

Application Note

Whole Genome Chromatin Interaction Analysis with Paired-End Tags (ChIA-PET)

Contributed by Dr. Melissa J. Fullwood and Prof. Yijun Ruan

Genome Technology and Biology Department, Genome Institute of Singapore, A*STAR Singapore

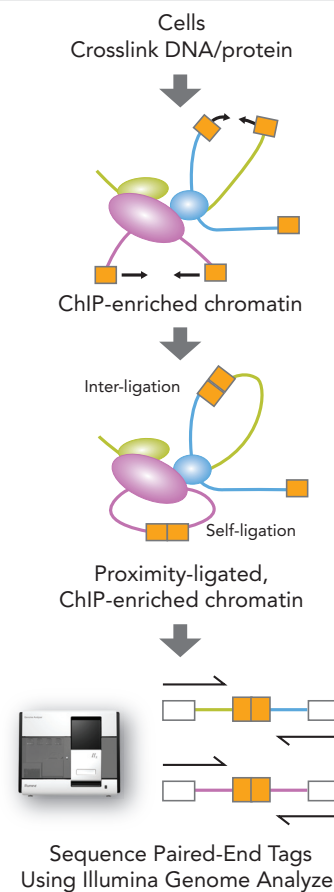
INTRODUCTION

Researchers at the Genome Institute of Singapore (GIS) of the Agency for Science, Technology, and Research (A*STAR), led by Yijun Ruan, have mapped abundant estrogen receptor alpha (ER- α) binding sites and chromatin interactions between the binding sites using a novel method of whole-genome, *de novo* chromatin interaction analysis with paired-end tags (ChIA-PET)^{1,2}.

Although genome sequence information is usually presented as a linear series of bases, genomic DNA is dynamically associated with protein factors and folded to form chromatin fibers and higher-order three-dimensional structures. Genetic elements separated by long distances may come into close proximity as a result of chromosome conformation and undergo a number of interactions, including transcriptional regulation. Distal transcription factor binding sites have been characterized through a number of molecular interaction mapping approaches. These include chromatin immunoprecipitation followed by hybridization of the bound DNA to a solid surface covered with DNA probes (ChIP-chip)³, direct sequencing of the bound DNA by paired-end tags (ChIP-PET)⁴, or high-throughput, single-tag sequencing (ChIP-Seq)⁵. These approaches have revealed important potential mechanisms of transcriptional regulation⁶.

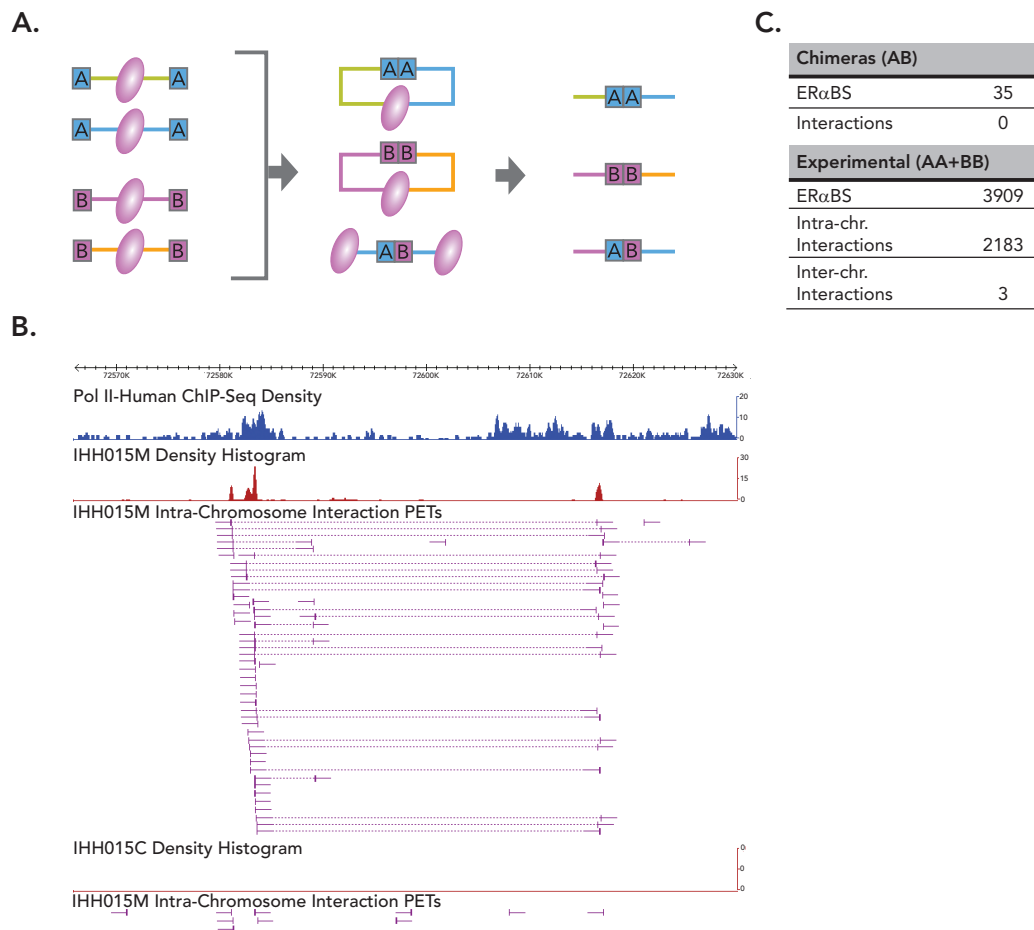
Paired-end tag (PET) analysis is a powerful technique that extracts short tag signature information (20–30 bp) from the two ends of target DNA fragments, pairs the two tags for sequencing analysis, and maps the paired tag sequences to reference genomes to demarcate the boundaries of the target DNA fragments in the genome landscape⁷. ChIA-PET is a novel PET-based method that enables unbiased, genome-wide *de novo* detection of chromatin interactions. First, a cross-linking step is performed to immobilize protein/DNA interactions, followed by immunoprecipitation with antibodies specific for the protein factors of interest⁸, followed by sonication to break up the DNA. Within

FIGURE 1: WHOLE-GENOME CHROMATIN INTERACTION ANALYSIS WITH PAIRED-END TAGS (ChIA-PET)



High-throughput sequencing enables a novel, paired-end tag-based method for *de novo* detection of genome-wide chromatin interactions.

FIGURE 2: LINKER BARCODING SCHEME FOR DETECTING TRUE BINDING AND INTERACTION SITES



A. Linker barcoding is used to identify chimeras and experimental PETs for further analysis. B. High peaks from density histograms constructed from overlapping, unique self-ligation PETs are taken to be binding sites, and overlapping, unique inter-ligation PETs are taken to be interactions. C. Table of putative, bona fide binding sites and interactions from chimeric and experimental data sets.

the resulting DNA/protein complexes, a linker sequence is proximity ligated at the junction of two DNA fragments. Linker-connected ligation products are extracted by digestion and analyzed by ultra high-throughput PET sequencing on the Illumina Genome Analyzer (Figure 1).

Using the ChIA-PET approach, the GIS group determined that most remote ER- α binding sites are anchored at gene promoter regions through long-range chromatin interactions¹. These data suggest that ER- α functions by physically bringing genes together through intensive chromatin looping into transcription foci. This application note explains the ChIA-PET experimental design and how it was used to map ER- α binding sites within the genome of a breast cancer cell line.

METHODS

Breast Cancer Cell Preparation

Cultured human breast adenocarcinoma (MCF-7) cells were deprived of hormones for three days, and then treated with estrogen for 45 minutes. Cells were cross-linked using 1%

formaldehyde, lysed, and sonicated. Sonication shears the chromatin into manageable sizes, and also breaks weak chromatin interactions where two DNA strands floated into close spatial proximity by chance. To obtain the most efficient sequencing results, ChIP was performed using anti-ER- α antibodies for higher specificity and reduced library complexity^{1,2}.

Proximity Linker Ligation

DNA fragments within ChIP complexes were ligated to biotinylated half-linkers (linker ligation) containing flanking MmeI restriction sites. The complexes were further ligated under dilute conditions (proximity ligation) such that the half-linkers on two or more DNA strands within the same chromatin complex would be ligated (called an inter-ligation) or the DNA strands would self-circularize via the half-linkers on the ends (called a self-ligation, Figure 1)^{1,2}.

Paired-End Tag (PET) Sequencing

Paired-end tags (PETs) were extracted from the ligation prod-

ucts by MmeI digestion. MmeI is a type IIs restriction enzyme that cuts 18–20 bp downstream of its recognition site⁹. PETs containing 18–20 bp of biotin-tagged sequence on each end of the ligation site were purified by immobilization to streptavidin-coated magnetic beads, ligated to sequencing adaptors, and PCR amplified. The gel-purified, amplified PETs were then sequenced on the Illumina Genome Analyzer and mapped to the hg18 genome assembly using ELAND. PETs were analyzed using an in-house ChIA-PET analysis pipeline¹⁰. PETs were classified as self-ligation if the two tags mapped close to each other (< 3 kb) on the same strand in head-to-tail orientation. Otherwise, they were classified as inter-ligation PETs^{1,2}.

Linker Barcoding Analysis

To identify whether chimeric ligations would result in false positive interactions, we incorporated barcoded linkers—half-linkers that contain differences in their sequences—allowing them to be distinguished from each other upon sequencing. Before proximity ligation, samples were divided into two groups. To one, half-linkers A were added. To another, half-linkers B were added. The samples were mixed and proximity ligated such that the half-linkers ligated to each other (Figure 2A). We found chimeric sequences with linkers AB (chimeras) and compared them with a library of AA and BB sequences that had been sequenced to a similar depth (experimental) (Figure 2B). We found that chimeric sequences were distributed randomly throughout the genome as inter-ligation PETs, and did not cluster to form multiple overlapping self-ligation PETs or inter-ligation PETs. As such, multiple overlapping self-ligation PETs are likely to be real binding sites and multiple overlapping inter-ligation PETs are likely to be real interactions (Figure 2C)¹.

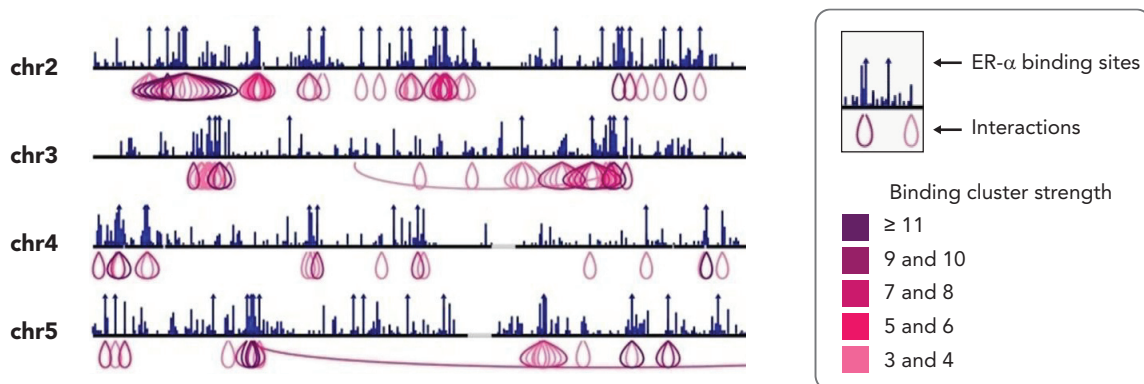
Binding Site Identification

To identify ER- α binding sites, we examined the sequence of multiple overlapping self-ligation PETs. Because the DNA was sonicated, the likelihood that multiple overlapping unique PETs will occur by random chance is low, and only ChIP enrichment would result in multiple PETs. This method is similar to that described by Wei *et al.*⁴ A threshold equivalent to false discovery rate < 0.01 was used, corresponding to ~4–5 self-ligation PETs. Binding sites that overlapped with satellites (highly repetitive regions of the genomes found in chromosome structures such as centromeres and telomeres), as well as binding sites that overlapped with regions comprising known structural variants were removed based on published information as well as genome structural variation studies using paired-end tag sequencing¹¹. Manual curation of individual sites on the genome browser verified that the automated analyses were performed correctly. As shown in Figure 3, we mapped abundant presumptive ER- α binding sites^{1,2}.

Interaction Identification

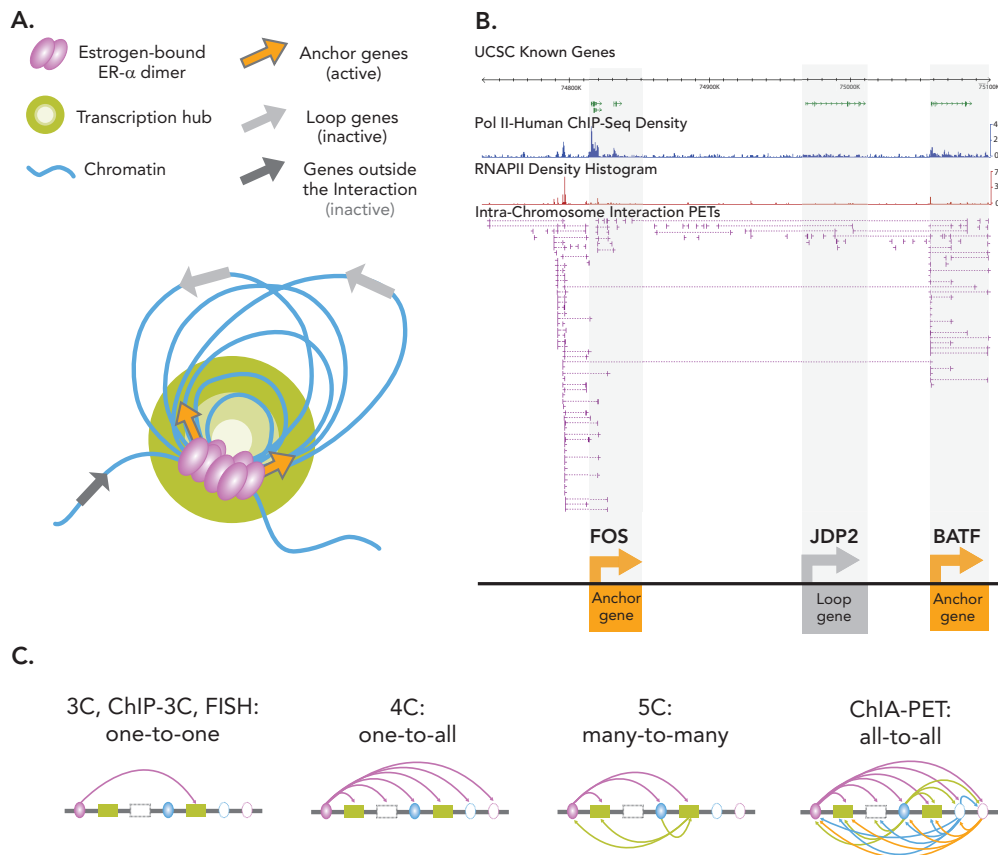
To identify binding site proximity interactions and correct for potential ChIP enrichment bias, we formulated a statistical analysis framework to calculate the probability of inter-ligation formation between two regions if ligations occurred by chance, and assigned each interaction a confidence score. An interaction consists of two anchor regions connected by a loop. Only interactions with three or more inter-ligation PETs between the two regions and a false discovery rate of < 0.05 were analyzed. We removed interactions with anchors that overlapped with satellites and structural variants that were found by genomic structural variation analyses studies using paired-end tag sequencing of genomic DNA. Manual curation of all inter-chromosomal interactions by visualization of the

FIGURE 3: WHOLE-GENOME INTERACTION VIEW OF ER- α BINDING SITES



All ER- α binding sites and interactions were visualized as a whole-genome interaction view, represented here as a partial view of chromosomes 2–5. ER- α binding site intensities correlate with the height of the blue arrows, and chromatin interaction intensities are shown by the color indicating binding cluster strength. Most interactions are within < 1 MB, represented as short loops. Complex interaction regions appear onion-shaped because there are many interactions at those loci.

FIGURE 4: GENE ANALYSIS AND CHROMATIN INTERACTION MODEL



A. Example of a complex interaction comprising anchor and loop genes displayed in the genome browser with RNA polymerase II annotations. B. Chromatin interaction model. Complex, local, and intrachromosomal chromatin interactions bound by ER create boundaries for inactive and active genes. C. Unlike other methods based on prior knowledge and either target-specific PCR assays (3C, ChIP-3C, 4C, 5C) or fluorescent probes (FISH), ChIA-PET allows discovery and analysis of all DNA/protein interactions across the genome.

interactions on genome browsers revealed that many interactions could not be trusted because they occurred in regions with genomic structural variations just below the threshold used for automated removal. These were filtered out. We curated a similarly sized subset of the intra-chromosomal interactions, and found no such issues with genomic structural variations just below the threshold for automated removal. As shown in the example in Figure 3, we found 2,183 intra-chromosomal interactions^{1,2}.

Frequent Complex Interactions

We observed that many chromatin interactions are connected to other interactions, forming so-called daisy chain-connected interactions. These connected interactions were defined as complex, whereas those that did not interconnect were annotated as standalone, duplex interactions (Figure 3).

Validation

ER- α binding sites were validated by ChIP-qPCR. We validated 11 putative intra-chromosomal interaction sites (both interactions and negative controls) using 3C, ChIP-3C, 4C, and

FISH methods (data not shown), and in all cases, we could reproduce the ChIA-PET results¹. Gene expression levels were validated by RT-qPCR.

Loop and Anchor Gene Identification

UCSC Known Gene transcriptional units were associated with interactions using the following definitions: A gene was called an anchor gene if a transcription start site (TSS) was within +20 kb of the middle of an interaction anchor. Loop genes were defined if a TSS was within an interaction but not near an anchor. Background genes occur outside of anchor or loop genes (Figure 4A). We characterized gene activity in MCF-7 by ChIP-Seq for RNA Polymerase II (RNAPII) using Illumina single-pass sequencing. The ChIP-Seq data was mapped to the hg18 genome build, and enrichment peaks were derived using a previously-described *de novo* motif discovery algorithm¹¹. We found that anchor genes appeared to be more activated compared with loop and background genes (Figure 4B)^{1,2}.

CONCLUSIONS

Complex, local, intrachromosomal chromatin interactions bound by ER- α constitute a common mechanism by which estrogen-regulated gene promoters are brought into transcriptional hubs for coordinated gene activation. ChIA-PET is a novel *de novo*, high-throughput method for identifying global chromatin interactions, and is a valuable starting point for detailed analyses of three-dimensional chromatin interactions. Unlike other methods such as 3C, FISH, 4C, and 5C that examine the connectivity of one or more DNA fragments tethered by protein factors, ChIA-PET is a high-throughput method of examining all connections across the genome (Figure 4C). Illumina paired-end sequencing enables easy and cost-effective acquisition of high numbers of paired-end tags for ChIA-

REFERENCES

1. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor-alpha-bound human interactome. *Nature* 462:58–64.
2. Fullwood, MJ et al., *Current Protocols in Molecular Biology*, Manuscript accepted.
3. Collas P, Dahl JA (2008) Chop it, ChIP it, Check it: The current status of chromatin immunoprecipitation. *Front Biosci* 1,13 929–934.
4. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P et al. (2006) A global map of p53 transcription factor binding sites in the human genome. *Cell* 124, 207–219.
5. Wold B, Myers RM (2008) Sequence census methods for functional genomics. *Nat Methods* 5,1 19–21.
6. Massie CE, Mills IG (2008) ChIPing away at gene regulation. *EMBO Rep* 9,4 337–343.
7. Fullwood MJ, Wei C-L, Liu ET, Ruan Y (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analysis. 19,4 521–532.
8. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 113, 1106–1117.
9. Morgan RD, Bhatia TK, Lovasco L, Davis TB (2008) MmeI: A minimal type II restriction-modification system that only modifies one DNA strand for host protection. *Nucleic Acids Res* 36, 6558–6570.
10. Li et al. (2009) manuscript in preparation.
11. Ruan, Y., personal communication.
12. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, et al. (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biology* 11:R22

ACKNOWLEDGMENTS

The authors wish to thank Dr. Wei Chia-Lin, Dr. Edwin Cheung, Dr. Edison Liu, Dr. Valere Cacheux-Rataboul, Dr. Ruan Xiaolan, Dr. Henk Stunnenberg, Dr. Ken Sung, Dr. Liu Mei-Hui, Dr. You-Fu Pan, Mr. Yusoff Bin Mohamed, Mr. Han Xu, Mr. Hong-Sain Ooi, Mr. Willem-Jan Welboren, Dr. Roy Joseph, Mr. Phillips Huang, Ms. Yuyuan Han, Dr. Guoliang Li, and the Genome Technology and Biology Sequencing team. This work was funded by an NIH ENCODE grant and the Agency for Science, Technology, and Research (A*STAR).

We are committed to providing you with the content you want as a member of the Illumina community. Please email us with comments and suggestions for topics at icomunity@illumina.com.

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.

Illumina, illuminaDx, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, and HiSeq are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.

Pub. No. 070-2010-004 Current as of 3 March 2010

illumina[®]