PAIRED-END TAGS FOR UNRAVELLING GENOMIC ELEMENTS AND CHROMATIN INTERACTIONS

MELISSA JANE FULLWOOD

(BSc. (Hons.), STANFORD UNIVERSITY)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2009

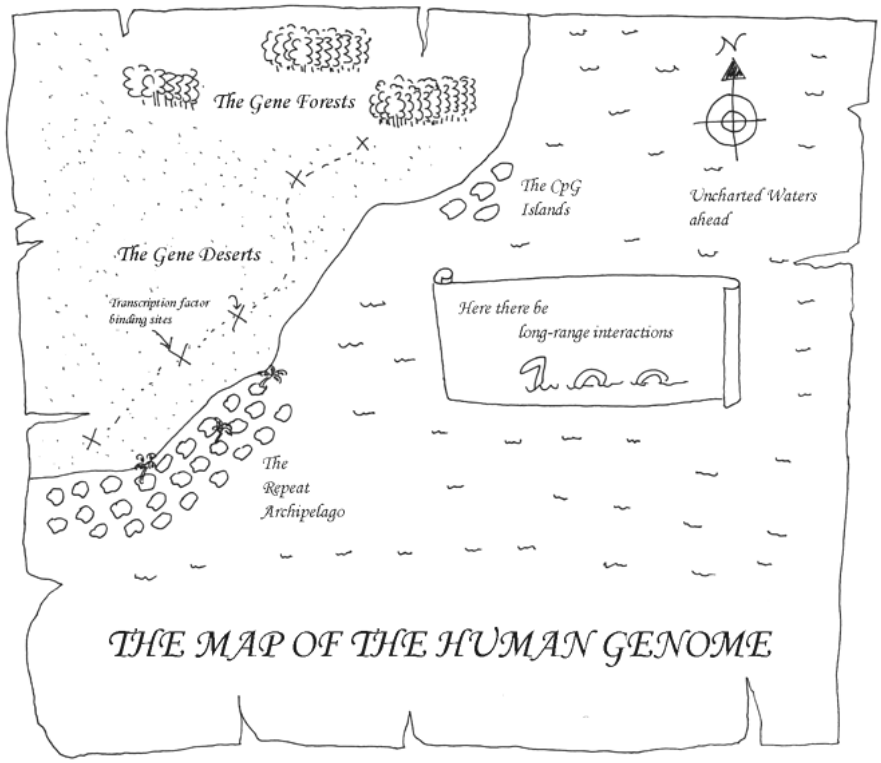
Table of Contents

Acknowledgements v
Summary viii
List of tables ix
List of figures x
List of abbreviations and symbols xi
Chapter One: Paired-End Tag Technologies 1
Introduction 1
The development of the Paired-End Tag (PET) strategy 4
Construction of PET structures 9
Sequencing analysis of PET constructs 12
Insights from PET applications to transcriptome studies 16
Insights from PET applications to genome structure analysis 18
Insights from PET applications to identify regulatory and epigenetic elements 23
New developments in PET technology 27
Proposal: Finding chromatin interactions with PETs 29
Chapter Two: Selection-MDA for amplifying complex DNA libraries 34
Introduction 34
Results 37
Discussion 44
Chapter Three: Whole Genome Chromatin Interaction Analysis using Paired-End Tag Sequencing 47
Introduction 47
Results 48
Construction and mapping of ChIA-PETs 48
ERα binding sites and interactions determined by ChIA-PETs 56
Discussion 76
Chapter Four: The Estrogen Receptor α-mediated Human Chromatin Interactome 79
Introduction 79
Results 79
ERα-mediated chromatin interactome map 79
ERαBS association with interactions and other DNA elements 94
Chromatin interaction and transcription regulation 100
Discussion 109
Chapter Five: Conclusions 112

Summary 112
The future of chromatin interactome biology 112
The future of the PET technology 115
Chapter Six: Materials and Methods 119
Materials and Methods used in Chapter 2 119
Cell culture 119
Full length cDNA library construction 119
GIS-PET library construction 120
Selection-MDA GIS-PET library construction 120
Data analysis 121
Materials and Methods for Chapter 3 122
Cell culture and estrogen treatment 122
Chromatin immunoprecipitation (ChIP) 123
ChIA-PET library construction and sequencing 123
ChIA-PET barcoding 124
RNAPII ChIP-Seq 125
Cloning-free ChIP-PET library construction and sequencing 125
Library saturation analysis 126
DNA-PET 10 Kb insert data 126
PET extraction and mapping 127
PET classification 127
Identification of ERα binding sites 128
Identification of ChIP enrichment levels 129
ERE motif analysis of ERα binding sites 129
Comparative analysis of ERα binding sites 130
ChIA-PET data visualization 131
Using inter-ligation PETs to identify ERα-mediated interactions 131
Manual curation 133
Assignment of genes to high confidence interactions 133
Chromosome Conformation Capture (3C) 134
Chromatin Immunoprecipitation Chromosome Conformation Capture (ChIP-3C) 134
RT-qPCR 135
ChIP-qPCR 136
Materials and Methods for Chapter 4 136
ChIA-PET library construction and sequencing 136

H3K4me3 ChIP-Seq data 137
RNAPII ChIP-Seq data 137
DNA-PET 10 Kb insert data 137
Microarray gene expression data to identify estrogen-regulated genes 137
PET sequence analysis 138
Interaction complexes 138
ERαBS association with relevant genomic features 139
TRANSFAC analysis 141
Association of ERα-mediated chromatin interactions with genes 142
Gene expression visualization and analysis 143
Circular Chromosome Conformation Capture (4C) 144
Fluorescence in-situ hybridization (FISH) 145
siRNA knockdown 147
References 148
Appendices 159

Acknowledgements



Genomics research appears to be a very high-tech endeavor. But our understanding of the human genome is still in early days, and frequently, we seem to be using extremely rough maps. In this thesis, I have hunted the elusive long-range interactions (which sometimes do resemble dragons indeed), and sailed the often-stormy uncharted waters of the human genome with technologies that I've had to improvise. Of course, this journey would not have been possible without the help of many people. And so, I'd like to thank…

My parents, family, and friends, for supporting me always.

Ruan Yijun, for being my PhD supervisor, and providing me with a lot of support.

Edison Liu, for mentoring me for 7 years and working with me on the ChIA-PET papers.

Edwin Cheung, for mentoring me during my lab rotation, and also working with me on the ChIA-PET papers.

Cagan Sekercioglu, Arthur Kornberg, Martha Cyert, Paul Ehrlich, Gretchen Daily, and Cresson Fraley, for being my undergraduate mentors.

Wei Chialin, Edwin Cheung, Liu Jun, Lee Yen Ling, Zhao Bing, Vinsensius Vega, Patrick Ng, Lee Yew-Kok, and everyone else who has taught me.

Phillips Huang, Brenda Han Yuyuan, and Andrea Chavasse, for working with me, helping me, and letting me teach them.

Members of Genome Technology and Biology, especially Audrey Teo, members of Cancer Biology 3, and members of Information and Mathematical Sciences, for their friendship and help.

All paper coauthors and people who have contributed to this thesis in one way or another (names are not in any particular order): Herve Thoreau, Melvyn Tan, Yow Jit Sin, Dawn Choi, Low Hwee Meng, Eleanor Wong, Ong Chin Thing (Jo), Neo Say Chuan, Yap Zhei Hwee, Poh Tong Shing, Leong See Ting, Adeline Chew, Jeremiah Decosta, Alexis Khng Jiaying, Lim Kian Chew, Ruan Yijun, Wei Chia-Lin, Ruan Xiaoan, Edwin Cheung, Edison Liu, Audrey Teo, Phillips Huang, Han Yuyuan (Brenda), Andrea Chavasse, Liu Jun, Patrick Ng, Lee Yen Ling, Jack Tan, Yao Fei, James Ye, Lim Yan Wei, Isnarti Bte Abdullah, Haixia Li, R. Krishna Murthy Karuturi, Pan You Fu, Guillaume Bourque, Valere Cacheux-Rataboul, Wing-Kin (Ken) Sung, Hong-Sain Ooi, Mei Hui Liu, Han Xu, Vinsensius Vega, Yusoff Bin Mohamed, Pramila Ariyaratne, Peck Yean Tan, Pei Ye Choy, Yanquan Luo, K. D. Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Charlie Lee, Lusy Handoko, Sim Hui Shan, Axel Hillmer, Goh Yu Fen, Christina Nilsson, Zhang Yu Bo, Ngan Chew Yee, Christine Gao, Andrea Ho, and Poh Huay Mei Chiu Kuoping, Roy Joseph, Yew Kok Lee, Kartiki Desai, and Jane Thomsen.

The GIS community, for support and friendly advice.

∞o∞o∞o∞o∞o∞

To my parents

∞o∞o∞o∞o∞o∞

In memoriam: Guy Grazier G'Sell

∞o∞o∞o∞o∞o∞

Summary

Comprehensive understanding of functional elements in the human genome will require thorough interrogation and comparison of individual human genomes and genomic structures. In particular, one of the most important questions in gene expression regulation is how remote control of transcription regulation in a complex genome is organized. The Paired-End Tag (PET) strategy involves extraction of paired short tags from the ends of linear DNA fragments for ultra-high-throughput sequencing. In addition to new methods of constructing PETs, here I show a novel application of PET in understanding molecular interactions between distant genomic elements. Using this Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing method, I present the first-ever global estrogen receptor α-mediated human interactome chromatin map. I show that chromatin interactions are important in gene regulation. With its versatile and powerful nature, the PET sequencing strategies and the new application, ChIA-PET, have a bright future ahead.

List of tables

Table 1: PET technology applications for the study of genomes and transcriptomes. 3
Table 2: Analysis of GIS-PET library quality control measures. 41
Table 3: Identities of the Top 20 transcriptional units of each library. 44
Table 4. Statistics of library datasets used in this chapter 53
Table 5. Genes associated with ERα binding and interactions identified in previous studies and in this chapter. 66
Table 6. Statistics of overlaps between ChIA-PET library 1 and 2 interactions 68
Table 7. Statistics of inter-ligation PET clusters in all libraries 69
Table 8. Summary statistics of PET sequences and mapping to reference genome (hg18). 80
Table 9. Upregulated and downregulated genes near ERαBS. 100
Table 10. Association of ERα-mediated chromatin interactions with genes. 102

List of figures

Figure 1. Sequencing-based methods for understanding genetic elements in genomes. 5
Figure 2. Schematic view of PET methodology. 10
Figure 3. PET applications to address genome biology questions. 15
Figure 4. Schematic of a GIS-PET library prepared by the Selection-MDA method. 36
Figure 5. Full-length cDNA and GIS-PET library quality controls. 37
Figure 6. Data analysis method. 40
Figure 7. Analysis of length bias between the MDA approach and the bacterial amplification approach. 43
Figure 8. Differences between the GIS-PET method with classic approach and the GIS-PET method with the new Selection-MDA approach. 45
Figure 9. The ChIA-PET method. 49
Figure 10. ChIA-PET structures allow inference of self-ligation and inter-ligation status. 51
Figure 11. Control libraries. 55
Figure 12. The TFF1 positive control chromatin interaction. 58
Figure 13. The GREB1 (also known as KIAA0575) positive control chromatin interaction. 59
Figure 14. A novel chromatin interaction at CAP2. 60
Figure 15. ERα binding sites and interactions determined by ERα ChIA-PET. 62
Figure 16. ChIP-qPCR validation of new ERα binding sites identified by ChIA-PET. 63
Figure 17. Library sequencing saturation analyses. 67
Figure 18. Validation of ChIA-PET interaction data by ChIP-3C analysis. 71
Figure 19. 3C and ChIP-3C validation of a novel chromatin interaction at P2RY2. 73
Figure 20. Chromatin interactions and target gene expression. 75
Figure 21. Transcriptional activity at the GREB1 chromatin interaction locus. 76
Figure 22. A whole genome view of the human chromatin interactome map mediated by ERα binding. 81
Figure 23. Illustration of structural components of ERα-mediated interactions. 82
Figure 32. Different classes of involvements of ERαBS with chromatin interactions. 95
Figure 33. Numbers of ERαBS in different classes of interaction association. 95
Figure 34. Association of binding sites with interactions and genomic elements. 96
Figure 35. ERα-mediated chromatin interaction regions are associated with gene upregulation 103
Figure 36. Example of an enclosed anchor gene on chr 5 (CXXC5). 105
Figure 37. Example of an enclosed anchor gene on chr 2 (MLPH). 106
Figure 38. ERα-mediated chromatin interactions are required for transcription of estrogen-regulated genes. 109
Figure 39. A model for ERα function via chromatin interactions. 110

List of abbreviations and symbols

3C	Chromosome Conformation Capture
4C	Circular Chromosome Conformation Capture, or Chromosome Conformation Capture with chip
5C	Chromosome Conformation Capture Carbon Copy
ACT	Associated Chromatin Trap
BAC	Bacterial Artificial Chromosome
CAGE	Cap-associated Analysis of Gene Expression
cDNA	Complementary DNA
ChIA-PET	Chromatin Interaction Analysis using Paired-End Tag Sequencing
ChIP	Chromatin Immunoprecipitation
ChIP-3C	ChIP Chromosome Conformation Capture
ChIP-PET	Chromatin Immunoprecipitation with Paired-End Tags
ChIP-Seq	Chromatin Immunoprecipitation with Sequencing
DGS	Ditag Genome Scanning
DNA	Deoxyribonucleic Acid
ERα	Estrogen Receptor α
ERαBS	Estrogen Receptor α Binding Site(s)
EST	Expressed Sequence Tag
FDR	False Discovery Rate
FISH	Fluorescence *In-Situ* Hybridization
FlcDNA	Full-length cDNA
GIS-PET	Gene Identification Signature with Paired-End Tags
DNA-PET	Genomic DNA analysis with Paired-End Tags
GSC-PET	Gene Scanning CAGE with Paired-End Tags
GST	Genomic Signature Tags
iPET	Inter-ligation PET
mRNA	Messenger RNA
PAS	Polyadenylation Site
PCR	Polymerase Chain Reaction
PE-GST	Paired End Genomic Signature Tags
PEM	Paired End Mapping
PES	Paired End Sequencing
PET	Paired-End Tag
qPCR	Quantitative PCR
RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
siRNA	Short interfering RNA
STS	Short Tag Sequencing
SV	Structural Variant
TF	Transcription Factor
TFBS	Transcription Factor Binding Site(s)
TSS	Transcription Start Site
TU	Transcription Unit

Chapter One: Paired-End Tag Technologies

Introduction

Genomics holds much promise for huge improvements in human healthcare. However, genomics faces several practical challenges. Human genomes are read out as linear sequences, but in the cell, there are many complex interactions and mechanisms that operate around human DNA to transduce DNA information into biological function (Birney et al. 2007). Conventional DNA sequencing has been used to extensively explore genetic elements and structures (Birney et al. 2007); however, high sequencing costs and low throughputs have historically limited in-depth analysis of a broad range of genomic elements, making the development of new sequencing strategies necessary.

The Paired-End Tag (PET) sequencing strategy consists of extracting paired tags from the two ends of DNA fragments. The target DNA fragments may come from a variety of sources: cDNA reverse transcribed from mRNA, ChIP enriched DNA, and randomly sheared genomic DNA fragments. The end signatures, or "tags", consist of short DNA fragments (approximately 20-50bp) that are sequenced and mapped to the genome for accurate demarcations of the locations of the targeted DNA fragments in the genomic landscape.

The PET strategy has many benefits (Table 1). First, PET constructs can be easily sequenced by cheaper, massively parallel next-generation sequencing technologies. While these new technologies have much promise to transform biological exploration (Schuster 2008), they have shorter read lengths than Sanger capillary sequencing instruments and hence cannot sequence long templates (Wold et al. 2008). PETs are short enough to fit within this read length and yet contain sufficient information to identify the fragment through genome mapping. Another benefit is the higher mapping specificity of PETs over single tags. This is because PETs from long source fragments can span repeat regions which would otherwise lead to multiple, ambiguous mappings, as well as bridge unknown DNA sequences such as gaps in the genome assembly. Also, sequencing quality might drop as longer stretches are

sequenced, such that two sequenced tags each might have higher sequencing quality than a single sequenced tag that is twice as long. Hence, the PET sequencing strategy can double the amount of high quality sequencing data that can be obtained from a single template than might be otherwise possible using single tags. A further benefit is the decreased costs of sequencing a PET as opposed to sequencing one long single tag that spans the same genomic distance as the two ends of a PET, while retaining information regarding the defined distance and relationship between the two different ends. While just one end is insufficient to characterize a linear structure, a linear structure can be accurately and definitively defined using two points on either end. A caveat is that what is inside the linear structure, such as internal alternative splicing, would not be identified by PETs.

Table 1: PET technology applications for the study of genomes and transcriptomes.

Application	Benefits of PET	Techniques and References
General sequencing	PET template is compatible with next-generation machines Higher mapping specificity of PETs over single tags Decreased sequencing costs Retains information regarding the distance and relationship between the ends	Paired-End Tag (PET) (Ng et al. 2005; Wei et al. 2006) Paired End Sequencing (PES) (Holt et al. 2008; Lander et al. 2001) Paired End Mapping (PEM) (Korbel et al. 2007) Mate-pairs (Shendure et al. 2005)
Transcriptome	Identify 5' and 3' ends of transcription units Identify alternative TSS and PAS Enables ultra-high-throughput genome-wide identification of gene fusion events, which is not possible with other methods	GIS-PET (Ng et al. 2005) GSC-PET (Carninci et al. 2005)
TFBS and Epigenetic Sites	Improved specificity and demarcation of fragments containing sites of interest	ChIP-PET (Wei et al. 2006) Paired End Genomic Signature Tags (Dunn et al. 2007)
Chromatin interactions	Enables ultra-high-throughput genome-wide identification, which is not possible with other methods	ChIA-PET
Genomic structural variations	Paired readout of DNA sequence for accurate genome assembly Span repeats and gaps Enables ultra-high-throughput genome-wide identification of even small insertions, deletions and translocations, which is not possible with other methods	Ditag Genome Scanning (Chen et al. 2008a) DNA-PET Paired End Mapping (PEM) (Korbel et al. 2007) Paired End Sequencing (PES) (Holt et al. 2008; Lander et al. 2001) Mate-pairs (Shendure et al. 2005)

PET technology has been applied to the characterization of genetic elements and structures (Table 1). The advantages of PETs for transcriptome characterization are the abilities to quantitatively detect transcripts, detect transcript start and end points simultaneously, and identify fusion transcripts. When applied to the characterization of fragmented genomic DNA of a specific size, PETs can help to identify misassemblies and structural variants as well as provide valuable genome sequence data. Genomic regions containing repeats that cannot be independently mapped can be oriented and positioned by their connectivity to sequence-specific regions As applied to the analysis of chromatin, PETs can be used to identify transcription factor binding sites and epigenetic marks, as well as interactions between genetic elements.

In the future, PET technologies will continue to improve and expand to cover a greater range of applications in medical genomics. Eventually, PET technologies may help to overcome the challenges of personal genomics to make personal genomics a reality. Here, I provide a retrospective of the development of the PET sequencing strategy and its recent applications in transcriptome, epigenome, interactome and genome structure analyses. I also discuss the challenges faced by PET technologies. In this thesis, I propose several new solutions that may be offered by further developments in PET technologies, for answering novel biological questions.

The development of the Paired-End Tag (PET) strategy

The intellectual traces of the development of this PET strategy converged from two important technological concepts: conventional paired end sequencing and short tag sequencing (Figure 1).

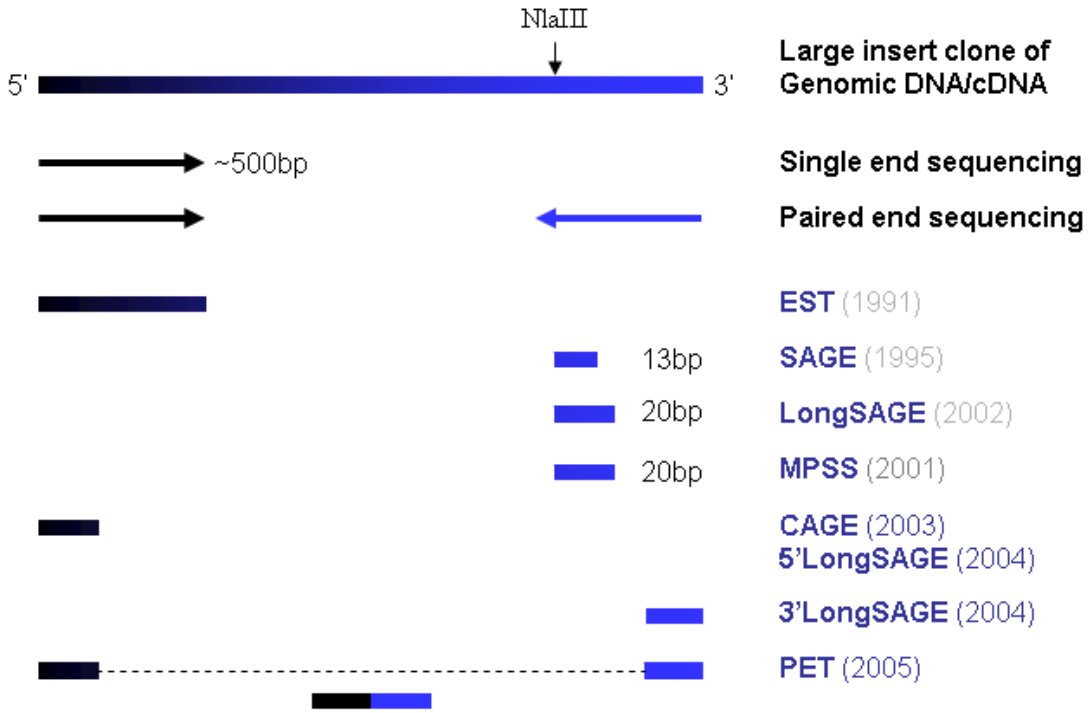


Figure 1. Sequencing-based methods for understanding genetic elements in genomes.

DNA fragments can be read from one end (single end) and/or both ends (paired end). EST was the first tag-based approach, generating one tag per sequencing read, used for characterizing expressed genes. The original SAGE tag was 13bp, and used for tagging transcripts. SAGE tags are concatenated for sequencing analysis with increased efficiency of 20-30 tags per sequencing read. LongSAGE and MPSS using MmeI as the tagging enzyme to generate 20bp tags that can be specifically aligned to reference genome sequences. The CAGE and 5' SAGE tags are derived from the 5' end of DNA fragments. 5' and 3' Long SAGE tags are derived from the two ends of DNA fragments, and can mark the 5' end or 3' end of the represented DNA fragments. PET combines the 5' and 3' signature tags of the same DNA fragment covalently into one ditag unit. When mapped to a reference genome sequence, a PET sequence can demarcate the boundaries of DNA elements in the genome landscape.

The first straightforward description of Paired End sequencing was reported by Hong (Hong 1981) using DNA inserts cloned into bacteriophage vectors and sequenced from both ends, thus reading twice as much sequencing data from long inserts. Then, in 1994, so-called "mate-pairs" consisting of sequencing reads from both ends of 2kb and 16 kb DNA inserts were used to help assemble the genome of *Haemophilus influenzae*, which was the first genome of a free-living organism to be sequenced (Fleischmann et al. 1995). Turning to

larger genomes, paired end sequencing was an important component of early proposals (Venter et al. 1996; Weber et al. 1997) and actual sequencing efforts such as the Drosophila genome, the public and the Celera human genome sequencing efforts (Adams et al. 2000; Lander et al. 2001; Myers et al. 2000; Rubin et al. 2000; Venter et al. 1998) . Later efforts to close up gaps in assemblies also employed paired end sequencing (Bovee et al. 2008). The benefits of paired end sequencing were similar to PETs, and in addition, cost savings from sequencing both ends of a plasmid prep rather than sequencing two single ends from two different plasmid preps could be substantial. Recently, many studies have employed paired fosmid (Kidd et al. 2008; Tuzun et al. 2005) or Bacterial Artificial Chromosome (BAC) end sequencing (Volik et al. 2006; Volik et al. 2003) to uncover structural variations in individual human genomes as well as chromosomal aberrations in cancer genomes. However, conventional Paired End Sequencing requires laborious cloning and expensive sequencing as it typically involves two full Sanger sequencing reads per Paired End Sequence.

The "chromosome jumping" method introduced by Collins and Weissman in 1984 was a novel approach that did not simply perform paired end sequencing from both ends of an insert, but instead first cloned the junctions formed by circularized ligation of the two DNA ends of large fragments, and then sequenced the junctions to reveal the two paired end sequences of large DNA segments (Collins et al. 1984). As this "chromosome jumping" method creates physical junctions between the two paired ends, it can be thought of as a direct precursor to later PET techniques which rely on the creation of physical junctions between two paired ends. The "chromosome jumping" method was designed to enable big "jumps" of hundreds of kilobases of DNA, as opposed to little "steps" across the genome, to aid positional cloning of disease genes. High molecular weight DNA was circularized under dilute ligation conditions to include a marker gene, such that the two ends of the DNA fragment were connected to the two sides of the marker gene. Digestion with another restriction enzyme would generate shorter DNA fragments, some of which consist of the junctions between the marker gene and the two DNA ends from the large fragments. These

shorter DNA fragments including the junction constructs were cloned into vectors for selection of the marker gene. Junctions containing the DNA of interest as well as DNA from a large jump away could be isolated and sequenced. This method was applied to efforts in cloning the disease gene for cystic fibrosis (Collins et al. 1987).

Around the same time, short tag methodologies were developed to overcome the prohibitively high costs of sequencing. The idea behind short tags was that not all of a DNA fragment had to be sequenced to identify it: a sequenced short tag from a particular fragment could be mapped to the reference genome, thus revealing the identity of the fragment. Expressed Sequence Tags (EST) were the first example of the tag-based sequencing concept, by using single direction Sanger sequencing reads to tag cDNA sequences reverse transcribed from mRNA, instead of sequencing full length cDNAs (Adams et al. 1991; Milner et al. 1983; Putney et al. 1983). Many cDNA libraries were characterized by EST sequencing, which led to the discovery of many genes (Adams et al. 1992) and the characterization of cancer transcriptomes (Brentani et al. 2003). Despite instant success and recognition, the high costs both in time and in resources for DNA sequencing promoted the desire to further shorten the sequenced tags, leading to the development of Serial Analysis of Gene Expression (SAGE) (Velculescu et al. 1995), and Massively Parallel Signature Sequencing (MPSS) (Brenner et al. 2000). In SAGE and MPSS, a special type of restriction enzyme, called a "tagging" enzyme, is employed. The tagging enzyme cuts DNA at a certain distance away from the restriction enzyme site. Examples include type IIS restriction enzymes (Velculescu et al. 1995). Adaptors with flanking tagging restriction enzyme sites are attached to the target DNA, and then libraries of short SAGE or MPSS tags are created by cutting these constructs with the type IIS restriction enzyme, thus resulting in a population of tags from different fragments (Velculescu et al. 1995). Because only short tags which represent a complete RNA fragment need to be sequenced as opposed Expressed Sequence Tags (ESTs) (Adams et al. 1991), the costs of sequencing SAGE tags to a depth necessary to adequately characterize transcriptomes are much lower than EST, and in turn flcDNA experiments. LongSAGE

featured the type IIS restriction enzyme, MmeI, that can cut DNA 18/20bp downstream of its recognition site, to produce 20bp SAGE tags, and was used for *de novo* identification of expressed genes (Saha et al. 2002). By contrast, the original SAGE method used enzymes that cut shorter tags, which often could not be mapped uniquely to the genome. The SuperSAGE method, introduced later, used the type III restriction enzyme, EcoP15I, which cuts 25/27bp downstream of its recognition site, allowing for the extraction of even longer SAGE tags (Matsumura et al. 2003). However, EcoP15I only cleaves head-to-head orientated recognition sites in supercoiled DNA, and does not turnover (Raghavendra et al. 2005). However, recently, it has been shown that the incorporation of sinefungin into EcoP15I allows cleavage at all recognition sites regardless of DNA topology (Raghavendra et al. 2005). In addition, prior methylation of EcoP15I sites within the target sequences prevents these internal EcoP15I sites from being cut and thus reducing the effective concentration of EcoP15I in the reaction. Taken together, these new results show promise in making EcoP15I a useful laboratory tool. The 27bp tags generated by this enzyme will be very useful for improving short tag mapping rates and mapping accuracies.

Besides extracting tags near the 3' side of cDNA fragments, SAGE and MPSS methods have been used in many other applications, including digital karyotyping (Dunn et al. 2002; Wang et al. 2002), mapping ChIP-enriched DNA fragments to identify transcription factor binding sites (Bhinge et al. 2007; Kim et al. 2005), and DNAseI-digested DNA to identify DNAseI-hypersensitive sites (DACS) (Sabo et al. 2004a; Sabo et al. 2004b). In order to characterize 5' transcription start sites and hence identify gene promoters, Cap Analysis of Gene Expression (CAGE) was introduced based on the Cap-trapper method (Carninci et al. 1999) to retain 5' intact transcripts for cDNA synthesis with modified linkers containing the type IIS restriction enzyme recognition sequence at the 5' ends, followed by enzymatic digestion and the standard LongSAGE method on these 5' CAGE tags (Shiraki et al. 2003). Two other groups (Hashimoto et al. 2004; Wei et al. 2004) also independently developed similar approaches such as 5'LongSAGE to map transcription start sites and infer the

locations of gene promoters. In addition, the companion 3'LongSAGE method was simultaneously developed, so as to map both 5' transcription start sites and the exact 3' polyadenylation sites to define the boundaries of expressed genes using two end tags as opposed to a single tag (Wei et al. 2004). Expanding from such a capacity, the Paired-End Tag (PET) method that covalently links the 5' tag and 3' tag of a DNA fragment into a ditag structure for cost-efficient sequencing analysis of linked structures was then developed (Ng et al. 2005).

Construction of PET structures

Construction of a PET structure is necessary because many next generation technologies are only compatible with short templates in specific formats. Hence, libraries need to be prepared which covalently link the two DNA ends to each other, remove the rest of the DNA, and adapters containing priming sites for universal primers need to be incorporated into the PET structure (Figure 2).

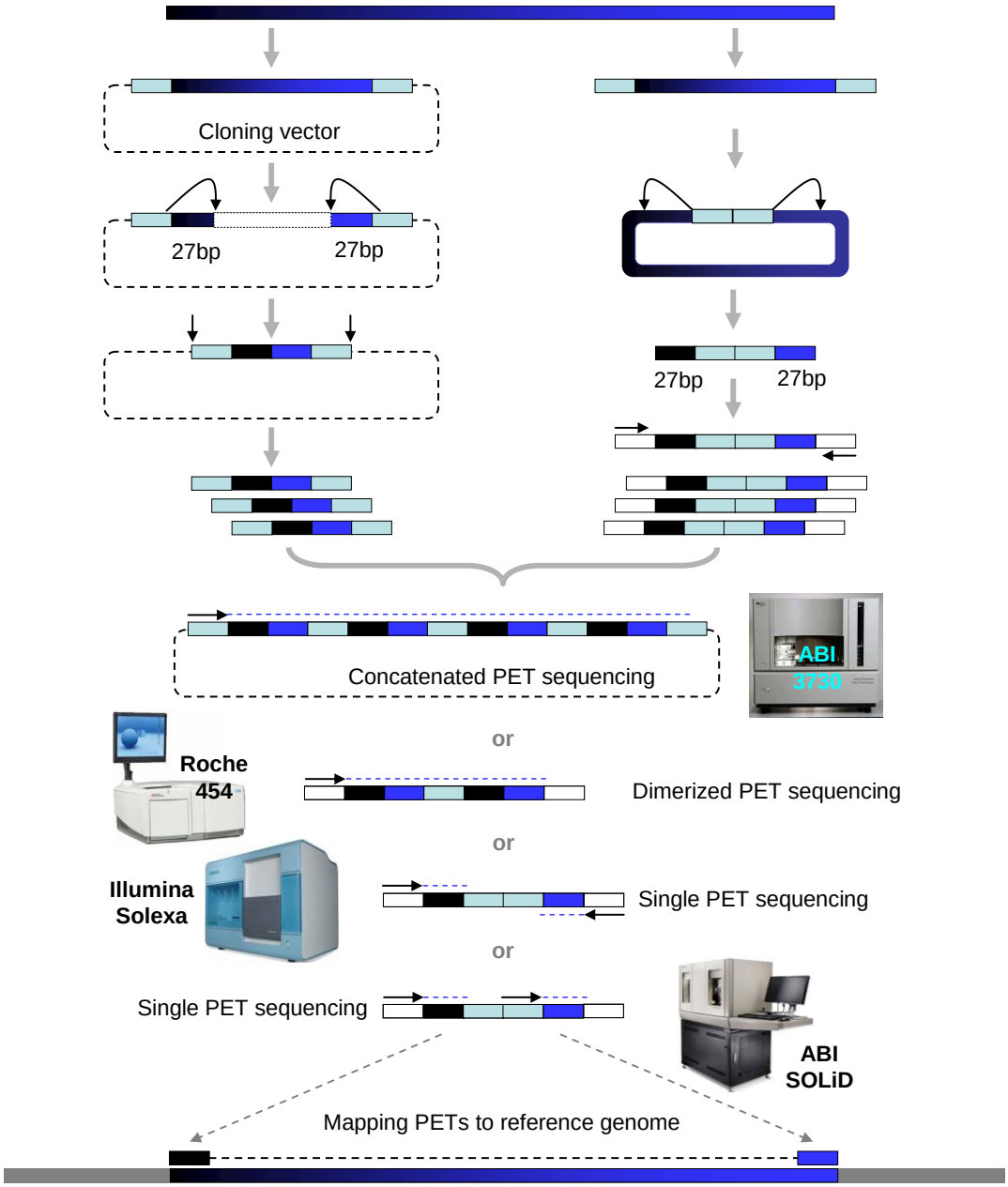


Figure 2. Schematic view of PET methodology.

The PET concept is the extraction of paired end signatures from the ends of target DNA fragments. These end signatures, or "tags" are short DNA fragments that are sequenced and mapped to the genome for the accurate demarcations of the locations of the targeted DNA fragments in the genomic landscape. The PET method may be carried out through cloning-based or cloning-free procedures. The PET structures may be analyzed through high-throughput sequencing of clones containing concatemers of tags using conventional Sanger capillary sequencing instruments or diPET constructs using 454 GS20/GS FLX or single PET constructs using Illumina GA/GAII and ABI SOLiD. The sequenced PETs can then be mapped to the reference genome for the identification of genetic elements.

The original PET method was a "cloning-based" approach: it used plasmid vectors to link 5' and 3' tags. It was implemented as Gene Identification Signature analysis using PETs (GIS-PET) for studying transcriptomes, in which the starting mRNA was converted into full-length cDNA with flanking adaptor sequences containing MmeI restriction sides immediately next to both cDNA ends. The full-length cDNA fragments were then ligated to linearized plasmids, and transformed into *Escherichia coli (E. coli)* cells as a full-length cDNA library. The purified plasmids of this full-length cDNA library are then digested with MmeI, which cuts into the cDNA insert to result in two 18/20bp tags attached to the vector backbone. The tag-vector-tag structures are gel-purified and re-circularized under intra-molecular ligation conditions, so that the two tags are joined covalently. The resulting single PET library can be amplified in bacteria cells and the PET constructs are then excised by a restriction digestion from purified PET library plasmids (Ng et al. 2007). A similar strategy was applied to characterize ChIP enriched DNA fragments for genome-wide identification of transcription factor binding sites in human cancer cell genomes (Wei et al. 2006) and mouse embryonic stem cell genomes (Loh et al. 2006). The strategy has been since extended to epigenetic modifications (Dunn et al. 2007; Zhao et al. 2007).

We and others (Shendure et al., 2005) developed a linker-based methodology (further described in Chapters 3 and 4). This methodology involves direct circularization of the target DNA fragments with linker oligonucleotides that covalently join the two ends of a DNA fragment. As the linker sequence linker sequence is typically designed to contain two MmeI or EcoP15I sites flanking the two ends of the circularized DNA fragment, restriction digestion with these enzymes would release the tag–linker–tag structure for sequencing. This strategy was first demonstrated in resequencing an *E. coli* genome using the polony sequencing method (Shendure et al., 2005). Besides tagging enzymes such as MmeI and EcoP15I that generate uniform sizes (18/20 bp and 25/27 bp) of PET constructs for easy manipulation, frequently cutting restriction enzymes and physical shearing by nebulization are also choices for generating randomly sized tag–linker–tag constructs. As reported (Korbel

et al., 2007), circularized DNA was randomly sheared by nebulization, and the fragments with biotinylated linkers were isolated using streptavidin. This method produces tags with a median size of 106 bp and is very useful for obtaining long tags because no type IIS or III restriction enzyme is currently known to produce tags more than 30 bp; however, many PETs prepared this way are unbalanced with tags of lengths under 15 bp, which would mean that these sequences would have to be discarded.

A benefit of the cloning-based method is that it preserves the original full-length cDNA or ChIP DNA fragments in a sustainable format of library clones. However, the construction process is long (2-4 weeks) and can be technically challenging. By contrast, the cloning-free method is rather straightforward and can avoid many biases related to cloning. In both cases, care needs to be taken to ensure that every step is done efficiently and accurately, such that the resulting libraries are accurate and of high complexity. If the library has low complexity, which might happen if too many PCR cycles are used to amplify the DNA, many redundant sequencing reads will be obtained.

Sequencing analysis of PET constructs

Here I review the multiple sequencing options for PET constructs (Figure 2), focusing on the specific method and benefits of each sequencing technology (Holt et al. 2008) with respect to PET sequencing.

PETs can be sequenced by Sanger sequencing. PETs can be concatenated into long stretches of DNA followed by cloning into a sequencing vector. An average Sanger sequencing read of several hundred base pairs would read out 20-30 PETs. This concatenation sequencing strategy was applied to PET sequencing with great success, demonstrating the value of PETs for transcriptome analysis (Ng et al. 2005) and genome functional analysis (Loh et al. 2006; Wei et al. 2006). However, the costs of conducting PET

experiments were still relatively high due to the high costs involved in DNA sequencing using conventional sequencing platforms.

One of the first successful next-generation sequencing methods was published in 2005 (Margulies et al. 2005) by 454 Corporation. In 2006, when it was first introduced to the research community, the GS20 instrument could generate about 200,000 sequence reads with average read lengths of approximately 100bp. It was straightforward to sequence single PET templates of about 40 bp with 454/Roche pyrosequencing. However, such an approach cannot fully utilize the sequencing capacity of each GS20 read; hence, we conceived a one-step ligation method to allow two units of PET constructs ligate to one another and to form a diPET template that is approximately 80 bp, perfectly fitting within the read length of the GS20 pyrosequencer. Using this approach we instantly doubled the output of GS20 for PET sequencing (Ng et al. 2006a). A single run of diPET templates in 4-hour of GS20 machine time can generate half a million PET sequences. This advance represented an immediate 100-fold increase in efficiency for PET sequencing when compared to the use of Sanger sequencing method to read PET concatemer clones which requires more than a month (Ng et al. 2006a).

Towards the end of 2006, the Illumina Genome Analyzer (GA) sequencing machine was introduced to the market. The most impressive feature of this method is its massively parallel capacity for reading up to 80 million DNA template clusters simultaneously, even though it reads only approximately 36-50bp from each template (Barski et al. 2007; Johnson et al. 2007). There are three ways to use the Illumina platform to obtain PET information. First, a PET construct can be read from both directions, one at a time, to cover the two tags in a PET construct, respectively. One strand of the PET template is read from one direction, the second strand is synthesized *in situ* to replace the first strand, and then read from another direction. The second way is simply to sequence the entire length of the PET construct using the improved GAII's maximum read length of 50 bp. A third way is to bypass the

construction of the PET, and simply sequence paired ends from the DNA of interest using the two directional sequencing method wherein one strand of a template of less than 1 kb is read from one end to give one tag, and then the second strand is synthesized *in situ* to replace the first strand, and then read from another direction to give the second paired tag (Campbell et al. 2008). This last method requires the least effort in constructing the library, but is limited to the analysis of short DNA fragments. Bridging repeats and gaps is difficult using short DNA fragments.

SOLiD is another massively parallel short tag sequencing platform introduced in late 2007 by Applied Biosystems. This sequencing platform was adapted from the polony ligation-based sequencing method (Shendure et al. 2005). The current version of SOLiD is designed for paired end sequencing, and can read about 200 million tags for 25bp from each end per machine run in two weeks of time.

After sequencing, the PETs have to be mapped to a suitable reference genome (Figure 2). The millions of PET sequences generated from each machine run have imposed immense challenges on how to efficiently process the data and accurately map the PET sequences to reference genomes. The companies that are developing the new sequencing technologies have been also developing software for base calling and tag mapping. More efforts in this area would be expected from end users as well as bioinformatics-based companies. To process PET sequences specifically, we developed PET-Tool, a user-friendly software package that does all steps, from PET extraction from raw sequence reads, to mapping the PET sequences to reference genomes, as well as provide a management system for hosting different PET experimental datasets (Chiu et al. 2006). PET-Tool maps efficiently using compressed suffix arrays , such that searching the human genome is within the capabilities of personal computers (Hon et al. 2007). A different method was described by Korbel et al., which uses a fleet of over 400 multiple processors employing Megablast in the first pass analysis and then the Smith-Waterman sequence alignment methods for further

refinement (Korbel et al. 2007). Roche/454 GSFLX has developed its own *de novo* genome assembler (*GS de novo assembler*) and mapping algorithm (*GS reference mapper*) which are capable of taking a combination of Sanger long reads, 454 shotgun and paired end reads to generate contigs and scaffolds or map to a reference genome. ELAND (Efficient Large-Scale Alignment of Nucleotide Databases) and SXOligoSearch (http://www.synamatix.com) were developed by Illumina and Synamatix for aligning Illumina short tag reads to mammalian genomes quickly and accurately. These different methods use the same stringency (up to 2 mismatches), and closely agree in terms of performance and time. Furthermore, Illumina and SOLiD have now independently developed pair end analysis pipelines for analyzing PETs based on their mapping coordinates and orientations.

In summary, the steps of the PET technique have been well developed, from PET construction to sequencing and data analysis. In the following sections, we review the applications of the PET technology in genome analysis and future perspectives (Figure 3).

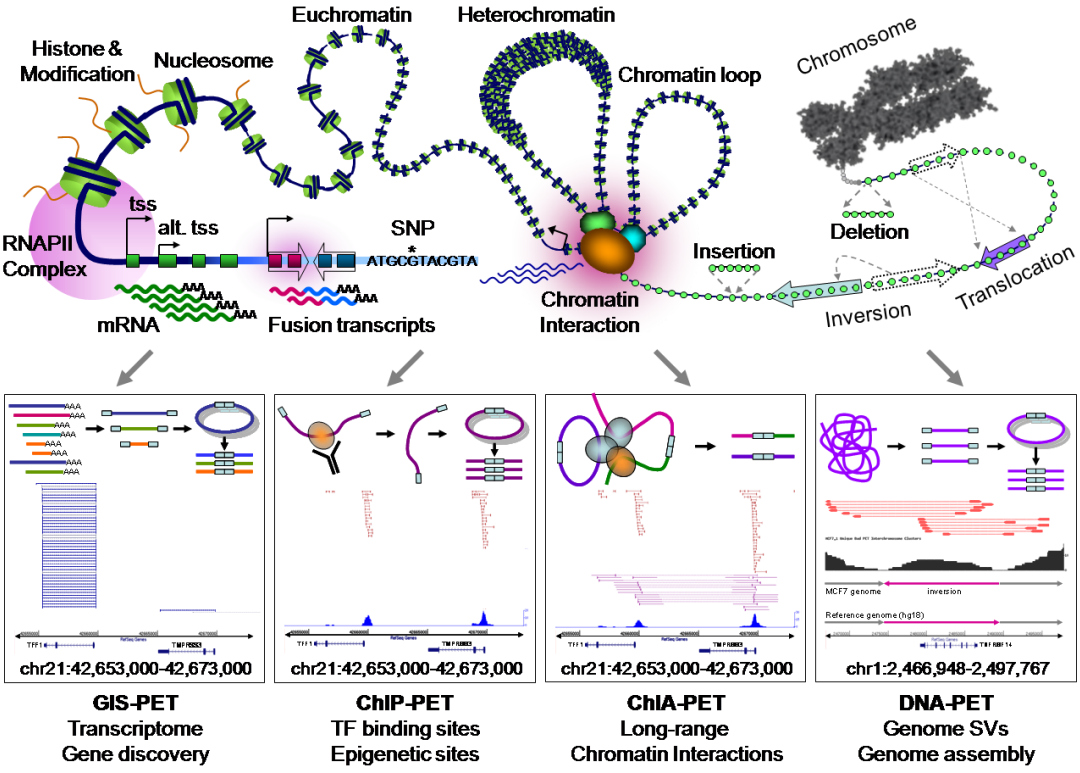


Figure 3. PET applications to address genome biology questions.

The cell has many different mechanisms for modifying, controlling, and transducing information encoded in the genome. The PET technology can be applied to investigate many questions regarding nuclear processes, such as transcriptomes, transcription factor binding sites, epigenetic modification sites, long range chromatin interactions, regulation mechanisms in 3-dimensional spaces, and genome structural variants (SVs). Examples of PET data from GIS-PET, ChIP-PET and ChIA-PET experiments of human breast cancer MCF-7 cells with estrogen induction treatment at the TFF1 locus (chr21:42,653,000-42,673,000) are shown: the high level of expression of the TFF1 gene and the low level of expression of the TMPRSS3 gene, the ERα binding at the TFF1 promoter sites and enhancer site, and the interaction of these two ERα binding sites. An example of DNA-PET data at the TNFRSF14 locus in the genome of MCF-7 cells shows an inversion event detected by two clusters of discordant DNA-PET mapping.

Insights from PET applications to transcriptome studies

Transcriptome studies include understanding gene structures encoded in the genome and gene transcription dynamics (Figure 3). The structural elements of genes include exons, introns, transcription start sites (TSS), polyadenylation sites (PAS) and transcription end sites. The gold standard for uncovering gene structure is the use of flcDNA sequencing to obtain complete gene structure information (Carninci et al. 1999). However, this is a very expensive and laborious approach. Whole genome tiling arrays have proved effective for identifying exons and measuring transcription dynamics (Birney et al. 2007; Kapranov et al. 2002); however arrays can be ambiguous in defining the exact boundaries of transcription units particularly in gene dense regions, because arrays lack connectivity information between exons identified by array hybridization. Mono-tag based approaches such as CAGE or 5'SAGE are effective in defining and quantifying alternative usage of transcription start sites, but only transcription start sites and no other aspects of gene structure (Hashimoto et al. 2004; Shiraki et al. 2003). Recently, shotgun sequencing of transcripts (RNA-Seq) has been used to profile genes, and has generated an unprecedented wealth of information about gene structures, particularly alternative splicing (Marioni et al. 2008; Morin et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Sultan et al. 2008; Wilhelm et al. 2008). However, as RNA-Seq requires many reads to characterize a transcript, it is rather expensive, even with the use of next-generation sequencing methods.

By contrast, the GIS-PET approach is a high-throughput method most suited for efficiently demarcating the boundaries of transcription units and defining transcription start sites and polyadenylation sites (Figure 3). The GIS-PET method is uniquely able to detect unconventional fusion genes because GIS-PET reads out the sequences of paired 5' and 3' ends from the same transcript, thereby delineating the relationship between two ends of the mRNA transcript. Human cancer cell lines are known to contain extensive chromosomal aberrations. Fusion genes created through chromosomal rearrangements could play roles in oncogenesis. Several successful diagnostic methods and therapies target fusion gene products (Mitelman et al. 2007); for example, Gleevec targets the *BCR/ABL* fusion in chronic myelogenous leukemia (Mauro et al. 2002). Although GIS-PET is very efficient and accurate in identifying the first and last exons of transcription units, an obvious limitation is that it does not generate information regarding internal exons. GIS-PET is therefore a complementary tool to tiling array RNA data and RNA-Seq.

In GIS-PET, flcDNA is prepared using the PET method: the capped 5' ends and the polyA-tailed 3' ends are captured in a pairwise manner by 20bp signature tags, and these paired end sequences may then be mapped to the genome, allowing the complete transcriptional unit to be inferred from the genome sequence in between the paired 5' and 3' tags. GIS-PET is designed to contain a residual AA dinucleotide from the mRNA polyA tail that indicates the orientation of the PET. In the Gene Scanning CAGE variant (GSC-PET), the PET sequences are generated from normalized flcDNA libraries in which highly abundant cDNA clones are removed, thus enriching for rarer clones, and hence allowing for more efficient discovery of rare genes (Carninci et al. 2005).

GIS-PET has been applied to the studies of transcriptomes in E14 mouse embryonic stem cells (Ng et al. 2005), various mouse tissues as part of the FANTOM3 project (Carninci et al. 2005), and a number of human cells as part of the ENCODE project (Consortium 2004). Many isoforms of transcripts with alternative transcription start sites and polyadenylation

sites were characterized, and large numbers of novel transcription units were identified. From E14 mouse embryonic stem cells, a trans-splicing fusion mRNA between *Ppp2r4* and *Set* was found, in which the first exon of *Ppp2r4* was joined to the second exon of *Set*. This fusion gene is preferentially expressed in embryonic as opposed to adult tissues, and the fusion gene might encode a new functional protein, suggesting that the fusion might play a role in early development in mice (Ng et al. 2005). Additionally, two human cancer cell lines, MCF-7 (breast cancer) and HCT116 (colon cancer), were characterized with GIS-PET to understand unconventional fusion transcripts in cancer cells (Ruan et al. 2007). From an analysis of 865,000 GIS-PETs from MCF-7 and HCT-116, 70 fusion genes were found including a fusion between *BCAS3*/*BCAS4* that had been previously identified in MCF-7 cells. Other fusion genes identified and validated by RT-PCR included *CXorf15/SYAP1* and *RPS6KB1/TMEM49* (Ruan et al. 2007). Interestingly, *SYAP1* has been implicated in chemotherapy response (Al-Dhaheri et al. 2006), and *RPS6KB1* is an oncogenic marker (van der Hage et al. 2004), suggesting a possible role for these fusion genes in cancer progression.

In conclusion, GIS-PET is the most efficient and accurate approach to demarcate the boundaries of transcription units of genes and complements other methods for transcriptome studies. The most unique benefit of GIS-PET is that it is the only efficient system for large scale investigations of unconventional fusion gene transcripts. A large scale GIS-PET program to investigate unconventional fusion gene transcripts could lead discovery of new candidates as biomarkers for diagnostic and therapeutic options.

Insights from PET applications to genome structure analysis

Genomes are variable at both nucleotide level and large structural levels (Figure 3). Genome variations at nucleotide level such as SNPs and mutations are well understood to have functional roles in normal traits and diseases (Shastry 2007). However, our understanding of large structural variations in the human genomes is very limited. SAGE-based digital karyotyping (Dunn et al. 2002; Wang et al. 2002) and array comparative genomic hybridization (a-CGH) (Pinkel et al. 1998) have contributed to this field by identifying large

chunks of deletions and assessing copy numbers of amplified regions in disease genome compared to normal or reference genome. However, both the mono-tag based sequencing approach and a-CGH cannot identify balanced structural variations such as insertions, inversions, and translocations in genome rearrangement. Although paired end sequencing of large genomic DNA inserts in fosmid and BAC clones using conventional sequencing technique have been used to generate highly valuable information regarding human genome structural variations (Kidd et al. 2008; Tuzun et al. 2005), the costs of such efforts is prohibitive.

DNA-PET is an ideal method for sequencing and assembling genomes as well as studying genome structural variations (Korbel et al. 2007). DNA-PET provides linked 5' and 3' tag sequences from genomic DNA fragments of specific sizes, for example, 400bp (Campbell et al. 2008) or 3 kb (Korbel et al. 2007) (Figure 3). To accomplish this, genomic DNA is sheared by nebulization and purified to a specific size range. Paired end 5' and 3' tags are then obtained from the genomic DNA fragments, which are then sequenced and mapped to the reference genome sequence to infer the size of DNA fragments. Most PET sequences would match well to the reference genome with correct orientation and specific size range. PETs with discordant mapping orientation and distance between the two tags would be located at the breakpoints of structural variations between the reference genome and the genome under study.

The DNA-PET method was first demonstrated in resequencing an evolved *E. coli* genome using the polony sequencing-by-ligation method (Shendure et al. 2005). The early polony sequencing method was very limited in terms of tag lengths (6-7 bp), but because a PET structure contains 4 different places for sequencing to begin (1 end from the left, 2 ends from the center linker region, and 1 end from the right), the PET structure allowed for the acquisition of approximately 26 bp per amplicon. In addition to high PET mapping accuracy,

Shendure et al. found nucleotide changes and genomic rearrangements that had been engineered into the sequenced genome (Shendure et al. 2005).

In an effort to study human genomic structural variation (Korbel et al. 2007), genomic DNA from an African and a European individual were sheared into 3 kb fragments, PETs of the DNA fragments were sequenced by 454/Roche, and the PET sequences were mapped to the reference human genome. Simple deletions were predicted from PET mapped spans that were much larger than 3 kb, and simple insertions were predicted from PET mapped spans that were much shorter than 3 kb, while inversions were predicted from altered end orientations. More complex structural variations were also found from PET mapping patterns that did not match expected mapping patterns. Through this analysis, 1,297 structural variations were found. 45% of structural variations were shared between the two individuals, suggesting that some structural variations might be common. Hotspots of structural variations were found, which turned out to be regions that have been found to be involved in genomic disorders. Additionally, many structural variations could affect gene functioning by either removing exons, creating gene fusions, being present in introns, altering gene orientation, or by amplifying the genes. Interestingly, genes with protein products that were associated with interactions with the environment contained more structural variants than expected by chance (Korbel et al. 2007). This observation suggests a possible role for differences in these genes in order to cope with differences in environments.

The DNA-PET approach has also been applied to map cancer genome variations (Campbell et al. 2008). The authors took an even simpler approach to generate PET sequences from two cancer cell line genomes, in genomic DNA was sheared to an average size of 200 bp, isolated, and 29-36bp at either end were sequenced by Illumina paired end sequencing methods. About 7 million PET sequences from each of the two cell lines were uniquely mapped to reference genome and more than 400 rearrangements were identified to base pair resolution. Because of the high density of the tag sequence data, accurate copy

numbers of amplified regions in the human cancer genome were also obtained. Further analysis of the data allowed the authors to identify 103 somatic rearrangements and 306 germline structural variations. It has suggested that many somatic variations are associated with amplicon regions of the genome, while most germline rearrangements are mediated by retrotransposition elements such as AluY and Line. This work demonstrates the feasibility of systematic genome-wide efforts to characterize the architecture of complex human cancer genomes. It should be noted that the authors had to discard 48% of the sequenced reads as they did not map to the reference genome. These results suggest that inefficiencies in the library construction steps, or the new Illumina Paired End sequencing method, reduced the amount of data that might otherwise have been obtained from the sequencing run. Moreover, of the reads that did map well, the authors excluded 38% because they precisely duplicated other sequences from the same library. The authors suggest that these sequences might have been preferentially amplified during the PCR step. Increased amounts of starting genomic DNA, reduction in the number of PCR cycles used, and PCR amplification of the entire ligation mix as opposed to a small aliquot, are measures that could increase the complexity of the resulting library. In addition, care should be taken during library preparation such that all steps go to completion, to ensure that the resulting library is of high quality.

Recently, a variation of DNA-PET called Ditag Genome Scanning (DGS) used restriction enzymes to digest the genomic DNA instead of shearing. As a proof of principle, Chen et al. applied this method to the study of normal human GM15510 and human leukemia Kasumi-1 DNA, and demonstrated that DGS could uncover DNA fragments that vary from the reference human genome sequences (Chen et al. 2008a). The use of restriction enzymes has the advantage of higher mapping rates as well as faster mapping times (minutes on a regular desktop computer) because a smaller database consisting of sequences near the particular restriction enzyme site can be used as a reference (Chen et al. 2008a). However, a limitation is that structural variants or regions of the genome that are not near any restriction enzyme sites cannot be analyzed. Multiple libraries may be constructed using different

restriction enzymes to circumvent this problem, but this approach also increases the laboriousness of the procedure.

The power of connectivity provided by DNA-PET may be used to facilitate the assembly of whole genome shotgun sequence reads for *de novo* genome sequencing and resequencing. With the current dramatic increase of DNA sequencing capacity, getting enough coverage of shotgun reads is no longer a serious issue. Using the massively parallel short tag sequencing platforms, 10-20X fold base-pair coverage of a human genome can be generated with a fairly small budget and within weeks. However, assembling such short tag sequences alone would result in large numbers of contigs that cannot be joined up with each other. The real challenge is how to connect and orientate these contigs into the complete assembly of a complex genome such as the human genome. DNA-PET experiments (Campbell et al. 2008; Korbel et al. 2007) and computer simulations (Shendure et al. 2005) suggest that short tag (20-30bp) PET sequences could be used for *de novo* complex genome sequencing.

A critical aspect in developing such a DNA-PET based strategy is the construction of PETs for large DNA insert fragments, such as 10 kb or even 100 kb fragments. One reason for this is that mammalian genomes have many repeat elements that are greater than 3 kb long. PETs that are longer than the repeat element length are needed to assemble fragments, by crossing over the repeated elements. Another reason is that longer DNA fragments will enable the discovery of insertions and translocation events greater than 3 kb, which is the upper limit of the current DNA-PET approaches. In our lab, we are able to generate PET sequences from up to 15Kb genomic DNA inserts. Our preliminary data shows that large insert DNA-PET is clearly better than short insert DNA-PET, because large insert DNA-PET gives higher physical coverage. *In silico* analyses support this finding: as the length of the insert DNA increases, the physical coverage increases, and hence the probability of detecting a fusion point increases (Bashir et al. 2008). With these improvements, the DNA-PET

method combined with ultra-high-throughput sequencing platforms will become a very powerful strategy for *de novo* genome sequencing and individual genome resequencing. Just as the human genome sequencing experiments were performed with paired end sequences from inserts of multiple sizes, a combination of multiple DNA-PET sizes could be useful in resequencing the human genome as well as in *de novo* sequencing. Small structural variants might be detected and small repeats might be crossed using 10 kb DNA-PET approaches, and large structural variants might be detected and large repeats might be crossed using 100 kb DNA-PET approaches. If this strategy proves successful, this development in DNA-PET will pave the way for personal genomic approaches to resequence many individual human genomes.

In conclusion, the DNA-PET strategy for genome structure analysis has immediate value and long term promise. Already, DNA-PET with the current sequencing capacity can provide comprehensive characterizations of human structural variations associated with genetic diseases. Further development of DNA-PET with improved speeds, reduced costs, and the ability to use clinical samples would create a new digital cytogenetics platform for clinical implementation. In the long term, DNA-PET can become a vital part in the concept of personal genomics for personal medicine.

Insights from PET applications to identify regulatory and epigenetic elements

Besides gene coding sequences, genomes contain many non-coding elements that have important regulatory functions through interaction with protein factors (Figure 3). Thus, mapping protein factor binding sites in the genome is an important starting point for understanding regulatory circuits. The traditional mainstream approach for mapping such protein/DNA interactions is ChIP-chip, a method in which chromatin is formaldehyde-fixed, sonicated to randomly fragment the DNA, and enriched for desired protein target regions by Chromatin ImmunoPrecipitation (ChIP). The enriched DNA fragments are then detected by whole genome microarray (chip) hybridization (Ren et al. 2000). Although ChIP-chip has had phenomenal success, array-based detection methods are limited to partial genome coverage

using tiled probes with gaps in between probes, and repeats and unhybridizable regions left out.

ChIP-PET represents the first serious sequencing-based alternative approach to characterize ChIP enriched DNA fragments (Wei et al. 2006) (Figure 3). The ChIP-PET method provides linked 5' and 3' sequences for ChIP-enriched DNA molecules, which are mapped to the genome such that the complete ChIP DNA fragment can be inferred from the genome sequence in between the 5' and 3' tags, and transcription factor binding sites can be determined. ChIP-PET analysis depends on several principles. First, as the chromatin is sonicated, the probability of generating exactly identical DNA fragments is low; hence any redundant PETs are considered to be copies amplified during the cloning and/or PCR amplification processes. Therefore, only nonredundant distinct PETs are used for further analysis. Next, while ChIP enriches for transcription factor binding sites, there is still a lot of non-specific noise in the ChIP DNA, as a result of nonspecific antibody binding. Hence, the "multiple overlaps" concept is used to distinguish true signals from noise. The principle of this concept is that we expect PETs derived from nonspecific fragments to be randomly distributed in the genome as background PETs, whereas PETs derived from the same ChIP-enriched transcription factor binding site will overlap with each other to form a cluster of PETs. The region of maximum PET overlap in this PET cluster is taken to define the transcription factor binding site at base pair level resolution (Wei et al. 2006). Further, some cell lines have amplified regions in their genomes as compared with the reference human genome. Amplified regions would be sequenced more and hence some amplified regions might be mistaken for binding sites when the sequenced enrichment is due to genome amplification rather than ChIP enrichment. Thus, a method was developed for making corrections on the basis of the numbers of non-specific fragment noise PETs (Lin et al. 2007).

The ChIP-PET method was used to examine p53 transcription factor binding sites in HCT116 colon cancer cells, and found 542 high confidence binding sites (Wei et al. 2006).

Over 99% of these high confidence binding sites could be verified by ChIP-qPCR validation experiments, and PET-defined binding regions could be narrowed down to as little as 10 bp. These binding sites are clinically relevant to p53-dependent pathways in primary cancer samples. Interestingly, in addition to 5' promoter proximal regions of genes, many transcription factor binding sites can be found in gene introns, 3' ends of genes, and also far away from any genes. However, no transcription factor binding sites were found in exons. This observation is statistically significant, and not due to random chance (Wei et al. 2006). ChIP-PET was then used to map whole genome binding profiles for a number of important transcription factors, including Oct4 and Nanog (Loh et al. 2006), cMyc (Zeller et al. 2006); ERα (Lin et al. 2007); and NF-KB (Lim et al. 2007). We also applied ChIP-PET to map epigenetic marks for epigenomic profiles of histone modifications in human embryonic stem cells (Zhao et al. 2007).

Recently, a similar method called Paired End Genomic Signature Tags (PE-GST) has been independently developed. It has been used to identify transcription factor binding sites in a similar manner as ChIP-PET, as well as DNA methylation patterns (Dunn et al. 2007). Cancer cells exhibit aberrant methylation, and further understanding of cellular methylomes could help in the development of new diagnostic and treatment modalities (Feinberg et al. 2006). To investigate 5' methylation of cytosine in CpG dinucleotides, Dunn et al. describe a method involving the digestion of genomic DNA using MseI, which cuts rarely in CpG islands. Following this, DNA containing methylated cytosines is enriched by affinity purification, and these fragments are then subjected to the PE-GST procedure (Dunn et al. 2007). Alternatively, the genomic DNA may be digested with SmaI, a methylation-sensitive restriction enzyme which only cleaves unmethylated CpG islands present in its recognition sequence (Toyota et al. 2002). These fragments are then subjected to the PE-GST procedure (Dunn et al. 2007).

Comparisons of binding sites found by ChIP-chip and ChIP-PET technologies have concluded that both methods agree well on strong binding sites. However, there is less overlap with respect to weak binding sites, and ChIP-chip and ChIP-PET are frequently complementary in being able to detect true binding sites that the other method misses (Euskirchen et al. 2007). Microarrays do not typically include sequences with repeats; however, many true binding sites contain repeats, which will be missed by ChIP-chip methods (Euskirchen et al. 2007). There is a conceptual disadvantage of ChIP-PET that it has to read out all the non-specific sequence noise to identify true binding signals. Even in the best ChIP experiments, the majority of sequences in a library are non-specific. However, the ChIP sequence noise can also be useful. As ChIP fragments are randomly sampled from the genomes of the cells under investigation, a ChIP-PET experiment does not only generate a global map of transcription factor binding sites, but can also provide enough tag sequences for digital karyotyping of the genome (Dunn et al. 2002; Wang et al. 2002). Such an approach can be used to understand copy number variations in the cell genomes (Lin et al. 2007).

The arrival of next-generation sequencing is critical to further advance the sequencing-based measurement of ChIP DNA. The 454 sequencing platform has been used for ChIP-PET sequencing (Ng et al. 2006a), particularly with regards to the characterization of epigenomic profiles of histone modifications in mouse embryonic stem cells (Zhao et al. 2007). Recently, the ChIP sequencing strategy has been further extended by taking the advantage of the Illumina sequencing platform. In this new ChIP-Seq method, randomly sheared ChIP DNA is ligated to adaptors, and optionally amplified by PCR. A narrow size range, for example 200-300 bp, is gel-excised and sent for single direction Illumina sequencing. Many ChIP experiments yield very little DNA, therefore the low sample amount requirements of Illumina (10 ng), combined with high-throughput and low cost, make this option very attractive. ChIP-Seq has been used to generate exciting results in mapping histone modifications, transcription factor binding sites, and other DNA binding proteins (Barski et al. 2007; Chen et al. 2008b; Johnson et al. 2007). Even more recently, Illumina has

developed a Paired End sequencing method, which can be used to sequence PETs from the 5' and 3' ends of adaptor-ligated and gel-excised ChIP DNA, instead of only single tags. The PETs that define the two ends will then unambiguously infer the genome sequence content of ChIP DNA fragments. Collectively, ChIP-PET and ChIP-Seq powered by Illumina and other massively parallel short tag sequencing platforms have generated and will continue to generate valuable maps of protein factors interacting with genomic DNA in the genomic landscape. From these analyses, general pictures of transcription factor binding have started to emerge. Many transcription factors show complex binding patterns with relation to target genes (including p53 (Wei et al. 2006), Oct4 and Nanog (Loh et al. 2006), cMyc (Zeller et al. 2006); ERα (Lin et al. 2007); and NF-KB (Lim et al. 2007)). Many transcription factor binding sites are far away from transcription start sites and the promoters of target genes. How remote transcription factor binding sites function, if at all, is still largely unknown.

New developments in PET technology

The unique feature of building connectivity between two points of DNA from linear and non-linear structures in PET analysis has tremendous value in many aspects of genomic analysis that cannot be simply and easily replaced by just improving sequencing capacity in near future. The PET concept is versatile allowing for ready adaptation to new sequencing technologies. In the future, one way by which PET technology will grow is by finding new applications for answering biological questions and overcoming limitations.

One such limitation lies in the cloning step, which is a tedious affair that involves large scale plating, scrapping of bacteria from solid surface agar plates, and plasmid maxiprep. In this thesis, I present two proposed methods for overcoming the requirements for large scale scrapping. One method, called Selection-MDA, involves the use of a new Phi29 polymerase to amplify DNA after a short period of selection of circular, non-chimeric DNA in bacteria. This method is able to replace tedious solid-phase agar scraping steps used for the amplification of complex cloning-based libraries, while still maintaining high accuracy and efficiency. These advantages go beyond use in PET library construction methods: all complex

libraries, such as full-length cDNA libraries, that typically involve library scrapping may use Selection-MDA to replace library scrapping steps yet still maintain low levels of chimerism. The development of Selection-MDA is described in Chapter 2.

Another method is the use of alternative methods for PET library construction involving direct circularization of the target DNA fragments with linker oligonucleotides that covalently join the two ends of a DNA fragment. As the linker sequence is typically designed to contain two MmeI or EcoP15I sites flanking the two ends of the circularized DNA fragment, restriction digestion with these enzymes would release the tag-linker-tag structure. These PET templates can be further manipulated by adding flanking adaptors and PCR amplification before sequencing analysis. This strategy was first demonstrated in resequencing an *E. coli* genome using the polony sequencing method (Shendure et al. 2005). Another unique feature in linker design is the inclusion of a biotin group in the oligonucleotide, which allows efficient separation of the biotinylated tag-linker-tag structures from unwanted DNA debris by streptavidin-biotin based purification before and after restriction digestion. Besides tagging enzymes such as MmeI and EcoP15I that generate uniform sizes (18/20bp and 25/27bp) of PET constructs for easy purification, frequently cutting restriction enzymes and physical shearing by nebulization are also choices for generating randomly sized tag-linker-tag constructs. As reported (Korbel et al. 2007), circularized DNA was randomly sheared by nebulization and the fragments with biotinylated linkers are isolated using streptavidin. This method produced tags with a median size of 106 bp, and is very useful for obtaining long tags because no type IIS or III restriction enzyme is currently known to produce tags more than 30bp; however, many PETs prepared this way are unbalanced with tags of lengths under 15 bp, which would mean that these sequences would have to be discarded. In this thesis, I demonstrate the use of linker sequences to ligate DNA fragments followed by MmeI digestion in a new procedure to analyze chromatin DNA, which is a new application of the "cloning-free" approach. This new application is described in the proposal below, and in Chapter 3.

Proposal: Finding chromatin interactions with PETs

The applications described above have concentrated on finding genetic elements in linear DNA. However, thinking of genomic information in a one-dimensional form is far less than sufficient to elucidate the complexity of genome functions implemented through 3-dimensional organization structures in the limited nuclear space. Evidence suggests that DNA molecules are packaged with protein factors to form chromatin fibers and are folded into higher-order structures and eventually chromosomes as organizational units (Woodcock 2006) (Figure 3). Genetic elements may interact by coming into close proximity as a result of chromosome conformation to produce spatial-based functions (Figure 3). Genome functions such as transcription and replication could be closely associated with this higher-order genome organization (Fraser et al. 2007); however, we are still in early stages of understanding the complex structure-function interplay of the human genome.

Much of our current understanding of genome organization and function has come from two categories of technologies: molecular probing and molecular interaction mapping. The molecular probing technology enables us to visualize the 3-dimensional structure of genome organization at the nuclear compartment level and monitor the dynamics and functions of genomic structures in living cell nuclei. Electron Microscopy has been used to directly visualize DNA loops (Mastrangelo et al. 1991; Su et al. 1990), but Electron Microscopy requires harsh fixation and staining conditions, which could disrupt looping structures to be visualized. Atomic Force Microscopy does not have these limitations, and works by measuring forces between the scanning probe and the sample under study. It has been applied to studies of DNA looping (Yoshimura et al. 2004). Fluorescence in situ hybridization (FISH) and variants such as Cryo-FISH use fluorescently labeled DNA or RNA probes to visualize specific regions of chromatin, and has been used to generate much valuable data regarding very long interactions and chromatin conformation in the entire nucleus (Branco et al. 2006; Cremer et al. 2001; Osborne et al. 2004). However, FISH is limited by low resolution. RNA-TRAP, an extension of FISH methods capable of studying

local chromatin loops near genes in high resolution (Carter et al. 2002); however, it is limited by its inability to study multiple loci at the same time.

Molecular interaction mapping approaches identify functional DNA elements that are in close spatial proximity and hence are likely to be potential interaction points in spatial genomic organization. One of the first experiments in this area was the Nuclear Ligation Assay (Cullen et al. 1993), which sought to understand the potential of enhancer sites to form looping interactions. The enhancer sites were cloned into minichromosomes that were stably transfected into a rat cell line. The chromatin was then digested with restriction enzymes, and ligated under dilute conditions to join the sticky ends. This ligation product was then inspected using PCR with specific primers for the presence of particular known interactions that bring together target genomic regions and the transfected minichromosomal regions. If the interaction was not present, then the sequences would not be brought into close proximity, and PCR with specific primers would show no products as the primers were not on the same template. If the interaction is present, then the sequences would have been in close proximity, and the PCR would yield specific products. This Nuclear Ligation approach was further optimized in the Chromosome Conformation Capture (3C) protocol (Dekker et al. 2002). 3C was the first application to investigate *in vivo* chromatin interactions in yeast cells without the need to insert DNA sequences into minichromosomes. In 3C, chromatin is formaldehyde cross-linked, digested with restriction enzymes, diluted, ligated in a dilute manner, reverse cross-linked, and interactions are detected by PCR similar to the Nuclear Ligation Assay (Dekker et al. 2002). 3C was subsequently applied to the study of long range chromatin interactions between the ß-globin locus and locus control regions (LCR) in mammalian cells (Tolhuis et al. 2002). Further, the 3C method had been combined with ChIP in the ChIP-loop assay to identify long range interactions mediated by MECP2 at the *Dlx5-Dlx6* locus (Horike et al. 2005). However, 3C or ChIP-3C methods are limited to the detection of specific interactions using prior knowledge or perception of the existence of such interactions. To overcome this limitation, a number of groups have developed Associated Chromatin Trap

(ACT) (Ling et al. 2006), Chromosome Conformation Capture using Chip (4C) (Simonis et al. 2006), Circular Chromosome Conformation Capture (also called 4C) (Zhao et al. 2006), Open-ended Chromosome Conformation Capture (Wurtele et al. 2006) and Chromosome Conformation Capture Carbon Copy (5C) (Dostie et al. 2006) methods to expand the scope of detection for chromatin interactions. Starting from 3C or ChIP-3C ligation products, 4C, Open-ended Chromosome Conformation Capture and ACT methods all essentially use PCR to prime on known target sites and extend into unknown DNA fragments. The amplified products can then be characterized by either microarrays or cloning and sequencing analysis. Hence, these methods have the potential to detect many chromatin interactions between a known site in one location and all known and unknown counterparts in other locations; however, they are still constrained to at least one known location. The 5C method starts with a 3C template, and uses multiplex primers based on all possible combination of the restriction sites used for chromatin digestion to interrogate many interactions at the same time. However, the multiplexing is currently limited; hence 5C studies have focused on specific genomic regions (Dostie et al. 2006). In conclusion, although the currently available technologies are valuable for providing insights of chromatin interactions at limited loci or limited resolutions, they are constrained by their inability to provide a whole genome view of chromatin interactions (Simonis et al. 2007). Therefore new genome-wide technologies are needed to advance the field to provide comprehensive views and datasets of whole genome interactions and the 3-dimensional structure of chromosomes.

We have previously applied the PET approach to identify unconventional fusion genes originating from chromosomal re-arrangement events such as deletions, insertions, inversions and translocations, through mapping of 5' tags to the first exon of one gene in a genomic locus and the 3' tags to the last exon of another gene in a different genomic locus. The same concept can also be extended to characterize artificially fused DNA fragments, such as nuclear proximity ligation products. With this in mind, we propose a new strategy for whole genome Chromatin Interaction Analysis using Paired-End Tag sequencing (ChIA-

PET) (Figure 3). The basic concept of ChIA-PET is to introduce a linker sequence in the junction of two DNA fragments during nuclear proximity ligation to build connectivity of DNA fragments that are tethered together in same chromatin complex. Therefore, all linker connected ligation products can be extracted for the tag-linker-tag constructs that can be analyzed by ultra-high-throughput PET sequencing. When mapped to the reference genome, the ChIA-PET sequences are read out to detect the relationship of two DNA fragments in chromatin interactions captured by chromatin proximity ligation. As this strategy is not dependent on any specific sites for detection like 3C or 4C, ChIA-PET has the potential to be an unbiased genome-wide approach for *de novo* detection of chromatin interactions.

We anticipate that the complexity of potential substance for proximity ligation is high, the non-specific noise can be excessive; hence the cost of sequencing such material to the required depth to find true proximity ligation products can be prohibitive even for the most advanced sequencing technology currently available. To reduce the complexity and background level, we propose to use ChIP against specific protein factors to enrich specific chromatin fragments of interest before proximity ligation (Fullwood et al. 2009a). This enrichment approach would not only make the ChIA-PET sequencing approach practical by reducing the complexity of the system to be examined, but also add specificity to the identified interaction points. Depending on the protein factors used for ChIP enrichment, ChIA-PET analysis can be applied to the detection of all chromatin interactions involved in a particular nuclear process. For instance, the use of general transcription factors or RNA Polymerase II components for ChIP enrichment and ChIA-PET analysis would identify all chromatin interactions involved in transcription regulation, and the use of protein factors involved in DNA replication or chromatin structure would allow identification of all chromatin interactions due to DNA replication and chromatin structural modification. More specifically, the use of specific transcription factors for ChIA-PET analysis would further reduce ChIA-PET library complexity and add specificity to chromatin interactions, and enable examination of specific chromatin interactions mediated by particular transcription

factors. Our preliminary experimental data (Chapter 3) has demonstrated that ChIA-PET can generate PET sequences that identify transcription factor binding sites and interactions between remote binding sites.

Through large-scale application of ChIA-PET to the system of estrogen receptor α-mediated chromatin interactions in human MCF-7 breast cancer cells, we have generated the first global chromatin interactome map (Chapter 4). We found that the great majority of ERα binding sites are anchored at promoter regions of target genes through long-range chromatin interactions. Our data suggests that ERα functions in transcription regulation by bringing genes together through intensive looping of chromatin interactions into transcription foci. These findings suggest that chromatin interaction is a primary mechanism for regulating transcription in mammalian genomes particularly in transcriptional induction.

With further development and optimization of the ChIA-PET prototype protocol, and with our new findings from this first application of ChIA-PET, we expect that this whole genome approach for unbiased and *de novo* discovery of long range chromatin interactions and these chromatin interaction maps will help to establish an emerging field for studying genome interaction and regulation networks in 3 dimensions.

In conclusion, the PET technology is a versatile method which can couple methods for asking biological questions with next-generation sequencing (Fullwood et al. 2009b). In this thesis, I present new methods which can improve current library construction methods as well as ask new biological questions, thus furthering both PET technology and our understanding of biological systems.

Chapter Two: Selection-MDA for amplifying complex DNA libraries

Introduction

A mainstay of genomic technologies to interrogate genomes and functional genomic elements is the generation of complex cloning-based DNA libraries. Examples of such libraries include genomic DNA libraries used in the sequencing of the human genome (Lander et al. 2001) as well as other genomes (Waterston et al. 2002); full-length cDNA (flcDNA) libraries (Strausberg et al. 1999) and Gene Identification Signatures with Paired-End Tags (GIS-PET) libraries used for elucidating the transcriptome (Ng et al. 2005); as well as Chromatin Immunoprecipitation with Paired-End Tags (ChIP-PET) libraries used for elucidating transcription factor binding sites (Wei et al. 2006).

In constructing such libraries, the starting DNA samples are often limited, and therefore DNA amplification is often necessary. The method of choice has been bacterial propagation of DNA fragments in plasmid vectors. To ensure accurate representation, the bacteria must not be allowed to compete with each other for nutrients. Therefore, growth and scraping from solid-surface agar is commonly used because colonies are spread out on solid-surface agar such that they will not encounter each other and compete. As the libraries are complex and contain many different DNA molecules, a large number of colonies must be scraped from the agar to ensure that the resulting library contains sufficient coverage of the different DNA molecules present in the original pool. Plating and scraping large numbers of solid-surface agar bacteria clones then results in methods that are tedious, time consuming, and difficult to scale up.

Multiple Displacement Amplification (MDA) has been recently developed as a method for *in vitro* amplification of DNA. MDA is a method for amplifying plasmids and long strands of DNA in a cell-free system using phi29 polymerase, a newly discovered polymerase enzyme that has very high fidelity (Esteban et al. 1993), proof-reading activity (Garmendia et al. 1992), and processivity (Blanco et al. 1989). Such a system would be ideal

for replacing the tedious solid-phase agar scraping steps used for the amplification of complex cloning-based libraries. The use of MDA would remove this bottleneck, as MDA is able to amplify complex mixtures with high accuracy and efficiency.

However, one obstacle to the use of MDA for the amplification of complex cloning-based libraries is the fact that cloning ligation reactions into vectors typically results in multimers of plasmid vectors and DNA fragments. Bacterial propagation can remove multimers because replication constructs that contain multiple origins of replication will not survive during bacterial replication, while MDA alone is not capable of such selection to eliminate multimers during amplification.

To overcome this problem, we developed a method, called Selection-MDA, which combines the selection capability of bacterial replication for single vector/insert constructs with the efficiency and convenience of MDA. In this method, we first transfer the vector/insert ligation into electrocompetent *E. coli* for a short period of replication and selection in liquid media. Because the bacteria are harvested after a short period of growth in liquid media, the bacteria would not have multiplied to such an extent that they begin to compete for nutrients, yet plasmids with multiple origins of replication would be selected out. The multimer-free pool of plasmids is then purified from liquid media and used for MDA, which amplifies large quantities of multimer-free DNA, thus eliminating tedious and time-consuming plating and scraping of solid-surface agar. As such, the selective advantage of bacterial propagation can be combined with the efficiency convenience of the MDA method without the disadvantages of sample bias or chimeras. The end result is an MDA-amplified library of the same quality as a similar library prepared by bacterial propagation.

To validate the Selection-MDA method in a complex library, we prepared a GIS-PET library (Ng et al. 2005) with the Selection-MDA method, and compared it with the same library prepared by conventional bacterial amplification on solid surface agar (Zhao et al. 2007). Short Paired-End Tag (PET) libraries, including GIS-PET, were conceived of in order

to improve sequencing efficiency. In GIS-PET, the 5' and 3' signatures of each full-length cDNA are covalently linked into structures in which the 5' and 3' tags were paired together, and then sequenced, allowing a 20-30 fold increase in efficiency compared with bidirectional sequencing of DNA (Ng et al. 2006a). The paired end nature of the method also allows the use of GIS-PET to study unconventional fusion transcripts (Ruan et al. 2007). The same concept has also been applied to ChIP DNA characterization (ChIP-PET) (Wei et al. 2006). The PET analysis method involves the construction of two libraries: the original DNA insert library (flcDNA library for GIS-PET), and the single PET library, which is derived from the original DNA insert library. The amplification of the libraries using bacteria propagation is time consuming and labor intensive. To further improve PET analysis, we applied the Selection-MDA method to replace the single PET library amplification step (Figure 4) (Fullwood et al. 2008).

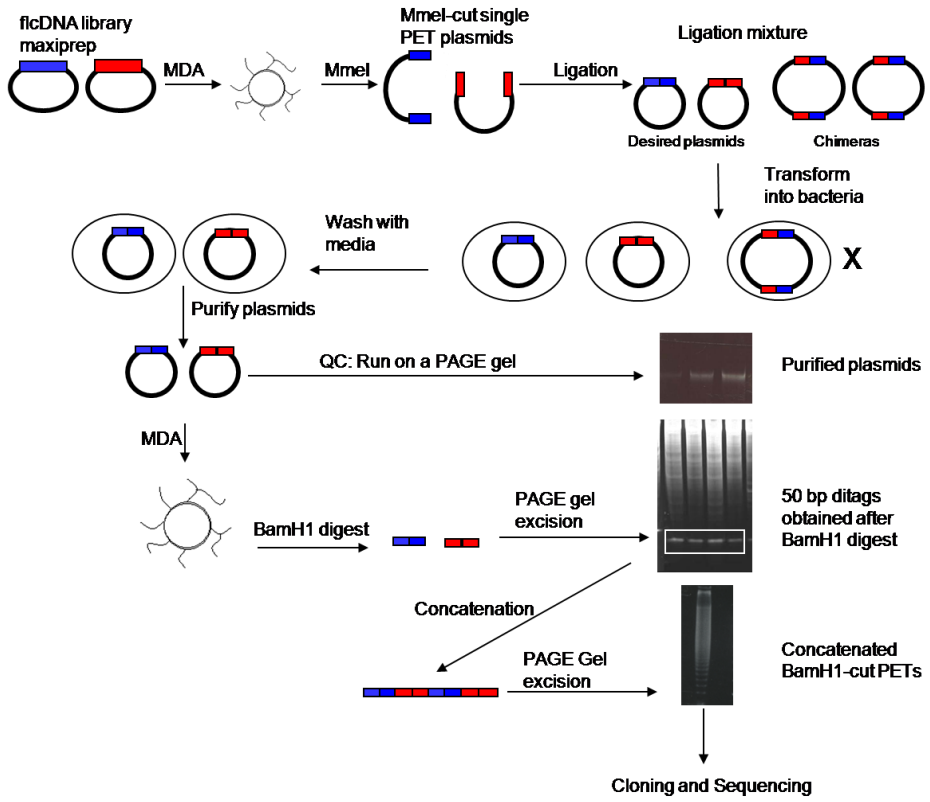


Figure 4. Schematic of a GIS-PET library prepared by the Selection-MDA method.

FlcDNA maxiprep was cut with MmeI, self-ligated, and transformed into bacteria, which were recovered for 4 hours. After this, cells were washed with media, plasmids were extracted. MDA was then performed, followed by enzymatic digestion, concatenation, and then cloning and sequencing. We ran quality control aliquots of the reactions on PAGE gels after the plasmid purification. Clean plasmids of the correct size, 2,800 bp, were obtained. After BamH1 digestion, 50 bp PETs were successfully recovered, as may be seen from the PAGE gel which shows a band of 50 bp (marked by a white box) separated from a high molecular weight smear from the plasmid backbone. PETs were successfully excised and concatenated, as may be seen from the smear from the concatemers, which was seen on a third PAGE gel. The concatemers were excised from the PAGE gel and prepared for subsequent cloning and sequencing. *(Note: Selection-MDA GIS-PET library was prepared by Jack Tan).*

Results

The starting point for this analysis was HES3 human embryonic stem cell RNA, from which we generated a full-length cDNA (flcDNA) library (Figure 5A, B, and C). We then generated two libraries: (1) a GIS-PET library by the standard approach, called SHE001 (Figure 5D), which comprised 613,905 unique PETs that were collapsed into 25,845 transcriptional units; and (2) a GIS-PET library prepared by the Selection-MDA approach, called SHE002 (Figure 4), which comprised 12,888 unique PETs which were collapsed into 3,584 transcriptional units.

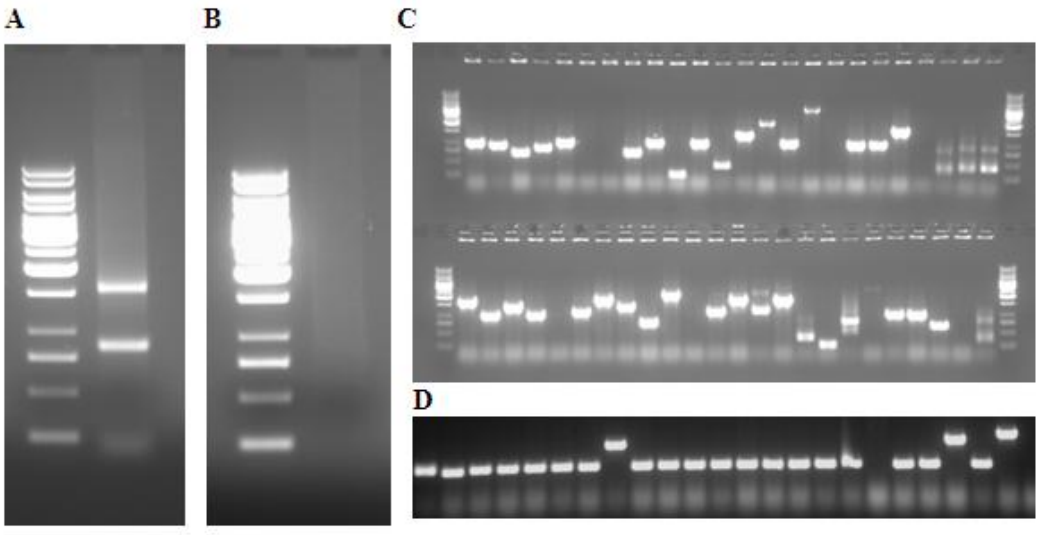


Figure 5. Full-length cDNA and GIS-PET library quality controls.

A. HES3 Human embryonic stem cells were grown and prepared as described. Total RNA was prepared by the Trizol isolation method. A smear of RNA with

two bright bands corresponding to the 28S and 18S rRNA was obtained. The ladder used in all panels is Generuler 1 Kb (Fermentas) (http://www.fermentas.com/catalog/electrophoresis/images/generuler031123.jpg) . B. The mRNA prepared by the use of the μMACS mRNA isolation kit on total RNA showed no bright bands corresponding to the rRNA. C. A full-length cDNA library was prepared by the Captrapper method, which had a titer of 4.6×10^6 cfu. Colony PCR Quality Control of the library was performed. An empty vector will produce a PCR product of size 260 bp (corresponds to the first band of the ladder); insert sizes were therefore calculated by subtracting off the size of the empty vector. Colony PCR therefore showed a range of insert sizes from 250-2,000 bp (corresponds to the second to seventh bands of the ladder). This is expected, as a full-length cDNA library is expected to give a range of different sized inserts, with no single dominant size. Given that the library was of good quality, as can be seen from the colony PCR, the library was used to prepare two libraries: A single-PET library by the classic method, and a single-PET library by the Selection-MDA method. D. A single-PET library was prepared from the full-length library as per the classic bacterial propagation method. Colony PCR Quality Control of this library showed a single predominant fixed size of 300 bp in many colonies, which is expected, as single-PET plasmids all have a fixed size of 2,800 bp, and hence upon PCR, will give a band of 300 bp. Certain clones do not show this fixed size, which could be the result of the incorporation of foreign DNA, or other factors. Colony PCR QC showed an insert ratio of 75% based on the number of wells that had PCR products of the correct size (300 bp). *(Note: HES3 cells were prepared by Andrew Choo and Steve Ho, and the flcDNA and GIS-PET libraries were constructed by Liu Jun).*

To construct the MDA-amplified library (schematic in Figure 4), a single-PET ligation mixture was generated from the maxiprep of the flcDNA library, transformed into bacteria, and recovered for 4 hours in the "Selection" part of the procedure. The short 4 hour growth in liquid media, which allows for the selection of single insert clones because multiple insert clones have multiple origins of replication and cannot survive. However, the time is not long enough to result in crowding of bacteria in liquid media, such that size bias is minimized. To investigate whether the bacteria would have multiplied such that they crowd, we analyzed and plotted the optical density of the liquid media at 0 h, 1 h, 2 h and 4 h. The optical density absorbance at 600 nm (OD_{600}) of the media increased from 0.728 at 0 h to 0.897 over 4 h. Using the approximation that 1 OD_{600} is approximately 1×10^9 cells, our bacteria increased from 7.3×10^8 to 9.0×10^8 cells over 4 h. Hence, our bacteria are still in log growth, and the increase in cell number should not be sufficient to cause crowding. At the end of 4 h, the bacteria were washed well and harvested. Plasmids were prepared by

miniprep and DNAse cleanup. A quality control check showed that clean plasmids (Figure 4) were obtained. PETs were then released by BamHI digestion (Figure 4). Released PETs were concatenated for Sanger sequencing (Figure 4). These quality controls indicate that the Selection- MDA procedures were successful in producing PETs for sequencing.

We analyzed the library of PET sequences derived from the MDA approach using standard GIS-PET quality control measures (Ng et al. 2005), to investigate whether libraries prepared by the MDA approach are of good quality. Of a total 12,888 unique PETs sequenced, the number of PETs that could not be mapped to the human genome was 22.9%. This number is comparable to the percentage of unmappable PETs (26%) shown in a mouse embryonic stem cell library (Ng et al. 2005), and indicates that the MDA approach has a low percentage of chimeras due to multimers as well as high accuracy amplification, which allows the amplified sequences to map well to the genome. In addition, the mapping accuracy (percentage within $\pm$ 100 bp of the transcription start site or polyadenylation site) for all known PETs in SHE002 was 92.5% for 5' tags and 91.9% for 3' tags, comparable to the mouse embryonic stem cell GIS-PET (Ng et al. 2005), which showed results of 90.7% for 5' tags and 86.9% for 3' tags. Overall, the percentage of PETs with both 5' and 3' tags that map accurately is 88.4% for the entire library. While high, this measure includes mRNAs that have alternative splicing and alternative transcription start sites and hence represents a lower bound. The 12,888 unique PETs were collapsed into 3,584 transcriptional units. To more accurately measure the mapping accuracy of the library, we examined PET sequences from the top 20 most abundant transcriptional units, which are well-annotated. The overall mapping accuracy is 98.5% for the top 20 transcriptional units of SHE002. This high level of mapping accuracy indicates that Selection-MDA method can accurately capture gene identification signatures.

In order to directly compare the performance of the Selection-MDA protocol with the standard protocol, we wanted to compare the quality control measures of the MDA-prepared

GIS-PET library with those of a GIS-PET library (SHE001) prepared by conventional bacterial amplification. As the size of the data sampled from library SHE001 (the total number of PETs is 613,905) is almost 50-fold larger than the size sampled from library SHE002 (the total number of PETs is 12,888), a direct comparison of these two library will not be meaningful. Therefore, in order to compare the two libraries at the same number of PETs, we created 3 smaller virtual libraries, SHE004, SHE005, and SHE006 (Table 1), by randomly selection of data from bacterial propagation library SHE001, such that the virtual libraries had the same approximate size as that of the MDA-prepared SHE002. Differences within the set of these 3 virtual libraries would reflect sampling variation. Hence, if the differences between the MDA approach and the conventional approach are significant, then the differences between SHE002, and SHE004, SHE005 and SHE006 should be much larger than the differences between SHE004, SHE005, and SHE006 (Figure 6).

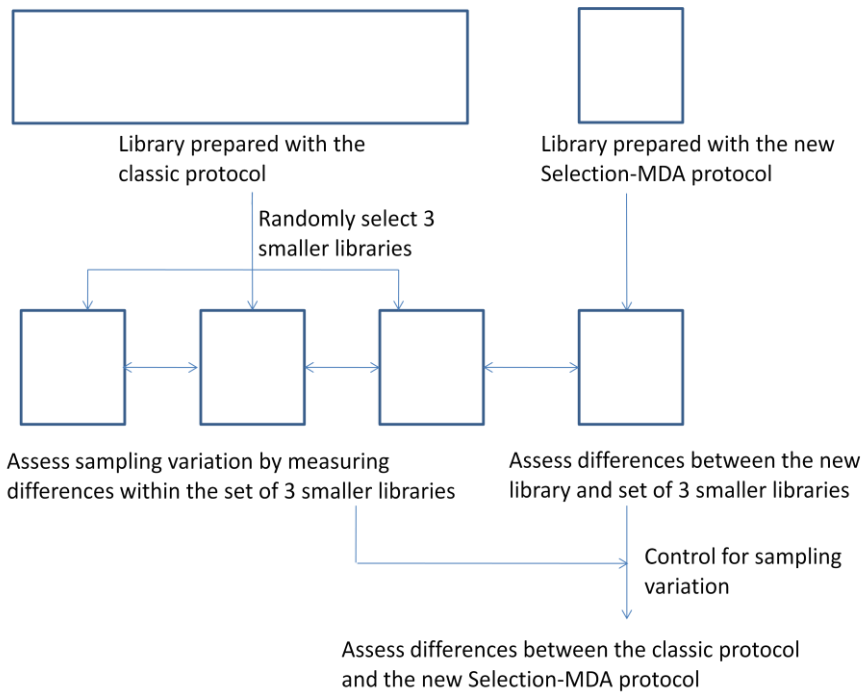


Figure 6. Data analysis method.

A Selection-MDA GIS-PET library was created. This library was compared with a GIS-PET library created from the same source material, but which was prepared with the classic bacterial propagation protocol. We wished to assess whether the differences between the new protocol and the classic protocol are the

result of sampling differences between the libraries or actual biases in the protocol. Therefore, we randomly chose 3 smaller libraries from the bacterial propagation library. Each smaller library is of the same size as the library prepared by Selection MDA. If the differences between the new protocol and the classic protocol are the result of sampling differences, then the differences between the three libraries randomly selected from the same parental library should be the same as the libraries between each of the libraries and the library prepared by the new method. Otherwise, the differences are the result of actual biases, either on the part of the classic bacterial propagation protocol, the new Selection-MDA protocol, or both.

Table 2: Analysis of GIS-PET library quality control measures.

Category		**SHE002 (Selection-MDA)**	**SHE004 (Classic)**	**SHE005 (Classic)**	**SHE006 (Classic)**
PET Sequences					
Total number of unique PETs		12,888	13,196	12,988	13,102
PET matches to the genome	0 matches	2,953 (22.9%)	2,903 (22.0%)	2,895 (22.3%)	2,925 (22.3%)
	1 match	9,641 (74.8%)	8,266 (62.5%)	9,851 (75.8%)	9,936 (75.8%)
	>1 match	294 (2.3%)	2,027 (15.4%)	242 (1.9%)	241 (1.8%)
Mapping accuracy	All PETs	88.4%	89.1%	88.2%	88.4%
	PETs from the top 20 transcriptional units	98.5%	97.9%	98.5%	99.2%
GC percentage		49.7%	48.9%	48.2%	48.3%
Categories of PETs with 1 match to the genome	Known	5,697 (59.1%)	5,253 (63.6%)	6,080 (61.7%)	6,083 (61.2%)
	ESTs	3,512 (36.4%)	2,678 (32.4%)	3,291 (33.4%)	3,385 (34.1%)
	Gene predictions	380 (3.9%)	303 (3.7%)	431 (4.4%)	420 (4.2%)
	Novel	52 (0.5%)	31 (0.4%)	48 (0.5%)	48 (0.5%)
Transcriptional units					
Total number of transcriptional units		3,584	3,362	3,780	3,776
Categories of Transcriptional units	Known	2,278 (63.6%)	2,309 (68.7%)	2,490 (65.9%)	2,506 (66.4%)
	ESTs	997 (27.8%)	817 (24.3%)	965 (25.5%)	956 (25.3%)
	Gene predictions	265 (7.4%)	209 (6.2%)	287 (7.6%)	280 (7.4%)
	Novel	44 (1.2%)	27 (0.8%)	38 (1.0%)	34 (0.9%)

The percentages of PET matches to the genome, numbers of transcriptional units, as well as mapping accuracies of SHE004, SHE005, and SHE006 are comparable to that of SHE002, indicating that the MDA-prepared library is of similar quality as that of the conventionally-prepared library constructed from the same starting material (Table 2).

Next, we checked whether the MDA procedure caused any biases in the sample. Because MDA is a different amplification method from bacterial amplification, we wished to investigate if there was any base bias. Base bias was measured by calculating the GC percentage of the library. There is minimal base bias between the MDA method and the conventional method (Table 2).

Again because MDA is a different amplification method, we investigated whether there is any bias towards any category of genes, such as novel genes. We grouped the PETs into "known genes", "gene predictions", "ESTs" and "novel genes". All libraries showed similar distributions, indicating minimal category bias (Table 2).

The Selection-MDA step could not have introduced a length bias in this particular library, because Selection-MDA was performed on single PET clones, which are all of a fixed size. Therefore, we could not test whether Selection-MDA would result in length biases or not. However, given that MDA was performed on the full-length cDNA library maxiprep to obtain more material for the construction of the single-PET library in the MDA procedure, we reasoned that this step might have introduced a length bias, and hence investigated whether there was a length bias. We tested for the presence of length bias by investigating the mRNA lengths of the best-matching Known Genes, ESTs, or Gene Predictions, and found there was a length bias towards shorter mRNAs on the part of Selection-MDA, but the bias is small (Figure 7). Given that the bias is small, it is possible that the apparent bias could still be the result of sampling variation.

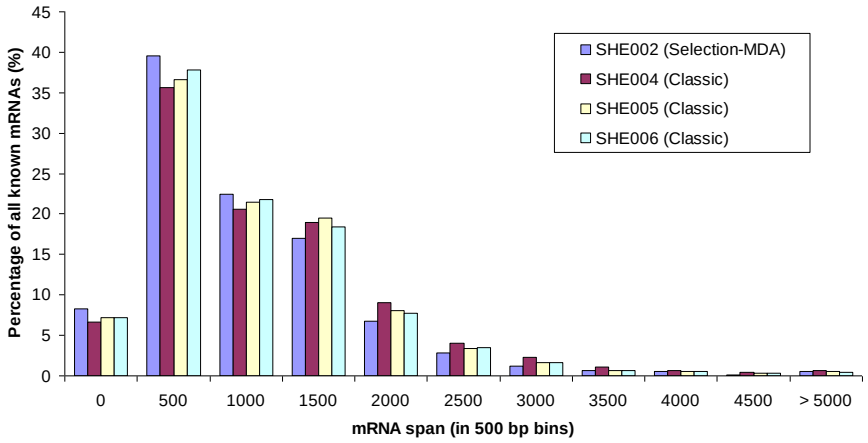


Figure 7. Analysis of length bias between the MDA approach and the bacterial amplification approach.

We tested for the presence of length bias by classifying the mRNA lengths of the best-matching Known Genes, ESTs, or Gene Predictions from each library into 500-bp bins, which were then plotted on a graph. There is a small length bias. Because the length bias is small, it is possible that the apparent bias is due to sampling variation.

Next, we reasoned that the contents of the SHE002, SHE004, SHE005, and SHE006 libraries should be similar, because the same starting full-length cDNA library was used for the preparation of the two libraries. Hence, we compared the top 20 most abundant transcriptional units of each library with each other (table 3). The average number of transcriptional units that are the same between SHE002 (the MDA-prepared library) and any randomly selected library from a bacterial propagation library is 13. The average number of transcriptional units that are the same between the bacterial propagation libraries is 14, suggesting that the agreement between the MDA method and the bacterial amplification method is similar to the agreement between randomly selected libraries chosen from the same bacterial propagation library. This analysis thus indicates that the contents of the MDA-prepared library show a good match to those of the conventionally-prepared library.

Table 3: Identities of the Top 20 transcriptional units of each library.

Rank	SHE002 (Selection-MDA)	SHE004 (Classic)	SHE005 (Classic)	SHE006 (Classic)
1	FTL	FTL	FTL	FTL
2	GAPDH	MIF	ENO1	MIF
3	MIF	ENO1	MIF	ENO1
4	TPI1	PRDX1	RPL13	LOC388817
5	ENO1	IFITM1	RPS2	RPS2
6	LOC388817	C14orf172	TPI1	RPL13
7	RPL13	K-ALPHA-1	LOC388817	H3F3A
8	OAZ1	PGK1	RPL9	PRDX1
9	FTH1	RPL13	K-ALPHA-1	TMSL3
10	TMSL3	PFN1	FTH1	TPI1
11	H3F3A	LOC388817	MDK	H2AFZ
12	IFITM1	RPL18	H2AFZ	K-ALPHA-1
13	H2AFZ	IFITM3	H3F3A	FTH1
14	PRDX1	ACTG1	PGK1	PRDX4
15	C14orf172	PRDX4	RPL18	PFN1
16	PFN1	MDK	TMSL3	C14orf172
17	RPL15	OAZ1	IFITM1	IFITM1
18	TPT1	RPL8	PRDX1	RPL9
19	RPL9	RPLP0	C14orf172	RPL18
20	RPL10	STOML2	OAZ1	PGK1

Discussion

Taken together, we have shown the method of inserting plasmids into bacteria for a short selection interval followed by MDA is a feasible method for the construction of a complex library. We have successfully applied Selection- MDA to the construction of a complex GIS-PET library and found that the Selection-MDA method results in a library with similar content and quality control statistics as compared with a library constructed from the same starting material that was amplified with bacteria and harvested through scraping bacterial colonies from solid surface agar.

Comparing the steps between the MDA version and the bacterial propagation method, it is clear that the MDA version requires much less hands-on labor (Figure 8). In terof the physical handling, the MDA version uses small scale 1.5 ml tubes of material whereas the bacterial propagation method uses 10 large Q-trays and many maxiprep columns. The approximate times for each step that differed between the two protocols was estimated (Figure 2A). Comparing the absolute times required, the MDA method requires 4 h less time

than the bacterial propagation method. Considering the fact that many of the time-consuming steps in MDA do not require hands-on activities and hence allows other projects to be carried out in parallel, the time requirement of the MDA method is much less than the bacterial propagation method. With recent improvements in the MDA method (for example, the Illustra Genomiphi V2 DNA Amplification kit from GE Healthcare), further time savings could be possible.

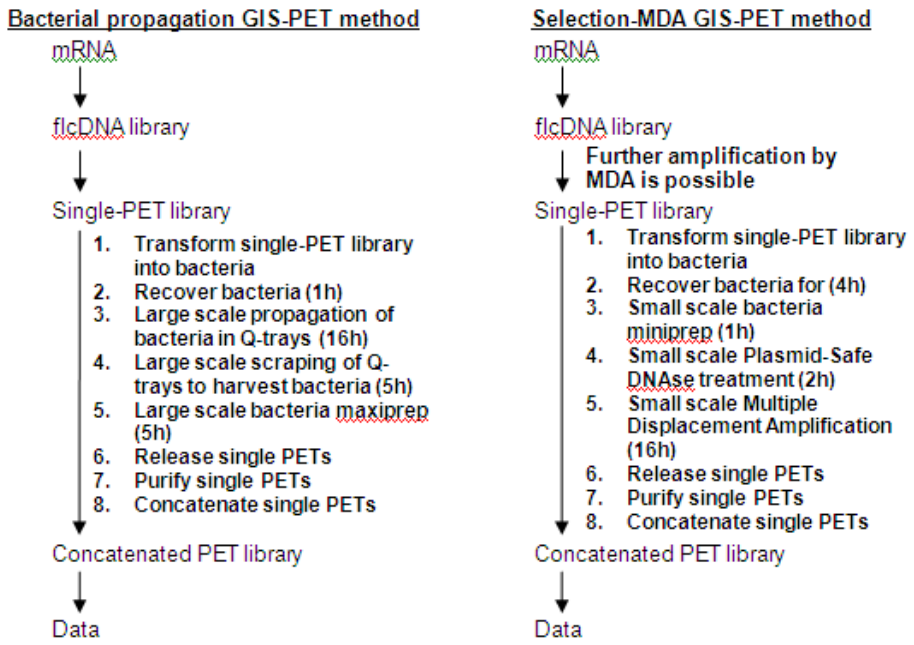


Figure 8. Differences between the GIS-PET method with classic approach and the GIS-PET method with the new Selection-MDA approach.

The Selection-MDA version allows for further amplification of the flcDNA library maxiprep by MDA, as well as amplification of the single-PET library solely by Selection-MDA without the need for tedious plating and scraping of large numbers of bacterial colonies from solid surface agar. Approximate times required for steps that are different between different protocols are given in brackets. Comparing the steps between Selection-MDA and the bacterial propagation method, it is clear that Selection-MDA requires much less hands-on labor and time, and also, in terof absolute time, is at least 4 hours shorter.

The concept of performing bacterial selection followed by MDA (Selection-MDA) may be used to replace amplification steps in complex libraries, and represents a substantial improvement to existing cloning-based protocols. The Selection-MDA method is an effective and simple method for the unbiased amplification of a pool of complex clones, which allows

scale-up and elimination of tedious scraping steps in library preparation protocols. The method may be readily integrated and applied to current cloning-based protocols.

In conclusion, Selection-MDA is a novel method for the amplification of cloned libraries consisting of complex DNA. We applied Selection-MDA to a GIS-PET library, an example of a cloned, complex DNA library, to illustrate the benefits of Selection-MDA. Library preparation was made simpler, and differences between the MDA-prepared library and a library prepared by the classic protocol were minimal. Hence, Selection-MDA is an effective and useful improvement to current cloning-based protocols.

Chapter Three: Whole Genome Chromatin Interaction Analysis using Paired-End Tag Sequencing

Introduction

While genomic information is usually presented as a linear series of bases, genomes are known to be organized into three-dimensional structures *in vivo* (Woodcock 2006). Genome-wide studies of transcription factor binding sites (TFBS) using chromatin immunoprecipitation (ChIP) followed by microarray detection (ChIP-chip) (Cawley et al. 2004), paired end sequencing (ChIP-PET) (Lim et al. 2007; Lin et al. 2007; Loh et al. 2006; Wei et al. 2006; Zeller et al. 2006; Zhao et al. 2007) or single end sequencing (ChIP-Seq) (Johnson et al. 2007), particularly for estrogen receptor α (ERα) (Lin et al. 2007), have shown that many TFBS are not located 5' proximal to genes, suggesting extensive remote regulation in many systems. Possible models of remote regulation include looping and sliding (West et al. 2005). Various methods can investigate looping interactions, such as Chromosome Conformation Capture (3C) and variants including ChIP-3C, 4C and 5C (Cai et al. 2006; Carroll et al. 2005; Dekker et al. 2002; Dostie et al. 2006; Simonis et al. 2006; Wurtele et al. 2006; Zhao et al. 2006), as well as RNA-Trap (Carter et al. 2002) and FISH (Cremer et al. 2001), which have provided many insights into higher level organization of chromatin structures. However, these methods are limited to one-point oriented or partial genome detection of interactions, and are incapable of *de novo* detection of genome-wide interactions. A global strategy for investigating higher-order chromatin structures is needed to understand mechanisms for the remote control of transcription regulation in 3-dimensional nuclear space. We therefore developed a genome-wide, high-throughput, and unbiased approach called ChIA-PET with the incorporation of the original concept of "nuclear proximity ligation" (Cullen et al. 1993) that has been applied in the 3C approach (Dekker et al. 2002) to capture interacting DNA segments bound by protein factors, the exploitation of the Paired-End Tag (PET) strategy (Ng et al. 2006a; Ng et al. 2005; Wei et al. 2006), as well as the utilization of

next generation sequencing technologies (Margulies et al. 2005) for *de novo* detection of chromatin interactions. Here, we demonstrate this method using the system of ERα-mediated transcription regulation.

Results

Construction and mapping of ChIA-PETs

The basic principle of detecting chromatin interactions is the use of "proximity ligation" to capture DNA elements that are in close spatial distances as a result of juxtaposition by protein factors but which are located far away from each other in the linear genome (Cullen et al. 1993; Dekker et al. 2002). In "proximity ligation", chromatin is diluted, and spatially proximate DNA fragments within the same chromatin complex connect to each other through ligation, while chimeric ligations between different chromatin complexes are minimized. One of the major challenges of developing an unbiased whole genome approach for *de novo* detection of chromatin interactions is to find a method for manipulation of the connected DNA fragments. An even bigger challenge is the expected high level of complexity of chromatin interactions in the compacted nuclear space crowded by masses of DNA and related proteins. Consequently, any region of the genome could potentially interact with multiple segments of the genome, specifically or non-specifically. Moreover, such interactions may act transiently and proximately (Misteli 2007). Further challenges arise when a population of non-synchronized cells is studied, in which specific interactions may occur only in a small portion of the cell population (Misteli 2007; Simonis et al. 2007). Hence, analyses of chromatin interactions are expected to be very noisy, and the question of how to reduce the complexity for detection of specific interactions is a critical issue. The 3C and variant methods use sequence-specific approaches to reduce the complexity by detecting interactions that are only related to the targeted genome locations, but exclude interactions in all other regions(Simonis et al. 2007).

To overcome these issues, we devised a strategy to introduce a specific oligonucleotide sequence into the junction of all proximity ligation products. We coupled this strategy with Chromatin Immunoprecipitation (ChIP) to enrich specific chromatin interactions, as well as ultra-high-throughput sequencing technology for deep coverage, and thus formulated the ChIA-PET analysis procedure (Figure 9).

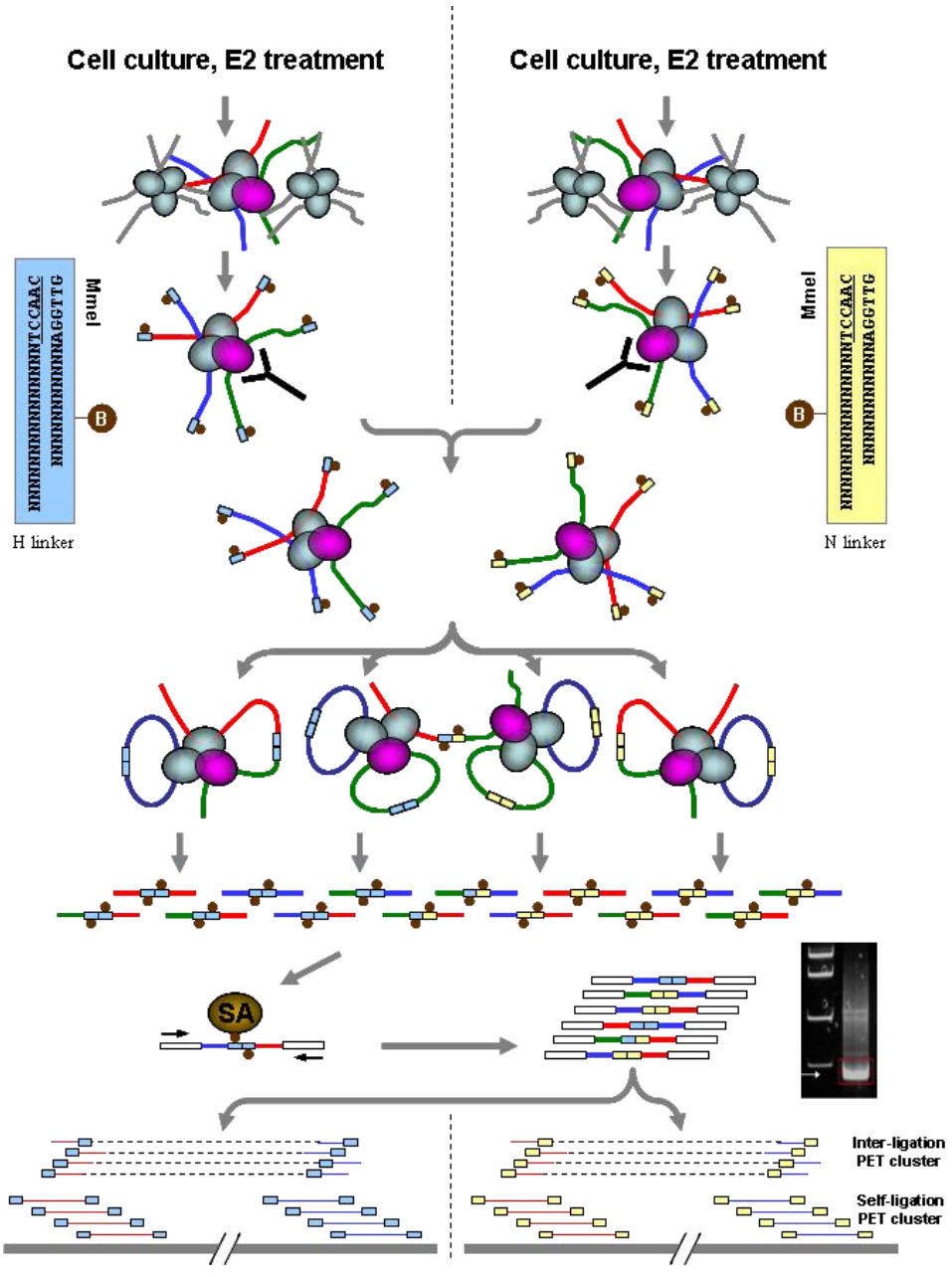


Figure 9. The ChIA-PET method.

> Two biological replicates were analyzed simultaneously in the same ChIA-PET experiment, using barcoded linkers (H and N). ChIP-enriched chromatin was labeled with different linkers to represent different replicates, pooled, and ligated in a dilute manner. PETs were obtained from these products after MmeI digestion. The biotinylated PETs were bound to streptavidin beads, ligated to universal adapters, PCR-amplified, gel-purified, sequenced, and mapped to the genome. *(Note: ChIA-PET was performed together with Liu Jun. All sequencing was done by the Genome Technology and Biology Sequencing Team, led by Wei Chialin. All names are given in the Materials and Methods, Chapter 6).*

Briefly, after ChIP enrichment, tethered DNA fragments in chromatin complexes are first ligated to excess half-linker oligonucleotides; the intermediate molecules are then circularized under dilute conditions for "proximity ligation", resulting in two kinds of ligation products: the "self-ligation" where a complete linker joins the two ends of one DNA fragment, and the "inter-ligation" where a complete linker connects two different DNA fragments. The linker sequence is designed to contain a MmeI site flanking each end of the ligated DNA fragments, so as to allow type IIS restriction digestion to release a PET structure (20bp tag – linker – 20 bp tag) from each of the linker-ligated products. In addition, the linkers are biotinylated, allowing for easy manipulation of the PETs using streptavidin magnetic beads. The PET structures derived from both self-ligated and inter-ligated DNA fragments are then subjected to ultra-high throughput sequencing using a Roche 454 pyrosequencer (Margulies et al. 2005), and the PET sequences are mapped to the reference genome sequences. Therefore, the sequencing of one ChIA-PET library can generate two datasets: the self-ligation PETs from individual ChIP DNA fragments, and the inter-ligation PETs from interacting DNA fragments (Figure 10A).

We expect that real and specific chromatin interactions would be enriched, and this enrichment would be reflected as increased frequencies of multiple PETs occurring at, or between, specific regions in the ChIA-PET library sequence dataset, while non-specific sequence data would scatter randomly along the genome. Using this principle of multiple PET overlaps as a readout for real binding sites (Wei et al. 2006) and interactions, a ChIA-

PET experiment can identify precise transcription factor binding sites (TFBS) using the overlap density of the PET data as a measure of ChIP enrichment, and reveal true interactions between TFBS using overlapping inter-ligation PET data. Thus, we can distinguish true binding sites and interactions from random background noise (Figure 10B).

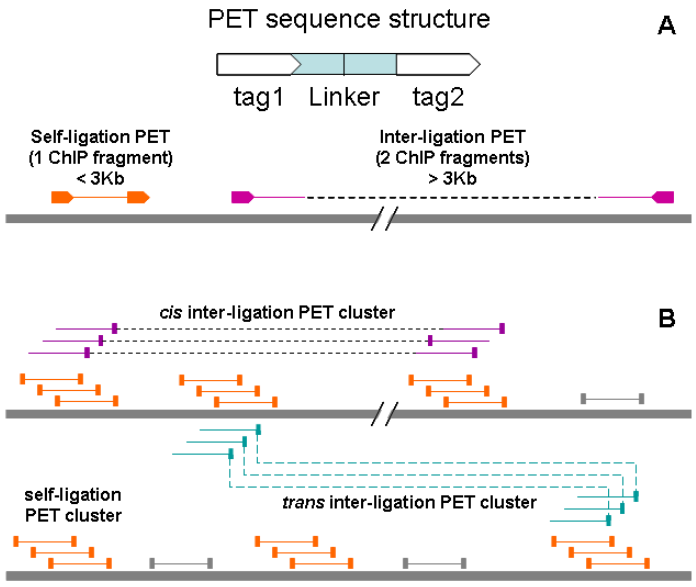


Figure 10. ChIA-PET structures allow inference of self-ligation and inter-ligation status.

A. Schematic of PET structure and mapping. PETs with both tags that map to the same chromosome with the genomic span in the range of ChIP DNA fragments (less than 3 kb), with expected self-ligation orientation and on the same strand, are considered to derive from the self-ligation of a single ChIP DNA fragment, and are therefore called "self-ligation PETs". The genomic span in the range of the ChIP DNA fragments was determined by examining the sizes of the sonicated products on an agarose gel and taking an upper limit. If a PET did not fit into these criteria, we considered that the PET most likely resulted from a ligation product between two DNA fragments, an "inter-ligation PET". B. Clustering method for determining binding sites and interactions. Self-ligation PET clusters involving multiple overlapping unique self-ligation PETs were taken to represent binding sites; intrachromosomal inter-ligation PET clusters involving multiple overlapping unique inter-ligation PETs were taken to represent intrachromosomal interactions; interchromosomal inter-ligation PET clusters involving multiple overlapping unique inter-ligation PETs were taken to represent interchromosomal interactions; and singleton PETs were taken to represent noise *(Note: PET processing was done by Hong-Sain Ooi, Pramila Ariyaratne, Han Xu, and Yusoff Bin Mohamed).*

We applied this method to characterize ERα-mediated chromatin interactions in human breast adenocarcinoma cells MCF-7. ERα is a ligand-dependent transcription factor that plays important roles in breast cancer and normal human physiology (Ali et al. 2000). Whole genome maps of ERα binding sites have been generated by ChIP-chip (Carroll et al. 2005; Lupien et al. 2008) and ChIP-PET (Lin et al. 2007) approaches. In addition, a few ERα-mediated long-range chromatin interactions have been characterized between the promoters and enhancers of the estrogen-responsive genes TFF1 (also known as pS2) (Carroll et al. 2005) and GREB1 (also known as KIAA0575) (Deschenes et al. 2007). Hence, the ERα system constitutes an excellent model for testing the ChIA-PET method in complex genomes.

We first generated two ERα ChIA-PET libraries as two biological replicates of MCF-7 cell cultures treated with estradiol (E2) using two linker sequences with different nucleotide barcodes. As a linker sequence can include a unique nucleotide barcode, multiple linkers with distinctive nucleotide barcode sequences can be used to specify different experiments or replicates, and monitor the non-specific inter-ligation (chimeric) rate between ChIP complexes. Hence, different biological samples or replicates may be analyzed under similar experimental conditions in a time and cost-effective manner to reduce technical variations of measurement. Using the Roche 454 pyrosequencing platform (Margulies et al. 2005), we generated 764,899 and 715,369 PET sequences for these two libraries, respectively (Table 4).

Table 4. Statistics of library datasets used in this chapter.

Library description	Raw PET sequences	Unique PET sequences	Mapped PETs	ERαBS (FDR<0.01)[A]	Intrachrom. Inter-ligation PET clusters
ERα ChIA-PET Library 1 (IHM001_NN)	867,751	715,369	497,979 (70%)	2,720	189
ERα ChIA-PET Library 2 (IHM001_HH)	941,151	764,899	514,192 (67%)	2,179	208
Chimeras[B]	40,165	30,808	18872 (61%)	34	2[C]
ERα ChIP-PET Library(Lin et al. 2007) (IHM043)	2,543,100	1,118,509	895,624 (80%)	1,211	2[D]
IgG ChIA-PET Library (IHM062)	508,211	436248	403,149 (92%)	0	0

Notes: A: "False Discovery Rate" is abbreviated "FDR". B: The chimeras are inter-ligation PETs between the two ChIA-PET library materials with hybrid linker H and N. C: Of these 2 interactions, 1 was in an amplicon. The other showed no abnormalities upon manual curation, but could be a random noise. D: These PET clusters have genomic spans of over 10 Mb and have only 2 overlapping PETs. They are therefore considered to be non-specific.

We also generated control libraries for comparison with the ChIA-PET libraries to validate the ChIA-PET library data. For a genome-wide negative control of proximity ligation, we constructed a ChIP-PET library using MCF-7 cells with the same E2 induction and ChIP treatment, and generated over 1 million PET sequences (Table 4). The ChIP-PET and ChIA-PET library procedures are almost identical. A key difference is that for the ChIP-PET method, DNA fragments are released from protein-bound chromatin complexes by reverse cross-linking before ligation is done under dilute volumes to circularize the linker-

ligated ChIP DNA (Figure 11), while for the ChIA-PET method, the proximity ligation under dilute volumes is done when the linker-ligated ChIP DNA fragments are still tethered together in chromatin complexes (Figure 11). As a result of this experimental design, the ChIP-PET data would only reveal the enrichment of ChIP DNA fragments, while the ChIA-PET data would reveal both ChIP enrichment and chromatin interaction events. Any "interactions" found in a ChIP-PET library would be the result of random *in vitro* chimeric ligations, mapping errors, and chromosomal aberrations in MCF-7 cells. For an even more general control, we used IgG, which binds to chromatin nonspecifically, to perform a mock ChIA-PET analysis and produced close to half a million PET sequences (435,973).

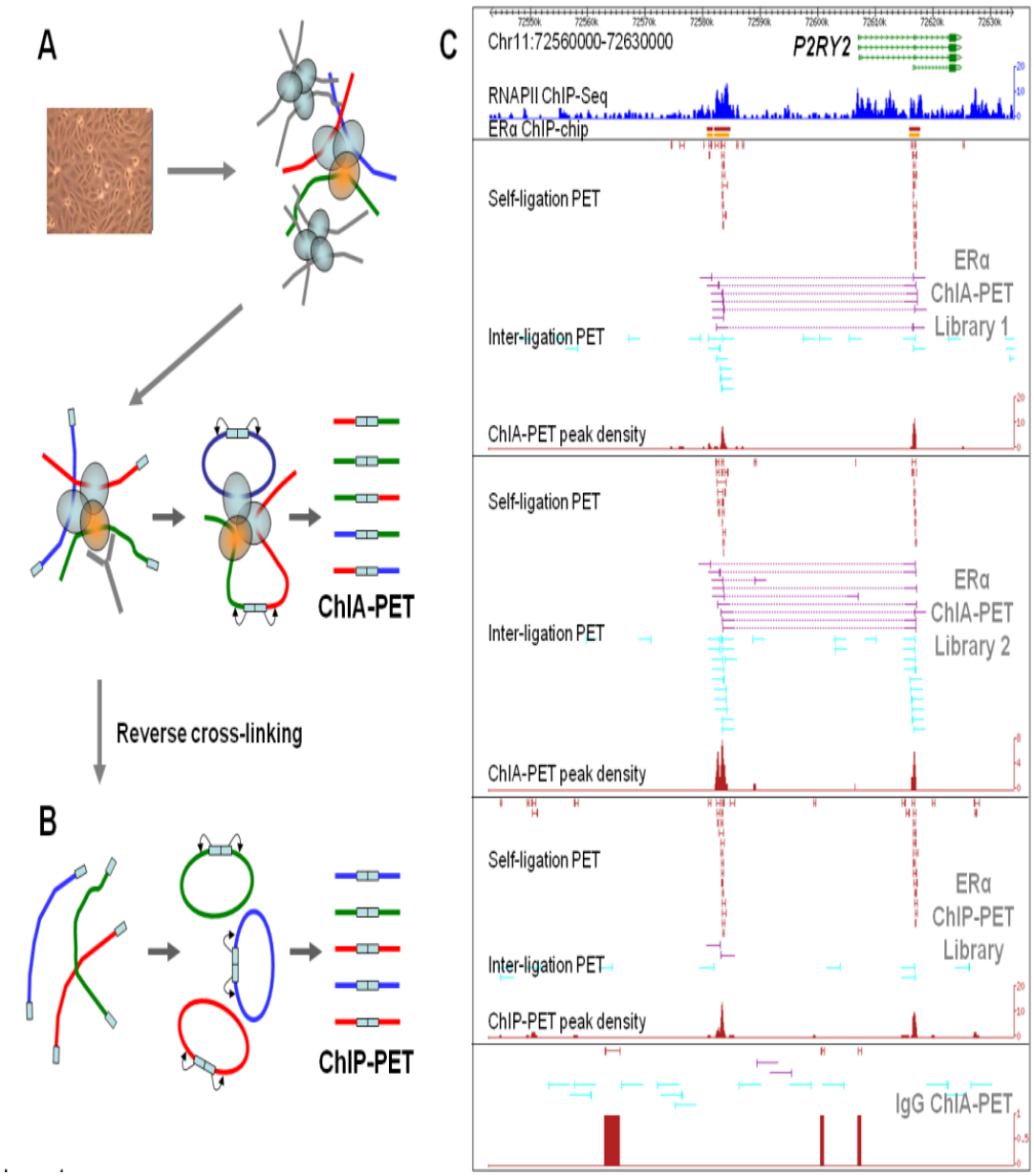


Figure 11. Control libraries.

A. ChIA-PET construction: Tethered DNA fragments in ChIP complexes were ligated to half-linkers containing flanking MmeI restriction sites (the first ligation). The DNA fragments were further ligated (the second ligation) under dilute conditions (further described in Materials and Methods) to produce two kinds of ligation products: "self-ligation" fragments through circularization of DNA fragments, or "inter-ligation" fragments between different DNA fragments in close proximity within the same chromatin complex. Paired-End Tags (PETs) were extracted from the ligation products by MmeI digestion. The released PETs were subjected to Roche 454 pyrosequencing analysis, and the PET sequences were mapped to the reference genome. B. ChIP-PET construction: After adding half-linkers, the tethered DNA fragments were released from protein-bound chromatin complexes by reverse cross-linking. The purified free DNA fragments

were then circularized by self-ligation (the second ligation), followed by PET extraction and sequencing similar to ChIA-PET analysis. C. Mapping and display of ChIA-PET and ChIP-PET data: The *P2RY2* locus on chromosome 11 shows the mapping of ChIA-PET and ChIP-PET sequences in the human genome ChIA-PET browser. The top box shows the *P2RY2* gene model including alternative isoforms, the RNAPII ChIP-Seq density track, and the ERα ChIP-chip data track. The four boxes below show different PET library data mapped in this locus: two ERα ChIA-PET replicates, one ERα ChIP-PET, and one IgG ChIA-PET control. Self-ligation PETs (orange) are shown as two vertical bars with a horizontal line in between. The inter-ligation PETs are presented as vertical bars with extended horizontal lines to represent the average length of ChIP DNA fragments. The intrachromosomal inter-ligation PETs are shown in purple and the interchromosomal inter-ligation PETs are in light blue. The dotted lines (purple) indicate the connection between the two paired interacting tags. The sum of PET-defined ChIP DNA fragments was converted into a ChIA-PET density peak track. *(Note: The ERα control ChIP-PET library was prepared by Ruan Xiaoan's team. The IgG control ChIA-PET library was prepared by Andrea Ho and Ruan Xiaoan. Genome Browser visualization was performed by Hong Sain Ooi, Pramila Ariyaratne, and Yusoff Bin Mohamed.)*

The PET sequences were mapped to the reference human genome sequence assembly (hg18). If a pair of tags aligned in same chromosome, in head-to-tail orientation, and within the upper size range of the ChIP DNA fragments (Figure 10), then this PET was most probably derived from a self-ligation product, and therefore originates from a single DNA fragment. Otherwise, if the paired tags of PET sequences mapped with genomic distances beyond the upper size range of the ChIP DNA fragments or to different chromosomes, they were assumed to be derived from inter-ligated products of two different DNA fragments, with each of the tags originating from one of the paired DNA fragments (Figure 10). The sum of overlapping ChIP DNA fragments reflects the ChIP enrichment and the most overlapped region of a PET mapping cluster indicates the core binding site at the nucleotide level.

ERα binding sites and interactions determined by ChIA-PETs

As expected, the PET sequences of ERα ChIA-PET libraries included both self-ligation PETs and inter-ligation PETs (intrachromosomal and interchromosomal) that are highly enriched at known ERα binding sites. The high numbers of multiple overlapping unique intrachromosomal inter-ligation PETs connecting the two ERα binding sites at the P2RY2

locus suggests a possible chromatin interaction between the two sites (Figure 11). The similar mapping patterns of the two ChIA-PET libraries at specific loci such as P2RY2 (Figure 11), TFF1 (Figure 12), and GREB1 (Figure 13), as well as CAP2 (Figure 14), also suggest that the interactions found by ChIA-PET are reproducible. By contrast, in control libraries, the ChIP-PET library data had only abundant self-ligation PETs at these two ERα binding sites, which supports the notion that the frequent intrachromosomal inter-ligation PETs observed here are specific and not due to random ligation by chance.

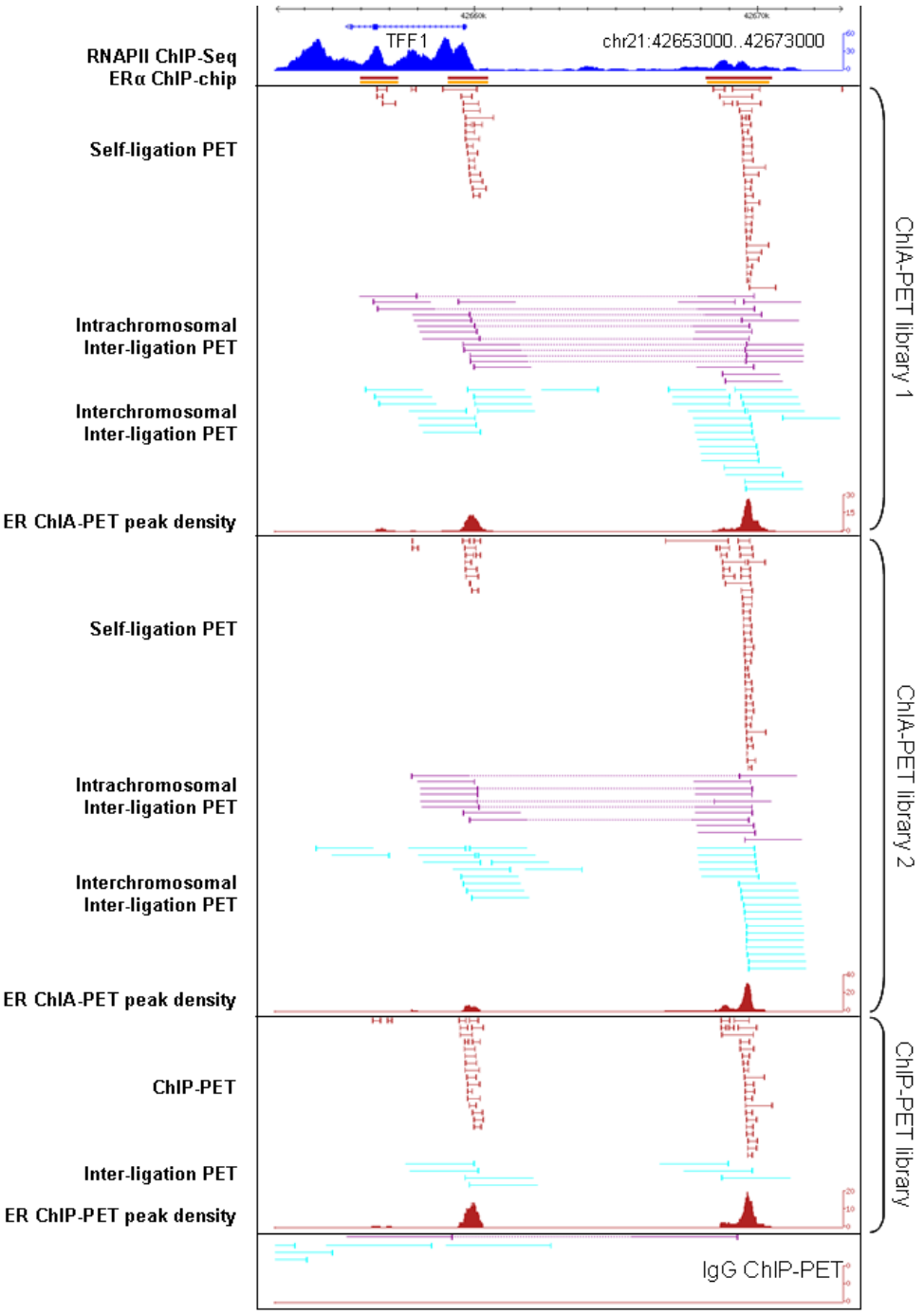


Figure 12. The TFF1 positive control chromatin interaction.

Genome browser views are provided of the TFF1 locus, which is known to have a chromatin interaction (Carroll et al. 2005). Views from the ChIA-PET library 1, 2, ChIP-PET library, and IgG ChIA-PET are provided.

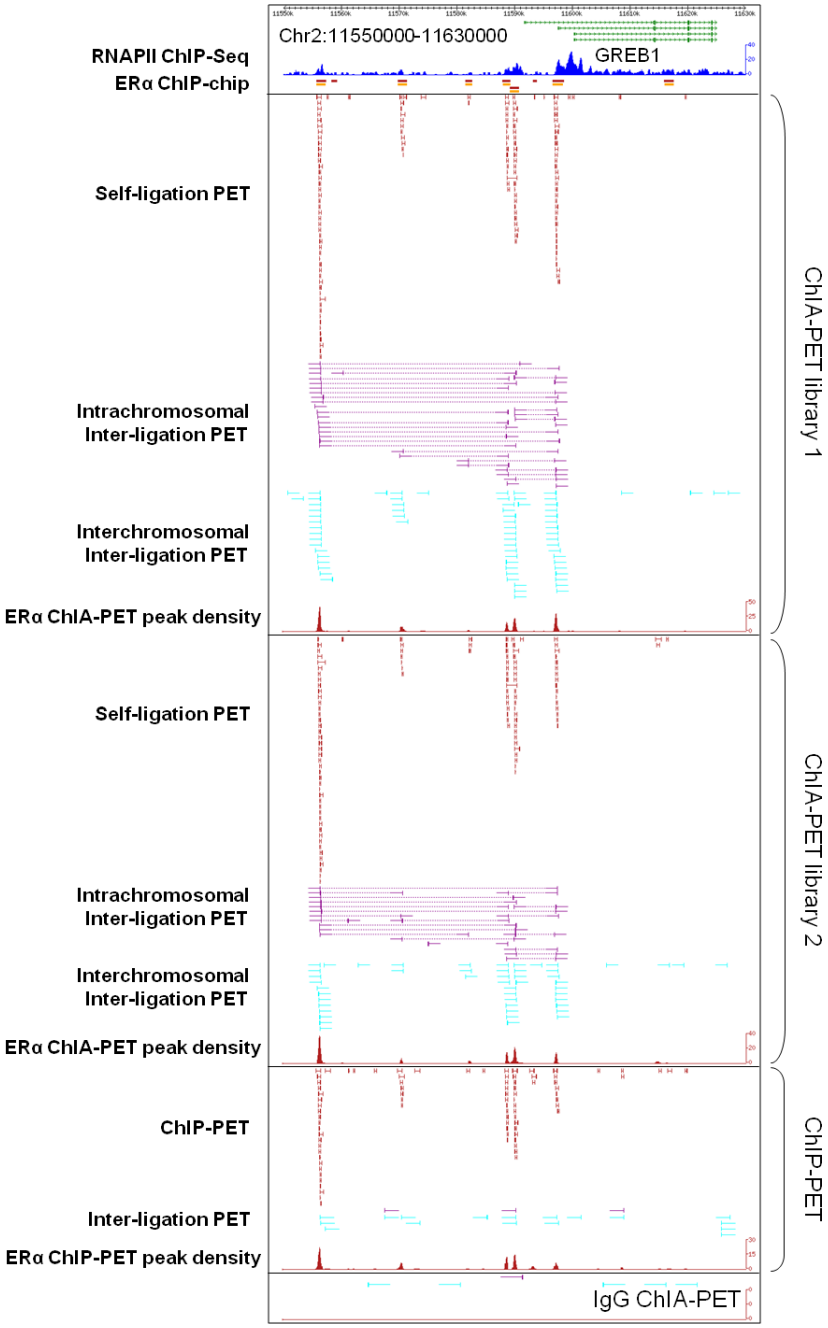


Figure 13. The GREB1 (also known as KIAA0575) positive control chromatin interaction.

Genome browser views are provided of the GREB1 locus, which is known to have a chromatin interaction (Deschenes et al. 2007). Views from the ChIA-PET library 1, 2, ChIP-PET library, and IgG ChIA-PET are provided.

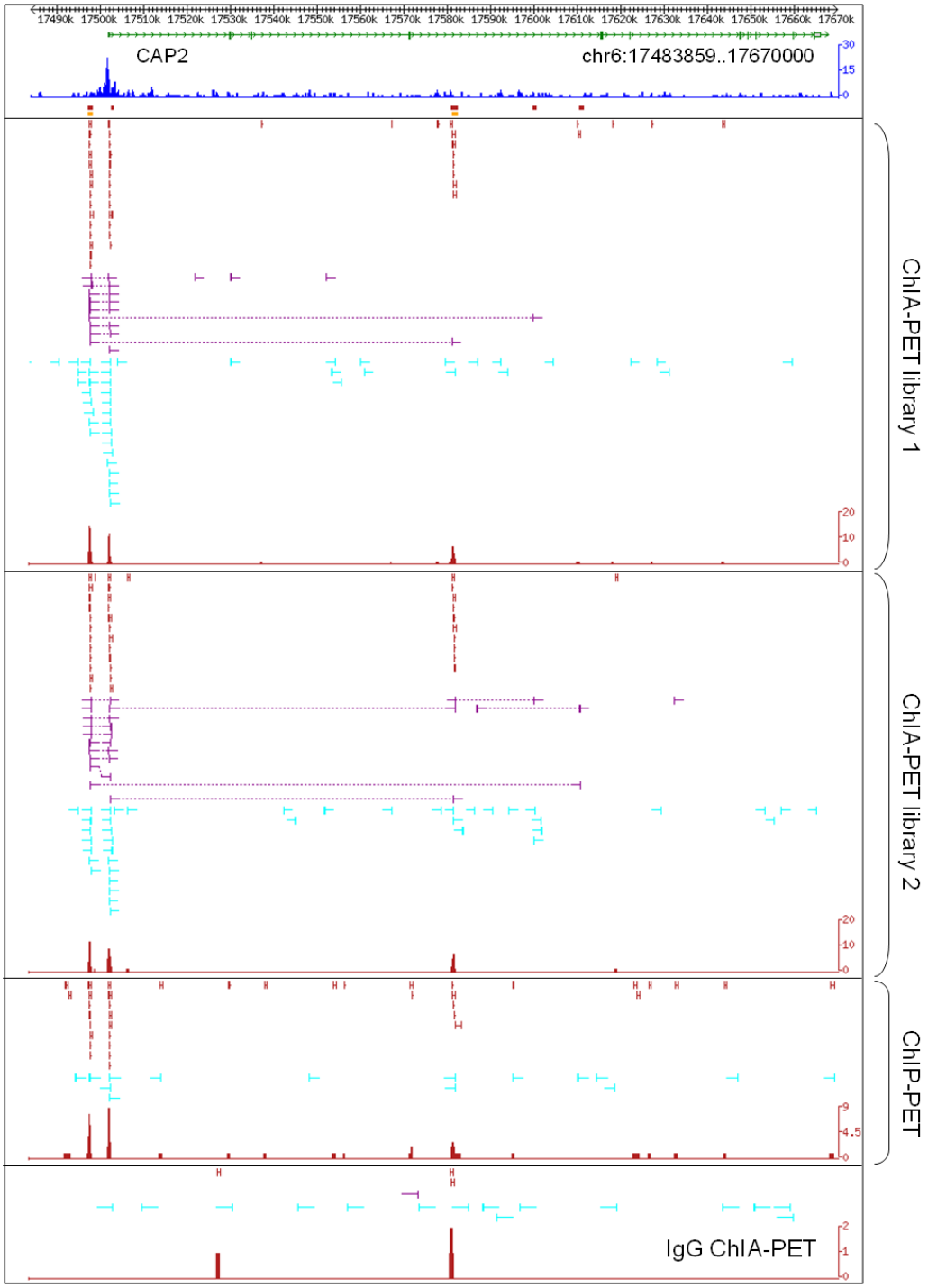


Figure 14. A novel chromatin interaction at CAP2.

Genome browser views are provided of the CAP2 locus, which has a novel interaction identified by ChIA-PET. Views from the ChIA-PET library 1, 2, ChIP-PET library, and IgG ChIA-PET are provided.

We identified 2,179 and 2,720 putative ERα binding sites (FDR≤ 0.01) from the two ERα ChIA-PET libraries. The majority of the binding sites are shared between both libraries

(Figure 15A): 1,459 out of 2,179 (67%) and 2720 (54%) binding sites overlapped, and most of the binding sites that did not overlap had low ChIP enrichment counts (Appendix). Of the shared binding sites, the Pearson correlation between the ChIP enrichment levels of the same sites in the two different replicates was 0.90, indicating that the ChIA-PET procedure is highly reproducible for quantitative measurement of transcription factor binding sites. We then combined the sequences from these two libraries, used the same false discovery rate as a cutoff (FDR≤ 0.01), and identified 4,124 putative ERα binding sites. We compared the ERα binding sites found in the ChIA-PET libraries with the sites identified by the ChIP-PET library data in this study and the ChIP-PET data of our previous study (Lin et al. 2007), as well as the ERα ChIP-chip data (Lupien et al. 2008). 48.6% and 71.9% of the ERα binding sites identified by ChIP-chip and ChIP-PET experiments overlapped with ChIA-PET data. Of the binding sites that were previously identified by ChIP-PET data and that did not overlap with ChIA-PET data, the majority had low PET counts, suggesting they were most likely low occupancy sites. In many examples (Figures 11-14), the self-ligation PETs overlapped at binding sites which correlated precisely with previously reported ChIP-chip (Lupien et al. 2008) and ChIP-PET (Lin et al. 2007) data.

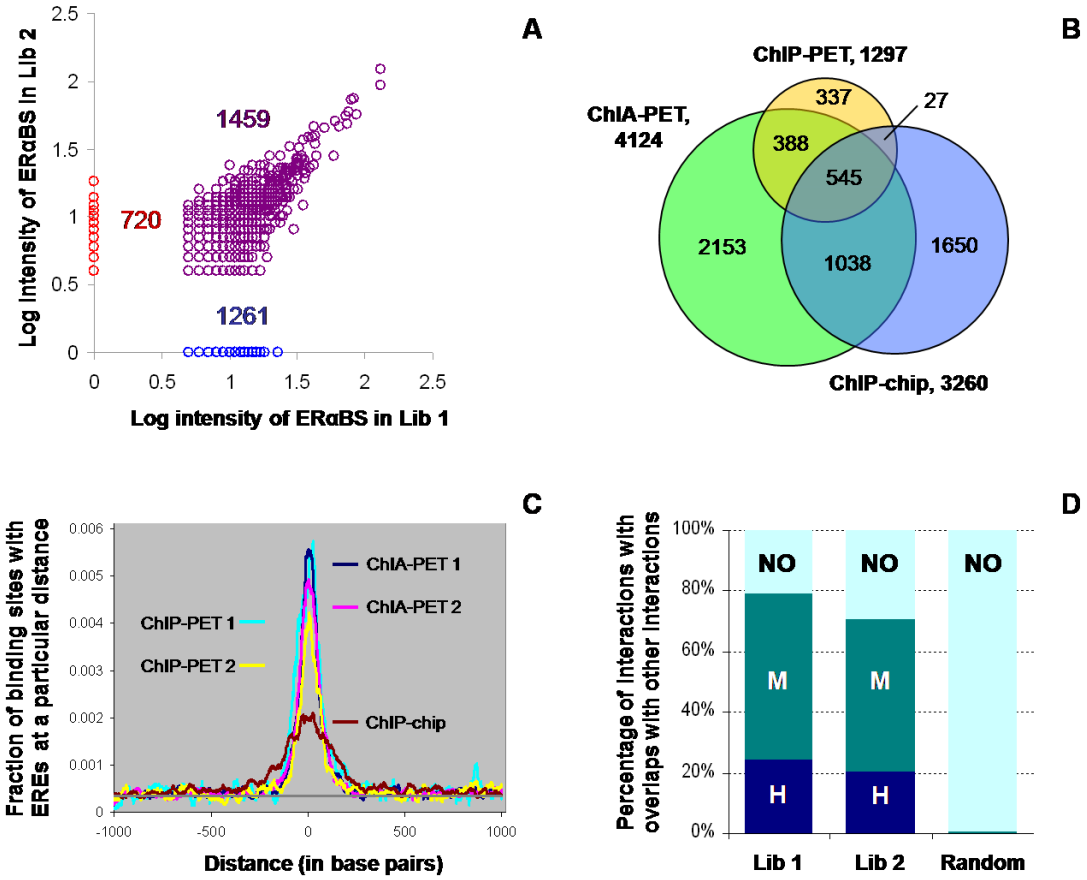
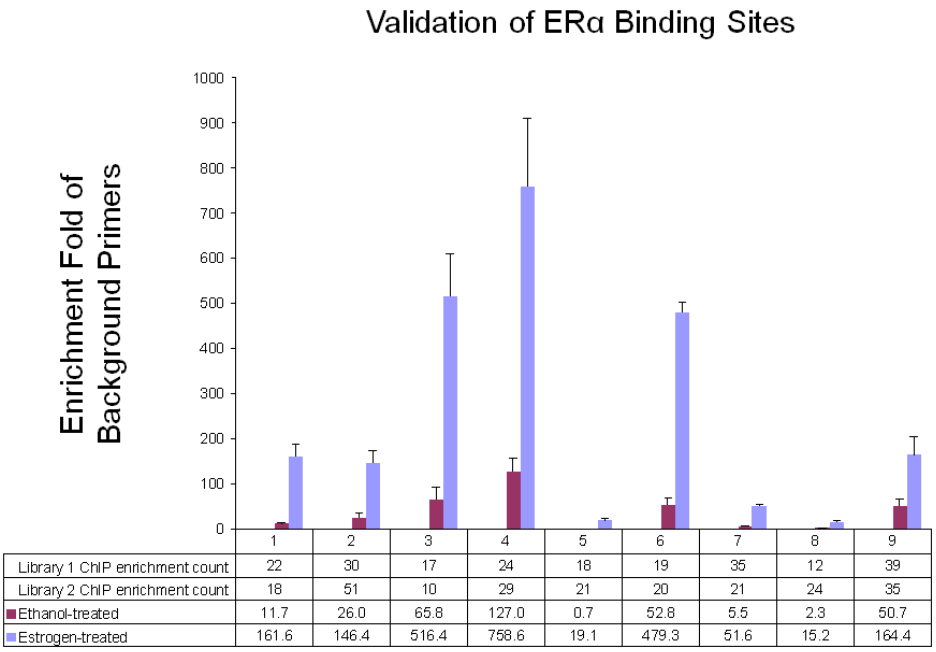


Figure 15. ERα binding sites and interactions determined by ERα ChIA-PET.

A. The ERα binding sites identified by ERα ChIA-PET experiments are largely reproducible. Most ERα binding sites found in one library can also be found in the other library. The enrichment intensities (as measured by overlapping PET counts) of the ERα binding sites in both libraries are highly correlated (Pearson correlation coefficient = 0.9). B. Venn Diagram of ERα binding sites found by different studies. The comparison was performed between the ERα binding sites found by ChIA-PET, ChIP-PET (Lin et al. 2007), and ChIP-chip (Lupien et al. 2008). The combined dataset from the two ChIA-PET libraries identified 4,124 binding sites. The ChIP-PET library in this study and a previous one found 1,297 binding sites. The ChIP-chip experiment found 3,260 binding sites (Lupien et al. 2008). Of the 1,297 ChIP-PET binding sites, 933 (72%) overlapped with the ChIA-PET study, whereas only 27 sites (0.65%) overlapped with the ChIP-chip data solely. Of the 3,260 sites in the ChIP-chip study, about half were overlapped with the ChIA-PET data. C. Distribution of ERE motifs in ERα binding sites identified by ChIA-PET, ChIP-PET, and ChIP-chip data. The background level is shown as a grey line. D. Reproducibility of ERα-mediated interactions from two ChIA-PET libraries. The high confidence interactions (29 and 34) identified in the two ChIA-PET libraries largely overlap. H= high confidence interactions, M= medium confidence and singleton interactions, NO= no interaction PETs found. The overlap percentage by random chance is less than 0.5%. *(Note: Analyses were performed by Han Xu).*

We selected 9 binding sites defined by ChIA-PET not found in ChIP-PET and ChIP-Chip experiments for validation testing by ChIP-qPCR, and all of them showed ChIP enrichment under estrogen induction (Figure 16). We also analyzed the ERα binding sites for the presence of the ERE motif, and found that 1,217 (55.9%) out of 2,179 and 1,456 (53.5%) out of 2,720 binding sites contain at least one ERE motif within 100 bp, which is significantly higher than random background (Fig. 15C). As noted before (Lin et al. 2007), the majority of the ERE motifs were located at the center of the ERα binding sites (Fig. 15C). These analyses collectively prove that binding sites identified by overlapping PET sequences of ChIA-PET data are *bona fide*.



	1	2	3	4	5	6	7	8	9
Library 1 ChIP enrichment count	22	30	17	24	18	19	35	12	39
Library 2 ChIP enrichment count	18	51	10	29	21	20	21	24	35
Ethanol-treated	11.7	26.0	65.8	127.0	0.7	52.8	5.5	2.3	50.7
Estrogen-treated	161.6	146.4	516.4	758.6	19.1	479.3	51.6	15.2	164.4

Figure 16. ChIP-qPCR validation of new ERα binding sites identified by ChIA-PET.

9 sites were selected for ChIP-qPCR validation. All sites show ChIP enrichment, indicating the new ERα binding sites identified by ChIA-PET are *bona fide*. *(Note: ChIP-qPCR was performed by Shi Chi Leow).*

From the two ChIA-PET libraries, we captured 422,813 inter-ligation PET sequences. These PETs could represent proximity ligation products of true chromatin

interaction events, but could also be derived from random *in vivo* interactions, chromosomal aberrations, chimeric ligation products between chromatin complexes, or incorrect tag sequence mapping. To distinguish the inter-ligation PETs representing real interactions from noise, we reasoned that if the proximity ligations of tethered DNA fragments occurred non-randomly, specific inter-ligation PETs between particular regions would be detected at higher frequencies than random background in the dataset. In addition, we rationalized that ERα-mediated interactions would be associated with ERα binding, and ERα ChIP would enrich ERα-mediated interactions for detection. Meanwhile, we were also concerned about the likelihood that ChIP enriched loci with more DNA fragments would result in proportionally higher chances of more inter-ligation PETs, leading to false positives between highly enriched sites. Hence, we conducted statistical analyses to calculate the probability for any overlapping clusters of inter-ligation PETs to occur if the ligations between DNA fragments occurred based on random chance. The assumption in this analysis is that in a ERα ChIP enriched DNA fragment population, if each of the DNA fragments has equal chance to interact with and be ligated to any other fragment randomly, the analysis would calculate the expected level of interaction frequency between two genomic loci and be able to estimate the p-value for the frequency of interactions observed by inter-ligation PETs. This statistical analysis would also neutralize the enrichment effect by ChIP that could potentially result in higher chances of finding overlapping inter-ligations among highly enriched ChIP DNA fragments (Materials and Methods, Chapter 6). In performing this analysis, we first identified 205 overlapping inter-ligation PET clusters (a cluster consists of 2 or more inter-ligation PETs) from one library and 228 clusters from another, suggesting putative chromatin interactions mediated by ERα. Approximately two thirds (64% in library 1 and 66% in library 2) of the putative interactions have ERα binding sites on both sides of the suggested interactions, more than 90% have at least one binding site, and less than 10% have no binding site at all (Appendix). Next, we applied the statistical analysis, and calculated confidence p-values for each putative interaction. We applied Bonferroni correction and used Bonferroni-

corrected p-values < 0.05 as the cutoff to determine high confidence interactions. After performing manual curation to remove some obvious false interactions due to chromosomal aberrations in MCF-7 cells, we identified 29 high confidence interactions from one replicate and 34 from another (Appendix). In total, we identified 56 high-confidence interactions from these two libraries. All of them have at least 3 overlapping inter-ligation PETs. Some of them have more than 10 overlapping PETs. For example, the interaction sites with the highest numbers of overlapping inter-ligation PET clusters at the GREB1 and P2RY2 loci have14 and 13 overlapping inter-ligation PETs in both libraries, respectively. We also observed that most of these high confidence interactions have ChIP enrichments in both of the interacting regions, suggesting genuine interactions mediated by ERα binding (Appendix). Previously characterized chromatin interactions including GREB1 and TFF1 (4 inter-ligation PETs and a Bonferroni-corrected p-value of 4.1E-03 in library 1; 8 inter-ligation PETs and a Bonferroni-corrected p-value of 2.33E-16 in library 2) (Figures 12 and 13) are included in this class. In addition, the category of high confidence interactions includes many new chromatin interactions identified by this study that are at the loci of genes that were previously described as ERα-responsive genes (Figures 11 and 14, Table 5, and Appendix). Interestingly, all high confidence interactions are between sites within an individual chromosome, suggesting that most strong interactions mediated by ERα binding are intrachromosomal. The detection of interchromosomal interactions by ERα, if any, would require much deeper PET sequencing. The high confidence interactions identified in the two ChIA-PET libraries showed a high percentage of overlaps between the two libraries (>70% of library 1 and >80% of library, Figure 15D and Table 6), suggesting that the detection of chromatin interactions by the ChIA-PET method is qualitatively reproducible. With the understanding that the sequencing coverages of these two libraries are still very modest (Figure 17), we believe that further sequencing would make the ChIA-PET library approach quantitatively reproducible.

Table 5. Genes associated with ERα binding and interactions identified in previous studies and in this chapter.

Gene	Identified previously (Carroll et al. 2005; Deschenes et al. 2007; Lin et al. 2007; Pan et al. 2008)			Identified in this study		
	ERαBS	ERα-mediated Interaction	EXP	ERαBS	ERα-mediated Interaction[A]	RNAPII ChIP-Seq
TFF1	Yes	Yes^	Yes	Yes	6.26082E-26	Yes
GREB1	Yes	Yes#	Yes	Yes	4.14542E-26	Yes
NAV2	Yes	No	No	Yes	1.67931E-18	No
SIAH2	Yes	No	Yes	Yes	2.82192E-16	No
P2RY2	Yes	No	No	Yes	1.87406E-24	Yes
TMPRSS3	Yes	No	Yes	Yes	6.26082E-26	No
SLC9A3R1	Yes	No	Yes	Yes	3.79477E-14	Yes
CXXC5	Yes	No	Yes	Yes	4.37714E-18	Yes
CDH26	Yes	No	Yes	Yes	8.07967E-14	No
ZMYND11	Yes	No	Yes	Yes	4.20366E-13	No

Notes: EXP= Expression of the gene is regulated by estrogen induction as detected by microarray experiments; A: Hypergeometric p-value of ERα mediated chromatin interactions detected in either or both ERα ChIA-PET library 1 and 2.

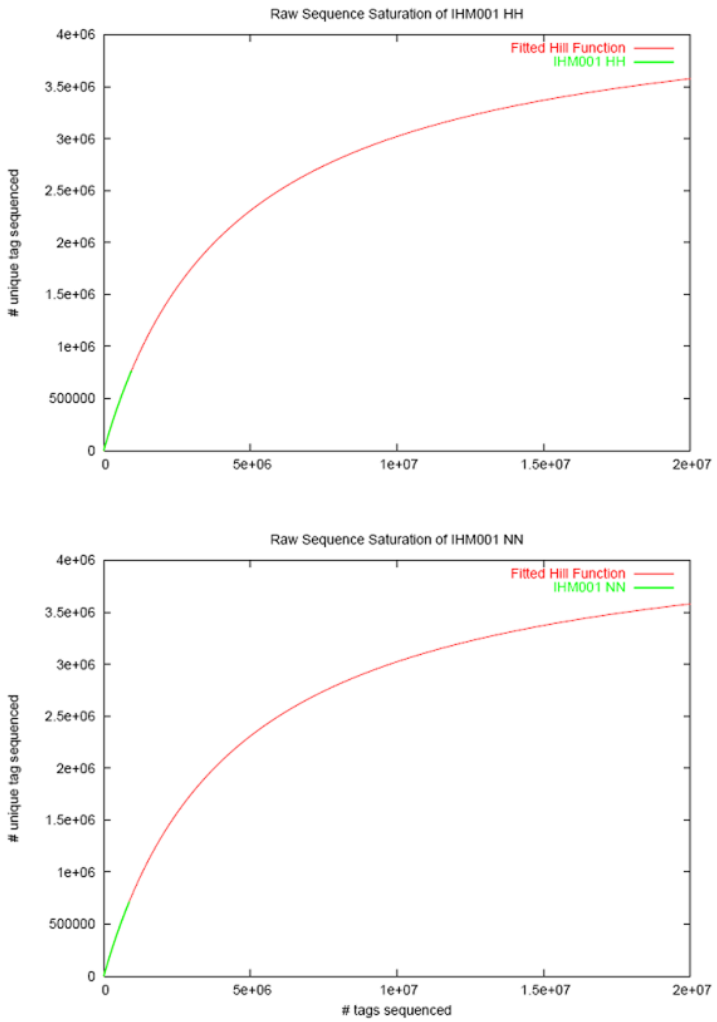


Figure 17. Library sequencing saturation analyses.

We carried out a saturation analysis on each library to assess the sequencing depth reached and to estimate the upper bound unique sequencing attainable. The saturation is modeled using the Hill Function. Based on the redundancy of the sequenced PETs, we found that the ChIA-PET library 1 and 2 were about 16.2% and 17.4% saturated. *(Note: Library saturation analyses were performed by Vinsensius Vega).*

Table 6. Statistics of overlaps between ChIA-PET library 1 and 2 interactions.

	Library 2 high confidence interactions	Library 2 medium confidence interactions	Library 2 inter-ligation singleton PETs
Library 1 high confidence interactions	7	9	7
Library 1 medium confidence interactions	5	24	49
Library 1 inter-ligation singleton PETs	12	43	1,913

Of the remaining putative interactions with less significance ("medium confidence"), the average number of inter-ligation PETs is 2.1 in one library and 2.2 in another library. As expected, the number of medium confidence interactions that involve ERα binding sites is high: 84 of 169 (49.7%) and 78 of 190 (41.1%) interactions from each library have good binding sites on both sides of the interaction. Again, the vast majority of medium confidence interactions are intrachromosomal: 154 of 169 (91.1%) interactions from one library and 170 of 190 (89.5%) interactions from another are intrachromosomal.

By contrast, in the control ChIP-PET libraries (libraries not subjected to proximity ligation, but submitted to ERα ChIP enrichment), we did not observe any high confidence overlapping clusters of inter-ligation PETs. Only abundant self-ligation PETs, singleton inter-ligation PETs scattered along the genome, and a few extremely long intrachromosomal or interchromosomal PET clusters were found in both of the ERα ChIP-PET control libraries (Table 7), indicating significant ERα ChIP enrichment but no detected interactions. The singleton inter-ligation PETs most likely reflect a level of possible technical noise due to artifactual chimeric ligation and mapping errors. Of the few inter-ligation PET clusters detected, most do not have ERα direct binding support (Table 7), suggesting that they were derived from either non-ERα mediated interactions (such as through other protein factors) or non-specific noise. Also as expected, the control IgG ChIA-PET library data had no significantly enriched sites and had no significant inter-ligation PET clusters. Together, the

low numbers of inter-ligation PET clusters in control libraries suggest that even medium confidence intrachromosomal interactions found through ChIA-PET approach in this dataset are likely to be specific, and not due to library construction errors, mapping errors or chromosomal aberrations in the MCF-7 genome. The presence of a few interchromosomal PET clusters in the control libraries suggests that the "medium confidence" interchromosomal PET clusters found in ChIA-PET are most likely non-specific random noise, further indicating that the vast majority of real interactions mediated by ERα binding are intrachromosomal.

Table 7. Statistics of inter-ligation PET clusters in all libraries

	Inter-ligation PET clusters	**Intra-chromosomal, > 5 kb (with ERαBS)**	**Inter-chromsomal (with ERαBS)**
ChIA-PET Library 1	198	183 (115)	15# (2^)
ChIA-PET Library 2	224	203 (133)	20# (0)
Chimeras	3	2* (1)	1 (0)
ChIP-PET Library 1	48	2** (0)	39 (0)
ChIP-PET Library 2 §	64	6** (0)	56 (0)
IgG ChIA-PET	0	0 (0)	0 (0)

Note: Except for the number of all inter-ligation PET clusters, all other numbers for other categories in ChIA-PET replicate 1 and 2 numbers include the manual curation of the high confidence interactions. The intrachromosomal numbers and interchromosomal numbers in bracket indicate those that have ERαBS. § This is the previous library (Lin et al. 2007) which was reprocessed to hg18. There were 635K raw sequences, of which 361K unique PET sequences were derived. 312K (87%) mapped to the genome. 501 ERαBS and 6 intrachromosomal inter-ligation PET clusters were found. * 1 interaction can be found in an amplicon region. The other interaction looks normal, but as it just occurred once, it could be a random noise. ** All these interactions were over 10 Mb. Based on what random mapping of tags would generate, this finding suggests that the interactions found were random. In ChIA-PET libraries 1 and 2, only 2 interactions were over 10 Mb. ^ These appear to be due to genomic structural variations. # These interchromosomals were subjected to manual curation as well. We found that 14 interchromosomal interactions had significant structural variations overlapping with or near the interactions, suggesting that the interactions are not reliable.

We conducted 3C and ChIP-3C analyses to validate selected high confidence and medium confidence interactions detected by ChIA-PET sequencing data (Figures 18 and 19). We found that the intrachromosomal interactions detected by ChIA-PET at the GREB1 and P2RY2 sites are *bona fide* and are ligand dependent by ERα binding. To further validate if the interacting ERα binding sites detected by ChIA-PET data specifically interact with each other, and not non-specifically with other ERα binding sites, we conducted ChIP-3C analysis between two strong ERα binding sites, one located at the GREB1 locus and the other one at the P2RY2 locus (Figure 18). The ChIP-3C result proved that there were no interactions between these two sites although they were both highly enriched by ERα ChIP, suggesting that the ChIA-PET detected interactions are locus-specific, and not solely dependent on high levels of enrichment of the ChIP fragments at the GREB1 and P2RY2 sites, or any enriched ERα binding sites. Moreover, we conducted 3C analysis with multiple points between the interacting fragments, and showed that the interactions found are not solely due to random flexing of the DNA polymer (Figure 19). In addition, we conducted 3C analyses in estrogen-treated and untreated conditions, and found that interaction levels were higher in the estrogen-treated conditions, showing interactions are ligand dependent (Figure 19).

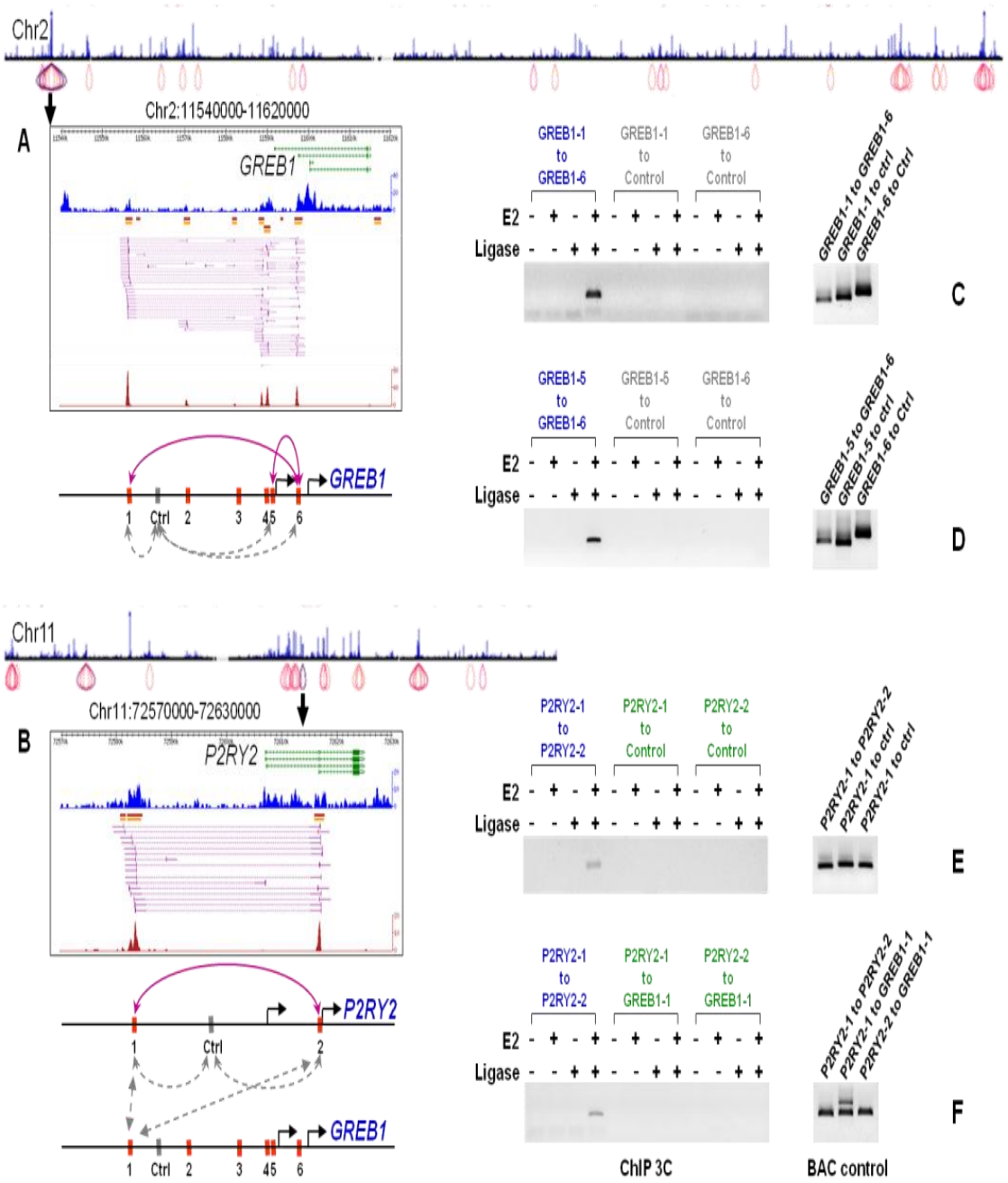


Figure 18. Validation of ChIA-PET interaction data by ChIP-3C analysis.

A. ERα ChIA-PET mapping on chromosome 2. ERα binding sites are shown as blue vertical bar and ERα-mediated interactions are shown as purple rings. The zoomed-in view on the GREB1 locus is shown in an 80Kb window. Six ERα binding sites were identified in this region. The binding sites #1, #5 and #6 were selected to represent long (40Kb) and short (8Kb) genomic distances of interactions for ChIP-3C validation tests (purple arrowed lines). In addition, internal non-interacting sites were chosen to be negative controls of ChIP-3C experiments (grey dotted arrowed lines). B. ERα ChIA-PET mapping on chromosome 11 and the zoomed-in view on the P2RY2 locus. Two ERα binding sites were identified in this region. The binding sites #1 and #2 were tested by

ChIP-3C experiments (purple arrowed line). Internal non-interacting sites were included in the ChIP-3C experiments as negative controls (grey dotted arrowed lines). C. Result of ChIP-3C analysis between the GREB1 binding sites #1 and #6 with negative controls and positive controls. D. Result of ChIP-3C between the GREB1 binding sites #5 and #6 with controls. E. Result of ChIP-3C between the P2RY2 binding sites #1 and #2 with controls. F. Result of ChIP-3C between the GREB1 and the P2RY2 loci. Positive controls of 3C PCR reactions using various primers were tested using digested, mixed and ligated BAC clones from the GREB1 and P2RY2 regions. *(Note: ChIP-3C was performed by the lab of Edwin Cheung, in particular by Pan You Fu).*

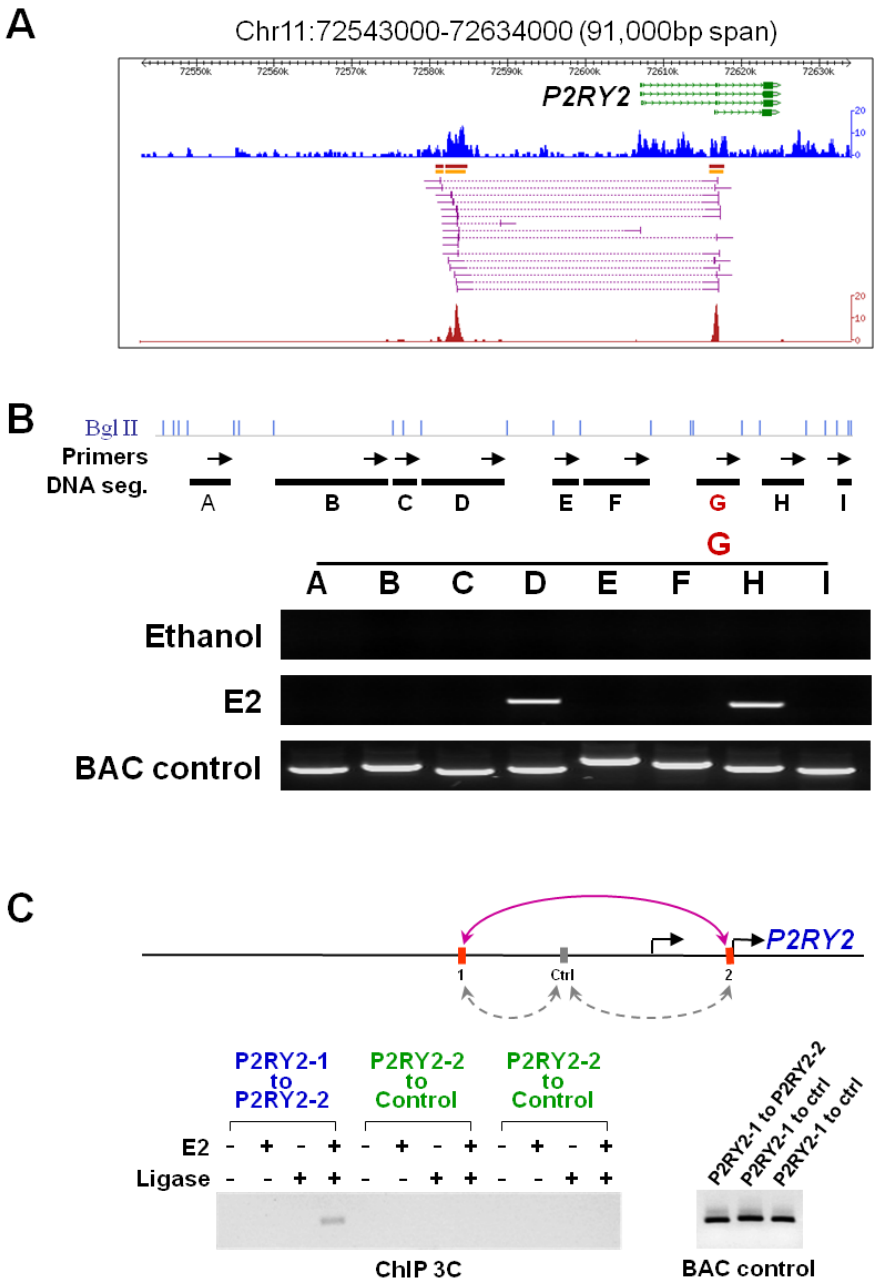


Figure 19. 3C and ChIP-3C validation of a novel chromatin interaction at P2RY2.

A. Genome browser views are provided of the P2RY2 locus, which has a novel interaction identified by ChIA-PET. B. A 3C validation experiment shows that the interaction is bona fide, but would not specify if the interaction is ERα dependent. C. A further ERα ChIP-3C validation experiment further shows that the interaction is bona fide, and also that it is bound by ERα. *(Note: 3C was performed by Mei Hui Liu).*

Further detailed analysis focusing on high confidence interactions revealed that many of the ERα mediated interactions have at least one interacting locus in close proximity to the promoters of putative target genes (Figure 20 and Table 5). RNAPII ChIP-PCR and ChIP-Sequencing data derived from the estrogen-induced MCF-7 cells indicate that these promoters and genes are transcriptionally active (Figure 21). We performed ChIP-3C analysis at these sites to validate the interactions, and conducted RT-qPCR analysis to show that the transcriptional levels of these genes are modulated over the time course of estrogen treatment. These results showing that many transcriptionally active genes are in close proximity to interacting loci suggest that the interaction structures identified by ChIA-PET analysis are functional in regulating the transcription of these genes.

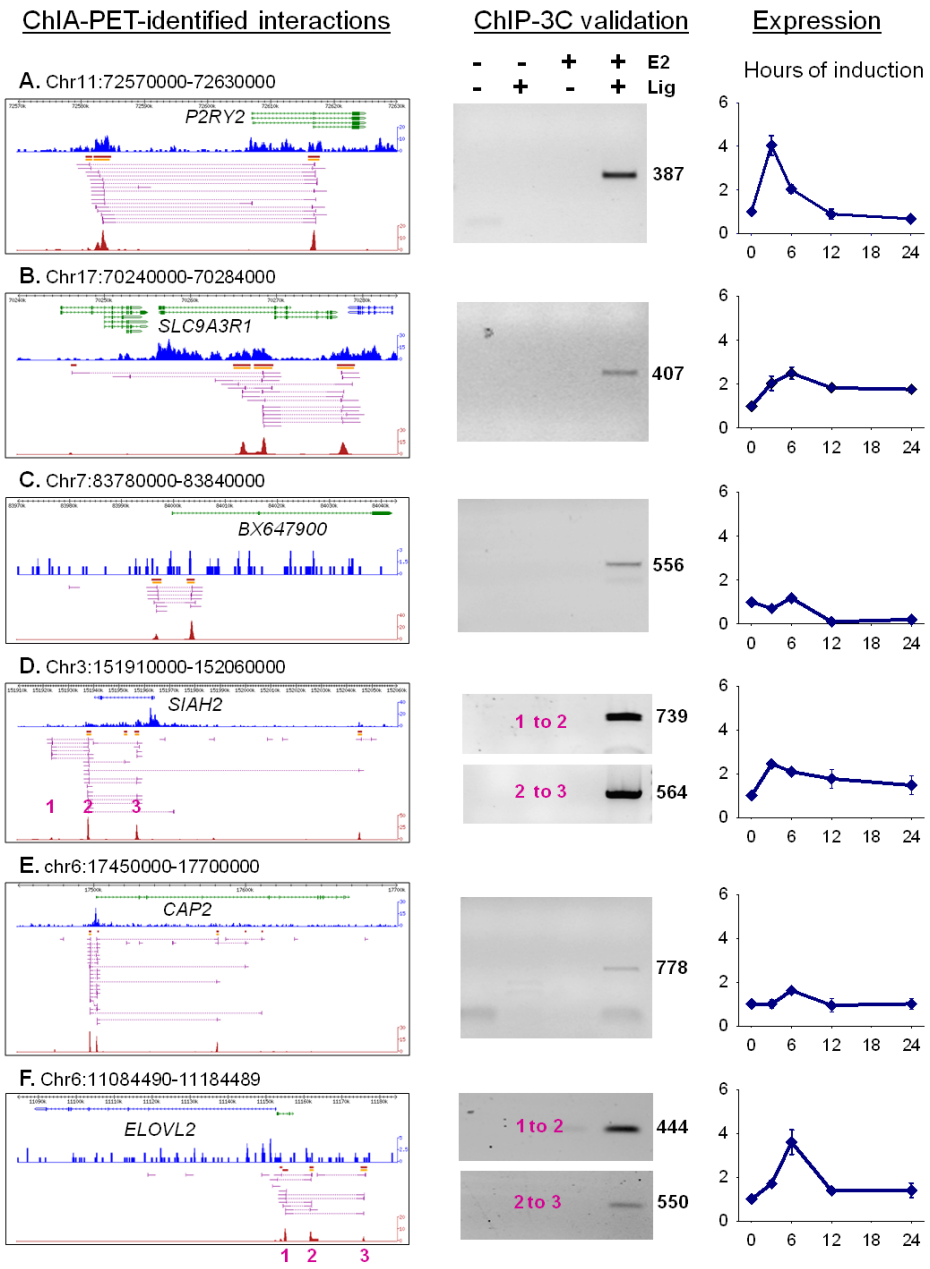


Figure 20. Chromatin interactions and target gene expression.

Examples of 6 loci mapped with inter-ligation PET sequences of the ChIA-PET experiments showing chromatin interactions. Each interaction locus tested in a ChIP-3C experiment is indicated by a number. ERα ChIP-3C experiments were performed with estrogen-treated and untreated MCF-7 cells, and the interactions were validated. The interacting loci are indicated by the numbers in pink. RT-qPCR experiments suggest that the target genes are modulated during estrogen induction. *(Note: ChIP-3C and RT-qPCR experiments were performed by the lab of Edwin Cheung).*

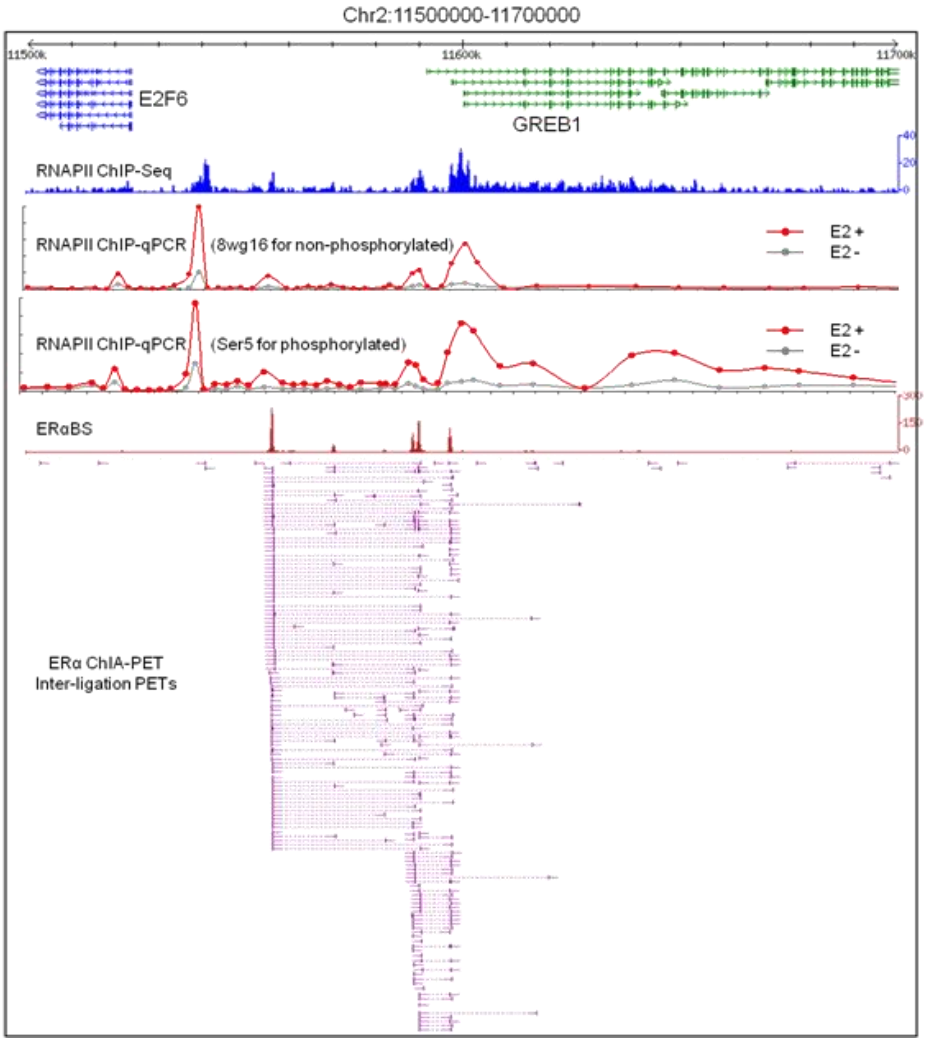


Figure 21. Transcriptional activity at the GREB1 chromatin interaction locus.

We performed ChIP-qPCR and ChIP-Seq analyses to investigate transcriptional activity mediated by the ChIA-PET interaction. The ChIA-PET interaction near GREB1 could recruit RNAPII, which then transcribes the GREB1 mRNA. *(Note: ChIP-qPCR and upstream ChIP-Seq analyses were performed by Pan You Fu, and downstream ChIP-Seq analyses were performed by Liu Jun).*

Discussion

In this study, we demonstrated that the ChIA-PET strategy combined with ultra high throughput sequencing is an unbiased, whole genome approach for *de novo* analysis of chromatin interactions, which represents a significant advance in our ability to study higher-order organization of chromosomal structures and functions. Because a single ChIA-PET experiment is capable of providing two global datasets: the protein factor binding sites and the interactions among the binding sites, this method is conceptually superior to all currently

available methodologies including the recently reported ChIP-Seq (Johnson et al. 2007) that provides only protein/DNA binding information for chromatin interactome analysis.

The most distinctive technical component that sets the ChIA-PET method apart from the 3C and its variants is the introduction of a linker sequence to the junction of two DNA ends in proximity ligation. With this common biotinylated oligonucleotide sequence ligated to all proximity ligation products, we can easily manipulate the proximity ligation products for efficient PET extraction and ultra high throughput sequencing analysis. In addition, the oligonucleotide linker can be used for barcoding purposes. In this way, similar technical conditions can be used to compare two ChIA-PET profiles of two different biological replicates, or even two different samples. This PET strategy is particularly suitable for sequencing analyses using massively parallel tag-based sequencing platforms (Fullwood et al. 2009b). Although in this study we used the Roche 454 sequencing technology, the ChIA-PET method is flexible and can be coupled with any of the tag-based next-generation sequencing systems such as Illumina/Solexa (Barski et al. 2007; Johnson et al. 2007) and SOLiD (Shendure et al. 2005).

The incorporation of ChIP into the ChIA-PET method is also critical. First, ChIP helps to reduce the complexity of proximity ligation products for sequencing analysis. The use of ERα ChIP allowed us to enrich ERα-mediated interactions for sequencing-based detection within the current sequencing capacity. Second, the use of ChIP also elucidates whether a particular protein factor is bound to the interactions, which is information that 3C cannot provide. The detection scope of a ChIA-PET analysis can be well defined by the choice of the protein factor for ChIP. The use of transcription factors such as ERα for ChIA-PET analysis will identify all ERα-mediated interactions. If the target is a general factor for transcription such as RNAPII or TAF, the ChIA-PET analysis would identify all transcription-related interactions. Similarly, if chromatin structure proteins are applied to ChIA-PET, then chromatin structure-related interactions will be revealed. Therefore, there

will be a balance between the specificity and the scope of interactions. Collectively, all specific interactions mediated by protein factors can be identified by the ChIA-PET approach.

The complexity of a ChIA-PET library is very high. In this ERα ChIA-PET experiment, although we generated close to one million PETs for each of the two libraries, they are still far from saturation (Figure 17). With the depth of sequencing analysis presented in this study, we probably only detected highly frequent interactions. Thorough analysis by much deeper sequencing of ChIA-PET libraries will provide comprehensive whole genome views of protein factor-mediated chromatin interactions.

Chapter Four: The Estrogen Receptor α-mediated Human Chromatin Interactome

Introduction

Encouraged by findings in Chapter Three indicating that ChIA-PET can find *bona fide de novo* chromatin interactions, we went on to comprehensively characterize ERα-mediated chromatin interactions in estrogen-treated human breast adenocarcinoma cells (MCF-7). Thus, we generated the first human chromatin interactome map. Using this map, we explored chromatin interactome biology. We asked whether most high quality ERα binding sites (ERαBS) are involved in chromatin interactions. Furthermore, using active promoter and transcriptional marks such as H3K4me3 and RNAPII from ChIP-sequencing as well as gene expression microarray data, we asked whether ERα-mediated chromatin interactions are functionally involved in regulating specific genes.

Results

ERα-mediated chromatin interactome map

Using the ChIA-PET method to examine ERα binding and chromatin interactions in estrogen-treated MCF-7 cells, we generated 5.9 million non-redundant ChIA-PET sequences using next-generation sequencing. 3.6 million (61%) PET sequences were mapped to the human reference genome (hg18), and 1.7 million uniquely aligned PET sequences were processed for further analysis (Table 8). Of the uniquely aligned PET sequences (1.7 million), 0.46 million (25.7%) were considered "self-ligation PETs" as the two tags of each PET mapped within 3 kb of each other (Lin et al. 2007; Wei et al. 2006). 32.1% self-ligation PETs formed overlapping PET groups, representing 9,015 putative ERαBS (False Discovery Rate, FDR$\leq$ 0.01) (Appendix). Besides self-ligation, the tethered DNA fragments in individual chromatin complexes could also ligate with each other. We found 0.11 million (6.2% of uniquely aligned PETs) intrachromosomal inter-ligation PETs (both tags of each PET are from the same chromosome) and 1.2 million (68%) interchromosomal inter-ligation PETs (the tags are from different chromosomes) (Table 8).

Table 8. Summary statistics of PET sequences and mapping to reference genome (hg18).

ERα ChIA-PET	No. of PETs (%)			
Total PET sequences	5,924,521 (100.0)			
Unmapped PETs	2,339,986 (39.5)			
Mapped PETs	3,584,535 (60.5)			
Multiple mapping	1,703,688 (28.8)			
Unique mapping	1,880,847 (31.7)			
	All PETs (%)	Singleton PETs (%)	Overlap PETs[2] (%)	PET Clusters
Uniquely aligned PETs[1]	1,772,119 (100.0)			
Self-ligation PETs	456,264 (25.7)	309,878 (67.9)	146,386 (32.1)	9,015[3]
Intrachromosomal inter-ligation PETs	110,007 (6.2)	101,358 (92.1)	8,649 (7.9)	2,496[4]
Interchromosomal inter-ligation PETs	1,205,848 (68.0)	1,205,185 (99.9)	663 (0.05)	303[4]

Notes: 1. The uniquely mapped PETs were further collapsed if different PET sequences with their two tags were aligned in same genomic location with a difference of only two base pairs. 2. If the alignment locations of the two tagged DNA fragments of a PET overlapped with the two tagged DNA fragment locations of another PET, these two PETs were considered to be "Overlap PETs". 3. The clusters of overlapping self-ligation PETs are based on FDR<0.01. 4. The clusters of overlapping inter-ligation PETs are based on 2 or more inter-ligation PETs.

After filtering out inter-ligation PETs that mapped as singletons (non-overlapping PET sequences) in the reference genome, which is presumed experimental background noise, we identified a set of 2,496 intrachromosomal and 303 interchromosomal overlapping clusters of inter-ligation PETs, representing paired inter-ligating ChIP fragments which indicate potential distant chromatin interactions bound by ERα (Appendix) (Figure 22).

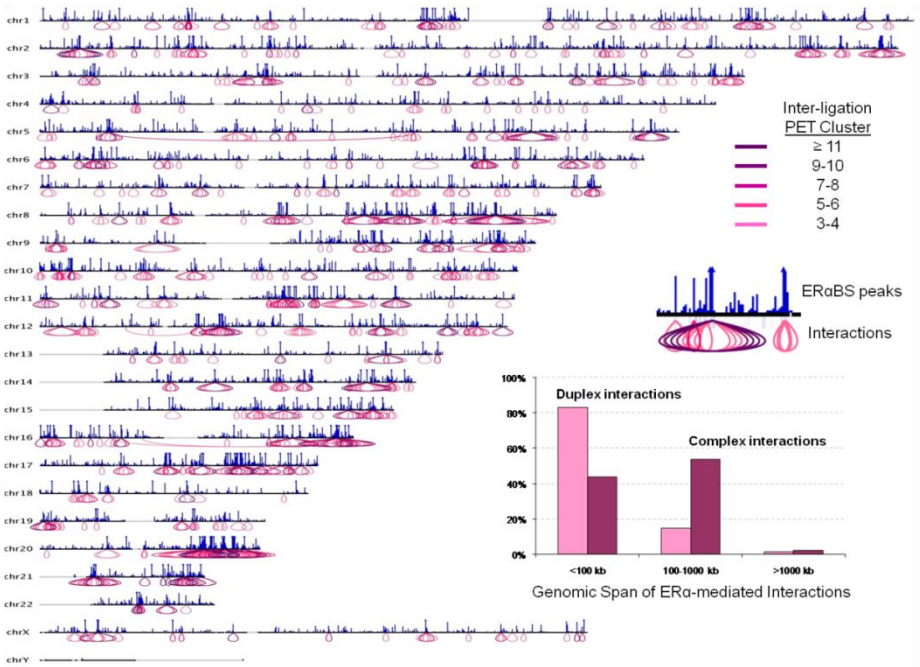


Figure 22. A whole genome view of the human chromatin interactome map mediated by ERα binding.

This map consists of ERαBS (blue vertical bars; the arrow heads indicate peak intensities above the display cutoff limit) and chromatin interactions with 3 or more inter-ligation PETs (purple circles with color gradient corresponding to PET count in each interaction) in the MCF-7 genome. Inset: The length distribution of the chromatin interactions. *(Note: Binding sites and interactions were found by Han Xu and Vinsensius Vega. The Whole Genome Interaction View was developed by Hong-Sain Ooi, Pramila Ariyaratne, and Yusoff Bin Mohamed).*

Each interaction detected by an inter-ligation PET cluster consists of two anchor regions (the two interacting loci) and a loop (the intermediate genomic region between the two anchors), and is therefore called a duplex interaction (Figure 23). While some of the interaction anchors showed weak ERα binding that did not reach an arbitrary cut-off to be

called a binding site, most of the anchors (4,378/5,598=78%) were identified as ERαBS (FDR< 0.01).

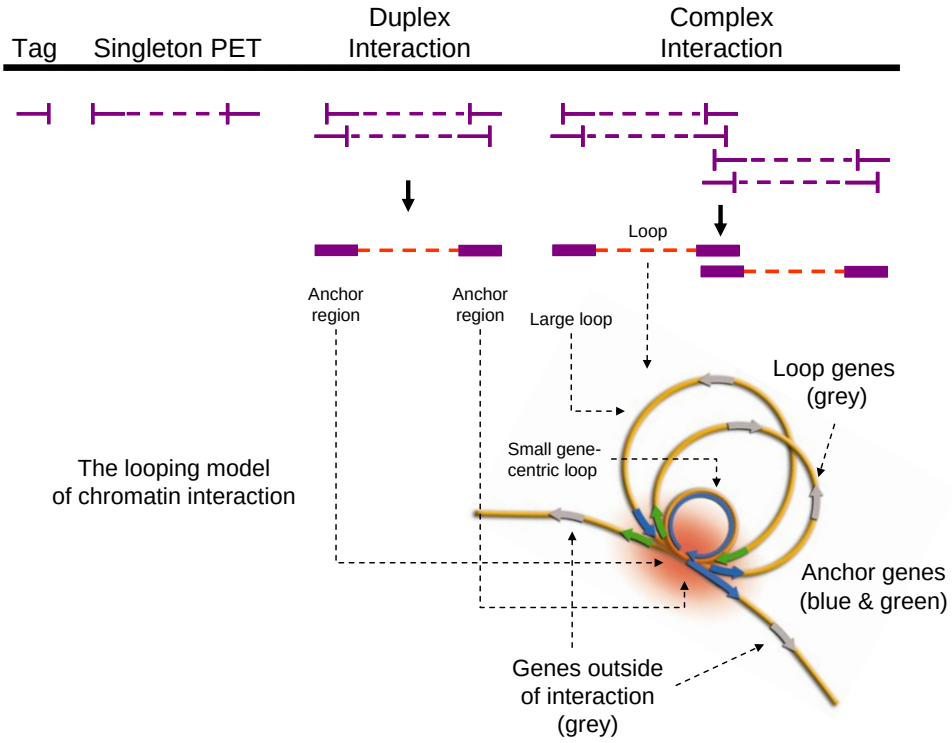


Figure 23. Illustration of structural components of ERα-mediated interactions.

Structures of tags, PETs, duplex interactions, complex interactions, anchors, loops, and genes associated with interactions.

Manual evaluation of the inter-ligation PET clusters revealed that most interchromosomal inter-ligation PET clusters (Appendix) derived from either highly repetitive or highly amplified genomic regions, representing mostly tag mapping artifacts. PET clusters located in these regions were filtered out. To reduce mapping noise, we used more stringent parameters by requiring 3 or more inter-ligation PETs to be present in an inter-ligation PET cluster (as depicted in Figure 22). This left 21 interactions, of which 3 had ERαBS on both sides. The 3 interchromosomal interactions are: chr1:121185663-121186957 to chr19:32423623-32426631 (4 inter-ligation PETs), chr8:126146208-126153795 to chrX:148959660-148960748 (3 inter-ligation PETs), and chr9:129848738-129853141 to

chr10:43408532-43416137 (3 inter-ligation PETs). We consider these 3 interchromosomal inter-ligation PET clusters to be of the highest confidence within the interchromosomal class, pending further validation studies. Nevertheless, all inter-ligation PET clusters were found to only have 2-4 inter-ligation PETs, and to date, have not yielded positive results in FISH validation tests. In contrast, most (2,287, 92%) of the 2,496 intrachromosomal inter-ligation PET clusters did not show such characteristics, and were taken to involve genuine interactions. Several intrachromosomal interactions could also be validated. Hence, ERα appears to primarily mediate intrachromosomal interactions. Our remaining analyses, therefore, focused on intrachromosomal interactions.

In all, we validated a number of selected putative intrachromosomal interactions (16 duplex interactions in 9 different interaction regions) by 3C, ChIP-3C, 4C, and FISH experiments (Chapter 3 shows some validations; others are shown in Figure 24). In each case, the intrachromosomal interactions could be repeatedly confirmed by alternative validation technologies. The 3 interchromosomal interactions were tested by FISH, but to date, have not given positive FISH results, suggesting the interactions are either too weak to be detected by FISH or are noise.

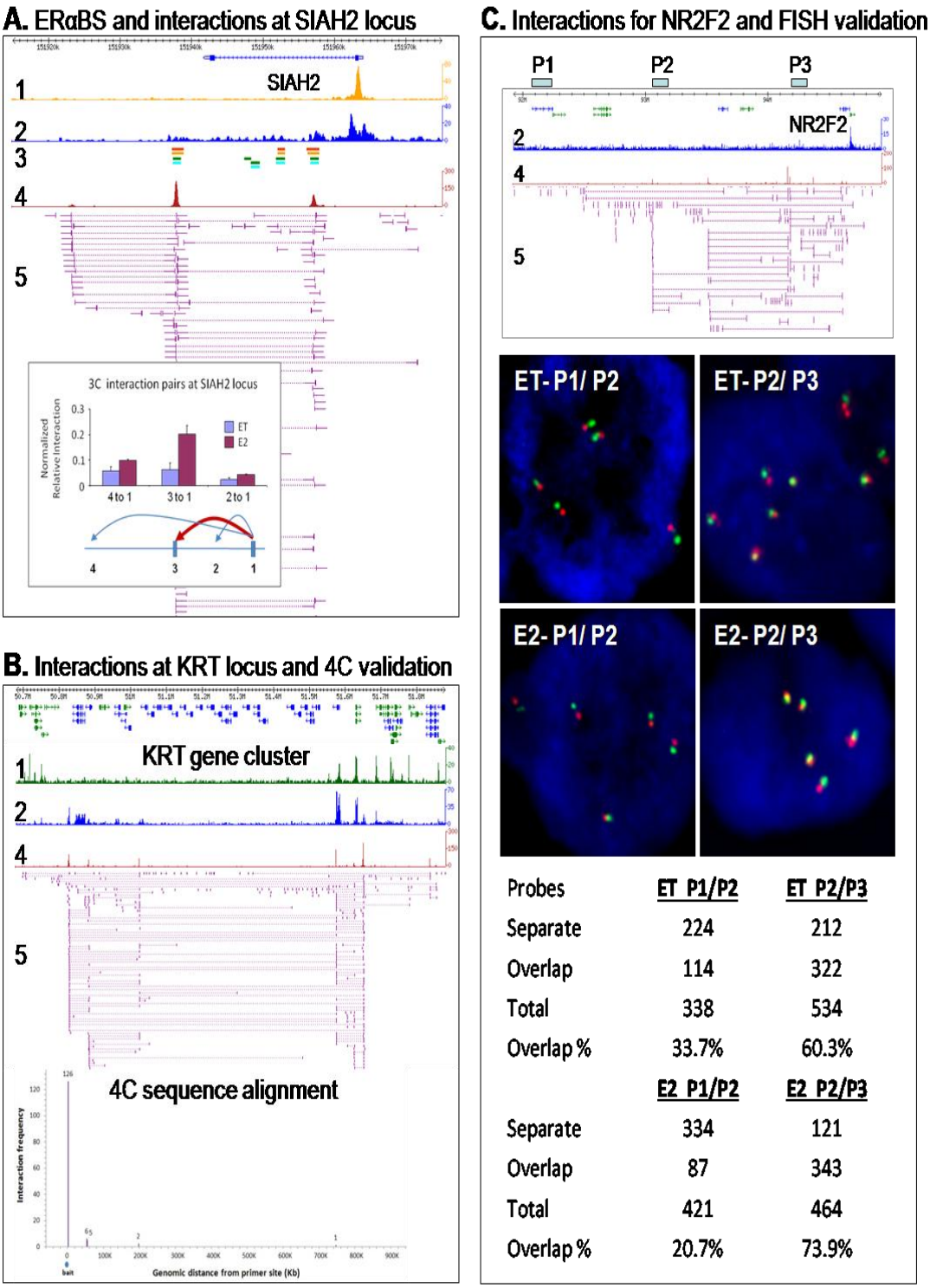


Figure 24. ChIA-PET interaction validations.

A. An example of ChIA-PET data as shown in a genome browser at the GREB1 locus. The data tracks below the gene model are: 1. Peak density of H3K4me3 ChIP-Seq data (green); 2. Peak density of RNAPII ChIP-Seq data (blue); 3. ChIP-chip data of ERα binding (red for high and orange for low confidence) and FoxA1 binding (green for high and light green for low confidence) (Lupien et al.

2008); 4. Density peaks of ERα ChIA-PET self-ligation PETs (brick red); 5. Intrachromosomal inter-ligation PETs (purple), in which tag alignments are shown as vertical bars with extended solid lines to represent DNA fragments and dotted lines for connecting the paired tags; 6. Compact density view of interchromosomal inter-ligation PETs (light blue). Inset: 3C using quantitative (qPCR) validation of chromatin interactions between ERαBS 1, 2, and 6 of the GREB1 interaction complex. B. ERα-mediated chromatin interaction complex at the keratin gene cluster and validation data by 4C. The vertical bar shows the "bait" anchoring detection site, and the horizontal bars show interacting fragments as determined by 4C sequencing. C. ERα-mediated chromatin interaction complex near NR2F2 and FISH validation. P1 represents a control BAC probe. P2 and P3 are test BAC probes near the two anchors of the interaction complex covering >1 Mb. The FISH experiments using the combined probes of P1/P2 and P2/P3 were done in ET (ethanol control) treated and E2 (estrogen) treated MCF-7 cells. The FISH images of the probe pairs show red and green spots when the probes are separated (ET-P1/P2 and E2-P1/P2), and yellow sections between red and green spots when the probes overlap (ET-P2/P3 and E2-P2/P3). The probe overlap rate in E2-P2/P3 (73.9%) is significantly higher than the overlap rate in control experiments of E2-P1/P2 (20.7%) and ET-P2/P3 (60.3%) by Fisher's Exact Test (Fisher's Exact Test 2-tailed p-value is $3.3e^{-59}$ and $9.8e^{-12}$ respectively). *(Note: 3C was performed by Mei-Hui Liu, 4C was performed together with Phillips Huang, and FISH was performed together with Valere Cacheux-Rataboul).*

Interestingly, manual evaluation revealed that many duplex interactions are connected to other duplex interactions, linking three or more anchors into "daisy-chain" aggregated complex interactions, each involving 2 or more duplex interactions (Figures 25-26). While 663 interactions were stand-alone duplex interactions (Figure 25), based on such connectivity, 1,684 of the 2,287 duplex interactions were further assembled into 406 complex interactions (Figure 26). Collectively, we identified 1,009 ERα-mediated interaction regions in this study.

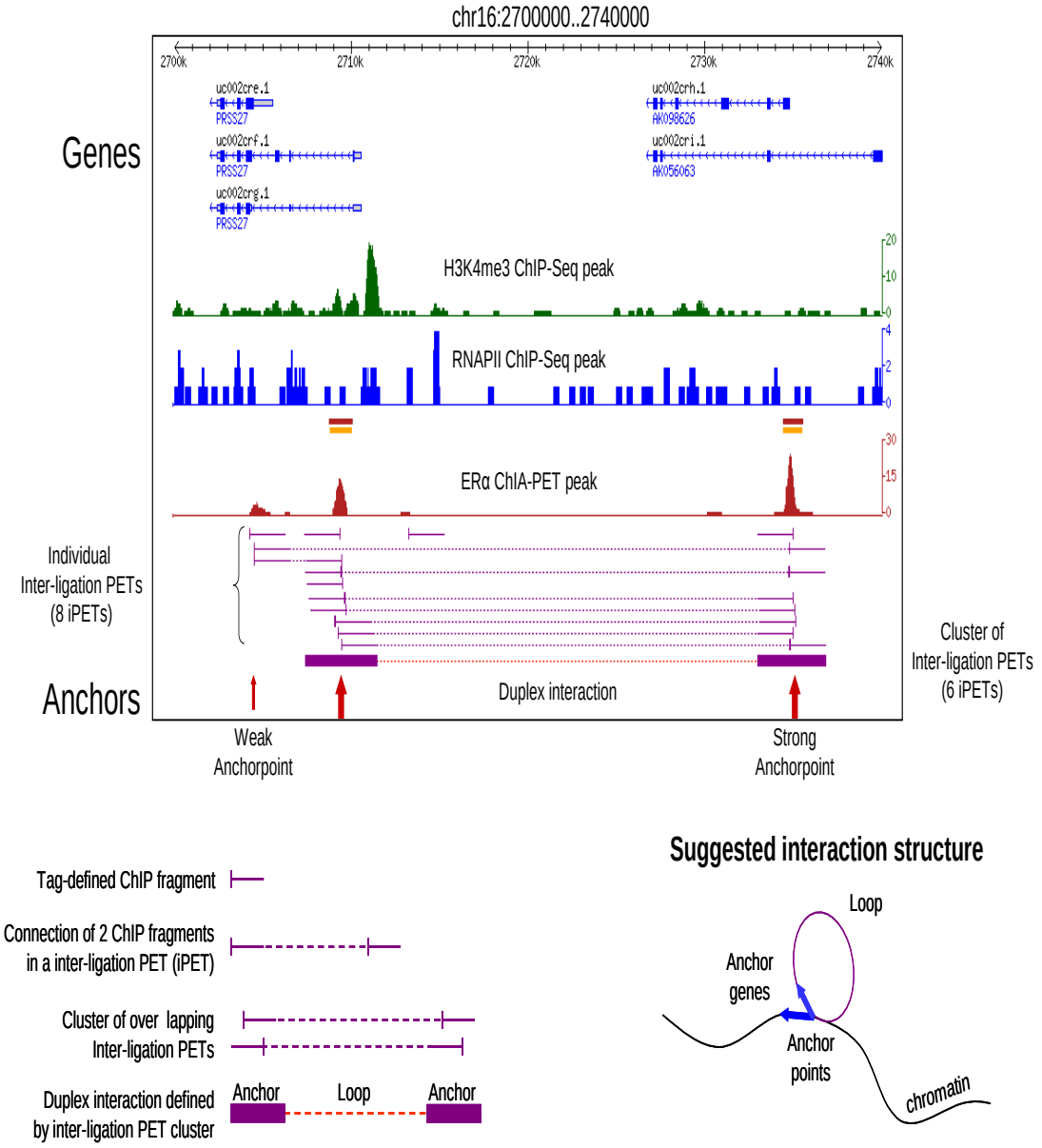


Figure 25. Example of a stand-alone duplex interaction structure.

The data tracks below the UCSC Known Gene isoforms are: 1. Peak density of H3K4me3 ChIP-Seq data (green); 2. Peak density of RNAPII ChIP-Seq data (blue); 3. ChIP-chip data of ERα binding (red for high and orange for low confidence) (Lupien et al. 2008); 4. Density peaks of ERα ChIA-PET self-ligation PETs (brick red); 5. Intrachromosomal inter-ligation PETs (purple), in which tag alignments are shown as vertical bars with extended solid lines to represent DNA fragments and dotted lines for connecting the paired tags; and clusters of overlapping inter-ligation PETs are shown as anchor (thick purple line) and loop (dotted line) representations.

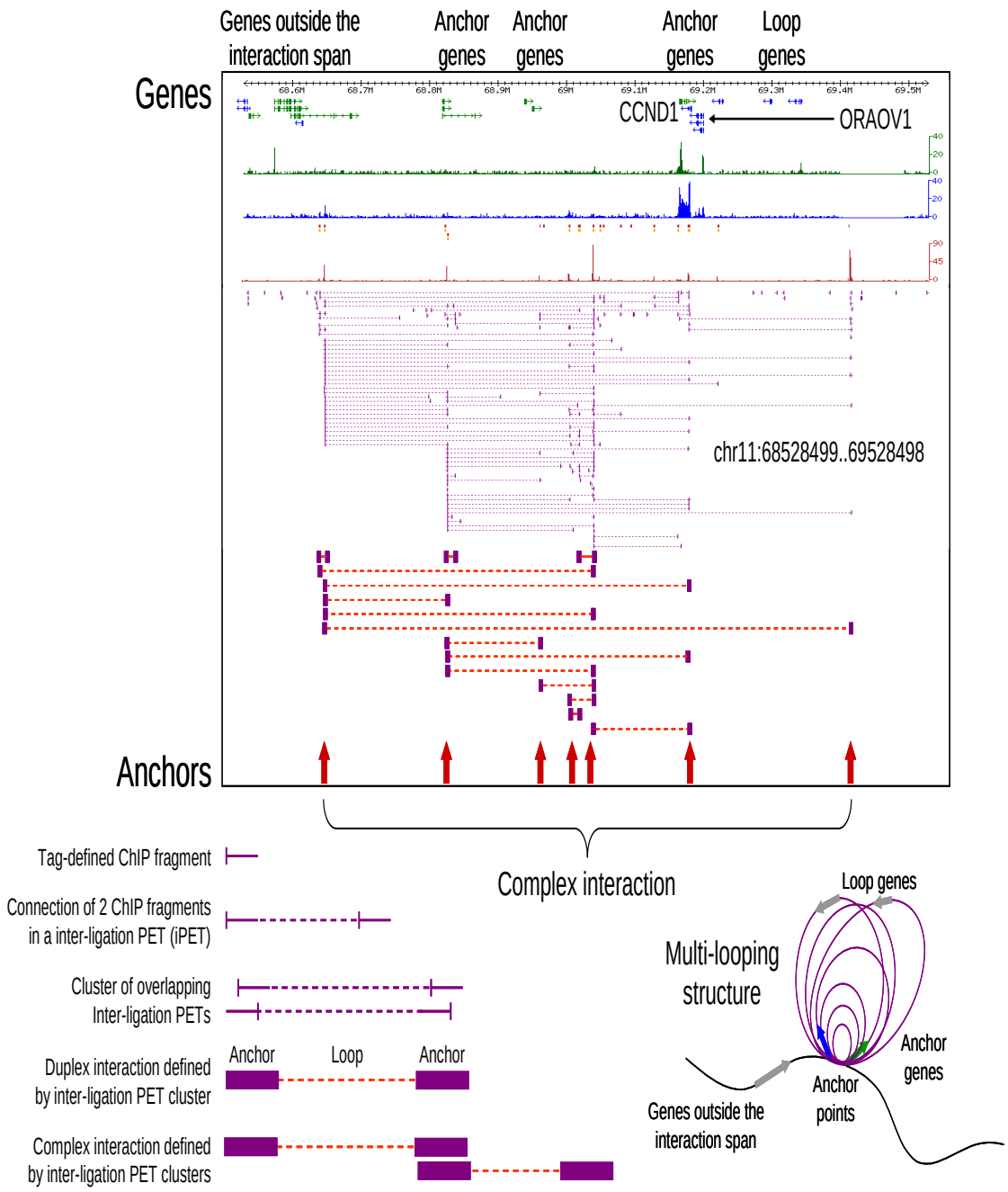


Figure 26. Example of a complex interaction structure.

(Note: Clustering of interactions was performed by Han Xu).

Often, the strongest ERαBS in a complex interaction is either far upstream of the TSS (Figures 27-28), or downstream of the polyadenylation signal sequence (PAS) or within introns (Figures 29-30), with each ERαBS linked through interactions with the promoter; and anchors adjacent to gene promoters may lack significant ERαBS but still have weak ERα

binding (Figures 31). These observations suggest that direct ERα binding might be initiated at multiple distal sites, and then recruit other binding sites as anchors to form an interaction complex that would ultimately engage the transcriptional machinery at gene promoter regions.

chr5:172650000..172970000

Figure 27. Example on chr 5 (STC2) showing stronger binding sites at distal regions than promoters.

Chr8:103670000..103745000

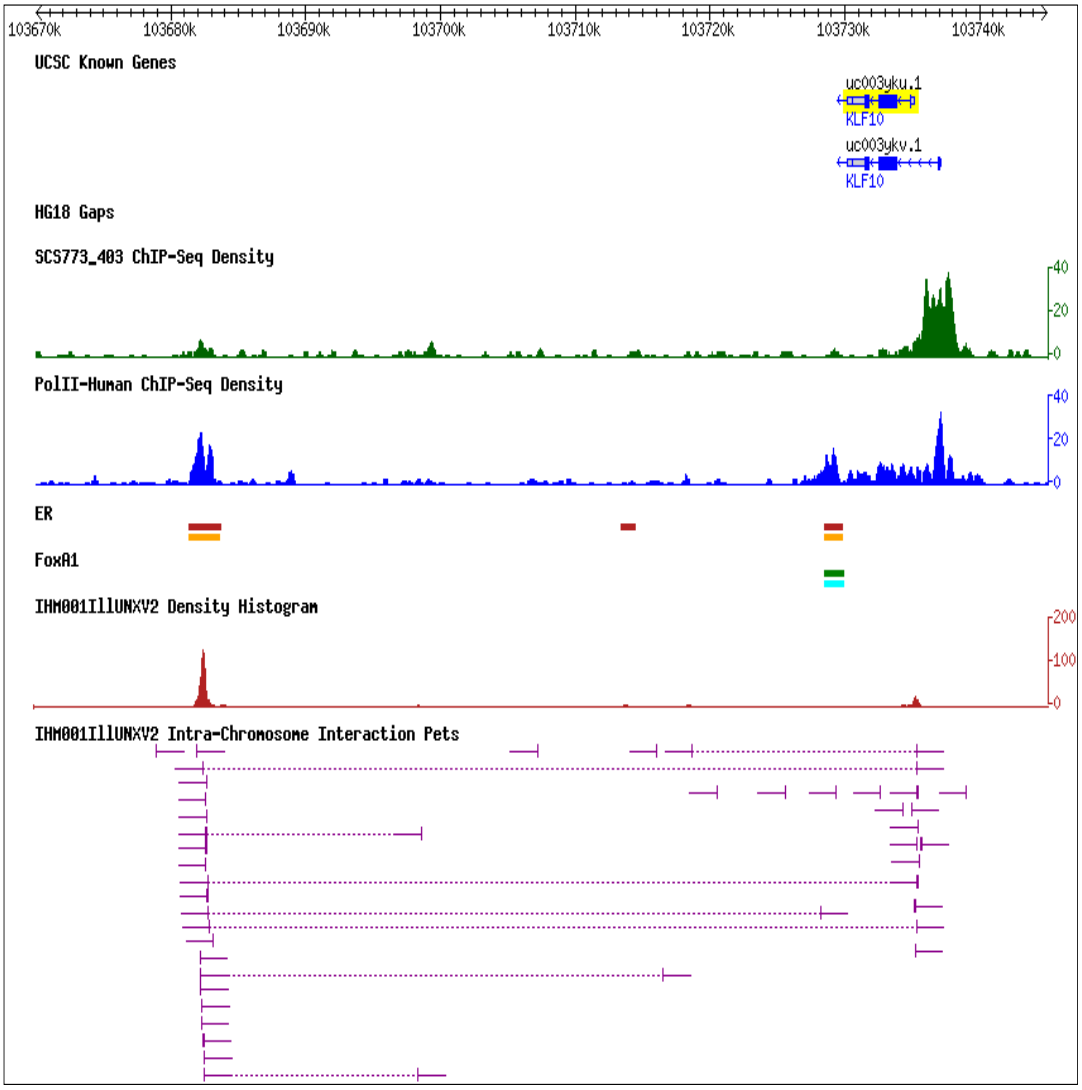


Figure 28. Example on chr 8 (KLF10) showing stronger binding sites at distal regions than promoters.

Chr11:70820000..70903582

Figure 29. Example on chr11 (DHCR7, NADSYN1) showing stronger binding sites at distal regions than promoters.

chr20:54613236-55081395

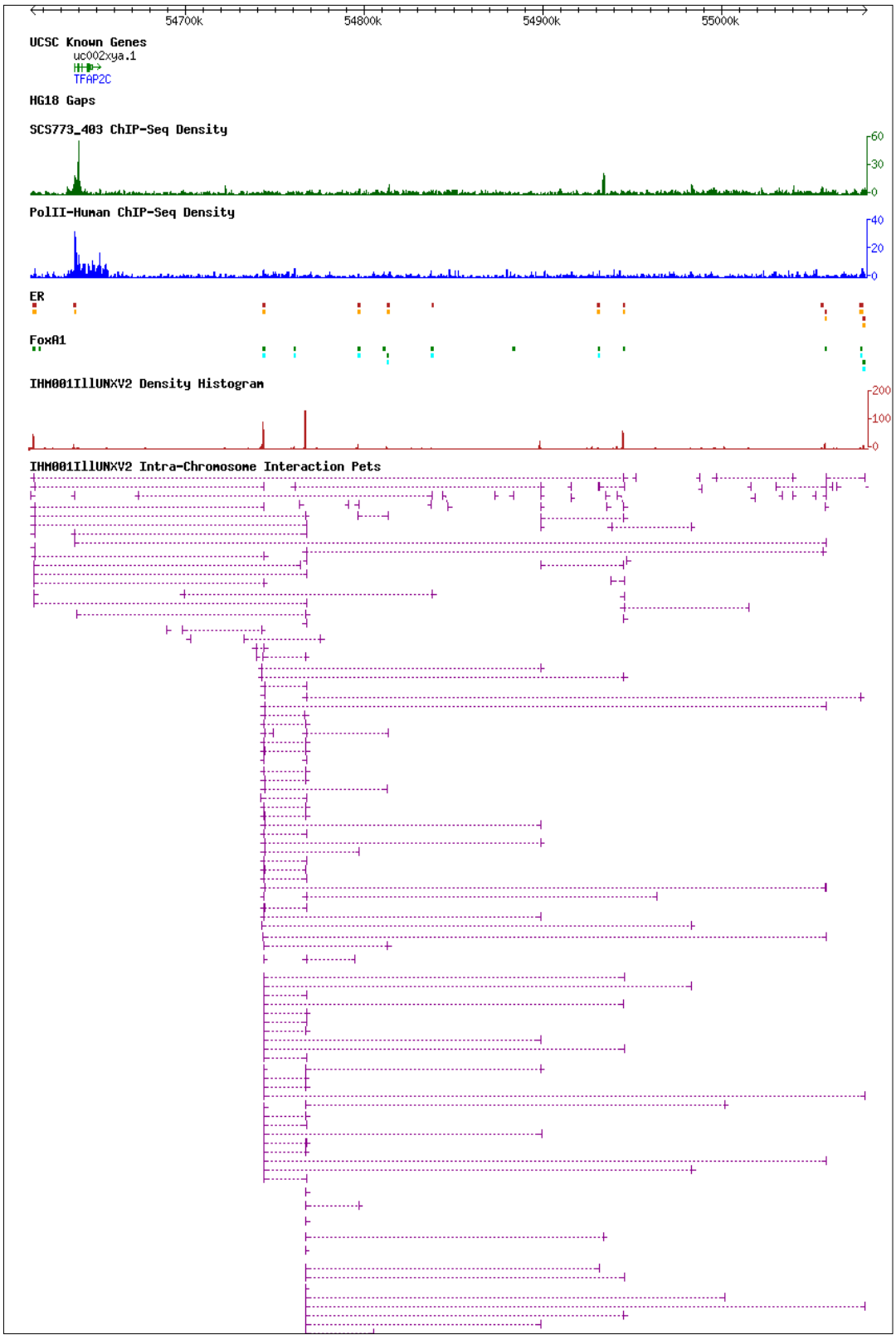


Figure 30. Example on chr20 (TFAP2C) showing stronger binding sites at distal regions than promoters.

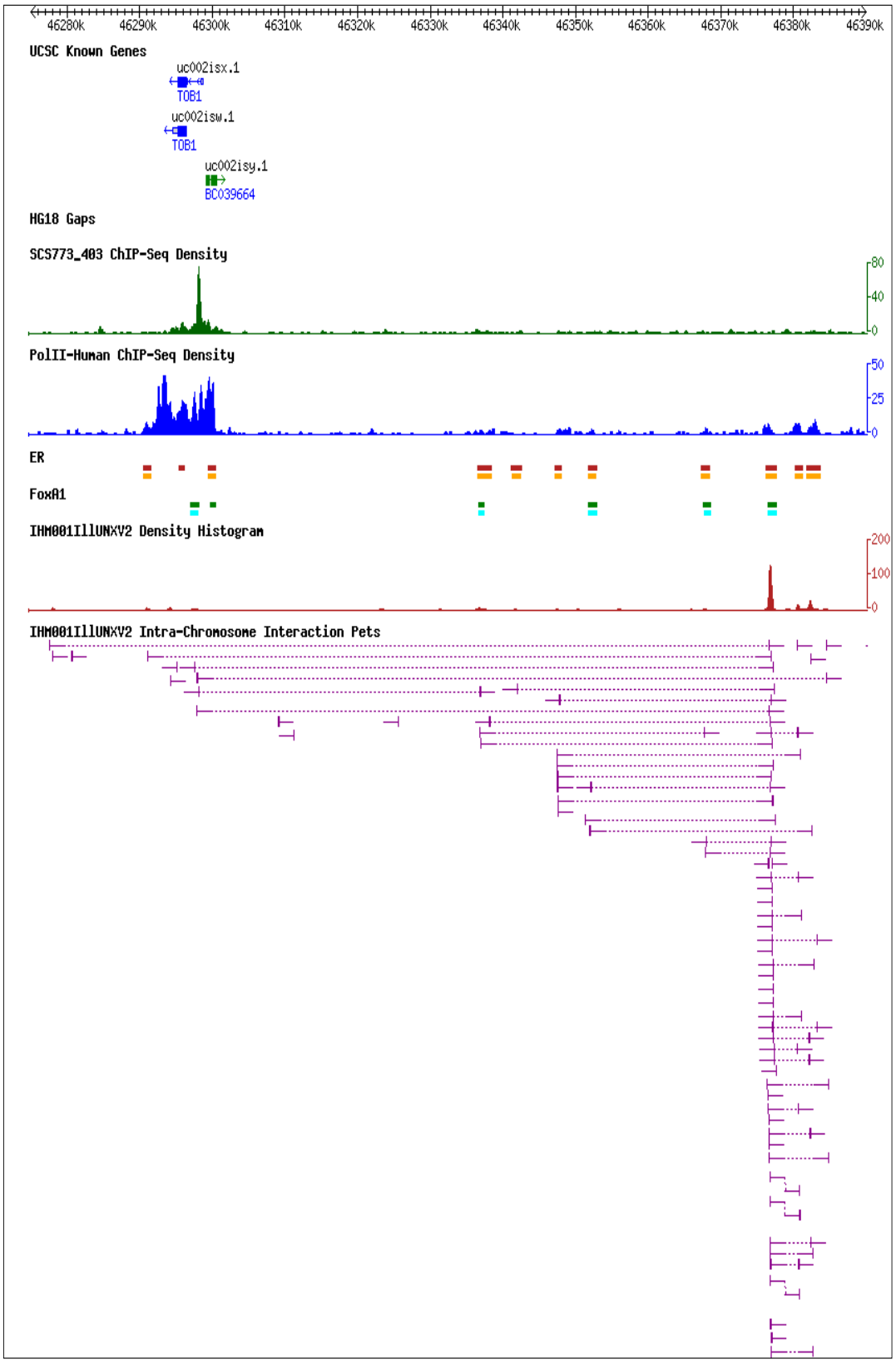


Figure 31. Example on chr17 (TOB1) showing stronger binding sites at distal regions than promoters.

Taken together, the ERαBS and chromatin interactions identified by ChIA-PET data constitute a whole genome chromatin interaction map mediated by ERα binding (Figure 22). The genomic span of most interactions (~80%) is less than 100 Kb, about 20% are in the range of 100-1000 Kb, and very few are over 1 Mb. Complex interactions extend genomic span by connecting multiple duplex interactions. Hence, most complex interactions (~60%) have genomic spans in the range of 100-1000 Kb, with a few that are over 1 Mb (Figure 22; Appendix).

ERαBS association with interactions and other DNA elements

In this interactome map, we asked how many ERαBS are involved in chromatin interactions. We classified the involvement of ERαBS with chromatin interactions into 4 levels: binding sites involved in complex interactions (strong-interactions) (Figures 25, 31 and 32A); in stand-alone duplex interactions (intermediate-interactions) (Figures 24 and 32B); with singleton inter-ligation PETs, which are regarded as "weak-interactions" (Figure 32C) and may require even deeper sequencing to distinguish whether they are signal or noise; and finally, binding sites showing no inter-ligation PETs, which are defined as "no-interactions" (Figure 32D)[1].

[1] Here we used 2 or more inter-ligation PETs to define a cluster as opposed to 3 or more inter-ligation PETs previously. The reason is that otherwise, we would have to create a further category, interactions with only 2 inter-ligation PETs, which would complicate results. This was only done for this section.

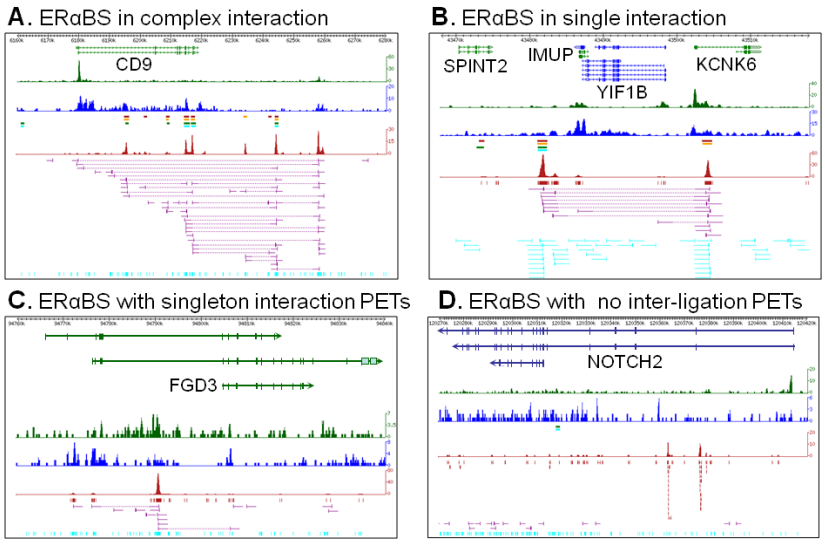


Figure 24. Different classes of involvements of ERαBS with chromatin interactions.

Of the 9,015 putative ERαBS (FDR < 0.01, PET count per ERαBS ≥5), 20% were involved in strong-interactions, 11% in intermediate-interactions, 65% in weak-interactions, and only 3% did not associate with any interactions at all (no-interaction) (Figure 33). ERαBS with low-enrichment (5-19 PET counts per site) are less involved in strong- and intermediate-interactions (16%), and less associated with ERα ChIP-chip data (Carroll et al. 2006; Lupien et al. 2008), while ERαBS with high-enrichment (≥20 PET counts per site) are more frequently involved in interactions (56%) and more associated with ERα ChIP-chip data (Figure 34).

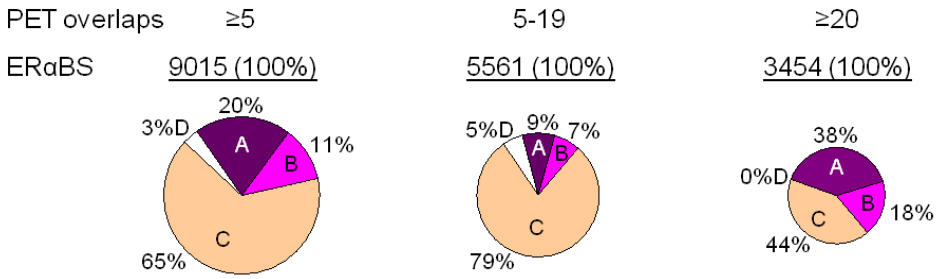


Figure 25. Numbers of ERαBS in different classes of interaction association.

These charts show the number of ERαBS in different classes of interaction association: ERαBS with strong-interactions (complex interactions) are shown in

dark purple; ERαBS with intermediate-interactions (stand-alone duplex interactions) are shown in light purple; ERαBS with weak-interactions are shown in yellow; and ERαBS with no-interactions are shown in white. ERαBS were also classified according to ChIP enrichment levels: all 9,015 ERαBS (left, FDR< 0.001, PET count per ERαBS ≥5); ERαBS with low-enrichment (middle, 5-19 PET counts per site); and ERαBS with high-enrichment (right, ≥20 PET counts per site).

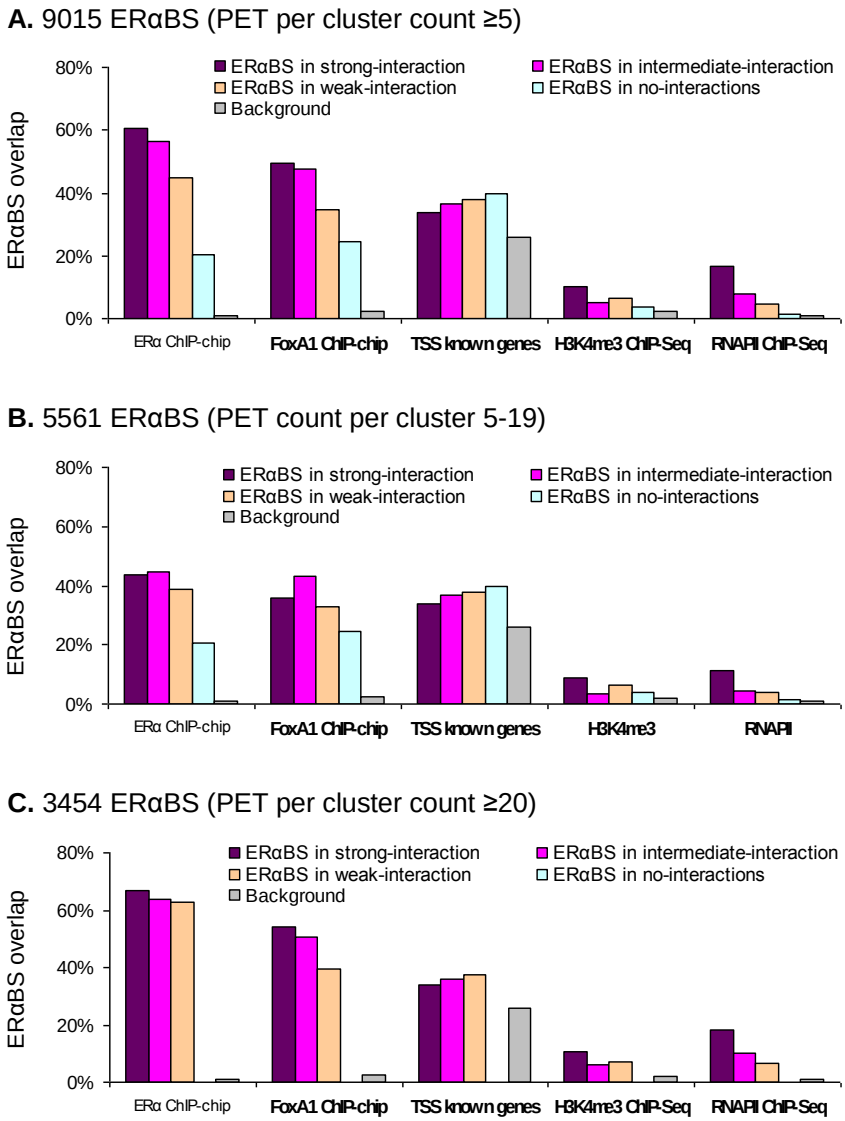


Figure 26. Association of binding sites with interactions and genomic elements.

Association of A. all ERαBS (≥5 PET counts per binding site), B. ERαBS with low ChIP enrichment (5-19 PET counts per binding site) and C. ERαBS with high ChIP enrichment (≥20 PET counts per binding site) involved in strong-interactions (dark purple), intermediate-interactions (purple), weak-interactions (yellow), no-interactions (light blue) and background controls (grey) with ERα and FoxA1 binding sites identified by ChIP-chip (Lupien et al. 2008), Transcription Start Sites (TSS) of known genes (UCSC known genes (Hsu et al. 2006)), and H3K4me3 and RNAPII ChIP-Seq peaks. The background controls

used were singleton ChIA-PETs, and these were also associated with different genomic elements. As the singletons are random in the genome, they show the expected level of association by random chance. The ratios of the number of ERαBS in strong-interactions vs. weak-interactions were calculated for each genomic element. *(Note: H3K4me3 ChIP-Seq was performed by Roy Joseph).*

Furthermore, FoxA1 (a known "pioneer factor" to ERα-chromatin binding (Carroll et al. 2006; Lupien et al. 2008)) binding sites are significantly enriched in association with ERαBS involved in strong- and intermediate-interactions as compared to ERαBS involved in weak-interactions or no-interactions (Fisher's Exact Test 2-tailed p-value = $5.4e^{-15}$, Materials and Methods in Chapter 6, Figure 34). These results suggest that ERαBS with high PET counts are more reliable, and most *bona fide* ERαBS are engaged in chromatin interactions.

Besides FoxA1, we were interested in finding potential co-factors of ERα from our data and assessing whether they were significantly involved in the chromatin interaction detected using the ChIA-PET assay. We used the presence of binding motif (as defined using TRANSFAC weight matrices and criteria) as a proxy to the transcription factor binding. First, we looked for motifs that were enriched in the datasets of binding sites. This analysis was performed in a similar manner as previously described (Lin et al. 2007), except hg18 was used instead of hg17. In addition to finding the ERα motif, we also found many motifs that had previously been found to be associated with ERα binding, such as FoxA1 (Lin et al. 2007). Next, we looked for motifs that were enriched in ERαBS with high- and intermediate-interactions as compared to ERαBS with weak-interactions. To do this, we began with motifs that were enriched in binding sites within interaction regions, and filtered out non-vertebrate motifs, motifs without FDR< 0.05, and motifs with fewer than 50 sequences with at least 1 hit (called "hits"). On the remaining motifs, we employed Fisher's Exact Test to determine which motifs were significantly enriched in the dataset of binding sites with interactions as opposed to those that do not. We also performed Bonferroni correction for multiple hypothesis testing. Vitamin D Receptor (VDR; V$VDR_Q3, Bonferroni-corrected 1-tailed p-value = 0.00095614) was significantly enriched within the pool of binding sites with

interactions. Vitamin D receptor has been shown to be involved with estrogen receptor functioning (Lee et al. 2007). While the FoxA1 motif was found to be enriched in binding sites, it was not found to be significantly enriched between binding sites with and without interactions (Bonferroni-corrected 1-tailed p-value = 1). A possible explanation for the disparity between the experimental findings and the findings from motif predictions with respect to FoxA1 is that not all motifs are occupied by FoxA1 proteins, such that even though the number of motifs might be similar between the two datasets, the levels of occupation by FoxA1 protein are higher in the binding sites with interactions category. In addition, while the ERE motif was found to be enriched in binding sites, it was not found to be significantly enriched between binding sites with and without interactions (ER; V$ER_Q6, Bonferroni-corrected 1-tailed p-value =1). This finding suggests that the significant enrichment of VDR within the pool of binding sites with interactions is not due to similarities between it and the ERE motif. TRANSFAC analyses are given in the Appendix.

Next, we analyzed the relationship between ERαBS and gene promoters, and found that ERαBS are rarely at the transcription start sites (TSS) of known genes (UCSC known genes (Hsu et al. 2006)) (Figure 34), which is consistent with early experimental data showing that most ERαBS are distal to gene promoters (Carroll et al. 2005; Carroll et al. 2006; Lin et al. 2007).

To further investigate the involvement of ERαBS in transcription activation, we generated genome-wide histone H3 lysine 4 trimethylation (H3K4me3) and RNA Polymerase II (RNAPII) ChIP-Seq data from MCF-7 cells under estrogen induction (Materials and Methods, Chapter 6). H3K4me3 is a histone modification mark specific to active promoters (Barski et al. 2007), and the presence of RNAPII is strong evidence for genes that are actively transcribed (Phatnani et al. 2006). H3K4me3 and RNAPII marks are significantly enriched in ERαBS with strong- and intermediate-interactions as opposed to ERαBS with weak-interactions (Fisher's Exact Test 2-tailed p-value = 0.00034827 and $2.5e^{-17}$ respectively)

(Figure 34), suggesting that ERα-mediated interactions are associated with transcriptional activation, by potentially employing long-distance looping to bring remote ERαBS close to gene promoters. Thus, this explains why H3K4me3 and RNAPII marks can be found even at distal ERαBS (Figure 31A, B).

To further understand transcriptional activation with respect to ERαBS with and without interactions, we examined the percentages of upregulated and downregulated genes. As previously described, genes in proximity to transcription factor binding sites such as ERα and p53 appear to be more likely to be upregulated than downregulated (Lin et al. 2007; Wei et al. 2006). Considering all 9,015 binding sites, more binding sites are associated with downregulated genes (418) than those associated with upregulated genes (378); however, such binding sites tend to be low-enrichment ERαBS. This association is not significant (Fisher's Exact Test 2-tailed p-value = 0.157) When high-enrichment ERαBS (≥20 PET counts per site; 3454 binding sites) are considered, more binding sites are associated with upregulated genes (176) than those associated with downregulated genes (128). This association is significant (Fisher's Exact Test 2-tailed p-value = 0.00575). Partitioning all binding sites, we found that ERαBS with strong- and intermediate- interactions tend to have significantly better association with upregulated genes than binding sites with weak- or no-interactions (Fisher's Exact Test 2-tailed p-value = 0.000809). This trend continues into the high-enrichment ERαBS, although significance testing failed to show significance (Fisher's Exact Test 2-tailed p-value = 0.640). By contrast, we found that ERαBS with strong- and intermediate interactions are significantly under enriched in downregulated genes compared with binding sites with weak- or no-interactions (Fisher's Exact Test 2-tailed p-value = 0.0000187). This trend continues into the high-enrichment ERαBS, and is significant (Fisher's Exact Test 2-tailed p-value = 0.000249).

Table 9. Upregulated and downregulated genes near ERαBS.

	Strong	Intermediate	Weak	No	Background
ERαBS (≥20 PET counts per site)					
UCSC known genes transcription units	33.8%	36.0%	37.7%	N.A.	25.8%
Downregulated gene transcription units	2.5%	3.0%	5.1%	N.A.	2.2%
Upregulated gene transcription units	6.0%	3.8%	4.9%	N.A.	1.7%
ERαBS (5-19 PET counts per site)					
UCSC known gene transcription units	34.0%	37.0%	37.7%	39.7%	25.8%
Downregulated gene transcription units	4.0%	5.4%	5.2%	7.2%	2.2%
Upregulated gene transcription units	6.1%	4.3%	3.4%	2.4%	1.7%
ERαBS (≥5 PET counts per site)					
UCSC known gene transcription units	33.9%	36.4%	37.7%	39.7%	25.8%
Downregulated gene transcription units	2.9%	3.9%	5.2%	7.2%	2.2%
Upregulated gene transcription units	6.0%	4.0%	3.8%	2.4%	1.7%

Note: ERαBS were partitioned into different categories and associated with UCSC Known gene transcription units, as well as up or downregulation information.

Chromatin interaction and transcription regulation

Subsequently, we examined the ERα-mediated chromatin interaction regions with respect to gene transcription. For added stringency, we focused on the 406 complex interactions and the 181 stand-alone duplex chromatin interactions that consist of 3 or more inter-ligation PETs (587 interaction regions). We envisage that multiple ERαBS may function as "anchor" regions generating looping structures in 3-dimensional space (Figures 24, 25 and 35A). We annotated the 587 interaction regions in relation to UCSC known gene database entries (Hsu et al. 2006). A gene was considered associated with a chromatin interaction region if a transcriptional unit has a TSS in or within 20 kb of the interaction boundaries. Most interaction regions (400/587=68%) were associated with genes (altogether, 3,957 UCSC known gene entries; Appendix). Many interaction regions include multiple genes, such as the keratin gene cluster (Figure 31B) and NR2F2 locus (Figure 31C). 1,490 entries (38% of 3,957 interaction-associated genes) have their TSS proximal (within 20 kb) to at least one anchor in

an interaction region. Genes with such transcriptional units are called “anchor genes”. The remaining 2,467 transcriptional units are far away from interaction anchors and reside in loops of the interactions; genes with such transcriptional units are therefore denoted “loop genes” (Fig. 35A).

We found that most interaction-associated genes have active promoter status (associated with H3K4me3 peaks) and are actively transcribing (associated RNAPII peaks) while non-associated genes are significantly less associated (Fisher’s Exact Test 2-tailed p-value = 0.0507 and $7.18e^{-18}$; Table 10). Moreover, in the interaction-associated genes, significantly higher percentage of anchor genes are actively transcribing (RNAPII marks) than the loop genes (Fisher’s Exact Test 2-tailed p-value = 0.00384 and 0.00001653 respectively; Table 10). We further analyzed the expression profiles of genes involved in chromatin interactions using microarray gene expression data over a time course of estrogen induction (Materials and Methods, Chapter 6), and validated selected examples using RT-qPCR (Fig. 35B-D).

Table 10. Association of ERα-mediated chromatin interactions with genes.

		Transcription marks		Expression status	Differentially expressed genes	
	Number of TUs^	**H3K4me3**	**RNAPII**	**Regulated**	**Up-regulated**	**Down-regulated**
All UCSC Known Genes	51123*	14116 (27.6%)¶	7101 (13.9%)	6082 (11.9%)§	2668 (43.9%)#	3414 (56.1%)
Interaction-associated genes	3957	1150 (29.1%)	754 (19.1%)	669 (16.9%)	375 (56.1%)	294 (43.9%)
Loop genes	2467	757 (30.7%)	418 (16.9%)	378 (15.3%)	187 (49.5%)	191 (50.5%)
Anchor genes	1490	393 (26.4%)	336 (22.6%)	291 (19.5%)	188 (64.6%)	103 (35.4%)
Non-enclosed anchor genes	394	99 (25.1%)	66 (16.8%)	62 (15.7%)	36 (58.1%)	26 (41.9%)
Enclosed anchor genes	1096	294 (26.8%)	270 (24.6%)	229 (20.9%)	152 (66.4%)	77 (33.6%)

Notes: ^All numbers are given in terof gene transcription units. “Transcription unit” is abbreviated as “TU”. *This number only shows the transcription units present on chromosomes 1-22 and chromosome X, and minus the genes involved in interactions. ¶ The percentages of genes with H3K4me3 and RNAPII marks are based on the number of TU in each category. § The percentages of genes differentially expressed are based on the number of TU in each category. # The percentages of up or down regulated genes are based on the number of differentially expressed genes in each category.

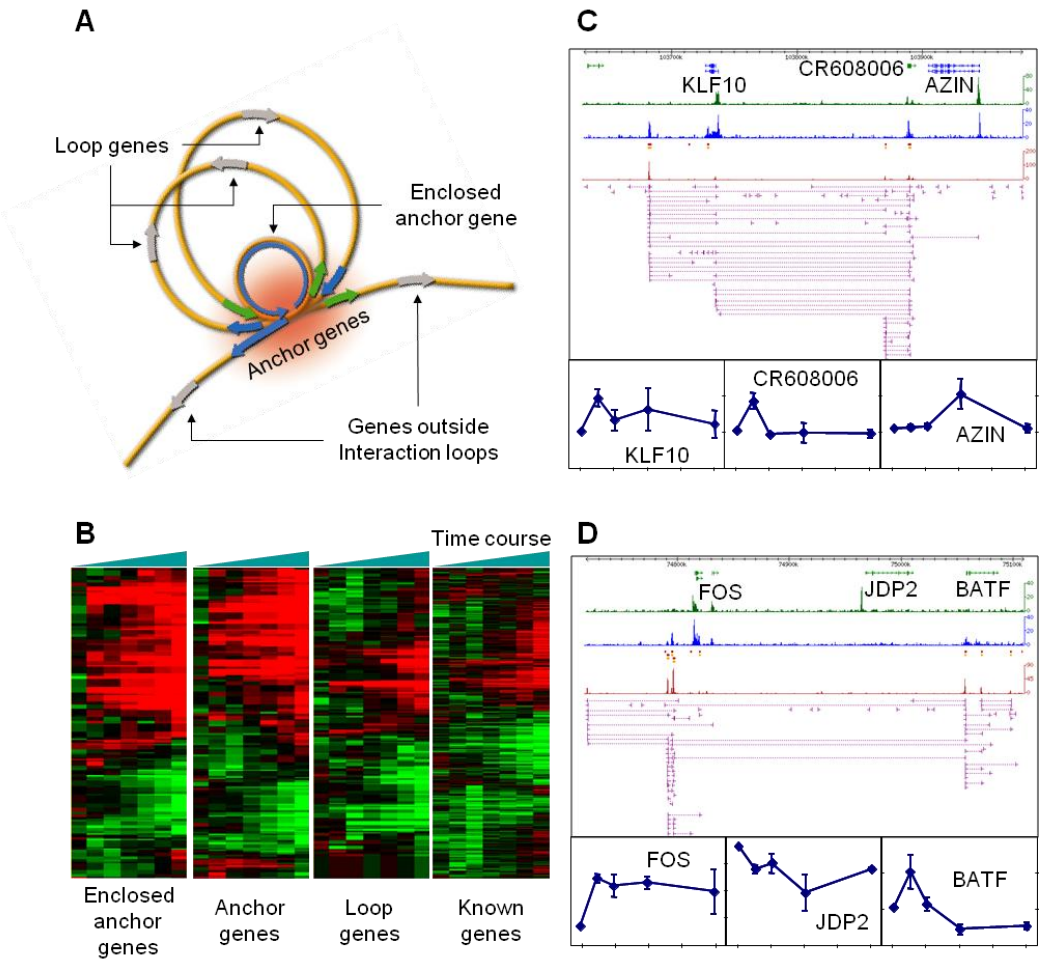


Figure 27. ERα-mediated chromatin interaction regions are associated with gene upregulation

A. Proposed model of multi-looping structure of chromatin interactions with multiple anchors by ERα binding. Anchor genes (green for top strand and blue for bottom strand) have promoters in close proximity to interaction anchoring center, where the transcription machinery are assumed in high concentration, and hence are active. Loop genes (gray) tucked inside the loop structure far away from the interaction anchoring center are less active. Similarly, genes outside the interaction structures (gray) may not be regulated through ERα binding. B. Gene expression microarray results over an estrogen induction time course (0, 3, 6, 9, 12, 24, 48h) of differentially expressed genes involved in chromatin interactions. Enclosed anchor genes, anchor genes and loop genes are presented. The UCSC Known Genes (Hsu et al. 2006) less the interaction-associated genes, "Known genes", were used as a general background control set for comparison. Red denotes estrogen-mediated activation and green denotes estrogen-mediated repression. C. and D. Examples of complex interactions involving multiple genes that are differentially transcribed as shown by H3K4me3 and RNAPII marks as well as RT-qPCR analysis. *(Note: Microarray data was prepared and analyzed by Kartiki Desai, Jane Thomsen, Yew Kok Lee, Haixia Li, and R. Krishna Murthy Karuturi).*

We found that interaction-associated genes are preferentially upregulated compared to non-associated genes (Fisher's Exact Test 2-tailed p-value = $2.73e^{-25}$; Table 10). Interaction-associated genes are weakly preferentially associated with downregulated genes compared to non-associated genes (Fisher's Exact Test 2-tailed p-value = 0.0701; Table 10). Moreover, anchor genes are preferentially upregulated compared to loop genes (Fisher's Exact Test 2-tailed p-value = $3.013e^{-17}$; Figure 35B, Table 10). By contrast, anchor genes are not significantly downregulated compared to loop genes (Fisher's Exact Test 2-tailed p-value = 0.349; Table 10).

Intriguingly, within the anchor gene category, we found that the majority (1,096 out of 1,490, 74%) of gene entries have 5' and 3' ends within the interaction boundaries. Such entries, called "enclosed anchor genes", frequently occupy the entirety of short interaction loops and are often found to engage multiple anchor sites within the gene structure as well. We observed that the "enclosed anchor genes" tend to have intense RNAPII marks covering the entire gene (examples in Figures 32A, B and 36-37), and are preferentially associated with RNAPII marks compared to non-enclosed anchor genes (Fisher's Exact Test 2-tailed p-value = 0.0012, Table 10).

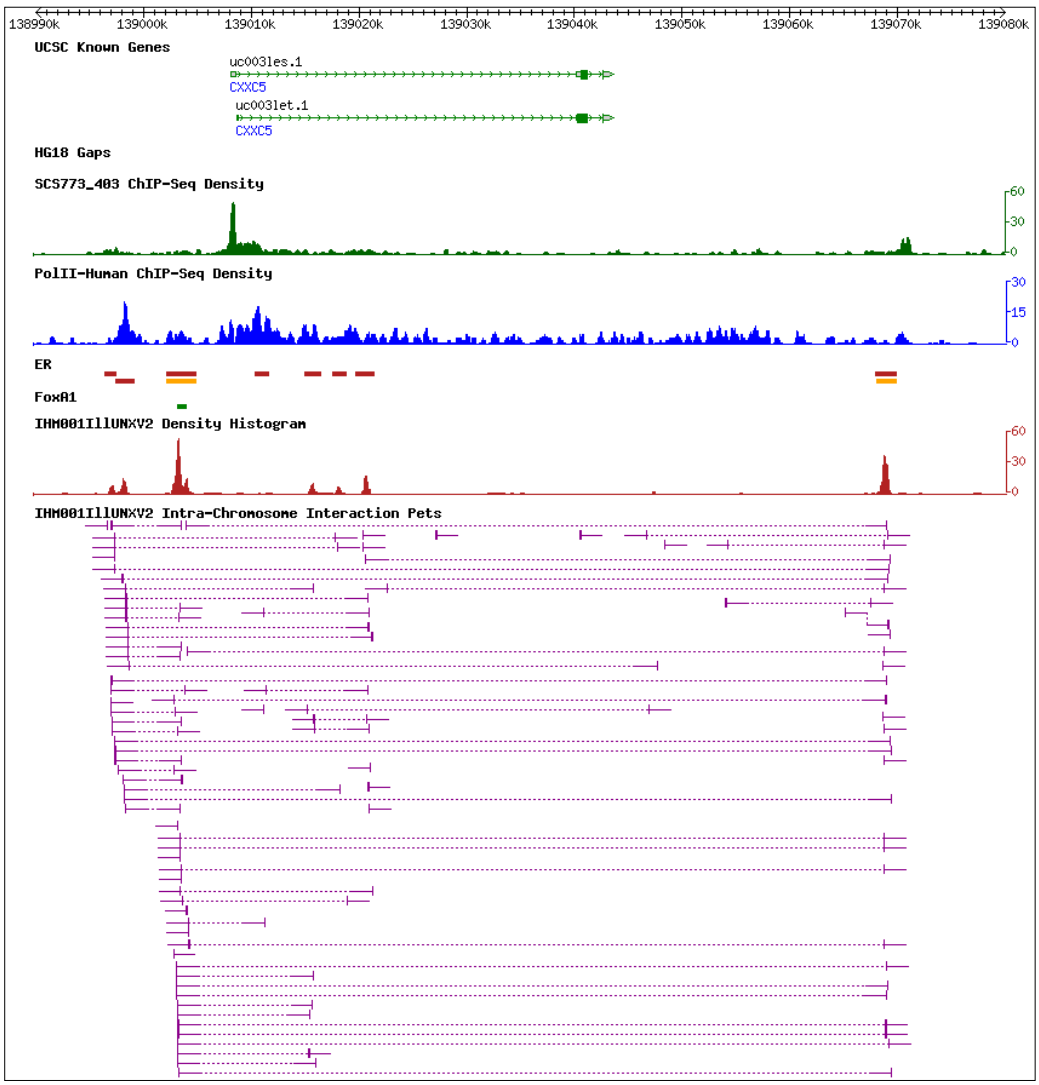


Figure 28. Example of an enclosed anchor gene on chr 5 (CXXC5).

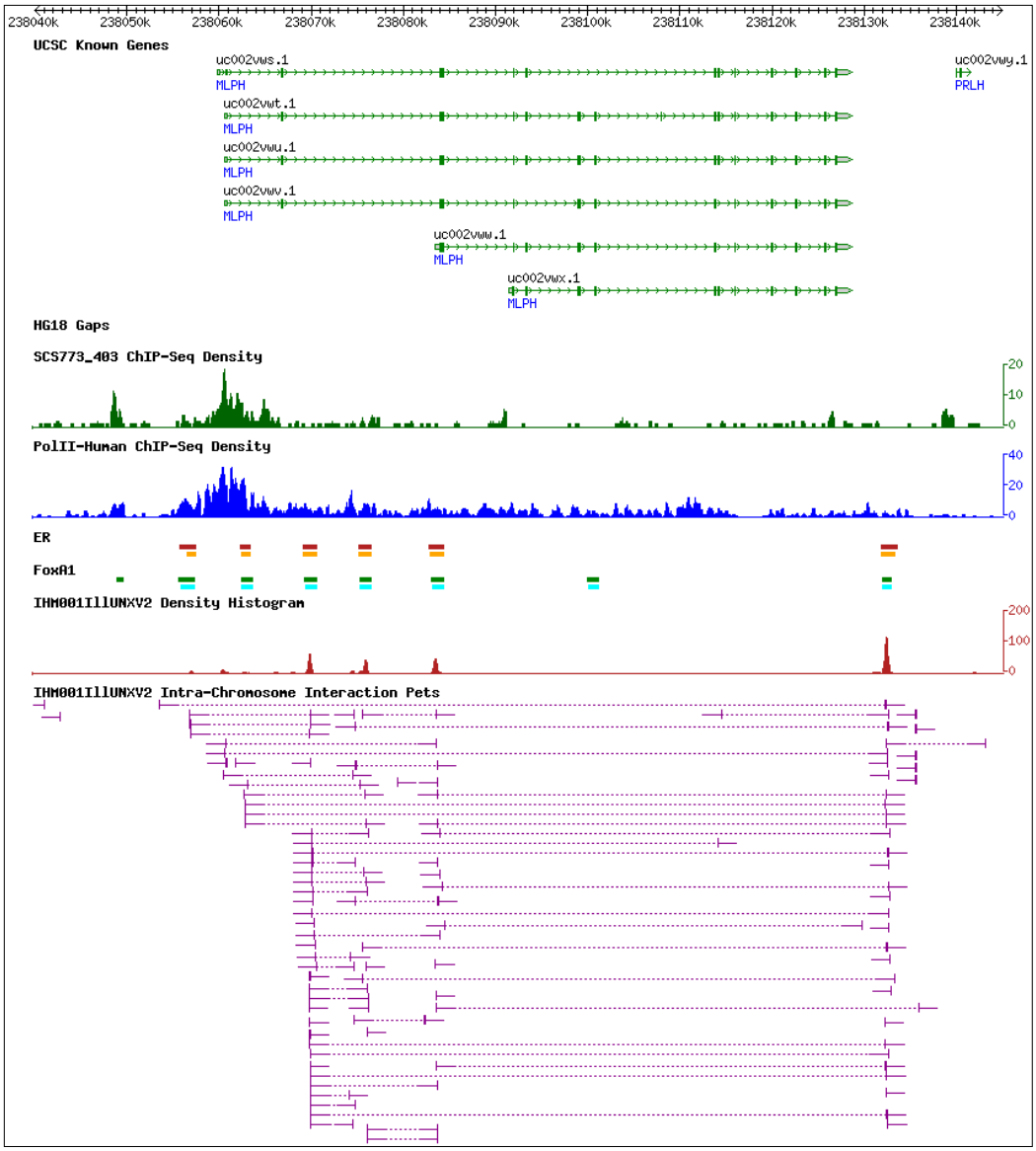


Figure 29. Example of an enclosed anchor gene on chr 2 (MLPH).

Moreover, enclosed anchor genes are preferentially upregulated compared to non-enclosed anchor genes (Figure 35B, Fisher's Exact Test 2-tailed p-value = 0.017; Table 10). Taken together, our data shows a gradient of functional association with ERα binding involved chromatin interactions and gene transcriptional activation: the enclosed anchor genes are closely correlated with upregulation as measured by gene expression microarray data and the RNAPII ChIP-Seq peaks, followed by non-enclosed anchor genes, loop genes,

and then genes not associated with interactions. Collectively, these results suggest that gene-centric interaction structures may provide an enclosed compartment for achieving higher local concentrations of ERα, transcription co-factors, and general transcriptional components at the target genes. We further speculate that transcriptional machinery could recirculate and cycle between transcription starting and ending sites of "enclosed anchor genes" tethered by ERα binding. This would represent a parsimonious strategy for transcriptional enhancement.

We also found evidence that ERα-mediated interactions may coordinate transcription regulation for genes involved in same functional pathways. One example is the complex interaction that encompasses 3 genes, FOS, JDP2, and BATF (Figure 35D) which encode the dimerization partners of JUN to form the AP-1 transcription factors. AP-1 is important in estrogen-mediated transcription, functioning either as a DNA tethering partner or as an ERα co-factor (Kushner et al. 2000). In this complex interaction, FOS and BATF are enclosed anchor genes, and are upregulated as shown by RNAPII marks and RT-qPCR; whereas JDP2 is a loop gene and is downregulated as shown by RT-qPCR and the lack of an RNAPII mark. We also noted that JDP2 has H3K4me3 marks, and that many loop genes are only marked by H3K4me3 (Table 10). It is conceivable that JDP2 and other loop genes could be "poised" and ready to be activated if it escapes from the interaction loop (Figure 35D).

Another very interesting example is the interaction region at the keratin gene cluster (Fig. 31B). Keratins play major structural roles in cells (Fuchs et al. 1994; Rogers et al. 2005; Steinert et al. 1988), and mutations give rise to various human hereditary keratin diseases, such as epidermolysis bullosa simplex (Moll et al. 2008). Keratins are also known to be involved in signaling and regulatory pathways (Moll et al. 2008). Keratins have very distinct expression patterns, and epithelial tumors frequently have the same patterns as the originating cells. This finding has led some genes, including KRT8, KRT18, and KRT7, to be used in immunohistochemistry analyses of cancers to identify tumor origins (Moll et al. 2008). Keratins are present in the human genome as two families: type I genes on chr17, and type II

genes on chr12 (Rogers et al. 2005). Keratins are unique in that type I genes and type II genes pair up by the formation of a heterodimer between one type I and one type II. Any keratin proteins that deviate from this rule are rapidly degraded (Lu et al. 1990). Therefore, gene expression in the keratin gene cluster has to be highly regulated in order to maintain distinct coexpression patterns. We hypothesize that chromatin interactions help in coordinating gene regulation and in maintaining coexpression patterns. We examined MCF-7 human breast adenocarcinoma cells, which are derived from ductal epithelial cells. Of the keratins used in immunohistochemistry diagnosis, breast adenocarcinomas typically express KRT8, KRT18, KRT19, KRT7, and occasionally KRT5, but not KRT20. Analysis of chromatin interactions in the keratin region suggests that chromatin interactions are correlated with gene expression coordination. Both ChIA-PET and 4C data shows that KRT7, KRT8, and KRT18 are all pulled into the "hub" of the same interaction complex. KRT7, 8, and 18 are known to be expressed in breast carcinomas. In particular, KRT8 and KRT18 are tightly coexpressed genes, and the gene products bind tightly to each other. These two genes are connected by many inter-ligations. By contrast, KRT5, 6, 1, 2, and the hair keratins are not expressed, and they are present in the "loop" of the interaction complex. Hence, chromatin interactions in the keratin region may bring together relevant genes into transcriptional foci, and loop out irrelevant genes, in order to achieve tightly coordinated gene expression regulation.

Taken together, our results suggest that long-range transcriptional regulation by ERα may be a fine-tuning mechanism that evolved to differentially regulate specific sets of related genes. To functionally determine whether such ERα-associated interaction regions are dependent on ERα, we used siRNA to knock down the level of ERα protein in MCF-7 cells (Materials and Methods, Chapter 6) and then measured if the interactions are disrupted and if gene transcription is affected. As shown in Figure 38, siRNA against ERα (siERα) efficiently reduced the amount of ERα protein compared to control siRNA, and effectively abolished the long-range chromatin interactions as demonstrated by a set of 3C assays at the GREB1 locus. Furthermore, siERα blocked GREB1 transcription as determined by measuring the mRNA

using RT-qPCR. This experiment was also previously conducted at the TFF1 site – resulting in a total of two sites examined in this manner, and similar results were obtained at the TFF1 site (Pan et al. 2008). In both cases, the chromatin loop and gene expression levels were greatly reduced, to close to zero. These data indicate that long-range chromatin interactions identified by ERα ChIA-PET data are dependent on ERα, and are required for the transcription regulation of estrogen target genes. Further work examining more sites would be desirable, and would help to substantiate the notion that ERα mediates chromatin interactions at most sites, as opposed to being a passive binder of chromatin interactions.

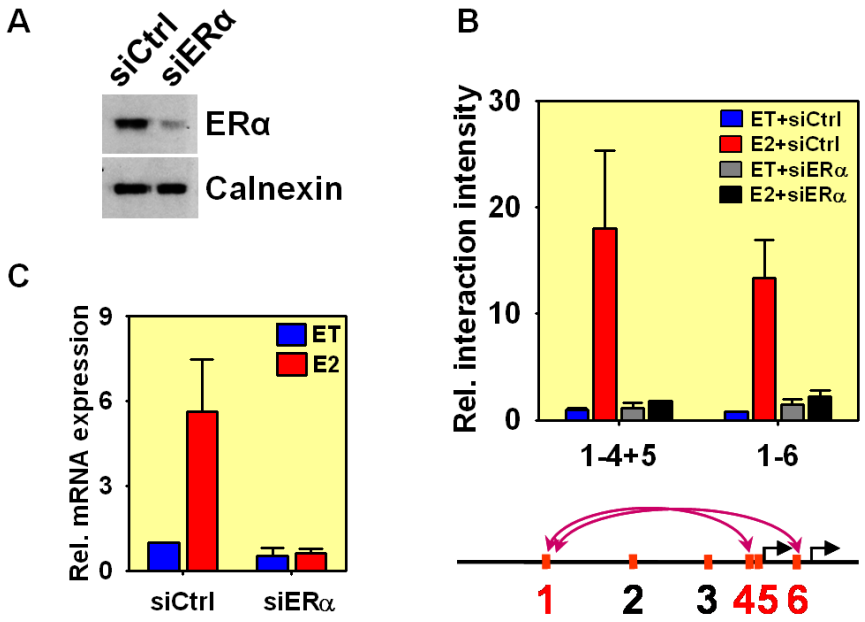


Figure 30. ERα-mediated chromatin interactions are required for transcription of estrogen-regulated genes.

MCF-7 cells were transfected with either control (siControl) or siRNA against ERα (siERα), respectively, and then analyzed by (A) western blot with antibodies directed against ERα and calnexin as a control, (B) 3C assays at the GREB1 locus, and (C) RT-qPCR to assess the mRNA levels of GREB1. (*Note: siRNA knockdown analysis was performed by the lab of Edwin Cheung).*

Discussion

Early genome-wide ChIP studies have found many more TFBSs than regulated genes and raised questions such as, why there are so many binding sites distal to gene promoters, are

these distal TFBSs functional, and if these TFBSs function at such distances, then which genes are regulated by these binding sites? Our results provided plausible answers to these questions. From this comprehensive map of a human chromatin interactome, we postulate a primary mechanism for ERα function in transcription regulation: ERα protein dimers are recruited to multiple ERαBS which may interact with one another to form looping structures around target genes; such topological architectures may partition individual genes in sub-compartments of nuclear space for differential transcriptional activation or repression.

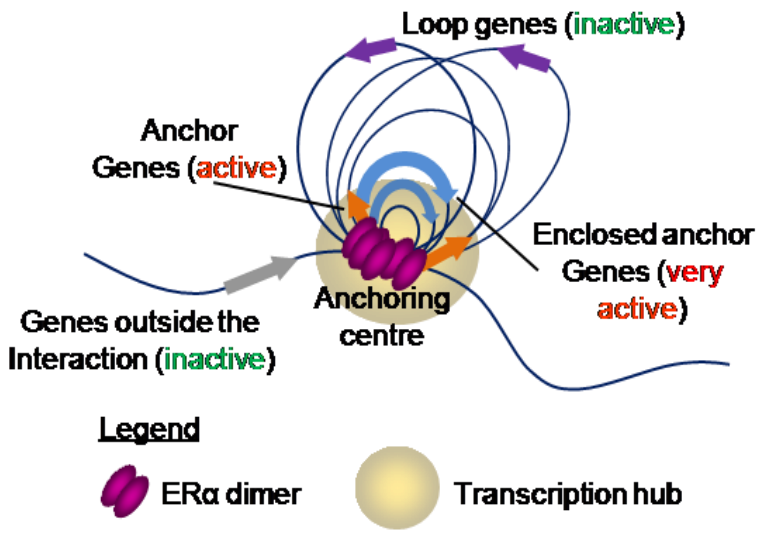


Figure 31. A model for ERα function via chromatin interactions.

In particular, anchor genes and enclosed anchor genes which are near the anchoring center especially in interactions with small loops, are packaged into a tight sub-compartment of chromatin looping structures, which could increase the local concentration of ERα, and transcriptional cofactors. Interactions may coordinate the regulation of different genes involved. Loop genes, especially those in large loops, may be separated from the transcriptional hub and thus be silenced.

An intriguing question is why a transcription factor such as ERα evolved to use such an extensive and intensive chromatin interaction mechanism for transcription regulation. When viewed in total, our data suggest that these chromatin interactions represent the most parsimonious use of binding sites constrained by an imposed linear distribution (order) in

several levels. First, the obvious redundancy in ERα binding and interactions is thought to enhance the robustness of ERα transcriptional control such that mutations at any one interaction site would not entirely eliminate regulatory control. Second, as a matter of topology, looping and anchor clustering provides greater degrees of advance for direct regulation of the transcriptional machinery than the proximity constraints of linear DNA. We further speculate that chromatin interaction centers involve many strands of chromatin coming together that could help achieve and maintain high local concentrations of transcriptional components. Loops that connect gene transcription start and end sites may allow for cycling of transcriptional machinery in a highly efficient manner. It is now known that ERα-DNA interactions at a defined ERαBS oscillate in an on-off state with periodicity, and oscillators use boundaries to change wave direction (Metivier et al. 2003). Given the extensive system of interaction complexes, ERα could oscillate between spatially proximate anchors of interaction regions, using the chromatin boundaries to provide oscillation dynamics to ERα behavior. Thus, the looping and anchor system we hypothesize represents a topological solution to a number of mechanistic observations of this transcription factor. Similar mechanisms may also be employed by other transcription factors in mammalian genomes.

We anticipate that this first-ever global chromatin interactome map and the ChIA-PET assay will constitute a valuable starting point for future studies into the 3-dimensional architecture of nuclear dynamics of transcription factor biology.

Chapter Five: Conclusions

Summary

In this thesis, I have demonstrated new methods for constructing PET libraries, and developed a new application of the PET method. This new application, Chromatin Interaction Analysis using Paired-End Tags (ChIA-PET), addresses a major issue in transcriptome biology: Are distal binding sites found in many whole genome transcription factor binding site ChIP experiments functional in gene regulation? If so, what is the mechanism of remote transcriptional control? Through the application of our tested and validated ChIA-PET to the system of ERα human breast cancer cells, we generated the first human chromatin interactome and showed that chromatin interactions are a primary mechanism by which ERα mediates transcriptional regulation. We proposed a new model for ERα functioning via chromatin interactions. In this model, we speculate that ERα protein dimers bind to distal regulatory elements and initiate long-range chromatin interactions involving promoter regions of target genes. These interactions form DNA loop structures with multiple ERα binding at the anchoring center. Multiple small and gene-centric loops could package genes near the anchoring center in a tight sub-compartment of chromatin looping structures, which could increase the local concentration of ERα, and therefore, attract and retain more molecules of cofactors as well as transcriptional machinery for enhanced transcriptional activation. This topological structure could also provide transcription efficiency, allowing RNAPII to cycle the tight circular gene templates. The large interaction loops, however, are more likely to link together distant genes at either end of the loop residing near anchor sites for coordinated regulation, and separate the genes in long loops from the active ERα regulation. This model may be used by other transcription factors in other systems.

The future of chromatin interactome biology

While we have developed the first global, high-throughput, *de novo* assay for chromatin interactions, and performing the ChIA-PET method is relatively straightforward, performing

multiple validations using FISH and 3C is laborious as site-specific BACs or PCR primers have to be chosen. Ideally, an alternative whole-genome chromatin interaction assay should be developed, such that both ChIA-PET and an alternative assay would complement each other to allow for validation of chromatin interactions found by each other. If the resolution of microscopy techniques could be refined while retaining the structure of chromatin interactions, microscopy would provide an ideal parallel approach. Atomic Force Microscopy and Electron Microscopy could be future candidates, given such further improvements.

While our findings have explained many questions in transcriptome genome biology, our new model of chromatin interactomes also raises many new questions. What factors are required for chromatin interaction formation and maintenance? Are there certain features of DNA that predict whether interactions will occur (Meaburn et al. 2007a)? Do other transcription factors employ similar mechanisms to regulate genes? Are transcription factor mediated chromatin interactions mainly intrachromosomal like ERα? How do chromatin structural proteins, such as histones, CTCF, and cohesin, contribute to the 3-dimensional structures of chromosomes? Do interaction locations tend to show translocations, as different DNA elements are brought together in close proximity (Meaburn et al. 2007a; Meaburn et al. 2007b)? Do chromatin interactions indeed coordinately regulate genes? In this regard, further work seeking evidence for cross-species conservation in the linear organization of the genes, particularly the coordinately-regulated genes, could help to support a conserved biological role. Going further into the question of conservation, are chromatin interactions at conserved genes themselves conserved in other organisms? How did chromatin interactions evolve? One possibility is that "junk DNA" could have separated out genes and enhancers. Chromatin interactions could then bridge the gap between these enhancers and their target genes. Furthermore, it would be interesting to explore the dynamics of chromatin interactions such as in response to cell cycle changes and different environmental factors which could perturb chromatin interaction profiles.

This first-ever global chromatin interactome map and the ChIA-PET assay constitute a valuable starting point for future studies into the unknown 3-dimensional space of the nucleus, investigating these questions. ChIA-PET can be readily applied to transcriptional cofactors such as RNA Polymerase II and p53, as well as chromosomal structural proteins such as cohesion. Together, these global chromatin interactome maps may be mined for deeper insights into chromatin interactome biology. Certain protein motifs may be found to be associated with "interaction status" of the binding site, allowing one to predict whether the binding site is likely to show interactions or not. These proteins may then be analyzed through knockdown studies to see whether knockdown abolishes interactions. Comparison of profiles of different factors can help to answer whether different factors show different patterns – for example, while ERα mainly employs local, intrachromosomal chromatin interactions, other factors could predominantly employ interchromosomal chromatin interactions. Certain proteins with chromosome structural roles might be expected to employ different mechanisms from ERα. It would also be interesting to see whether chromatin interactions could also "daisy-chain" towards gene promoters and then repress them, rather than activate them as was the case with ERα. ChIA-PET using the same factor in multiple different cell-lines can also help to identify whether chromatin interactome networks remain the same or different. Cell-specific chromatin interactions may be one method by which genomes are regulated to give rise to cell-specific changes. Moreover, ChIA-PET using different environmental conditions, such as using estrogen-treated and estrogen-untreated MCF-7 cells, might answer questions as to how environmental conditions are translated into genomic changes in cells (Meaburn et al. 2007a).

Chromatin interactome networks could possibly have important clinical implications. Studying drug-treated as opposed to drug-untreated cells, or virus-challenged and unchallenged cells, using ChIA-PET, may also reveal how such challenges affect the chromatin interactome network in cells, and demonstrate the mechanisms by which these exert their changes on the cells. Chromatin interactome changes could be important early

signals of cellular transformation, particularly as they might be involved in cancer-causing translocations (Meaburn et al. 2007b). FISH probes, or other markers, that reveal such changes could be used as early diagnostic markers (Meaburn et al. 2007a). Moreover, drugs which directly affect chromatin interactomes may even have clinical utility, as aberrant chromatin interactome networks may play critical roles in global dysregulation of genes.

Taken together, ChIA-PET, as the first method that can uncover chromatin interactions in a *de novo*, whole-genome manner, has helped to start a new field of chromatin interactome genomics, for understanding chromatin interactome networks. This new field could potentially prove to be important in the clinic.

The future of the PET technology

The unique feature of building connectivity between two points of DNA from linear and non-linear structures in PET analysis has tremendous value in many aspects of genomic analysis that cannot be simply and easily replaced by just improving sequencing capacity in near future. The PET concept is versatile allowing for ready adaptation to new sequencing technologies. In the future, PET technology will grow by incorporating new sequencing technologies, overcoming existing limitations, and finding new applications for answering biological questions.

One limitation arises from sequencing: while sequencing costs have dropped dramatically in recent years, it is still very high, and current next-generation sequencing methods have biases, and inaccuracies (Holt et al. 2008). Additional advancements in current and future next-generation sequencing machines promise to bring forth further improvements in costs, read lengths, through-puts, run times, preparation times, and accuracies (Metzker 2005). One example is Helicos sequencing, a very new sequencing technology for single molecule sequencing, which has the advantages of not requiring an additional clonal amplification step as well as allowing sequencing to operate in an asynchronous manner

which reduces the number of misincorporations (Harris et al. 2008); such a method is expected to result in lower costs and required sample amounts for sequencing. Once this platform becomes available for large scale data production, one such machine run would generate billions of PET sequences that could be enough paired sequences for *de novo* assembly of a complete individual human genome. Hence, with rapid development of next-generation sequencing machines, the PET techniques will also need to develop rapidly for easy, cost-effective and timely integration with the particular format that the sequencing machines use. As sequencing read lengths become longer, concatemers and length-controlled ligation methods such as the diPET method (Ng et al. 2006a) would become increasingly useful for making full use of the maximum read lengths of the machines.

Another limitation of the current PET technology is the library construction. The current protocols for making PET constructs using both cloning-based and cloning-free methods are still tedious, require large numbers of cells to start the experiment, and involve relatively short tags. Although optimizations of each step involved in PET construction could make incremental improvements, eventually, the PET method would have to be performed by robotic or miniaturized lab-on-a-chip systems in order to match the speed and efficiency of DNA sequencing machines. An important benefit of making PET constructs in a nanometer scale system is that this might allow PET analysis for smaller numbers of cells. Only with this nano-scale capability can PET analysis be applied to clinical samples that usually are not present in such large amounts. The use of microfluidics technologies to manipulate tiny amounts of fluids using tiny channels (Whitesides 2006) would be necessary for the development of such miniaturized assays. Emulsion technologies can also be used to create "microreactors" for partitioning reactions, by using water droplets dispersed in oil (Griffiths et al. 2006). 454 pyrosequencing relies on emulsions to separate amplicons when amplifying templates for sequencing (Margulies et al. 2005). Further developments promise to make contributions to library construction methods. Longer tags are more desirable because they give rise to increased mapping specificity, particularly when dealing with repeat regions.

However, PET preparation methods that use tagging enzymes are constrained by the restriction enzymes available. The current maximum tag length is 27 bp, from EcoP15I. Ideally, in the future, restriction enzymes that can cut longer tags would be found, and sequencing technologies would be able to accept longer templates.

Paired-End Tag sequencing is a fundamental concept, and can be implemented to any application that generates DNA fragments for analysis; for example, DNAse I hypersensitive sites could be mapped using PET technologies. Functional elements and transcription factor binding sites in the genome have been associated with open chromatin regions which can be easily digested with DNAse I, most likely due to nucleosome displacement during cell processes such as transcription activation. This feature has been used to obtain DNAseI-digested DNA, which is then sequenced in a high-throughput manner to identify these genetic elements (Sabo et al. 2004a; Sabo et al. 2004b). As an alternative, PETs could be obtained for identification of these genetic elements – the DNA could be sequenced in a bidirectional manner. Alternatively, FAIRE (Giresi et al. 2007) could be used to prepare DNA, which could then be processed by the PET method and sequenced. In other applications to look at nucleosome positioning, micrococcal nuclease (Schones et al. 2008) could be used. Micrococcal nuclease makes double-stranded cleavages between nucleosomes. The benefits of using PETs to analyze these genetic elements are that the exact 5' and 3' boundaries can be read out, to give precise positioning information. In addition, genetic elements that are associated with repeats may be more easily identified, because the additional information content in PETs leads to higher specificity, as well as the ability of PETs to cross over the boundaries of repeats.

With the capability to perform fast, cheap, and robust PET analyses on a wide variety of applications, we expect that PET-based methods will become the method of choice for many sequencing projects. Particularly, PET technologies have great potential to make big contributions to the field of personal genomics. In the near future, DNA-PET could be

combined with ultra-high-throughput sequencing technologies to give rise to a robust, cost-effective platform for individual personal human genome sequencing. In addition, the wide variety of PET applications for genome structure, transcriptome, and interactome characterizations will be useful in annotating the human genomes in great detail for functional and clinical implementations. With these new capacities, personal genome sequences combined with patient-specific transcriptomes and interactomes could become a practical reality, and greatly benefit human healthcare and society.

In conclusion, the PET technology is a versatile method which can couple methods for asking biological questions with next-generation sequencing. With sequencing improving rapidly and increasing demands for sequencing to interrogate biological and clinical questions, the future of PET technologies is very bright.

Chapter Six: Materials and Methods

Note: Except for a test run performed by Illumina, USA, all sequencing described here was performed by the Sequencing Team of Genome Technology and Biology led by Wei Chia-Lin, Genome Institute of Singapore, Singapore. The members of the Sequencing team are: Herve Thoreau (lab manager), Melvyn Tan, Yow Jit Sin, Dawn Choi, Low Hwee Meng, Eleanor Wong (now in the Research Team), Ong Chin Thing (Jo), Neo Say Chuan, Yap Zhei Hwee, Poh Tong Shing, Leong See Ting, Adeline Chew, Jeremiah Decosta (now in the Research Team), Alexis Khng Jiaying, and Lim Kian Chew.

Materials and Methods used in Chapter 2

Cell culture

HES3 Human Embryonic Stem (ES) Cells were grown and prepared as described (Zhao et al. 2007). ***Note: Cells were obtained from ES Cell International, Singapore, and grown and prepared by Andrew Choo from the lab of Steve Ho, Bioprocessing Technology Institute, Singapore.***

Full length cDNA library construction

A full length cDNA library was constructed from the human embryonic stem cells and PETs were prepared for sequencing as described in the classic bacterial propagation protocol (Ng et al. 2006b). Briefly, RNA was isolated from HES3 cells, and poly A+ RNA was isolated from RNA using the μMACS mRNA isolation kit. The poly A+ RNA was converted into cDNA by oligo-dT-primed reverse transcription. RNA ends were biotinylated. Cap-trapper selection was performed to select full-length first strand cDNA. 5' adapters were added to prime for second strand cDNA synthesis, and the material was then digested to give rise to sticky ends for cloning. The flcDNA was then ligated with pGIS4b vector cut with NotI (NEB) and GsuI (Fermentas). The flcDNA library was amplified by bacterial amplification at 37°C on solid surface agar Q-trays followed by scraping and plasmid extraction by Maxiprep

(Qiagen). ***Note: FlcDNA library construction was performed by Yao Fei, Genome Institute of Singapore, Singapore.***

GIS-PET library construction

An aliquot of the Maxiprep was used to prepare a GIS-PET library by the classic bacterial propagation GIS-PET protocol (Ng et al. 2006b). Briefly, MmeI digestion was performed, and the single-PET plasmids were end-polished with T4 polymerase (Promega). The single-PET plasmids were then self-ligated and amplified by bacterial amplification at 37°C on solid surface agar Q-trays followed by scrapping and plasmid extraction by Maxiprep (Qiagen). Single PETs were released with BseRI, purified, and concatenated. The concatemers were then blunted by T4 DNA polymerase (Promega), cloned into EcoRV-cut pZErO-1 vectors (Invitrogen), and 300 384-well plates were sequenced with Sanger capillary sequencing. This library was called SHE001. The library was analyzed, and the results were reported separately (Zhao et al. 2007). ***Note: This library was created by Liu Jun, Genome Institute of Singapore.***

Selection-MDA GIS-PET library construction

To construct the MDA-amplified library using the new Selection-MDA protocol, we took an aliquot of 8 ng of maxiprep from the GIS-PET full-length cDNA library and added it to 50 µl of Templiphi 500 sample buffer (GE Healthcare). The sample was denatured at 95°C for 3 min, and then cooled to 4°C. 2 µl of Templiphi 500 enzyme mix (GE Healthcare) was added to 50 µl Templiphi sample buffer on ice, and the mixture was then added to the 50 µl sample buffer with denatured template. The reaction was incubated at 30°C for 18h, and then heat inactivated at 65°C for 10 minutes. The material was quantitated with Picogreen Fluorimetry (Invitrogen), and an MmeI (New England Biolabs) digestion was performed following the Single PET construction method as described (Ng et al. 2006b). 800 ng of self-ligation reaction was purified to remove salts before electroporation by phenol/chloroform isopropanol precipitation as described (Ng et al. 2006b). The pellet was resuspended in 5 µl

of Elution Buffer (Qiagen). The entire ligation mix was transformed into 50 µl of Top10 *E. coli* electrocompetent cells (Invitrogen) and recovered in 1 ml of Lucigen Recovery Medium (Lucigen) with shaking at 37°C for 4 hours. Because recovery was for only 4 hours, the bacteria would not have multiplied sufficiently so as to compete with each other; hence the library should contain no size bias. To monitor bacterial growth, the optical density at 600 nm (OD_{600}) of aliquots were taken at various time points by Nanodrop. Cells were spun down at 10,000 g for 5 min and washed twice with 750 µl of Lucigen Recovery Medium to remove free floating DNA that was not introduced into the cells. Next, plasmids were extracted by performing Miniprep (Qiagen). 40 µl of Elution Buffer was used for the elution, and the DNA was quantitated with Picogreen fluorimetry. 1 µl was run on a PAGE gel to check that plasmids were prepared correctly. Plasmid-Safe DNAse (Epicenter) treatment was then performed to remove any linear species, such as bacterial genomic DNA, that might be present. Phenol/chloroform ethanol precipitation was then performed and pellets were resuspended in 20 µl of Elution Buffer (Qiagen). MDA was performed on aliquots of 8 ng of material as described above. The material was quantitated with Picogreen Fluorimetry, and digested with BamHI (New England Biolabs) according to the manufacturer's protocols. The PETs were PAGE gel purified, then cloned, concatenated, partially digested with BamHI, cloned into BamHI-cut pZErO-1 vectors (Invitrogen), and prepared for sequencing as described (Ng et al. 2006b). 10 plates of 384 colonies consisting of concatenated PETs were sequenced as a GIS-PET library, SHE002. A more detailed protocol is in the Appendix. ***Note: Jack Tan, Genome Institute of Singapore, Singapore conceived of and performed the experiments on Selection-MDA.***

Data analysis

Data analysis was performed using PET-Tool for PET extraction and genome mapping (Chiu et al. 2006), followed by visualization in the T2G browser, a specially designed visualization system for Paired-End Tags mapped to genome assemblies (Ng et al. 2005). Calculations were performed with Microsoft Excel. Categories of the genes were identified using RefSeq

(Pruitt et al. 2007), UCSC Known Genes (Hsu et al. 2006), Genbank mRNA (http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide), MGC (Gerhard et al. 2004), Ensembl (Hubbard et al. 2007), ESTs (Boguski et al. 1993), Twinscan (Korf et al. 2001), SGPGene (Guigo et al. 2003; Parra et al. 2003), and Genescan (Burge et al. 1997) databases.

Materials and Methods for Chapter 3

Note: The Materials and Methods for Chapter 3 and 4 have a number of overlaps; where this occurs, Chapter 4 refers to Chapter 3, and the description in Chapter 3 includes slight modifications used in Chapter 4.

Cell culture and estrogen treatment

MCF-7 cells were grown to at least 80% confluence in DMEM/F12 (Invitrogen/Gibco) supplemented with 5% FBS (Invitrogen/Gibco), penicillin (Invitrogen), streptomycin (Invitrogen), and gentamycin (Invitrogen). In preparation for the 17 beta-estradiol ("estrogen"; Sigma) treatment, cells were grown in hormone-free media: they were washed with PBS and incubated in phenol red-free medium (Invitrogen/Gibco) supplemented with 5% charcoal-dextran stripped FBS (Hyclone), penicillin, streptomycin, gentamycin, and L-glutamine (Invitrogen) for a minimum of 72 hours. Hormone-depleted cells were treated with estrogen (17 beta-estradiol, E2, (Sigma) to a final concentration of 100 nM for 45 min before the ChIP procedure. The control cells were treated with an equal volume and concentration of vehicle, ethanol (Merck), for 45 min. For a ChIA-PET experiment, we routinely use approximately 1 x 10^8 cells from 6 150-mm diameter cell culture plates. ***Note: Starter cultures and some batches of MCF-7 cells kindly provided by the lab of Edwin Cheung, Genome Institute of Singapore, Singapore, or the lab of Edison Liu, Genome Institute of Singapore, Singapore.***

Chromatin immunoprecipitation (ChIP)

ChIP protocol was performed as described previously (Lin et al. 2007). Briefly, we used 1% formaldehyde to crosslink the cells, and sonication to break the chromatin fibers. ERα specific antibody (HC-20, Santa Cruz) was used to enrich ERα bound chromatin fragments. IgG specific antibody (sc-2027, Santa Cruz) were also used for ChIP analyses. ChIP material bound on the antibody beads was subjected to ChIA-PET library construction. ***Note: Some ERα and IgG ChIP preparations were performed by the lab of Edwin Cheung, in particular Pan You Fu, Genome Institute of Singapore, Singapore.***

ChIA-PET library construction and sequencing

The DNA fragments tethered in chromatin fragments were end-repaired using T4 DNA polymerase (NEB), followed by overnight ligation of biotinylated half-linkers that contain a flanking MmeI site (IDT), using T4 DNA ligase (Fermentas) at 16ºC, with mixing. The linker added DNA fragments were then phosphorylated with T4 polynucleotide kinase (NEB), and followed by a second ligation reaction overnight at 22ºC under dilute conditions. The conditions for ligation were based on previous PET protocols for self-circularization of plasmids in a complex library (Ng et al., 2005; Wei et al., 2006). The cross-links in the DNA/protein complexes were then reversed by incubation at 65ºC overnight with 0.2% SDS (Ambion) and proteinase K (Ambion), and the DNA fragments were purified by phenol/chloroform isopropanol precipitation. Any nicks present were subsequently repaired by incubation with *E. coli* DNA ligase (NEB) and *E. coli* DNA polymerase I (NEB) at 16ºC overnight. The purified DNA was then digested by MmeI (NEB) for at least 2h at 37ºC to release the tag-linker-tag structure (Paired-End Tag, PET). The biotinylated PETs were then immobilized on streptavidin-conjugated magnetic Dynabeads (Invitrogen) and the ends of each PET structure were then ligated to an adapter by T4 DNA ligase (Fermentas) at 22ºC overnight with mixing followed by 20 cycles of PCR reaction to amplify the PETs. This PCR product was the template for sequencing analysis using Roche 454 pyrosequencer (GS20) following the manufacturer's protocol. For two ERα ChIA-PET libraries, we conducted 5

GS20 runs and generated a total of 1.8 million raw PET sequences for further analysis. More details are available in the Appendix. In addition, as a genome-wide control, we prepared an IgG ChIA-PET library, conducted 1 GS20 run, and generated a total of 0.52 million raw PET sequences. ***Note: ERα ChIA-PET construction was performed together with Liu Jun, Genome Institute of Singapore, Singapore. The IgG ChIA-PET library was prepared by Andrea Ho and Ruan Xiaoan, Genome Institute of Singapore, Singapore.***

ChIA-PET barcoding

We constructed the two ERα ChIA-PET libraries with two biological replicates of MCF-7 cell cultures treated with estradiol (E2) using two linker sequences with different nucleotide barcodes. As a linker sequence can include unique nucleotide barcode, multiple linkers with distinctive barcode sequences can be used to specify different experiments or replicates. Advantages of PET barcoding are that different biological samples or replicates may be analyzed within the same experiment, leading to time and cost savings, as well as reductions in technical variations of measurement. The barcoding was performed as follows: The two biological replicates were kept separate throughout the ChIP procedure and the first ligation. In the first ligation, half-linker 1 was introduced to replicate 1 in a microfuge tube, and half-linker 2 was introduced to replicate 2 in a separate microfuge tube. After the first ligation, the samples were washed well to remove any unligated half-linkers, and combined. The second ligation was then performed. More details are available in the Appendix. Another benefit of barcoding in this manner was that the number of chimeric ligations in the second ligation could be estimated. Any PETs with combinations of half-linker 1 and half-linker 2 into full linker sequences would have to result from chimeric, random ligations. The two different half-linker sequences are reported in the Appendix. We generated 941,151 and 867,751 unique PET sequences from the two libraries. We found very few chimeras, only 40,165 unique PET sequences (2.17% of the total unique PET sequences).

RNAPII ChIP-Seq

Illumina single-read sequencing was used to analyze serine-5 phosphorylated RNAPII (ab5131, Abcam) ChIP material. Examples of this data are shown in the GREB1 locus as well as in Table 2. ***RNAPII ChIP-Seq library construction was performed by Pan You Fu and Liu Jun, Genome Institute of Singapore, Singapore.***

Cloning-free ChIP-PET library construction and sequencing

As a genome-wide control, we compared the ERα ChIA-PET data with the ERα ChIP-PET data generated by a cloning-free method similar to the ChIA-PET method. As the ChIP-PET library did not use proximity ligation to capture the relationship of DNA fragments tethered by chromatin complex, we do not expect to see many inter-ligation PETs. If we see any, these inter-ligation PETs should be from non-specific ligations. A key difference between the methods involved the second ligation. In detail, in the ChIP-PET procedure, after the first ligation (ligation of the half-linkers) and the chromatin phosphorylation with T4 polynucleotide kinase, cross-links were reversed by incubation at 65ºC overnight with 0.2% SDS (Ambion) and proteinase K (Ambion). DNA was purified by phenol/chloroform isopropanol precipitation. Subsequently, overnight dilute ligation was performed on the ChIP DNA with T4 DNA ligase (Fermentas) at 22ºC without agitation. Nick repair was then performed by incubation with *E. coli* DNA ligase (NEB) and *E. coli* DNA polymerase I (NEB) at 16ºC overnight, followed by DNA purification which included a Plasmid-Safe Enzyme (Epicenter) step for removing uncircularized products. MmeI digestion and subsequent steps were performed as per the ChIA-PET protocol. For the ERα ChIP-PET library, we conducted 4 GSFLX runs and generated a total of 2.82 million raw sequences for further analysis. In addition, we reprocessed SHC007, the ERα ChIP-PET library described previously (Lin et al. 2007), in order to convert it from hg17 to hg18 genomic assembly. 635K raw sequences were generated. ***Note: The ERα control ChIP-PET library was prepared by Ruan Xiaoan's team, from the Genome Institute of Singapore, Singapore. A full list of people involved in the preparation and analysis of SHC007 can be found in the***

journal reference. Reprocessing was performed by Han Xu, Genome Institute of Singapore, Singapore.

Library saturation analysis

We carried out a saturation analysis on each library to assess the sequencing depth reached and to estimate the upper bound unique sequencing attainable. The saturation is modeled using the Hill Function:

$$f(x) = \frac{ax^b}{x^b + c^b}$$

with x as the number of PETs sequenced and $f(x)$ as the number of distinct PET sequences obtained. Using the Marquardt-Levenberg nonlinear least-square fitting algorithm, we fitted the Hill Function to each library, with the order of sequencing randomly permuted. Based on the redundancy of the sequenced PETs, we found that the ChIA-PET library 1 and 2 were about 16.2% and 17.4% saturated. The combination of these two libraries was about 16.7% saturated. ***Note: Library saturation analyses were performed by Vinsensius Vega, Genome Institute of Singapore, Singapore.***

DNA-PET 10 Kb insert data

It is known that the MCF-7 genome involves lots of rearrangements (Volik et al. 2006). Therefore, ChIA-PET data generated from this genome for detecting long-range interactions could be complicated by genome structural differences between this and the reference genome (hg18). To avoid such complications, we constructed DNA-PET libraries with insert sizes around 10 Kb in span. We generated 35 million DNA-PET sequences, which is a 100-fold physical coverage of the MCF-7 genome. This dataset provides comprehensive karyotyping information regarding deletions, inversions, translocations, and insertions in the MCF-7 genome, and identifies rearranged genomic regions. We used this information to filter out inter-ligation PET clusters located in these genome aberration regions, and therefore

reduce false positive calls. ***MCF-7 DNA-PET libraries were prepared by Yao Fei. Data analysis was performed by Wing-Kin Ken Sung's lab.***

PET extraction and mapping

The raw sequence reads generated by the Roche 454 pyrosequencer were processed through the 'PET-Tool' program (Chiu et al. 2006) for extraction of PET sequences and mapping of the PETs using compressed suffix arrays (Hon et al. 2007) to the reference human genome sequence (hg18). The PET sequence was extracted based on the basic unit of tag/linker/tag with defined parameters such as the linker sequence and the tag length. As PET barcoding was used in the ChIA-PET procedure, we identified PETs belonging to either replicate 1, 2 or "chimeras" by examining the linker sequences in each PET, and assigning each PET containing a particular linker to that particular category. Up to 1 mismatch was allowed in the linker sequences of the PETs. The mismatch could be an insertion, deletion or substitution. The average length of the tag is 20bp with +/- one nucleotide variation due to a known characteristic of plasticity by MmeI enzyme (Dunn et al. 2002). The tag sequences were then aligned to the human reference genome sequence (hg18), and the two tags of the same PET were paired for their mapping coordinates. Each tag had a tag length, n. For every tag, first we attempted to map all n bases of the tag. If the mapping found a hit or several hits, mapping stopped on the particular tag. If not, then we tried n-1, n-2, and so on until n = 18. If this failed, the tag was transferred to the "unmapped" category. This set of PETs represented the "uniquely mapped PETs". The "uniquely mapped PETs" were further merged if any PETs shared the same mapping locations (as up to 1 mismatch was tolerated, any two PETs might be unique before mapping, but after mapping they might be found to match the exact same locations). This set of merged data was called "uniquely mapped and merged PETs". ***Note: PET mapping was performed by Pramila Ariyaratne, Hong-Sain Ooi, Yusoff bin Mohamed, and Chiu Kuo Ping, Genome Institute of Singapore, Singapore.***

PET classification

Based on mapping characteristics, each PET sequence can be classified by whether it was derived from one DNA fragment or two DNA fragments. If the two tags of a PET were mapped on the same chromosome with the genomic span in the range of ChIP DNA fragments (less than 3 kb), with expected self-ligation orientation and on the same strand, we considered that this PET was most likely derived from a self-ligation of a single ChIP DNA fragment (Lin et al. 2007), and therefore called the PET a "self-ligation PET". We chose to use 3 kb as the cutoff for "self-ligation PETs" because it is the upper range of ChIP DNA fragments in this experiment. If a PET did not fit into these criteria, we considered that the PET most likely resulted from a ligation product between two DNA fragments, therefore we called the PETs "inter-ligation PETs". The two tags of the "inter-ligation PETs" do not have fixed tag orientations, might not be found on the same strands, might have any genomic span, and might not map to the same chromosome. In addition, specifically for the "inter-ligation PETs", if the two tags of a PET mapped in same chromosome but with a span > 3 kb in any orientation or if the two tags mapped with spans of less than 3 kb but not with expected orientation or to the same strands, these PETs were called "intrachromosomal inter-ligation PETs". PETs which mapped to different chromosomes were called "interchromosomal inter-ligation PETs". Further analysis was performed on these PETs to determine whether they were the result of specific ChIP, ligation and mapping, or the result of non-specific processes.

Note: PET classification was performed by Hong-Sain Ooi, Han Xu, and Yusoff bin Mohamed, Genome Insitute of Singapore, Singapore.

Identification of ERα binding sites

Binding site peaks were found based on self-ligation PETs (Lin et al. 2007). Self-ligation PETs were converted into a density histogram representing enrichment density, and local maxima represent peaks that indicate ChIP enriched binding sites. Binding sites were found using a threshold of FDR $\leq$ 0.01. We identified whether peaks correspond to any satellites using the "RepeatMasker" (Smit et al. 1996-2004) track in the UCSC Genome Browser

(Karolchik et al. 2003). We removed binding sites found in satellite regions, as manual curation showed these to be the result of nonspecific ChIP pulldown. Accuracy of the automatic analyses was also double-checked using manual curation. From 1 replicate of the ChIA-PET experiment, we found 2,179 binding sites; from another, we found 2,720. From the ChIP-PET experiment, we found 1,211. As expected, from the IgG control ChIA-PET experiment, 0 binding sites could be found. All binding sites are listed in the Appendix. ***Note: Identification of binding sites was performed together with Han Xu, Genome Institute of Singapore, Singapore.***

Identification of ChIP enrichment levels

As ChIP enrichment of a given DNA-binding protein target can be reflected by overlapping virtual ChIP DNA fragments represented by ChIP-PETs (Lin et al. 2007), or ChIP-Seq (Johnson et al. 2007) fragments, and regions with higher numbers of ChIP-PETs are more likely to be true binding sites (Lin et al. 2007), similarly, multiple virtual ChIP DNA fragments represented by "self-ligation PETs" and "inter-ligation PETs" derived from a particular region will indicate the ChIP enrichment of that region. ***Note: ChIP enrichment level identification was conceived of by Ruan Yijun, Genome Institute of Singapore, Singapore.***

ERE motif analysis of ERα binding sites

We analyzed the presence of the Estrogen Response Element (ERE) motif in the ERα binding sites identified in this study, according to the method in our previous publication (Lin et al. 2007). Briefly, we looked for the presence of the full consensus ERE motif (GGTCA-nnn-TGACC), allowing for a maximum of 2 mismatches (Lin et al. 2007). The distribution of ERE motifs relative to the binding sites was plotted in Figure 15C. ERE motifs are enriched at the center of the ChIA-PET identified ERα binding sites. We also investigated the distribution of ERE motifs in other datasets, and found them to be similar. ***Note: ERE***

analysis was performed by Han Xu and Vinsensius Vega, Genome Institute of Singapore, Singapore.

Comparative analysis of ERα binding sites

To understand whether the binding sites identified by ChIA-PET were valid, first we compared the two library replicates with each other. We extended ± 200 bp from the mid-point of each binding site to do the comparison. 1459 binding sites overlapped between the two ChIA-PET replicates, out of 2,720 (54%) and 2,179 (67%) binding sites found in each replicate. While the overlap is good, a reason why the overlap is not even higher could be that the libraries are not yet saturated. For the binding sites that did overlap, the correlation of intensity was very high, with a Pearson correlation of 0.90, indicating strong correlation (Figure 2A). Next, we compared the two ChIP-PET experiments, one of which was previously published (Lin et al. 2007) and cloning-based (called SHC007, "old"), and one of which was based on a cloning-free procedure very similar to the ChIA-PET method ("new"). For consistency, we re-mapped the previous cloning-based ChIP-PET to hg18 and processed the library with the same pipeline to identify high confidence (FDR $\leq$ 0.01) binding sites. Because the FDR was more stringent than previously used, therefore fewer binding sites were found: 501 as compared with the previous 1,234. We extended ± 200 bp from the mid-point of each binding site of the two datasets for the comparison. 231 binding sites were found to overlap, out of 501 (46%) in the old dataset and 1,211 (19%) in the new dataset. Again, a reason why the overlap is not higher could be that the libraries are not yet saturated. Next, we compared the combined ChIA-PET datasets with the combined ChIP-PET datasets and the 3,665 binding sites (p-value < E-05) found by ChIP-chip experiment (Carroll et al. 2006; Lupien et al. 2008). We extended ± 200 bp from the mid-point of each binding site of the ChIA-PET and ChIP-PET datasets, and used the entire binding site region reported in the ChIP-chip region, to look for overlaps. Accuracy of the automatic analyses was double-checked using manual curation. Multiple numbers are given in the overlaps, as sometimes, 2

or more peaks in one dataset might overlap to a single peak in a different dataset. We observed that most of the ChIP-PET binding sites that did not overlap had low peak values. Using independent ChIP-qPCR, we experimentally validated a subset of 9 highest ChIA-PET sites that did not have any overlaps with ChIP-chip and cloning-based ChIP-PET datasets. 3 of the 9 sites overlapped with the cloning-free ChIP-PET dataset, and many were in repeat regions, supporting the idea that sites that did not overlap could be sites difficult to identify with ChIP-chip due to repeats, or were unclonable such that the cloning-based ChIP-PET would not have been able to pick them up. ***Note: Library comparison was performed together with Han Xu, Genome Institute of Singapore, Singapore.***

ChIA-PET data visualization

We adopted the "Generic genome browser" system (Stein et al. 2002) and developed the "ChIA-PET Genome Browser" to organize and visualize the ChIA-PET data. The "self-ligation PETs" and the "inter-ligation PETs" are displayed in separate tracks to show transcription factor binding sites and interactions, respectively. This browser also includes a custom 'Whole Genome Interaction Viewer' which provides a macroscopic picture of binding sites and interactions along with a whole genome landscape (http://cms1.gis.a-star.edu.sg). The username is "guest" and the password is "gisimsgtb". A manual is provided in the Appendix. ***Note: ChIA-PET visualization was performed by Hong-Sain Ooi, Pramila Ariyaratne, and Yusoff bin Mohamed, Genome Institute of Singapore, Singapore.***

Using inter-ligation PETs to identify ERα-mediated interactions

As each inter-ligation PET was derived from two ChIP DNA fragments, the majority of which were less than 1,500 bp in size, we extended the mapped 20 bp tags to 1,500 bp along the reference genome to represent the virtual DNA. Multiple overlapping virtual DNAs are merged into DNA regions. To determine the DNA regions that were bound together in close spatial proximity by an ERα-mediated protein complex, we made the following assumptions.

First, if an interaction between two DNA regions is specific, it would be enriched by the ChIP procedure, and hence the inter-ligation PETs that "link" these regions would be over-represented in the ChIA-PET data; while if it is non-specific and occurs randomly, it would be sampled much less frequently than real interactions and at the level expected by chance. We modeled the non-specific interactions such that each DNA fragment has an equal chance to interact with and be ligated to any other fragments.

Consider a library with N inter-ligation PETs, the total number of sampled DNA fragments is $2N$. We denote R_A and R_B as representing two DNA regions with c_A and c_B virtual DNAs, where $c_A, c_B << N$. Under the random model, the number of inter-ligation PETs that link R_A and R_B, denoted $I_{A,B}$, approximately follow a hypergeometric distribution:

$$\Pr(I_{A,B} \mid N, c_A, c_B) = \frac{\binom{c_A}{I_{A,B}}\binom{2N - c_A}{c_B - I_{A,B}}}{\binom{2N}{c_B}}$$

By this, we are able to compute a p-value to test if $I_{A,B}$ is over-represented. Note that the p-values were Bonferroni-corrected for multiple hypothesis test. We found 228 and 205 interactions with multiple inter-ligation PETs from each replicate of the ChIA-PET libraries. Interactions with satellite repeats were filtered out, as these tend to be non-specifically enriched by ChIP. In addition, interactions with genomic distances of < 5 kb were filtered out automatically (chapter 3) or subjected to manual curation (in chapter 4), because we reasoned that these interactions could result from multiple unusually long ChIP-PET fragments. As this filtering is based on the genomic span, it is not expected to carry any bias leading to interactions close to gene promoters being dropped at a greater frequency than interactions that are far from gene promoters. Using the Bonferroni correction, we looked for high confidence interactions with Bonferroni-corrected p-value < 0.05. However, because analyses of the medium confidence interactions suggested that they could also contain many *bona fide*

interactions, in Chapter 4, we used both high and medium confidence interactions. All interactions are in the Appendix. ***Note: Interactions were identified by Han Xu and Vinsensius Vega (chapter 4), Genome Institute of Singapore, Singapore.***

Manual curation

To understand the characteristics of interactions, subsets of interactions were manually curated by visualizing them on the ChIA-PET browser (described in a separate section) in order to (1) examine the binding sites to see whether they were present and if they formed well-shaped peaks, (2) check if the interactions found by automatic methods indeed had inter-ligation PETs between them, and (3) check whether the interactions could be found in both libraries, and (4) check whether the interactions involved amplicon regions and rearrangements. ***Note: Manual curation was performed together with Phillips Huang and Brenda Yuyuan Han, Genome Institute of Singapore, Singapore.***

Assignment of genes to high confidence interactions

We assigned UCSC Known Gene transcription units (Hsu et al. 2006) to high confidence interactions. Although the UCSC Known Gene browser has some "redundancies" in the sense that the same gene has multiple different transcription units, we used the database on an "as-is" basis because different transcription units might have different characteristics (some might have RNAPII marks but not others, and some might be within the interactions but not others), so we wanted to capture all these features in an unbiased manner. We assigned genes if they were present within the genomic span of the interactions or if they were within 20 kb of the loci of the interactions. In addition, RNAPII ChIP-sequencing data from estrogen-treated MCF-7 cells was used provide information regarding transcription status for genes involved in interactions. A transcript was said to be marked by RNAPII if the promoter (a region of $\pm$ 1 kb from the transcription start site) contained an RNAPII peak. This information is given in the Appendix.

Chromosome Conformation Capture (3C)

3C was performed as described previously (Hagege et al. 2007) with modifications. Briefly, MCF-7 cells were treated as mentioned in the ChIP protocol up to the crosslinking step with 1% formaldehyde. Nuclei were resuspended in 500 µl of 1.2 x restriction enzyme buffer at 37°C for 1 hr, 7.5 µl 20% SDS for 1 hr, followed by 50 µl 20% Triton X-100 for additional 1 hr. Samples were then incubated with 400 units of selected restriction enzyme at 37°C overnight. After digestion, 40 µl 20% SDS was added to the digested nuclei and incubated at 65°C for 10 min. 6.125 ml of 1.15x ligation buffer and 375 µl 20% Triton X-100 was added and incubated at 37°C for 1 hr prior to the addition of 2000 units of T4 DNA ligase (NEB) at 16°C for 4 hr. Samples were then de-crosslinked at 65°C overnight followed by phenol-chloroform extraction and ethanol precipitation. Primers and restriction enzymes for the 3C procedure were chosen based on the ChIA-PET interactions. All primers had to be within a region of ±150 bp from the restriction enzyme digestion site. Primers (1stBase) were designed using Primer3 software available from: http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi (Rozen et al. 2000). PCR products were amplified with AccuPrime Taq High Fidelity DNA Polymerase (Invitrogen) for 40 cycles. PCR products were run on a 2% agarose gel. Each validation experiment was repeated at least twice. ***Note: 3C was performed by Mei Hui Liu, Genome Institute of Singapore, Singapore.***

Chromatin Immunoprecipitation Chromosome Conformation Capture (ChIP-3C)

ChIP-3C was performed as described previously (Hagege et al. 2007) with modifications. Briefly, chromatin immunoprecipitation was performed overnight as described in the ChIP protocol. Beads were then washed twice with PBS, and restriction enzyme digestion was performed overnight in 100 µl of 1x buffer at 37ºC with nutation (all from NEB). The beads were then spun down, and the buffer removed. A further restriction digest was performed with fresh buffer and enzyme at 37ºC for half a day. The beads were then spun down, and the buffer removed. The beads were then washed 3x with PBS, and ligation was performed using

1x ligation buffer and T4 DNA ligase (NEB) in 100 μl at 16ºC. A further ligation was performed by adding 100 μl of fresh buffer and enzyme to the mixture and incubating at 16ºC for half a day. 100 μl of Elution Buffer containing 1% SDS was then added to the beads, and the beads were incubated at 65ºC for at least 6 hours. The supernatant was purified with a PCR purification kit (Qiagen). Primers and restriction enzymes for the ChIP-3C procedure were chosen based on the ChIA-PET sequences. All primers and restriction enzymes had to be within a region of ±100-500 bp from the targeted ERα binding site peak. Primers (1stBase) were designed using Primer3 software available from: http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi (Rozen et al. 2000). PCR products were amplified with AccuPrime Taq High Fidelity DNA Polymerase (Invitrogen) using an MJ thermocycler (GMI). The PCR program used was (1) 94°C for 2 min, (2) 94°C for 30s, (3) 56-60°C for 40s, (4) 68°C for 40s (5) 68°C 5 min, (6) 4°C forever. Steps (2) to (4) were run for 35-47 cycles. PCR products were run on a 1% agarose gel with ethidium bromide. PCR products were sequenced to verify the long-range ligation product. Each validation experiment was repeated at least twice for confirmation. ***Note: ChIP-3C was performed by the lab of Edwin Cheung, in particular by Pan You-Fu, Genome Institute of Singapore, Singapore.***

RT-qPCR

Total RNA was prepared from MCF-7 cells induced with estrogen for 0, 3, 6, 12 and 24 hours using an RNA purification kit (Qiagen), following the manufacturer's protocols. 1 μg of total RNA was incubated with 50 ng of random primer (Roche) at 70ºC for 10 min and then cooled on ice for 1 min. To the mixture, first strand buffer (Clontech) was added to a final concentration of 1x, DTT (Clontech) was added to 0.01 M, dNTP mix (Invitrogen) was added to 1 mM, and 1 μl of Powerscript RT enzyme (Clontech) was added. The mixture was heated to 42ºC for 90 min, and heat inactivated at 70ºC for 15 min. Real-time quantitative PCR was performed using an ABI Real-time PCR 7500 system. PCR was performed with a 10 μl reaction volume consisting of substrate, 0.5 μM of primer pairs (1stBase) and 1x SYBR

Green PCR Master Mix (ABI). Reactions were incubated at 95ºC for 10 min, and then 40 cycles (95ºC for 15s, 60ºC for 1 min) were carried out. Fluorescence was acquired at the end of each cycle at 60°C during the amplification step. The control pair of primers used was that of 36B4 (ribosomal protein mRNA). All experiments were repeated at least twice. ***Note: RT-qPCR was performed by the lab of Edwin Cheung, Genome Institute of Singapore, Singapore.***

ChIP-qPCR

ChIP-qPCR experiments were performed against ERα, unphosphorylated RNAPII (8WG16, Covance), and serine-5 phosphorylated RNAPII (ab5131, Abcam). ChIP material was prepared from MCF-7 cells induced with estrogen for 45 min ("estrogen-treated"), as well as negative control MCF-7 cells induced with an equal volume of ethanol for 45 min ("ethanol-treated"), as described earlier. ChIP material was reverse cross-linked under conditions of 1% SDS and 65°C, and purified using a PCR purification kit (Qiagen). Real-time PCR quantification was performed as described earlier. The control primer used was from Zhao et al., 2007 (Zhao et al. 2007). All experiments were repeated at least twice. ***Note: ChIP-PCR was performed by Pan You Fu and Shi-Chi Leow, Genome Institute of Singapore, Singapore.***

Materials and Methods for Chapter 4

Note: There is some overlap between the Materials and Methods for Chapter 3 and 4. Where they are the same, reference is made to Chapter 3. The descriptions in Chapter 3 include slight modifications used for Chapter 4.

ChIA-PET library construction and sequencing

As described in "Materials and Methods for Chapter 3". 454 pyrosequencing and Illumina paired end sequencing analysis was performed. In total we generated 7.4 million raw PET

sequences that passed Illumina's filtering for quality base calling. We also generated 1.8 million raw PET sequences from 454 pyrosequencing analysis. We combined these two libraries and removed redundant PETs, which resulted in 5.9 million total PET sequences for further analysis.

H3K4me3 ChIP-Seq data

H3K4me3 antibody (ab8580, Abcam) was used to generate ChIP-enriched DNA fragments for Illumina single read sequencing analysis. The H3K4me3 ChIP-Seq data was mapped to hg18 genome, and enrichment peaks for H3K4me3 binding were identified using ChIP-Seq peak calling algorithm as previously described (Chen et al. 2008b). 37,542 H3K4me3 binding sites were identified in the MCF-7 genome from this dataset. This dataset characterize the promoter status of genes in MCF-7 cells during estrogen induction, which were used to annotate the genes involved in ERα-mediated chromatin interactions. ***Note: H3K4me3 ChIP-Seq was prepared by Roy Joseph, Genome Institute of Singapore, Singapore.***

RNAPII ChIP-Seq data

As described in "Materials and Methods for Chapter 3".

DNA-PET 10 Kb insert data

As described in "Materials and Methods for Chapter 3".

Microarray gene expression data to identify estrogen-regulated genes

A comprehensive dataset of time-course microarray experiments was performed to investigate the effects of estrogen treatment on gene expression profiles and identify estrogen responsive genes. Estrogen treated (10 nM) and DMSO-mock MCF-7 cells (negative control) for 0, 3, 6, 9, 12, 24, and 48 hours were collected for RNA extraction and the labeled probes were hybridized to microarrays (HG-U133 Plus). 3 replicates were performed for each time point. The data was analyzed using two different time-course differential expression analysis methods: Pooled Variance Meta Analysis (2) and LIMMA (3) and ranked by their scores. The

top 5,000 probes or ~10% of all probes were obtained from each ranking and combined resulting in ~7,500 probes. The set was further filtered using mean inclusive Data-driven Smoothness Enhanced Variance Ratio Test (dSEVRAT) with dSEVRAT score > 200 resulting in ~3700 probes. Up and down regulation for each gene was decided based on their trend using hierarchical clustering carried out using Eisen software(Eisen et al. 1998) (http://rana.lbl.gov/EisenSoftware.htm). ***Note: Microarray data was prepared and analyzed by Kartiki Desai, Jane Thomsen, Yew Kok Lee, Haixia Li, and R. Krishna Murthy Karuturi, Genome Institute of Singapore, Singapore.***

PET sequence analysis

The pipeline for processing PET sequences (PET extraction and mapping) is described in Chiu et al., 2006 (Chiu et al. 2006). PET classification was performed as described in "Materials and Methods for Chapter 3". Identification of binding sites was performed as described in "Materials and Methods for Chapter 3". We have 9,015 binding sites with a false discovery rate (FDR) $\leq$ 0.01. Identification of ERα-mediated chromatin interactions using inter-ligation PETs was also performed in a similar manner as described in "Materials and Methods for Chapter 3".

Interaction complexes

Many of the putative chromatin interactions (duplex interactions involving two anchors) connect to each other by overlapping anchors (anchors can be thought of the base of the loop). Based on such connectivity, multiple individual interactions were collapsed together into interaction regions. We evaluated each genomic locus using manual curation to double-check the automatic procedure, and also determine if that particular region has structural rearrangements, based on the DNA-PET library data that characterized the genomic aberrations in MCF-7 cells. We also identified the inter-ligation PET clusters located in amplicon regions where complicated rearrangements often happen, and filtered them out. We then required that the resulting interaction regions must have 3 or more inter-ligation PETs,

for further specificity. The resulting 406 complex interactions and 181 duplex interactions are listed in the Appendix. ***Note: Interaction complex analysis was performed together with Han Xu. Manual curation was performed together with the labs of Ruan Yijun, Wei Chialin, and Ruan Xiaoan.***

ERαBS association with relevant genomic features

First, ERα binding sites were grouped into categories based on their involvement with interaction characteristics. A. binding sites involved in complex interactions ("strong-interactions"), B. in single interactions ("intermediate-interactions"), C. with singleton inter-ligation PETs that may likely represent weak interactions or random background noise ("weak-interactions"), and D. with no inter-ligation PETs ("no-interactions"). These data are listed in the Appendix.

Next, we performed association of ERαBS with ChIP-chip data of ERα binding. We associated ERαBS involved in the 4 categories with the 12,193 ChIP-chip defined ERα binding sites (Lupien et al. 2008). For analyses of the 9,015 ERαBS identified by ChIA-PET, a region of $\pm$100bp from the middle of the ChIA-PET-identified binding sites was used, and overlapped with a region of $\pm$ 100 bp from the middle of the ChIP-chip-identified binding sites. For a random noise reference, we used regions of $\pm$ 100bp from the middle of a random sampling of 9,015 ChIA-PET singleton mapped loci. 0.87% of the PET singletons overlapped with the ERα ChIP-chip defined loci. We used the Fisher's exact test to see if there is significant enrichment in the association levels with ERα ChIP-chip data between high-enrichment ERαBS (≥20 PET counts per site) with high- and intermediate- interactions as compared with high-enrichment ERαBS (≥20 PET counts per site) with weak-interactions. The 2-tailed p-value is 0.08487, which is weakly significant.

In association of ERαBS with ChIP-chip data of FoxA1 binding, we associated binding sites and interaction loci with FoxA1 binding sites generated using ChIP-chip (Lupien et al. 2008). There are 23,745 FoxA1 binding sites in the human genome. For

analyses of the 9,015 binding sites found by ChIA-PET, a region of ± 250bp from the middle of the ChIA-PET-identified binding sites was used, and overlapped with a region of ± 250 bp from the FoxA1 binding sites (FoxA1 binding sites are reported as 1 bp in size). We chose ± 250bp from the middle because FoxA1 is a different protein that might bind to ERα directly, or other proteins that bind to ERα in a complex; hence the FoxA1 peak might be fairly far away from the ERα peak. For a random noise reference, we used regions of ± 250bp from the middle of a random sampling of 9,015 ChIA-PET singleton loci. We used the Fisher's exact test to see if there is significant enrichment in the FoxA1 content between high-enrichment ERαBS (≥20 PET counts per site) with high- and intermediate- interactions as compared with high-enrichment ERαBS (≥20 PET counts per site) with weak-interactions. The 2-tailed p-value is $5.4e^{-15}$, which is very significant.

In association of ERαBS with H3K4me3 ChIP-Seq mapping sites, we associated binding sites and interaction loci with H3K4me3 binding sites generated using Illumina sequencing. There are 37,542 H3K4me3 binding sites in the human genome. For analyses of the 9,015 binding sites found by ChIA-PET, a region of ± 250bp from the middle of the ChIA-PET-identified binding sites was used, and overlapped with a region of ± 250 bp from the H3K4me3 binding site peaks (H3K4me3 peaks are 1 bp in size). We chose ± 250bp from the middle because H3K4me3 is a histone mark that might be fairly far away from the ERα protein binding site peak. For a random noise reference, we used regions of ± 250bp from the middle of a random sampling of 9,015 ChIA-PET singleton loci. We used the Fisher's exact test to see if there is significant enrichment in the H3K4me3 content between high-enrichment ERαBS (≥20 PET counts per site) with high- and intermediate- interactions as compared with high-enrichment ERαBS (≥20 PET counts per site) with weak-interactions. The 2-tailed p-value is 0.021032, which is significant.

In association of ERαBS with RNAPII ChIP-Seq mapping sites, we associated ERα-mediated interaction loci with RNAPII binding sites generated using Illumina sequencing.

We generated 13,132 RNAPII peaks in the human genome. For analyses of the 9,015 binding sites found by ChIA-PET, a region of ± 250bp from the middle of the ChIA-PET-identified binding sites was used, and overlapped with a region of ± 250 bp from the RNAPII binding site peaks (RNAPII peaks are 1 bp in size). We chose ± 250bp from the middle because RNAPII is a different protein that might bind to ERα directly, or other proteins that bind to ERα in a complex; hence the RNAPII peak might be fairly far away from the ERα peak. For a random noise reference, we used regions of ± 250bp from the middle of a random sampling of 9,015 ChIA-PET singleton loci. We used the Fisher's exact test to see if there is significant enrichment in the RNAPII content between high-enrichment ERαBS (≥20 PET counts per site) with high- and intermediate- interactions as compared with high-enrichment ERαBS (≥20 PET counts per site) with weak-interactions. The 2-tailed p-value is $2.5e^{-17}$, which is significant.

In association of ERαBS with TSS of known genes UCSC Known Genes(Hsu et al. 2006) (hg18) were annotated to binding sites and interaction loci if the 5' transcription start sites (TSS) was located within ± 20 kb from the binding site. For a random noise reference, we also annotated TSS located within ± 20 kb from the middle of a random sampling of 9015 ChIA-PET singleton loci. We used the Fisher's exact test to see if there is significant enrichment in the RNAPII content between high-enrichment ERαBS (≥20 PET counts per site) with high- and intermediate- interactions as compared with high-enrichment ERαBS (≥20 PET counts per site) with weak-interactions. The 2-tailed p-value is 0.17652, which is not significant.

TRANSFAC analysis

We were interested in finding potential co-factors of ERα from our data and assessing whether they were significantly involved in the chromatin interaction detected using the ChIA-PET assay. We used the presence of binding motif (as defined using TRANSFAC

weight matrices and criteria) as a proxy to the transcription factor binding. First, we looked for motifs that were enriched in the datasets of binding sites. This analysis was performed in a similar manner as previously described (Lin et al. 2007), except hg18 was used instead of hg17. In addition to finding the ERα motif, we also found many motifs that had previously been found to be associated with ERα binding, such as FoxA1 (Lin et al. 2007). Next, we looked for motifs that were enriched in ERαBS with high- and intermediate- interactions as compared to ERαBS with weak-interactions. To do this, we began with motifs that were enriched in binding sites within interaction regions, and filtered out non-vertebrate motifs, motifs without FDR < 0.05, and motifs with fewer than 50 sequences with at least 1 hit (called "hits"). On the remaining motifs, we employed Fisher's Exact Test to determine which motifs were significantly enriched in the dataset of binding sites with interactions as opposed to those that do not. We also performed Bonferroni correction for multiple hypothesis testing. The TRANSFAC analysis data is in the Appendix. ***Note: TRANSFAC analysis was performed together with Vinsensius Vega, Genome Institute of Singapore, Singapore.***

Association of ERα-mediated chromatin interactions with genes

Genes (UCSC Known Genes, hg18) (Hsu et al. 2006) were assigned to complex and standalone duplex interactions (collectively called "interaction regions"). Some genes have multiple alternative transcripts and thus are reflected in the genome as different gene models (transcription units), which are each given a different unique gene ID. These different transcription units may share the same gene name, but can have different features, for example, some transcripts might have RNAPII marks but not others. An example of such a gene with different gene models is GREB1. In addition, some transcription start sites from a particular gene might be near the interactions but not other transcription start sites belonging to the same gene. In order to fully capture all features of all transcript units, and obtain the most accurate mapping of interactions to genes, we used all gene IDs as given in the UCSC

Known Genes database. In the text, these different gene models which each have unique gene IDs as given by the UCSC Known Gene database, are called "transcription units".

If the 5' transcription start site of a transcription unit falls anywhere within the interaction boundaries of the interaction complex plus 20 kb (20 kb upstream of the middle of the 5'-most anchor to 20kb downstream of the middle of the 3'-most interaction anchor), then we assigned the associated gene as an **"interaction-associated gene"**. If the TSS of a transcription unit was within ± 20 kb of the middle of any anchor in an interaction unit, the associated gene was assigned as an **"anchor gene"** otherwise it was assigned as a **"loop gene"**. If the transcription unit was not just within ± 20 kb of the middle of any anchor in an interaction unit but also had the entire transcription unit (5' transcription start site to 3' transcription end site for that particular transcription unit) entirely wrapped up within interaction boundaries of the interaction unit, then the associated gene was further called an **"enclosed anchor gene"**. Otherwise, if the gene was an anchor gene but not classified as an "enclosed anchor gene" because none of the associated transcription units were entirely wrapped up within the interaction boundaries of the interaction unit, it was called a **"non-enclosed anchor gene"**. The gene was marked as upregulated or downregulated based on whether it showed such microarray expression probes. The gene was marked as H3K4me3 associated if the promoter (1 kb upstream and downstream of the gene transcription start site) had such a peak. Similarly, it was marked as RNAPII associated if the promoter (1 kb upstream and downstream of the gene transcription start site) had such a peak. 27 genes have multiple transcription units wherein one transcription unit is defined as "anchor" and one transcription unit is defined as "loop". This means that one transcription unit for one gene had a TSS within 20 kb of an anchor, whereas another transcription unit for the same gene had a TSS that was not within 20 kb of an anchor. All genes are listed in the Appendix.

Gene expression visualization and analysis

Gene transcription units in different categories were clustered using *Cluster* version 2.11 (http://rana.lbl.gov/eisen/?page_id=42) and visualized using *TreeView* version 1.60 (November 2002) (http://rana.lbl.gov/eisen/?page_id=42) (Eisen et al. 1998). If two or more probes could be assigned to the same transcription unit, one probe was chosen randomly.

Circular Chromosome Conformation Capture (4C)

We developed a new sonication-based method for performing Circular Chromosome Conformation Capture (4C) (Zhao et al. 2006). Briefly, MCF-7 cells were treated as mentioned in the ChIP protocol up to the crosslinking step with 1% formaldehyde. An additional centrifugation step was performed to further clarify the supernatant by removing cellular debris. Aliquots were removed and diluted 10 times with Tris-HCl buffer (Qiagen, Buffer EB) containing 1x Protease Inhibitor Cocktail (Roche). The chromatin was incubated for 1h at 37°C. 1 % (final concentration) Triton X-100 was added and the chromatin material was allowed to stand for a further hour at 37 °C. End-blunting was performed at room temperature for 45 min, using the End-It DNA End-Repair Kit (Epicentre). The chromatin samples were diluted to 10 ml with sterile water containing 1 x Complete Protease Inhibitor Cocktail, and we performed ligation by adding 1000 units of T4 DNA ligase (Fermentas) and letting the reaction stand at 16°C overnight. 0.15 µg/µl (final concentration) of Proteinase K (Invitrogen) was added, and the chromatin material was reverse cross-linked at 65 °C overnight. The DNA was purified by phenol extraction and isopropanol precipitation, and treated with RNase A (Qiagen) at 37°C for 30 min. Non-circularized DNA was digested away by incubation with Plasmid-safe DNase (Epicentre) at 37°C overnight, and the DNA was re-purified by phenol extraction and isopropanol precipitation. The DNA samples were amplified using nested inverse PCR. Primers (1st Base) had to be within 100 bp of the targeted ERα binding site peak and were designed using Primer3 software available from: http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi(Rozen et al. 2000). The RepeatMasker track (Smit et al. 1996-2004) in the UCSC Genome Browser

(http://genome.ucsc.edu/) (Karolchik et al. 2003) was used to ensure that the primers did not lie in repeat regions. An MJ thermocyler (GMI) and the high-fidelity DNA polymerase Phusion (Finnzymes) were used for the PCR reactions. The PCR program used for first-round amplification was: (1) 98°C for 30 s; (2) 25 cycles of 98°C for 10 s, 70°C for 30 s and 72°C for 30 s; (3) 72°C for 10 min; and (4) 4 °C forever. The PCR program used for second-round amplification was: (1) 98 °C for 30 s; (2) 25 cycles of 98 °C for 10 s and 72 °C for 1 min; (3) 72 °C for 10 min; and (4) 4 °C forever. The resulting amplification product was run in a 6 % PAGE gel, and the fraction of the smear band above about 500 bp in size was excised. The DNA samples were sequenced using a 454 GSFLX long reads kit. ***Note: 4C analysis was performed together with Phillips Huang, Brenda Han and Charlie Lee, Genome Institute of Singapore.***

Fluorescence in-situ hybridization (FISH)

For FISH studies, we chose one of the longest intrachromosomal interaction complexes, chr15:93128663-94685818, which is about 1.5 Mb in genomic span. This interaction involves many genes, including NR2F2, AK000872, AK307134, AK057337, and BC040875. For convenience, we refer to this interaction as the "NR2F2 interaction". BAC probes P1, P2, and P3 were chosen from the list of available BACs (http://www.ncbi.nlm.nih.gov/projects/mapview/). P1 and P2 span a region of about 756K, and do not involve interactions. This is the "negative control" region. P2 and P3 span a region of about 966K, and involve interactions. This is the "experimental" region. MCF-7 nuclei were harvested by treating cells with 0.75 M KCl for 20 min at 37°C. The cells were fixed in Methanol/Acetic acid (3/1), and nuclei were dropped on slides for FISH. Following overnight culture in LB media, DNA's BAC were extracted with Nucleobond PC500 (Macherey-Nagel), and then labeled by nick translation in the presence of biotin-16-dUTP or digoxigenin-11-dUTP using Nick translation system (Invitrogen). In presence of 1µg/µl of Cot1DNA (Invitrogen), DNAs BAC clones were resuspended at a concentration of 5ng/µl in

hybridization buffer (2SSC, 10% dextran sulfate, 1X PBS, 50% formamide). Prior to hybridization, MCF-7 nuclei slides were treated with proteinase K (Sigma) at 37°C for 2 min followed by 2 1X PBS rinses (5 min at room temperature) and dehydratation through ethanol series (70%, 80% and 100%). Denaturated probes were applied to these pretreated slides and codenaturated at 75°C for 5min and hybridized at 37°C overnight. Two posthybridization washes were performed at 45°C in 2SSC/50% formamide for 7 min each followed by 2 washes in 2SSC at 45°C for 7 min each. After blocking, the slides were revealed with avidin-conjugated fluorescein isothiocyanate (FITC) (Vector Laboratories, CA) for biotinylated probes and anti-digoxigenin- Rhodamine for digoxigenin-labeled probes (Roche). After washing, slides were mounted with vectashield (Vector Laboratories, CA) and observed under an epifluorescence microscope (Nikon). Between 100-200 interphase nuclei were analyzed for each mix of probes. Fusion and colocalization spots were counted in each nuclei. Fisher's Exact Test was used to evaluate whether the number of fusions were significantly higher when comparing the various types of cells. Comparing control probes (P1/P2) with experimental probes (P2/P3) in ethanol-treated (ET) cells, there is a very significant (Fisher's Exact Test 2-tailed p-value = $2.39277e^{-14}$) enrichment in the number of fusions when experimental probes are used, indicating the interaction is present in ethanol-treated cells. Comparing control probes (P1/P2) with experimental probes (P2/P3) in estrogen-treated (E2) cells, there is an extremely significant (Fisher's Exact Test 2-tailed p-value = $3.33981e^{-59}$) enrichment in the number of fusions when experimental probes are used, indicating the interaction is present in estrogen-treated cells. Comparing control probes (P1/P2) in ethanol-treated (ET) cells with control probes (P1/P2) in estrogen-treated (E2) cells, there is a very weakly significant difference between the two datasets (Fisher's Exact Test 2-tailed p-value = 0.044127). The control site is therefore weakly estrogen-dependent. By contrast, comparing experimental probes (P2/P3) in ethanol-treated (ET) cells with control probes (P2/P3) in estrogen-treated (E2) cells, there is a significant difference between the two datasets (Fisher's Exact Test 2-tailed p-value = $9.7873e^{-12}$). The experimental site is therefore strongly

estrogen-dependent – that is, the interaction is present in more of the estrogen-treated cells than the ethanol-treated cells. ***Note: FISH analysis was performed together with Valere Cacheux-Rataboul, Genome Institute of Singapore, Singapore.***

siRNA knockdown

MCF-7 cells were seeded in hormone depleted medium for 1 day prior to transfection. 100 nM siGENOME Non-Targeting siRNA Pool #1 or ERα ON-TARGETplus SMARTpool siRNA (Dharmacon) was then transfected into MCF-7 cells using Lipofectamine 2000 (Invitrogen) according to manufacturer's protocol. 48 hrs following transfection, the cells were treated with either E2 or ethanol for 45 min (for western blot analysis, 3C and ChIP assays) or 8 hrs (for mRNA analysis). Total RNA was isolated with TRI® Reagent (Sigma) and purified using QIAGEN RNeasy. The RNA was reverse transcribed with oligo $(dT)_{15}$ primer (Promega), dNTP Mix, and M-MLV RT (Promega). Real-time PCR quantification was performed as described earlier. All experiments were repeated at least twice. ***Note: siRNA knockdown analysis was performed by the lab of Edwin Cheung.***

References

Adams, M.D. S.E. Celniker R.A. Holt C.A. Evans J.D. Gocayne P.G. Amanatides S.E. Scherer P.W. Li R.A. Hoskins R.F. Galle et al. 2000. The genome sequence of Drosophila melanogaster. *Science* 287: 2185-2195.

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* 355: 632-634.

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

Al-Dhaheri, M.H., Y.M. Shah, V. Basrur, S. Pind, and B.G. Rowan. 2006. Identification of novel proteins induced by estradiol, 4-hydroxytamoxifen and acolbifene in T47D breast cancer cells. *Steroids* 71: 966-978.

Ali, S. and R.C. Coombes. 2000. Estrogen receptor alpha in human breast cancer: occurrence and significance. *J Mammary Gland Biol Neoplasia* 5: 271-281.

Barski, A., S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823-837.

Bashir, A., S. Volik, C. Collins, V. Bafna, and B.J. Raphael. 2008. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* 4: e1000051.

Bhinge, A.A., J. Kim, G.M. Euskirchen, M. Snyder, and V.R. Iyer. 2007. Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Res* 17: 910-916.

Birney, E. J.A. Stamatoyannopoulos A. Dutta R. Guigo T.R. Gingeras E.H. Margulies Z. Weng M. Snyder E.T. Dermitzakis R.E. Thurman et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816.

Blanco, L., A. Bernad, J.M. Lazaro, G. Martin, C. Garmendia, and M. Salas. 1989. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem* 264: 8935-8940.

Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev. 1993. dbEST--database for "expressed sequence tags". *Nat Genet* 4: 332-333.

Bovee, D., Y. Zhou, E. Haugen, Z. Wu, H.S. Hayden, W. Gillett, E. Tuzun, G.M. Cooper, N. Sampas, K. Phelps et al. 2008. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat Genet* 40: 96-101.

Branco, M.R. and A. Pombo. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* 4: e138.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18: 630-634.

Brentani, H. O.L. Caballero A.A. Camargo A.M. da Silva W.A. da Silva, Jr. E. Dias Neto M. Grivet A. Gruber P.E. Guimaraes W. Hide et al. 2003. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc Natl Acad Sci U S A* 100: 13418-13423.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268: 78-94.

Cai, S., C.C. Lee, and T. Kohwi-Shigematsu. 2006. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet* 38: 1278-1288.

Campbell, P.J., P.J. Stephens, E.D. Pleasance, S. O'Meara, H. Li, T. Santarius, L.A. Stebbings, C. Leroy, S. Edkins, C. Hardy et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*.

Carninci, P. and Y. Hayashizaki. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol* 303: 19-44.

Carninci, P. T. Kasukawa S. Katayama J. Gough M.C. Frith N. Maeda R. Oyama T. Ravasi B. Lenhard C. Wells et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309: 1559-1563.

Carroll, J.S., X.S. Liu, A.S. Brodsky, W. Li, C.A. Meyer, A.J. Szary, J. Eeckhoute, W. Shao, E.V. Hestermann, T.R. Geistlinger et al. 2005. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 122: 33-43.

Carroll, J.S., C.A. Meyer, J. Song, W. Li, T.R. Geistlinger, J. Eeckhoute, A.S. Brodsky, E.K. Keeton, K.C. Fertuck, G.F. Hall et al. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* 38: 1289-1297.

Carter, D., L. Chakalova, C.S. Osborne, Y.F. Dai, and P. Fraser. 2002. Long-range chromatin regulatory interactions in vivo. *Nat Genet* 32: 623-626.

Cawley, S., S. Bekiranov, H.H. Ng, P. Kapranov, E.A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A.J. Williams et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116: 499-509.

Chen, J., Y.C. Kim, Y.C. Jung, Z. Xuan, G. Dworkin, Y. Zhang, M.Q. Zhang, and S.M. Wang. 2008a. Scanning the human genome at kilobase resolution. *Genome Res* 18: 751-762.

Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, J. Jiang et al. 2008b. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133: 1106-1117.

Chiu, K.P., C.H. Wong, Q. Chen, P. Ariyaratne, H.S. Ooi, C.L. Wei, W.K. Sung, and Y. Ruan. 2006. PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. *BMC Bioinformatics* 7: 390.

Collins, F.S., M.L. Drumm, J.L. Cole, W.K. Lockwood, G.F. Vande Woude, and M.C. Iannuzzi. 1987. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* 235: 1046-1049.

Collins, F.S. and S.M. Weissman. 1984. Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. *Proc Natl Acad Sci U S A* 81: 6812-6816.

Consortium, T.E. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306: 636-640.

Cremer, T. and C. Cremer. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2: 292-301.

Cullen, K.E., M.P. Kladde, and M.A. Seyfred. 1993. Interaction between transcription regulatory regions of prolactin chromatin. *Science* 261: 203-206.

Dekker, J., K. Rippe, M. Dekker, and N. Kleckner. 2002. Capturing chromosome conformation. *Science* 295: 1306-1311.

Deschenes, J., V. Bourdeau, J.H. White, and S. Mader. 2007. Regulation of GREB1 transcription by estrogen receptor alpha through a multipartite enhancer spread over 20 kb of upstream flanking sequences. *J Biol Chem* 282: 17335-17339.

Dostie, J., T.A. Richmond, R.A. Arnaout, R.R. Selzer, W.L. Lee, T.A. Honan, E.D. Rubio, A. Krumm, J. Lamb, C. Nusbaum et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16: 1299-1309.

Dunn, J.J., S.R. McCorkle, L. Everett, and C.W. Anderson. 2007. Paired-end genomic signature tags: a method for the functional analysis of genomes and epigenomes. *Genet Eng (N Y)* 28: 159-173.

Dunn, J.J., S.R. McCorkle, L.A. Praissman, G. Hind, D. Van Der Lelie, W.F. Bahou, D.V. Gnatenko, and M.K. Krause. 2002. Genomic signature tags (GSTs): a system for profiling genomic DNA. *Genome Res* 12: 1756-1765.

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95: 14863-14868.

Esteban, J.A., M. Salas, and L. Blanco. 1993. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem* 268: 2719-2726.

Euskirchen, G.M., J.S. Rozowsky, C.L. Wei, W.H. Lee, Z.D. Zhang, S. Hartman, O. Emanuelsson, V. Stolc, S. Weissman, M.B. Gerstein et al. 2007. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res* 17: 898-909.

Feinberg, A.P., R. Ohlsson, and S. Henikoff. 2006. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* 7: 21-33.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269: 496-512.

Fraser, P. and W. Bickmore. 2007. Nuclear organization of the genome and the potential for gene regulation. *Nature* 447: 413-417.

Fuchs, E. and K. Weber. 1994. Intermediate filaments: structure, dynamics, function, and disease. *Annu Rev Biochem* 63: 345-382.

Fullwood, M.J. and Y. Ruan. 2009a. ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem*.

Fullwood, M.J., J.J. Tan, P.W. Ng, K.P. Chiu, J. Liu, C.L. Wei, and Y. Ruan. 2008. The use of multiple displacement amplification to amplify complex DNA libraries. *Nucleic Acids Res* 36: e32.

Fullwood, M.J., C.L. Wei, E.T. Liu, and Y. Ruan. 2009b. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* 19: 521-532.

Garmendia, C., A. Bernad, J.A. Esteban, L. Blanco, and M. Salas. 1992. The bacteriophage phi 29 DNA polymerase, a proofreading enzyme. *J Biol Chem* 267: 2594-2599.

Gerhard, D.S. L. Wagner E.A. Feingold C.M. Shenmen L.H. Grouse G. Schuler S.L. Klein S. Old R. Rasooly P. Good et al. 2004. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* 14: 2121-2127.

Giresi, P.G., J. Kim, R.M. McDaniell, V.R. Iyer, and J.D. Lieb. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17: 877-885.

Griffiths, A.D. and D.S. Tawfik. 2006. Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol* 24: 395-402.

Guigo, R., E.T. Dermitzakis, P. Agarwal, C.P. Ponting, G. Parra, A. Reymond, J.F. Abril, E. Keibler, R. Lyle, C. Ucla et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci U S A* 100: 1140-1145.

Hagege, H., P. Klous, C. Braem, E. Splinter, J. Dekker, G. Cathala, W. de Laat, and T. Forne. 2007. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* 2: 1722-1733.

Harris, T.D., P.R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J.W. Efcavitch et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320: 106-109.

Hashimoto, S., Y. Suzuki, Y. Kasai, K. Morohoshi, T. Yamada, J. Sese, S. Morishita, S. Sugano, and K. Matsushima. 2004. 5'-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol* 22: 1146-1149.

Holt, R.A. and S.J. Jones. 2008. The new paradigm of flow cell sequencing. *Genome Res* 18: 839-846.

Hon, W.K., T.W. Lam, K. Sadakane, K.W. Sung, and S.M. Yiu. 2007. A space and time efficient algorithm for constructing compressed suffix arrays. *Algorithmica* 48: 23-36.

Hong, G.F. 1981. A method for sequencing single-stranded cloned DNA in both directions. *Biosci Rep* 1: 243-252.

Horike, S., S. Cai, M. Miyano, J.F. Cheng, and T. Kohwi-Shigematsu. 2005. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet* 37: 31-40.

Hsu, F., W.J. Kent, H. Clawson, R.M. Kuhn, M. Diekhans, and D. Haussler. 2006. The UCSC Known Genes. *Bioinformatics* 22: 1036-1046.

Hubbard, T.J., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts et al. 2007. Ensembl 2007. *Nucleic Acids Res* 35: D610-617.

Johnson, D.S., A. Mortazavi, R.M. Myers, and B. Wold. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316: 1497-1502.

Kapranov, P., S.E. Cawley, J. Drenkow, S. Bekiranov, R.L. Strausberg, S.P. Fodor, and T.R. Gingeras. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296: 916-919.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* 31: 51-54.

Kidd, J.M., G.M. Cooper, W.F. Donahue, H.S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* 453: 56-64.

Kim, J., A.A. Bhinge, X.C. Morgan, and V.R. Iyer. 2005. Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat Methods* 2: 47-53.

Korbel, J.O., A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318: 420-426.

Korf, I., P. Flicek, D. Duan, and M.R. Brent. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* 17 Suppl 1: S140-148.

Kushner, P.J., D.A. Agard, G.L. Greene, T.S. Scanlan, A.K. Shiau, R.M. Uht, and P. Webb. 2000. Estrogen receptor pathways to AP-1. *J Steroid Biochem Mol Biol* 74: 311-317.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Lee, J. and S. Safe. 2007. Coactivation of estrogen receptor alpha (ER alpha)/Sp1 by vitamin D receptor interacting protein 150 (DRIP150). *Arch Biochem Biophys* 461: 200-210.

Lim, C.A., F. Yao, J.J. Wong, J. George, H. Xu, K.P. Chiu, W.K. Sung, L. Lipovich, V.B. Vega, J. Chen et al. 2007. Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation. *Mol Cell* 27: 622-635.

Lin, C.Y., V.B. Vega, J.S. Thomsen, T. Zhang, S.L. Kong, M. Xie, K.P. Chiu, L. Lipovich, D.H. Barnett, F. Stossi et al. 2007. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet* 3: e87.

Ling, J.Q., T. Li, J.F. Hu, T.H. Vu, H.L. Chen, X.W. Qiu, A.M. Cherry, and A.R. Hoffman. 2006. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science* 312: 269-272.

Loh, Y.H., Q. Wu, J.L. Chew, V.B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu et al. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* 38: 431-440.

Lu, X. and E.B. Lane. 1990. Retrovirus-mediated transgenic keratin expression in cultured fibroblasts: specific domain functions in keratin stabilization and filament formation. *Cell* 62: 681-696.

Lupien, M., J. Eeckhoute, C.A. Meyer, Q. Wang, Y. Zhang, W. Li, J.S. Carroll, X.S. Liu, and M. Brown. 2008. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* 132: 958-970.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.

Marioni, J.C., C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*.

Mastrangelo, I.A., A.J. Courey, J.S. Wall, S.P. Jackson, and P.V. Hough. 1991. DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. *Proc Natl Acad Sci U S A* 88: 5670-5674.

Matsumura, H., S. Reich, A. Ito, H. Saitoh, S. Kamoun, P. Winter, G. Kahl, M. Reuter, D.H. Kruger, and R. Terauchi. 2003. Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc Natl Acad Sci U S A* 100: 15718-15723.

Mauro, M.J., M. O'Dwyer, M.C. Heinrich, and B.J. Druker. 2002. STI571: a paradigm of new agents for cancer therapeutics. *J Clin Oncol* 20: 325-334.

Meaburn, K.J. and T. Misteli. 2007a. Cell biology: chromosome territories. *Nature* 445: 379-781.

Meaburn, K.J., T. Misteli, and E. Soutoglou. 2007b. Spatial genome organization in the formation of chromosomal translocations. *Semin Cancer Biol* 17: 80-90.

Metivier, R., G. Penot, M.R. Hubner, G. Reid, H. Brand, M. Kos, and F. Gannon. 2003. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* 115: 751-763.

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome Res* 15: 1767-1776.

Milner, R.J. and J.G. Sutcliffe. 1983. Gene expression in rat brain. *Nucleic Acids Res* 11: 5497-5520.

Misteli, T. 2007. Beyond the sequence: cellular organization of genome function. *Cell* 128: 787-800.

Mitelman, F., B. Johansson, and F. Mertens. 2007. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* 7: 233-245.

Moll, R., M. Divo, and L. Langbein. 2008. The human keratins: biology and pathology. *Histochem Cell Biol* 129: 705-733.

Morin, R., M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, and M. Marra. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 45: 81-94.

Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5: 621-628.

Myers, E.W., G.G. Sutton, A.L. Delcher, I.M. Dew, D.P. Fasulo, M.J. Flanigan, S.A. Kravitz, C.M. Mobarry, K.H. Reinert, K.A. Remington et al. 2000. A whole-genome assembly of Drosophila. *Science* 287: 2196-2204.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-1349.

Ng, P., J.J. Tan, H.S. Ooi, Y.L. Lee, K.P. Chiu, M.J. Fullwood, K.G. Srinivasan, C. Perbost, L. Du, W.K. Sung et al. 2006a. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res* 34: e84.

Ng, P., C.L. Wei, and Y. Ruan. 2006b. Paired-End diTagging for Transcriptome and Genome Analysis. In *Current Protocols in Molecular Biology, 2006, Unit 21.12* (eds. F.M. Ausubel R.

Brent R.E. Kingston D.D. Moore J.G. Seidman J.A. Smith, and K. Struhl). John Wiley and Sons, Inc.

Ng, P., C.L. Wei, and Y. Ruan. 2007. Paired-end diTagging for transcriptome and genome analysis. *Curr Protoc Mol Biol* Chapter 21: Unit 21 12.

Ng, P., C.L. Wei, W.K. Sung, K.P. Chiu, L. Lipovich, C.C. Ang, S. Gupta, A. Shahab, A. Ridwan, C.H. Wong et al. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2: 105-111.

Osborne, C.S., L. Chakalova, K.E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J.A. Mitchell, S. Lopes, W. Reik et al. 2004. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* 36: 1065-1071.

Pan, Y.F., K.D. Wansa, M.H. Liu, B. Zhao, S.Z. Hong, P.Y. Tan, K.S. Lim, G. Borque, E.T. Liu, and E. Cheung. 2008. Regulation of estrogen receptor-mediated long-range transcription via evolutionarily conserved distal response elements. *J Biol Chem*.

Parra, G., P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett, and R. Guigo. 2003. Comparative gene prediction in human and mouse. *Genome Res* 13: 108-117.

Phatnani, H.P. and A.L. Greenleaf. 2006. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* 20: 2922-2936.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 20: 207-211.

Pruitt, K.D., T. Tatusova, and D.R. Maglott. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61-65.

Putney, S.D., W.C. Herlihy, and P. Schimmel. 1983. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* 302: 718-721.

Raghavendra, N.K. and D.N. Rao. 2005. Exogenous AdoMet and its analogue sinefungin differentially influence DNA cleavage by R.EcoP15I--usefulness in SAGE. *Biochem Biophys Res Commun* 334: 803-811.

Ren, B., F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290: 2306-2309.

Rogers, M.A., L. Edler, H. Winter, L. Langbein, I. Beckmann, and J. Schweizer. 2005. Characterization of new members of the human type II keratin gene family and a general evaluation of the keratin gene domain on chromosome 12q13.13. *J Invest Dermatol* 124: 536-544.

Rozen, S. and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds. S. Krawetz and S. Misener), pp. 365-386. Humana Press, Totowa, NJ.

Ruan, Y., H.S. Ooi, S.W. Choo, K.P. Chiu, X.D. Zhao, K.G. Srinivasan, F. Yao, C.Y. Choo, J. Liu, P. Ariyaratne et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* 17: 828-838.

Rubin, G.M. and E.B. Lewis. 2000. A brief history of Drosophila's contributions to genome research. *Science* 287: 2216-2218.

Sabo, P.J., M. Hawrylycz, J.C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, M.O. Dorschner et al. 2004a. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U S A* 101: 16837-16842.

Sabo, P.J., R. Humbert, M. Hawrylycz, J.C. Wallace, M.O. Dorschner, M. McArthur, and J.A. Stamatoyannopoulos. 2004b. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc Natl Acad Sci U S A* 101: 4537-4542.

Saha, S., A.B. Sparks, C. Rago, V. Akmaev, C.J. Wang, B. Vogelstein, K.W. Kinzler, and V.E. Velculescu. 2002. Using the transcriptome to annotate the genome. *Nat Biotechnol* 20: 508-512.

Schones, D.E., K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132: 887-898.

Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nat Methods* 5: 16-18.

Shastry, B.S. 2007. SNPs in disease gene mapping, medicinal drug development and evolution. *J Hum Genet* 52: 871-880.

Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-1732.

Shiraki, T., S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 100: 15776-15781.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38: 1348-1354.

Simonis, M., J. Kooren, and W. de Laat. 2007. An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* 4: 895-901.

Smit, A.F.A., R. Hubley, and P. Green. 1996-2004. RepeatMasker Open-3.0.

Stein, L.D., C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva et al. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res* 12: 1599-1610.

Steinert, P.M. and D.R. Roop. 1988. Molecular and cellular biology of intermediate filaments. *Annu Rev Biochem* 57: 593-625.

Strausberg, R.L., E.A. Feingold, R.D. Klausner, and F.S. Collins. 1999. The mammalian gene collection. *Science* 286: 455-457.

Su, W., S. Porter, S. Kustu, and H. Echols. 1990. DNA-looping and enhancer activity: association between DNA-bound NtrC activator and RNA polymerase at the bacterial glnA promoter. *Proc Natl Acad Sci U S A* 87: 5504-5508.

Sultan, M., M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321: 956-960.

Tolhuis, B., R.J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* 10: 1453-1465.

Toyota, M. and J.P. Issa. 2002. Methylated CpG island amplification for methylation analysis and cloning differentially methylated sequences. *Methods Mol Biol* 200: 101-110.

Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* 37: 727-732.

van der Hage, J.A., L.J. van den Broek, C. Legrand, P.C. Clahsen, C.J. Bosch, E.C. Robanus-Maandag, C.J. van de Velde, and M.J. van de Vijver. 2004. Overexpression of P70 S6 kinase protein is associated with increased risk of locoregional recurrence in node-negative premenopausal early breast cancer patients. *Br J Cancer* 90: 1543-1550.

Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler. 1995. Serial analysis of gene expression. *Science* 270: 484-487.

Venter, J.C., M.D. Adams, G.G. Sutton, A.R. Kerlavage, H.O. Smith, and M. Hunkapiller. 1998. Shotgun sequencing of the human genome. *Science* 280: 1540-1542.

Venter, J.C., H.O. Smith, and L. Hood. 1996. A new strategy for genome sequencing. *Nature* 381: 364-366.

Volik, S., B.J. Raphael, G. Huang, M.R. Stratton, G. Bignel, J. Murnane, J.H. Brebner, K. Bajsarowicz, P.L. Paris, Q. Tao et al. 2006. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* 16: 394-404.

Volik, S., S. Zhao, K. Chin, J.H. Brebner, D.R. Herndon, Q. Tao, D. Kowbel, G. Huang, A. Lapuk, W.L. Kuo et al. 2003. End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci U S A* 100: 7696-7701.

Wang, T.L., C. Maierhofer, M.R. Speicher, C. Lengauer, B. Vogelstein, K.W. Kinzler, and V.E. Velculescu. 2002. Digital karyotyping. *Proc Natl Acad Sci U S A* 99: 16156-16161.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.

Weber, J.L. and E.W. Myers. 1997. Human whole-genome shotgun sequencing. *Genome Res* 7: 401-409.

Wei, C.L., P. Ng, K.P. Chiu, C.H. Wong, C.C. Ang, L. Lipovich, E.T. Liu, and Y. Ruan. 2004. 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. *Proc Natl Acad Sci U S A* 101: 11701-11706.

Wei, C.L., Q. Wu, V.B. Vega, K.P. Chiu, P. Ng, T. Zhang, A. Shahab, H.C. Yong, Y. Fu, Z. Weng et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124: 207-219.

West, A.G. and P. Fraser. 2005. Remote control of gene transcription. *Hum Mol Genet* 14 Spec No 1: R101-111.

Whitesides, G.M. 2006. The origins and the future of microfluidics. *Nature* 442: 368-373.

Wilhelm, B.T., S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C.J. Penkett, J. Rogers, and J. Bahler. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453: 1239-1243.

Wold, B. and R.M. Myers. 2008. Sequence census methods for functional genomics. *Nat Methods* 5: 19-21.

Woodcock, C.L. 2006. Chromatin architecture. *Curr Opin Struct Biol* 16: 213-220.

Wurtele, H. and P. Chartrand. 2006. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res* 14: 477-495.

Yoshimura, S.H., H. Maruyama, F. Ishikawa, R. Ohki, and K. Takeyasu. 2004. Molecular mechanisms of DNA end-loop formation by TRF2. *Genes Cells* 9: 205-218.

Zeller, K.I., X. Zhao, C.W. Lee, K.P. Chiu, F. Yao, J.T. Yustein, H.S. Ooi, Y.L. Orlov, A. Shahab, H.C. Yong et al. 2006. Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc Natl Acad Sci U S A* 103: 17834-17839.

Zhao, X.D., X. Han, J.L. Chew, J. Liu, K.P. Chiu, A. Choo, Y.L. Orlov, W.K. Sung, A. Shahab, V.A. Kuznetsov et al. 2007. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* 1: 286-298.

Zhao, Z., G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K.S. Sandhu, U. Singh et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38: 1341-1347.

Appendices

Note: Appendices include a statement of work performed by myself, detailed protocols, manuals for using software, papers, and raw data. All appendices, and a PDF version of this thesis, are included in an attached CD-ROM.

Thank you for reading!

∞∞∞∞∞∞∞