin immunoprecipitation (ChIP). Protein-DNA complexes are crosslinked using formaldehyde, and chromatin is isolated and sheared into small fragments. The protein of interest is then immunoprecipitated using a specific antibody. The DNA crosslinked to that protein is co-precipitated, yielding a fraction enriched for DNA bound by the protein of interest. After reversal of the crosslinks, DNA is purified and used as input for subsequent analysis to identify the enriched DNA.

The application of massively parallel sequencing to ChIP has opened up new avenues to elucidate entire regulatory networks and pathways at the whole-genome scale. To date, genome-wide profiling of DNA-associated transcription factors or individual histone modifications has been performed using microarray platforms. In so-called ChIP-on-chip experiments, the identity of co-precipitated genomic DNA fragments is determined through hybridization to immobilized oligonucleotides or PCR-amplified DNA fragments. However, a drawback of hybridization-based analysis is its inherent bias. Probes must be designed from a known genome sequence, and differ in melting temperature, nucleotide composition, and uniqueness, resulting in differential annealing and cross-hybridization.

In experiments described below, we have used massively parallel sequencing enabled by the Illumina Genome Analyzer to perform a comprehensive analysis of estrogen receptor alpha (ERα) binding sites in MCF-7 cells, a model cell line for breast cancer. The Illumina sequencing platform provides the unique and exciting possibility to identify interaction sites by direct large-scale sequencing of the co-precipitated genomic DNA fragments.

ERα is a member of the nuclear receptors superfamily, which are ligand-dependent transcription factors. ERα is thought to regulate transcription of target genes by either directly binding to specific cis-acting DNA sequences—estrogen response elements (EREs)—or indirectly via protein-protein interactions with other transcription factors such as AP-1, Sp1, or NF-κB that are bound to DNA at their cognate regulatory sites. The unsurpassed accuracy and sensitivity of the Illumina Genome Analyzer enabled us to comprehensively identify and catalog the ligand-dependent genomic interaction sites of ERα. A brief overview of the experimental design and results are reviewed in this application note.

Methods

MCF-7 cells were hormone deprived for 48 hours, followed by induction for one hour with either 10 nM 17β-estradiol (E2) or solvent control. Proteins and DNA were crosslinked for 30 minutes at room temperature using 1% formaldehyde, quenched with 0.125 M glycine, and washed at 4°C. Chromatin was sheared into fragments (~500bp–1kb) using the Bioruptor (Diagenode). ChIP was performed using a Red ChIP kit (Diagenode) as described by the manufacturer using 2.5 µl of an anti-ERα monoclonal antibody (Diagenode Mab-NRF3A6-050). Then, the chromatin was incubated overnight with protein A/G beads and secondary antibody at 4°C with slow rotation. Subsequently, the beads were washed six times at increasing stringency to remove nonspecifically bound chromatin. To reverse the crosslinks, precipitated chromatin was incubated at 65°C for 4 hours with the addition of 200 mM NaCl, then phenol extracted and precipitated overnight. Three parallel ChIP products using approximately 3 x 10^6 cells were pooled. DNA was purified using the Qiagen reaction cleanup kit. The total amount of DNA was measured using PicoGreen (Invitrogen) and a NanoDrop Spectrophotometer.

To assess the quality of the ChIP, ERα occupancy was detected by qPCR at the pS2/TFF1 and GREB1 promoters and a known enhancer region on chromosome 1. Primers were designed using the primer3 algorithm[1] with a product size of 50–150bp, and verified to produce one amplicon by in silico PCR. The formation of a single specific amplicon was also assessed by qPCR on genomic DNA. Primers that amplified more than one product were discarded. The qPCR was performed on a BioRad MyIQ light cycler. The quality of the ChIP was deduced from the recovery (yield), calculated as the percentage of chromatin input that was co-precipitated in the ChIP assay, and from the occupancy (specificity), calculated as fold enrichment over an arbitrarily chosen non-binding control region. ERα occupancy at the pS2/TFF1 promoter, GREB1 promoter, and an enhancer region on chromosome 1 were 15-, 86-, and 221-fold enriched (respectively) compared to the control (exon 2 of the myoglobin gene) with recoveries greater than 3–4%.

Finally, for genome-wide readout, the purified co-precipitated DNA was processed for analysis on the Illumina Genome Analyzer using the Illumina ChIP-Seq Library Preparation Kit (IP-102-1001). Sequencing reactions were performed with system components and reagents available in September 2007.

Table 1: Summary of Sequenced and Annotated Tags

Condition	Total tags	Two mismatches	No mismatch
- E2 (control)	3,375,602	1,736,369 (51%)	977,618 (28%)
+ E2	7,500,354	5,594,488 (75%)	4,395,434 (59%)

Illumina Sequencing Technology

Illumina DNA sequencing technology leverages clonal cluster amplification and reversible terminator nucleotides to generate high-density, high-throughput sequencing runs. The fully automated Illumina Cluster Station amplifies adapter-ligated ChIP DNA fragments on a solid flow cell substrate to create clusters of approximately 1000 clonal copies each. The resulting high-density array of template clusters on the flow cell surface is sequenced by the Illumina Genome Analyzer. Tens of millions of template clusters present on a flow cell undergo sequencing by synthesis in parallel.

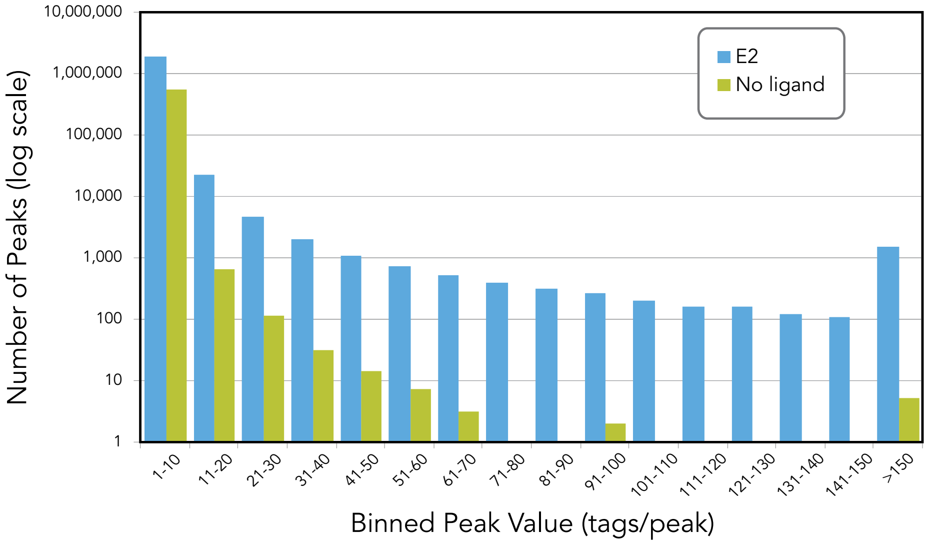
Due to this capacity for high oversampling and potential for read-signal redundancy, binding event signals are readily detectable above background. Sensitivity and statistical certainty can be tuned by adjusting the total number of sequence reads to provide an even wider dynamic range and greater sensitivity to detect rare or weak DNA-protein interaction sites. Illumina's data collection and analysis software aligns DNA sequence reads to a reference genome sequence, allowing determination of all of the binding sites for a factor of interest. Sequence read lengths of only 25–32 bases are sufficient to accurately align and identify millions of fragments per run. Unlike microarray-based ChIP readout methods, the positional accuracy of ChIP-Seq is ±50bp or less.

Identification of Enriched Regions

Using two flow cell lanes each for ChIP DNA from E2-induced (+E2) and non-induced (-E2) cells, a total of ~7.5 million and ~3.3 million 32-base sequence reads were obtained. All 32-base sequence reads were mapped to the human reference genome sequence (NCBI36 - HG18) using the Illumina ELAND algorithm at its default setting (two or no mismatches allowed). Table 1 shows the total number of sequenced tags and the number of tags that could be aligned to the genome for the E2-induced and the non-induced data sets. The percentage of annotated tags is significantly higher in the +E2 data set compared to the no ligand control (-E2). This is expected because in the presence of E2 ligand, ERα targets are significantly enriched in the ChIP, and consequently the background noise is underrepresented. In the absence of ligand, ERα does not bind to DNA. Hence, enrichment-specific genomic fragments are not obtained and the population of sequenced DNA fragments is then a representation of the full human genome with its high proportion of repetitive elements that cannot be mapped unambiguously to the genome.

To determine ERα peaks, the aligned 32-base sequence tags were computationally extended to 200bp, since DNA fragments of ~200bp were excised from the original agarose gel and processed for sequencing. The 200bp sequence strings were sorted and the

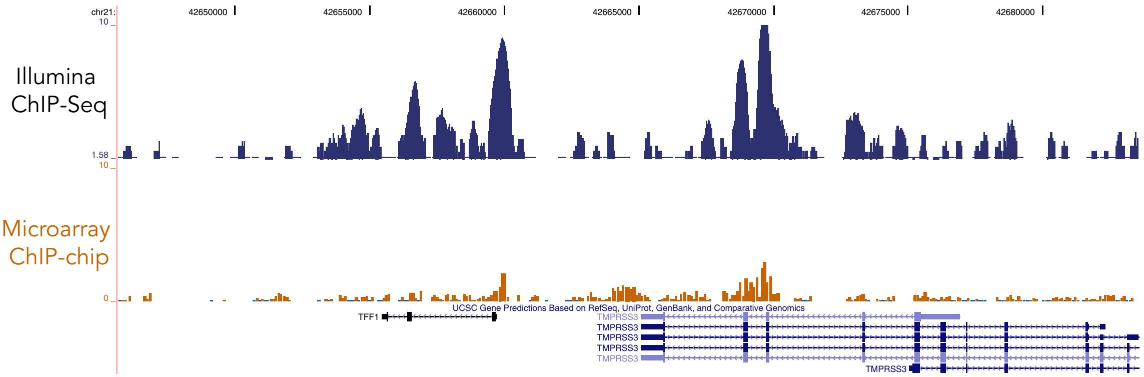
Figure 1: Histogram of the Distribution of Peak Value (Sequence Tags Per Peak)



Peaks were binned according to peak value (number of overlapping sequence tags). The number of peaks within each bin is plotted on a log10 scale. For the no-ligand control data set (green) the vast majority of the peaks are in the bins with the lowest peak value, indicating that only very few high-confidence ERα interaction sites (> 20 overlapping sequences) were detected. The E2-induced ERα interaction site data (blue) show peaks throughout the entire range (up to 10,000 peaks), indicating the presence of high-confidence interaction sites that are occupied in a strictly ligand-dependent manner.

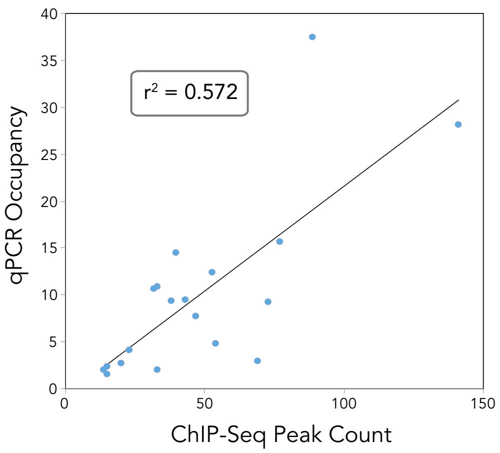
AATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTA
AACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAC
ACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGT
CGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACA
AATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTA
TTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACGAA
TATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTAAC

Figure 2: Comparison Between ChIP-Seq and ChIP-On-ChIP



Screenshots of the UCSC genome browser[8] showing ERα binding sites around the well known ERα target gene, pS2/TFF1. ChIP-Seq results for ERα binding following E2 induction are shown in the upper track in blue. ChIP-on-chip results are in the lower track in orange. The UCSC known gene track is on the bottom. For the Illumina ChIP-Seq data, the number of overlapping fragments in a 10bp fixed window is counted and plotted on a log2 scale. For ChIP-on-chip, the ratio between ChIP/input is shown on a log2 scale. The highest peaks in ChIP-Seq as well as ChIP-on-chip are observed over the promoter region and the upstream enhancer at -10kb. The high resolution and signal-to-noise ratio of the ChIP-Seq analysis reveal two distinct peaks at the upstream enhancer.

Figure 3: Validation of Randomly Selected ERa Binding Sites



Twenty ERα binding sites identified by ChIP-Seq were randomly chosen to be validated by ChIP-qPCR. Relative occupancy (average of 3 biological replicates) assessed by qPCR (y-axis) is plotted against sequence reads per peak (x-axis) for each site, demonstrating the high correlation between the two different methods.

overlap between fragments was determined. Overlapping fragments were combined into one peak-ID, resulting in a list with one or more fragments per peak (peak value). For each data set, the peaks were binned according to peak value and the number of peaks within each bin were counted. Figure 1 shows the distribution of sequence tags across peak value bins. The distribution of peaks in the non-induced set is skewed left and almost all are in the bins with the fewest tags, whereas the +E2 set has peaks across the entire range of peak values. Based on qPCR validation data (false positive rate < 5%), we applied a threshold of 20 overlapping fragments to identify true binding sites. This resulted in the identification of 13,173 ERα interaction sites following ligand treatment. In contrast, in the absence of ligand only 191 apparent interaction sites were detected at this threshold, and the majority of these apparent peaks mapped to regions rich in repetitive sequences (145) or overlapped with amplified regions in MCF-7[2].

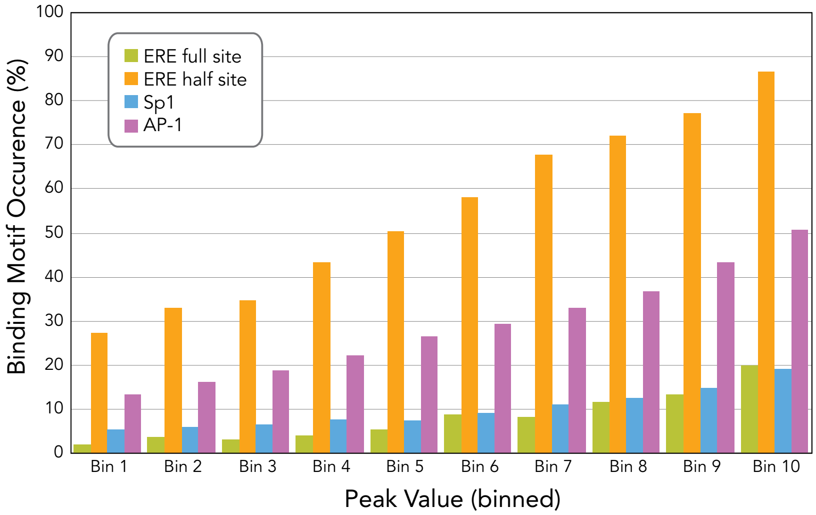
Our ChIP-Seq analysis resulted in the unambiguous identification of genomic ERα interaction sites at high signal-to-noise ratio and dynamic range: peaks of 20 to ~10,000 overlapping sequence tags against a background of fewer than 20 sequence tags. These data confirm that the interaction of ERα with chromatin is strongly dependent on the presence of ligand.

The landscape of ERα interaction sites across the genome was determined by counting the number of aligned fragments in a fixed window of 10bp and plotted on a log^2 scale. A comparison between ChIP-on-chip and ChIP-Seq profiles is shown in Figure 2 for the classical ERα target gene, pS2/TFF1, showing the superior signal-to-noise ratio, read depth, and sensitivity of ChIP-Seq.

Validation of Several ERa Binding Sites

To confirm the accuracy of identified ERα binding sites, 20 peaks were randomly selected for validation using ChIP-qPCR. ChIP-qPCR is commonly accepted as the gold standard to determine the occupancy of ERα at an individual site. Primers were designed and a targeted ChIP was performed on three biological replicates as described above. The average relative occupancy values obtained by qPCR were compared to their the peak values. Figure 3 shows that the peak values obtained by ChIP-Seq correlated well with the ChIP-qPCR data ($r=0.76$, $r^2=0.57$), demonstrating the semi-quantitative character of the ChIP-Seq approach.

Figure 4: Searching for Sequence Motifs



Using the Transfac 11 database and a Jaspar Estrogen Responsive Element (ERE_weight matrix), the sequences underlying identified peaks were scanned for the presence of ERα, Sp1, and AP-1 binding sites. Specific transcription factor binding motifs in the sequence underlying a peak are shown for various peak values (x-axis). Peaks were divided into 10 equal bins according to peak value. The percentage of peaks in a bin containing a specific motif is plotted on the y-axis. Peak value correlates well with the presence of ERE half-site, ERE full-site, AP-1, and Sp1 motifs.

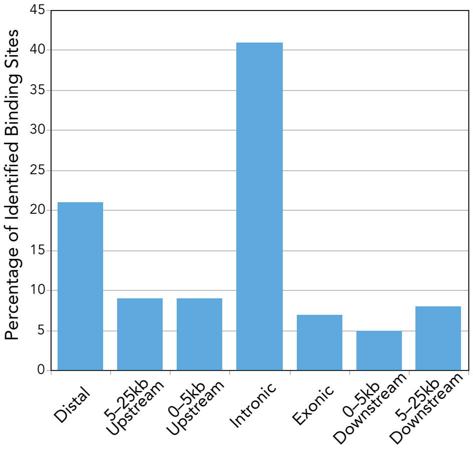
Motif Search

To determine cis-acting sequence elements that facilitate binding of ligand-loaded ERα (either directly or via protein-protein interaction), the entire ChIP-Seq results were searched for consensus estrogen response elements (half-site or full-palindrome), AP-1, and Sp1 motifs. The search was done using the MATCH program and weight matrices from the Transfac 11 database[3]. For the ERE full site, the Jaspar weight matrix[4] was used. Peaks were divided into 10 equal bins according to their peak value. For each bin, the percentage of peaks that contain a particular motif was calculated. Figure 4 shows that there is a good correlation between peak value and the occurrence of specific motifs. This demonstrates that peaks identified by a high number of sequence tags are more likely to contain an ERE half-site, AP-1 motif, ERE full site, or Sp1 motif.

Visibility into Other Interactions

Results from chromatin conformation capture assays—known as 3C or 4C[5,6]—imply that genomic transcription factor interaction sites identified by our highly sensitive ChIP-Seq analysis may comprise not only primary sites of ERα binding (cis-acting elements such as EREs or AP-1 sites) but also secondary genomic sites to which ERα was crosslinked due to looping, such as between promoters and enhancers. Given the sensitivity and depth of ChIP-Seq, it seems likely that not only stable short- and long-range intra- and inter-chromosomal interactions, but also short-lived transient interactions will be uncovered with the ChIP-Seq approach.

Figure 5: Location of Binding Sites Relative to Ensembl Genes



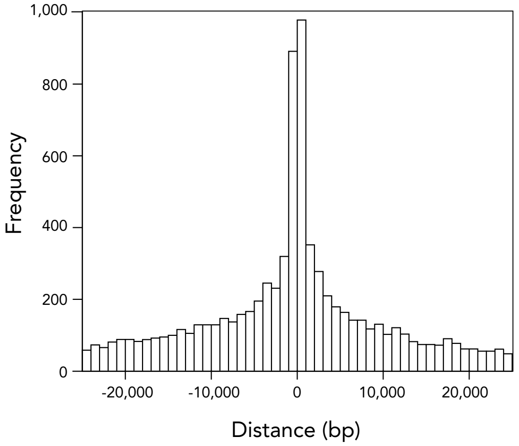
For each ERα-binding peak detected by ChIP-Seq, the nearest gene and the distance to that gene were determined. Binding sites were divided into 7 classes of functional regions. The majority of binding sites are located in introns (41%) or more than 25kb from a gene (21%). 9% of all sites are located in promoter regions (i.e., within 5kb upstream of a gene).

Locations of Binding Sites

The location of binding sites relative to the closest gene was determined using the Ensembl 47 database[7]. Peaks were divided into 7 classes based on their relative location.   shows that the majority of ERα interaction sites (41%) are located in introns while 9% are located in promoter regions. In addition, a large number of sites (21%) are located distal from any annotated gene. These data emphasize that ERα can function as an enhancer, as well as a classical transcription factor.

To further characterize binding site locations relative to a gene start site, the distribution of binding sites relative to the transcription start site (TSS) was determined in a window of 30kb upstream and downstream of the TSS. As shown in Figure 6, this distribution reveals that ERα binding is highly enriched in close proximity to the TSS.

Figure 6: Distance From Detected Binding Sites to Closest TSS



Histogram depicting the binding of ERα relative to the closest transcription start site. The large peak centered at 0bp indicates that ERα binding is enriched close to transcription start sites.

Conclusion

Genome-wide ChIP-Seq analysis enables characterization of the interactome, or entire interaction network, of a transcription factor of interest. We chose to examine the interactome of ERα, and with a single experiment have substantially confirmed and expanded on the decades of research that have focused on this factor. With data from only two lanes of a flow cell (~7 million tags), we were able to identify thousands of precise genomic locations of direct, as well as indirect, ERα binding. With these data, we confirmed known ERα binding sites, revealed novel enriched regions, and provided new information on cis-acting sequences that facilitate the binding of ERα.

With high sensitivity, low noise, and no hybridization bias, Illumina ChIP-Seq provides major advantages over microarray-based detection. In addition, it is possible to obtain data on binding in repeat regions, which are excluded in microarray-based ChIP-on-chip approaches. These data facilitate future experiments to focus on further analysis of novel discovered binding sites as well as the analysis of genome-wide binding patterns of other ERα-related transcription factors.

References

1. http://frodo.wi.mit.edu
2. Shadeo A, Lam WL (2006) Comprehensive Copy Number Profiles of Breast Cancer Cell Model Genomes. Breast Cancer Res 8: R9.
3. http://www.biobase-international.com
4. http://jaspar.cgb.ki.se
5. Dekker J (2006) The Three 'C's of Chromosome Conformation Capture: Controls, Controls, Controls. Nature Methods 3: 17-21.
6. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R et al. (2006) Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture-on-Chip (4c). Nature Genetics 38: 1348-1354.
7. http://www.ensembl.org
8. http://genome.ucsc.edu

ADDITIONAL INFORMATION

Visit our website or contact us at the address on the back page to learn more about ChIP-Seq or other Illumina Sequencing or epigenetics analysis applications.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

Pub. No. 770-2007-016 Current as of 3 March 2008