

**Differential Global Effects of Selective Estrogen Receptor
Modulators on Estrogen Receptor Binding and Transcriptional
Regulation**

Lee Yew Kok

(B.Eng.(Hons.),NUS

NUS Graduate School for Integrative Sciences and Engineering

NATIONAL UNIVERSITY OF SINGAPORE

A Thesis submitted

For the degree of Doctor of Philosophy

2010

Acknowledgements

Here I sincerely thank my main supervisor – Professor Edison Liu for his excellent guidance, patience and sharing of knowledge. I am very grateful for his direct supervision and one-to-one meetings despite his busy schedule as a director of the Genome Institute of Singapore. I will always remember the paper reading sessions when he personally coached me. He has trained me and also provided me many opportunities to learn and acquire all the essential skills and thinking in doing research. Dr Jane Thomsen is another great supervisor who was always there to help and to show concern on my Ph.D work. Her enthusiasm and knowledge in research has also greatly inspired me. Throughout my Ph.D studies, she really helped to build up my knowledge on biology. Dr Jane is also a great friend to me, who listened to all my joys and woes in the laboratory and institution. Another great supervisor is Dr Krishnamurthy, who was very encouraging and imparted lots of bioinformatics knowledge to me. Being very approachable and intelligent, he was always ready to provide valuable solutions. I really appreciate that he always cares very much about my Ph.D progress. I had a wonderful time with him – I learnt many things under little pressure and the discoveries revealed from the data analysis were so exciting. I would also like to thank Dr Kartiki, whose suggestions and advice were always very helpful and right to the point. I also admire her management of the laboratory and profound knowledge in many areas from bioinformatics to biology. Lastly, I like to thank my beloved wife for her constant encouragement and love, and my family members for their understanding and support.

Table of Content

Acknowledgements.....	I
Table of Content	II
List of Tables	IX
List of Figures.....	XI
List of Illustrations.....	XVII
List of Equations.....	XVII
List of Acronym.....	XVIII
List of Acronym.....	XVIII
Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Scope and Strategy.....	11
1.3 Report Layout	13
Chapter 2 Construction of Customized Estrogen Receptor Binding Sites Array.	14
2.1 Selection of input regions	17
2.2 Array design considerations.....	28
2.3 Quality control of arrays	29
2.4 Checking the quality of reused arrays.....	33
2.5 Construction and quality control of HD2.1 Nimblegen chips for chromosome 21, 22 and additional regions	35
Chapter 3 Dynamics of Estrogen Receptor Binding in a Genome Wide Scale	39
3.1 ChIP-chip analysis mapped 6482 ER binding sites	39
3.2 Most ER binding sites are pre-occupied by ER and they have a greater ER recruitment upon E2 treatment	54

3.3	SERMs impose small changes to ER binding locations but greatly reduce ER binding affinity.....	58
3.4	ER-SERMs utilize tethering mechanism much more than ER-E2 and de novo motif predictions indicate shifting of preferential binding motif.....	65
3.5	Binding sites with basal occupancy are more accessible to TF than those without basal occupancy.....	68
3.6	Binding sites with basal occupancy show greatest FAIRE signals and highest H3K4Me1 enhancer marks, indicative of more accessible DNA regions...	71
3.7	FOXA1 does not play major role as a pioneering factor but largely attributed to constriction while GATA3 functions as co-factor.....	74
3.8	H3K4Me1 is the most predictive factor for identifying ER binding sites	80
3.9	ER regulates distinctive promoters and enhancers in Ishikawa cell line from MCF-7	82
3.10	Concluding remarks.....	84
Chapter 4	Integrative Analysis of SERMs on ER Responses on a Genome Wide Scale	86
4.1	Identification of regulated genes in SERMs and E2 treatments	86
4.2	E2-regulated genes use more of Pol II preloading mechanism and mechanism of down-regulation involves Pol II pausing or stalling	94
4.3	Strongly regulated genes in E2 treatment have ER binding sites in closer proximity than non-regulated genes	96
4.4	Strong E2-ER binding sites with basal occupancy associates with E2 up-regulated genes	97

4.5	Higher occurrence of ERE in E2-induced binding sites associated with higher regulated genes and higher binding sites fold change	98
4.6	Modulating effects of SERMs on gene expression.....	100
4.7	Differential trends of SERMs modulation on E2 up-regulated or down-regulated genes.....	105
4.8	Discovery of unique novel genes to SERMs and E2, exclusive of one another	108
4.9	ER tends to remain occupied across SERMs conditions for up-regulated genes.....	112
4.10	SERMs alter ER's spatial binding characteristics in promoter-context and cell environment.....	116
4.11	Revealing Spatiotemporal Expression Profiles of ER-responsive Genes in Different Tissues Upon E2 And SERMs Treatments	120
4.12	Concluding remarks	130
Chapter 5	Functional Analysis of Transcription Factor Binding Site Variants in Human Population	132
5.1	Identification and Genotyping Analysis of SNP.....	132
5.2	Molecular characterization of the p53 binding site within <i>PRKAG2</i> and its germ-line polymorphism (rs1860746)	134
5.3	Binding affinity by reporter assay analysis.....	140
5.4	Transcription activity by real-time PCR analysis	141
5.5	Polymorphism's impact on the protein levels by western blot analysis	143
5.6	Genetic association analysis of the p53 binding motif SNP (rs180746) with cancer susceptibility.....	145

5.7	Concluding remarks.....	146
Chapter 6	Conclusion	149
Chapter 7	Materials and Methods.....	156
7.1	Material and Methods for Binding Sites Array.....	156
7.2	Materials and Methods for Affymetrix Array.....	169
7.3	Materials and Methods for Functional Studies	175
	Appendices.....	188

Summary

Selective estrogen receptor modulators (SERMs) are used clinically to treat breast cancer as they inhibit estrogen both in promoting cell proliferation and expressing ER-mediated gene expression. SERMs are compound that block the effect of estrogen on estrogen receptor (ER). However, the complexity of estrogen receptor biology hinders an effective drug design. Our lab is interested in examining the global ER binding sites and the corresponding gene expression profiles upon treatment of ER by different SERMs.

Chromatin Immunoprecipitation assay (ChIP) was performed on MCF-7 breast tumor cells in the presence or absence of E2/SERMs or a combination of E2+SERMs and immunoprecipitated with ER α antibodies. We tested a panel of 24 validated binding and 27 non-binding control sites by real-time PCR analysis. Overall, our studies indicated that binding site variations were associated with differences in ER binding dynamics and intensity as a function of the ligand used. Subsequently, global studies investigating genome-wide binding sites through customized tiling array containing more than 40,000 mapped and putative ER binding sites from Nimblegen were initiated. The design issues and considerations for a customised array including the selection of probes were discussed. Various validations to assess the performance of the customised array were carried out.

Genome-wide binding sites profiles with the customised array were obtained for different drug treatments (E2 and SERMs), different antibodies (ER α , H3K4Me1, FOXA1 and GATA3), different experiments (ChIP and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)) and different cell lines (MCF-7 and Ishikawa cell lines). In order to mine out all the biological information, numerous

computational approaches are attempted and developed for data exploration and analysis. Various parameters and algorithms are being fine-tuned for extracting the best representative biological information. Variable Factor Linear Model (VFLM) was developed and implemented, which detected 6482 ER binding sites for CHIP-chip experiment immunoprecipitated with ER α antibody in E2 treatment. The VFLM peak-finding method was also implemented in the entire customised array data.

We also obtained genome-wide gene expression profiles with the Affymetrix array (HG-U133 Plus) for different drugs treatments (E2, T, R and I) at different time points (0, 3, 6, 9, 12, 24 and 48 hours). For Affymetrix experiments, the Pooled Variance Meta-analysis methods was used, followed by applying a Data-driven Smoothness Enhanced Variance Ratio Test (dSEVRAT) method for assessing the smoothness of the expression of gene across time point. We selected regulated genes based on 3 criterias: P-value \leq 0.05, smoothness score \geq 200 and fold change \geq 1.5.

In literature, correlations between binding and expression profiles to decipher the complex process of gene regulation have been made. Affymetrix experiments have been performed with different E2/SERMs treatments across varying time points in both MCF-7 and Ishikawa cells. The 2 cell lines allow comparison on the tissue-specific sensitivity, which is of therapeutical value. The comprehensive expression data were correlated with the binding profiles to infer direct target genes and functional binding sites.

Binding and transcription regulation are also substantially influenced by the chromatin structure. The presence of nucleosomes will limit the accessibility of transcription factors and their partners. Through joint effort, studies on the positioning of nucleosomes were carried out whereby all the nucleosome experiments were performed by other while the author was assigned the computational analysis. Two

Nimblegen high-density arrays with 2.1 millions probes have been designed, tiling the entire chromosome 21 & 22 arrays with additional selected regions throughout human genome.

Lastly, emerging evidences show that regulatory genetic variations have an influence on gene regulation like changing binding site recognition. For the functional studies, we have concentrated on Single Nucleotide Polymorphisms (SNPs) present within transcriptional binding sites and its biological functions. Since breast cancer cell lines for different binding sites polymorphism are not readily available, the polymorphism studies were performed in lymphoblastoid cell lines. ChIP analysis was performed on 8 different cell lines with different genotypes to assess the binding affinity. Furthermore, the binding characteristics associated with homogygous genotype were also carried out in an allele-specific Taqman assay. Interestingly, the SNPs have different regulation of the target gene PRKAG2 through expression studies and AMPK proteins through western blots. These studies above discovered and confirmed a functional SNP within binding site that exhibits an allele-specific transcription factor binding.

Together, the information from binding sites, gene expression profiles upon drug treatments, nucleosome profiles and the information from studying regulatory genetic variations will help to decipher the mechanism of Estrogen Receptor gene regulation over time and over several pharmacologic interventions: E2 and SERMS.

List of Tables

Table 1 Characteristic of Histones	6
Table 2 Regions selected for customized binding sites array include binding sites reported in literatures, ChIP-PET(Lin, 2007), ERE Prediction(Vega, 2006), ChIP-chip (Carroll, 2006) and negative controls	18
Table 3 ER binding sites validated in literatures or in-house	21
Table 4 ER non-binding sites validated in literatures or in-house.....	22
Table 5 Coverage of probes in input regions show good coverage as the regions not covered are due to repetitive, low complexity regions	36
Table 6 Majority of probes are less than 100bps spacings	36
Table 7 Correlation between arrays for Normalized Values shows that the correlation between biological replicates was about 0.42~0.50.....	37
Table 8 Selection of SERMs.....	40
Table 9 Results by Variable Factor Linear Model in MCF-7	46
Table 10 Distribution of Peaks Detected in Current Studies in Different Categories of Input Regions	47
Table 11 Distribution of detected 953 sites and 281 missed sites in high-confidence Lin (1234)	50
Table 12 Table comparing VFLM peaks and ChIP-Seq data. VFLM peaks have higher percentage of coverages in Lin and Carroll than the ChIP-Seq data.....	53
Table 13 Overlap between TE1, TE2 and TE3 peaks.....	54
Table 14 Distribution of full ERE, half ERE and no ERE.....	65
Table 15 Distribution of full ERE, half ERE and no ERE for unique binding sites to SERMs.....	66

Table 16 VFLM results for FOXA1 and GATA3	75
Table 17 Summary of AUC for all epigenetic marks alone and in combinations	82
Table 18 Linear model results on Ishikawa cell line	83
Table 19 Gene ontology for E2-regulated genes	94
Table 20 Probesets or genes enriched in each treatment w.r.t DMSO	103
Table 21 Binding sites detection across different treatments and down-regulated E2 genes	113
Table 22 Binding sites detection across different treatments and up-regulated E2 genes	113
Table 23 Genotype and allele frequencies for refSNP rs1860746.....	133
Table 24 Analysis of the association of SNP with cancer susceptibility under a recessive model of inheritance.....	146

List of Figures

Figure 1 Domain of Estrogen Receptor	3
Figure 2 Distribution of ChIP-PET cluster sizes	18
Figure 3 Distribution of ChIP-chip (Carroll, Meyer et al. 2006) cluster sizes	19
Figure 4 Profile of ChIP Enrichment after Drug Treatment for Binding Sites.....	23
Figure 5 Profile of ChIP Enrichment for Non-binding Sites	24
Figure 6 Binding Profiles for SERMs with E2	24
Figure 7 Binding Profiles for SERMs only.....	25
Figure 8 Remote regions > 100kb from all input regions.....	26
Figure 9 An example of an isolated ditag belongs to the PET1 classification	27
Figure 10 Histogram of all probes' melting temperature shows similar melting temperature which ensures similar hybridization specificity.....	29
Figure 11 Histogram of all probe lengths show about 53% of them are 45~47 nucleotides	30
Figure 12 Contour map of melting temperature plotted on GC and probe length axes illustrates that higher melting temperature corresponds to larger GC and longer probe length	31
Figure 13 Histogram of all probe spacings indicate almost 97% of probes have probe spacings less than 100bps	31
Figure 14 Histogram of Variance shows that 77% of all the probes had variances less than \log_2 ratios of E2-ERalpha(Cy5) over Input DNA(Cy3) = 0.3. It guarantees that the error of the mean < 30% over 3 replicates.....	32
Figure 15 Scatter Plot of 1 st and 2 nd Technical Replicates of E2 treatment show great reproducibility of the array	33

Figure 16 Scatter plots between the different reuses on stripped arrays show that same array can be reused up to 3 times.....	34
Figure 17 Histogram of All Probes' Melting Temperature for Nucleosome Array shows that about 39% of probes have melting temperature between 73°C to 77°C since this is not an isothermal array	36
Figure 18 Scatter plot between ratio for E1 and E2 that has a correlation of 0.49	37
Figure 19 Assessment of experiment data using GREB1, PTGES and IL6ST	41
Figure 20 Binding site profile for PTGES and IL6ST under SERMs condition	42
Figure 21 Distribution of Peaks Detected in Overlapped Input Regions. Highest percentages of peaks found to be common binding sites with ERE motif	48
Figure 22 Percentage of detected peaks increases with higher PET numbers of Lin category.....	49
Figure 23 Scatter plot on ER ChIP-on-chip intensity and qPCR.....	51
Figure 24 Detected 77 Peaks and Real-time Fold Change	51
Figure 25 Good Overlap between VFLM Peaks and ChIP-Seq Results	52
Figure 26 Linear model for MCF-7 (Takes any 2 out of 3).....	53
Figure 27 Overlap between binding sites found in E2 and DMSO treatment shows 65% of all E2 binding sites are pre-occupied	55
Figure 28 Histogram of Difference (E2 – DMSO).....	55
Figure 29 Scatter plot between E2 and DMSO treatment.....	56
Figure 30 GREB1 has basal occupancy while PTGES has no basal occupancy	57
Figure 31 Box plots between categories for E2_DM and E2_only	58
Figure 32 Peaks detected in MCF-7 by Linear Model.....	59
Figure 33 Overlap in peaks between E2 and SERMs (Percentage of SERMs).....	60
Figure 34 Overlap in peaks between E2 and SERMs (Percentage of E2).....	60

Figure 35 Overlap between E2, Tamoxifen and Raloxifene peaks	61
Figure 36 Peaks in Difference (Treatment – DMSO) detected in MCF-7 by Linear Model.....	62
Figure 37 Overlap in difference (treatment – DMSO) between E2 and SERMs (Percentage of E2).....	62
Figure 38 Overlap in difference (treatment – DMSO) between E2 and SERMs (Percentage of SERM)	63
Figure 39 Distribution of ratio intensities for E2, T, R and I Binding Sites.....	64
Figure 40 Distribution of ratio intensities for E2, TE, RE and IE Binding Sites.....	64
Figure 41 De novo motif prediction on top 500 E2 binding site	67
Figure 42 De novo motif prediction on unique SERMs binding site	67
Figure 43 Nucleosome profiles for E2_DM and E2_only	70
Figure 44 Profiles of nucleosome (Categorized w.r.t 1.5 fold change)	71
Figure 45 Profiles of FAIRE, K4Me1 and Nucleosome Signals.....	73
Figure 46 Overlap between FOXA1 E2 peaks and FOXA1 DM peaks	75
Figure 47 Overlap between GATA3 E2 peaks and GATA3 DM peaks.....	75
Figure 48 Overlap between FOXA1 peaks and ER peaks.....	76
Figure 49 Boxplot of Change in FOXA1 occupancy in sites co-occupy ER binding sites vs. those that are not used as ER binding sites from DM to E2 condition ..	77
Figure 50 Overlap between GATA3 peaks and ER peaks.....	78
Figure 51 Boxplot of Change in GATA3 occupancy in sites co-occupy ER binding sites vs. those that are not used as ER binding sites from DM to E2 condition ..	79
Figure 52 Plot of ROC curve for K4ME1 epigenetic mark.....	81
Figure 53 Overlap in binding sites between MCF-7 and Ishikawa with E2 treatment	83

Figure 54 Overlap in binding sites between MCF-7 and Ishikawa with tamoxifen treatment	84
Figure 55 Overlap in binding sites between MCF-7 and Ishikawa with R treatment..	84
Figure 56 Schematic of Affymetrix gene expression analysis.....	86
Figure 57 Selection of probesets of genes based on 3 criterias	89
Figure 58 Examples for classification of genes	90
Figure 59 Heatmap for E2-regulated genes in MCF-7	91
Figure 60 Heatmap for a panel of 15 well-known E2 responsive genes (true positives)	92
Figure 61 Heatmap for false negative genes and GAPDHS house-keeping gene	92
Figure 62 Fold Changes of Genes across time points for E2.....	93
Figure 63 Fold Changes of TFF1 Gene across time points for E2 and SERMs	93
Figure 64 Percentage of E2 up-regulated genes in proximity to Pol II binding sites ..	95
Figure 65 Correlation between gene expression and E2 binding sites	96
Figure 66 Ratio of Up/Down E2 Genes across different fold change	97
Figure 67 Average number of ERE per Binding Sites across different fold changes	99
Figure 68 Average number of regulated genes per Binding Sites across different fold change	99
Figure 69 Effects of SERMs on E2-regulated genes in MCF-7	101
Figure 70 Effects of SERMs on E2-regulated genes in MCF-7 with reference to E2	102
Figure 71 E2/SERMs – DMSO (In MCF-7).....	103
Figure 72 Suppression of E2-regulated Genes in Different Treatments.....	104
Figure 73 Summary of expression changes in terms of genes.....	105
Figure 74 Gene expression profiles within 5kbs of E2 Binding Sites	106

Figure 75 Boxplot for gene expression profiles within 5kb of E2 binding sites	107
Figure 76 Venn-diagram for intersections between E2, T, R and I	109
Figure 77 Heatmap for unique genes in E2	109
Figure 78 Heatmap for unique genes in T	110
Figure 79 Heatmap for unique genes in R	110
Figure 80 Heatmap for unique genes in I.....	111
Figure 81 E2 up-regulated genes within 5kb of ER binding sites	114
Figure 82 E2 down-regulated genes within 5kb of ER binding sites	114
Figure 83 Decision tree for classifying up- and down-regulated genes.....	116
Figure 84 Profile of FAIRE signal for E2 and Tamoxifen	117
Figure 85 Profile of FAIRE signal for E2 and Raloxifene	118
Figure 86 Profile of H3K4Me1 signal for E2 and Tamoxifen.....	119
Figure 87 Profile of H3K4Me1 signal for E2 and Raloxifene.....	119
Figure 88 No. of regulated genes across treatments in MCF-7 and Ishikawa cell lines	121
Figure 89 Up-regulation and down-regulation of genes across treatments in MCF-7 and Ishikawa cell lines	121
Figure 90 Heatmap for E2-regulated genes in Ishikawa.....	122
Figure 91 Effects of SERMs on E2-regulated genes in Ishikawa.....	123
Figure 92 Tissue-specific effects shown on MCF7 on E2-regulated genes in Ishikawa cell line.....	124
Figure 93 Tissue-specific effects shown on MCF-7 on E2-regulated genes in Ishikawa cell line (Boxplot)	125
Figure 94 Intersection between MCF-7 and Ishikawa cell lines in E2 regulated genes	126

Figure 95 Intersection between MCF-7 and Ishikawa cell lines in SERMs regulated genes	127
Figure 96 Tissue-specific effects shown on Ishikawa on E2-regulated genes in MCF-7 (Tree-view)	128
Figure 97 Tissue-specific effects shown on Ishikawa on E2-regulated genes in MCF-7 (Boxplot).....	129
Figure 98 refSNP rs1860746 is located within the consensus p53 motif	133
Figure 99 Significant enrichment of p21 binding site sequence after 5-Fu treatment	134
Figure 100. Preliminary Study on the Influence of rs1860746 on p53 Binding.....	136
Figure 101 Effect of SNP on Enrichment of binding sites after ChIP Assay	137
Figure 102 Validation of Taqman probes using Allelic Discrimination of Plot.....	138
Figure 103 Difference in Ct values across different DNA amount for Taqman Assay	139
Figure 104 Taqman Assay result to show allele-specific enrichment of ChIP DNA with C Allele	140
Figure 105 Functional analysis of the binding site sequence (226 bp fragment) and its polymorphism (rs184672) by reporter gene assay in wild-type and p53-null HCT116 cells with or without 5FU treatment	141
Figure 106. Preliminary Study on the Influence of rs1860746 on Gene Expression	142
Figure 107. Real-time PCR results for gene expression change of PRKAG2.....	143
Figure 108 Western blot analysis on AMPK sub-units, p53 and actin.....	144

List of Illustrations

Illustration 1 Batch effect correction in VFLM	44
Illustration 2 Sliding window approach to determine VFLM peaks.....	46
Illustration 3 How average nucleosome profiles are obtained.....	68
Illustration 4 Decision Tree for rules governing ER binding	80
Illustration 5 Expression values of E2/SERMs with reference to DM or E2	101

List of Equations

Equation 1 VFLM equation	45
Equation 2 VFLM equation for finding peaks for the difference	61

List of Acronym

ChIP	chromatin immunoprecipitation
ER	estrogen receptor
ER α	estrogen receptor α
ERE	estrogen response element
KG	known gene
moPET	maximum overlap PET
PET	paired end diTag
TFBS	transcription factor binding sites
TSS	transcriptional start site
VFLM	variable factor linear model
BAC	bacterial artificial chromosome
cDNA	complementary DNA
ChIP-Seq	chromatin immunoprecipitation with sequencing
DNA	deoxyribonucleic acid
FDR	false discovery rate
mRNA	Messenger RNA
PCR	polymerase chain reaction
qPCR	quantitative PCR
RNA	ribonucleic acid
TF	transcription factor
E/ E2	estradiol
D/ DM/ DMSO	dimethyl sulfoxide
T	4-hydroxytamoxifen
R	raloxifene hydrochloride
I	ICI 182,780
TE	4-hydroxytamoxifen + estradiol
RE	raloxifene hydrochloride + estradiol
IE	ICI 182,780 + estradiol
FAIRE	formaldehyde-assisted isolation of regulatory elements

Chapter 1 Introduction

1.1 Background

The important role of estrogen receptor in breast cancer

The most common form of malignant cancer faced by women worldwide is breast cancer. According to the Singapore Cancer Society, about 1000 women of all ethnic groups are diagnosed with breast cancer annually and the rate of the number of affected women is increasing at 3%. Breast cancer has a high incidence rate of about 12% for American and 4~5 % for Singaporean women (<http://www.singaporecancersociety.org.sg>). Breast cancer usually originates from the uncontrolled, excessive cell divisions occurring at the milk ducts and glands, thus forming a lump or tumour. If untreated, the cancer will invade the nearby stroma which consists of blood and lymphatic vessels and metastases to lung, bones and liver (Sledge and Miller 2003). The end result is death. Beside death, there is also great emotional impact on the well-being of the women as the treatments may involve lumpectomy or mastectomy which may represent a loss of femininity and beauty to the woman involved. There are undesirable side-effects of drugs that add to the discomfort and stresses on the patient. In both premalignant and malignant breast cancers, estrogen receptor α (ER α) is often found with higher protein levels than in normal tissue presence. Estrogen receptor can be used as one of the factors for predicting and diagnosing breast cancer (Ali and Coombes 2000) .

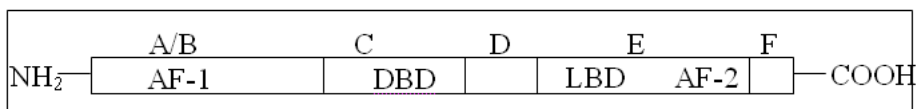
Estrogen receptor belongs to the family of steroid receptor and is activated by the hormone estrogen. Estrogen exerts its effects through growth and proliferation of breast tissue. When estrogen binds to an estrogen receptor, the receptor dissociates from its cytoplasmic chaperones, the receptor-associated proteins. The hormone-

receptor complex then moves to the nucleus, binds to the estrogen-response element (EREs) (5'-GGTCAnnnTGAXX-3') (Klinge 1999) through the DNA-binding domain (DBD) of the receptor and stimulates transcription. Alternatively, ER binds to DNA indirectly by tethering to other TFs like AP-1 (Kushner, Agard et al. 2000) and Sp1 (Porter, Saville et al. 1997). Preinitiation complex is formed by the assembly of RNA polymerase II (POL II), TATA-box-binding protein (TBP), associated factors and transcription factors. The ER-E2 complex also interacts with co-factors such as 160-kD steroid-receptor coactivator protein (P160) and p300-cyclic AMP response-element-binding protein (CBP). Some of the recruited factors have histone modifying activities that decondense the chromatin for accessibility of transcription factor to chromatin. The above description portrays a classical pathway of estrogen signal transduction. A modified pathway of estrogen signaling requires the pioneer factor of forkhead protein FoxA1 that opens the chromatin to allow ER accessibility (Carroll, Liu et al. 2005).

Two estrogen receptor subtypes - ER α and ER β

ER α was discovered and cloned in 1986 (Greene, Gilna et al. 1986) and only after about 10 years later, ER β was also discovered (Kuiper, Enmark et al. 1996). Both ER α and ER β have the following modular structure in Figure 1. The domain A/B located at the NH₂- terminal contains the ligand-independent activation function (AF-1) which has very low sequence homology (~18%) between ER α and ER β . AF-1 serves the purpose of cell- and promoter-specific transactivation. On the other hand, the DNA binding domain (DBD) has very high sequence homology (~97%) between ER α and ER β . The DBD located in C domain comprises of two zinc-fingers (Green, Kumar et al. 1988) that recognises specific hormone response elements and contains both the dimerisation and nuclear localisation signal (NLS). A hinge region (domain

D) serves to connect the domain C to domain E. In domain E, the ligand binding domain (LBD) and the ligand-dependent activation function (AF-2) are located. The sequence homology between ER α and ER β in domain E is around 60%. Domain E also contains the dimerisation and NLS.



ER α and ER β also have vastly different expressions in different tissues. Comparing the relative level of ER α and ER β when both ER subtypes are detected in the tissues, the relative level of ER α is found to be much higher than ER β in mammary gland, kidney, pituitary and uterus tissues. On the contrary, ER β is much higher than ER α in lung, bladder and prostate tissues. Both ER α and ER β are present in roughly equal distribution in bone, ovary, testes and thymus tissues. In liver tissue, only ER α is present. Since estrogen receptors are distributed in different levels in many tissues and estrogen enhances the activity of estrogen receptor, estrogen also exhibits differential effects in different tissues and examples include osteoporosis, prostate and colon cancer. It is found that the effects from estrogen are not only governed by the levels and the sub-types of the estrogen receptor found, but the estrogenic effects are also both promoter and cell-specific to the particular environments found in different tissues.

Selective Estrogen Receptor Modulator and ER

One of the major strategies to treat and prevent breast cancer is to inhibit the agonistic property of ER. Selective Estrogen Receptor Modulators (SERMs) have been developed as compounds with a mixed agonist/ antagonist activity on estrogen receptors. Ideally, SERMs should have an antagonist effect on breast and uterus tissue

but agonist effect on bone. SERMs are used clinically as drugs to treat and prevent breast cancer or osteoporosis despite we understand very little on their mechanism of actions and there are numerous side effects of drug administration. Side effects may include diarrhoea, pain at back and abdominal, vomiting, headache and constipation. Hormone therapy using Tamoxifen increases the risk of endometrial cancer, premature menopause in woman and even stroke. There are also cases that the drugs were ineffective and patients developed resistance to the drugs. If we understand more about the molecular basis of SERM action, better SERMs can be designed with negligible side effects and to be more effective with higher specificity. Tamoxifen was developed and used for treating early breast cancer since 1980s. Now Tamoxifen is used in advanced breast cancer, as an adjuvant for early breast cancer or post-operation in late breast cancer. However, Tamoxifen has good efficacy mostly in ER-positive breast cancer patients and most recurring breast cancer develops resistance to Tamoxifen. Based on the results provided from in-vitro assays which showed that both p300 and histone acetyl transferases (HATs) were recruited to the promoter of TFF1 E2-induced gene in Tamoxifen-resistant MCF-7 cell line, Shou, J. et al. suggests that HATs is involved in the mechanism of Tamoxifen resistance (Shou, Massarweh et al. 2004). Another SERM is ICI 182,780 (FaslodexTM), which was the first steroidal estrogen antagonist. It worked by degrading estrogen receptor, and thus removed the effects of estrogen. It maintained its efficacy even in breast carcinoma resistant to Tamoxifen therapy. The drug was shown to exhibit anti-proliferative effects on both breast and endometrium in both pre-clinical and clinical trial (Howell, Osborne et al. 2000). There are many SERMs that exhibit different degrees of agonist/antagonist effects in different tissues. Tamoxifen displays partial agonistic /antagonistic effects while ICI is a pure antagonist in breast tissues. Raloxifene is one

of the SERMs that exhibits both estrogenic and anti-estrogenic effects depending on the tissues. Raloxifene acts as an agonist in bone and lipid levels, but behaves as an antagonist in breast and uterine tissues (Fitzpatrick, Berrodin et al. 1999). Raloxifene has been shown to be effective in treating osteoporosis. Raloxifene also reduces the growth of breast cancer cell in vitro (Wolczynski, Surazynski et al. 2001). The knowledge on the mechanism of SERMs in terms of the interactions with ER and the affinity of the resulting ER-SERM complexes with DNA are as follows: ICI binds to ER with similar affinity as E2 whereas Tamoxifen has lesser affinity than E2 to ER. In both cases, chaperone proteins are also released. Unlike binding to E2 where both AF1 and AF2 are active, only AF1 is active in Tamoxifen binding that gives it partial agonist activity while both AF1 and AF2 are not active in ICI binding and ER is also rapidly degraded (Howell, Osborne et al. 2000). Tamoxifen competes with E2 through displacing the carboxy-terminal helix (H12) from co-activator docking site in the LBD domain. ICI eliminates completely any interaction between H12 and the LBD domain that H12 becomes very flexible and are not placed in any particular position like ER-E2 or ER_SERMs (Pike, Brzozowski et al. 2001). Different conformation changes will be induced by different ligands complexes with ER and these in turn cause differential ER stability. Wu, Yang et al. reported that GW5668 ligand similarly dislocates H12 like Tamoxifen but also decreases the ER stability, which may explain that GW5668 works in Tamoxifen resistance breast cancer (Wu, Yang et al. 2005). Besides opposing the action of ER in breast cancer treatment, depletion of the natural hormone E2 is an alternative treatment. This is implemented by the use of aromatase inhibitor which prevents the conversion of androgen to estrogen especially in post-menopausal woman. The inhibitor can also hinder the production of estrogen during

the productive years of premenopausal woman. As with most cancer, chemotherapy treatment also applies to breast cancer.

As can be seen above, ER exerts its diverse effects in many cell types through the ligand structures, the concentration of ER α and ER β , promoter context and the proportion of co-activators and co-repressors. For the subsequent sections, estrogen receptor will refer to ER α unless specified otherwise.

ER α has a half-life of 4-5 hour in both breast cancer and uterine tissue when ligand is not present (Eckert, Mullick et al. 1984; Monsma, Katzenellenbogen et al. 1984; Nardulli and Katzenellenbogen 1986). The protein stability of transcriptional factor is inversely proportional to its rate of transcriptional activities (Philips, Chalbos et al. 1993; Imhof and McDonnell 1996).

Transcriptional controls from histones

Another significant transcriptional control is the discovery of histones and their functions to control the accessibility of chromatin to transcription factors. Histones are lysine (K) and arginine (R) rich proteins, which are also highly conserved and basic. Table 1 shows the characteristic of histones of their molecular weight, number of amino acids and amino acid composition.

Table 1 Characteristic of Histones

<i>Histones</i>	<i>Molecular Weight (kDa)</i>	<i>No. of Amino Acids</i>	<i>Amino Acid Composition</i>	
			<i>Lysine(%)</i>	<i>Arginine(%)</i>
<i>H1</i>	<i>17.0~28.0</i>	<i>200-265</i>	<i>27</i>	<i>2</i>
<i>H2A</i>	<i>13.9</i>	<i>129-155</i>	<i>11</i>	<i>9</i>
<i>H2B</i>	<i>13.8</i>	<i>121-148</i>	<i>16</i>	<i>6</i>
<i>H3</i>	<i>15.3</i>	<i>135</i>	<i>10</i>	<i>15</i>
<i>H4</i>	<i>11.3</i>	<i>102</i>	<i>11</i>	<i>4</i>

Histones and other nuclear proteins tightly bound DNA and they form the simple 'beads on a string' structure, which is then packed into very compact chromatin. The

chromatin is also organised into two domains- euchromatin and heterochromatin. Euchromatin represents loosely-packed chromatin which also contains high concentration of active genes. Heterochromatin is tightly-packed chromatin which is usually saturated with highly repetitive DNA. There are also various proteins present for the assembly and packaging of chromatin, DNA repair, DNA transcription and replication, DNA and histone modifications and DNA recombinant. Nucleosome is the most basic structural component of chromatin, 180-200 base-pair DNA of which 146bps wraps to histone octamer consists of a pair of H2A, H2B, H3 and H4. The linker histone that ties the nucleosomes together is called H1. The nucleosome is a stable structure that controls the access of transcription factor to DNA through the interaction of positively-charged histone tails with negatively-charged DNA. Nucleosome is known for its function in DNA packaging and in carrying epigenetic information. The presence of nucleosomes strongly decreases the rate of transcription by preventing the access to promoters. Histone acetylations have a main role in remodeling chromatin structure, changing many nuclear processes. Acetylation has the effect of neutralizing the positive charge of histones, causing it to release the negatively-charged DNA-now accessible to transcription factor. Hypoacetylated histones cause genes to be transcriptionally inactive. Two groups of proteins involved in chromatin remodeling. The first group is histone acetyl transferases (HATs) which acetylates the N-terminal tails of histones and the second group is ATP-dependent proteins such as SWI/SNF complex [Mating type SWItching; Sucrose NonFermenting]. SWI/SNF complex consists of 11 proteins and they can reposition the nucleosome by sliding it down the DNA. RNA polymerase can move along the stretch of DNA by translocating the histones sequentially by 75 to 80 base pairs within the nucleosome thereby nucleosomes' positioning remains undisturbed. At

transcriptionally active genes, histone variants H2A.Z and H3.3 are found instead of H2A and H3. H2A.Z is incorporated into nucleosomes by the catalytic exchanger of SWR1 protein (Mizuguchi, Shen et al. 2004) while the replacement of histone H3 with histone variant H3.3 requires CHD1 motor protein (Konev, Tribus et al. 2007). Histone modification is more conserved than DNA methylation and is universal among the eukaryotic organism from yeast to human. On the other hand, DNA methylation is more common only in complex genomes of higher eukaryotic organisms.

Epigenetics is the study of processes that establish metastable (i.e. somatically heritable) states of gene expression without altering the DNA sequence. DNA methylation has been proven to play important role in epigenetics of the genome by regulating the higher-order chromatin structures. There are at least two development periods, in which genome-wide methylation occurs, resulting in proliferation and differentiation of many types of cells. The periods are during the development of germ cells and preimplantation embryo. Imprinting of genes and stem cell differentiation are also affected by the epigenetics reprogramming. Epigenetics is also likely to regulate the regeneracy of a whole organism from a small part in normal development and in cloned animals. Histones modifications also play similar roles as DNA methylation.

Distribution of ER binding sites

Many recent studies seem to suggest that promoter-proximal regions do not account for the majority of the estrogen-response genes. As such, it seems that there is a long distance regulation from ERE sites acting as enhancers such as looping effects, but the transcriptional regulation is carried out by the basic transcription machinery. Based on the above findings on genome distribution of transcription factors, the use of

promoter array targeting upstream region of 5' end of genes or CpG islands array targeting CG rich regions would miss a large percentage of loci. This is even more relevant in mammalian cells in which genes constitute only a small portion of the whole genome and transcriptional factors can function at variable distance and up to 25kb as proven in β -globin gene regulation experiment (Horak, Mahajan et al. 2002). Given the close proximity of tandem loci spacing on average of as little as a few bps, the use of PCR or BAC arrays are not appropriate as the probes are too long and do not provide resolution though they have greater genomic nucleotides coverage. Moreover, strand specificity cannot be determined and it is also labour intensive. Oligonucleotides-based array that have 25-60 bps probes and high-density features is the best type of array for surveying the binding sites. A good example is the studies on the unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 using all the non-repetitive sequences (Cawley, Bekiranov et al. 2004).

Single Nucleotide Polymorphisms (SNPs) and its roles in diseases and gene regulation

One of the main motivations in deciphering the detailed mechanism of transcription regulation is to provide cures to diseases and cancers. There is evidence that numerous diseases are highly associated with a particular allele while complex diseases are also affected by variants through subtle mechanisms. As such, there is emerging interest on regulatory genetic variations within expression promoter and transcription factor binding region. Genome wide studies were carried out to identify functional SNPs in the promoter regions of human gene (Mottagui-Tabar, Faghihi et al. 2005). There are also attempts to characterise the promoter polymorphism (Buckland, Hoogendoorn et al. 2005). Examples of association studies on promoter

polymorphism and diseases are BCL2 gene for chronic lymphocytic leukemia (Nuckel, Frey et al. 2006), Interferon regulatory factor-1 for hepatitis C virus infection (Cheong, Cho et al. 2006) and insulin VNTR polymorphism with type 1 diabetes (Durinovic-Bello, Jelinek et al. 2005). Evidence has shown that SNPs occur in higher frequency in gene promoter regions than those further upstream. Additionally, SNPs are also highly associated with transcription factor binding sites as compared with non-binding site sequences (Guo and Jamison 2005). Useful information from SNPs found in the binding sites can be extracted and used to tailor patient specific therapy. Numerous papers have shown that binding sites occur in many different places and different binding sites serve different functions. For example, after genome-wide analysis of CREB transcription factor on rat genome, only 40% of the total unique genomic loci were within 2kb of the transcriptional start site of an annotated gene and 49% were within 1kb of CpG islands (Impey, McCorkle et al. 2004).

Analysis of ChIP-on-chip data

The invention of ChIP technique enables the in-vivo capturing of the actual physical interaction between transcription factors and actual genome location through cross-linking the proteins and DNA together. Microarray technology has also been matured to an advanced level of chip manufacture with smaller features, accurate printing of probes and higher probe density per array. Nimblegen has synthesized 2.1 billions probes on a single chip. The use of microarray following ChIP experiment, which is also known as ChIP-on-chip allows the snapshots of genome-wide physical interactions of protein-DNA to be interrogated at the same time. Much of the initial focus has been to identify transcription factor binding sites across the whole genome. However, there is an increasing trend for quantifying the binding site affinity using

the microarray probe intensities. Comparison has been made to correlate the fold change obtained from real-time PCR with the microarray probe intensity. Much of the interests are on comparing the changes in the binding site affinity upon drug treatments as the level of binding site affinity is generally proportional to transcription activity, which in turn affects the gene regulation network. It is preferred to have a new algorithm that can directly quantify the change in the binding site intensity from drug A to drug B.

Moreover, the new algorithm should also be able to make full use of biological and technical replicates since the array technology is much cheaper and commonly available. Many replicates are made feasible as the arrays are now reused. Many protocols are invented and tested to strip the array and have successfully reused the stripped arrays. The most commonly CHIP-on-chip methods are MAT, ChIPOTle and MACS. These are single array methods that do not take into account of many factors. Hence there is a need for a new algorithm that considers many treatments, many technical or biological replicates, batches. As such, a new algorithm called Variable Factor Linear Modeling (VFLM) is developed in this report. VFLM works on the basis of a moving sliding window with linear modelling and it can be used for quantifying absolute affinity as well as differential affinity.

1.2 Scope and Strategy

My research focus is to understand the mechanisms of regulation by estrogen receptor and how it is modulated by SERMS. This is mainly achieved by eliciting and examining the modulation of the global ER binding sites and the gene expression profiles upon treatment of ER by different SERMs – Tamoxifen, Raloxifene and ICI. My strategy involves both the wet lab and the computational approaches, covering

several aspects of transcriptional regulation from chromatin structure to binding site profiling to gene expression profiling. Gene expression profiling experiments using Affymetrix chips and data analysis were performed in MCF-7 and Ishikawa cell lines for 8 different treatments: DMSO (control), E2, Tamoxifen+E2, Tamoxifen, Raloxifene +E2, Raloxifene, ICI+E2 and ICI with a time course. I designed and validated a customised Nimblegen array to profile ER binding. The customised nimblegen array put together all true and putative ER binding sites from various sources. The arrays were utilised to profile the ER binding sites and their modulation by SERMS by doing ChIP with different treatments and antibodies. Subsequently, different types of correlations between the binding profiles and gene expressions profiles were carried out to explore new interesting information on Transcriptional Regulation by ER.

As the binding and transcription regulation are substantially influenced by the chromatin structure which may be greatly defined by variations in the positioning/occupancies of nucleosomes, the presence of nucleosomes will limit the accessibility of transcription factors and their partners. To study the influence of nucleosomes, we have designed and validated a Nimblegen tiling array of 2.1M probes spanning the entire chromosomes 21 & 22 along with additional selected regions throughout human genome. Using this array, we profiled the nucleosome occupancies and their component modifications under E2 treatment. Cross-comparisons among binding site, gene expression and nucleosome positioning profiles have been made.

A novel method was developed to analyse the Nimblegen array data both from binding site profiling experiments as well as the nucleosome profiling experiments based on linear model.

Lastly, functional studies in binding site were carried out. For polymorphism studies in binding sites, different breast cancer cell lines that have vast variations in the polymorphism pattern in binding sites are not readily available. As such, we chose lymphoblastoid cell lines as the model system to study and characterise functional binding sites with polymorphism. This confirmed a functional SNP within binding site that exhibits an allele-specific transcription factor binding.

1.3 Report Layout

The first chapter of the report starts with the introduction and research significance. The background related the research project is explained and relevant literatures are reviewed. The scope and strategy in querying the hypotheses are also outlined. Chapter 2 describes all the design considerations and validations in constructing a customised ER binding sites array. Chapter 3 describes the biological experiments performed to examine and compare the binding sites' profiles with different treatments and different antibodies. Chapter 4 describes the results and the comparison of the gene expression landscape obtained after SERMs treatments. Chapter 5 is an exciting chapter where the linkage and correlation between binding sites and gene expressions are scrutinised in detail to unravel further new mechanism of gene regulation. Chapter 6 describes the functional studies on SNP found within a functional binding site. Chapter 7 is the concluding chapter, drawing summaries and linking the findings across the previous chapters. Lastly, detailed tables and graphs are provided in the appendices.

Chapter 2 Construction of Customized Estrogen Receptor Binding Sites Array

We have designed a specially customized Nimblegen tiling array which contains all-inclusive mapped & putative ER binding sites to investigate the modulation of ER binding at a genome scale under different treatments in a cost- and time-effective manner. We have collated more than 40,000 binding sites from the literature and ERE prediction algorithm (Vega, Lin et al. 2006).

The inception of high-throughput technologies such as ChIP-on-chip and ChIP-Sequencing enable many thousands of binding sites or the whole human genome to be interrogated. Earlier the detection of ER binding sites were limited to a few genomic locations at a time through experiments such as gel mobility shift assay, luciferase assay system and lastly ChIP experiments coupled with quantitative PCR. DNA microarray is a multiplex innovation that contains as large as millions of microscopic spots of DNA oligonucleotides, which target short section of a gene or DNA region of interests such as binding sites. The first high-throughput detection of ER binding sites were focused on chromosome 21 and 22 with Affymetrix Genechip Chromosome 21/22 1.0F Array Set (P/N 900545), which consists of 3 arrays set using 25-mer oligos at 35bps spacing tiling on all the unmasked regions specified by RepeatMasker . Altogether there were 57 ER binding sites detected across chromosome 21 and 22, in which majority of them were not found proximal to promoter regions of genes (Carroll, Liu et al. 2005). The first unbiased genome-wide survey on all ER binding sites were carried out in Affymetrix Human tiling 1.0 microarrays, which consists of 14 chips set using 25-mer oligos at 35bps spacing that span the entire non-repetitive regions defined by RepeatMasker . The study detected 3665 ER binding sites at

higher threshold of which only 4% of the binding sites were within 1kb promoter regions (Carroll, Meyer et al. 2006). Another unbiased method for genome-wide analysis of ER binding sites was the construction and sequencing of CHIP-PET library (Lin, Vega et al. 2007). Signature sequences of 18bps were extracted with MmeI restriction enzymes from both the 5' and 3' of CHIP DNA and self-ligated together to form a single PET (Paired-End diTags). Many different single PETs were concatenated and subsequently cloned into a plasmid for sequencing. This significantly reduced the amount of DNA to be sequenced yet the enriched CHIP DNA can be mapped to genome as the boundary of the sonicated fragments is demarcated by the Paired-End diTags. The binding sites are determined through comparing the tag counts generated from CHIP DNA relative to input DNA as the background. The CHIP-PET approach mapped 1234 high confidence ER binding sites of which only 4% of the binding sites were within 5kb from TSS of all genes. Taken together, the studies from Carroll, Meyer et al. and Lin, Vega et al. have unanimously shown that the majority of binding sites were far away.

Building on the vast information gained on the location characteristics of ER binding sites, a suitable customized CHIP-on-chip platform needs to be chosen. The above studies in interrogating the ER binding sites clearly prove that promoter-proximal regions do not account for the majority of the estrogen receptor binding sites. As such, the use of promoter array targeting upstream region of 5' end of genes or CpG islands array targeting CG rich regions would miss a large percentage of ER binding sites. Given that a sizeable number of ER binding sites are in close proximity to one another, the use of PCR or bacterial artificial chromosome (BAC) arrays are not appropriate as the probes are too long and do not provide resolution though they have greater genomic nucleotides coverage. Moreover, strand specificity cannot be

determined for PCR arrays as both the target and its complementary sequences are present at each feature. It is also labour intensive to create the individual ~1kb PCR fragment. Oligonucleotides-based array that have 25-75 bps probes and high-density features is the best type of array for surveying the binding sites. Nimblegen array system was chosen because it can accommodate about 385,000 isothermal probes with variable probe lengths 45-75 bps.

The customized Nimblegen array incorporates all the mapped and highly-putative ER binding sites identified from all the aforementioned studies. Another source of the probed regions came from ERE motif-finding program (Vega, Lin et al. 2006). There are also additional binding sites from the literature and in-house studies validated through quantitative PCR. The selection of all putative ER binding sites effectively shrinks the DNA regions to be tiled with probes yet still comprehensively surveys the whole genome. This also brings about a significant reduction in the number of chips required as well as saving in cost and time. Instead of 3 chips across chromosome 21 and 22 which detected 57 ER binding sites or 14 chips across whole genome which detected 3665 highly probable ER binding sites, a single chip is designed which can survey more than 40,000 highly-putative ER binding sites at once. One of the major limitations in ChIP-on-chip studies is that a few μg of DNA is required for array hybridization yet ChIP assay typically yield less than 50ng from a single 150mm round plate of cultured cells. In summary, a single customized array requires less sample, lower reagents cost and labor hour, but yet comprehensively examined many thousands of experimentally validated ER binding sites genome-wide.

Besides, the customized array also offers several improvements and advantages than the platforms used in previous studies. Nimblegen platform offers greater specificity as the 45-75mers probes are longer than the 25-mer probes in Affymetrix

system. Nimblegen platform utilized isothermal probes which ensure similar specificity in hybridization. A large number of important regions may be missing out from the non-repetitive regions. The customized Nimblegen array makes use of frequency of occurrence of probes in the whole genome that will increase the regions covered. The average probe spacings in customized array are also comparable to the average spacings in the Affymetrix chip. Nimblegen platform uses maskless synthesis process so there is flexibility in modifying the probes in the array. Lastly, the comparison using single chip compared to a set of multiple chips would not be subjected much to batch effects or human variation errors such as the use of master mix in reagents. Therefore, the chosen Nimblegen platform provides greater specificity, design flexibility, improved analysis, more uniform hybridization, and coverage of repetitive regions.

2.1 Selection of input regions

Our approach used to interrogate the binding sites profiles of a transcription factor affected by different drugs across the genome, was achieved through careful selection of probed regions. Since the regions are derived from previously validated binding sites, there is greater experimental confidence inherently that the binding sites detected on the customized chip would have been experimentally determined twice. Table 2 shows the summary of the regions selected for the customized ER binding sites array. The paragraphs that follow will describe the selected regions and explain the selection of regions in details.

Table 2 Regions selected for customized binding sites array include binding sites reported in literatures, ChIP-PET(Lin, 2007), ERE Prediction(Vega, 2006), ChIP-chip (Carroll, 2006) and negative controls

Categories of Regions	ChIP-PET (Lin,2007) 7574			ChIP-chip (Carroll,2006)	ERE Prediction (Vega,2006)	Binding Sites from Literature	Negative Control	
	moPET1-2	moPET3-5	moPET6+				Non-binding Sites from Literature	Exons Probes- 4997 Prokaryotes Probes- 100 PET1 Probes- 10036 Negative Probes - 5001
No. of Regions	6100	1171	303	10599	37499	55	68	20134 probes

Maximum overlap PET (moPET) is defined as the maximum number of PETs of all sub-regions in a cluster.

Selection of input regions from ChIP-PET Data (Lin, Vega et al. 2007)

ChIP-PET data will be referred to as Lin for the remaining sections of thesis. Criteria in choosing the probed regions from Lin include: (1) Ditags that specifically bind to 1 genomic location; (2) Clusters that contain more than 2 ditags. In total, an approximate 10 million bps long regions were selected. The ChIP-PET technology and the term moPET are described in the appendix. High confidence Lin data are from moPET 3+ while low confidence Lin data are from moPET2 category.

Since the selected ChIP-PET clusters were of variable lengths, we examined the distribution of clusters sizes (Figure 2). There were not many clusters longer than 4000bps. The bulk of the cluster sizes centre around 600 to 1200, which is closed to the optimal shearing size of about 500~1000bps for ChIP experiment.

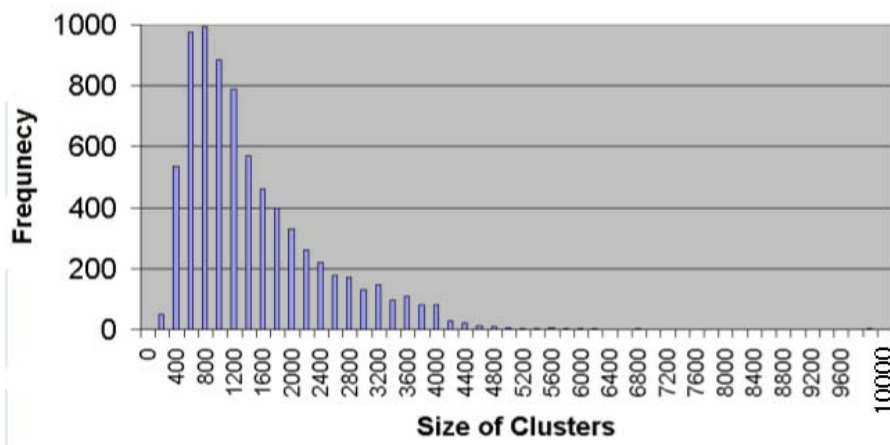


Figure 2 Distribution of ChIP-PET cluster sizes

Selected ChIP-PET clusters from specified criterias have variable lengths that the majority of the cluster sizes are closed to the optimal shearing size for ChIP experiment.

Selection of input regions from ChIP-chip (Carroll, Meyer et al. 2006)

10599 unique ER binding sites at the lower FDR threshold of 5%, including the higher confidence 3665 binding sites at FDR of 1% were included in the array. The binding sites were obtained with three biological replicates through the intersection of a nonparametric generalized Mann-Whitney U-test (P-value < 10^{-5}) and MAT algorithm. BLAT was performed to remove identical binding sites. The clusters were also of variable length. Greater than 80% of the clusters have sizes between 800 to 1000 bps (Figure 3).

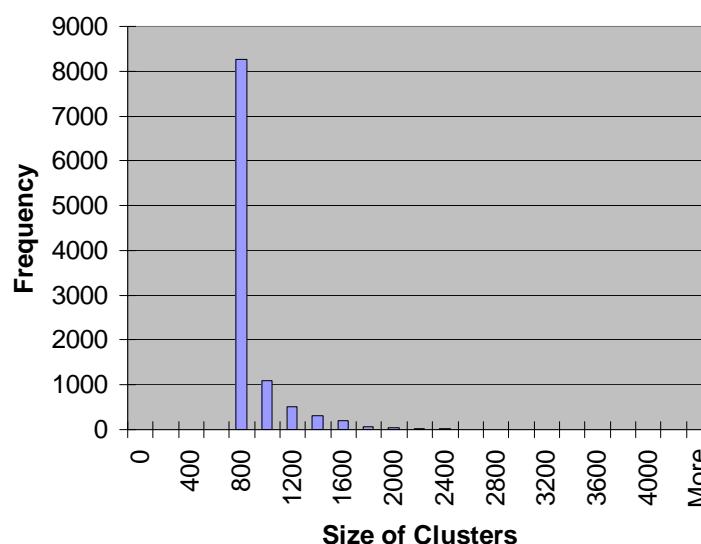


Figure 3 Distribution of ChIP-chip (Carroll, Meyer et al. 2006) cluster sizes

Selection of predicted binding sites

The predicted binding sites were obtained using the h-ERE algorithm which utilizes machine learning with training sets from both the validated binding and non-binding motifs collected from various experimental sources (Vega, Lin et al. 2006). A decision tree was used for classifying the potential binders and non-binders based on the binding and non-binding scores. The 'stringent' criteria of the algorithm were

used to define 38,024 putative ERE sites in the human genome. Each ERE has 19bps long sequence and was extended 250bps on both sides. Upon expanding each ERE site to 519 bps long, the number of input predicted regions has reduced to 37499 due to some ERE sites are in close proximity. A total of 220458 probes were tiled for the 37499 predicted binding (ERE) sites. When this number of probes added to the probes obtained in other categories, the total number of probes far exceeded what an array can accommodate. So the number of probes in the predicted regions was reduced to 182970 since other categories of input region are of higher importance. The predicted regions were further reduced by removing all the probes in the regions that have ≥ 1 non-unique probe (1529 regions removed); keeping only the “central” 6 probes for regions that have > 6 probes and lastly randomly removed 4546 regions.

Selection of ER binding sites from literature references and in-house studies

Validated ER binding sites and non binding sites were collected and probes had been designed and incorporated into the array. Validated ER binding sites and non binding sites are tabulated in Table 3 and Table 4 respectively.

Table 3 ER binding sites validated in literatures or in-house

Acc Num/Promoter ID*	Name	Genomic Location	Pattern	Reference
PROM0307S00001656	PDZK1	chr1:143,215,756-143,215,768	GGTCAcccAGTCC	Internal
NM_000674	ADORA1	chr1:199,790,269-199,790,281	GGTAgggTGACC	Bourdeau et al. 2004, Internal
PROM0307S00002573	ADORA1	chr1:199,790,414-199,790,426	GGTGTcttTGACC	Internal
NM_001227	CASP7	chr10:115,428,398-115,428,410	GGTCAgggTGAAC	Bourdeau et al. 2004
NM_001227	CASP7	chr10:115,428,492-115,428,504	GGTCCGgggTGAAC	Bourdeau et al. 2004
NM_001227	CASP7	chr10:115,428,572-115,428,584	GGTCAgggTGAAC	Bourdeau et al. 2004
NM_001227	CASP7	chr10:115,428,612-115,428,624	GGTCAgggTGAAC	Bourdeau et al. 2004
NM_001227	CASP7	chr10:115,428,652-115,428,664	GGTCAgggTGAAC	Bourdeau et al. 2004
NM_001227	CASP7	chr10:115,428,689-115,428,701	GGTCAgggTGAAC	Bourdeau et al. 2004
NM_001227	CASP7	chr10:115,428,743-115,428,755	GGTCAgggTGAAC	Bourdeau et al. 2004
PROM0307S000020873	FLJ30973	chr15:55,670,850-55,670,862	GGGCAgtgTGCC	Internal
PROM0307S000020873	FLJ30973	chr15:55,671,545-55,671,557	GGTCAcccTGCTC	Internal
NM_001089	ABCA3	chr16:2,319,793-2,319,805	GGTCAcggTGTC	Lin et al. 2004
NM_005082	EFF	chr17:52,323,321-52,323,333	GGTCAgtgTGACC	Klinge Review 2001, Bourdeau et al. 2004
PROM0307S000025715	MGC26694	chr19:19,035,118-19,035,130	GTTCAgagTGACC	Internal
PROM0307S000025862	KIAA1533	chr19:40,182,519-40,182,531	GGCCTggcTGACC	Internal
PROM0307S000026011	ACTN4	chr19:43,897,093-43,897,105	GGTCActgTGACT	Internal
PROM0307S000026372	GPR77	chr19:52,532,131-52,532,143	GGTCActcTGACA	Internal
NM_000064	C3	chr19:6,671,884-6,671,902	GGTGGcccTGACC	Klinge Review 2001
NM_148903	GREB1	chr2:11,603,634-11,603,646	GGTCAaaaTGACC	Bourdeau et al. 2004
NM_148903	GREB1	chr2:11,615,324-11,615,336	GGTCActcTGACC	Bourdeau et al. 2004
NM_148903	GREB1	chr2:11,621,861-11,621,873	AGTCAgtgTCACC	Internal
NM_148903	GREB1	chr2:11,623,258-11,623,270	GGTCActcTGACC	Bourdeau et al. 2004, Lin et al. 2004
PROM0307S00003333	CYP1B1	chr2:38,214,993-38,215,005	GGTCCgctTGCCC	Internal
PROM0307S00003333	CYP1B1	chr2:38,215,049-38,215,061	GGTCAaagCGGCC	Internal
NM_003489	NRIP1	chr21:15,359,833-15,359,845	GGTCAaagTGACC	Lin et al. 2004
PROM0307S000027970	TFF1	chr21:42,659,626-42,659,638	GGTCTctgTGTC	Internal
PROM0307S000027970	TFF1	chr21:42,659,906-42,659,918	AGCCAagaTGACC	Internal
NM_003225	TFF1	chr21:42,660,106-42,660,118	GGTCAcggTGCC	Klinge Review 2001
PROM0307S000028194	CRKL	chr22:19,595,695-19,595,707	AGTCAatcTAACC	Internal
NM_002343	LTF	chr3:46,481,739-46,481,751	GGTCAaggCGATC	Bourdeau et al. 2004
NM_001657	AREG	chr4:75,676,340-75,676,352	GGACAaggTGTC	Internal
NM_017770	ELOVL2	chr6:11,154,748-11,154,760	GGTCActcTGATG	Internal
NM_003376	VEGF	chr6:43,844,381-43,844,393	AATCAgacTGACT	Klinge Review 2001
NM_002346	LY6E	chr8:144,170,802-144,170,814	GGACAagaTGACC	Bourdeau et al. 2004
NM_004878	PTGES	chr9:129,597,654-129,597,666	GGACAgccTGCC	Internal
882	FLJ32833	chr1:108,492,542-108,492,560	ttagGTCAgctTGTCcag	Internal
137	C1orf21	chr1:181,327,606-181,327,624	ctgGGTCAgcaTGACCttc	Internal
1221	1221	chr1:64,942,548-64,942,566	ctgGGCAtgctCACTca	Internal
23b	SLC38A1	chr12:44,881,783-44,881,801	cagAGTGAactTGACCtga	Internal
23a	SLC38A1	chr12:44,881,800-44,881,818	gagGGTCAtccCAACCcca	Internal
1616	1616	chr16:2,781,142-2,781,160	ccaGGTCCgctTGCCctta	Internal
323	323	chr16:743,678-743,696	atgGGTCActgTGACCcag	Internal
1305	1305	chr17:46,382,536-46,382,554	cccGGACAagaTGTCccc	Internal
239	TEX14	chr17:54,072,183-54,072,201	caacGGTCAtggtTGACCtga	Internal
401	SEC15L2	chr2:72,713,948-72,713,966	ggaGGTCAggTGACCctcg	Internal
17	17	chr20:54,945,262-54,945,280	gggAGACAcccTGACCtaa	Internal
1607	1607	chr3:132,571,914-132,571,932	aggGGTCAtggtTGACAta	Internal
1259	SLC6A6	chr3:14,429,604-14,429,622	ctgGGTCActgTGTCcga	Internal
1361	SIAH2	chr3:151,957,126-151,957,144	acaGGTCAccaTGACCtgg	Internal
484	SNX24	chr5:122,216,372-122,216,390	cagGGTTAtctTAACCaac	Internal
242	PKIB	chr6:122,985,938-122,985,956	tttGGTCAtggtGGCCtga	Internal
1272	1272	chr6:23,720,183-23,720,201	tcgGGTCAtgctGGCCtggg	Internal
795	BTBD9	chr6:38,337,561-38,337,579	tggGGTCAtggtTGACTcct	Internal
95	SHB	chr9:37,943,504-37,943,522	gcaGGTGGgctTGCCcca	Internal

Table 4 ER non-binding sites validated in literatures or in-house

Acc Num/Promoter ID*	Name	Genomic Location	Pattern	Reference
NM_000549	TSHB	chr1:115,283,928-115,283,940	GGTCAgctTGACA	Bourdeau et al. 2004
NM_006472	TXNIP	chr1:142,927,222-142,927,234	GGTCAgtgGGATC	Internal
NM_000427	LOR	chr1:150,045,850-150,045,862	GGTCcaaaGGACC	Internal
NM_022365	DNAJC1	chr10:22,333,030-22,333,042	GTCaactTGTC	Internal
NM_000818	GAD2	chr10:26,545,037-26,545,049	GGTCGcagTGACC	Bourdeau et al. 2004
NM_000609	CXCL12	chr10:44,202,437-44,202,449	GGTCcagcTGCC	Internal
NM_000609	CXCL12	chr10:44,203,283-44,203,295	TGTCaaaaTGCC	Internal
NM_000926	PGR	chr11:100,509,203-100,509,215	AGTCatgtTGACA	Internal
NM_003646	DGKZ	chr11:46,321,832-46,321,844	GGCCAtgcTGCC	Internal
PROM0307S00016490	CTSW	chr11:65,403,499-65,403,511	GACCagccTGACC	Internal
PROM0307S00020376	C14orf131	chr14:101,872,078-101,872,090	GGCCAacaTGACA	Internal
PROM0307S00019823	DLG7	chr14:54,727,987-54,727,999	GGTCgtccAGACC	Internal
NM_001437	ESR2	chr14:63,876,354-63,876,366	GACCagccTGACC	Internal
NM_003246	THSB1	chr15:37,657,943-37,657,955	GGTCaatcCCACC	Internal
NM_024817	FLJ13710	chr15:69,737,514-69,737,526	AGTCatgtTGACC	Internal
NM_024817	FLJ13710	chr15:69,738,257-69,738,269	GGTCaatgTGCC	Internal
NM_024817	FLJ13710	chr15:69,738,459-69,738,471	GcTCActtTGTC	Internal
PROM0307S00021235	SH3GL3	chr15:82,077,053-82,077,065	GATCttgctTGACC	Internal
PROM0307S00021337	SMAP-1	chr15:89,278,745-89,278,757	AGTCaatcTGTC	Internal
NM_001089	ABCA3	chr16:2,321,166-2,321,178	GGTCttttTGACC	Internal
PROM0307S00021613	HCFC1R1	chr16:3,015,149-3,015,161	GACCagccTGACC	Internal
NM_001116	ADCY9	chr16:4,107,737-4,107,749	GGTCagcctGGTC	Internal
NM_001116	ADCY9	chr16:4,108,935-4,108,947	GGTCAaaaTGTC	Internal
PROM0307S00022205	CAPNS2	chr16:54,100,244-54,100,256	GGTCcgtcCGACC	Internal
NM_000430	PAFAH1B1	chr17:2,441,502-2,441,514	CCCATgtTGACC	Internal
NM_001552	IGFBP4	chr17:35,851,519-35,851,531	GATCAcgtTAACC	Internal
NM_001552	IGFBP4	chr17:35,853,510-35,853,522	GGTCatgtGCC	Internal
NM_002894	RBBP8	chr18:18,766,140-18,766,152	GGTCatctTGCTC	Internal
NM_017572	MKNK2	chr19:2,382,491-2,382,503	GGGCAgagTGAGC	Internal
PROM0307S00026367	BBC3	chr19:52,426,840-52,426,852	TGTCatgtTGTC	Internal
PROM0307S00026367	BBC3	chr19:52,427,249-52,427,261	GGTCagcctGGTC	Internal
NM_148903	GREB1	chr2:11,622,443-11,622,455	TCCAcccaTGACC	Internal
NM_148903	GREB1	chr2:11,625,143-11,625,155	TGTCaatcTGTC	Internal
PROM0307S00003992	EN1	chr2:119,322,563-119,322,575	GGTTAccctGAAC	Internal
NM_020120	UGCGL1	chr2:128,563,200-128,563,212	TGTCaaaaTGTC	Internal
NM_020120	UGCGL1	chr2:128,565,292-128,565,304	TGTCaatTGAGC	Internal
NM_002665	PLGL	chr2:87,884,778-87,884,790	GGTCagtgTGCCA	Internal
NM_005067	SIAH2	chr3:151,966,545-151,966,557	GcTCatagTGCCC	Internal
NM_024524	AFURS1	chr3:195,656,453-195,656,465	GGTCat taTGACC	Internal
NM_013324	CISH	chr3:50,626,609-50,626,621	GGCCAgagGACC	Internal
NM_014583	LMCD1	chr3:8,517,591-8,517,603	GGCCtgcTGACC	Internal
NM_032219	FLJ22269	chr4:673,249-673,261	GGGCAgagTGACT	Internal
NM_004354	CCNG2	chr4:78,433,176-78,433,188	GGACAactTGATC	Internal
NM_003714	STC2	chr5:172,689,912-172,689,924	GGCAaatgTGAAAC	Internal
NM_002184	IL6ST	chr5:55,327,909-55,327,921	GGTAGcaTGATC	Internal
NM_006622	PLK2	chr5:57,792,972-57,792,984	GGTTAcagCGACC	Internal
PROM0307S00010481	OLIG3	chr6:137,857,308-137,857,320	CGTCatccTAACC	Internal
PROM0307S00009635	FKBPL	chr6:32,206,228-32,206,240	GGCCAgccCGACC	Internal
PROM0307S00009635	FKBPL	chr6:32,206,311-32,206,323	CCCAcccaTGACC	Internal
NM_000602	SERPINE1	chr7:100,361,980-100,361,992	GACCagccTGACC	Internal
NM_000602	SERPINE1	chr7:100,362,938-100,362,950	GGCAagcTGCC	Internal
NM_000602	SERPINE1	chr7:100,363,852-100,363,864	TGTCaagaAGACC	Internal
PROM0307S00010895	TM4SF13	chr7:16,566,080-16,566,092	GATAAgctTGACC	Internal
NM_000712	BLVRA	chr7:43,570,289-43,570,301	GGTCactTGCGT	Internal
NM_000712	BLVRA	chr7:43,570,774-43,570,786	AGTCAaccTTACC	Internal
NM_001497	B4GALT1	chr9:33,157,593-33,157,605	GCTCAacCGACC	Internal
NM_001497	B4GALT1	chr9:33,158,622-33,158,634	GATCAgaaGGACC	Internal
PROM0307S00029895	PGPL	chrY:169,893-169,905	GCTCAcgaTGACG	Internal
229	SORCS1	chr10:108,692,194-108,692,212	cacAGTCatgtTGACCcca	Internal
1448	1448	chr14:38,648,346-38,648,364	attGGTCagagTGACAgaa	Internal

The binding sites were validated in-house using both LightCycler System and Applied Biosystems real-time PCR.

Real-time PCR validations on selected binding sites: 24 validated binding and 27 non-binding sites for estrogen receptor were randomly selected and examined using real-

time PCR. Enrichment on the above binding sites was obtained using primers flanking the specific ERE to detect and quantify the ChIP-ER samples and ChIP-GST control DNA obtained from each SERM. Detailed tables for the real-time PCR validations for the validated binding and non-binding sites can be found in Appendix A and Appendix B respectively.

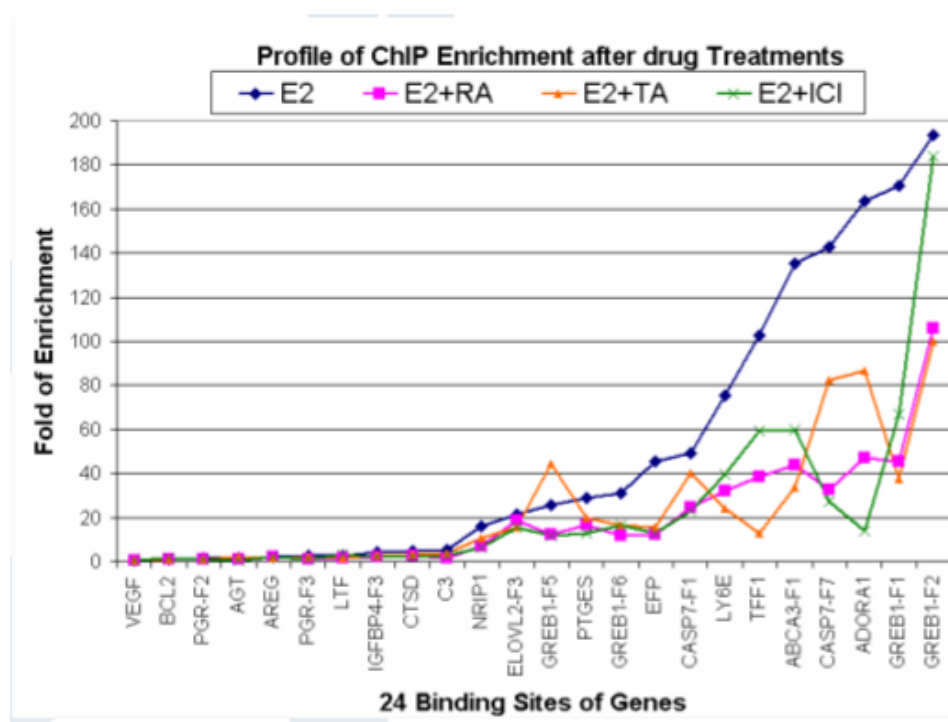


Figure 4 Profile of ChIP Enrichment after Drug Treatment for Binding Sites

Figure 4 is the profile of the enrichment for ERE binding sites. For all sites except 1, addition of SERM to E2 decreases the fold of enrichment to different degrees.

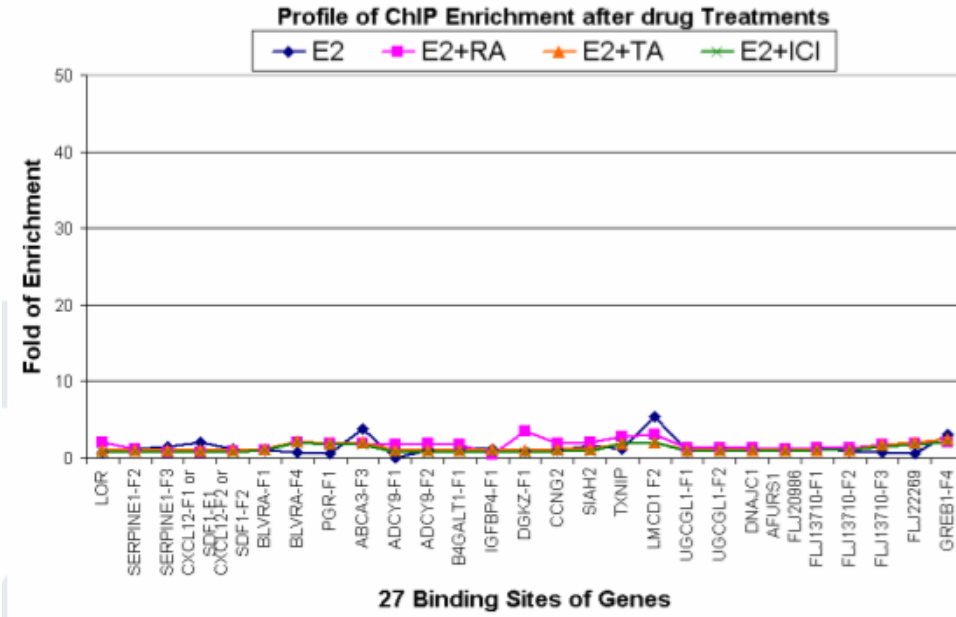


Figure 5 Profile of ChIP Enrichment for Non-binding Sites

Figure 5 is the profile of the enrichment for ERE non-binding sites. The fold of enrichments was low in general.

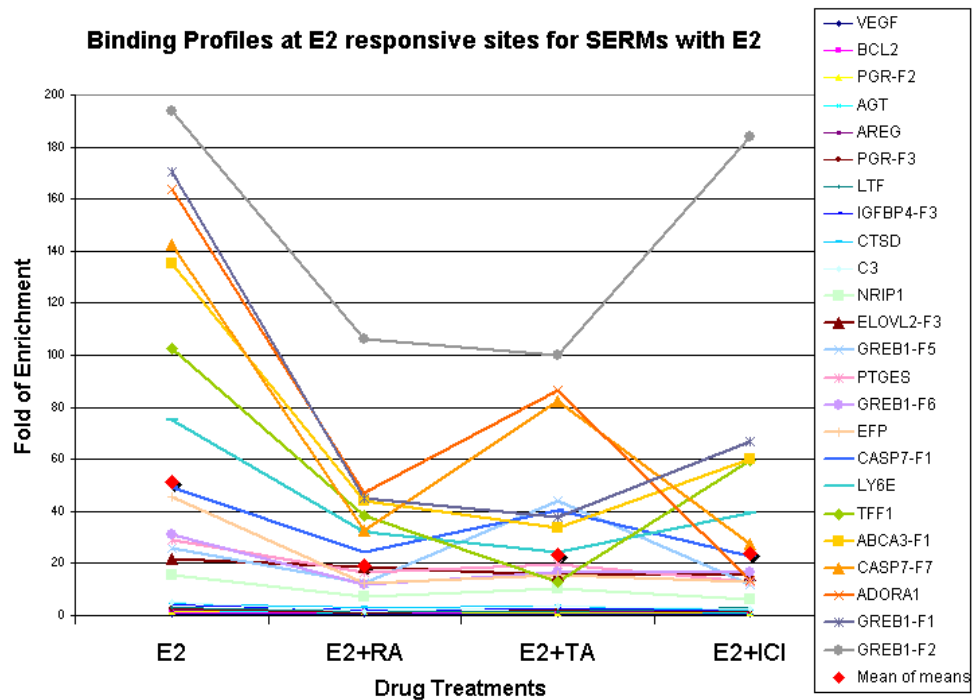


Figure 6 Binding Profiles for SERMs with E2

Y-axis is the fold of enrichment. X-axis represents the different treatments. Each line represents different binding sites.

This is a different view for all the ERE-binding sites and E2 + SERM decrease the enrichment in general (Figure 6).

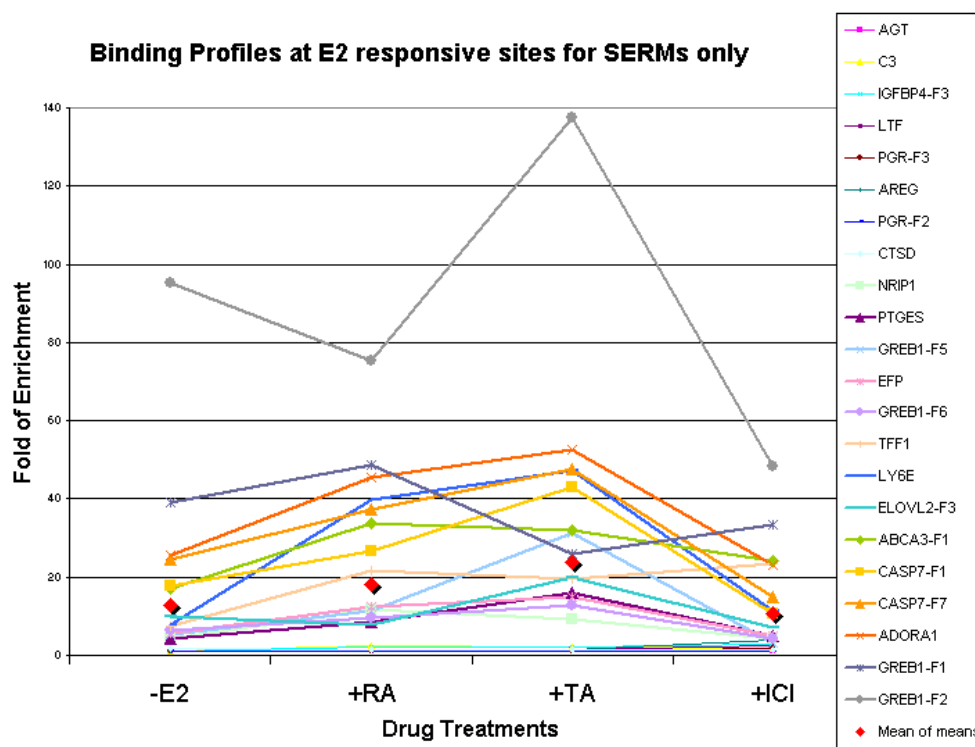


Figure 7 Binding Profiles for SERMs only

Figure 7 visualises the results for treatment with SERMs only. It is seen that RA or TA treatments bring up the enrichment in 14 and 15 binding sites respectively.

Selection of negative controls

Various negative controls were utilised in the arrays for functioning as hybridization controls and assessing the background signals in extraction of positive results. There were 4 types of negative controls that serve different functions and are likely to have different hybridization signals: 1) Remote regions >100kb from all input regions; 2) PET1 regions; 3) Exonic regions and 4) prokaryote sequences. We would also expect differential probe signals from closed chromatin, open chromatin not bounded with transcription factor and open chromatin with occupied transcription factor.

- 1) **Remote regions >100kb from all input regions:** 5001 negative probes selected from remote regions at least >100kb from all input regions (Figure 8). Sequences 100kb away from all input sequences would mostly contain closed chromatin.

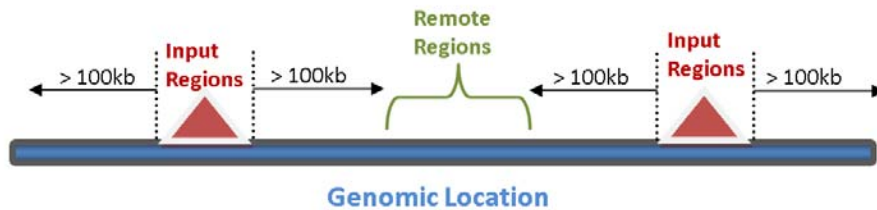


Figure 8 Remote regions > 100kb from all input regions

- 2) **PET1 regions:** Another category of negative probes came from unique tag with single isolated ditags classified as PET1 excluding all the predicted binding sites. An example of an isolated ditag belongs to the PET1 classification is depicted in Figure 9. About half of all PET1s regions were within 15kb of all our input regions. The filtering was based on PET1s >100kb from all our input regions and 3539 regions were extracted. Since PET1 is likely to contain more open chromatin with unbounded transcription factor, the signals from PET1 probes can be used to compare between open chromatin and open chromatin with bound transcription factor.

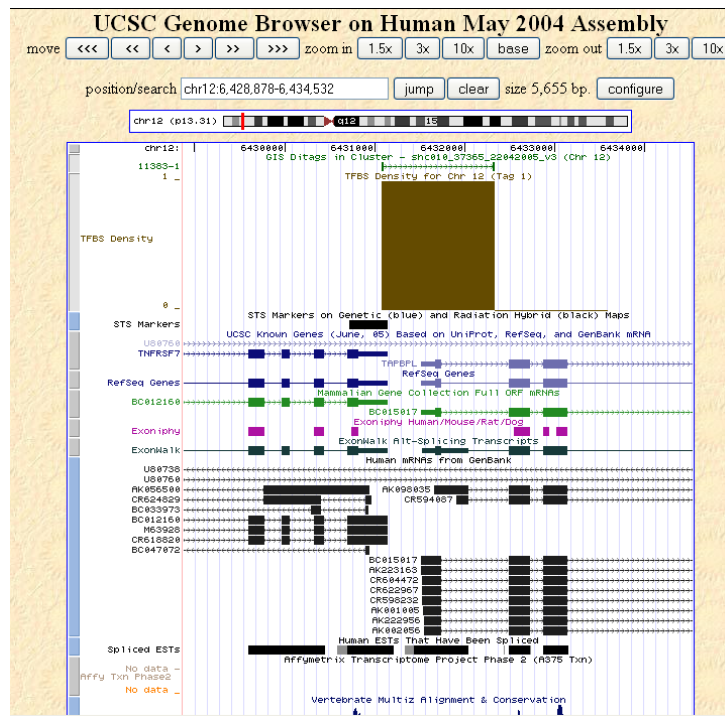


Figure 9 An example of an isolated ditag belongs to the PET1 classification

- 3) **Exonic regions:** There were also 4997 negative probes from translated exon regions in human genome that were expected to give the background signal. In terms of hybridization signals, probes from exons regions probably give no or little hybridization signal.
- 4) **Prokaryote Sequences:** There were 100 probes from prokaryotes sequences that have no homology and no similarity of up to 5 base pairs with the human genome. The prokaryotes probes are expected to give the lowest hybridization signals.

In summary, we have included negative probes that have variable range of expected hybridization signals and from different chromatin configurations.

2.2 Array design considerations

Tiling array design must strike a balance between various conflicting criteria for optimal design and usage of array. Criteria may include fixed-length or variable-length probes, probe uniqueness, probe spacings and probe melting points. The objectives are to tile the targeted regions with probes unbiasedly at the highest resolution and to facilitate optimal data analysis for discovering new biology.

Probes are designed using a sliding window and scoring on the composition within that window. Probes that have long runs (>5) of homopolymers or that are too GC rich or AT rich are avoided. The ability of the probes to self-anneal is also taken into account. No ambiguous bases must be present and probes should not require more than 148 cycles to be synthesized.

Variable-length probes with 45-75 mers will be easier than 50-mer fixed length probes in passing the design criteria that consist of no ambiguous bases and not more than 148 cycles in probe synthesis. Variable length probes are adopted so it is easier than using fixed-length probes to minimize the variance of the melting points of all the probes.

Besides, designing probes on only the unmasked regions based on Repeatmasker (Cawley, Bekiranov et al. 2004), another alternative is to make use of the frequency of the occurrence of probes in the whole genome. We may allow up to a certain frequency of occurrence such as 80 and mark those probes with high frequency. The masked regions are thought not to have biological significance. However, more evidence has illustrated the important roles these masked regions signify (Plohl, Luchetti et al. 2008).

It is pertinent to check the distribution and coverage of the probes on the designed regions. All the input regions have been tiled with at least 6 probes. Besides ensuring

the sufficient number of probes in a region, the probes should be consecutively close to one another and distribute evenly throughout the binding region.

2.3 Quality control of arrays

Quality of the arrays was firstly assessed by the properties of probes designed: melting temperature (T_m), probe lengths and probe spacings. The specificity of the hybridization of CHIP DNA samples to the targeted probes is very much dependent on the melting temperature of probes, i.e. the higher the melting temperature, the greater the specificity. Longer probe length also gives rise to higher melting temperature. The actual melting points of all probes are reasonably close to the ideal melting temperature of 76 °C. The design is an isothermal array that the about 70% of the probes have similar melting temperature between 73 to 77 °C which ensures similar specificity in hybridization and avoids any hybridization artifacts or bias (Figure 10).

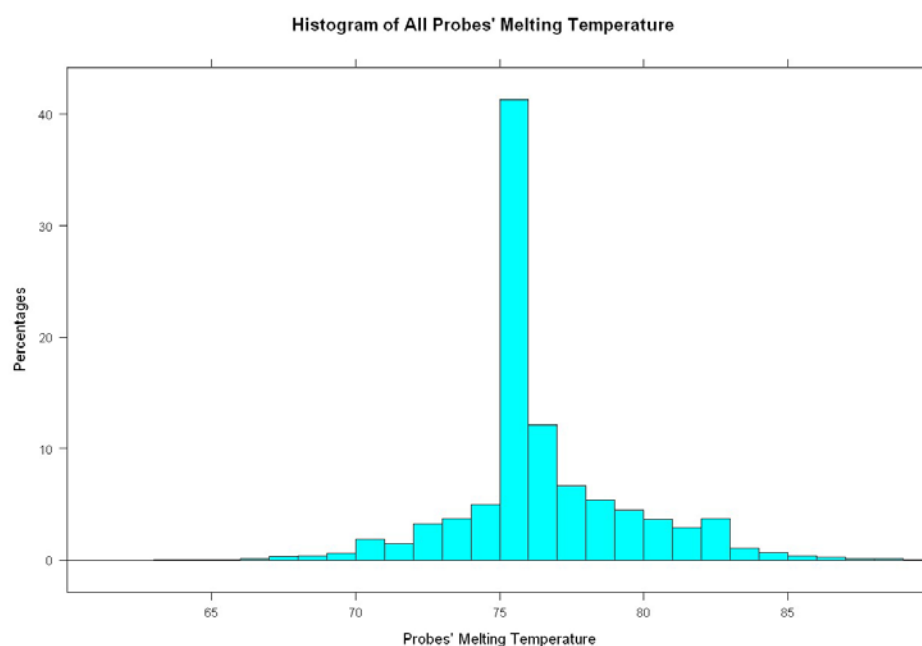


Figure 10 Histogram of all probes' melting temperature shows similar melting temperature which ensures similar hybridization specificity

Such a stringent control of the melting temperature was achieved by variable length probe design as can be seen in Figure 11. The probe lengths are variable and range from 45 to 75 nucleotides. There are about 53% of probes between 45 and 47 nucleotides long, significantly nearly half (47%) of the probes are of 48-65bp long.

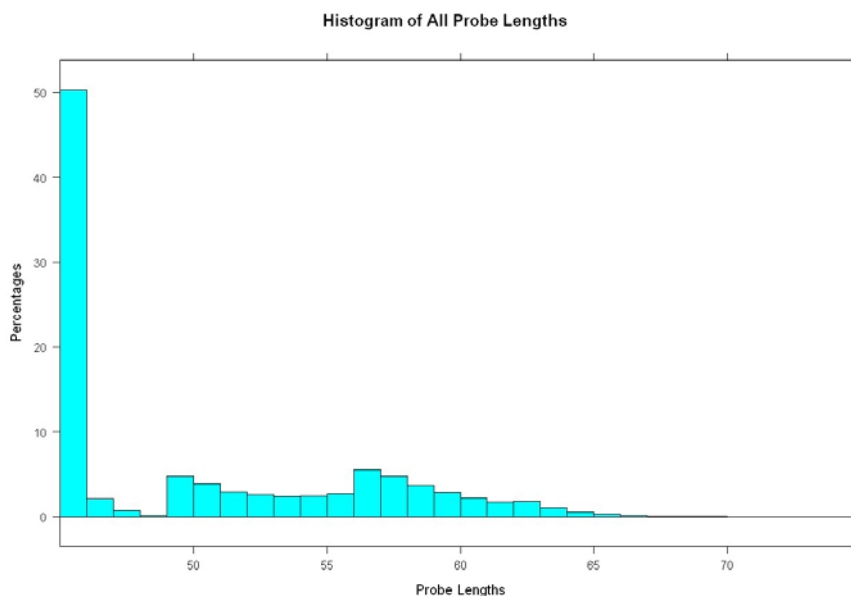


Figure 11 Histogram of all probe lengths show about 53% of them are 45~47 nucleotides

The relationship between melting temperature, probe length and GC content is summarized in Figure 12. In general, higher melting temperature corresponds to larger GC and longer probe length.

We also checked the probe spacings which are defined as from 5' end of 1 probe to the 5' end of the adjacent probe within the input regions. 76% of the probes are having probe spacings between 0 to 60 bps. Almost 97% probes have probe spacing less than 100bps (Figure 13).

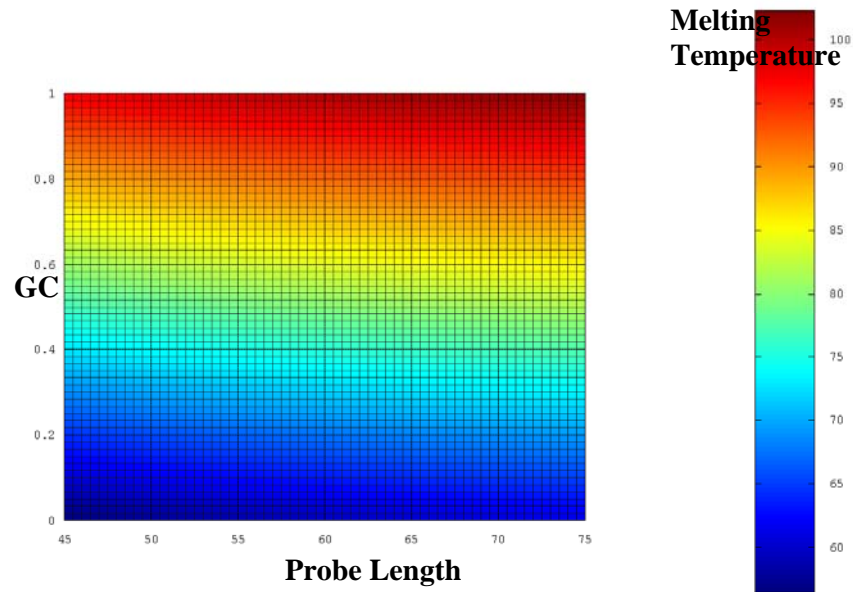


Figure 12 Contour map of melting temperature plotted on GC and probe length axes illustrates that higher melting temperature corresponds to larger GC and longer probe length

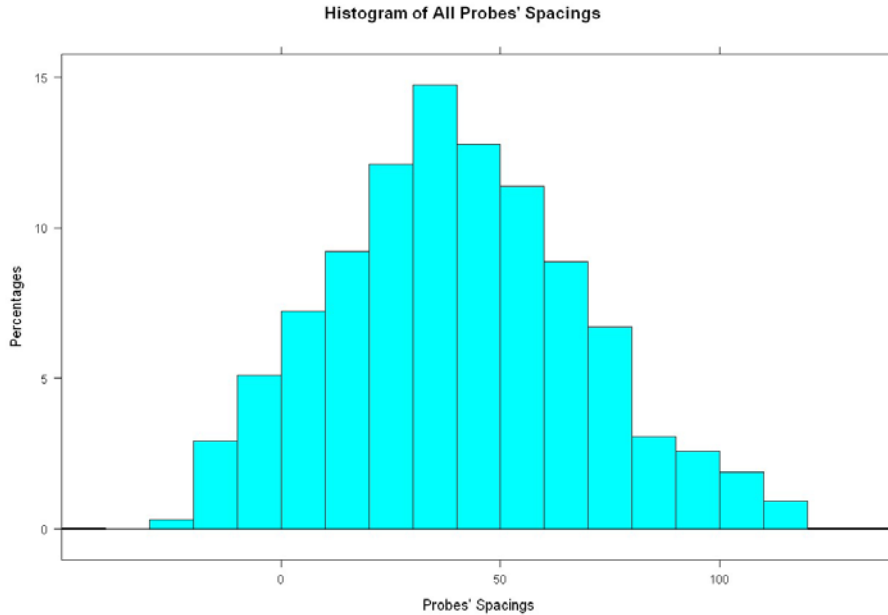


Figure 13 Histogram of all probe spacings indicate almost 97% of probes have probe spacings less than 100bps

Due to some overlapping input regions, the probes extracted from these regions had some negative probe spacings.

Quality control using 3 new arrays hybridized with the same samples

The objectives are to investigate the reproducibility and consistency of the array data generated across 3 different arrays with the same ChIP-E2 sample. Firstly, the variance across all the individual probes based on the \log_2 ratios of E2-ERalpha(Cy5) over Input DNA(Cy3) on 3 different new arrays was computed.

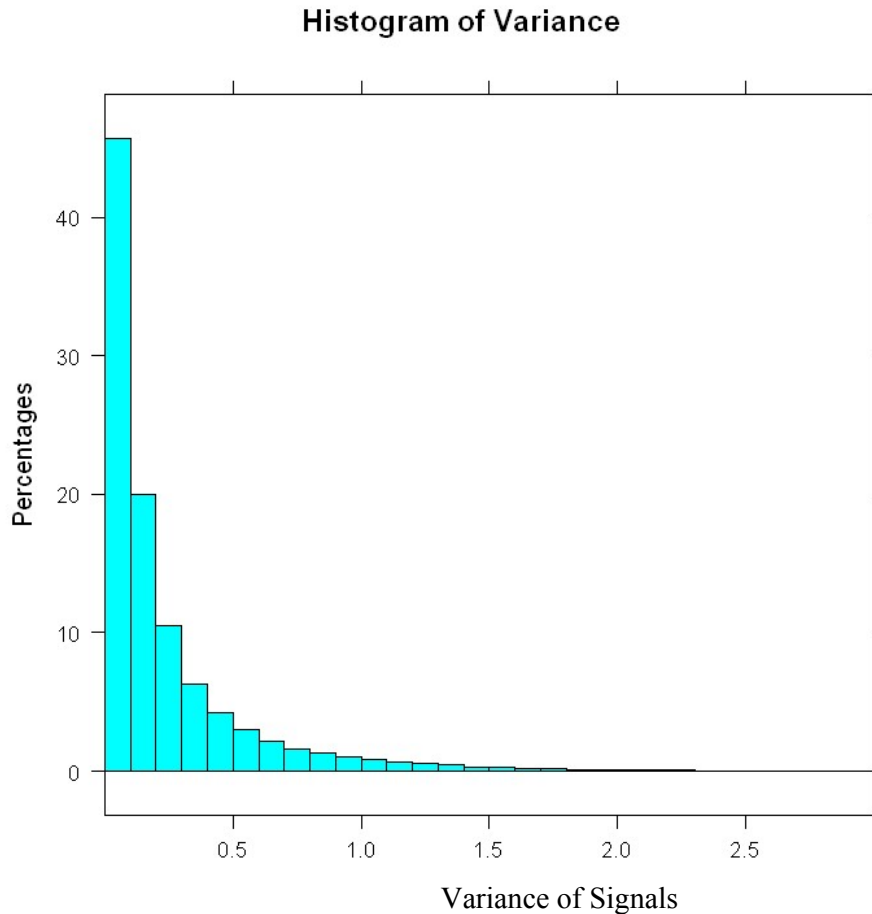


Figure 14 Histogram of Variance shows that 77% of all the probes had variances less than \log_2 ratios of E2-ERalpha(Cy5) over Input DNA(Cy3) = 0.3. It guarantees that the error of the mean < 30% over 3 replicates

The result showed that about 77% of all the probes had variances less than 0.3 (Figure 14). This means the array is capable of producing consistent probe ratios with little variations from array to array using the same sample.

Besides checking the variances of the arrays, we also investigated the reproducibility of the data from the same samples. The scatter plot in Figure 15 indicates that the 1st and 2nd technical replicates correlates very well that the data from the platform or arrays are reproducible.

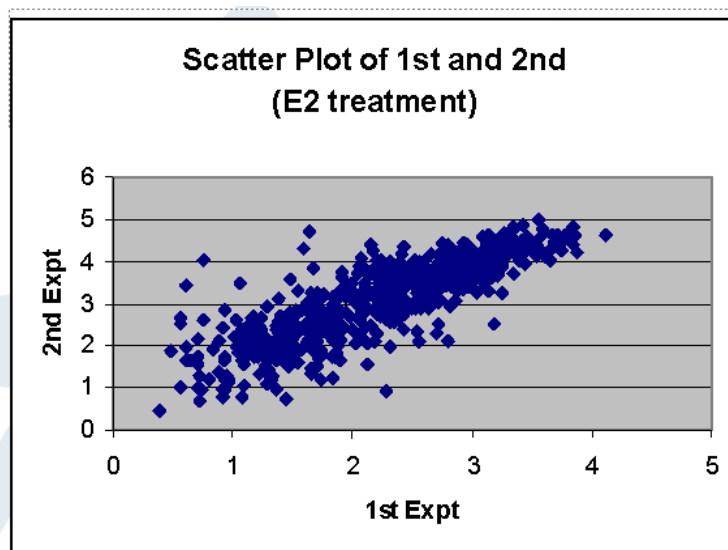


Figure 15 Scatter Plot of 1st and 2nd Technical Replicates of E2 treatment show great reproducibility of the array

2.4 Checking the quality of reused arrays

We checked the quality of reused arrays by hybridizing the same sample to the same arrays for four times that had been stripped 3 times. Figure 16A shows the scatter plots between first uses but scanned with different PMTs. Similarly, we plotted the scatter plots between 1st and 2nd uses in Figure 16B, between 1st uses and 3rd uses in Figure 16C and lastly between 1st uses and 4th uses in Figure 16D. Our study showed that the arrays are reusable up to 3rd use, the data from 1st and up to 3rd uses correlates very well but not the 1st and 4th uses.

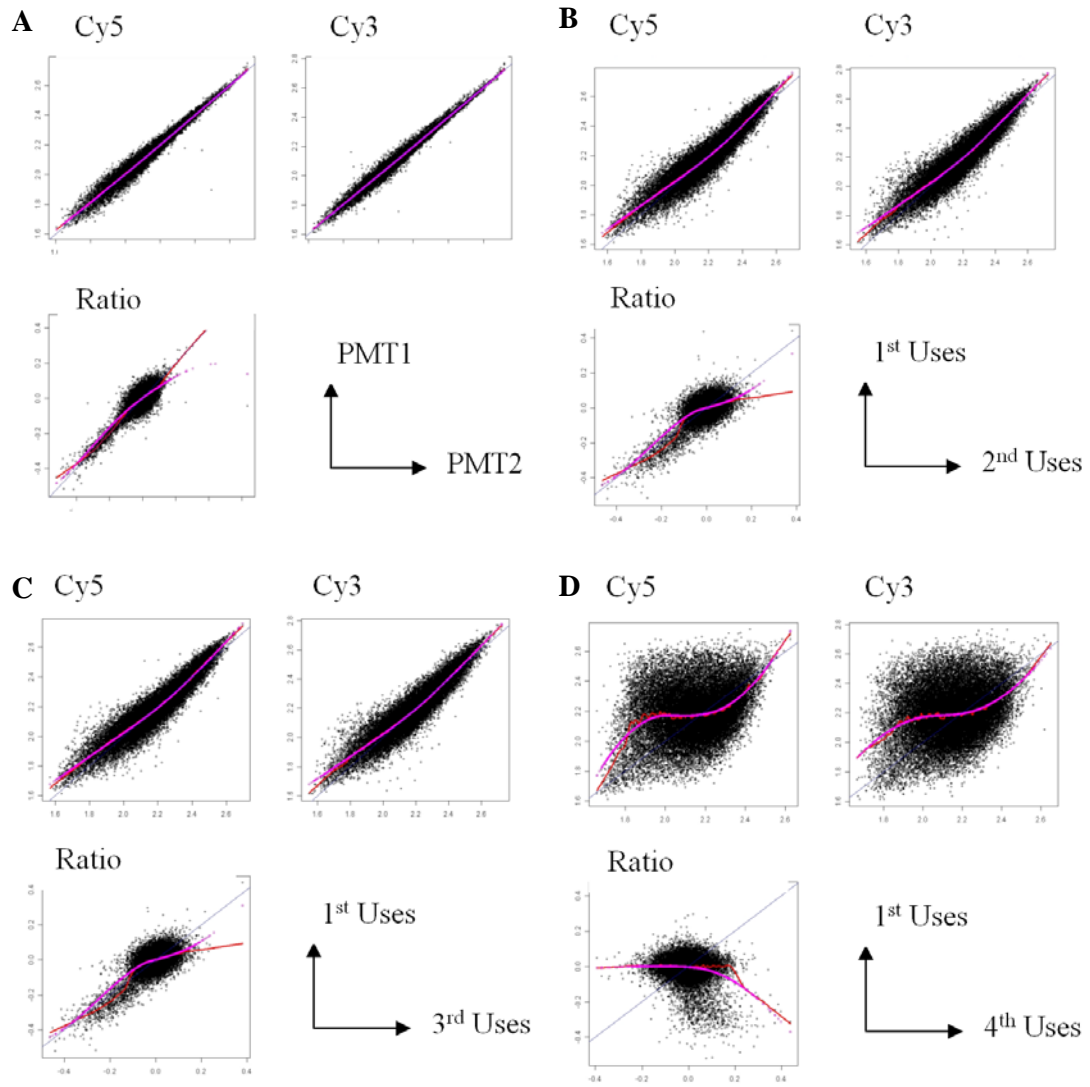


Figure 16 Scatter plots between the different reuses on stripped arrays show that same array can be reused up to 3 times

The panel shows the normalized data for Cy5 intensity, Cy3 intensity and the ratio of Cy5/Cy3 intensity. Normalization for the probes is based on normalized ratio = $\log(\text{ratio}) - \text{tukey.biweight}(\log(\text{ratio}))$ where $\log(\text{ratio}) = \log_2(R) - \log_2(G)$. The red line corresponds to local regression fitting by lowess method while the purple line corresponds to fitting smoothing curve by loess method.

2.5 Construction and quality control of HD2.1 Nimblegen chips for chromosome 21, 22 and additional regions

Nucleosomes play an important role in gene regulation by controlling the access of transcription factor to DNA. The presence of nucleosomes usually implies physical inhibition of transcription factor access. We designed two customised high density arrays which contain 2.1 millions probes to tile chromosome 21 & 22 and additional selected 206 regions with the aims of studying the dynamic changes of nucleosome locations before and after drug treatment. The 178 regions are each 200kbs long and selected from greatest confidence ER binding sites from ChIP-PET experiment conducted by Lin. In total there are approximately 4.2 millions unique probes. 3 Biological Replicates on different batches of cells have been performed on nucleosome samples with DM, E2 (2 treatments) on MCF-7 using the customized nucleosome arrays. In this design, a tightly constant overlapping 50bps probes were used for the purpose of defining the nucleosome positions.

The input regions were examined for the number of probes and coverage. Coverage here refers to the regions covered with probes. For example, two 50-mers probes overlapping 20bps will have coverage of 80bps. Given that there are large regions masked in chr21 and chr22, the percentage of coverage for chr21 and chr22 at 55.9% and 48.5% coverage are reasonably good. For the additional regions, the coverage was 77.3%.

Table 5 Coverage of probes in input regions show good coverage as the regions not covered are due to repetitive, low complexity regions



Since this is not an isothermal array, Figure 17 shows that only about 39% of probes have melting temperature between 73°C to 77°C. This is typical of many microarrays.

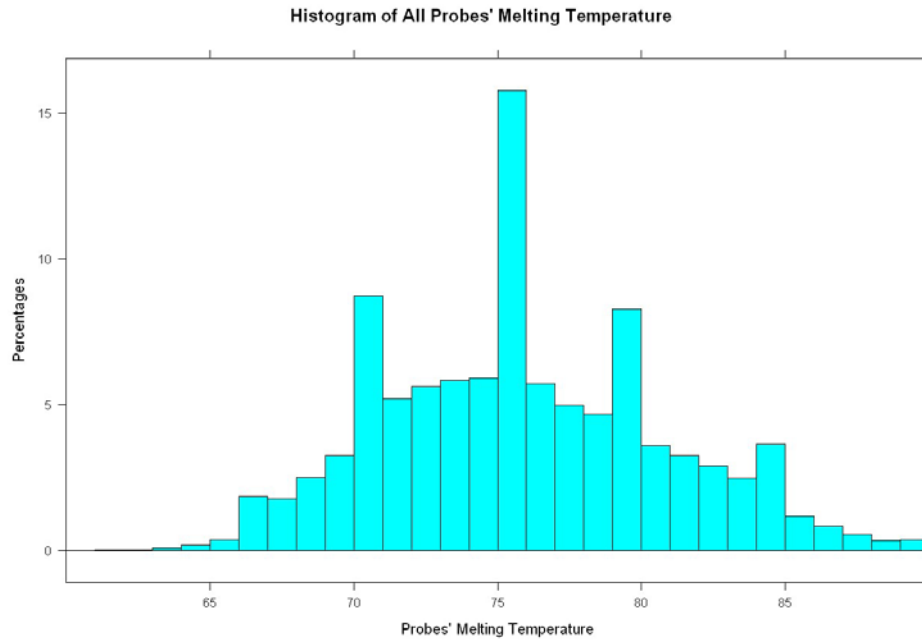


Figure 17 Histogram of All Probes' Melting Temperature for Nucleosome Array shows that about 39% of probes have melting temperature between 73°C to 77°C since this is not an isothermal array

The probe spacings on the nucleosome arrays are checked. Majority of the probes have spacing less than 100bps (Table 6).

Table 6 Majority of probes are less than 100bps spacings

	chr21:1-46944323	chr22:1-49554710	Additional Regions
Total Probes	1188230	1098358	2024881
Failed probes(Both Sides>100)	~2025	~2034	1618
Failed probes(Both Sides>150)			985
Failed probes(Both Sides>500)			136

The correlations between the 3 biological triplicates were assessed (Table 7). Since it is the biological replicates, the correlation between biological replicates in the same treatment was about 0.42~0.50.

Table 7 Correlation between arrays for Normalized Values shows that the correlation between biological replicates was about 0.42~0.50

Correlation	E1	E2	E3
E1	1.00		
E2	0.50	1.00	
E3	0.43	0.42	1.00

Correlation	D1	D2	D3
D1	1.00		
D2	0.50	1.00	
D3	0.47	0.47	1.00

E1 to E3 denotes the 3 biological triplicates in E2 treatment while D1 to D3 denotes the 3 biological triplicates in DMSO treatment.

With correlation of 0.49, i.e between E1 and E2 (Table 7), the scatter plot is shown in Figure 18.

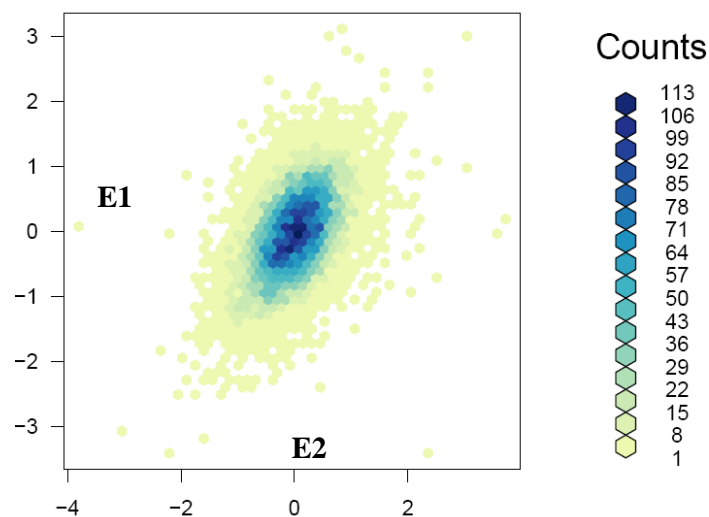


Figure 18 Scatter plot between ratio for E1 and E2 that has a correlation of 0.49
E1 to E3 denotes the 3 biological triplicates in E2 treatment

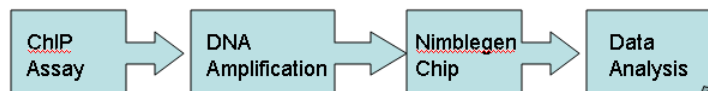
In conclusion, more than 40,000 mapped and putative ER binding sites have been incorporated into the customised array. We designed an isothermal array with

~380,000 probes that each region was tiled with at least 6 probes. The ER α specific binding site array followed all the important design considerations covered in this chapter. Briefly, variable-length probes were adopted to ensure similar hybridisation specificity that 70% of all probes have similar melting temperature between 73 to 77 °C. Various quality controls of the customised arrays were carried out, which included assessments based on probe properties, technical and biological reproducibility and reuses of stripped arrays. For examples, 77% of all probes have variances less than 0.3 within technical replicates; good data correlation up to the 3rd reuses of array and 97% of all probes have probe spacings less than 100bps. The above validations assessed the performance of the array and confirmed that the array is capable for querying the biology questions in this thesis.

With this customised ER chip, genome-wide binding sites profiles can be conveniently and comprehensively surveyed for different CHIP assays performed with virtually any antibodies, cell lines, drug treatments and concentrations in time course experiments. The array has broad applications in comparing the modulating effects of SERMs, the co-occupancy and interactions of various factors with ER, the chromatin states and histone modifications at those ER binding sites constructed on the array.

Chapter 3 Dynamics of Estrogen Receptor Binding in a Genome

Wide Scale



3.1 ChIP-chip analysis mapped 6482 ER binding sites

Chromatin Immunoprecipitation (ChIP) assay was performed on MCF-7 breast tumor cells subjected to different treatment conditions: Vehicle (DMSO), Estradiol (E2), Selective Estrogen Receptor Modulator (SERM) and SERM + E2. MCF-7 breast tumor cells were estrogen deprived for 3 days, followed by treatment for 45 minutes. Chromatin Immunoprecipitation (ChIP) assay is an *in vivo* method to obtain enriched DNA fragments bound by a particular transcription factor through cross-linking with formaldehyde and immunoprecipitated with a specific antibody. Formaldehyde cross-linking and sonication were carried out before immunoprecipitation was performed for ChIP DNA pull-down with ER α antibody (HC-20) or a non-specific GST antibody (Z-5). The cross-linking of proteins such as TFs to DNA allows the *in-vivo* studies on the interactions between TFs and DNA (Orlando, Strutt et al. 1997). TFs are recruited to binding sites which consist of both promoters and enhancers. Promoters are located within the proximity of TSS of genes whereas enhancers are located upstream, downstream or intragenic that is not in proximity to TSS. Promoters assemble the essential transcriptional machinery for initiation of transcription while enhancers are believed to enhance the assembly process by the looping mechanism which stabilises the preinitiation complex (Dion and Coulombe 2003). SERMs used in this study included 4-hydroxytamoxifen (T),

Raloxifene hydrochloride (R) and ICI 182,780 (I) and their specific agonist and antagonist properties are described in Table 8.

Table 8 Selection of SERMs

Index	SERMs Properties	agonist	antagonist	Company	Remark
1	ICI 182, 780 (I)	No	Yes	TOCRIS	No partial agonism
2	4-Hydroxytamoxifen (T)	Yes	Yes	SIGMA	More potent
3	Raloxifene hydrochloride (R)	Bone	Breast and uterine	SIGMA	Efficacy against estrogen-sensitive cancers

ICI is a pure antagonist while Tamoxifen exhibits partial agonist/antagonist properties. Raloxifene displays tissue-specific characteristics, i.e. as an agonist in bone but as an antagonist in breast and uterine. (Note: 4-hydroxytamoxifen is abbreviated to Tamoxifen, Raloxifene hydrochloride to Raloxifene and lastly ICI 182,780 to ICI in this report.)

To assess the quality of the experimental data, we plotted the average intensity for the biological triplicates along the locus for PTGES, GREB1 and IL6ST. Figure 19A and B shows that 2 well-known ER binding sites – GREB1 and PTGES have high ER occupancy. Lastly, Figure 19C shows IL6ST a non-binding site that has negligible occupancy, i.e. low intensity in E2 condition.

Subsequently, the corresponding changes in the enrichments for various SERMs on the above selected binding sites were examined in the heatmap (Figure 20). In general, SERMs attenuates ER recruitment, but ICI seems to have the strongest inhibition as shown in Figure 20A. For a non-binding site, the intensity for all treatments is very low and similar to one another as depicted in Figure 20B.

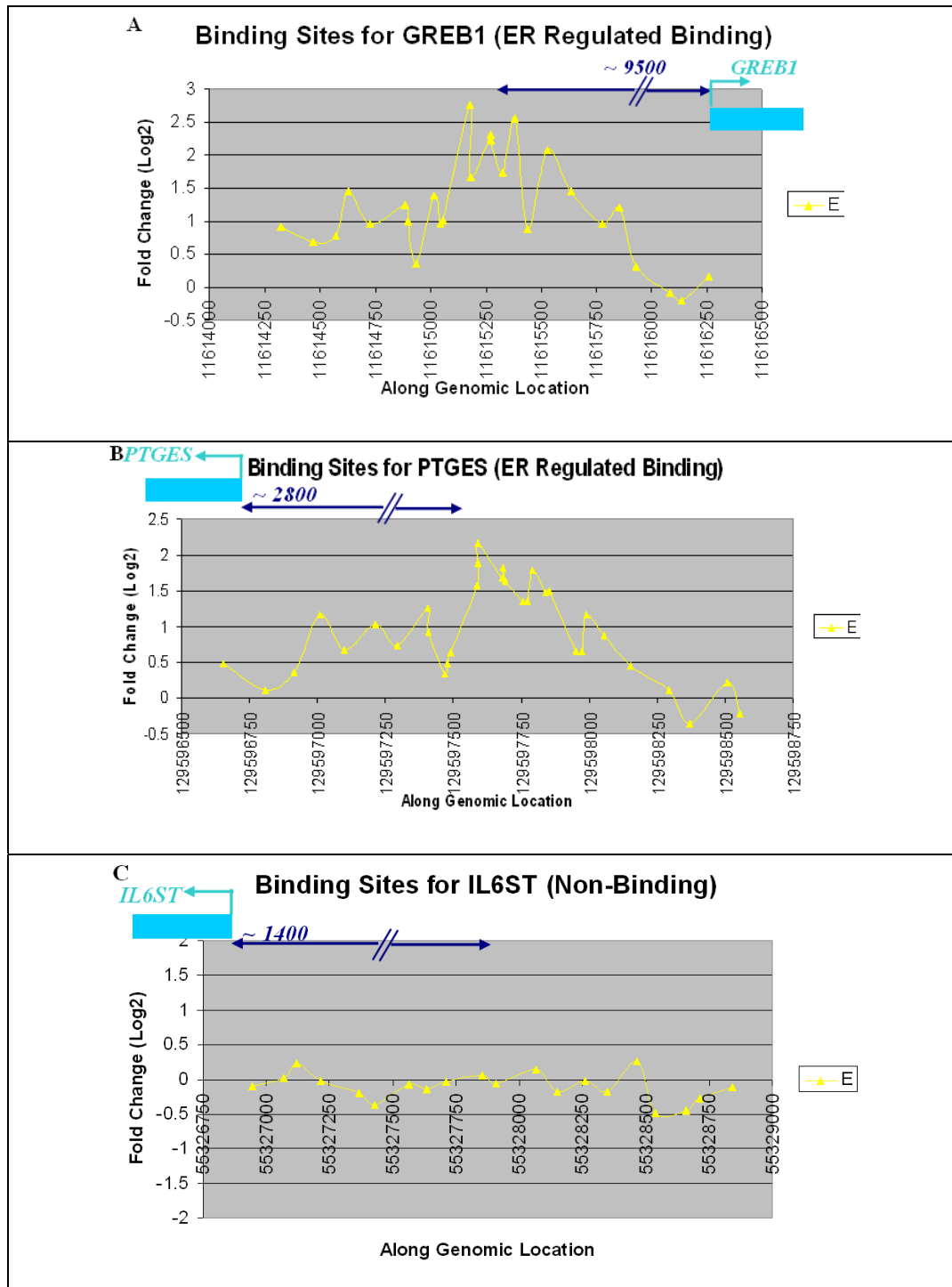


Figure 19 Assessment of experiment data using GREB1, PTGES and IL6ST

A) Binding site for GREB1 – ER binding site with basal occupancy; B) Binding site for PTGES – ER binding site without basal occupancy; C) Binding site for IL6ST – ER non-binding site. The average values represent the ratio of ChIP-ER/Input DNA and are derived from a minimum of two independent experiments. Concentration of E2 is 10nM. (E: E2 treated for 45 minutes)

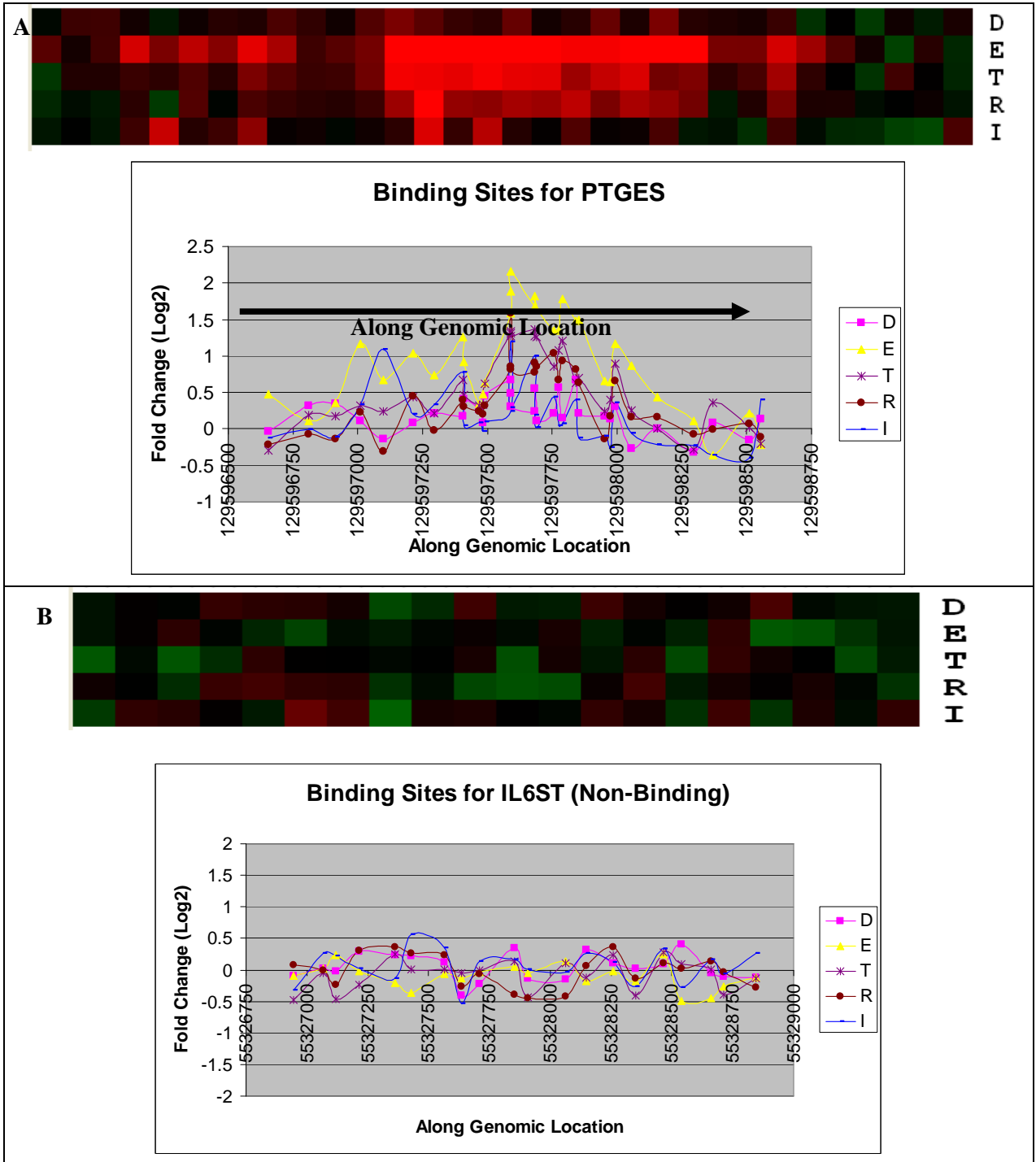


Figure 20 Binding site profile for PTGES and IL6ST under SERMs condition

A) Binding site for PTGES – ER binding site without basal occupancy under SERMs condition. Heatmap is also drawn to show the differential binding intensities since the plot is too clustered with lines; B) Binding site for IL6ST – ER non-binding site under SERMs condition. The average values represent the ratio of CHIP-ER/Input DNA and are derived from a minimum of two independent experiments. Concentration of E2 is 10nM while that of SERMs is 1 μ M. (E: E2 treated for 45 minutes; D: DMSO used a vehicle; T: Tamoxifen treated for 45 minutes; R: Raloxifene treated for 45 minutes; I: ICI treated for 45 minutes)

The above binding profiles for the known ER binding sites showed expected property of E2 and SERMs. This confirmed that the quality of experimental data may be good for surveying the rest of ER binding sites. The following paragraphs introduce a new algorithm to determine all the ER binding sites from the CHIP-chip data. Various validations were also carried out to assess the statistical significance and accuracy of the data. Lastly, comparisons were also made with the previously reported putative ER binding sites.

Results by Variable Factor Linear Model

To identify the dynamics of ER binding in genome-wide scale, a new peak-calling algorithm called *Variable Factors Linear Modeling (VFLM)* has been developed. All CHIP-on-chip analysis methods such as MAT, ChIPOTle and MACS obtain peaks from single hybridisation experiment. As they cannot handle replicates directly, we are forced to analyze the replicates independently and identify final peaks using standard non-parametric meta-analysis procedures. Alternatively, we can average the data from different arrays before peaks calling and apply one of the above peak calling algorithms. In both cases we are ignoring the multiple factors affecting the measurements which will reduce power and sensitivity to detect peaks. In contrast, the VFLM peak calling strategy performs linear modelling using all replicate data in one analysis. The advantage is that we can estimate the sources of variations that contribute to the errors in peak calling. Moreover, this approach provides estimate of variations from factors such as probe effects, batch effects of biological replicates and the reuse effect which the previous methods do not take into account. The importance of linear modelling can be demonstrated in the illustration 1A. If the experiments are carried out in two batches b1 and b2 for two different treatments E2 and DM. The top figure shows the distribution of data of a binding site for the two treatments and it can

be easily seen that the variations of measurements are large due to batch variations. Whereas the bottom figure shows the same measurements, but with the batch effects corrected, with much lower variations leading to potential increase in the power of the tests.

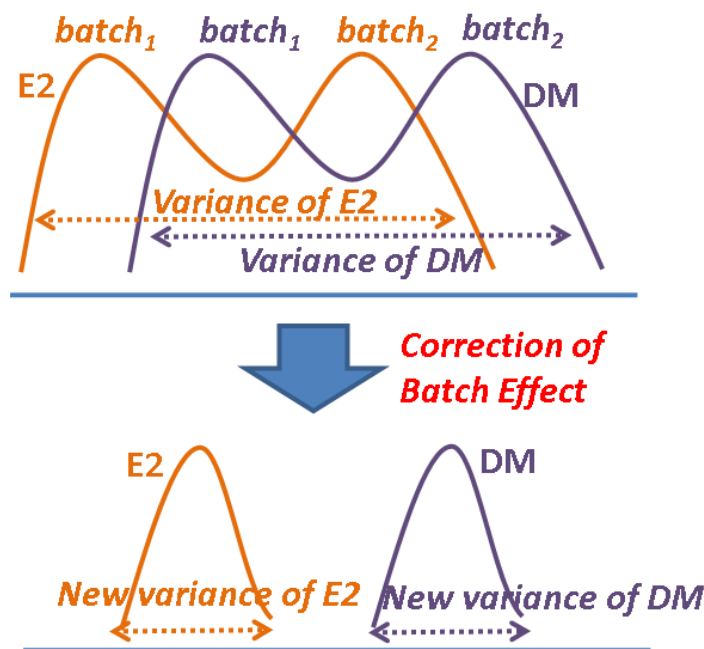


Illustration 1 Batch effect correction in VFLM

Another advantage of VFLM is its capability to measure directly the change in binding site intensity from one treatment to another by incorporating treatment as a factor in the linear model. This feature is not available in the existing peak calling methods.

The fundamental idea in VFLM, similar to the other peak calling algorithms, is to use sliding window approach. At each location (x), a linear model shown in Eq1 (for single treatment analysis) below is fit for all probes in the window on all replicates. P_i is i^{th} the probe in the window of size w at position x , b_i is the batch to which the replicate was made, R_i is the number of the reuse of the array of the replicate and e is the error not accounted by any of these factors. In our analysis the window size (w) was chosen to be 500bps and the window was slid with a step-size of 50bps.

$$y(x) = \sum_{i=x-w}^{i=x+w} P_i (\text{probe}) + \sum_{i=1}^{i=B} b_i (\text{batch}) + \sum_{i=1}^{i=r} R_i (\text{Reuse number}) + e (\text{error})$$

..... (Eq1)

Assumption for error follows $e \sim N(0, \sigma^2)$

Upon fitting model at a location x , binding enrichment is computed as the average of the co-efficients of all probes in the window $[x-w, x+w]$ i.e. enrichment is the average of probe effects.

$$FC_x = \frac{1}{2w + 1} \sum_{i=x-w}^{i=x+w} P_i$$

The p-value of the FCx (p_x) is obtained by combining the individual p-values of the probe effects (p_i) from the same $[x-w, x+w]$ using Fisher’s inverse Chi-Square method since they assume the same overall hypothesis, i.e.

$$p_x = \int_{S_x}^{\infty} X_{2n_x}^2(u) du; \quad \text{where} \quad S_x = \sum_{i=x-w}^{i=x+w} -2 \log(p_i)$$

Where n_x is the number of probes in the window $[x-w, x+w]$ and $X_{2n_x}^2$ is the chi-square distributios with $2n_x$ degrees of freedom.

VFLM calls peaks when the p-value falls below P_T and the fold changes are simultaneously above FC_T (Illustration 2).

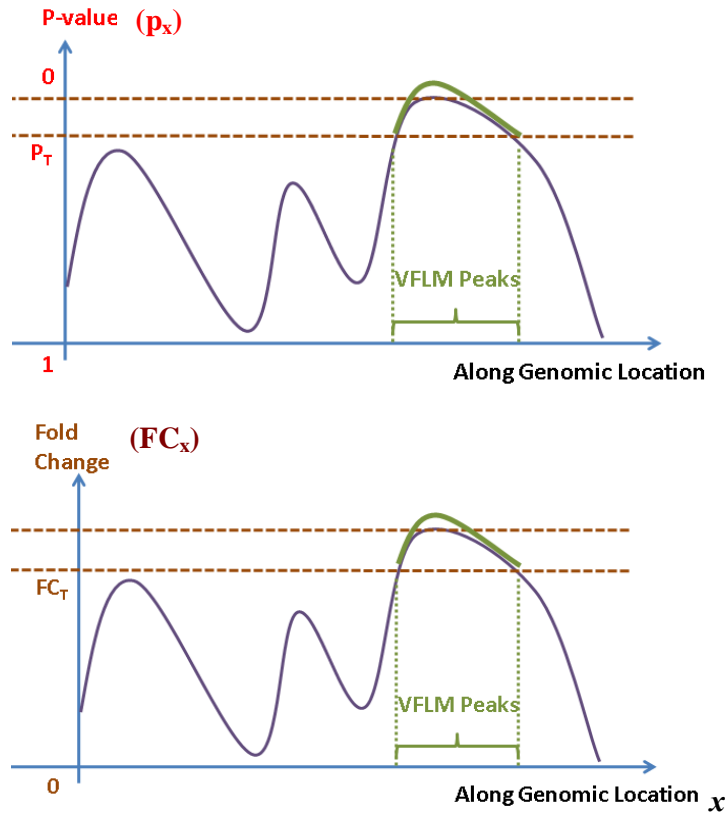


Illustration 2 Sliding window approach to determine VFLM peaks

Using VFLM, the number of binding sites for each treatment was obtained for the 4 categories below: 1) $P\text{-value} \leq 0.05$, Fold change ≥ 1.5 ; 2) $P\text{-value} \leq 0.05$, Fold change ≥ 2 ; 3) $P\text{-value} \leq 0.01$, Fold change ≥ 1.5 ; 4) $P\text{-value} \leq 0.01$, Fold change ≥ 2 as shown in Table 9.

Table 9 Results by Variable Factor Linear Model in MCF-7

MCF-7 (E2)	E2 Peak		DM Peak	
	$P \leq 0.01$	$P \leq 0.05$	$P \leq 0.01$	$P \leq 0.05$
1.5 Fold	6482	7165	4729	4926
2 Fold	4748	4966	2682	2706

Upon examining Table 9, the parameters of $P\text{-value} \leq 0.01$, Fold change ≥ 1.5 were chosen to determine the peaks in each treatment. The array platform has lesser sensitivity than real-time PCR system so a fold change of 1.5 was chosen instead of 2. Comparing the p-value of 0.05 and 0.01 at the fold change of 1.5, there was only a little difference in the number of peaks called, so p-value of 0.01 was used as the

peaks would have a higher statistical significance (Table 9). Hence, 6482 E2-peaks were detected. We proceeded to validate the VFLM results by comparing it to other known databases published by Lin and Carroll.

Validation against known databases

We evaluated our results by checking the reproducibility of mapped binding sites (Table 10). There was high reproducibility of Lin's and Carroll's high-confidence data at 79.5% and 80% respectively. True positive rate was assessed to be 83.6%. False positive rate was low as only 0.9% of the 1819 negative control regions were wrongly identified. The reproducibility of Lin's and Carroll's low-confidence data was reasonably good at 30% and 51.1% respectively. There were only 5.6% of the predicted EREs were truly binding. This might show that ERE motif presence alone is not sufficient to accurately predict a true binding.

Table 10 Distribution of Peaks Detected in Current Studies in Different Categories of Input Regions

Input Regions	Positive Regions						Negative Regions	
	Binding (55)	Lin High-confidence (1234)	Lin Low-confidence (7574) Include high-confidence set	Carroll High-confidence 1E-5 (3665)	Carroll Low-confidence 1E-3 (10599) Include high-confidence set	pERE (30832 ERE Predictions)	Non-Binding (68)	Negative Controls (1819)
Fold-Change=1.5; P-values=0.05	46 (83.6%)	953 (77.2%)	2273 (30%)	3114 (85%)	5413 (51.1%)	1716 (5.6%)	5 (7.4%)	17 (0.9%)

Since there were overlaps among the 3 major groups of binding sites (Lin, Carroll and pERE) besides the additional binding sites from literature references, it would be interesting to study their reproducibility. The overlaps among Lin, Carroll and pERE are shown in Figure 21A while the overlaps among pERE, high-confidence binding sites of Lin and high-confidence binding sites of Carroll are shown in Figure 21B. Interestingly, detected 6482 peaks overlap better with Carroll's. Referring to Figure 21A, it can be observed that the overlapped binding sites between Lin and Carroll

without ERE motif have high reproducibility of 89% but those with ERE motif are higher (94%). However, Lin's and Carroll's high-confidence data with and without ERE motif are similarly highly reproducible at 98% (Figure 21B). Those unique Lin or Carroll binding sites with ERE motif are more reproducible than those without ERE. In general, those highly reproducible ER binding sites contain ERE motif.

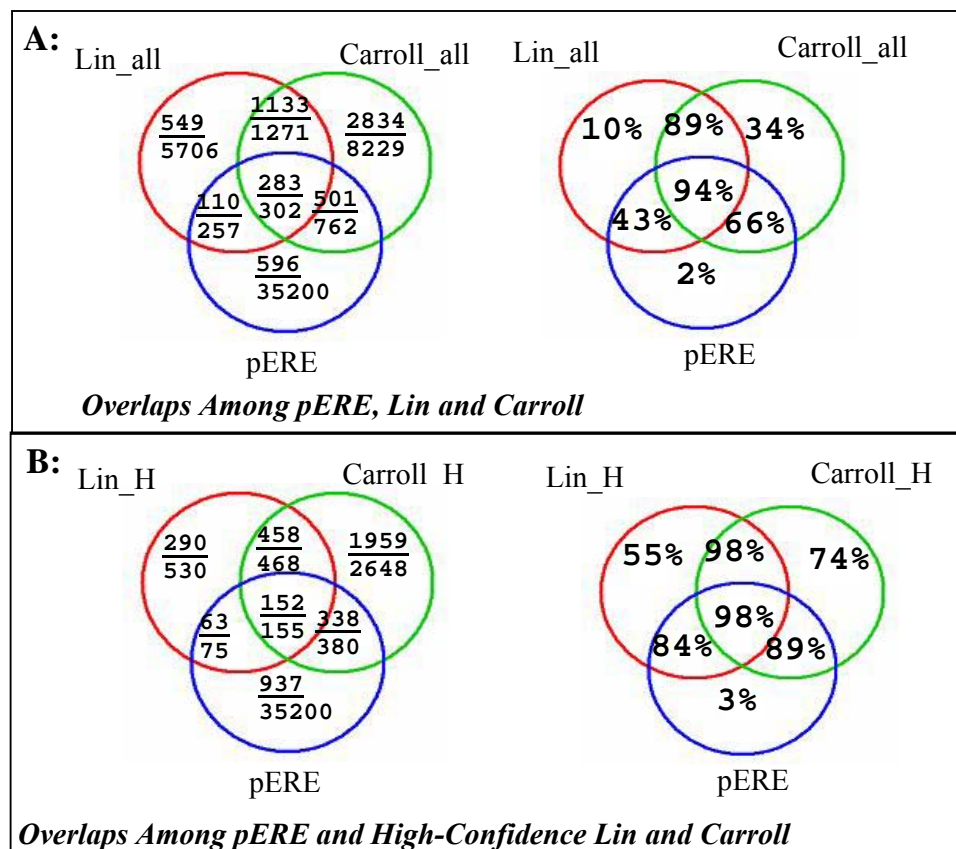


Figure 21 Distribution of Peaks Detected in Overlapped Input Regions. Highest percentages of peaks found to be common binding sites with ERE motif

The 1st Venn diagram shows the peaks found in this thesis as the numerator while the reported ER binding sites as the denominator in each category and their overlaps. Lastly, the percentage of peaks found over the probed binding sites regions were shown in the 2nd Venn diagram. Lin_all and Carroll_all denote the low-confidence binding sites for Lin and Carroll respectively while Lin_H and Carroll_H denote the high-confidence binding sites.

VFLM results were shown to be in good agreement with databases published by Lin and Carroll. Here we investigated in greater detail on the overlap between VFLM

results and different categories of Lin's data. It has distinct categories of varying confidence binding site sets. Figure 22 shows the percentage of peaks detected in different categories of Lin's data. As could be seen, the percentage of detected peaks increased with higher PET number. Higher PET numbers indicated greater confidence in the mapped binding sites. This shows that the accuracy of the detected binding sites increases with greater confidence in binding sites.

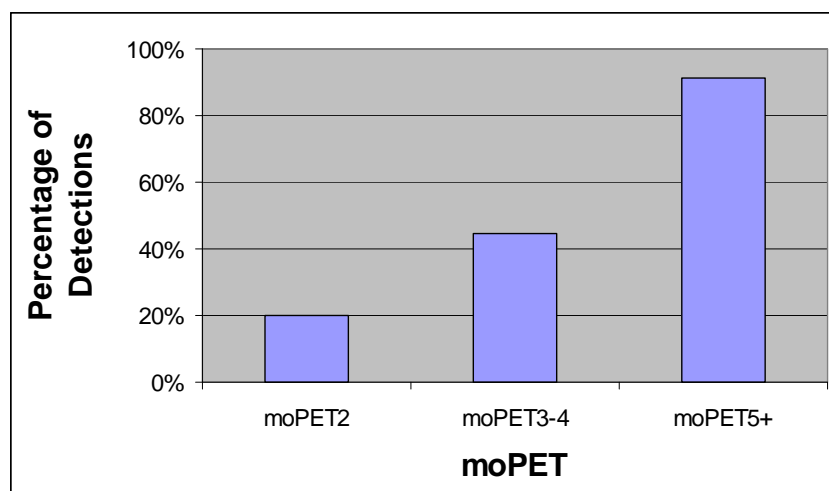


Figure 22 Percentage of detected peaks increases with higher PET numbers of Lin category

We further examined the distribution of the VFLM peaks for both the detected and missed 1234 high-confidence Lin data. For the detected 953 binding sites, the percentage of detection increases from 63.9% to 93.7% with increasing moPET numbers (Table 11). On the other hand, the distribution of the 281 missed sites showed that the percentage of missed sites decreases with increasing moPET numbers from 35.9% to 5.9%. 198 out of 281 missed sites (70.5%) belong to the moPET3 category while 93.7% (284/303) are detected in the moPET category ≥ 6 . Detected binding sites overlap substantially with higher moPET categories of 1234 binding sites (Lin, Vega et al. 2007). This shows that higher moPETs indicates higher confidence and associates with higher reproducibility.

Table 11 Distribution of detected 953 sites and 281 missed sites in high-confidence Lin (1234)

<i>moPET</i>	<i>All 1234 Sites</i>		<i>953 Detected Sites</i>		<i>281 Missed Sites</i>	
	<i>No. Sites</i>	<i>No. Sites</i>	<i>Percentage (1234)</i>	<i>No. Sites</i>	<i>Percentage (1234)</i>	
3	552	353	63.9%	198	35.9%	
4	245	198	80.8%	48	19.6%	
5	134	118	88.1%	17	12.7%	
>=6	303	284	93.7%	18	5.9%	

In summary, VFLM results were validated in databases published by Lin and Carroll with greater than 80% reproducibility for the high-confidence data. The next validation was carried out to compare the VFLM fold enrichments to the qPCR fold change.

Correlation of Real-time PCR results and E2-ER ChIP-on-chip binding sites

When correlating qPCR fold change of 77 sites obtained from Lin, C. Y., V. B. Vega, et al. (2007) to our ChIP-on-chip peaks, correlation of 0.54 and p-value of 2.359e-06 (Pearson's product-moment correlation) were obtained. The reason for the moderate correlation could be the ChIP-on-chip signals will be lower as there may not have probes exactly at the peak location of the targeted binding sites.

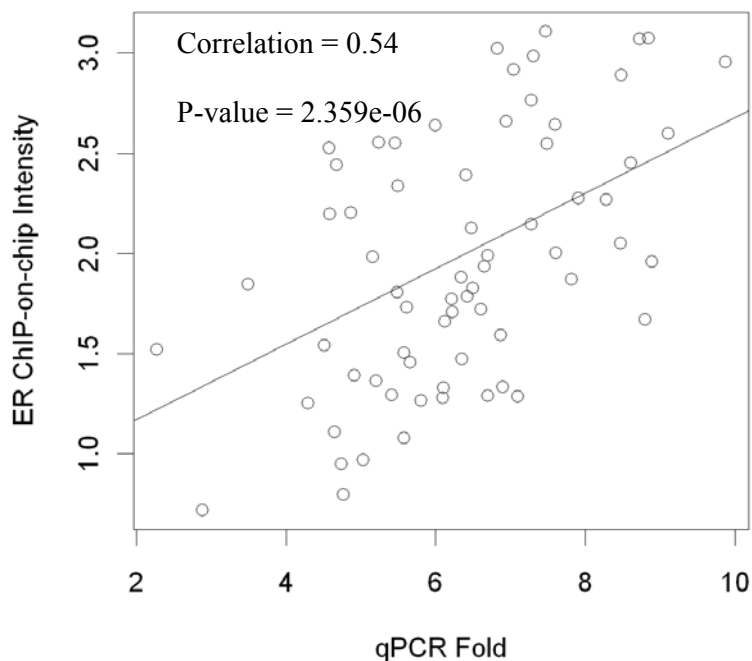


Figure 23 Scatter plot on ER ChIP-on-chip intensity and qPCR

We further examined the correlation between qPCR fold change and percentage of detection (Figure 24). The detection in the binding sites increased from 43% to 100% with increasing magnitude of qPCR fold change from 2-24 to 175-551. The biggest jump was from 2-24 to 25-49, i.e. 43% to 81%. Fold change of $\geq 25-49$ was guaranteed to be detected at the rate of $\geq 80\%$.

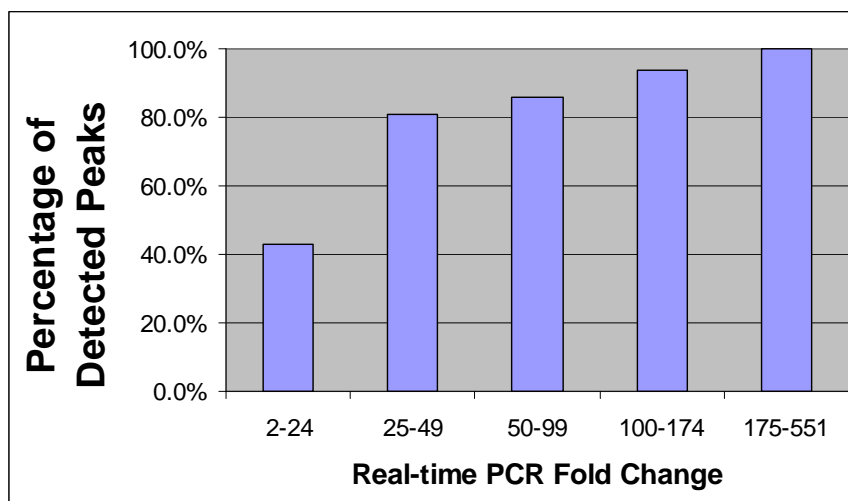


Figure 24 Detected 77 Peaks and Real-time Fold Change

In general, VFLM fold enrichments correlate well with qPCR fold change. Specifically, there was a very high percentage of detection (80%) in the array when the qPCR fold change was greater than 24.

Good agreement between VFLM peaks and ChIP-Seq results

Upon assessing VFLM results overlapped greater than 80% with known Lin's and Carroll's data and correlated well with qPCR fold change, we next compared our VFLM peaks with ChIP-Seq results (Unpublished data). ChIP-Seq is a most recent unbiased and precise approach to identify TFBS on a genome scale. This approach couples chromatin immunoprecipitation (ChIP) with the paired-end ditag (PET) sequencing strategy and the Illumina single read sequencing technology. The ChIP samples for ChIP-Seq experiments have the same experimental conditions as the ChIP-chip experiments. VFLM peaks showed good agreement – 80.0% and 53.6% of VFLM peaks overlap with all 17,418 and top 6500 ChIP-Seq peaks respectively (Figure 25). The 80% overlap was remarkably good as VFLM peaks were obtained from a focused array on ER binding sites and not an unbiased whole genome array.

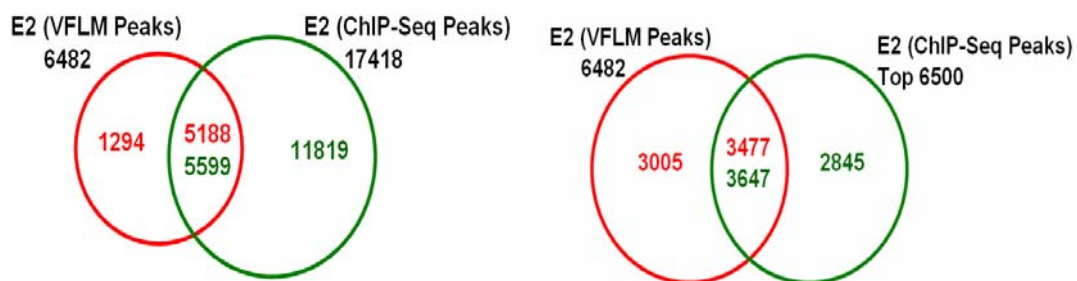


Figure 25 Good Overlap between VFLM Peaks and ChIP-Seq Results

Table 12 compares the reproducibility of various published databases in VFLM set and top 6500 ChIPSeq set. VFLM peaks have higher percentage of coverage in Lin and Carroll than the ChIP-Seq data, i.e. 12.8% more coverage in high-confidence Lin and 16.8% more coverage in high-confidence ChIP-Seq data.

Table 12 Table comparing VFLM peaks and ChIP-Seq data. VFLM peaks have higher percentage of coverages in Lin and Carroll than the ChIP-Seq data

Input Regions	Positive Regions						Negative Regions	
	Binding (55)	Lin High-confidence (1234)	Lin Low-confidence (7574) Include high-confidence set	Carroll High-confidence 1E-5 (3665)	Carroll Low-confidence 1E-3 (10599) Include high-confidence set	pERE (30832 ERE Predictions)	Non-Binding (68)	Negative Controls (1819)
VFLM Peaks (6482)	46 (83.6%)	953 (77.2%)	2273 (30%)	3114 (85%)	5413 (51.1%)	1716 (5.6%)	5 (7.4%)	17 (0.9%)
ChIP-Seq (6500)	46 (83.6%)	823 (66.7%)	1551 (20.5%)	2500 (68.2%)	3934 (37.1%)	1249 (4.1%)	5 (7.4%)	2 (0.1%)

In summary, VFLM peaks overlapped remarkably well with ChIP-Seq data at 56% and VFLM peaks had higher percentage of coverages in Lin and Carroll than the ChIP-Seq data.

Assessing the biological replicates

Lastly, we assessed the robustness of the pipeline (ChIP-chip-VFLM) by computing the peaks using leave-one-out strategy. As can be seen in Figure 26, TE peaks are the results from VFLM using all 3 biological replicates while TE1, TE2 and TE3 Peaks are the results from using any 2 biological replicates leaving out 1st, 2nd and 3rd replicate respectively. The numbers of peaks found were in general comparable to TE across TE1 to TE3. The standard deviation and mean for the TE1, TE2 and TE3 were also reasonably good.

VFLM Peaks P-value=0.01; Fold change ≥1.5 Number of Binding Sites	TE Peaks	TE1 Peaks	TE2 Peaks	TE3 Peaks	Mean for TE1 to TE3	Std. Deviation for TE1 to TE3
	6078	6219	5091	4817	5376	743

Figure 26 Linear model for MCF-7 (Takes any 2 out of 3)

Overlap study was performed between TE1, TE2 and TE3 peaks using the parameters Fold change ≥ 1.5 and p-value ≤ 0.01 . The percentage of overlaps varied from 89% to 93% with respect to the peaks of smaller size (Table 13).

Table 13 *Overlap between TE1, TE2 and TE3 peaks*

Overlap Between Peaks	TE1 (6219)	TE2 (5091)
TE2 (5091)	4593 (90.2%)	
TE3 (4817)	4459 (92.6%)	4274 (88.7%)

In summary, the standard deviation and mean for the TE1, TE2 and TE3 were reasonably good. The leave-one-out strategy showed good robustness of the pipeline (ChIP-chip-VFLM) as the percentage of overlaps were at least 88.7% between two leave-one-out computations.

3.2 Most ER binding sites are pre-occupied by ER and they have a greater ER recruitment upon E2 treatment

For years, researchers have been intrigued by the mechanism by which estrogen receptor selects and utilises its binding sites. In the past, most of the analysis on estrogen receptor binding concentrated on the ratio of Estradiol (E2) over DMSO treatment. Interestingly, the analysis on the individual treatments revealed a number of new aspects in the mechanism of estrogen receptor binding. Firstly, we observed that most ER α binding sites are occupied by ER α even before estradiol treatment. We detected 4729 sites in DMSO treated cells. The binding sites were obtained for each drug treatment through VFLM that required at least 1.5 fold enrichment at p-value \leq 0.01. In E2 treatment, there were 6482 binding sites. 4205 or 65% of which had a preferential binding to locations with basal occupancy, i.e. 90% of DMSO occupied remained to be bound by ER α . (Figure 27)

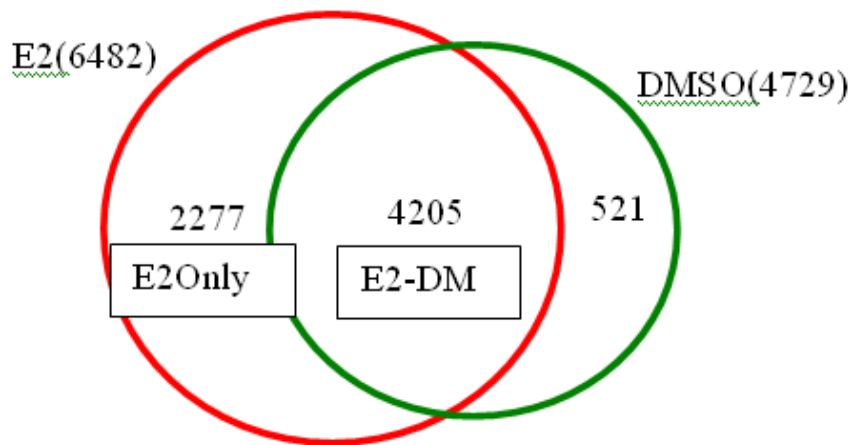


Figure 27 *Overlap between binding sites found in E2 and DMSO treatment shows 65% of all E2 binding sites are pre-occupied*

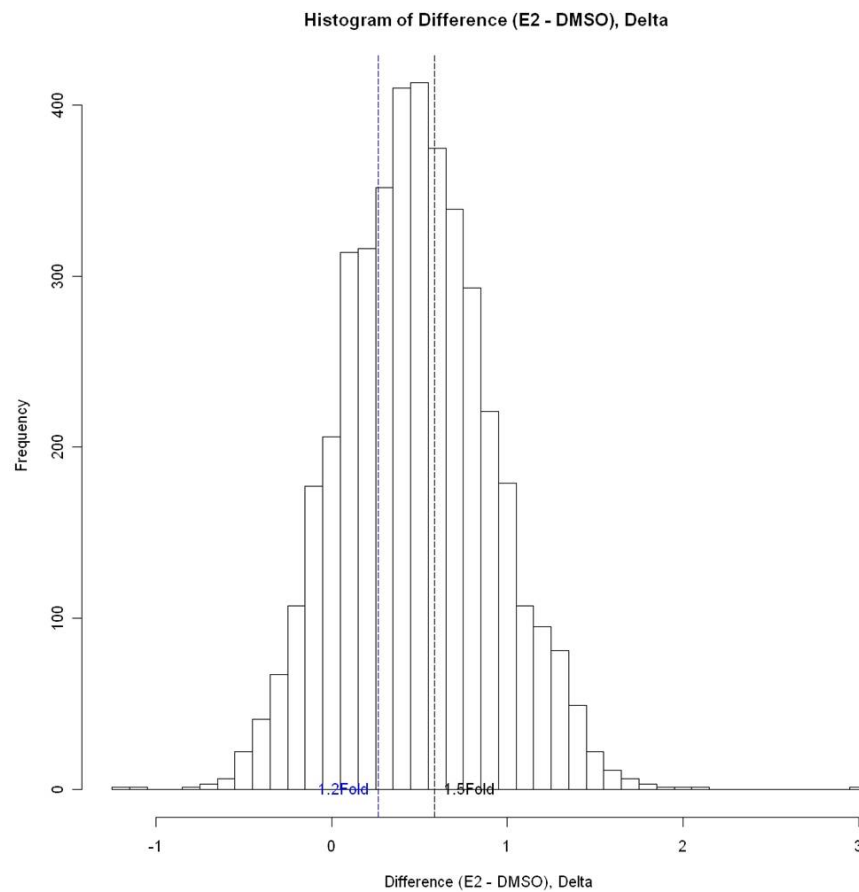


Figure 28 *Histogram of Difference (E2 – DMSO)*

Figure 28 shows the histogram of the difference (E2 – DMSO). 2 cut-offs of 1.2 folds and 1.5 folds are shown. With 1.2 fold cut-off, more than 50% of the binding

sites with fold change increases upon drug treatment were greater than 1.2 fold changes.

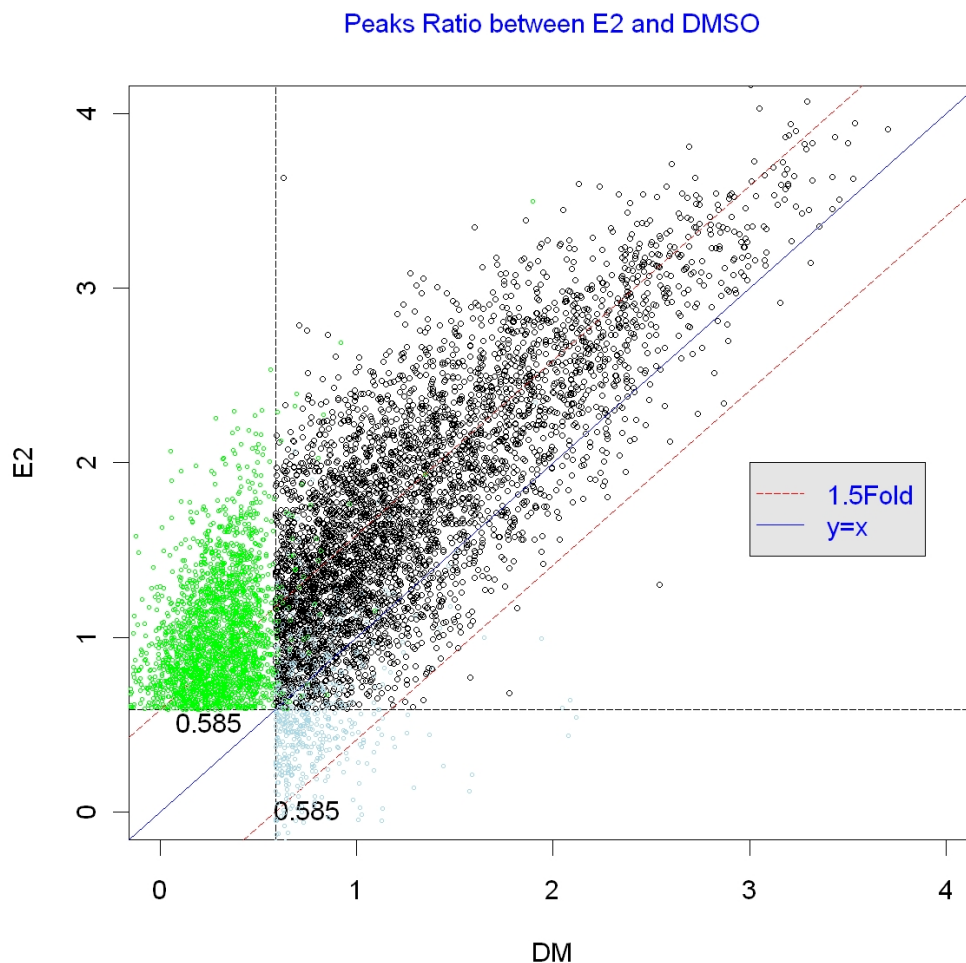


Figure 29 Scatter plot between E2 and DMSO treatment

The black-coloured points denote enrichments greater than 1.5 fold in both E2 and DMSO conditions; green-coloured denotes E2 greater than 1.5 fold while blue-coloured points denote DMSO greater than 1.5 fold.

Figure 30A showed a well-known ER binding site – GREB1. It was seen that the average intensities were quite high even in DMSO condition. This could be an example of binding site with high basal occupancy. On the other hand, Figure 30B shows a binding site of PTGES that has negligible basal occupancy, i.e. low intensity in DMSO condition but the intensity becomes high upon E2 treatment.

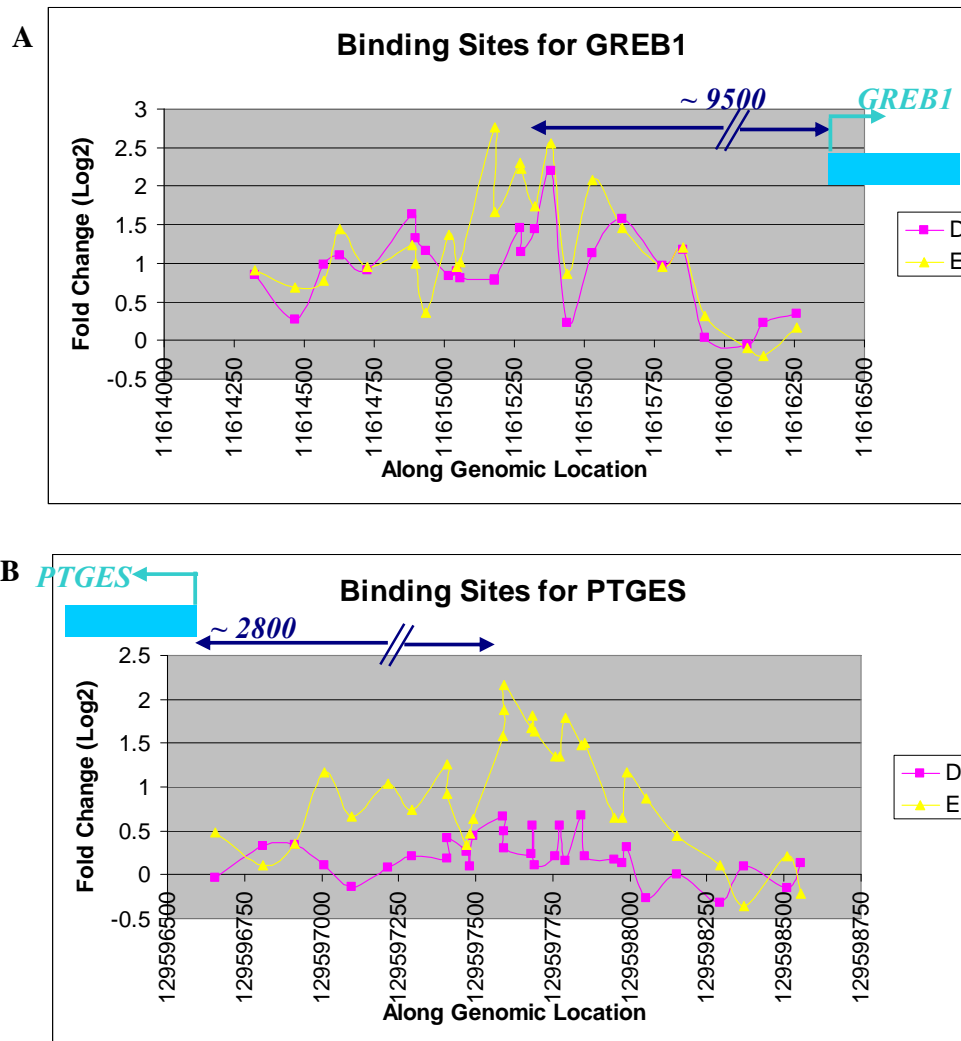


Figure 30 *GREB1* has basal occupancy while *PTGES* has no basal occupancy

A) Binding site for *GREB1* – ER binding site with basal occupancy; B) Binding site for *PTGES* – ER binding site without basal occupancy. The average values represent the ratio of ChIP-ER/Input DNA and are derived from a minimum of two independent experiments. Concentration of E2 is 10nM. (E: E2 treated for 45 minutes; D: DMSO used a vehicle)

Pre-occupied estrogen receptor binding sites have a greater ER occupancy upon E2 treatment

We would like to investigate whether pre-occupied estrogen receptor binding sites have a greater ER occupancy upon E2 treatment. We plotted the box plots for the E2 and DM ratios in the two categories of (1) Pre-occupied (E2_DM) and (2) not pre-occupied or new binding sites (E2Only). Interestingly, Figure 31 shows that both E2

and DM fold enrichments (log scale) showed a greater interquartile range of 1.01 and 0.80 respectively in the ER pre-occupied binding sites compared to the non-occupied binding sites of 0.46 and 0.27 for E2 and DM respectively. This implied that pre-occupied binding sites had higher fold enrichments. Interestingly, E2DM_DM had a higher ER enrichment than E2Only_E2, contrary to the notion that binding sites enrichments were typically higher in E2 treatment.

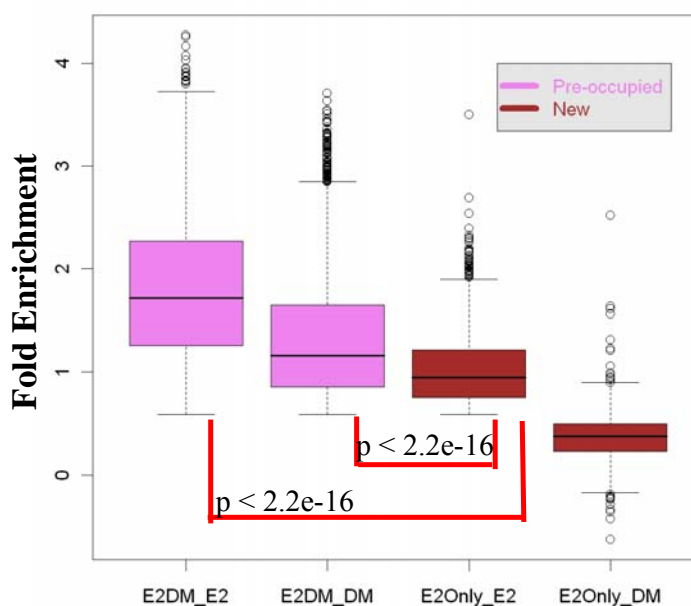


Figure 31 Box plots between categories for E2_DM and E2_only

Both the T-test (Significance for the difference in 2 means) and F test (Significance for the difference in 2 variances) between E2DM_E2 and E2Only_E2 shows p -value $< 2.2e-16$. Similarly for E2DM_DM and E2Only_E2.

In summary, 65% of all E2 binding sites were pre-occupied and those pre-occupied sites had higher fold enrichment.

3.3 SERMs impose small changes to ER binding locations but greatly reduce ER binding affinity

Although SERMs have been successfully used to treat breast cancer patients, the exact mechanisms by which SERMs modulate estrogen receptor binding are not well understood. Here we examined how SERMs affect the ER binding locations and the

corresponding binding intensity. Since the customized array is tailored towards all literature reported ER binding sites and is not a true whole-genome array, we are capturing the changes made by SERMs and counting the peaks of SERMs on the ER binding sites available on the array. Figure 32 shows the peaks detected in E2, SERMs alone and SERMs + E2. The number of peaks detected among the different treatments was fairly similar, ranges from around 5000 to 6000 peaks.

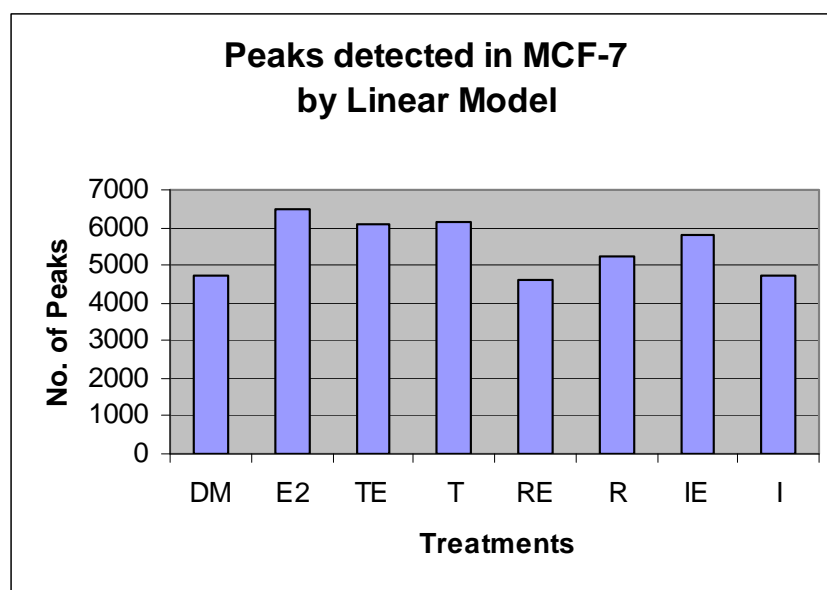


Figure 32 Peaks detected in MCF-7 by Linear Model

T, R and I indicate Tamoxifene, Raloxifene and ICI respectively while TE, RE and IE were combination of E2 and SERMs.

About 80% of the binding sites in SERMs overlapped with those from E2 treatment as depicted in Figure 33 while between 60~75% of binding sites in E2 remained intact with those from SERMs as indicated in Figure 34. The results showed that E2 or SERMs-liganded estrogen receptors still utilized very similar sites, and binding sites induced by SERMs were also very specific and similar to those induced by E2.

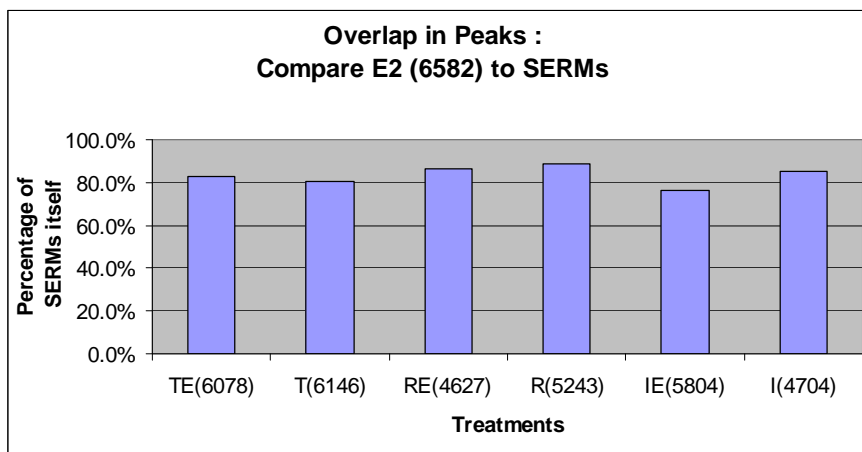


Figure 33 *Overlap in peaks between E2 and SERMs (Percentage of SERMs)*

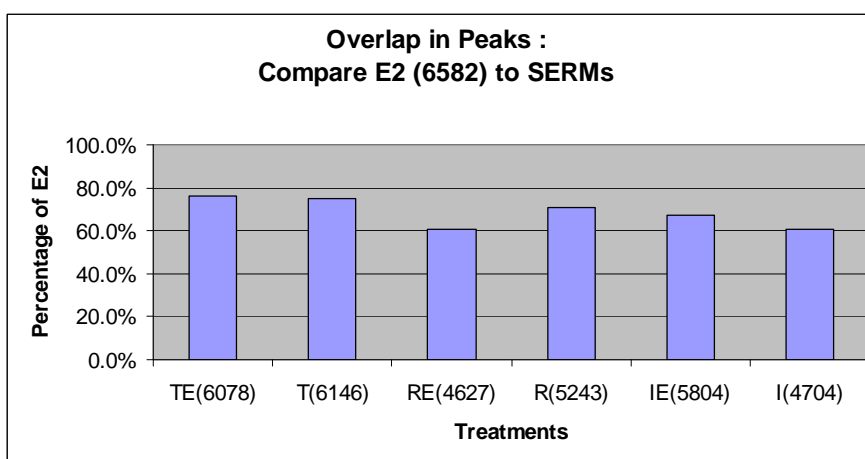


Figure 34 *Overlap in peaks between E2 and SERMs (Percentage of E2)*

Tamoxifen, Raloxifene and ICI shared a high percentage of binding sites with E2. Though ER under these different treatments occupied the same binding sites, the intensities were different and mostly attenuated under SERMs condition, compared to those of E2. Figure 35 shows that there are unique sites to Tamoxifen and Raloxifene. It was interesting that E2, Tamoxifen and Raloxifene shared 4427 common binding sites. This indicated the specificity of SERMs since they affected largely ER binding sites.

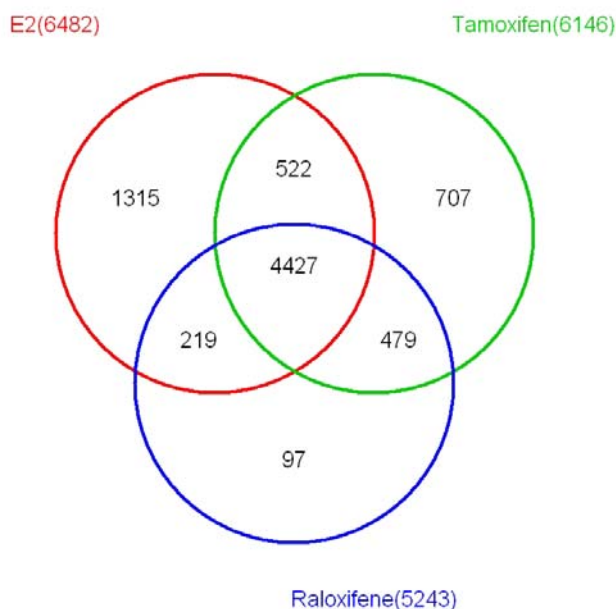


Figure 35 *Overlap between E2, Tamoxifen and Raloxifene peaks*

After we found that SERMs imposed small changes to ER binding location, we next asked whether the changes to ER binding recruitment were affected by SERMs. We obtained the peaks that did not just increase upon drug treatments but had a difference of greater than 1.5 fold changes and $p\text{-value} \leq 0.01$ with reference from DMSO (control). The differential enrichment (Treatment – DMSO) is given by the coefficient of T in the VFLM model as indicated in Eq2.

$$y(x) = T + \sum_{i=x-w}^{i=x+w} P_i (\text{probe}) + \sum_{i=1}^{i=B} b_i (\text{batch}) + \sum_{i=1}^{i=r} R_i (\text{Reuse number}) + e (\text{error})$$

Eq2 *Differential VFLM equation for finding peaks for the difference*

T denotes the differential fold change due to the different treatments. Peak fold changes are computed from average of probe effects. P-values are obtained from fisher inverse Chi-Square of p-values of probe effects.

The highest number of binding sites with significant difference from DMSO was seen in E2 treatment and number of different binding sites in SERMs from Tamoxifen to ICI far less than that of E2 (Figure 36).

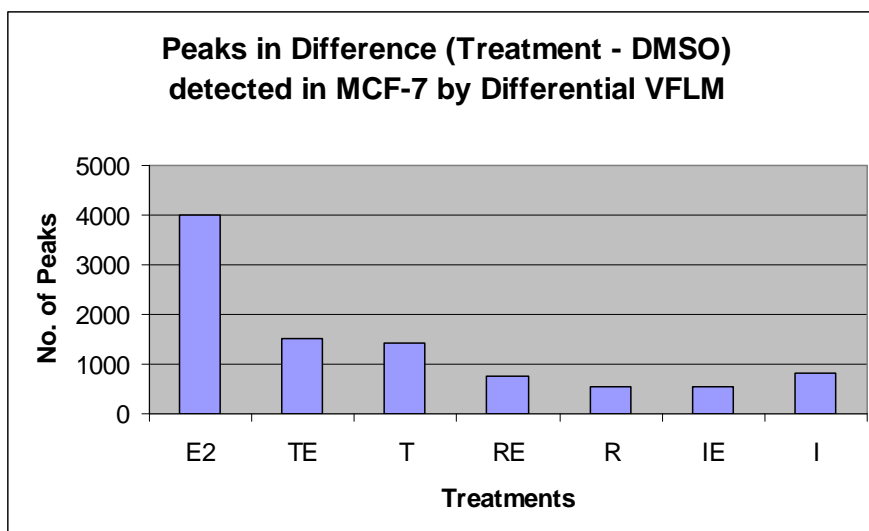


Figure 36 Peaks in Difference (Treatment – DMSO) detected in MCF-7 by Linear Model

The overlap in differential enrichment (Treatment – DMSO) was performed between E2 and SERMs. The overlaps were plotted in terms of percentage of E2 (Figure 37) and percentage of SERM (Figure 38). The overlap of E2 was lower than 20% while the overlap in SERMs was between 20 to 50%. The 20% overlap in percentage of E2 suggested that although the estrogen receptor occupied similar binding sites in both E2 and DMSO treatments, those sites were induced with different intensity. E2 treatment had the strongest binding sites intensity.

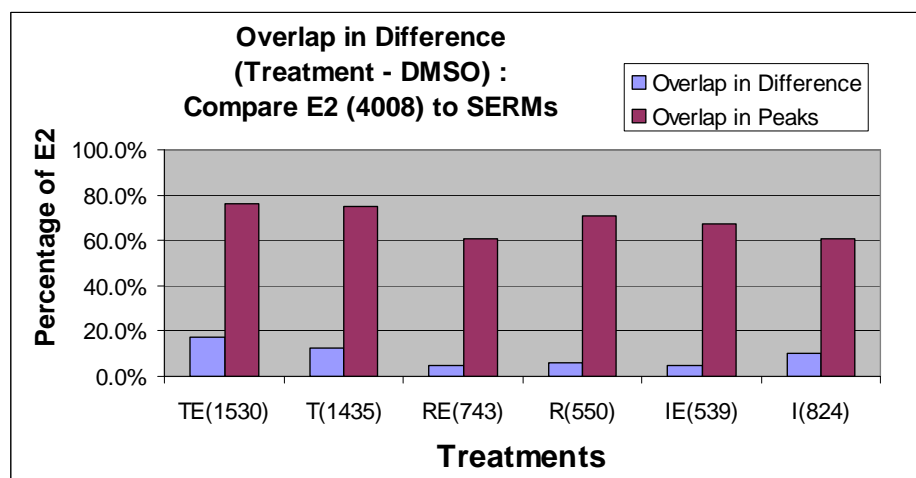


Figure 37 Overlap in difference (treatment – DMSO) between E2 and SERMs (Percentage of E2)

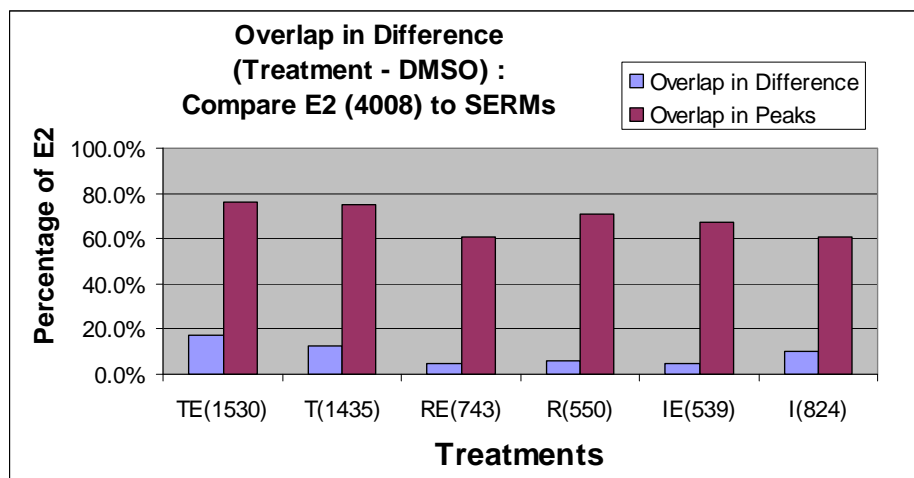


Figure 38 Overlap in difference (treatment – DMSO) between E2 and SERMs (Percentage of SERM)

To further confirm that the intensities for E2 binding are the strongest, we plotted the distribution for E2, T, R and I binding sites in Figure 39 and for E2, TE, RE and IE binding sites in Figure 40. In Figure 39, it can be seen that the order of magnitude in decreasing manner is E2, T, R and I. Tamoxifen tends to induce ER binding similar to E2 whereas ICI tends to inhibit the action of E2 the strongest. We observed the same descending order of E2, TE, RE and IE in Figure 40.

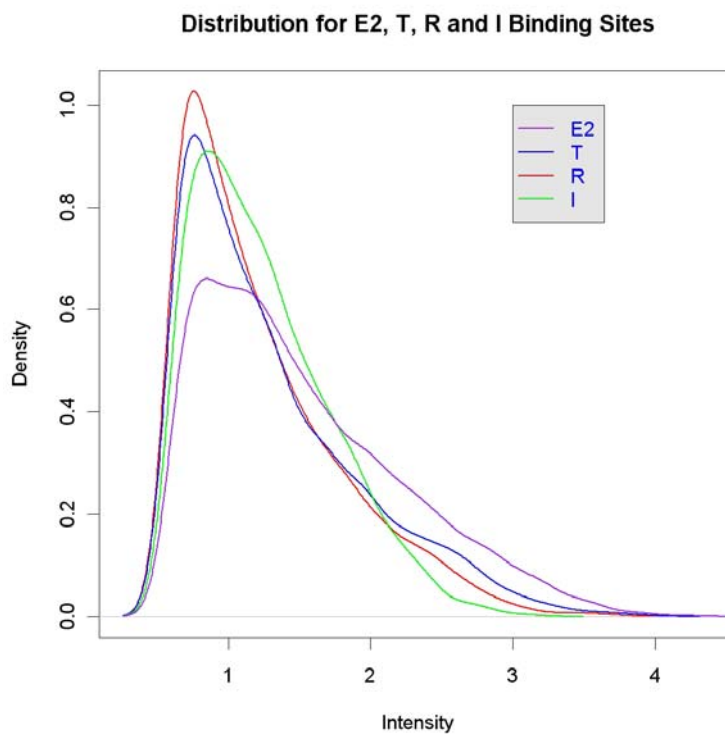


Figure 39 Distribution of ratio intensities for E2, T, R and I Binding Sites

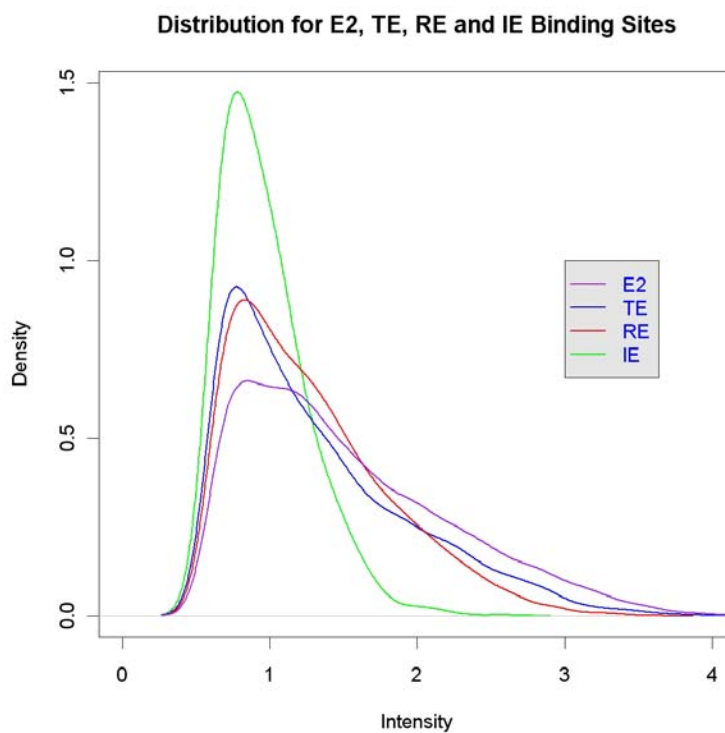


Figure 40 Distribution of ratio intensities for E2, TE, RE and IE Binding Sites

In summary, SERMs and E2 shared about 80% common binding sites but the intensities were much lower for SERMs.

3.4 ER-SERMs utilize tethering mechanism much more than ER-E2 and de novo motif predictions indicate shifting of preferential binding motif

Since ER binding sites are expected to predominantly contain full ERE sequence, we assessed the distribution of full ERE, half ERE and no ERE in the binding sites found in each treatment. As shown in Table 14, all treatments show a high percentage (60%) of binding sites carrying the full ERE. This percentage was significantly higher than the 4.1% of a randomly generated 100,000 DNA sites. Random sites showed 42.7% of fragments having half ERE which indicated the abundance of half ERE present in genome. The percentages of binding sites with no ERE for all treatments range from 5~8 % (see Table 14), which were significantly lower than the background of 53.2%.

Table 14 Distribution of full ERE, half ERE and no ERE

	E2	TE	T	RE	R	I	Random
Total BS	6482	6086	6146	4627	5243	4704	100000
BS with Full ERE	4232	3833	3852	2787	3285	2657	4100
BS with Half ERE	1938	1856	1892	1463	1585	1731	42700
BS with no ERE	312	397	402	377	373	316	53200
BS with Full ERE (Percentage)	65.3%	63.0%	62.7%	60.2%	62.7%	56.5%	4.1%
BS with Half ERE (Percentage)	29.9%	30.5%	30.8%	31.6%	30.2%	36.8%	42.7%
BS with no ERE (Percentage)	4.8%	6.5%	6.5%	8.1%	7.1%	6.7%	53.2%

The percentages of binding sites with no ERE was significantly lower (p -value = 0.037) than the percentages of the SERM-unique binding sites contain no ERE (Table 15) which range from 11~19%. This implies that the unique binding sites to SERMs more likely to use alternative binding mechanism such as the tethering mechanisms by anchoring to another transcription factor like Sp-1.

Table 15 Distribution of full ERE, half ERE and no ERE for unique binding sites to SERMs

	Unique to SERM (Exclusive from E2)					Random
	TE	T	RE	R	I	
Total BS	1036	1186	580	577	669	100000
BS with Full ERE	530	597	258	253	252	4100
BS with Half ERE	363	421	213	215	344	42700
BS with no ERE	143	168	109	109	73	53200
BS with Full ERE (Percentage)	51.2%	50.3%	44.5%	43.8%	37.7%	4.1%
BS with Half ERE (Percentage)	35.0%	35.5%	36.7%	37.3%	51.4%	42.7%
BS with no ERE (Percentage)	13.8%	14.2%	18.8%	18.9%	10.9%	53.2%

When doing de novo motif prediction on the top 500 E2 binding sites, an ERE-like motif was obtained (Figure 41). We further tested the motifs underlying unique sites. With reference to Figure 42, unique TE and T binding sites also give similar ERE-like sequences. For unique RE and R binding site, the obtained 13bps motifs were similar to the ERE consensus sequences except a 1 bp shift. Given that unique RE and R were different and obtained separately, this 1bp shift was likely not due to random. This could be due to a conformation change in ER+R complex (complex formed when ER binds with R) that the preferential binding to ERE site was shifted by 1 bp. This conformation change was similar for the mixture of ER+E2 / ER+R complexes when co-treatments of E2 and R were applied. Lastly, it seemed that ER+I complex had the greatest conformation changes that the de novo motif for ER+I complex comprised only 1 half-ERE “GGTCA”, i.e. the other half-site “TGACC” were not observed from the relative frequency of the nucleotides. The sequences flanking the half-ERE were also very different. Furthermore, the typical 3-spacer sequences were not present. In summary, ER+I complex had the greatest conformation change and the de novo motif was also the least similar to an ERE sequence.

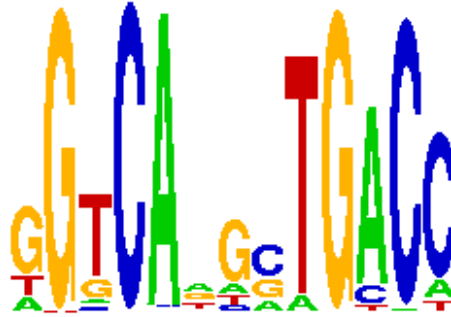


Figure 41 De novo motif prediction on top 500 E2 binding site

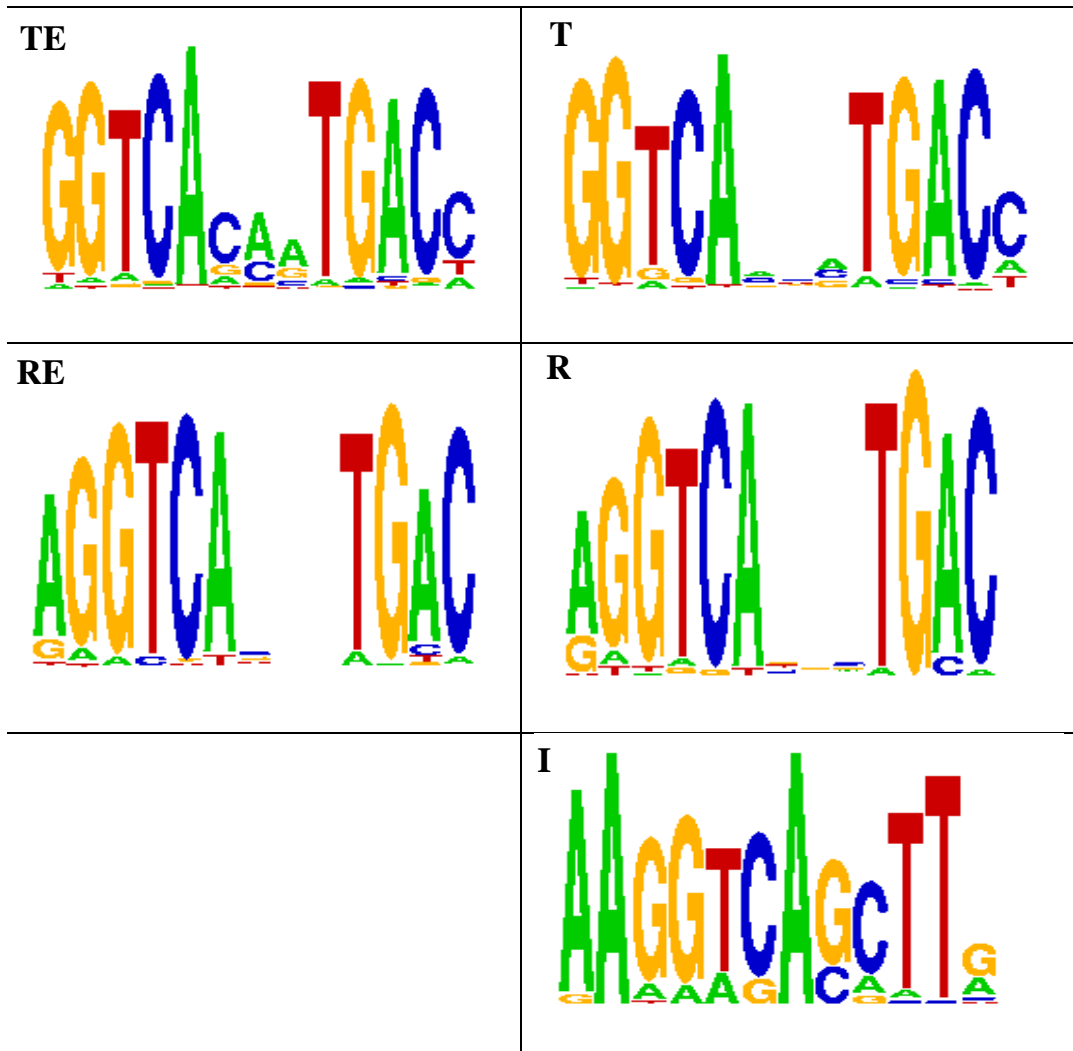


Figure 42 De novo motif prediction on unique SERMs binding site

3.5 Binding sites with basal occupancy are more accessible to TF than those without basal occupancy

Nucleosome positionings and dynamics play important roles in gene regulation and malfunctions in nucleosome regulations give rise to cancer and defective growth (Bu, Evrard et al. 2007). It has been well-established in yeast that active promoters have corresponding nucleosome depletion (Lee, Shibata et al. 2004). We examined how the nucleosome occupancy varies on different categories of binding sites that were induced in E2 with and without basal occupancy. It is posited that the opening-up of chromatin was greater in binding sites with basal occupancy than those without basal occupancy on average. ERE locations were computationally determined for both E2_DM and E2_only categories. Subsequently, the average nucleosome profiles were plotted from the center of ERE for each category as shown in Illustration 3.

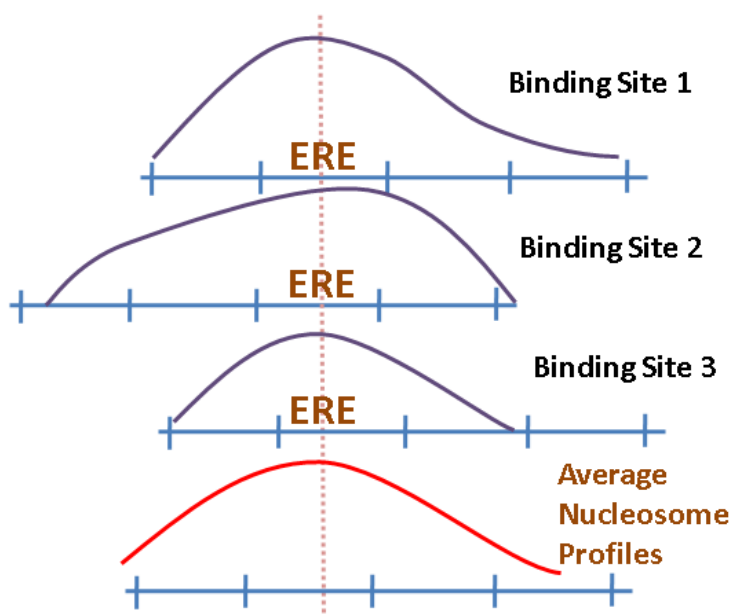


Illustration 3 How average nucleosome profiles are obtained

The average nucleosome profiles for about 1507 and 521 binding sites found with nucleosome probes for the E2_DM and E2_only categories respectively are shown in Figure 43. Using student's t-test with unequal sample sizes and unequal variances, the

p-value was $2.16E-8$ for the difference between the nucleosome signals from E2_DM and E2_only categories at the center of ERE. The nucleosome profile for E2_DM showed a clear depletion of nucleosome, indicative of open chromatin conformation. On the other hand, there was no clear nucleosome depletion for the E2_only category. The above figure may confirm a typical relationship interplayed between nucleosomes and transcription factors binding that nucleosomes were initially well-positioned at equal intervals and opened up to facilitate better transcription factor binding. The depletion in nucleosomes for E2_DM may also be attributed to constitutive accessible promoters that the binding sites are located in DNA regions not associated with any histones, i.e. in a linker region. The reasons for the lack of discernable nucleosome depletion in E2_Only could be that the accessibility to promoters is transient as mediated by the SWI2/SNF2 chromatin remodeling complexes. Hence, the nucleosomes may not be positioned permanently in the same place due to transient placement or eviction of histones and nucleosome sliding of histone octamer, which explains the low average nucleosome occupancy observed for the E2_Only.

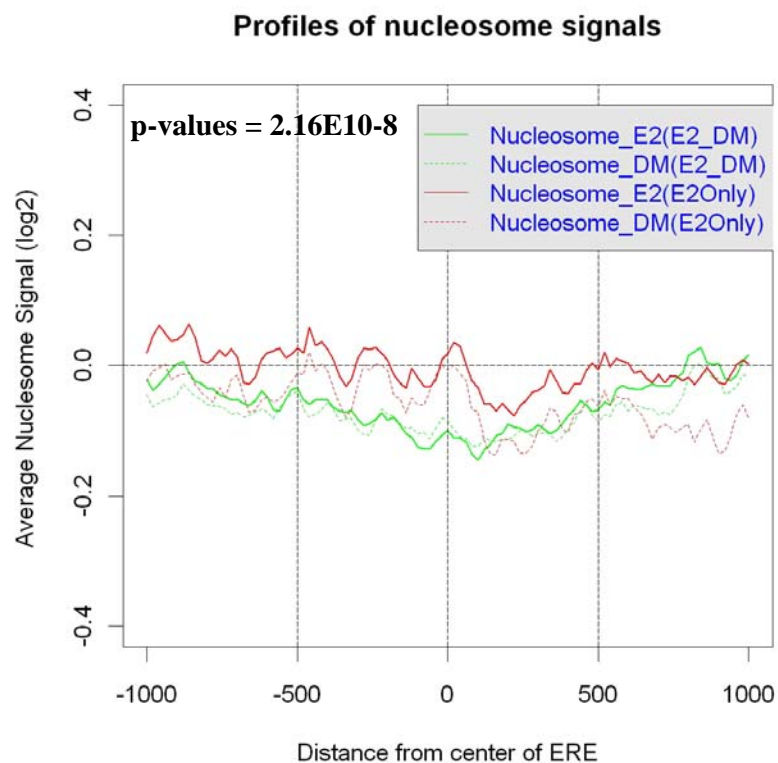


Figure 43 Nucleosome profiles for E2_DM and E2_only

We next examined whether the nucleosome profiles had any relation to the ER occupancy level. We subdivided E2_DM and E2_only categories by the fold enrichment of ER occupancy and plotted the respective nucleosome profiles. There were 878, 629, 412 and 109 binding sites found with nucleosome probes for the categories E2_DM_gt15 (Fold change \geq 1.5), E2_DM_lt15 (Fold change \leq 1.5), E2Only_gt15 (Fold change \geq 1.5) and E2Only_lt15 (Fold change \leq 1.5) respectively. For the category E2_DM, greater fold change corresponded to greater nucleosomes depletion. The nucleosome profiles for E2 were found to be significantly different between E2_DM_gt15 and E2_DM_lt15 (p-value = 3.649E-136 using student's t-test). For the category E2_only, lesser fold change seemed to have more fluctuations and more similar to the DMSO condition.

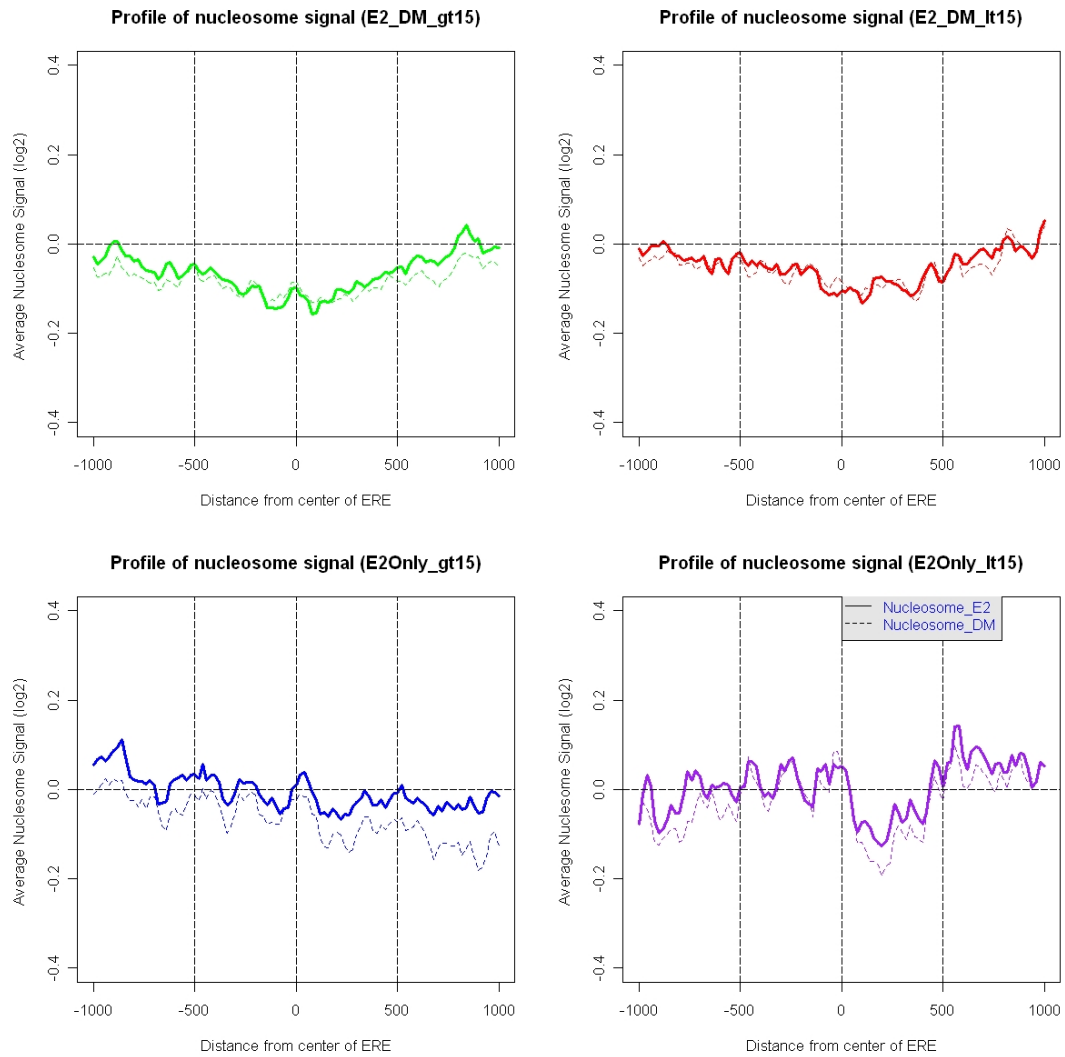


Figure 44 Profiles of nucleosome (Categorized w.r.t 1.5 fold change)

3.6 Binding sites with basal occupancy show greatest FAIRE signals and highest H3K4Me1 enhancer marks, indicative of more accessible DNA regions

Using ChIP-on-chip experiments covering 30Mb of human genome, it was found that active promoters have high signals in H3K4Me3 whereas enhancers have high signals in H3K4Me1 but not H3K4Me3 (Heintzman, Stuart et al. 2007). Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) is an experimental

procedure to isolate nucleosome-depleted DNA from chromatin. In FAIRE, DNA coincident with the location of DNaseI hypersensitive sites, transcriptional start sites, and active promoters in human cell lines is enriched. Both the FAIRE experiment and the ChIP experiment with H3K4Me3 antibodies were performed in MCF-7 cell line under both E2 and DMSO conditions.

With reference to Figure 45, the category of E2_DM was seen with the highest FAIRE signal, E2_only had the next higher FAIRE signal. Student t-test on the FAIRE signal between E2_DM and E2_only showed statistically significant difference (p-value = 3.392e-12). On the other hand, H3K4Me1 signal stayed the highest for the category E2_DM. Student t-test on the H3K4Me1 signal between E2_DM and E2_only also showed statistically significant difference (p-value < 2.2e-12). Only the category E2_DM with basal occupancy has observable high FAIRE and H3K4Me1 signals compared to E2_Only category without basal occupancy. High H3K4Me1 signals in the category E2_DM might indicate that the majority of the basal occupied binding sites were functioning as an enhancer. High FAIRE signals in the category E2_DM might indicate that the majority of the basal occupied binding sites were located in more opened chromatin configuration. This observation of E2_DM involved more in opened chromatin configuration was previously seen in the nucleosome profile (Figure 43) and is confirmed again here. On the other hand, the binding sites in E2_only category were involved in much lesser role as enhancers and situated in much less opened chromatin configuration.

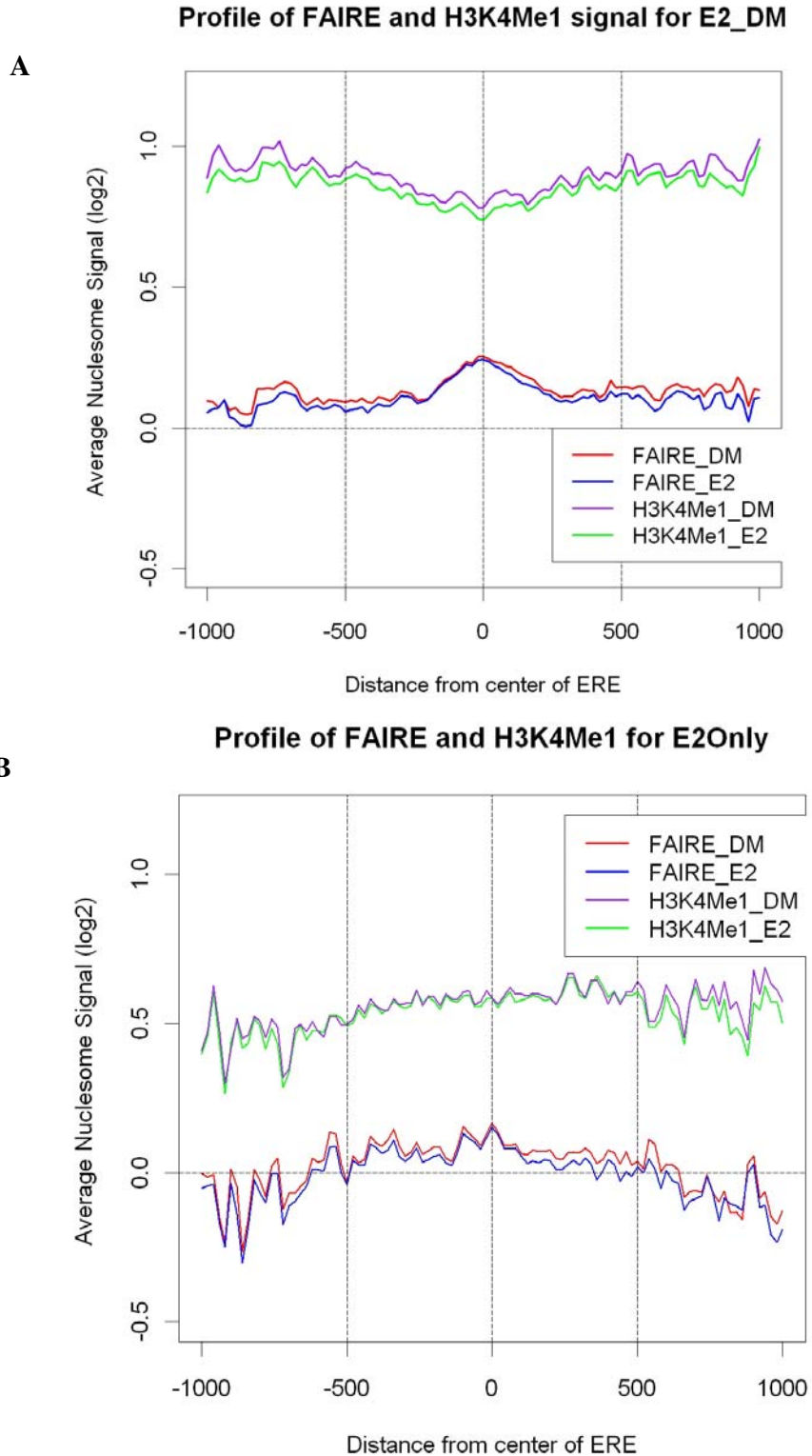


Figure 45 Profiles of FAIRE, K4Me1 and Nucleosome Signals

A) Profiles of FAIRE and K4Me1 Signals for E2_DM; B) Profiles of FAIRE, K4Me1 and Nucleosome Signals for E2_only; The list of perfect consensus ERE, ERE with 1 or 2 mismatches has been computationally obtained from whole genome using hg17 database.

In summary, binding sites with basal occupancy play greater roles as an enhancer due to the high H3K4Me1 signals and also situated in more opened chromatin configuration as shown by the high FAIRE signals. The above observations were not seen so much in binding sites without basal occupancy.

3.7 FOXA1 does not play major role as a pioneering factor but largely attributed to constriction while GATA3 functions as co-factor

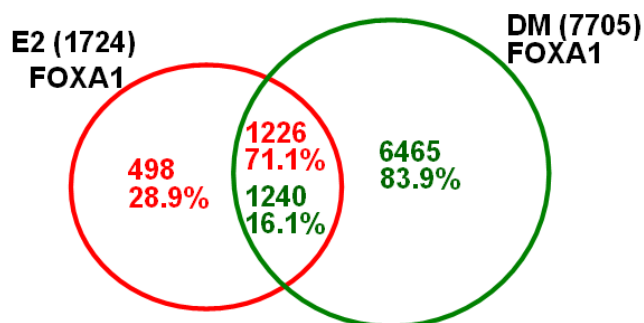
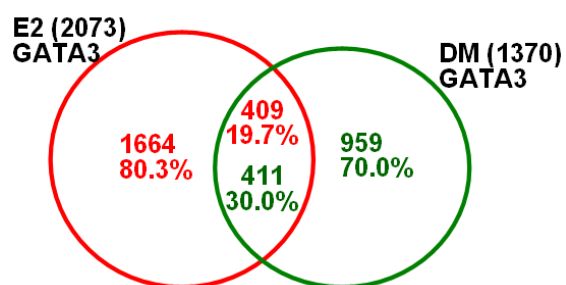
Earlier studies on Transfac analysis has detected the presence of enriched GATA3 and FOXA1 consensus sequences proximal to ER binding sites (Carroll, Meyer et al. 2006; Clark, Burch et al. 2007). CHIP-on-chip studies enable the actual mapping of the binding sites of these transcription factors – GATA3 and FOXA1. Here we correlated ER binding sites to FOXA1 and GATA3 binding sites to investigate in detail their co-occupancy as evidenced by Transfac analysis and how FOXA1 may function as pioneer factor, which opens the chromatin to allow ER accessibility, as previously reported (Carroll, Liu et al. 2005),.

Table 16 shows the VFLM results for FOXA1 and GATA3 based on the same criteria used for MCF-7 cell line, i.e. p-value ≤ 0.01 and fold change ≥ 1.5 , 1725 and 7706 FOXA1 peaks were found in E2 and DM treatments. Whereas GATA3 has 2074 and 1371 peaks in E2 and DM treatments. Interestingly, FOXA1 was found to have 4.5 times more peaks in DMSO condition compared to E2 condition. But GATA3 has similar number of peaks in both treatments.

Table 16 VFLM results for FOXA1 and GATA3

VFLM Peaks P-value=0.01; Fold change =1.5	Differential Enrichment (FOXA1)	FOXA1 Peaks (E2)	FOXA1 Peaks (DM)
Number of Binding Sites	676	1725	7706
VFLM Peaks P-value=0.01; Fold change =1.5	Differential Enrichment (GATA3)	GATA3 Peaks (E2)	GATA3 Peaks (DM)
Number of Binding Sites	136	2074	1371

Figure 46 shows the overlap between FOXA1 binding sites found in E2 and DMSO treatments, 71% of the peaks in E2 have basal occupancy i.e. present in DM treatment also. On the other hand, very few (~20%) GATA3 peaks in E2 have basal occupancy as shown in Figure 47.

**Figure 46 Overlap between FOXA1 E2 peaks and FOXA1 DM peaks****Figure 47 Overlap between GATA3 E2 peaks and GATA3 DM peaks**

We then examined the dynamics of the shifts in TF binding before and after E2 treatment. Of particular interest is whether FOXA1 identifies ER binding sites after

E2 stimulation which would be the case for a pioneering factor, or GATA3 may be found in more ER sites in E2 stimulation as would be the case for a co-factor.

We investigated the overlap between ER E2 peaks with both FOXA1 and GATA3 peaks under both E2 and DM conditions, see the 3-way venn diagrams between ER peaks and FOXA1 peaks in Figure 48.

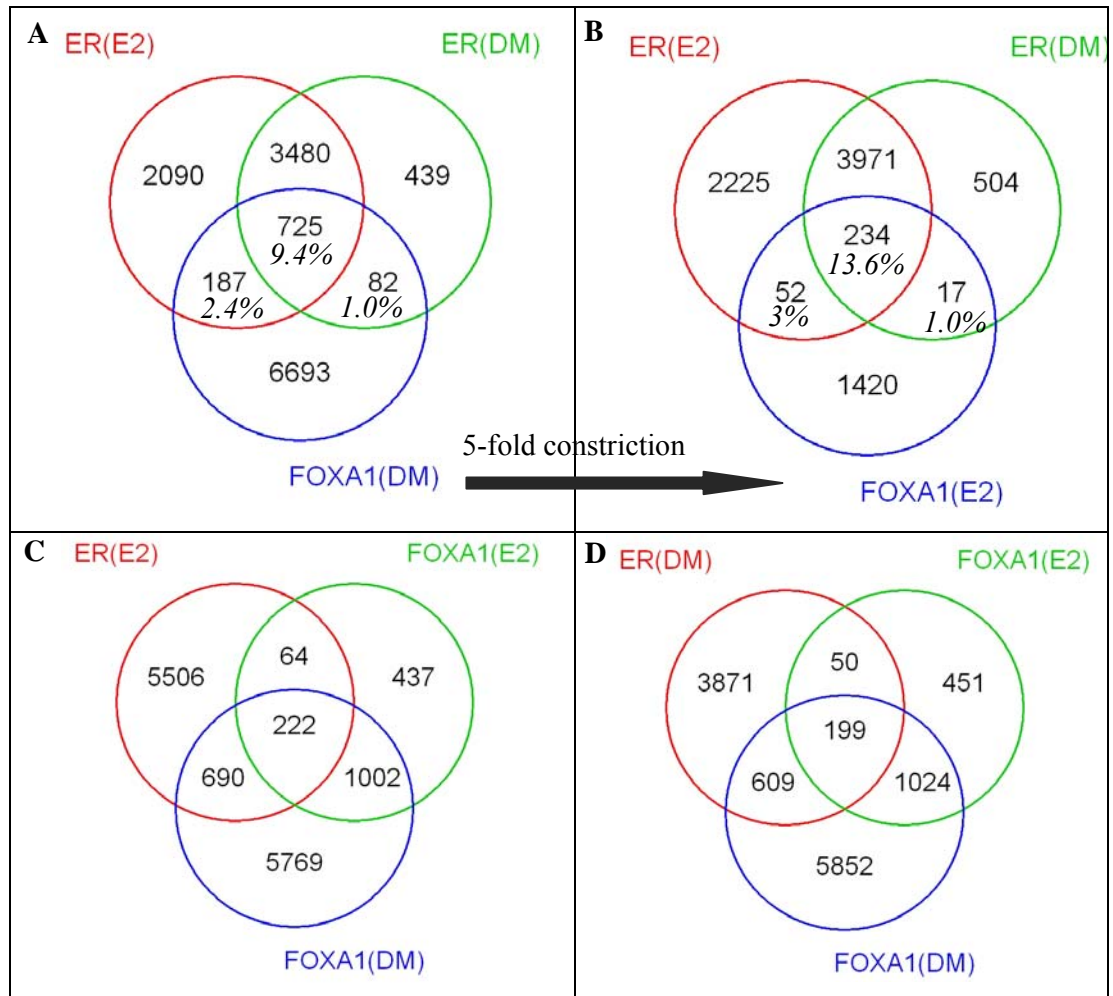


Figure 48 *Overlap between FOXA1 peaks and ER peaks*

A: Overlap between FOXA1 DM peaks, ER E2 and DM peaks ; B: Overlap between FOXA1 E2 peaks, ER E2 and DM peaks; C: Overlap between ER E2 peaks, FOXA1 E2 and DM peaks ; D: Overlap between ER DM peaks, FOXA1 E2 and DM peaks

FOXA1 binding was restricted in the diversity of sites by 5 fold. However, we observed that FOXA1 binding was lost at all sites regardless of subsequent ER binding in the same proportionality. This observation raises the possibility that

FOXA1 is not a pioneering factor (i.e. a factor that first binds leading to the enhancement of ER binding) but that all FOXA1 binding is reduced genome wide.

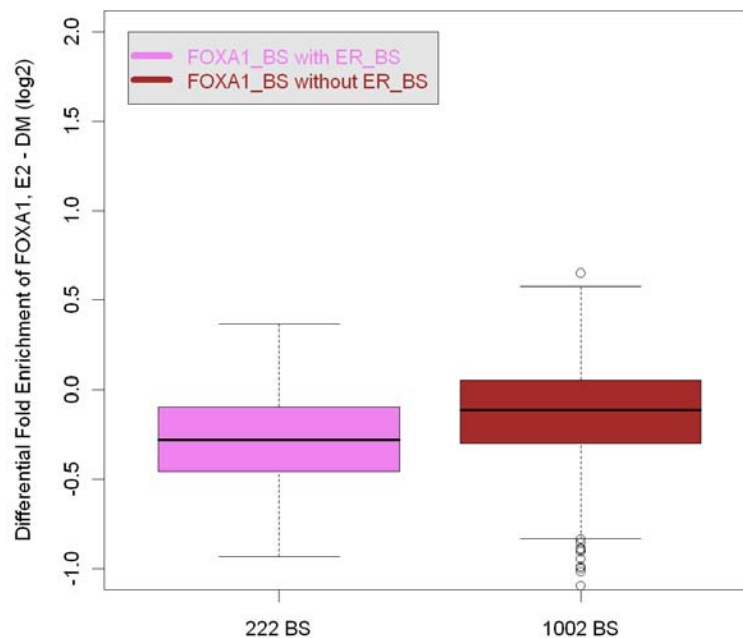


Figure 49 *Boxplot of Change in FOXA1 occupancy in sites co-occupy ER binding sites vs. those that are not used as ER binding sites from DM to E2 condition*

Figure 49 shows the boxplots of the differential enrichments of 222 and 1002 binding sites. Our results show that there is global decline in all FOXA1 binding but the decline in FOXA1 binding used by ER is steeper than those sites without ER binding (t-test p-value = $9.415e-13$). Therefore, FOXA1 binding site constriction is not the sole mechanism. Also, it is unlikely that the major role of FOXA1 is a pioneering factor for all ER binding sites but may still have a significant role in ER biology as a partner factor. The observation that ER recruitment is associated with falling off of FOXA1 at most of the sites is largely attributed to FOXA1 binding sites' constriction and partially due to FOXA1 role as a pioneering factor for a few ER binding sites.

GATA3 on the other hand appears to behave with different kinetics. While there is a considerable overlap between FOXA1 sites before and after E2 (with the only difference in the reduction in the number of other sites), GATA3 showed a modest but significant expansion of the number of binding sites (1370 to 2073) and a considerable shift in binding sites after E2 (Figure 50). The overlap with ER binding sites also increased after E2. This suggests that GATA3 functions as a co-occupying cofactor for ER.

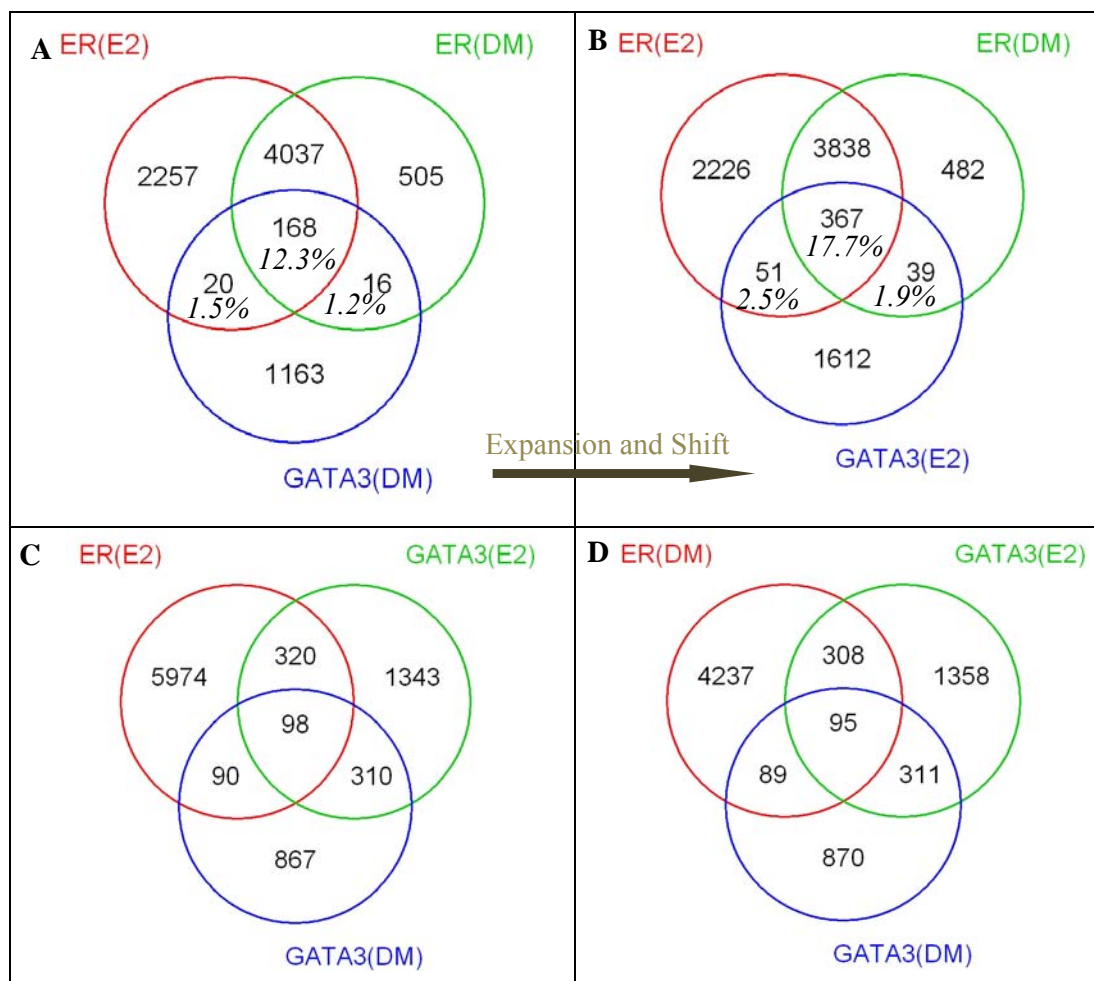


Figure 50 Overlap between GATA3 peaks and ER peaks

A: Overlap between GATA3 DM peaks, ER E2 and DM peaks ; B: Overlap between GATA3 E2 peaks, ER E2 and DM peaks; C: Overlap between ER E2 peaks, GATA3 E2 and DM peaks ; D: Overlap between ER DM peaks, GATA3 E2 and DM peaks

Similarly, we investigated whether GATA3 can potentially function as a pioneering factor for ER. To examine this, we assessed the differential enrichment of GATA3 binding sites common to both conditions that co-occupy ER binding sites vs. those do not i.e. 98 binding sites vs. 310 binding sites respectively as shown in Figure 50C. The boxplots of the differential enrichments of these two sets of binding sites is shown in Figure 51. There is no significant decline in GATA3 occupancy in the ER co-occupied sites and it is not different from the sites not used by ER. This indicates that GATA3 is not a pioneer factor for ER.

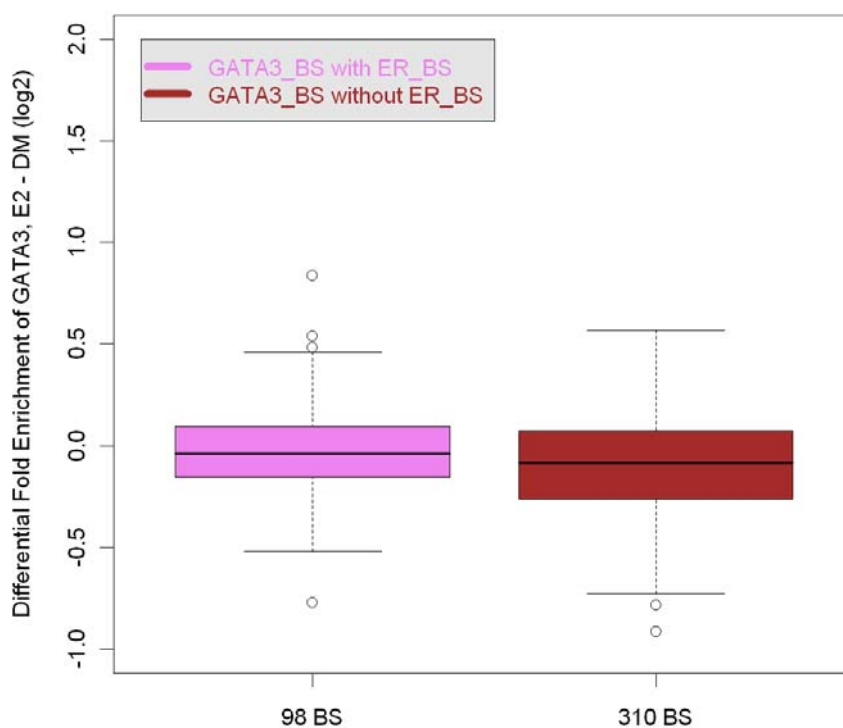


Figure 51 Boxplot of Change in GATA3 occupancy in sites co-occupy ER binding sites vs. those that are not used as ER binding sites from DM to E2 condition

In summary, FOXA1 does not play major role as a pioneering factor but the observation that ER recruitment associates with FOXA1 falling-off at most of the sites is largely attributed to FOXA1 binding sites' constriction. On the other hand, GATA3 do not function as a pioneer factor but acts as a co-factor.

3.8 H3K4Me1 is the most predictive factor for identifying ER binding sites

We would like to find rules that govern the binding sites occupancy by ER and make prediction on the binding sites utilization from the H3K4Me1 histone mark, FOXA1, GATA3 and FAIRE Signals. Using 6444 E2 binding sites and randomly selected 3333 non-binding sites, the corresponding signals at the binding sites location were extracted for the H3K4Me1 histone mark, FOXA1, GATA3 and FAIRE Signals. All binding sites that corresponded to known amplified regions were removed from the data analysis as amplified regions tend to exaggerate the binding sites enrichment which might not reflect true biological significance. Subsequently, the ‘rpart’ package in R, which is based on Classification and Regression Trees (CART) algorithms, was used for constructing the tree.

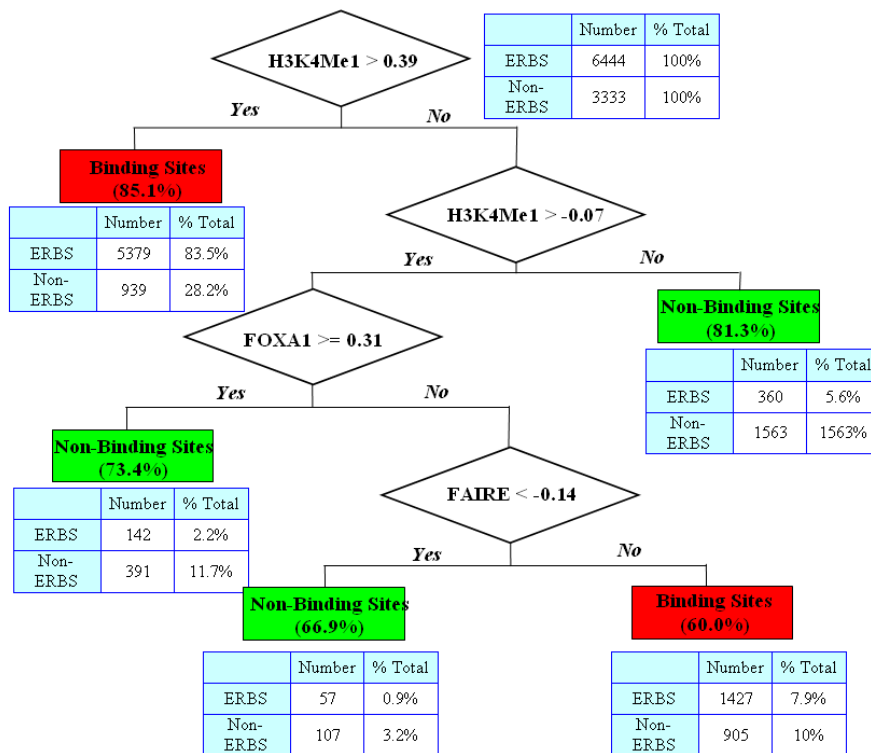


Illustration 4 Decision Tree for rules governing ER binding

By using the ChIP-Seq counts of individual epigenetic marks or the sum of the ChIP-Seq counts of a group of epigenetic marks, the receiver operating characteristic (ROC) curve was constructed (Figure 52). This is a graphical representation of sensitivity i.e. proportion of true positives versus $(1 - \text{specificity})$ i.e. proportion of false positives. The ideal situation is to obtain significant true positive fraction and negligible false positive fraction, which leads to an ideal plot of an upper triangle. To facilitate the comparison between different models, the Area Under the ROC curve (AUC) was computed. The higher the AUC values, the better the model in classifying the 2 classes of binding and non-binding sites.

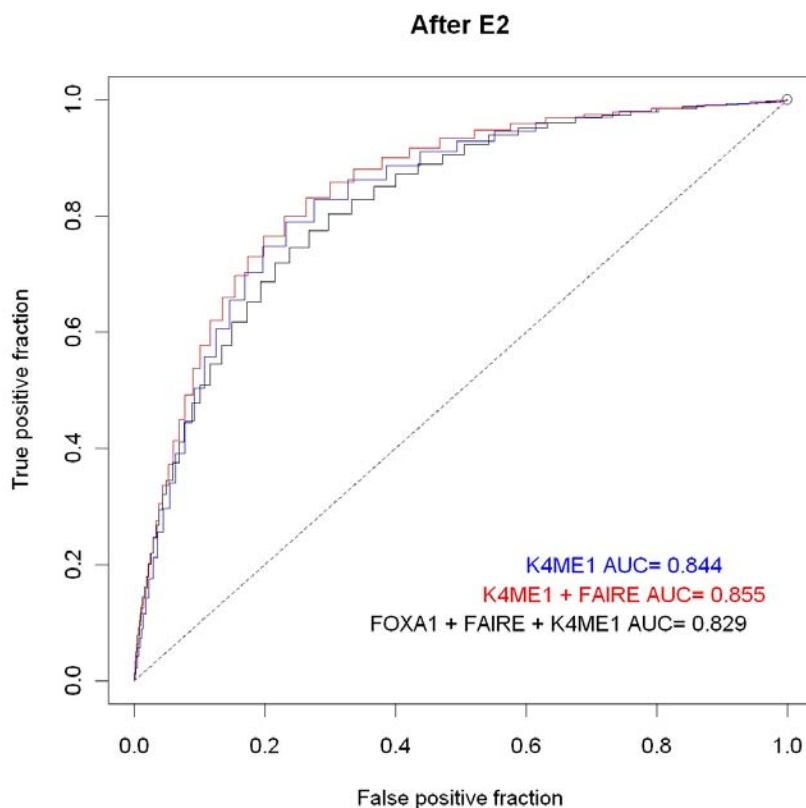
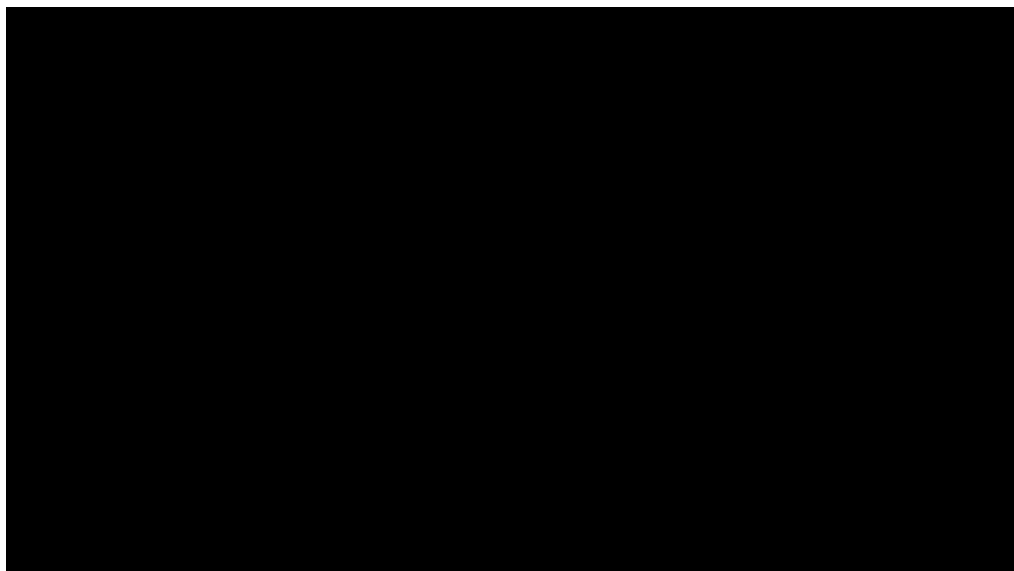


Figure 52 Plot of ROC curve for K4ME1 epigenetic mark

Table 17 Summary of AUC for all epigenetic marks alone and in combinations



Based on Table 17, the optimal model for classifying ER binding sites and non-binding sites is based on the combination consists of H3K4ME1 and FAIRE. H3K4ME1 alone was not far below. There was a slight improvement of 0.011 AUC over the prediction when combination of H3K4ME1 and FAIRE were used together compared to H3K4ME1 alone. Interestingly, FOXA1 and GATA3 marks alone were the worst predictors for the classification of putative ER binders.

3.9 ER regulates distinctive promoters and enhancers in Ishikawa cell line from MCF-7

ER is distributed with varying concentrations in different tissues. Possible explanations for the tissue-specific effects may be attributed to the differing levels of ER present and the differential transactivation properties of ER when binding to different enhancers. We hypothesized that the distinctive actions of ER in different cell lines is due to the distinctive and diverse binding sites profiles. Hence we obtained the binding site profile in uterus cell line (Ishikawa) and compared it to breast cell line (MCF-7). Table 18 shows the VFLM results on Ishikawa cell line.

Based on the same criteria used for MCF-7 cell line, i.e. p-value ≤ 0.05 and 1.5 fold change, 3661 E2-peaks, 4913 T-peaks and 5869 R-peaks were obtained. As seen in Figure 53, the overlap of MCF-7 and Ishikawa E2-peaks is only 12% of MCF-7 E2-peaks or 22% of Ishikawa E2-peaks. This indicates almost completely different utilization of ER binding sites in MCF-7 and Ishikawa cell lines.

Table 18 Linear model results on Ishikawa cell line

VFLM Peaks (E2) P-value=0.01; Fold change ≥ 1.5	Differential Enrichment (E2)	E2 Peak	DM Peak
Number of Binding Sites	3220	3661	710
VFLM Peaks (T) P-value=0.01; Fold change ≥ 1.5	Differential Enrichment (T)	T Peak	DM Peak
Number of Binding Sites	5910	4913	710
VFLM Peaks (R) P-value=0.01; Fold change ≥ 1.5	Differential Enrichment (R)	R Peak	DM Peak
Number of Binding Sites	8481	5869	710

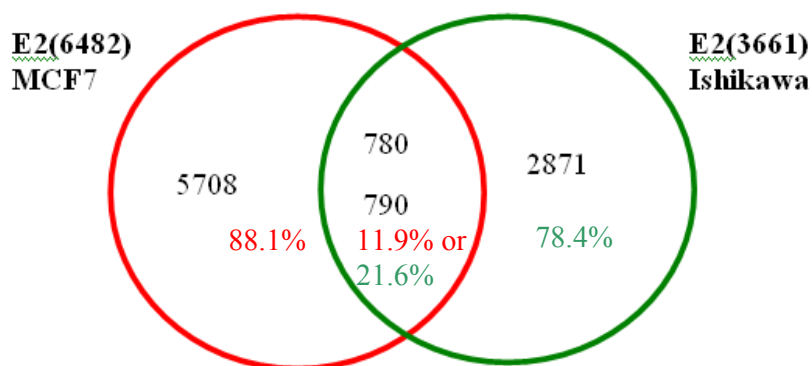


Figure 53 Overlap in binding sites between MCF-7 and Ishikawa with E2 treatment

When we compared the overlap between MCF-7 T-peaks and Ishikawa T-peaks as shown in Figure 54 and the overlap between MCF-7 R-peaks and Ishikawa R-peaks as shown in Figure 55, there were overlaps of 6.4% and 8.1% of Ishikawa peaks respectively. The low overlaps of less than 10% showed that the utilization of ER

binding sites in MCF-7 and Ishikawa cell lines will be more distinctly different when SERMs were used instead of E2. Previously, the overlap was 22% when E2 was used instead of SERMs. This suggests that the mechanisms of SERMs actions are unique in different cell lines, thus changing the overlap from 22% to less than 10%.

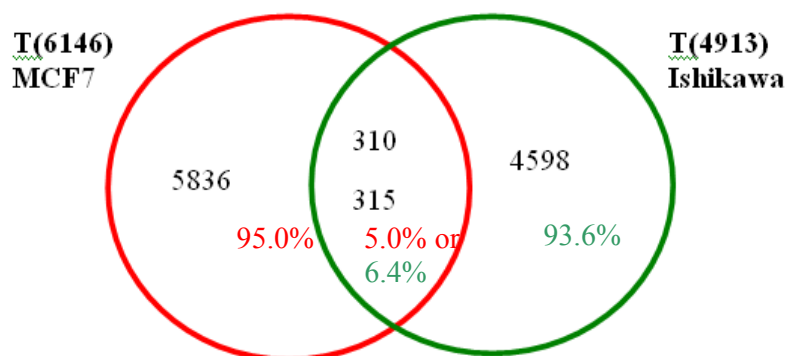


Figure 54 Overlap in binding sites between MCF-7 and Ishikawa with tamoxifen treatment

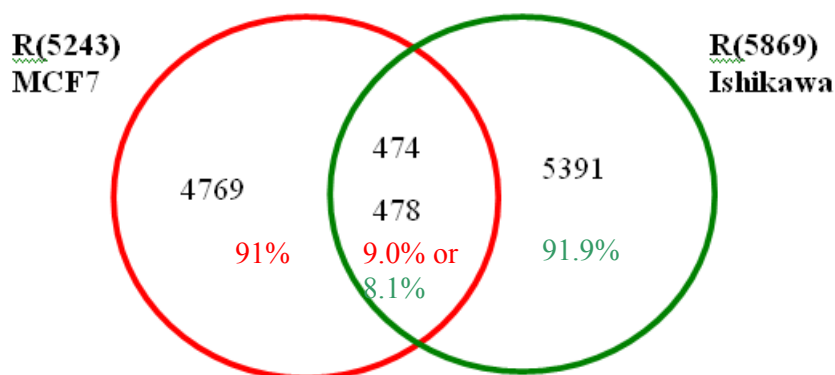


Figure 55 Overlap in binding sites between MCF-7 and Ishikawa with R treatment

3.10 Concluding remarks

In conclusion, we obtained genome-wide binding sites profiles with the customised ER chip for different drugs treatments (E2, T, R and I), different antibodies (ER, H3K4Me1, FOXA1 and GATA3), different experiments (ChIP and FAIRE) and different cell lines (MCF-7 and Ishikawa cell lines). A new algorithm called VFLM was designed and implemented, which detected 6482 VFLM peaks for

E2 treatment. Genome-wide binding sites analysis showed high prevalence of pre-occupied ER binding sites which also had a greater ER recruitment upon E2 treatment. Binding site profiles for SERMs showed similar location as E2 but the ER binding recruitments in SERMs were significantly much less than in E2. De novo motif analysis on the unique binding sites to different treatments showed that ER-SERMs might utilize tethering mechanism much more than ER-E2. ER-SERMs conformation was different in different SERMs with I seemed to have the greatest differential conformation change, followed by R and the least conformation change was T. Comparing E2 binding sites with and without basal occupancy, binding sites with basal occupancy showed more accessible chromatin configuration, had the greatest FAIRE signals and highest H3K4Me1 enhancer marks. H3K4Me1 was also found to be the most predictive factor for identifying ER binding sites through both decision tree and ROC curve. We also further confirmed the role of FOXA1 as pioneering factor while GATA3 functions as co-factor. Lastly, we showed that the ER binding were regulated differently in MCF-7 and Ishikawa cell line.

Chapter 4 Integrative Analysis of SERMs on ER Responses on a Genome Wide Scale

4.1 Identification of regulated genes in SERMs and E2 treatments

A comprehensive dataset of time-course microarray experiments was performed to investigate the effects of estrogen and SERMs treatment on gene expression profiles and to identify estrogen responsive genes. Estrogen treated (10 nM) or SERMs treated and DMSO-mock MCF-7 cells (negative control) for 0.5, 3, 6, 9, 12, 24, and 48 hours were collected for RNA extraction and the labeled cDNA were hybridized to microarrays (HG-U133 Plus). 3 biological replicates were performed for each time point. The whole process of Affymetrix data analysis procedure is illustrated in Figure 56.

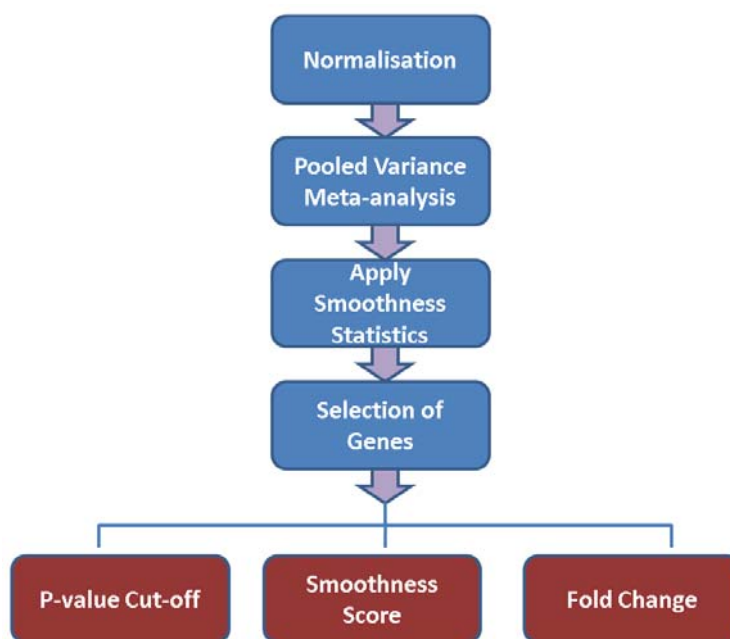


Figure 56 Schematic of Affymetrix gene expression analysis

Normalisation was performed for all microarrays using Robust Multichip Average (RMA) Normalisation (Irizarry, Ooi et al. 2003) from the AFFY package with the software R (Ihaka R. & Gentleman R. 1996). Subsequently, the normalized data was analyzed using Pooled Variance Meta Analysis¹ and ranked by their scores. The expression values are given by the average of (E2/SERMs at time t – E2/SERMs at 0 hr) – (DMSO at time t – DMSO at 0 hr). Regulated genes are obtained based on Data-driven Smoothness Enhanced Variance Ratio Test (dSEVRAT)² cut-off ≥ 200 , minimum PVMA p -value ≥ 0.05 in any time points from 3 hrs to 48 hrs and fold change greater than 1.5 fold (≥ 0.585) or less than 1.5 fold (≤ -0.585) in any time points from 3 hrs to 48 hrs.

Briefly, Pooled Variance Meta Analysis method defines a new T-statistics that incorporates a pooled variance from all the variances from different treatments. This

¹ Juntao li, Sneha Ravi, Jianhua Liu and R. Krishna Murthy Karuturi, Pooled Variance Meta-analysis for Small-sample Multi-group Microarray Expression Data Analysis, in the poster proc of *The 18th Intl Conf on Genomic Informatics (GIW'07)* , pp171-172, December, 2007, Singapore.

² Juntao Li, Jianhua Liu and R. Krishna Murthy Karuturi . Data-driven Smoothness Enhanced Variance Ratio Test to Unearth Responsive Genes in Unreplicated 0-time Normalized Time-course Microarray Studies, *Springer Verlag Lecture Notes in Bioinformatics series for ISBRA'07* , USA, May 2007.

unified or pooled variance will enable a better power to detect the responsive genes and comparison of the data across treatments and minimize the comparison errors arise from different variances of treatments due to differing variations of samples within the same treatment. Data-driven Smoothness Enhanced Variance Ratio Test (dSEVRAT) is able to assess the general smoothness of the expression of gene across time points. It is an accumulative measurement of the change in expressions between 2 consecutive time points. In general, most gene expressions should follow a trend with only one or two turning points for the time course we have. Thus the dSEVRAT score is able to filter out genes with high noises or fluctuations that are probably not having any biological significance.

Figure 57 below shows the number of probesets of genes filtered out by each criterion and their overlap for E2 treatment. Total number of probesets of genes selected was 12171, 16007 and 8009 for the criteria of $p\text{-value} \leq 0.05$, smoothness score ≥ 200 and fold change ≥ 1.5 respectively. When the criteria of both p-value and smoothness score were satisfied, the fold change criterion was also mostly fulfilled. Smoothness criterion alone also fulfilled either the criteria of P-value or fold change or both. There were 8117 probesets of genes that satisfied both p-value and fold change criteria but out of which 4172 might have irregular changes along the time points so they were not selected. Lastly, we selected 3945 probesets of genes that fulfilled all 3 requirements: $P\text{-value} \leq 0.05$, smoothness score ≥ 200 and fold change ≥ 1.5 .

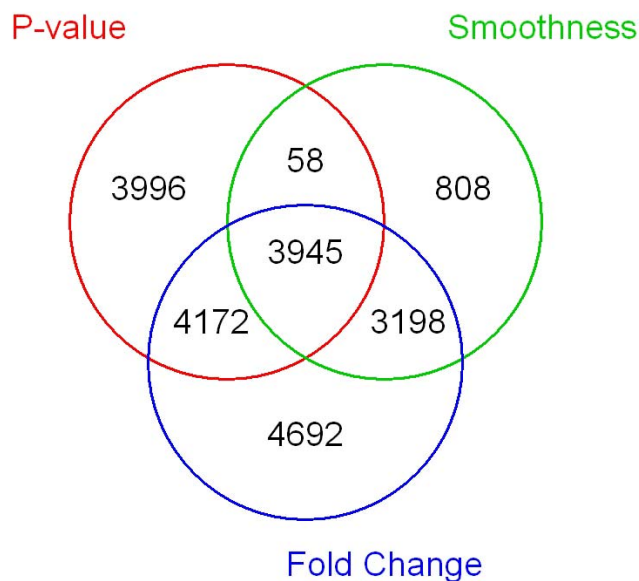


Figure 57 Selection of probesets of genes based on 3 criterias

Using the data analysis outlined above, 3945 probesets of genes or 3098 E2-regulated genes were obtained. Regulated genes are classified as up-regulated or down-regulated genes as follows: With reference to Figure 58, an up-regulated gene has positive expression values at all time point (*MYB* gene). If expression values are inconsistent, i.e. a mixture of positive and negative expression values, gradient is computed by fitting a straight line on the expression values. If the gradient is positive (*KLK11* gene), it is classified as up-regulated gene. For classification of down-regulated genes, the vice-versa applies; *CSTA* gene had all negative expression values while *REG4* gene had negative gradient with inconsistent expression across time-course.

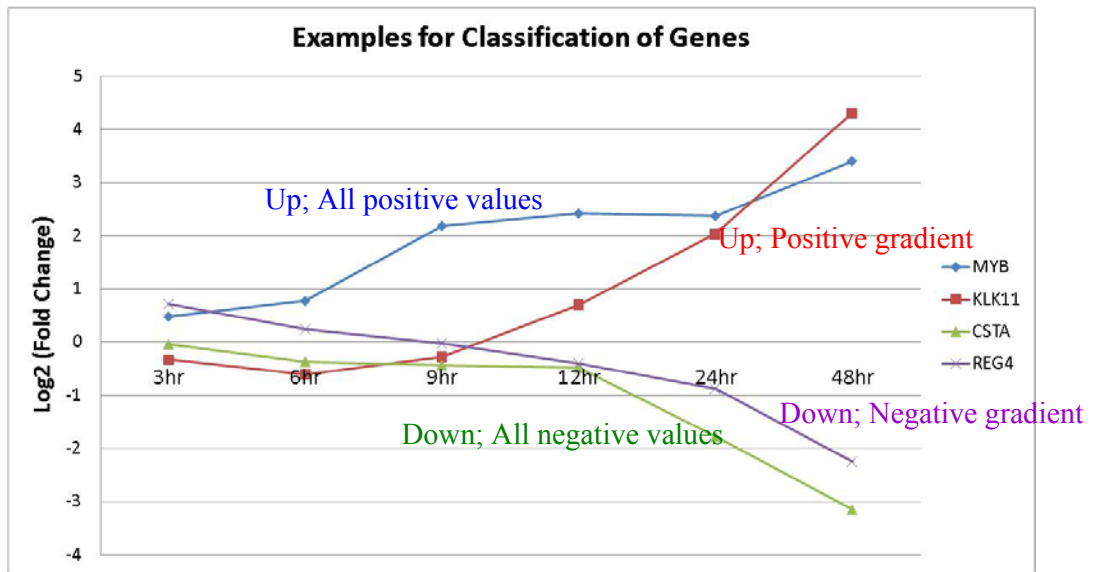


Figure 58 Examples for classification of genes

For the 3098 E2-regulated genes, 1886 genes were up-regulated while 1212 genes were down-regulated. Figure 59 showed the heatmap on the E2-induced genes. For gene expression in E2, a large majority of genes had a general increasing or decreasing trend for up-regulated and down-regulated genes respectively.

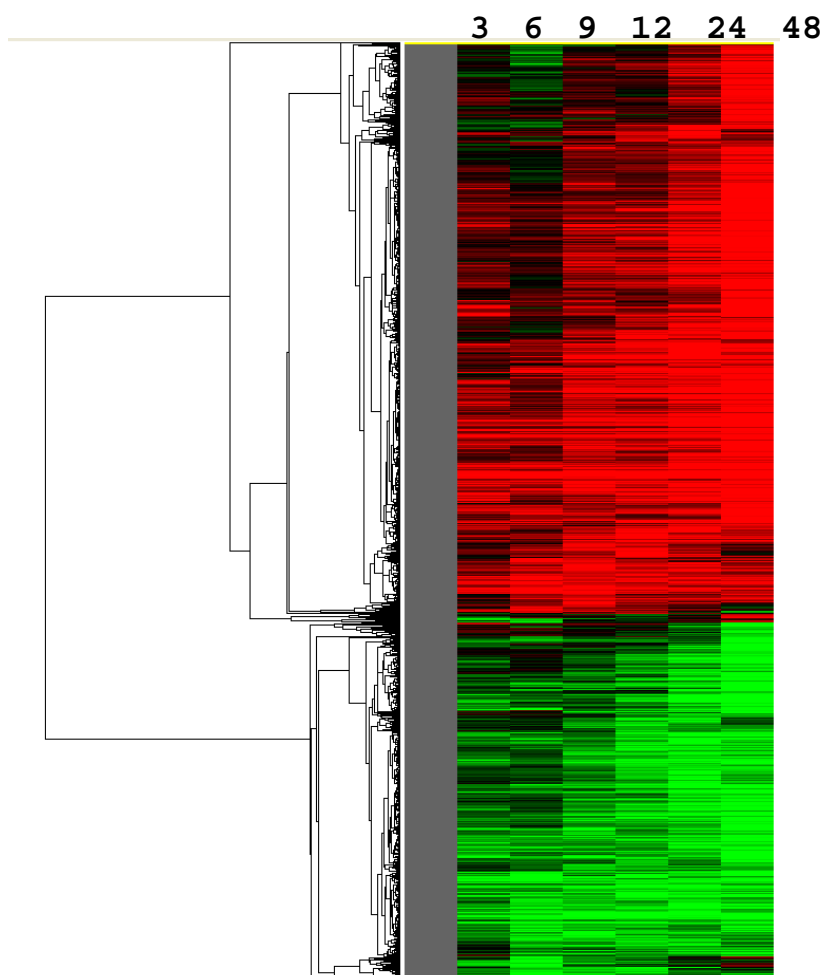


Figure 59 Heatmap for E2-regulated genes in MCF-7

In each treatment condition, the time points were arranged in order from 3, 6, 9, 12, 24 and 48 hours.

Validation on a panel of well-known E2 regulated genes across treatments and time

We examined the expression measurements of well-known E2 regulated genes across treatments and time on our data (Figure 60). The housekeeping gene also shows low expression values throughout (Figure 61). In general, the expression of E2 regulated genes increased with time in E2 treatment. SERMs also had differential antagonizing effects on the expressions of E2 genes.

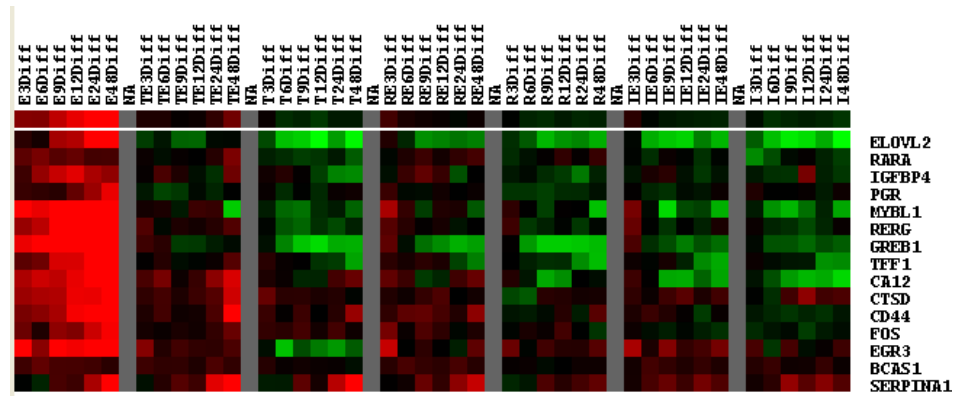


Figure 60 Heatmap for a panel of 15 well-known E2 responsive genes (true positives)

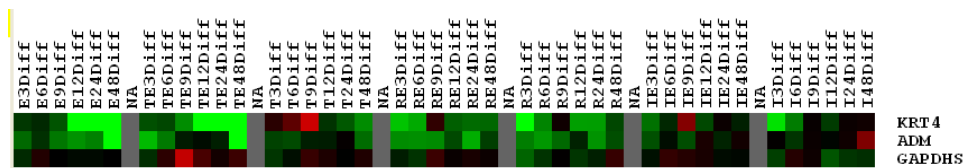


Figure 61 Heatmap for false negative genes and GAPDH house-keeping gene

Subsequently, we examined – *TFF1* (trefoil factor 1) gene for its response upon E2 treatment in greater detail. *TFF1* has been established as the validation target for the E2 responsiveness in breast cancer cells (Brown, Jeltsch et al. 1984). As can be seen in Figure 62, the expression increases with time, the highest induction is registered at 48 hours. Also plotted along was the house-keeping gene (*GAPDH*). Subsequently, it was of great interest to investigate the differential modulating effects of various SERMs, in combination with E2 or SERMs itself acting alone. Figure 63 below showed that all SERMs reduced the fold change or expression of *TFF1* gene. TE had the least reduction while all other SERMs reduced the expression greatly to similar level at 48 hours time point.

In summary, through examination of a panel of E2-genes and the scrutinization of *TFF1* gene across treatments and time, the expression changes were within the expectations on both E2's induction and SERMs' antagonizing properties. This shows that the array data is of good quality for subsequent exploration of gene expression

changes in a global manner and for identification of new genes in E2 and SERMs treatments.

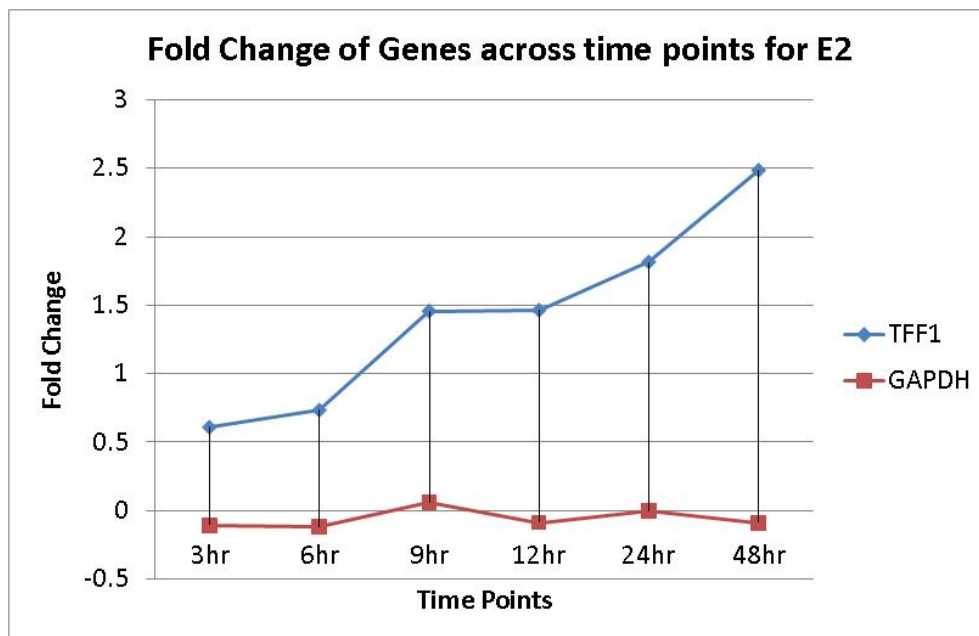


Figure 62 Fold Changes of Genes across time points for E2

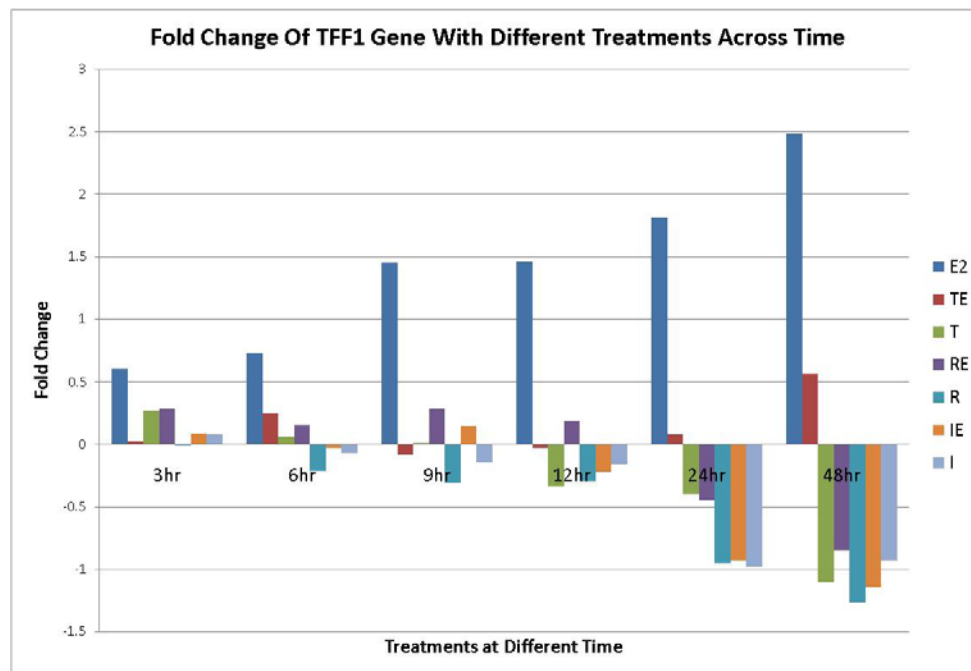


Figure 63 Fold Changes of TFF1 Gene across time points for E2 and SERMs

Gene ontology on all the E2-regulated genes revealed that the biological processes were largely related to cell proliferation such as cell cycle, metabolism, cell proliferation and differentiation (Table 19).

Table 19 Gene ontology for E2-regulated genes

Gene Ontology for regulated genes (E2)

Biological Process	NCBI: H. sapiens genes REFLIST (25431)	E2 Genes	E2 Genes (expected)	E2 Genes (over/under)	E2 Genes (P-value)
Biological process unclassified	11321	907	1228.66	-	1.73E-34
Nucleoside, nucleotide and nucleic acid metabolism	3343	546	362.81	+	4.98E-21
Cell cycle	1009	203	109.51	+	5.40E-15
Intracellular signaling cascade	871	170	94.53	+	9.32E-11
Protein modification	1157	200	125.57	+	2.90E-08
mRNA transcription	1914	298	207.72	+	6.68E-08
Protein metabolism and modification	3040	432	329.93	+	1.53E-07
Other metabolism	559	110	60.67	+	1.62E-07
Cell cycle control	418	91	45.37	+	1.64E-07
Cell proliferation and differentiation	1028	175	111.57	+	2.51E-07
mRNA transcription regulation	1459	234	158.34	+	7.43E-07
Protein phosphorylation	660	124	71.63	+	1.52E-06
Olfaction	198	1	21.49	-	1.92E-06
Developmental processes	2152	314	233.55	+	2.65E-06
Cell structure and motility	1148	184	124.59	+	6.08E-06
DNA metabolism	360	77	39.07	+	6.10E-06
Chemosensory perception	207	2	22.47	-	6.49E-06
Signal transduction	3406	461	369.65	+	1.31E-05
Oncogenesis	472	88	51.23	+	4.63E-05
Intracellular protein traffic	1008	159	109.4	+	9.90E-05
DNA replication	155	40	18.82	+	1.45E-04
Proteolysis	960	155	104.19	+	1.77E-04
Other carbon metabolism	82	25	8.9	+	9.95E-04
Ectoderm development	692	114	75.1	+	1.94E-03
Immunity and defense	1318	189	143.04	+	2.82E-03
Endocytosis	277	55	30.06	+	3.70E-03
Neurogenesis	587	99	63.71	+	3.98E-03
Coenzyme and prosthetic group metabolism	174	37	18.88	+	4.27E-03
Apoptosis	531	87	57.63	+	4.91E-03
Other intracellular signaling cascade	225	47	24.42	+	5.78E-03
Transport	1306	185	141.74	+	6.11E-03
Sensory perception	506	31	54.92	-	8.93E-03
Phosphate metabolism	117	27	12.7	+	9.51E-03
General vesicle transport	251	49	27.24	+	1.47E-02
Other transport	61	18	6.62	+	2.67E-02
Carbohydrate metabolism	592	90	64.25	+	3.79E-02

4.2 E2-regulated genes use more of Pol II preloading mechanism and mechanism of down-regulation involves Pol II pausing or stalling

We compared the proximity of regulated genes to Pol II binding sites in both E2 and DMSO conditions. Interestingly, the percentages of genes in each category of distances away from TSS of genes are similar for Pol II binding sites in E2 or DMSO conditions, i.e. between A and C; between B and D (Figure 64). This could indicate that E2 regulated genes may use more of Pol II preloading mechanism as opposed to the classical model of Pol II recruitment. 34.4% and 34.8% of E2 down-regulated genes were within 5-100kb of a Pol II binding site in E2 and DMSO conditions respectively compared to 20.3% and 22.2% of up-regulated genes respectively. This higher percentage of down-regulated genes within 5-100kb of a Pol II binding site

compared to up-regulated genes implies that more RNA Pol II binds along the gene body of down-regulated genes than up-regulated genes. These Pol IIs may possibly act as a hindrance to transcription process known as Pol II pausing or stalling, contributing to one of the mechanisms for down-regulation of genes.

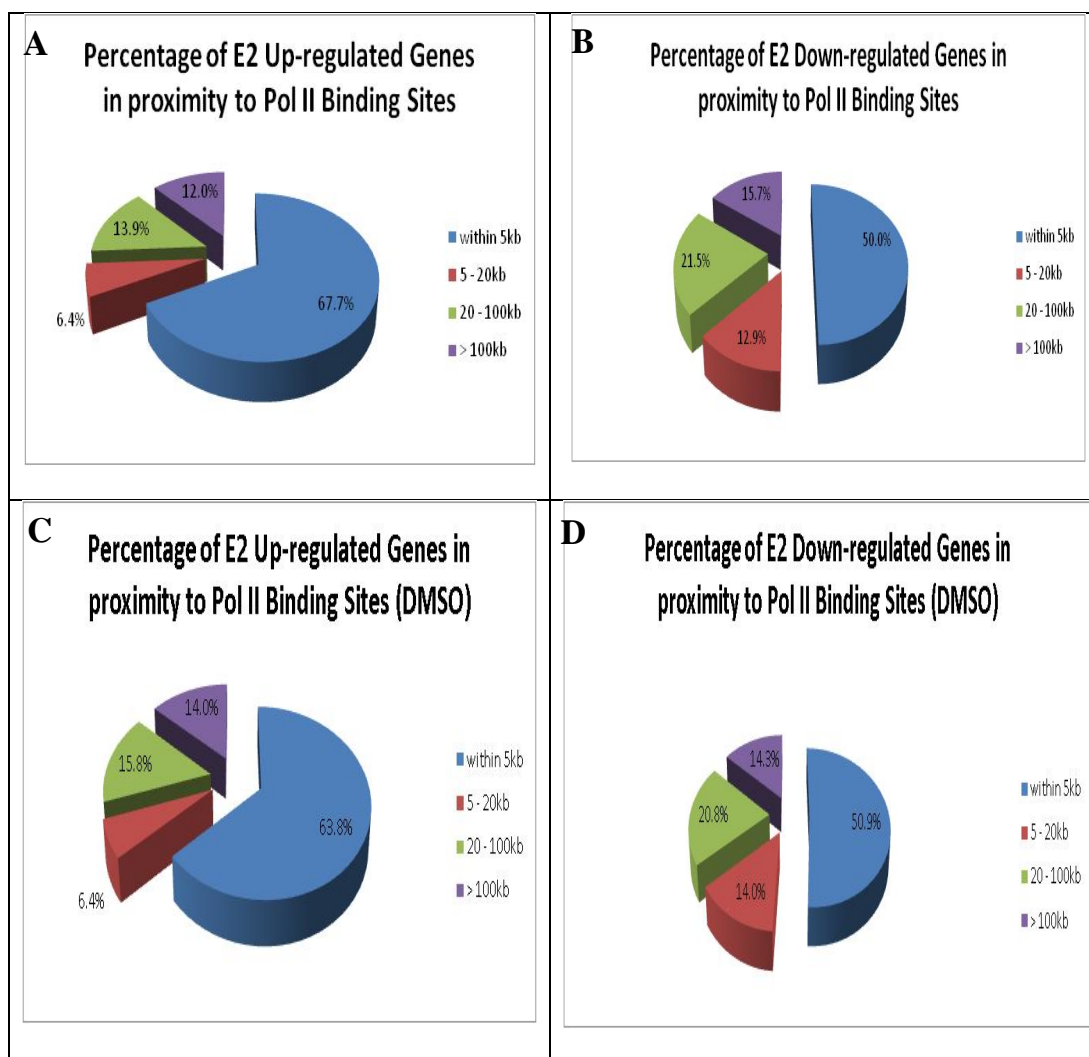


Figure 64 Percentage of E2 up-regulated genes in proximity to Pol II binding sites

A) Percentage of E2 up-regulated genes in proximity to Pol II binding sites in E2; B) Percentage of E2 down-regulated genes in proximity to Pol II binding sites in E2; C) Percentage of E2 up-regulated genes in proximity to Pol II binding sites in DMSO; D) Percentage of E2 down-regulated genes in proximity to Pol II binding sites in DMSO;

4.3 Strongly regulated genes in E2 treatment have ER binding sites in closer proximity than non-regulated genes

We correlated 6482 estrogen receptor binding sites in E2 with the gene expression data upon E2 treatment for all time points. The correlation shows a bias for up-regulated genes at all time points while the bias for down-regulated genes only occurs much later at 24 and 48 hours (Figure 65). Up-regulated genes are strongly correlated to ER binding sites from early time point of 3 hours suggest that these genes are highly directly regulated by ER. In contrast, down-regulated genes only correlated to ER binding sites at a much later time of 24 hours onwards suggest that these genes are indirectly regulated by ER and may be regulated by a secondary transcription factor induced by E2 or through co-operating binding with other transcription factors.

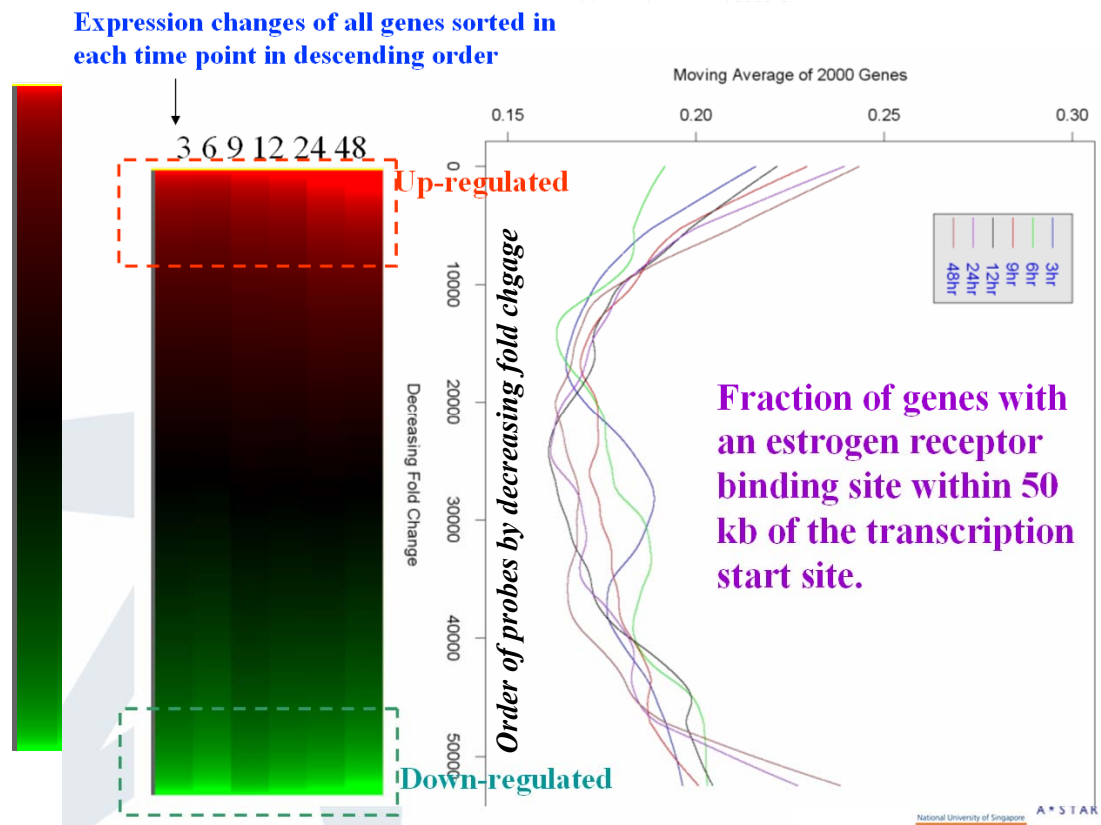


Figure 65 Correlation between gene expression and E2 binding sites

4.4 Strong E2-ER binding sites with basal occupancy associates with E2 up-regulated genes

The binding sites in Figure 27 were categorized in terms of differential enrichments (E2 – DMSO). Subsequently, numbers of E2 up-regulated or down-regulated genes that have an ER binding site within 100kb from TSS of genes were computed. The categories of binding sites were then correlated to the ratio of number of E2 up-regulated over number of down-regulated genes as shown in Figure 66.

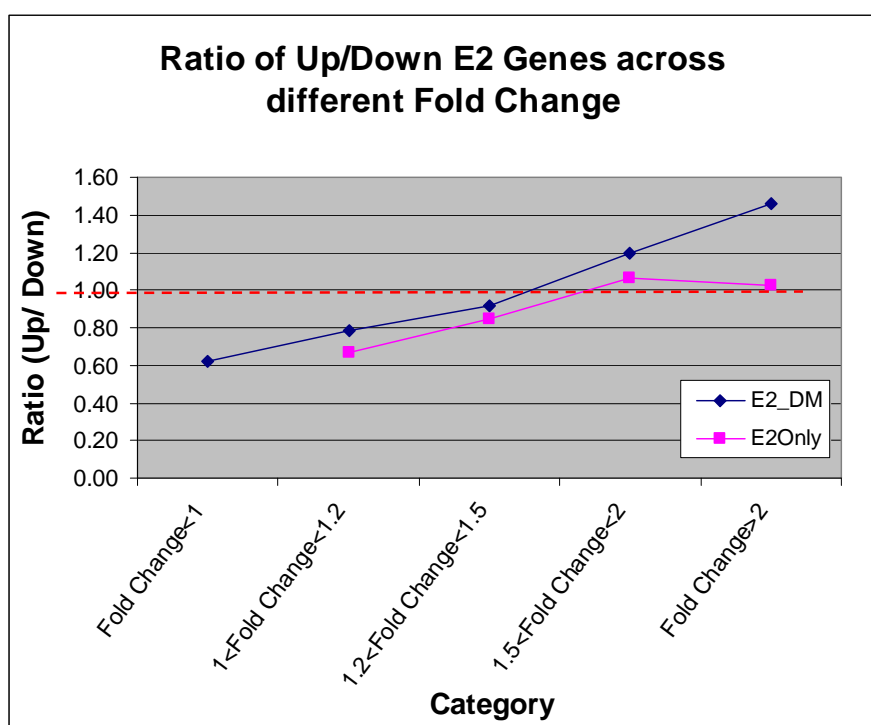


Figure 66 *Ratio of Up/Down E2 Genes across different fold change*

The blue line (E2_DM) refers to the 4205 binding sites (Overlap between binding sites in E2 and DMSO), i.e. basal occupied. The pink line (E2Only) refers to the 2277 binding sites appeared up on E2 treatment, i.e. without basal occupancy. Interestingly, the ratio of up-regulated genes over down-regulated genes increased for the E2_DM binding sites. Specifically, with fold change > 1.5, there were 20-50% more up-regulated genes than down-regulated genes. With fold change < 1, there were more

down-regulated genes, 67% more down-regulated genes compared to up-regulated genes, whereas for the E2Only binding sites, there were equal numbers of up and down-regulated genes at differential enrichment of ≥ 1.5 fold. In summary, when there was basal occupancy, there were more up-regulated genes with higher differential enrichment. On the other hand, when there was no basal occupancy, there were similar numbers of up-regulated and down-regulated genes in higher fold change. By Wilcoxon signed rank tests for all the difference in ratios between E2_DM and E2Only binding sites, the p-value of difference between E2_DM and E2Only was 0.03125, indicating that the difference was significant. The above finding postulates that basal occupancy plays a role in the regulation of genes.

4.5 Higher occurrence of ERE in E2-induced binding sites associated with higher regulated genes and higher binding sites fold change

The above categories of binding sites were also correlated to the ERE consensus sequence with up to 2 mismatches and the average number of regulated genes. The results are shown in Figure 67 and Figure 68 below.

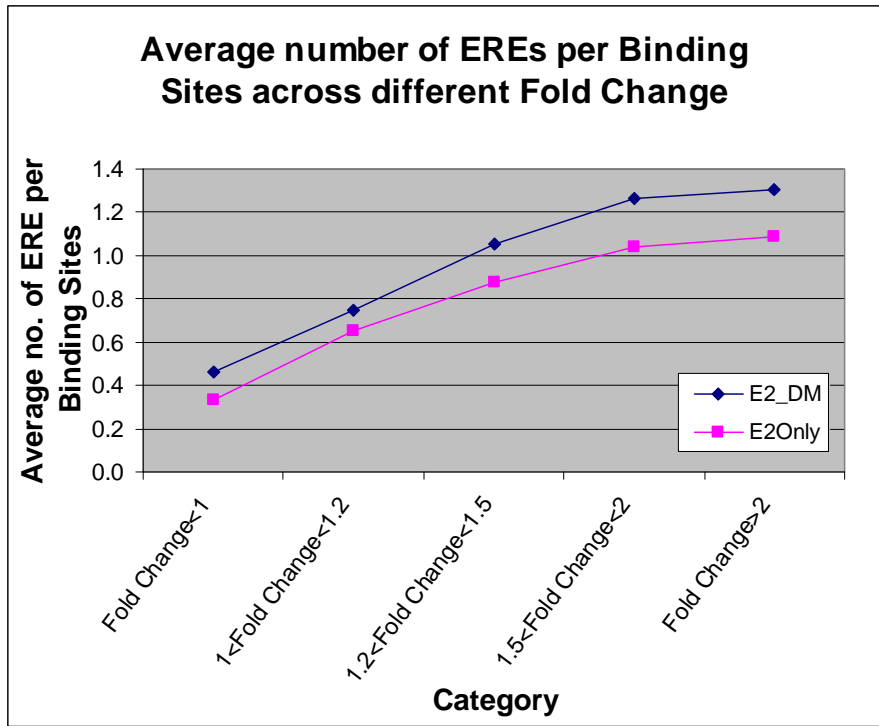


Figure 67 Average number of ERE per Binding Sites across different fold changes

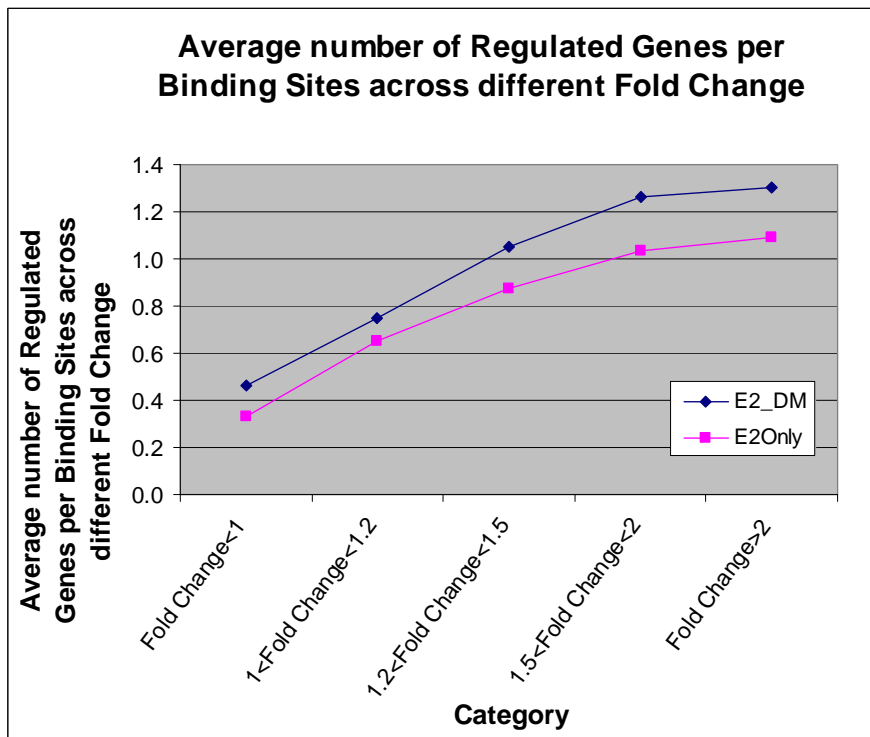


Figure 68 Average number of regulated genes per Binding Sites across different fold change

All binding sites with or without basal occupancy showed the same trend that with increasing fold change, both the average number of ERE per binding site (Figure 67) and regulated genes (Figure 68) increased. The above observations lead to the hypothesis that more genes were up-regulated when there was higher ER occupancy, i.e. greater fold change. Higher presence of EREs in turn associates with greater fold change. The finding above tallied with the report from the literature that the presence of ERE are commonly associated with up-regulated genes.

4.6 Modulating effects of SERMs on gene expression

Expression values indicated in this thesis are all with reference to DMSO unless it is stated otherwise (Illustration 5). Figure 69 showed the heatmap on the E2-induced genes with the corresponding gene expressions altered by SERMs. In order to show the relative effects of SERMs on E2, the changes from E2 to SERMs' expression values are depicted in Figure 70 (This is the only time the reference is made to E2 instead of DMSO). In general, SERMs had the effects on attenuating the intensity of either induced or repressed genes. For some down-regulated genes, effects of SERMs not just attenuated but reversed to up-regulation. The above observation similarly applied to some up-regulated genes, see illustration below.

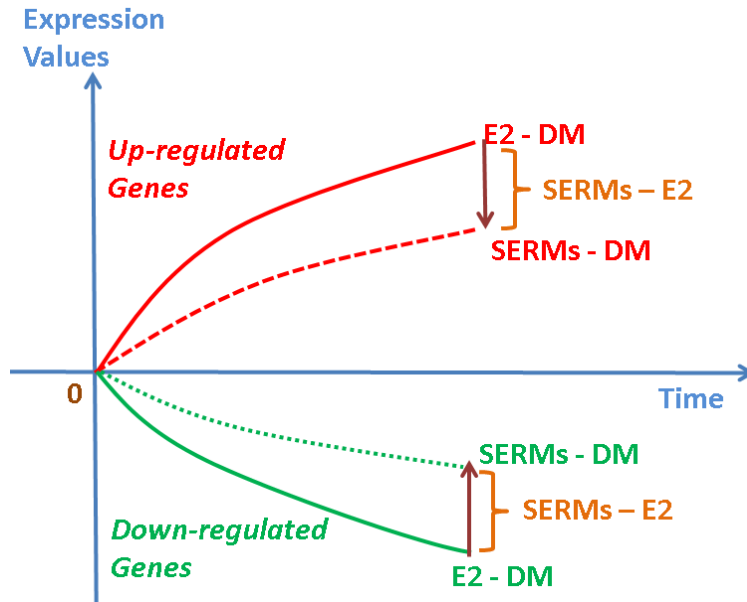


Illustration 5 Expression values of E2/SERMs with reference to DM or E2

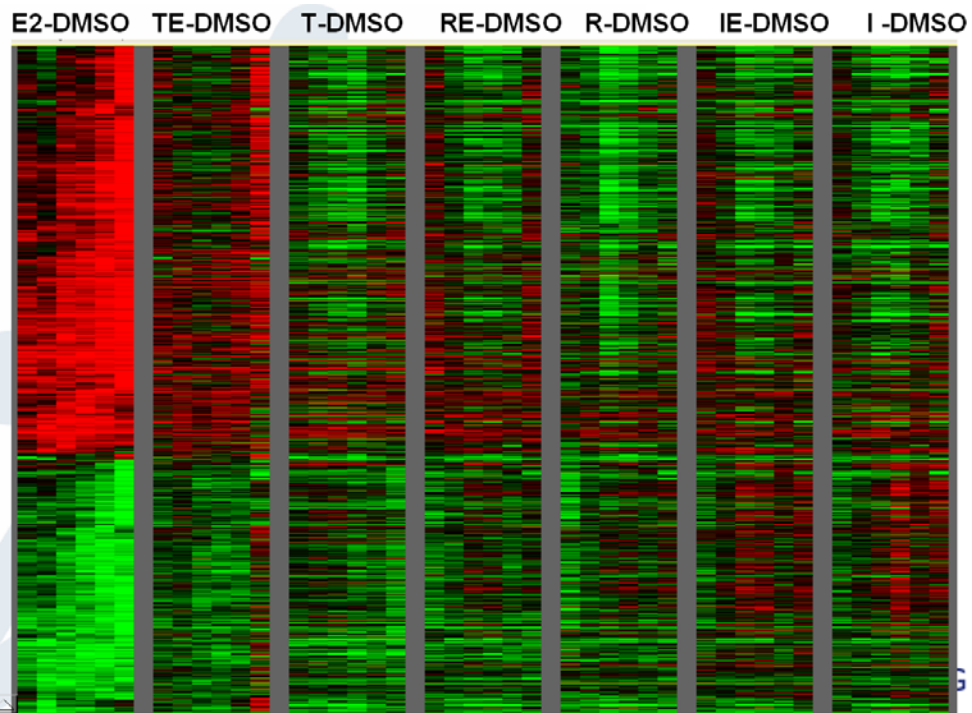


Figure 69 Effects of SERMs on E2-regulated genes in MCF-7

DM indicated the expression value in DMSO condition while E2 and SERMs indicated the change in expression with respect to DMSO. In each treatment condition, the time points were arranged in order from 3, 6, 9, 12, 24 and 48 hours.

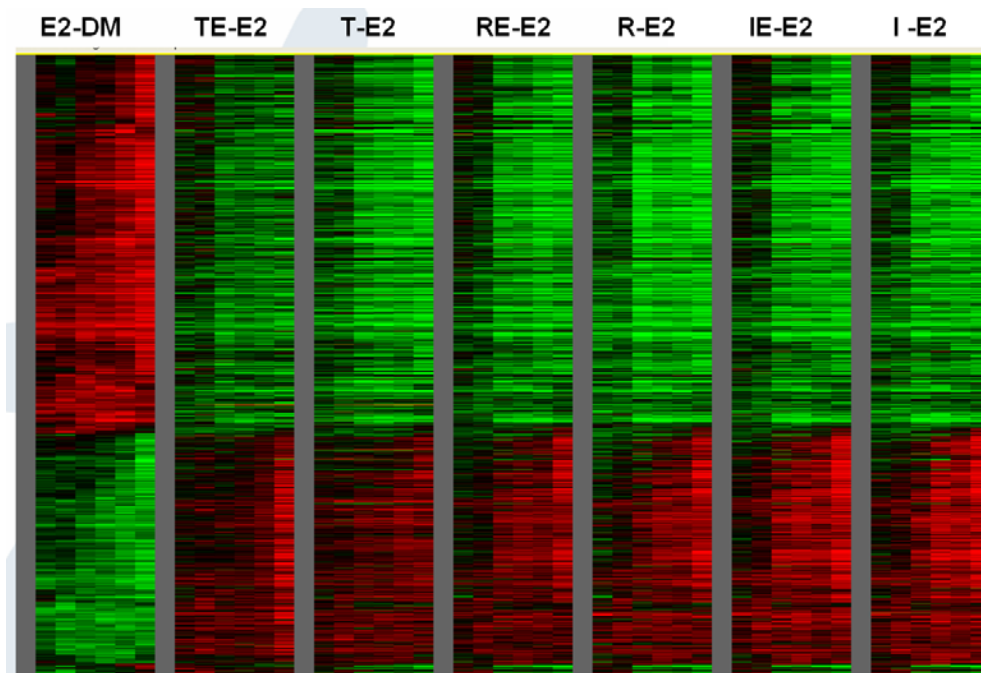


Figure 70 Effects of SERMs on E2-regulated genes in MCF-7 with reference to E2

In each treatment condition, the time points were arranged in order from 3, 6, 9, 12, 24 and 48 hours.

Global modulating effect of SERMs

This section examined the modulating effects of SERMs on gene expression induced by E2 treatment. Using the data analysis outlined in section 4.1, regulated genes were similarly defined for SERMs treatment. Table 20 summarises the number of probesets and genes expressed in each treatment using DMSO as the control. The total number of regulated genes in E2 was 3098, of which 1886 genes were up-regulated and 1212 genes were down-regulated. There were more up-regulated genes than down-regulated genes in E2 treatment. On the other hand, there were many more down-regulated genes than up-regulated genes in SERMs treatment which is expected due to SERMs' antagonizing properties. Intriguingly, besides modulating E2-regulated genes, SERMs also regulate their own unique set of genes. This may

partially explain why some SERMs treatments like TE, T, R and I treatments had larger total number of regulated genes than E2 treatments.

Table 20 Probesets or genes enriched in each treatment w.r.t DMSO

w.r.t DMSO (E2/SERMs - DMSO) in MCF-7						
Treatment	Probesets			Gene Symbol		
	Total	Up	Down	Total	Up	Down
E	3945	2450	1495	3098	1886	1212
TE	5361	3338	2023	4028	2548	1480
T	4964	1225	3739	3810	961	2849
RE	2930	658	2272	2330	550	1780
R	4734	968	3766	3466	774	2692
IE	3181	765	2416	2504	664	1840
I	4021	1356	2665	3085	1146	1939

The barchart in Figure 71 shows that both E2 and TE have higher proportion of up-regulated genes which reflect the agonistic property of E2 while the remaining SERMs or combination of SERMs+E2 have a high proportion of down-regulated genes that reflect the antagonistic property of SERMs, see Figure 71. Figure 72 shows the suppression of E2 genes in different SERMs.

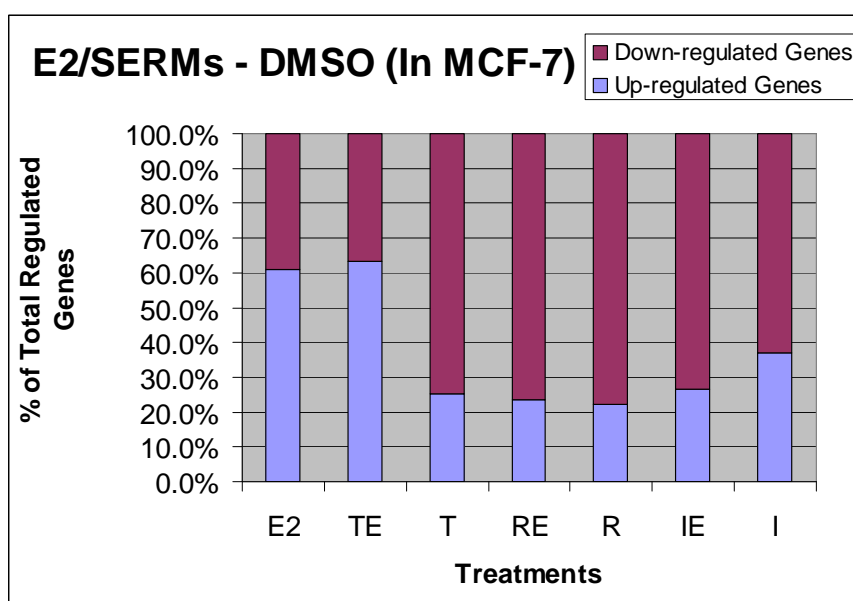


Figure 71 E2/SERMs – DMSO (In MCF-7)

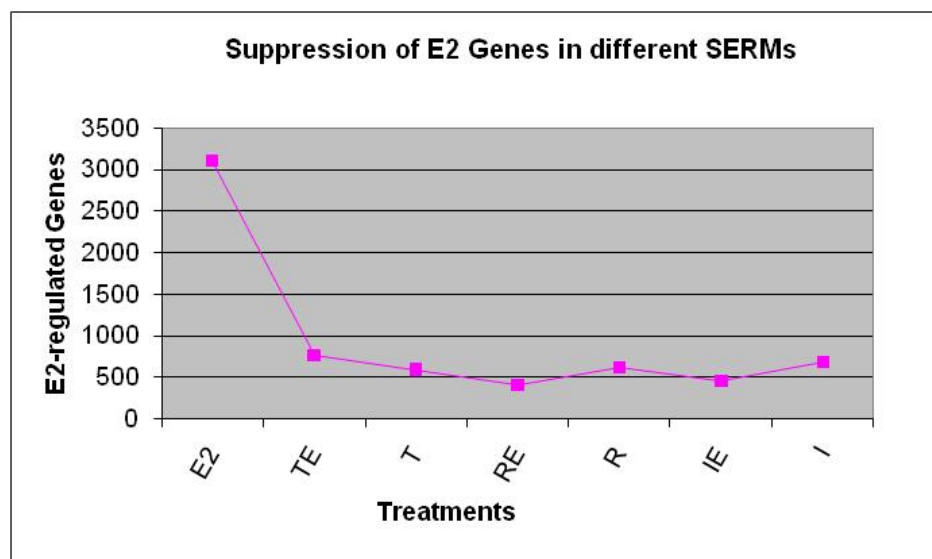


Figure 72 *Suppression of E2-regulated Genes in Different Treatments*

In order to quantify the changes better in gene regulation in MCF-7, the actual numbers of up-regulated and down-regulated genes were computed across individual time points in different treatments (Figure 73). Genes responsive to E2 treatment accumulated in number with time and were strongest at 48 hours. There were fluctuations in number of down-regulated genes for Tamoxifen, Raloxifene and ICI, which might indicate a mixture of primary and secondary effect of targeted genes. On the other hand, the regulated genes in E2 treatment did not have any fluctuation which suggested mainly primary targeted genes were involved. Taken together, SERMs appear to induce an initial wave of gene down-regulation.

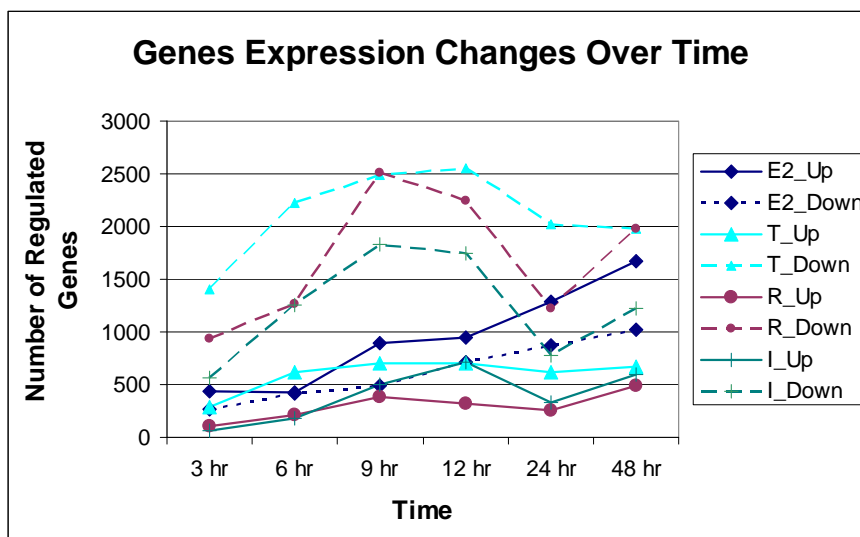


Figure 73 Summary of expression changes in terms of genes

4.7 Differential trends of SERMs modulation on E2 up-regulated or down-regulated genes

Both the binding sites and gene expressions profiles were further correlated to explore their relationship and roles in gene regulation with additional criteria of E2-regulated genes within 5kb of estrogen receptor binding sites. There were 216 E2-regulated genes fulfilled the above conditions and the corresponding heatmap had been constructed (Figure 74).

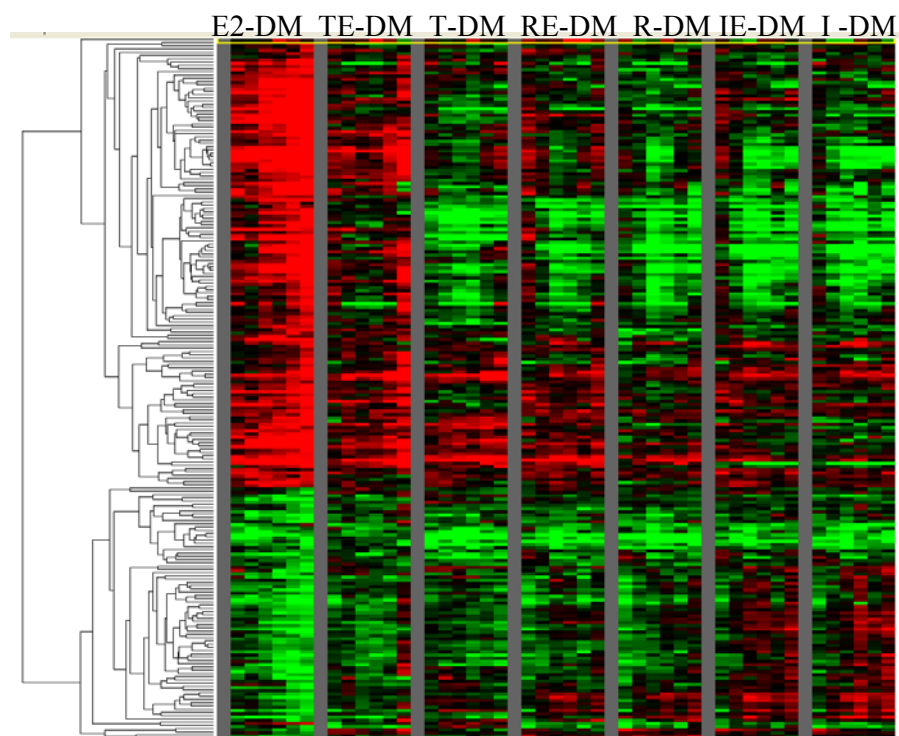


Figure 74 Gene expression profiles within 5kbs of E2 Binding Sites

In general, SERMs attenuated or reversed the magnitude of response of genes; up-regulated genes in E2 were suppressed or down-regulated while the down-regulated genes were also suppressed or up-regulated (Figure 74). For E2 up-regulated genes, the degree of suppression was the greatest in R, I and IE treatment whereas the TE has the lowest degree of suppression (Figure 75A). In general, the E2 up-regulated genes appear suppressed by SERMs for only 9-12 hours. Thereafter, . For E2 down-regulated genes, the degree of suppression was the greatest in I and IE treatment whereas T has the lowest degree of suppression (Figure 75B). In general, down-regulated genes exhibit a positive increasing trend for the change affected by SERMs.

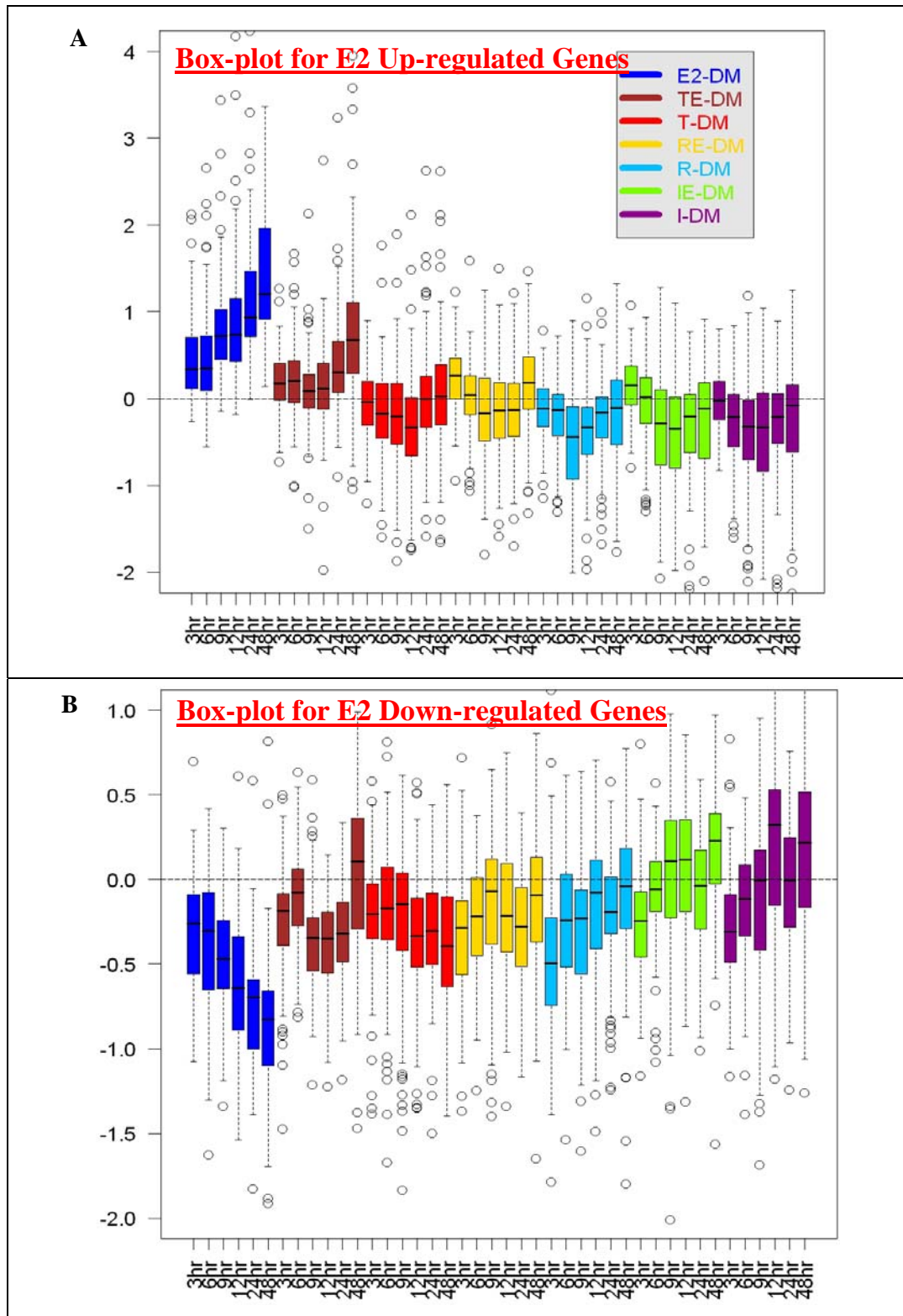


Figure 75 Boxplot for gene expression profiles within 5kb of E2 binding sites

A: Up-regulated genes in MCF-7 in E2 treatment and corresponding gene expressions affected by SERMs; B: Down-regulated genes in MCF-7 in E2 treatment and corresponding gene expressions affected by SERMs

With the above selection of genes, the modulating patterns of genes were classified into different groups. There were a number of distinct groups of E2 genes with different profile of SERMs' expression levels. Four distinct groups of profile of SERMS were selected and their gene ontology was studied using Ingenuity Pathway Analysis Software (<http://www.ingenuity.com/>). The four groups were 1) E2-up; SERMs-down, 2) E2-down; SERMs-up, 3) E2-up; SERMs-up and 4) E2-down; SERMs-down. For the first two groups where SERMs acts as a strong antagonists, the molecular and cellular functions have shown cellular growth and proliferation and cellular development which are expected of SERMs' action as being anti-proliferative as opposed to E2 ligand. Cellular proliferation will involve growth factors and cellular development involves cytokines as signaling proteins. Since the osteoclast and the osteoblast are targeted by estrogen putatively, this may confirm a unifying mechanism of estrogen actions in bone. For the last two groups where SERMs acts as weak antagonists or strong agonists like E2, the molecular and cellular functions are more on cancer and cellular movement. E2 and Tamoxifen have been shown inducing cytoskeletal remodeling and migration (Acconcia, Barnes et al. 2006).

4.8 Discovery of unique novel genes to SERMs and E2, exclusive of one another

Comparing the regulated genes between E2, T, R and ICI, unique genes were found pertaining to each treatment. For example, E2 unique genes are responsive only in E2 and not in any of T, R and I. The directionality of regulation was not considered in doing the intersection of regulated genes. There were 1450 unique E2 genes, 1368 unique T genes, 824 unique R genes and 658 unique I genes.

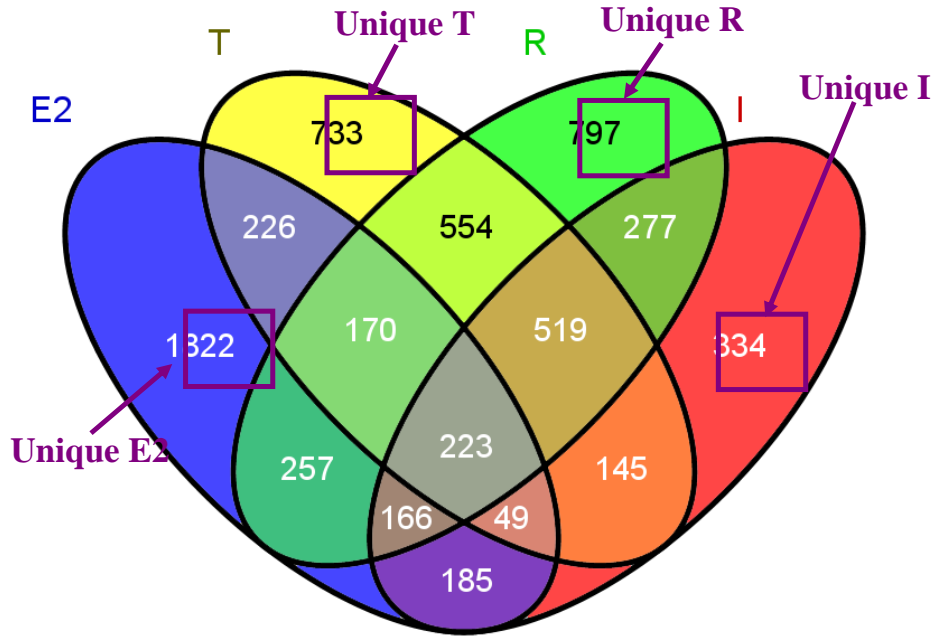


Figure 76 Venn-diagram for intersections between E2, T, R and I

The heatmap for the unique genes E2, T, R and I are shown in Figure 77, Figure 78, Figure 79 and Figure 80 respectively.

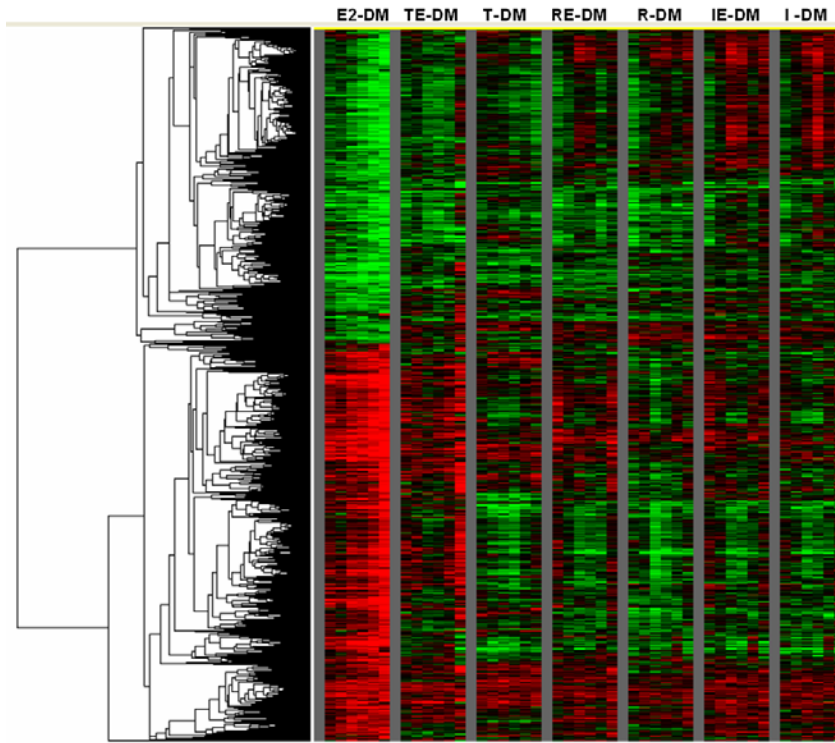


Figure 77 Heatmap for unique genes in E2

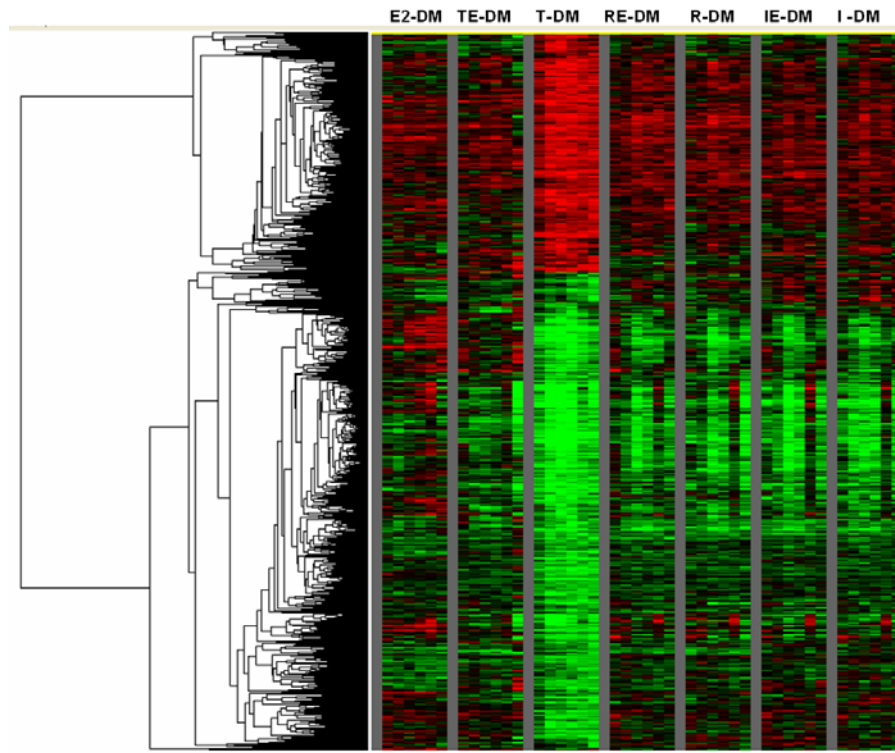


Figure 78 Heatmap for unique genes in T

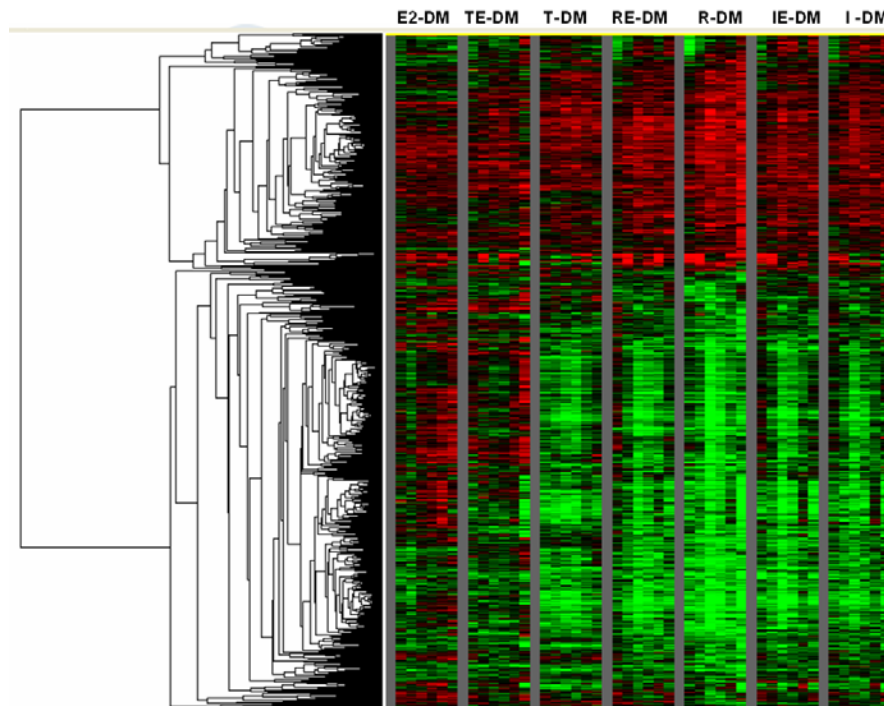


Figure 79 Heatmap for unique genes in R

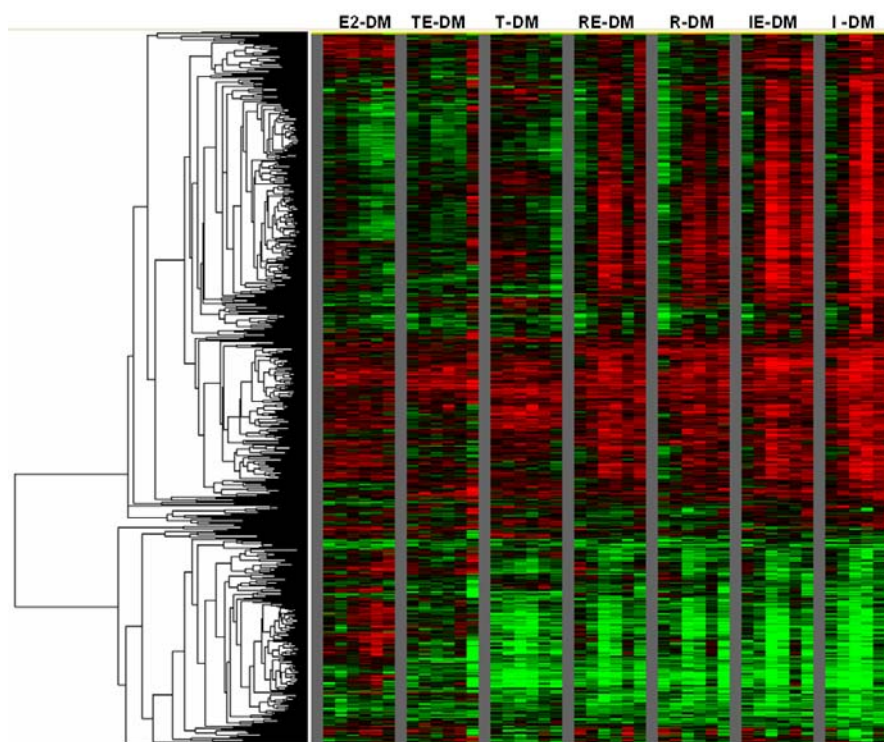


Figure 80 Heatmap for unique genes in I

Using gene ontology software, the unique biology of for the above groups of the unique genes was deciphered. The molecular and cellular functions for unique E2 genes were cell death, cell cycle, cellular development, cellular growth and proliferation, DNA replication, recombination and repair. The molecular and cellular functions for unique T were gene expression, lipid metabolism, small molecule biochemistry, cell cycle, vitamin and mineral metabolism. The molecular and cellular functions for unique R were gene expression, cell death, cellular function and maintenance, cellular assembly and organization, DNA replication, recombination and repair. The molecular and cellular functions for unique I were cell signalling, cellular development, cellular movement, post-translational modification, DNA replication, recombination, and repair. Common molecular and cellular functions between E2 and T were cell cycle; E2 and R were cell death; E2 and I were cellular development, DNA replication, recombination and repair.

4.9 ER tends to remain occupied across SERMs conditions for up-regulated genes

The binding sites occupancy at those E2-binding sites within 5kb of E2-regulated genes were examined across SERMs and listed in Table 21 and Table 22 for down-regulated and up-regulated genes respectively. From the 2 tables, ER was observed to remain present across different SERMs treatments for up-regulated genes.

Subsequently, the 2 tables were converted into trend lines depicting the proportion of binding sites occupied for individual SERMs treatment. The binding sites had been ordered according to the distance of binding site to TSS of genes. Keeping the same order, the proportion of binding sites occupied was calculated up to the site taken into account. For E2 up-regulated genes within 5kb of an ER binding site, all treatments except IE and RE had high proportion of binding sites occupied (> 60%) (Figure 81). IE and RE had low proportion of binding sites occupied (<35%).

Table 21 Binding sites detection across different treatments and down-regulated E2 genes

Position	E	D	TE	T	RE	R	IE	I	Distance from TSS	Gene Symbol	Regulation
chr3_114737988_114738841	P	P	P	P	P	P	P	P	4232	SIDT1	Down
chr1_112967955_112968614	P	P			P				3621	PPM1J	Down
chr12_54637517_54637731	P								3468	SILV	Down
chr4_6815720_6816334	P	P							2390	S100P	Down
chr16_30023120_30023849	P		P						-148	GDPD3	Down
chr16_31049520_31050104	P							P	-444	PRSS8	Down
chr7_72688028_72690097	P	P	P					P	-781	CLDN4	Down
chr5_355807_356136	P							P	-1319	AHRR	Down
chr12_108939839_108940183	P		P					P	-1691	FLJ40142	Down
chr17_35096087_35096301	P								-1724	ERBB2	Down
chr16_82766883_82767517	P	P	P		P	P			-1761	TAF1C	Down
chr16_83615353_83617845	P	P	P					P	-2311	KIAA0513	Down
chr5_354616_355115	P							P	-2425	AHRR	Down
chr15_88126246_88126898	P	P	P				P		-2558	ANPEP	Down
chr7_72686669_72687188	P								-2915	CLDN4	Down
chr6_3663379_3663736	P	P	P	P		P		P	-4295	C6orf145	Down
chr4_6808811_6809395	P								-4534	S100P	Down

Table 22 Binding sites detection across different treatments and up-regulated E2 genes

Position	E	D	TE	T	RE	R	IE	I	Distance from TSS	Gene Symbol	Regulation
chr12_14930387_14931027	P	P	P	P		P	P	P	4614	MGP	Up
chr16_86425368_86426082	P								4595	SLC7A5	Up
chr21_42668809_42670053	P	P	P	P		P		P	4363	TMPRSS3	Up
chr21_42659426_42660080	P	P	P	P	P	P	P	P	4294	TFF1	Up
chr1_133335773_133333542	P	P	P	P	P	P	P	P	3743	UBE2T	Up
chr3_129285783_129286507	P	P	P	P		P	P	P	3645	RUVBL1	Up
chr7_43574655_43574825	P								3229	BLVRA	Up
chr6_1558462_1559162	P	P	P	P	P	P		P	3133	FOXC1	Up
chr17_35590092_35590991	P	P	P	P				P	2403	RAPGEFL1	Up
chr11_108043181_108043619	P	P	P	P	P	P	P	P	2375	DDX10	Up
chr6_16238685_16239485	P	P	P	P				P	1790	MYLIP	Up
chr16_85998866_85999155	P		P						1647	ZCCHC14	Up
chr5_43639556_43639900	P					P		P	1147	NNT	Up
chr19_40223787_40224080	P				P			P	684	HPN	Up
chr3_133862288_133862724	P			P				P	663	UBE1DC1	Up
chr11_71176478_71176649	P								333	FLJ10661	Up
chr11_74811292_74811571	P	P	P			P	P	P	319	FLJ33790	Up
chr12_100593131_100595228	P	P	P	P		P		P	295	CHPT1	Up
chr15_60726862_60727296	P	P				P		P	278	TLN2	Up
chr6_152220412_152220996	P								-95	ESR1	Up
chr20_42776270_42777269	P	P	P	P				P	-529	WISP2	Up
chr1_143215501_143216460	P	P	P	P		P		P	-789	PDZK1	Up
chr21_45757581_45758092	P					P		P	-1220	SLC19A1	Up
chr2_11620976_11626117	P	P	P	P	P	P	P	P	-1293	GREB1	Up
chr18_73446765_73449266	P	P	P	P				P	-1314	SIN3A	Up
chr1_176929057_176929856	P	P	P	P	P	P		P	-1642	LHX4	Up
chr4_38867278_38867784	P	P	P	P				P	-1693	KLHL5	Up
chr16_15036997_15037339	P	P	P			P		P	-2049	NTAN1	Up
chr20_52198661_52201967	P	P	P	P			P	P	-3080	CYP24A1	Up
chr16_15035688_15036142	P								-3302	NTAN1	Up
chr14_74811741_74812183	P	P	P	P		P	P	P	-3321	ECS	Up
chr3_151937598_151938629	P	P	P	P	P	P	P	P	-3495	SIAH2	Up
chr10_104460207_104461044	P	P	P	P	P	P	P	P	-3662	SFXN2	Up
chr22_27511319_27511562	P	P	P	P					-3662	XBP1	Up
chr1_33143424_33143774	P					P	P	P	-4076	AK2	Up
chr11_100409714_100410585	P	P	P	P		P		P	-4163	PGR	Up
chr12_115429456_115429972	P	P	P	P		P		P	-4232	FLJ42957	Up

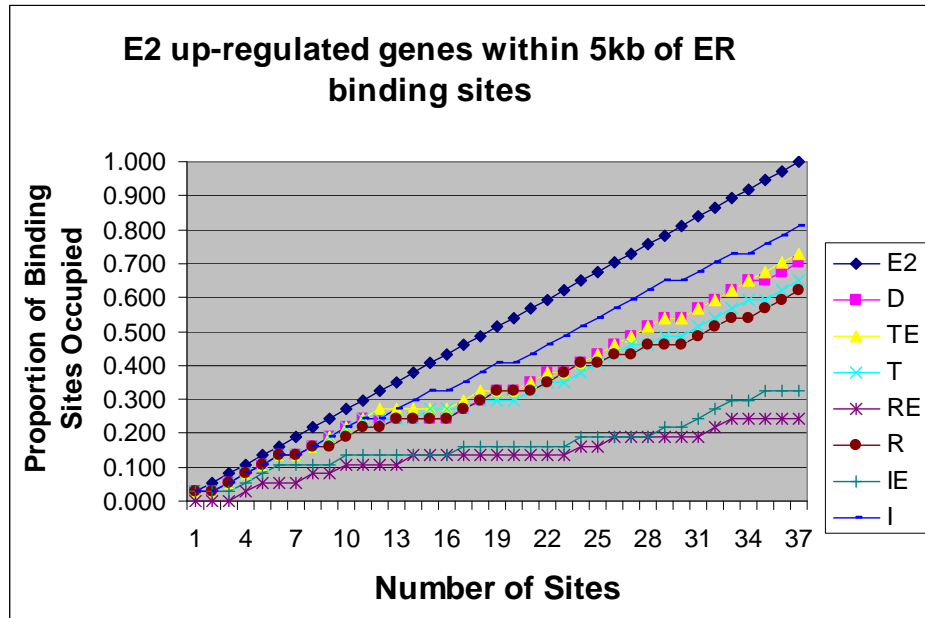


Figure 81 E2 up-regulated genes within 5kb of ER binding sites

For E2 down-regulated genes within 5kb of an ER binding site, all treatments except D, TE and I had low proportion of binding sites occupied (< 25%) (Figure 82). D and TE had only 50% proportion of binding sites occupied. E2 up-regulated genes would have higher proportion of binding sites that were still occupied by ER upon SERMs treatment than E2 down-regulated genes.

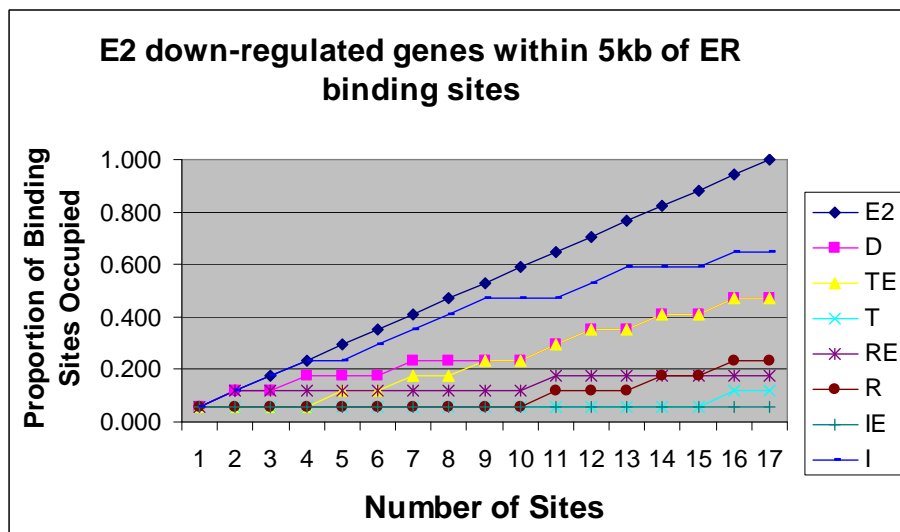


Figure 82 E2 down-regulated genes within 5kb of ER binding sites

Decision tree was used as the classifier for the 2 groups of regulated genes with the corresponding occupancy pattern in T, R and I (Figure 83). It was found that when T or R remains occupied, the probability that it is an up-regulated gene is very high (93% for T and 75% for R). From the decision tree, it can be observed that an up-regulated gene is likely associated when ER remained occupied in (1) T or (2) R when ER not occupied in T or (3) Not occupied in either T, R and I. Writing the above relationship in Boolean Algebraic form:

$U = T + T' R + T' R' I'$ where U denotes up-regulated genes; T denotes ER remains occupied in T treatment; T' denotes ER not occupied in T treatment; R denotes ER remains occupied in R treatment; R' denotes ER not occupied in R treatment; I' denotes ER not occupied in I treatment. For example, T' is complement to T, i.e. if $T = 1$, $T' = 0$ and vice versa.

Using the rules in Boolean Algebra, $A + A' B = A + B$,

$$U = T + T'R + T' R' I'$$

Can be reduced to

$$U = T + R + I'$$

The final equation above states that the likelihood of up-regulated genes is due to the continued occupancy of ER in T or R treatment but not in I treatment. With reference to section 3.4, unique I binding sites had a very different de novo motif prediction from consensus ERE whereas unique T and R binding sites had a similar motif to ERE. As such, unique I motif might provide an explanation for the association of ER remained occupied in I treatment to the down-regulated genes.

$$T = ?$$

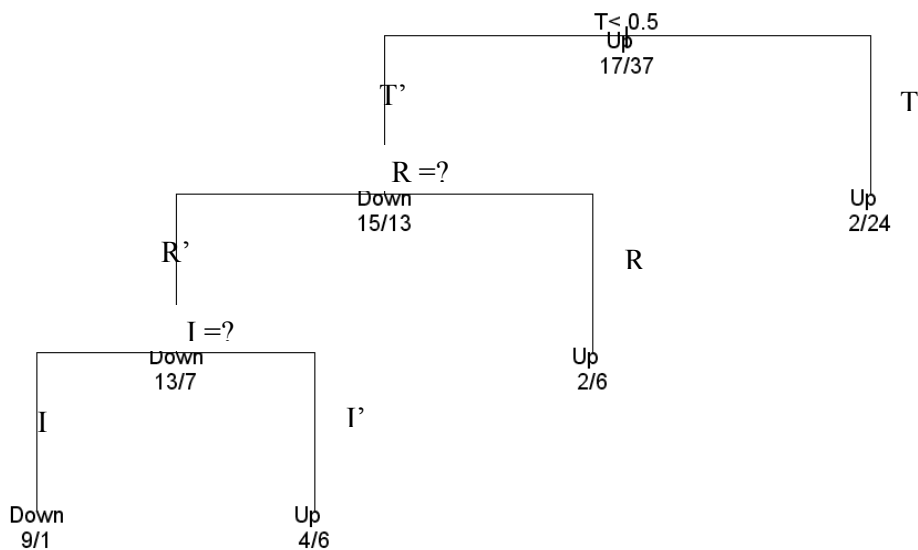


Figure 83 Decision tree for classifying up- and down-regulated genes

In summary, pattern of binding sites occupancy in E2 and across SERMs condition can predict gene regulation. When the ER continues to bind across different SERMs treatment, the gene regulation is likely to be for up-regulated genes, especially for T or R treatment.

4.10 SERMs alter ER's spatial binding characteristics in promoter-context and cell environment

We introduced in the beginning of this thesis that when ER complexed with E2, the conformation was different from that of when complexed with SERMs, i.e. differential displacements of the carboxy-terminal helix (H12). Here we hypothesize that the different conformations of ER-SERMs change the binding characteristics of ER. It is also hypothesized that ER-SERMs complexes have a lower binding affinity than ER-E2 complex that they preferentially select those binding sites with chromatin configuration that are more opened and accessible. As such, we investigated the profiles of both FAIRE and H3K4Me1 signals for the common and unique binding sites between E2 binding sites and those binding sites for Tamoxifen and Raloxifene.

Since ICI works by degrading the estrogen receptor, similar analysis was not carried out on ICI treatment.

The profiles of FAIRE signal for E2 and Tamoxifen, FAIRE signal for E2 and Raloxifene are shown in Figure 84 and Figure 85 respectively. Both FAIRE signals for unique Raloxifene and unique Tamoxifen binding sites showed that FAIRE signals were the strongest in DMSO condition while signals in E2 condition were the next highest. This implies that unique binding sites to either Tamoxifen or Raloxifene have the highest nucleosome free regions (Using t-test, p-value are 9.39×10^{-6} and 2.42×10^{-6} for Tamoxifen and Raloxifene respectively).

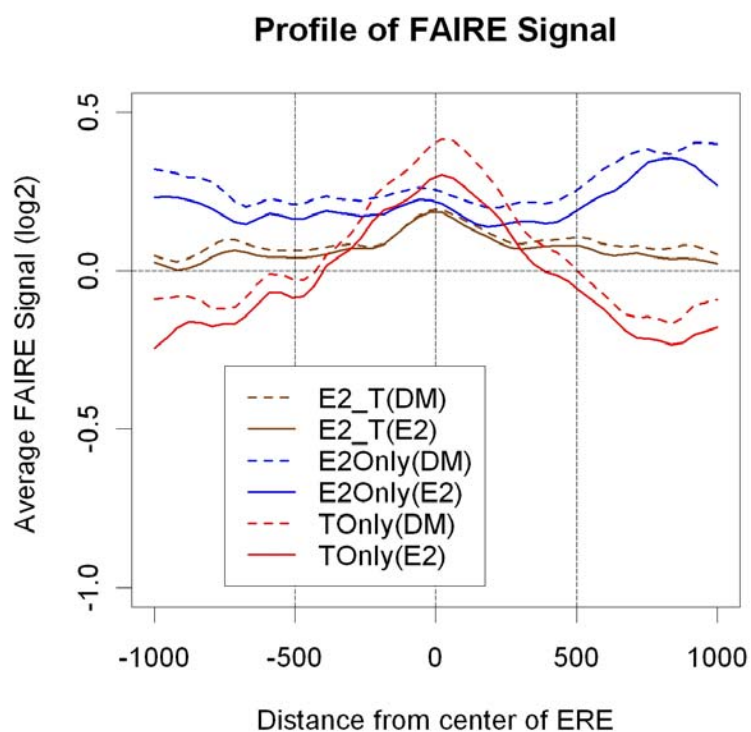


Figure 84 Profile of FAIRE signal for E2 and Tamoxifen

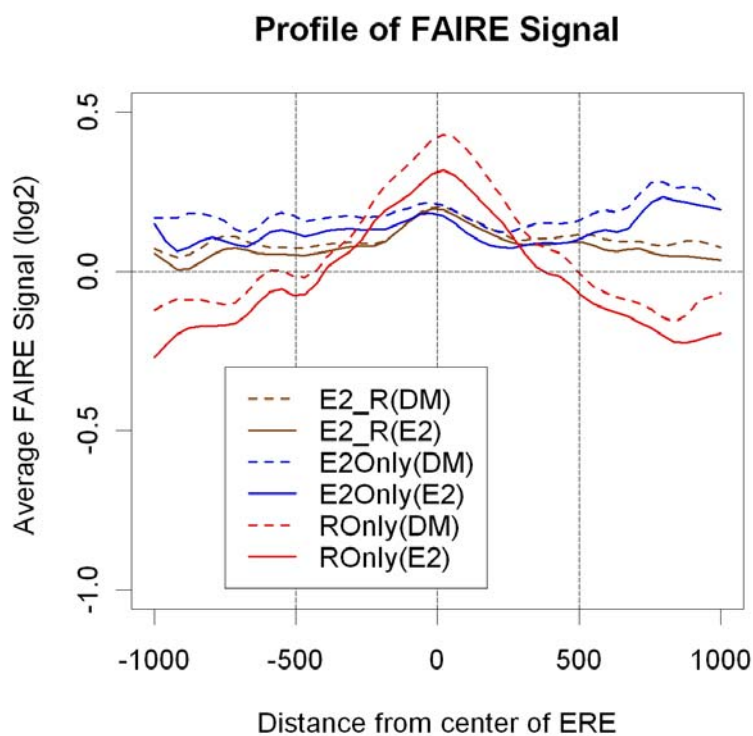


Figure 85 Profile of FAIRE signal for E2 and Raloxifene

Similarly, the profiles of H3K4Me1 signal for E2 and Tamoxifen, H3K4Me1 signal for E2 and Raloxifene are shown in Figure 86 and Figure 87 respectively. Both H3K4Me1 signals for unique Raloxifene and unique Tamoxifen binding sites showed that H3K4Me1 signals remained very high in both E2 and DMSO conditions (Using t-test, p-value are $4.25e-25$ and $5.0e-24$ for Tamoxifen and Raloxifene respectively). This may imply that unique binding sites to either Tamoxifen or Raloxifene do not function as enhancers as there H3K4Me1 signals remain high in the center of ERE.

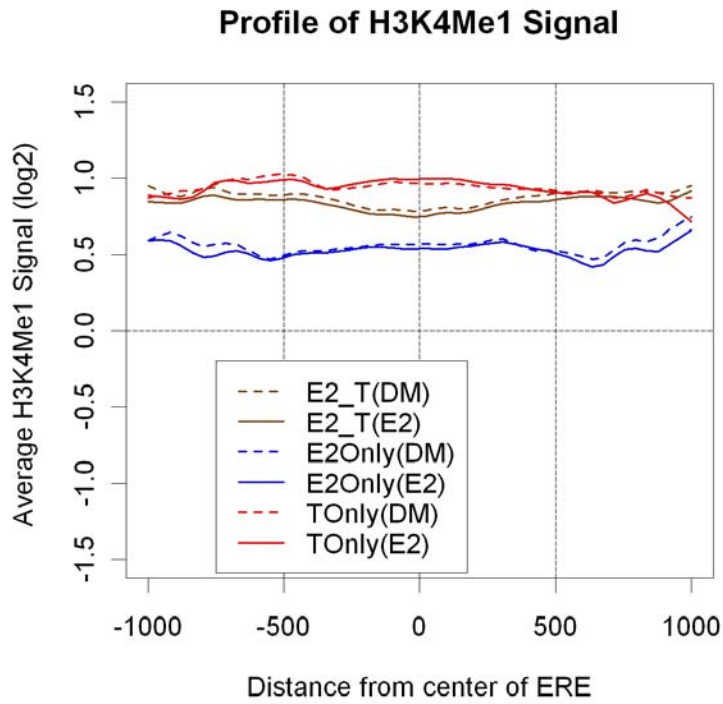


Figure 86 Profile of H3K4Me1 signal for E2 and Tamoxifen

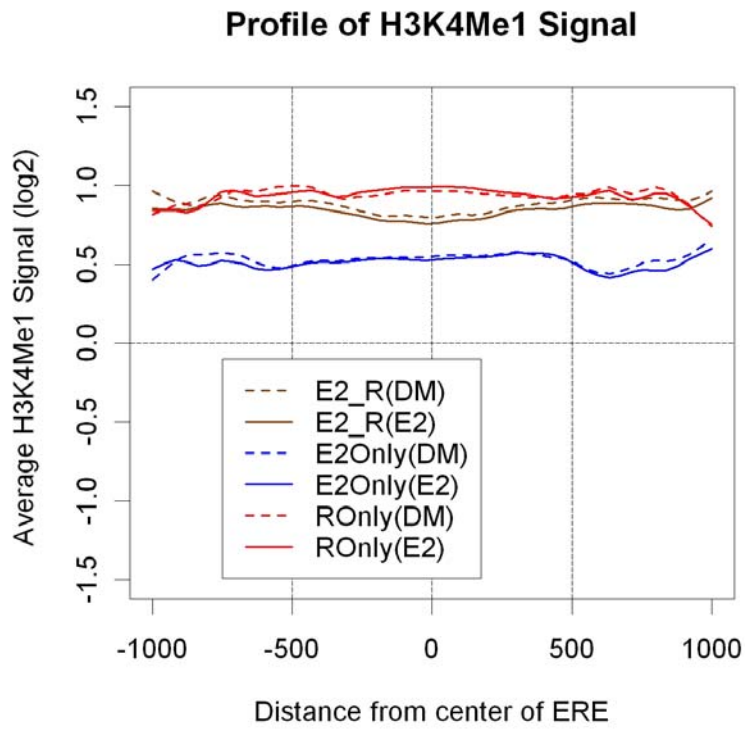


Figure 87 Profile of H3K4Me1 signal for E2 and Raloxifene

In summary, unique Tamoxifen or Raloxifene binding sites had distinctive FAIRE and H3K4ME1 profiles that suggest strongly SERMs alter ER's spatial binding characteristics.

4.11 Revealing Spatiotemporal Expression Profiles of ER-responsive Genes in Different Tissues Upon E2 And SERMs Treatments

Many factors contribute to tissue-specific effects such as DNA methylation, histone modifications, differential co-regulator recruitments and the relative abundances of various transcription factors. Here we examined the global expression profiles of two different tissues – breast and uterine tissues by using MCF-7 and Ishikawa cell lines. Ishikawa cell line was treated with the same drugs and time course i.e. expressions were measured for the same treatments and time course as that of MCF-7.

Using the data analysis outlined in section 4.1, the regulated genes in Ishikawa cell line under E2 and SERMs treatment were also obtained. Examining the number of regulated genes upon E2 and SERMs treatment across both MCF-7 and Ishikawa, there were much greater regulated genes in MCF-7 than Ishikawa cell line (Figure 88). For Ishikawa cell line, there were an exceptionally high number of genes induced in ICI treatment.

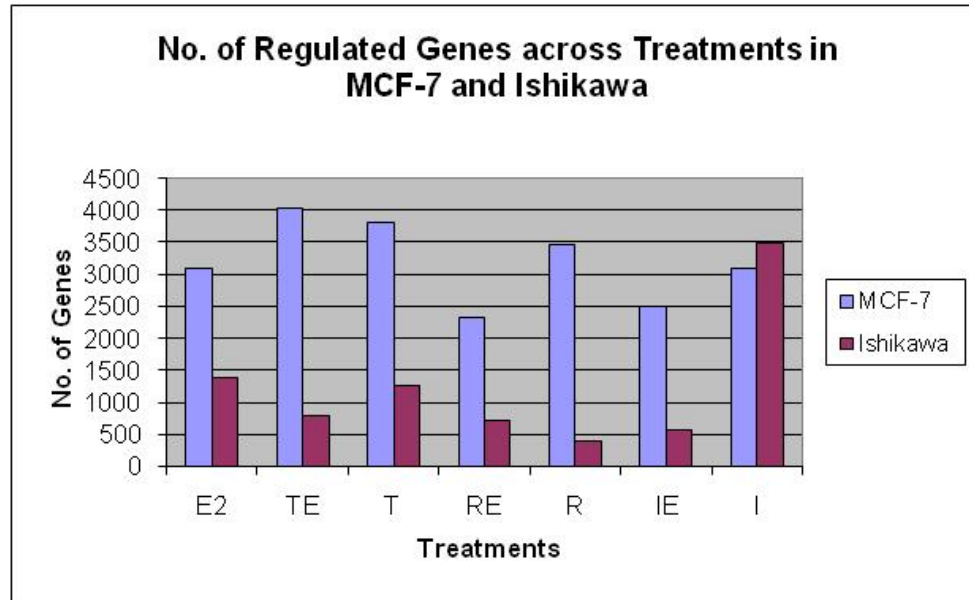


Figure 88 No. of regulated genes across treatments in MCF-7 and Ishikawa cell lines

While there were higher proportions of down-regulation in SERMs treatment for MCF-7, there were more up-regulated genes in SERMs treatment in Ishikawa cell line (Figure 89).

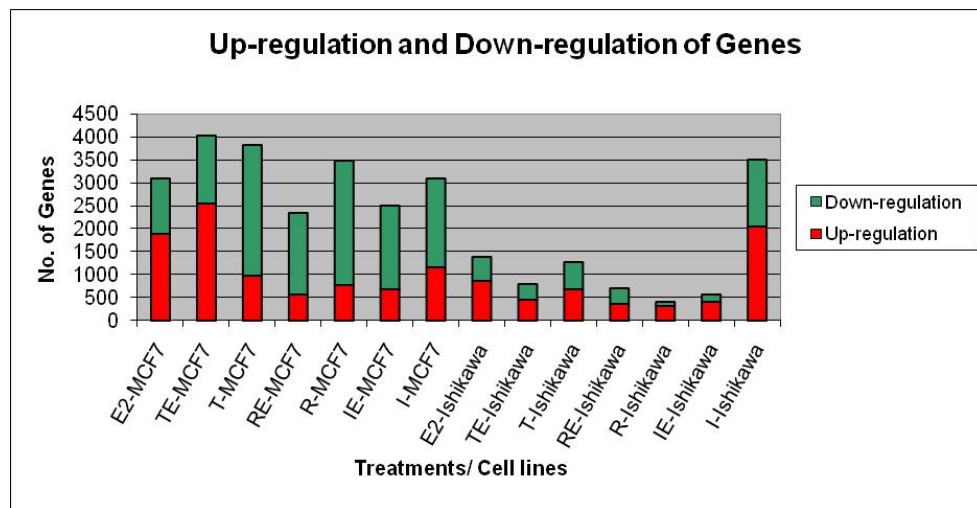


Figure 89 Up-regulation and down-regulation of genes across treatments in MCF-7 and Ishikawa cell lines

Next, we examined in detail the expression of Ishikawa regulated genes in MCF-7 cell line. The heatmap for the E2-regulated genes in Ishikawa (Figure 90) shows

that there are some oscillations in the expression for the up-regulated genes while the down-regulated genes are increasingly down-regulated with time.

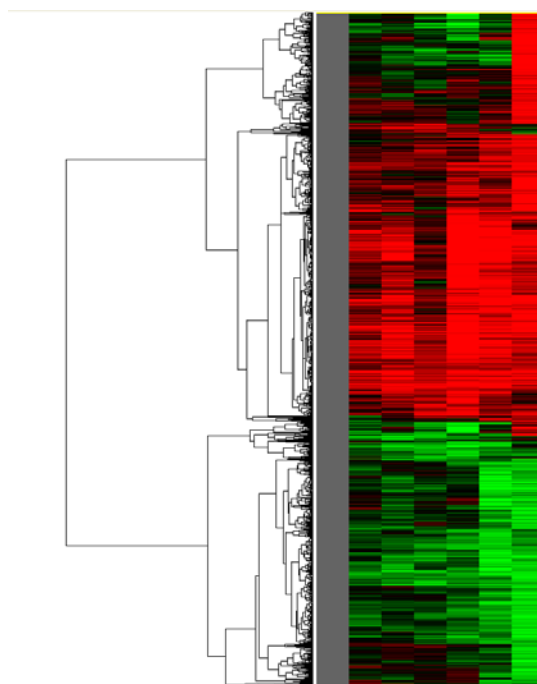


Figure 90 Heatmap for E2-regulated genes in Ishikawa

To investigate the effects of SERMs on the E2-regulated genes, the heatmap on the E2-regulated genes and the corresponding expression levels in SERMs treatments were constructed (Figure 91). The heatmap shows that SERMs decrease the magnitude of E2-repressed genes but the intensities of E2-induced genes are hardly antagonized by SERMs. The phenomenon seen in Ishikawa cell line was in great contrast to that seen in MCF-7 depicted in Figure 69. For example, up-regulation in MCF-7 was attenuated or even reversed it to down-regulation by SERMs instead of unchanged in Ishikawa; down-regulation of genes was also greatly reduced in MCF-7 than Ishikawa.

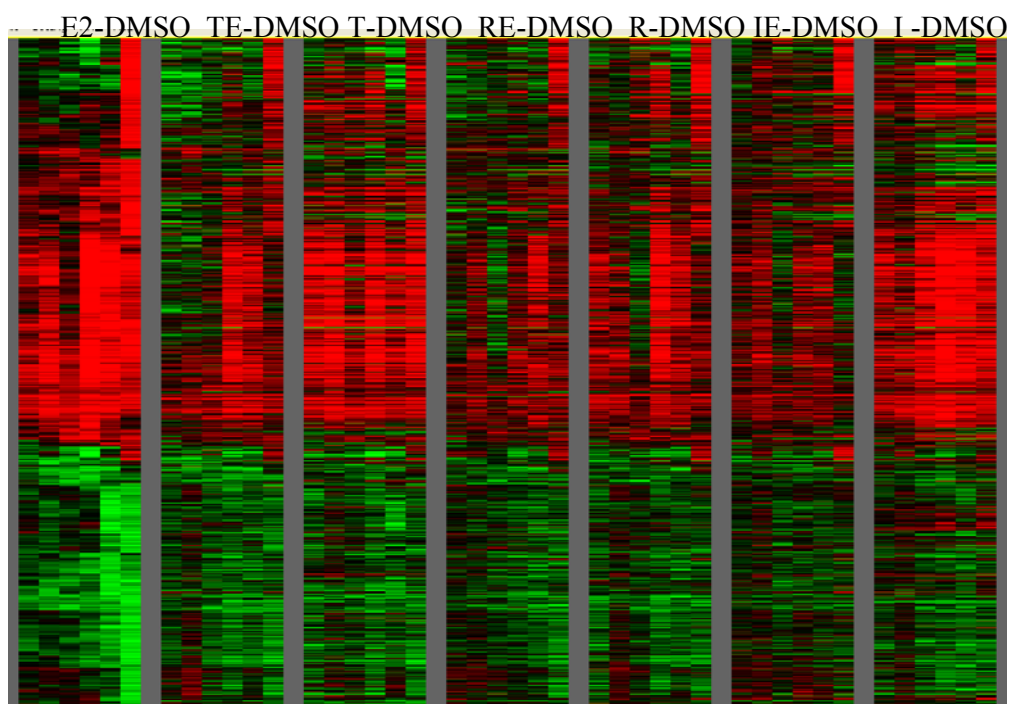


Figure 91 Effects of SERMs on E2-regulated genes in Ishikawa

Here we examined the change of E2-regulated genes in Ishikawa cell line on MCF-7 instead. The expression of genes regulated in Ishikawa cell line remained high for most of the genes in another cell line - MCF-7 (Figure 92). Majority of the up-regulated genes in Ishikawa continued to be up-regulated or have positive gene expression values in MCF-7 while about 50% of the down-regulated genes in Ishikawa were similarly down-regulated or had negative gene expression values in MCF-7. Figure 93A shows that up-regulation in MCF-7 also follows an increasing trend, albeit at a smaller magnitude unlike in Ishikawa cell line.

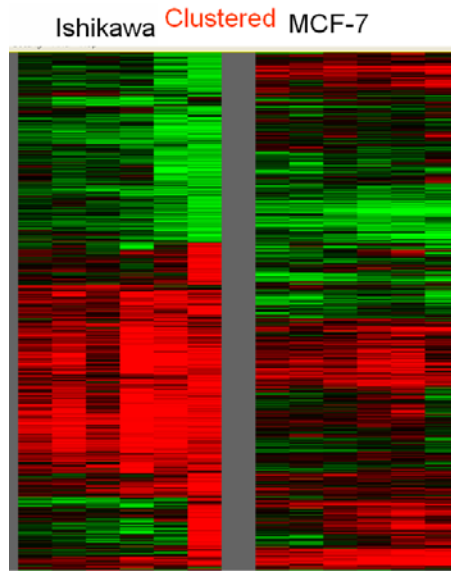
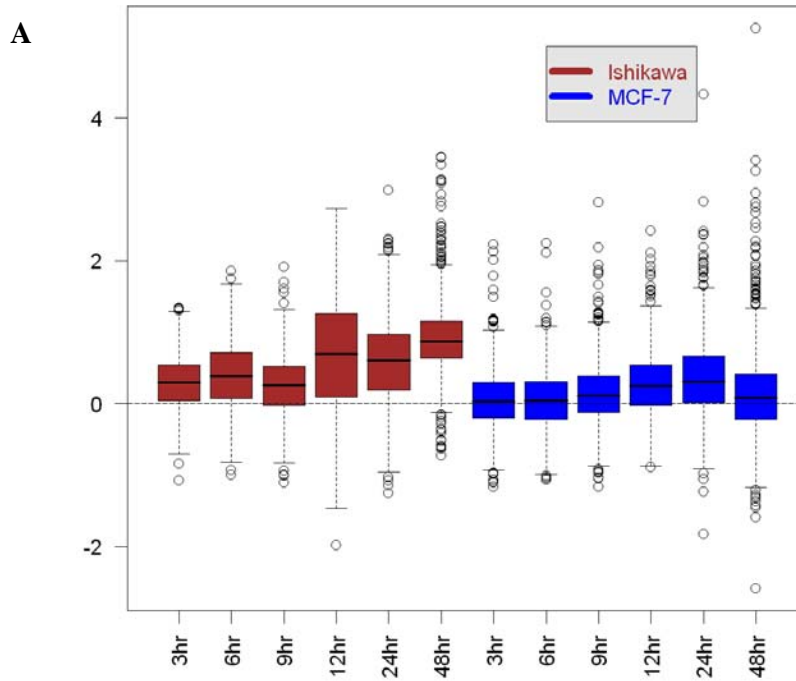


Figure 92 Tissue-specific effects shown on MCF7 on E2-regulated genes in Ishikawa cell line



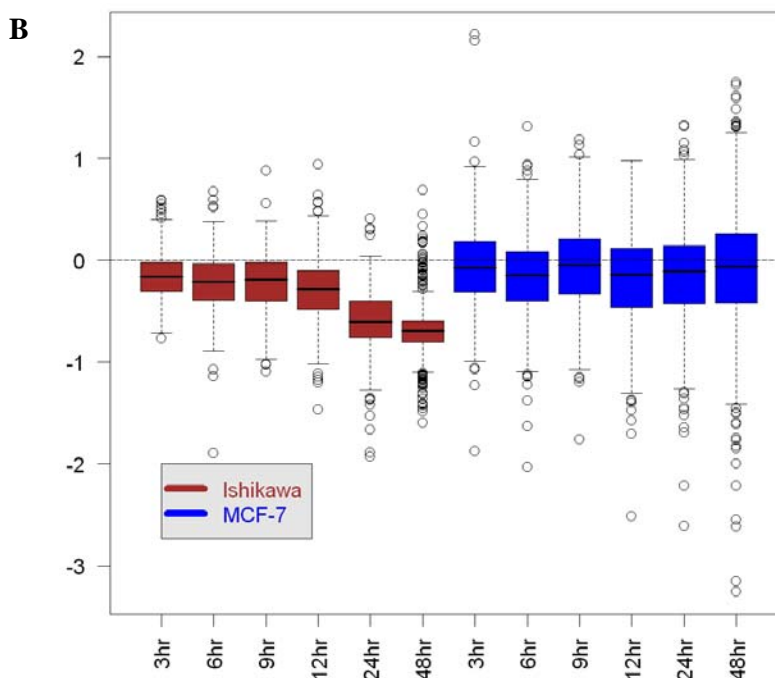


Figure 93 Tissue-specific effects shown on MCF-7 on E2-regulated genes in Ishikawa cell line (Boxplot)

A: Up-regulated genes in Ishikawa and corresponding gene expressions in MCF-7 cell line; B: Down-regulated genes in Ishikawa and corresponding gene expressions in MCF-7 cell line

The overlap in the E2-regulated genes in MCF-7 and Ishikawa cell lines was investigated. Only 10.9% of the genes responsive to E2 treatment in MCF-7 were also responsive in Ishikawa cell line for the same treatment (Figure 94). This small percentage of overlap in responsive genes in breast and uterine tissues was in accordance to the highly tissue-specific global gene expressions properties.

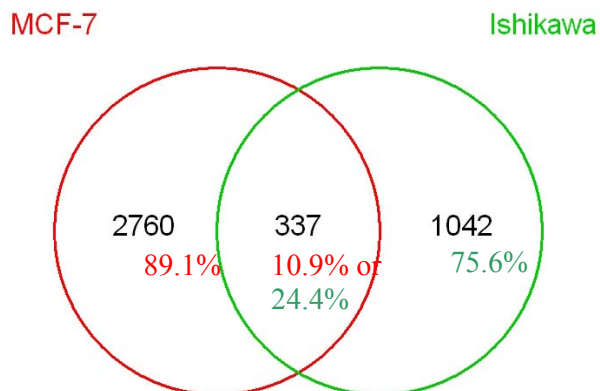


Figure 94 Intersection between MCF-7 and Ishikawa cell lines in E2 regulated genes

Similarly, we studied the overlap in regulated genes responsive to different SERMs treatment between MCF-7 and Ishikawa cell lines. The overlap in terms of percentages of genes in MCF-7 was small for all SERMs (Figure 95) as in E2 treatment (Figure 94). The only exception was in I treatment that the overlap was 33.4% in terms of regulated genes in MCF-7. This was because there were an exceptionally high number of I-regulated genes in Ishikawa, thus leading to the higher percentage of overlap.

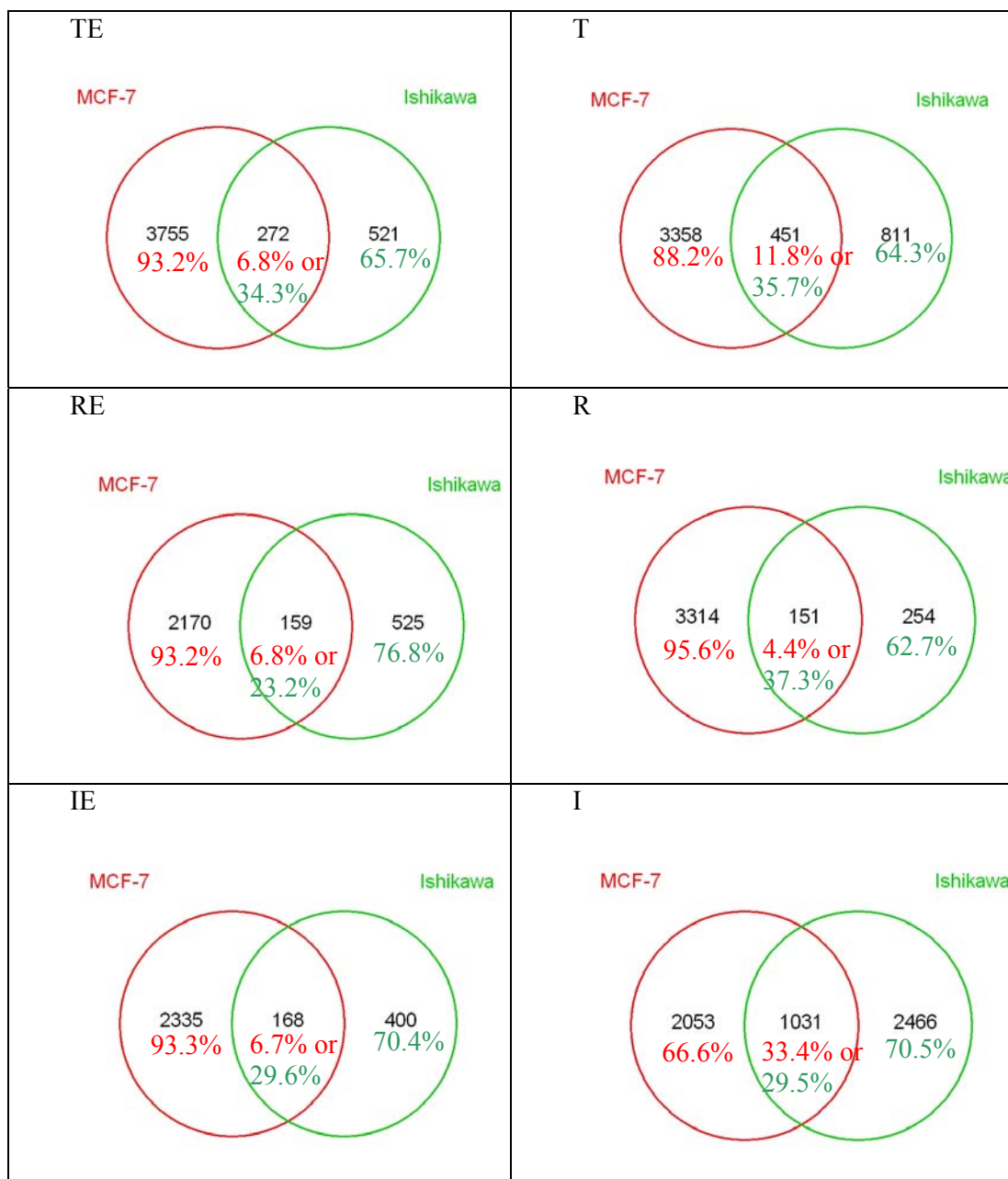


Figure 95 Intersection between MCF-7 and Ishikawa cell lines in SERMs regulated genes

Since we found that the overlap in the regulated genes between MCF-7 and Ishikawa cell line was low due to different tissues biology, we next examined in detail the corresponding changes in MCF-7 regulated genes in Ishikawa cell line. Figure 96 shows the clustered gene expression profile of MCF-7 and Ishikawa genes together. It

was interesting that the gene expression levels in Ishikawa cell line were generally much lower than that of MCF-7. Only about 50% of the up-regulated genes in MCF-7 have positive gene expression values in Ishikawa cell line (Figure 96). Figure 97A confirms that up-regulation of Ishikawa cell line is stronger at 12 and 48 hours. Majority of the down-regulated genes in MCF-7 have negative gene expression values in Ishikawa. Figure 97B shows that the down-regulation in Ishikawa cell line also follow a decreasing trend, albeit at a smaller magnitude unlike in MCF-7.

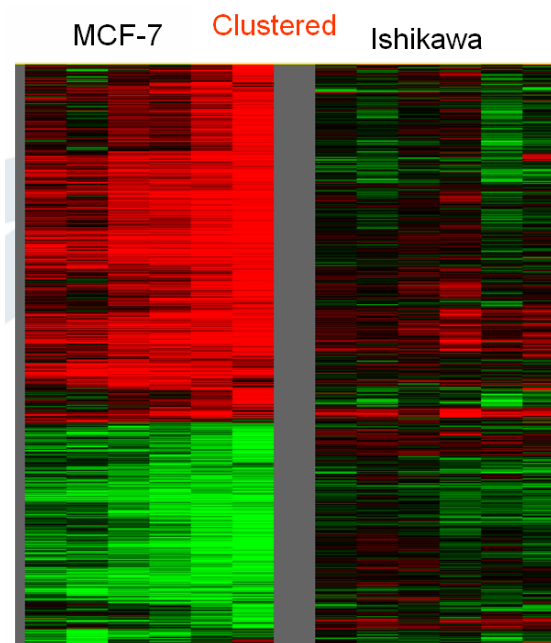


Figure 96 Tissue-specific effects shown on Ishikawa on E2-regulated genes in MCF-7 (Tree-view)

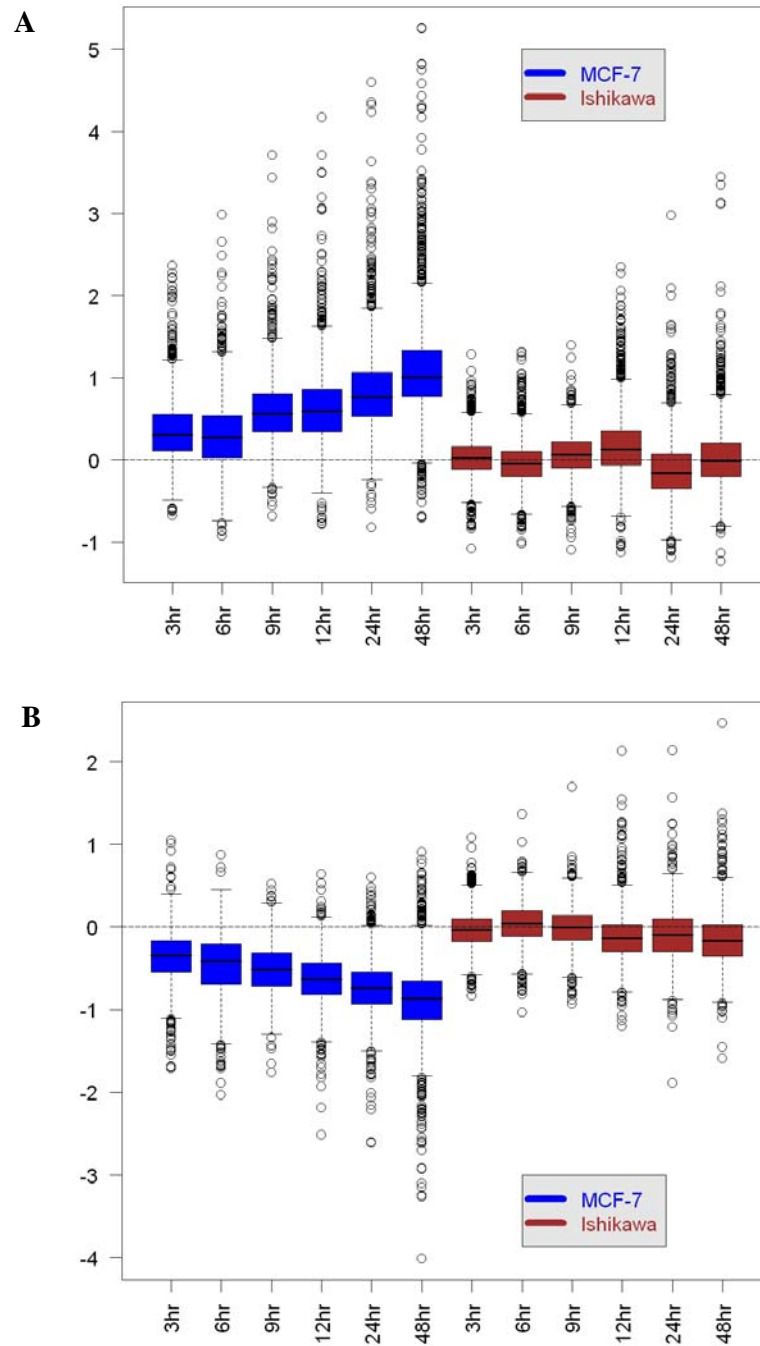


Figure 97 Tissue-specific effects shown on Ishikawa on E2-regulated genes in MCF-7 (Boxplot)

A: Up-regulated genes in MCF-7 and corresponding gene expressions in Ishikawa cell line; B: Down-regulated genes in MCF-7 and corresponding gene expressions in Ishikawa cell line

In summary, the small overlap in the number of regulated genes and different expression profiles in MCF-7 and Ishikawa cell lines show that ER is involved in completely different transcriptional programs in these cell lines.

4.12 Concluding remarks

In conclusion, we obtained genome-wide gene expression profiles with the Affymetrix array (HG-U133 Plus) for different drugs treatments (E2, T, R and I) at different time points (0, 3, 6, 9, 12, 24 and 48 hours). We combined published methods/approaches to gene expression analysis and obtained the regulated genes for each treatment. Using a panel of well-known genes, their gene expression patterns were examined in E2 and across SERMs conditions. The gene expression patterns were well within expectation which confirmed and validated the Affymetrix experiments and data. We correlated the E2-regulated genes to both ER and Pol II binding sites. There was a high prevalence of Pol II preloading mechanism for E2-regulated genes as the distribution of genes away from TSS was similar for both E2 and DMSO conditions. The high percentage of down-regulated genes in 5-100kb category away from TSS also suggests a potential mechanism for down-regulation which involves Pol II pausing or stalling. E2 genes that were strongly up- or down-regulated have a higher propensity to have ER binding sites in proximity than non-regulated genes. Strong E2-ER binding sites with basal occupancy associate with E2 up-regulated genes whereas weak binding associates with tethered mechanism. ER binding sites with greater fold change associate more with higher presence of ERE and larger number of regulated genes. SERMs have the modulating effects on E2 gene expressions by attenuating or even changing the directionality of gene regulation. SERMs besides modulating E2 genes, they also regulate their own set of genes. Only E2 and TE had higher proportion of up-regulated genes from their own set of

regulated genes while the remaining treatments had higher proportion of down-regulated genes. Unique genes to SERMs and E2 were found, exclusive of one another. The directionality of gene regulation, i.e. up-regulated or down-regulated, cannot be predicted based on ER occupancy in E2 and SERMs conditions. However, ER tends to remain occupied in E2 and across SERMs conditions for up-regulated genes. We explored the unique Tamoxifen or Raloxifene binding sites with reference to E2 binding sites by plotting the average FAIRE and H3K4Me1 profiles. The distinctive profiles and higher FAIRE and H3K4Me1 signals suggest strongly that SERMs alter ER's spatial binding characteristics. ER-responsive genes in different tissues upon E2 and SERMs treatments revealed spatiotemporal expression profiles. There were much more number of regulated genes in MCF-7 than Ishikawa cell line. Interestingly, there was greater proportion of down-regulated genes in MCF-7 while Ishikawa cell line had greater proportion of up-regulated genes. This indicated that the antagonizing properties of SERMs were stronger in MCF-7 than Ishikawa cell line. The overlap in the regulated genes between MCF-7 and Ishikawa cell line was low which suggests that the two cell lines had different tissues biology even at the gene level. The corresponding Ishikawa gene expressions for the MCF-7 E2-regulated genes show a much lower signal.

Chapter 5 Functional Analysis of Transcription Factor Binding Site Variants in Human Population

For polymorphism studies in ER binding sites, different breast cancer cell lines that have vast variations in the polymorphism pattern in binding sites are not readily available. As such, we chose a different cell line and transcription factor for studying and characterizing functional binding sites with polymorphism. Lymphoblastoid cell lines and p53 transcription factor were selected as the model system.

5.1 Identification and Genotyping Analysis of SNP

Identification of p53 Motif Polymorphisms

542 high-quality p53 binding sites obtained from ChIP-PET platform on HTC116 cell line (Wei, Wu et al. 2006), were searched for SNPs using dbSNP databases. Out of which 235 sites were found containing an unequivocal p53 consensus binding motif sequence (5'-RRRCWWGYYYRRRCWWGYYY-3') and selected for SNP mining. Using dbSNP database (version 115), 14 sites were found with SNPs identified within the binding motifs. We have successfully genotyped 12 SNPs in 76 anonymous germ-line DNA samples of Caucasian population. Lastly, 6 SNPs were confirmed to be polymorphic with a minor allele frequency (MAF) above 1%.

Of the 6 confirmed SNPs in p53 binding motif, rs1860746 was found to be located within the p53 motif (Figure 98) in the third intron of the PRKAG2 gene which had high p53 transcription factor occupancy (Wei, Wu et al. 2006). rs1860746 is a G/T substitution polymorphism located in the p53 consensus motif.

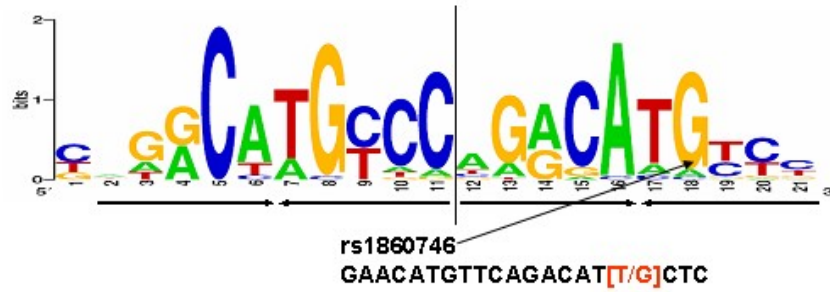


Figure 98 refSNP rs1860746 is located within the consensus p53 motif

The minor allele T causes a mismatch to the p53 consensus motif sequence: 5'-RRRCWWGYYYRRRCWW[G/T]YYY-3' while the major allele G has a perfect p53 consensus motif sequence. Therefore, sites containing p53 motif with major allele are expected to have high occupancy enrichments, whereas the sites carrying the minor allele T are associated with lower binding enrichment. Our genotyping analysis of the SNP in 76 CEPH germ-line DNA samples revealed its MAF to be 20%, which is consistent with the result from the HapMap project. Interestingly, according to the results from the HapMap project, the MAF of this SNP in Asian populations (Chinese and Japanese) is only about 1%, as compared to the higher MAF of 20% observed in African and Caucasian populations. The genotype and allele frequency for this SNP can be found in Table 23.

Table 23 Genotype and allele frequencies for refSNP rs1860746

ss#	Sample Ascertainment					Genotypes				Alleles		
	Population	Individual Group	Sample (2N)	Founder (N)	Source	G/G	G/T	T/T	HWP	G	T	Het. +/-std err
ss24459805	AFD EUR PANEL	European	48	24	IG	0.542	0.417	0.042	0.655	0.750	0.250	
	AFD AFR PANEL	African American	46	23	IG	0.565	0.435		0.200	0.783	0.217	
	AFD CHN PANEL	Asian	48	24	IG	1.000				1.000		
ss2734637	HapMap-CEU	European	120	60	IG	0.633	0.350	0.017	0.317	0.808	0.192	
	HapMap-HCB	Asian	90	45	IG	0.956	0.044		1.000	0.978	0.022	
	HapMap-JPT	Asian	90	45	IG	0.978	0.022		1.000	0.988	0.011	
	HapMap-YRI	Sub-Saharan African	120	60	IG	0.567	0.383	0.050	0.752	0.758	0.242	
Total Samples			562	281		0.744	0.238	0.018		0.863	0.137	0.236 +/- 0.250

PRKAG2 is a Homo sapiens protein kinase, AMP-activated, gamma 2 non-catalytic subunit. Since AMPK protein complex is a central sensor of energy stress, this germ-line p53 binding motif SNP may act as a cis-regulatory variant linking p53 and metabolic homeostasis. Furthermore, AMPK and p53 are known to involve in cancer development besides having interesting frequency pattern in different population. With the above observations, we characterized the molecular and physiological function of this germ-line p53 binding motif polymorphism in cancer development.

5.2 Molecular characterization of the p53 binding site within *PRKAG2* and its germ-line polymorphism (rs1860746)

p53 Response Analysis in Lymphoblastoid Cell Lines

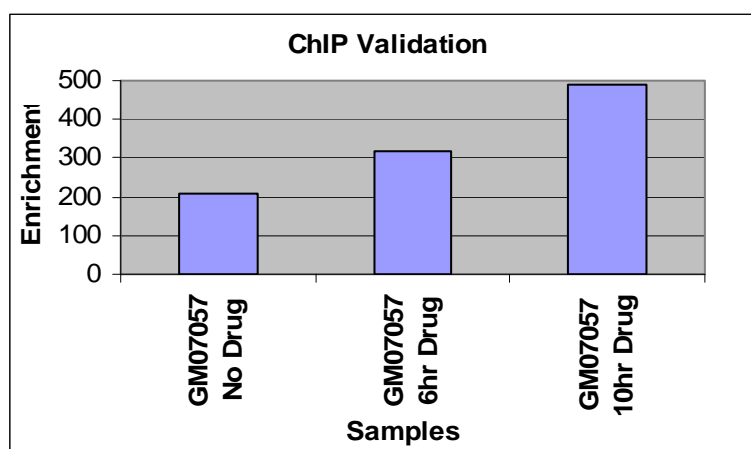


Figure 99 Significant enrichment of p21 binding site sequence after 5-Fu treatment

The results from the ChIP and real-time PCR analysis, showing the significant enrichment of the p53 binding motif sequence of the p21 promoter in the ChIP pull-down DNA from lymphoblastoid cells without or with 5FU treatment for 6 or 10 hrs.

We examined whether the putative binding site in Figure 98 is a true binding target of p53 transcription factor. We performed ChIP experiment validation using a well-characterized p53 target gene - CDKN1A(p21), which has a confirmed p53

binding site in its promoter region and encodes a cyclin-dependent kinase inhibitor (Kaeser and Iggo 2002). We also chose lymphoblastoid cell lines (LCLs) as *in-vitro* system because LCLs have a normal diploid genome and a large collection of cell lines where cells carrying different genotypes of germ-line SNPs are available for functional analysis. As seen in Figure 99, there is already high baseline in the absence of 5-FU drug in the LCLs (about 200 fold enrichment). After the activation of the p53 protein by 5-fluorouracil (5FU) treatment for 6 hrs and 10 hrs, the fold enrichments increased to about 300 fold and 500 fold higher respectively, i.e. 1.5 and 2.5 fold change. The cells were very responsive to 5-FU treatment and showed significant enrichment after drug treatment, which was probably due to the presence of p53 proteins. 5-FU is a well-established drug for increasing both the level of total and activated p53 proteins. LCL is therefore a good diploid cellular system for studying p53-mediated cellular response.

CHIP analysis of rs1860746 in LCLs

To investigate rs1860746's impact on p53's binding to its intronic binding site within *PRKAG2*, we performed the CHIP analysis on LCLs. We began with preliminary experiments on 2 cell lines and subsequently expanded to 8 cell lines.

Initial analysis of 2 cell lines

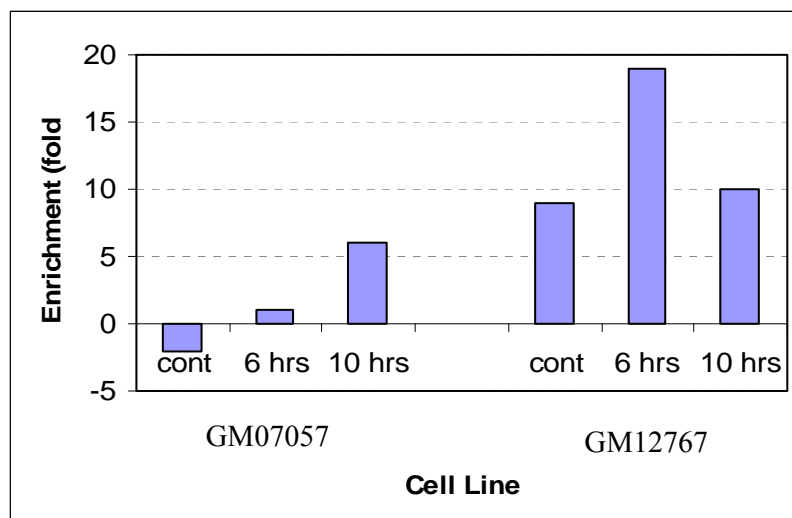


Figure 100. Preliminary Study on the Influence of rs1860746 on p53 Binding

Preliminary study in Figure 100 above showed that G allele increase binding of transcription factor. The study then moved on to three biological replicates and two technical replicates on 8 cell lines.

Validation analysis of 8 cell lines

The 8 cell lines consist of three homozygous for the mutant T allele; two homozygous for the wild-type G allele, and three heterozygous. As shown in Figure 101, A significant enrichment of the binding site sequence was observed at the baseline and further augmented after 5FU treatment (for 10hrs) in the five cell lines that carry either one or two copies of the wild-type G allele (12 fold enrichment in average), whereas the three cell lines carrying two copies of the mutant T allele showed little enrichment of binding sequence (2 fold enrichment in average). In summary, an allele-specific binding pattern was observed, i.e. genotypes associated with C allele increase binding of transcription factor.

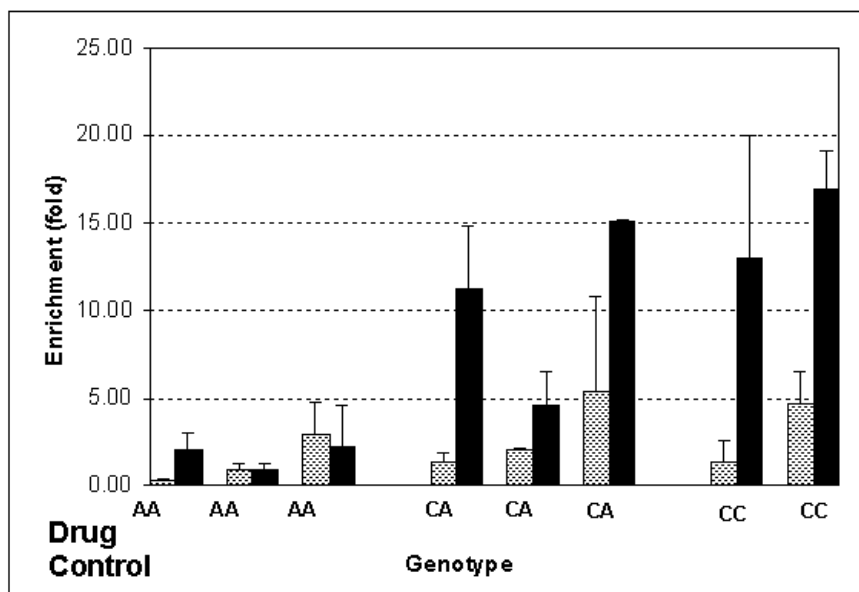


Figure 101 Effect of SNP on Enrichment of binding sites after ChIP Assay

The differential enrichment of the binding site sequence at the baseline and after 5FU treatment in the cell lines carrying either only wild-type allele (G/G) (two cell lines), or mutant (T/T) allele (three cell lines), or both alleles (G/T) (three cell lines)

In a separate study consisted of ChIP experiment pulled down with p53 antibodies and followed by a special technique of cloning and sequencing in HCT119 cell lines (Murphy, Weitsman et al. 2006), the above binding site had been identified as a high confidence p53 functional binding site with a cluster sizes of 8.

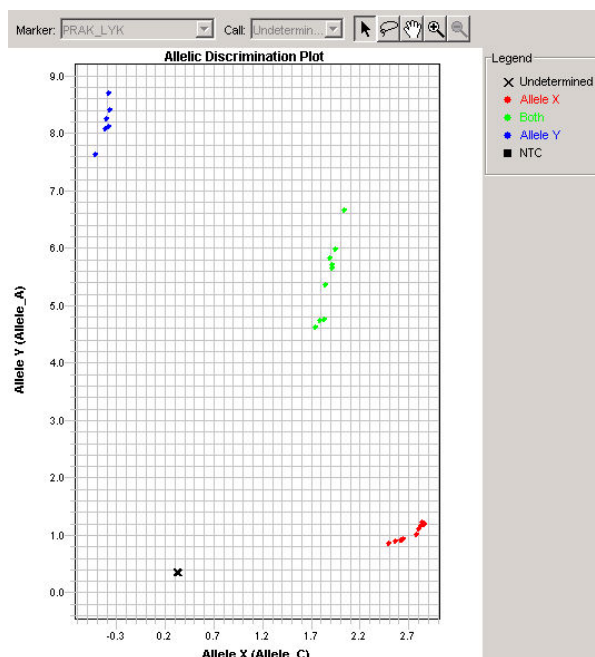
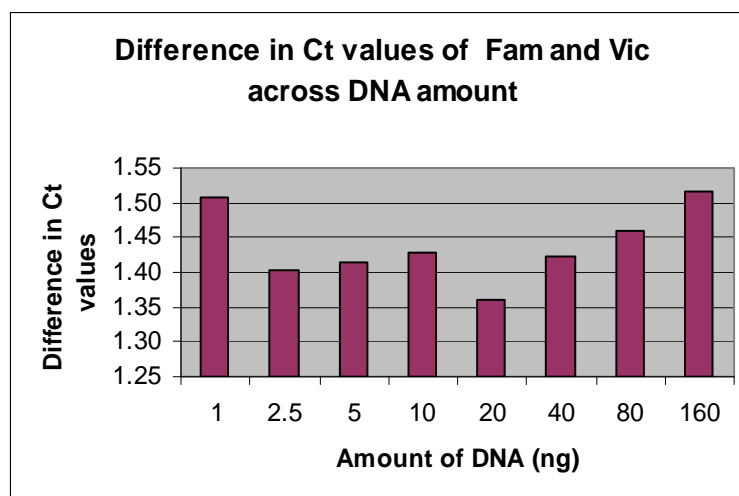
Confirmation analysis of allelic CHIP in 3 cell lines

Figure 102 Validation of Taqman probes using Allelic Discrimination of Plot

9 LCLs were used to validate the quality of Taqman probes. All the 3 different genotypes - AA, CC and AC had 3 different LCLs each. The non-template control (NTC) used was water. The three different genotypes were distinguished very well by the allelic discrimination program that the taqman probes design was validated (Figure 102).

Figure**103 Difference in Ct values across different DNA amount for Taqman Assay**

Since the difference in Ct values given by FAM and Vic was used to show the allele-specific enrichment after ChIP Assay, there was a concern that the difference between the Ct values was not constant across different DNA amount as the real time PCR template. As such, difference in Ct values for varying amount of DNA template amount was obtained. Figure 103 shows that the difference in Ct values is consistently constant for all DNA template amount.

To further demonstrate the stronger binding of p53 to the wild-type binding motif (G allele) than the mutant binding one (T allele), we directly measured the relative abundances of the wild-type (G allele) and mutant (T allele) motif sequences in the ChIP pull-down DNAs from the three heterozygous cell lines after 5FU treatment by real-time PCR analysis. This takes away the bias and uncertainty on any other unknown confounding factors. There are also three biological replicates and two technical replicates. After 5FU treatment for 6 or 32 hours, significantly more of the wild-type G allele sequences than the mutant T allele sequences were found in the ChIP pull-down DNAs (5 to 10 fold enrichment of wild-type over mutant alleles)

(Figure 104). The enrichment of the wild-type G over mutant T allele could also be observed at the baseline, although the enrichment is less prominent.

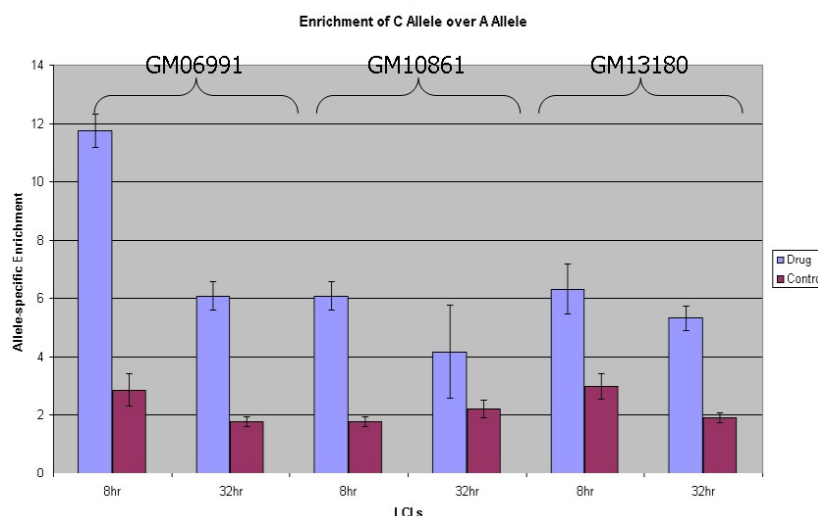


Figure 104 Taqman Assay result to show allele-specific enrichment of ChIP DNA with C Allele

The enrichment of the wild-type G allele over the mutant T allele in the ChIP pull-down DNAs from the three heterozygous cell lines (G/T) after 5FU treatment for 8 and 32 hours.

Our series of the ChIP analyses clearly show that the p53 protein has a higher binding affinity to the wild-type G allele than to the mutant T allele, although the single base substitution does not totally abolish p53's binding to this site.

5.3 Binding affinity by reporter assay analysis

To further investigate whether the suppressive transcriptional regulation was p53 dependent, the transcription regulatory activities of the wild-type and mutant binding site sequences were directly measured through a reporter assay analysis. In the reporter assay analysis, both wild-type and mutant binding site sequences were cloned into a TATA-luciferase reporter vector and then transfected into HCT116 cells with either wild-type p53 protein or with the p53 disrupted by homologous recombination (p53 null). It was found that the presence of the wild-type binding site sequence could

strongly induce the expression of the reporter gene (20 fold induction) in the p53 wild-type HCT116 cells, and this induction was augmented by the activation of p53 by 5FU treatment (about 30 fold induction) (Figure 105). Comparatively, in the p53 null HCT116 cells, this induction effect by the wild-type binding site sequence was largely abolished. In both p53 wild-type and null HCT116 cells, the mutant binding site sequence (T allele) showed a minimal induction of the report gene expression. This result gave direct evidence for this binding site sequence to be associated with a p53-dependent transcriptional regulatory activity.

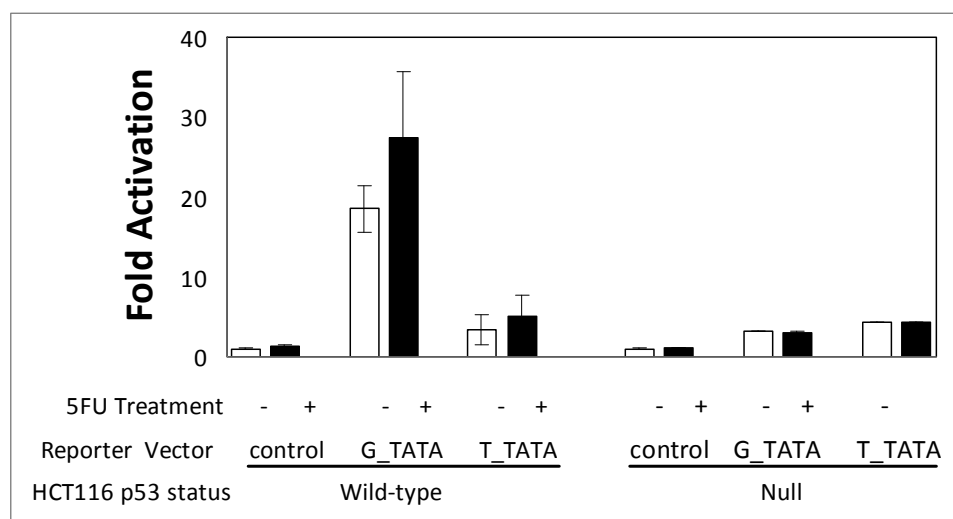


Figure 105 Functional analysis of the binding site sequence (226 bp fragment) and its polymorphism (rs184672) by reporter gene assay in wild-type and p53-null HCT116 cells with or without 5FU treatment

Control: TATA-luciferase pGL4 vector; G_TATA: TATA-luciferase pGL4 vector with a insert of the 226 bp binding site sequence of G allele; T_TATA: TATA-luciferase pGL4 vector with a insert of the 226 bp binding site sequence of T allele.

5.4 Transcription activity by real-time PCR analysis

To investigate whether the observed differential binding activity will lead to the difference in the expression activity of its putative target gene *PRKAG2*, we first analyzed the transcription of *PRKAG2* mRNA (with or without 5FU treatment) using real-time quantitative PCR (qPCR) in 2 cell lines.

Initial Analysis of 2 Cell Lines

The presence of Rs1860746 further down-regulates target gene (*PRKAG2*)'s expression regulation (Figure 106).

Sample Name	Allele	Treat	Norm Ct	Change (fold)	
GM07057	T	6 hrs	6.90	2.56	down
Gm07057	T	control	5.55		
GM07057	T	10 hrs	8.02	5.35	down
GM07057	T	control	5.60		
GM12767	G	6 hrs	9.69	2.49	down
GM12767	G	control	8.37		
GM12767	G	10 hrs	11.76	9.53	down
GM12767	G	Control	8.50		

Figure 106. Preliminary Study on the Influence of rs1860746 on Gene Expression

Validation Analysis of 13 Cell Lines

We next analyzed the transcription of *PRKAG2* mRNA (with or without 5FU treatment) using real-time quantitative PCR (qPCR) in 13 cell lines with different genotypes: three cell lines homozygous for the mutant T allele, five cell lines homozygous for the wild-type G allele, and five heterozygous cell lines (G/T). In most of the cell lines, there is a down-regulation of *PRKAG2* expression after 5FU treatment (Figure 107). Furthermore, the down-regulation of *PRKAG2* expression in the five homozygous cell lines for the wild-type G allele is significantly stronger than the down-regulation in the three homozygous cell lines for the mutant T allele ($p = 0.025$, t-test).

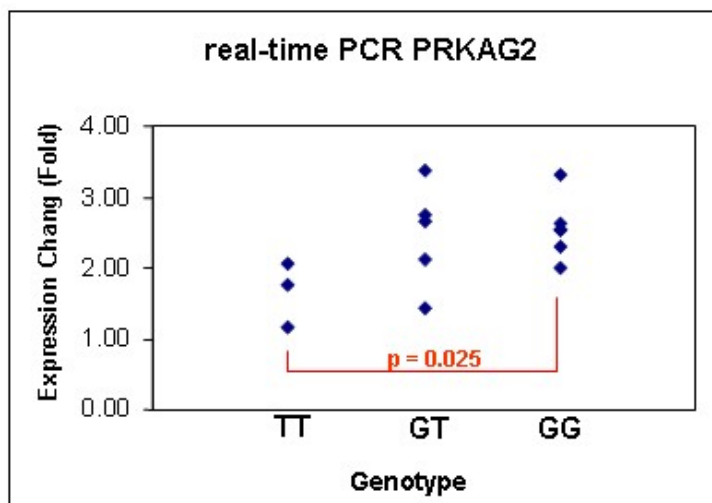


Figure 107. Real-time PCR results for gene expression change of PRKAG2

The results of the real-time gene expression analysis, showing the down-regulation of PRKAG2 expression after 5FU treatment in 13 cell lines carrying either only wild-type allele (G/G) (five cell lines), or mutant (T/T) allele (three cell lines), or both alleles (G/T) (five cell lines).

5.5 Polymorphism's impact on the protein levels by western blot analysis

AMPK protein complex consists of one catalytic (α) and two non-catalytic regulatory (β and γ) subunits, and the expression and activity of the AMPK protein complex depends on the co-regulation of its three subunits (Crute, Seefeld et al. 1998). We hypothesize that the interruption of p53's down-regulation effect on the transcriptional expression of the AMPK γ subunit, this germ-line p53 binding motif variant can have an impact on the expression and activity of the AMPK protein complex. As such, we investigated the polymorphism's impact on the protein levels of both AMPK γ and α subunits using western blot analysis. The western blot analysis was performed in two cell lines (among the 13 cell lines subjected to real-time PCR analysis) that show the most prominent difference in the p53-mediated down-regulation of PRKAG2 mRNA level (Refer to Figure 107). Protein levels of p53, total and phosphorylated AMPK α , AMPK γ and actin (endogenous control) were assessed in the two cell lines at baseline and after 5FU treatment for 8, 24 and 48

hours. As shown in Figure 108, the expression of p53 protein was induced in a time-dependent fashion by 5FU treatment in both cell lines. In contrast the levels of the AMPK γ and total and phosphorylated- AMPK α proteins after 5FU treatments differ significantly between the two cell lines. In the cell line carrying mutant binding site (T/T), the levels of the AMPK γ and total and phosphorylated- AMPK α proteins were largely unaffected by 5FU treatment, whereas in the cell line carrying wild-type binding site (G/G), a significantly decreased expression of the three proteins was seen after 5FU treatment, primarily of phosphorylated AMPK α and AMPK γ especially at 48 hours. These results are consistent with the regulatory effect of AMPK γ on the activity of AMPK α .

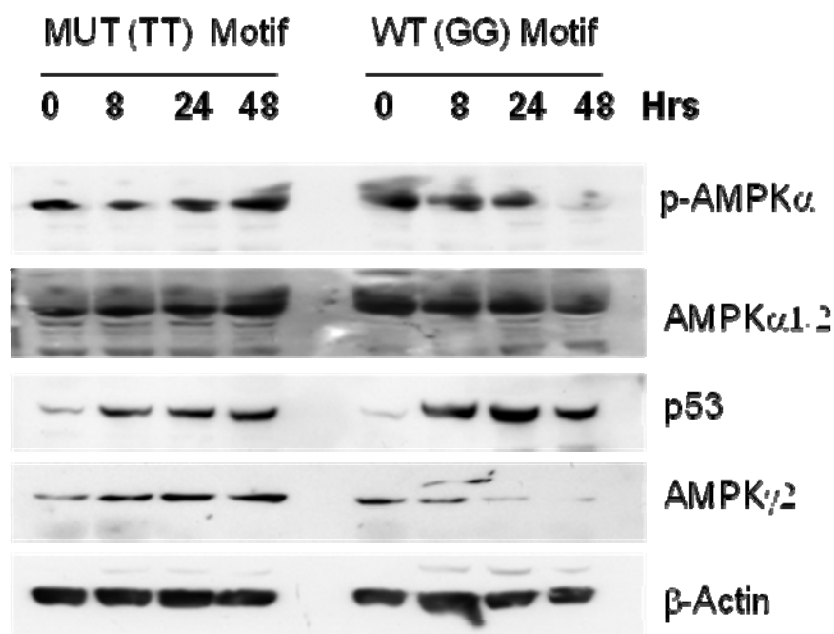


Figure 108 Western blot analysis on AMPK sub-units, p53 and actin

The results shows that the differential down-regulation effect of p53 activation by 5FU on AMPK protein complex (AMPK γ , total and phosphorylated- AMPK α proteins) in cells carrying either wild-type (G/G) or mutant (T/T) binding motif.

5.6 Genetic association analysis of the p53 binding motif SNP (rs180746) with cancer susceptibility

Given that both AMPK and p53 have been implicated in cancer development (Vousden and Lu 2002; Shaw 2006), it is postulated that this germ-line regulatory variant (rs180746) of the transcriptional link between p53 and AMPK may have an impact on cancer susceptibility. To test this hypothesis, the analysis of the SNP rs180746 in a large cancer sample from Sweden and Finland was carried out. The “worst case” assumption was that the effect of this SNP on cancer susceptibility would be low, as has been found for the recent identified breast cancer susceptibility loci (Easton, Pooley et al. 2007), and that only the homozygous TT genotype would show a phenotypic effect (as indicated by our *in vitro* functional analyses). The SNP in the cancer sample consisting of 1297 breast cancer patients, 579 endometrial cancer patients, 1637 healthy controls from Sweden, 2399 breast cancer patients and 1256 healthy controls from Finland were genotyped. The MAF of rs1860746 was 18.2% in our samples, which was very similar to the one detected in the 76 CEPH DNA samples as well as the one reported in Caucasians by the HapMap project. The genotype frequencies at this locus were in Hardy-Weinberger equilibrium.

Association analysis was performed using the χ^2 test under a recessive model of inheritance, and the results are summarised in Table 24. First, the association analysis was performed in the whole cancer sample of 4113 patients and 2983 controls. It was found that there was significant association of the homozygous mutant genotype (TT) with an increased risk for cancer development (OR = 1.36, $p = 0.024$). Second, the association was investigated in the breast cancer sample including 3541 patients and 2740 controls, and significant evidence was identified (OR = 1.34, $p = 0.043$) for the association with breast cancer susceptibility. Finally, the association of this

regulatory polymorphism in different clinical subgroups of patients stratified based on the menopausal status, family history and ER status, was investigated. Interestingly, significant association was evident only in the premenopausal patients (OR = 1.66, $p = 0.05$) and the patients with family history (OR = 1.48, $p = 0.04$) but not in sporadic (OR = 1.25, $p = 0.16$), postmenopausal (OR = 1.38, $p = 0.10$) cases or the subgroup analysis based on ER status.

Table 24 Analysis of the association of SNP with cancer susceptibility under a recessive model of inheritance

Sample Set	Sample	Size	Genotype Freq (%)		OR (95% CI)	P value
			GG or GT	TT		
Whole Sample	Controls	2893	97.11	2.89		
	Cases	4113	96.06	3.94	1.36 (1.04, 1.78)	0.02
Breast Sample	Controls	2740	97.08	2.92		
	All Cases	3541	96.07	3.93	1.34 (1.01, 1.77)	0.04
	ER+ Cases	2492	96.27	3.73	1.26 (0.92, 1.72)	0.15
	ER- Cases	588	95.58	4.42	1.48 (0.93, 2.35)	0.10
	Postmenopausal Cases	2447	96.24	3.76	1.30 (0.96, 1.76)	0.10
	Premenopausal Cases	502	94.82	5.18	1.66 (1.00, 2.75)	0.05
	Sporadic cases	2359	96.31	3.69	1.25 (0.92, 1.70)	0.16
	Familial cases	1098	95.72	4.28	1.48 (1.01, 2.17)	0.04

5.7 Concluding remarks

From the above functional analyses in LCL cells, we have proved that p53 is a suppressor on the $\gamma 2$ subunit (*PRKAG2* gene) of the AMPK protein complex. Upon exposure to a genotoxic agent – 5FU, the p53 proteins are both increased in levels and functionally activated, which subsequently down-regulates the expression of AMPK γ

(*PRAKG2* gene). The germ-line variant (rs1860746) found within the p53 binding sites for *PRAKG2* gene, reduced greatly the p53 binding, attenuating both the gene and protein expression.

AMP-activated protein kinase (AMPK) is an important central sensor of energy stress and stimulated AMPK are essential for maintaining normal metabolism through ATP-generating mechanism. AMPK activity is greatly controlled by the phosphorylation of Thr172 on AMPK. During conditions such as glucose deprivation and oxidative stress, the cells experience stresses that the AMP/ATP ratio is elevated, activating the AMPK. During AMPK protein activation, energy-consuming processes such as protein synthesis are inhibited while the energy-generating processes such as glucose uptake and fatty acid metabolism are promoted. In this way, the energy balance is restored in the cells and the stresses are gone. AMPK activates the p53 by phosphorylation on Ser-15 of p53, which in turns causes cell cycle arrest. (Jones, Plas et al. 2005) . Our findings lead to novel negative feed-back of p53 on AMPK that while p53 down-regulates AMPK, p53 itself is activated by AMPK.

Our study provides the first genetic evidence, at the population level, that conditionally higher AMPK activity is associated with increased susceptibility to cancer development. By genotyping this p53 binding site polymorphism in a large cancer sample of 4113 patients and 2893 controls, significant evidence was found for association of the homozygous mutant genotype with high susceptibility to cancer development (OR = 1.36, p = 0.024). Because the mutant allele was shown to interrupt p53's down-regulation of the *PRKAG2*/AMPK expression under conditions of genotoxic stress, our genetic analysis strongly suggests that AMPK more likely functions as a tumor promoter, at least at human population level. Interestingly, our subgroup analysis in breast cancer suggests that this germ-line p53 binding site

polymorphism may play a more prominent role in pre-menopausal breast cancer patients with a positive family history. This finding is consistent with the observations that patients with germ-line p53 mutations in the families affected with Li-Fraumeni syndromes (LFS) are at risk for early-onset breast cancer(Olivier, Goldgar et al. 2003) and that germ-line p53 mutations can be found in the patients of hereditary breast cancer who are negative for BRAC1 and BRAC2 mutations(Walsh, Casadei et al. 2006). Given that there is good evidence for p53 and metabolic stress to function in the aging process (Bensaad and Vousden 2007), we speculate that the germ-line p53 binding site polymorphism may also have an impact on some aspect of human longevity.

Chapter 6 Conclusion

In conclusion, more than 40,000 mapped and putative ER binding sites have been incorporated into the customised array which was an isothermal array with ~380,000 probes in which, each region was tiled with at least 6 probes. Mainly, the important design considerations of the ER α specific binding site array included the adoption of variable-length probes to ensure similar hybridisation specificity that 70% of all probes have similar melting temperature between 73 and 77 °C. It was also important to carry out various quality controls of the customised array such as assessments based on probe properties, technical and biological reproducibility and reuses of stripped arrays, so as to confirm that the array was capable of querying the biology questions in this thesis.

With this customised ER chip, the different CHIP assays performed with virtually any antibodies (ER, H3K4Me1, FOXA1 and GATA3), cell lines (MCF-7 and Ishikawa cell lines), drug treatments (E2, T, R and I) and concentrations in time course experiments (CHIP and FAIRE), the profiles of the genome-wide binding sites were comprehensively surveyed.

To detect enriched binding sites or peaks from the above profiles, a new algorithm called the VFLM was designed and implemented. It detected 6482 ER binding sites in E2 treatment. Genome-wide binding sites analysis showed a high prevalence of pre-occupied ER binding sites which also had a greater ER recruitment upon E2 treatment. ER binding profiles for SERMs showed a similar location as in E2 but the ER binding recruitments in SERMs were significantly much less than in E2. Distribution of Full ERE, half ERE or no ERE on both the SERMs or unique binding sites to different treatments showed that ER-SERMs might utilise tethering mechanism much more than ER-E2. De novo motif analysis showed that ER-SERMs

conformation was dissimilar in different SERMs with I seemed to have the greatest differential conformation change, followed by R and the least conformation change was in T. Comparing E2 binding sites with and without basal occupancy, the former showed a more accessible chromatin configuration, had the greatest FAIRE signals and the highest H3K4Me1 enhancer marks. H3K4Me1 was also found to be the most predictive factor for identifying ER binding sites through both decision tree and ROC curve. It was further confirmed that the role of FOXA1 was a pioneering factor while GATA3 served as a co-factor. Lastly, it was shown that ER binding were regulated differently in MCF-7 and Ishikawa cell line.

Under different drugs treatments (E2, T, R and I) at different time points (0, 3, 6, 9, 12, 24 and 48 hours), genome-wide gene expression profiles with the Affymetrix array (HG-U133 Plus) were obtained. With the combination of published methods, approaches and gene expression analysis, the regulated genes for each treatment were obtained. The regulation of genes (up-regulated or down-regulated) was classified using a novel approach whereby consistency in expression values and its gradient were used. The gene expression patterns of a panel of well-known genes were examined in E2 and across SERMs conditions for assessing the array data quality. The gene expression patterns were well within expectation which confirmed and validated the Affymetrix experiments and data. The E2-regulated genes were correlated to both ER and Pol II binding sites. There was also a high prevalence of Pol II preloading mechanism for E2-regulated genes as the distribution of genes away from TSS was similar for both E2 and DMSO conditions. The high percentage of down-regulated genes in 5-100kb category away from TSS also suggested a potential mechanism for down-regulation which involved Pol II pausing or stalling. E2 genes that were strongly up- or down-regulated had a higher propensity to have ER binding

sites in proximity than non-regulated genes. Strong E2-ER binding sites with basal occupancy associated with E2 up-regulated genes whereas weak binding ones associated with tethered mechanism. ER binding sites with greater fold change associated more with higher presence of ERE and a larger number of regulated genes. SERMs were found to have modulating effects on E2 gene expressions by attenuating or even changing the directionality of gene regulation. Besides modulating E2 genes, SERMs also regulated their own set of genes. From their own set of regulated genes, only E2 and TE had higher proportion of up-regulated genes while the remaining treatments resulted in a higher proportion of down-regulated genes. Unique genes to SERMs and E2 were found, exclusive of one another. The directionality of gene regulation, i.e. up-regulated or down-regulated, could not be predicted based on ER occupancy in E2 and SERMs conditions. However, ER remained occupied in E2 and across SERMs conditions for up-regulated genes. The unique Tamoxifen or Raloxifene binding sites with reference to E2 binding sites were also explored by plotting the average FAIRE and H3K4Me1 profiles. The results showed that the profiles were distinctive with unique SERMs binding sites showing higher FAIRE and H3K4Me1 signals, which strongly suggested that SERMs altered ER's spatial binding characteristics. ER-responsive genes in different tissues upon E2 and SERMs treatments revealed spatiotemporal expression profiles. There were a greater number of regulated genes in MCF-7 than Ishikawa cell line. Interestingly, there were a greater proportion of down-regulated genes in MCF-7 while Ishikawa cell line had a greater proportion of up-regulated genes. This indicated that the antagonising properties of SERMs were stronger in MCF-7 than Ishikawa cell line. The overlap in the regulated genes between MCF-7 and Ishikawa cell line was low which suggested that the two cell lines had different tissues properties even at the gene level. The

corresponding Ishikawa gene expressions for the MCF-7 E2-regulated genes showed a much lower signal. A key observation is that Ishikawa cell line is so different from MCF-7 cell line in their ER-dependent transcriptional programs.

Our proposed model failed to explain all the relationships between ER binding, chromatin state, histone modifications and FoxA1. This could be due to a number of reasons. Since ER β was not considered in the analysis, some of the results that deviated from the proposed model could be due to the involvement of ER β . Macaluso and his colleagues reported the co-immunoprecipitation of both ER β and pRb2/p300 in MCF7 (Macaluso, Montanari et al. 2006). Instead of purely ER α homodimers, some ER α may dimerise with ER β to form heterodimers. Together with ER β homodimers, these different dimers shaped and defined the complex observations obtained from the various experimental approaches. ER β has been shown to attenuate the transcriptional activities of ER α (Lindberg, Moverare et al. 2003) by binding to estrogen-responsive promoters and changing the association with c-Fos and c-Jun (Matthews, Wihlen et al. 2006).

E2 was shown to modify the micro-ribonucleic acid (miRNA) expression profiles (Cohen, Shmoish et al. 2008). Since miRNA has the ability to repress translation, it can down-regulate mRNA by degradation or cleavage without the requirement of perfect complimentary targeted sequences (Lim, Lau et al. 2005).

Even in DMSO only treated samples which served as a control, there was already high basal occupancy of 4729 binding sites albeit at low enrichments. These sites were probably occupied with estrogen receptor coupled with endogenous estrogen that was present despite the hormone starvation for 3 days. The presence of minute amount of endogenous estrogen may contribute to some of the binding sites profiles

which were observed in SERMs only treatments. Alternatively, a minority of ER could dimerise and bind to the DNA without forming ER-E2 complexes.

A number of binding sites were found located within genes. A significant proportion of binding sites within the gene was even located within the first intron. Most of the time, it also coincided to the 5' untranslated region (UTR). The observation that many binding sites were found inside genes often corresponded to down-regulated genes that led to the hypothesis that the binding sites may serve as a function to stop transcription. The TF may continue to recruit the associated factors and help to stop the transcription process.

The data analysis on the nucleosomes array indicated that the majority of nucleosome positions remained relatively the same before and after E2 treatment. However, there was greater nucleosome depletion when the genes became actively transcribed upon drug treatment.

The ideal study would be to determine the functional importance of ER binding to its cognate DNA sites in the whole tissue for active human populations. This however is not possible so we want to use alternative system – p53 binding in LCL cell lines. From the functional analyses in LCL cells, we have proven that p53 was a suppressor on the γ 2 subunit (*PRKAG2* gene) of the AMPK protein complex. Upon exposure to a genotoxic agent – 5FU, the p53 proteins were both increased in levels and functionally activated, which subsequently down-regulated the expression of AMPK γ (*PRAKG2* gene). The germ-line variant (rs1860746) found within the p53 binding sites for *PRAKG2* gene, reduced the p53 binding greatly, attenuating both the gene and protein expression.

AMP-activated protein kinase (AMPK) was an important central censor of energy stress and stimulated AMPK was essential for maintaining normal metabolism

through ATP-generating mechanism. AMPK activity was greatly controlled by the phosphorylation of Thr172 on AMPK. During conditions such as glucose deprivation and oxidative stress, the cells experience stress that the AMP/ATP ratio was elevated, activating the AMPK. During AMPK protein activation, energy-consuming processes such as protein synthesis were inhibited while the energy-generating processes such as glucose uptake and fatty acid metabolism were promoted. In this way, the energy balance was restored in the cells and the stresses were gone. AMPK activated the p53 by phosphorylation on Ser-15 of p53, which in turned caused cell cycle arrest. (Jones, Plas et al. 2005) . Our findings led to novel negative feed-back of p53 on AMPK that while p53 down-regulated AMPK, p53 itself was activated by AMPK. More importantly, our study has further demonstrated that this modulation of p53-AMPK transcriptional link by the germ-line polymorphism will increase the risk for cancer development. As an proof-in-principal study, our study has highlighted that combining the genome-wide discovery of transcription regulatory elements (such as transcription factor binding sites) with the forward genetic analysis in both model and human systems can greatly advance our understanding on the molecular and physiological functions of regulatory genetic variation. We further posit that a ‘marriage’ between the new genome-wide knowledge of various regulatory sequences and the rapidly accumulated disease association data on germ-line polymorphisms will bring a paradigm shift to regulatory variation research.

Further studies could be carried out by obtaining the binding sites and gene expression profiles on the actual clinical breast cancer samples. Subsequently, a relational map of ER and gene expressions could be constructed that may help to predict pharmacologic effects. Since a pool of ER binding sites has been obtained, it would be of high interest to study what the consequences of human variation at the

validated ER binding sites are. Besides functional studies on the polymorphism on those binding sites, ER association study of the binding sites could be carried out. We should treat CHIP with different antibodies to co-factors such as N-Cor and SRC-1 and examine the binding sites profiles in the same drug combination used in the studies. ERbeta binding sites could also be mapped to this customised array and assess the binding sites patterns under different conditions. In order to study the binding dynamics of ER, a time course CHIP could be performed. Lastly, all of the above could be performed and the studies could be carried out in this thesis using the latest technology – ChIPSeq to have an unbiased genome-wide study.

In conclusion, the novel ER customised array platform and the comprehensive data on the ER-binding sites and the transcription activities affected by SERMs provided in this thesis would aid in the discovery of better SERMs and the improved understanding of their mechanism of actions.

Chapter 7 Materials and Methods

7.1 Material and Methods for Binding Sites Array

Cell Culture and Treatments

MCF7 cells were in 20ml Dulbecco's modified Eagle's medium (DMEM/F-12) from Invitrogen/Gibco supplemented with 10% fetal bovine serum (FBS) from Hyclone at 37°C under 5% CO₂. The cells were split to more plates before they reach > 90% confluence to prevent the cells from changing in characteristics when they reached full confluence.

Chromatin Immunoprecipitation (CHIP) Assay

Day 1: Starvation of cells

MCF7 cells were grown to ~50% confluence. The cells were washed 3 times in each plate with 10ml of 1X PBS, and 20ml phenol-red free DMEM / F12 medium (Invitrogen/Gibco) supplemented with 5% charcoal-dextran stripped FBS(Hyclone) to starve them. The cells were then incubated for 3 days in preparation for 17E2-estradiol (E2; Sigma) treatment.

Day 2: (A) Estrodiol Treatment

20µl of 10µM E2 (estrodiol in DMSO) was added into each plate to obtain a final concentration of 10nM, while 20µl of DMSO was added to other plates as negative control. All treated dishes were incubated for another 45 minutes before harvesting.

(B) Harvesting Cells

MCF7 cells were harvested under the fume hood. 675 μ l of formaldehyde was added in each plate for cross-linking and rotated for 10 minutes on a rotator at room temperature (RT). 2ml of glycine was then added to stop the cross-linking followed by rotation for 5 minutes at RT. The medium was poured, washed 2 times with 10ml of cold 1 X PBS and finally the PBS was poured. Costar brand 3cm blade scraper was used to scrap the cells. The cells were then collected from one plate into 15ml tube individually.

The cells in the 15ml tube were centrifuged at 3000 rpm for 15 minutes at 4°C using Sorvall Legend RT (Heraeus). After the supernatant was removed, the pellets were re-suspended and each was washed with 5ml 1 X PBS. Next, the pellet cells were centrifuged at 3000 rpm for 5 minutes at 4°C. Subsequently, the supernatant was removed.

(C) Isolation of the Nuclei by Lysis of the Cell Membrane

The pellets were re-suspended and washed individually with 5ml of TritonX-100 lysis buffer. They were then incubated in cold room with 10 minutes gentle agitation and spun at 3000 rpm for 5 minutes at 4°C. This process of re-suspension, washing with the lysis buffer, incubation and centrifugation was repeated twice. The supernatant was removed. The pellet was re-suspended with 300 μ l SDS lysis buffer and the nuclear lysates were transferred into a 1.5ml microfuge tube and kept in ice.

(D) Sonication of the Nucleus Lysate (Set up Branson digital sonifier in the cold room)

2 sets of 15-second pulses is performed on the nucleus lysate using a sonicator, set to 20% of maximum power. Nucleus lysate was sonicated to shear genomic DNA to lengths between 500 and 1200 bps. The sample was chilled on ice for 15 seconds between sonications. The sample is kept on ice at all times to prevent DNA from denaturing by the heat generated by sonication.

The pellet cell debris was centrifuged at 13 X 1000 rpm for 15 min at 4 °C and the supernatant was transferred into a 15ml tube. ~ 50µl of nuclear lysate was aliquoted for checking the sizes of the DNA fragments and stored at -80°C.

Day 3: Set Up of the CHIP Assay

The bottle containing 50 % Protein A Sepharose (PAS or beads) slurry was shaken before use. 1.3ml of beads were placed into a 15ml tube and washed 2 times with 10ml of 1 X PBS with 0.1% TritonX-100 and 1 time with 10ml of TSE I buffer. They were spun down each time at 800rpm for 2 minutes at 4°C using Sorvall Legend RT (Heraeus) centrifuge. After the supernatant was removed, the beads were equilibrated in 1.3ml of TSE I buffer (to obtain 50% beads slurry).

300µl of the 50% beads slurry were transferred into 1.5ml microfuge tube where the beads were treated by adding 100µl of BSA (20 mg/ml). The beads were mixed well and incubated in cold room with rotation for ~ 2 hours.

Sonicated samples (from Day 2, section D) were thawed and each sample was diluted with 4ml dilution buffer. Samples were pre-cleared by adding ~ 250µl of the 50% beads slurry and incubated in the cold room with rotation for ~ 2 hours.

After 2 hours, pellet beads were spun down at 800rpm for 2 minutes at 4°C and the supernatant was removed. The beads were then washed 3 times with 1ml of TSE I buffer. Finally the beads were equilibrated in 300µl of TSE I to obtain 50% beads.

After pre-clearing on the sonicated samples, the pellet beads were spun down at 800rpm for 2 minutes at 4°C. The supernatant was collected, 0.5ml of it was taken out as total input, labeled and stored at -80°C, and the rest of the supernatant was distributed evenly into a new 15ml tube.

The immunoprecipitation (IP) was set up by using ~ 4.5ml of pre-cleared lysate, 75µl of 50% beads and 5µl of polyclonal anti-ER α AB (HC-20) (200 µg/ml) or polyclonal anti-GST AB (Z-5) (200µg/ml) were added into each 15ml tubes. The tubes were then incubated in the cold room with gentle rotation overnight or for at least 12 hours. Parafilm was used to seal the cap properly.

Day 4: Washing of the Bead and Eluting of the DNA

The overnight IP was spun in a 15ml tube at 800rpm for 2 minutes at 4°C, in Sorvall Legend RT (Heraeus) centrifuge. Next, the supernatant was removed with 10ml serological pipette and P1000 without disrupting the beads.

The beads were washed with 5ml of cold TSE I buffer by gentle rotation in cold room for 10 minutes. The pellet beads were spun at 800 rpm for 2 minutes at 4°C and the supernatant was removed.

The beads were washed again with 5ml of cold TSE I buffer by gentle rotation in cold room for 10 minutes and the pellet beads were spun at 800 rpm for 2 minutes at 4°C. Thereafter, the supernatant was removed.

Beads were washed with 5ml of cold buffer III by gentle rotation in cold room for 10 minutes. Pellet beads were spun at 800rpm for 2 minutes at 4°C and the supernatant was removed.

Finally, the beads were washed with 5ml of cold TE buffer by gentle rotation in cold room for 10 minutes. Pellet beads were spun at 800rpm for 2 minutes at 4°C and the supernatant was removed.

150µl of room temperature elution buffer was added to the beads in the 15ml tube. About 2mm of the yellow pipette tip was cut and used to re-suspend the beads and for transferring to a new 2ml screw cap microfuge tube. Another 100µl of elution buffer was used to rinse the 15ml tube. The remaining beads were transferred to the 2ml screw cap microfuge. Incubation was done in the Thermomixer at 700rpm and at 65°C for 30 minutes. The beads were spun down at 800rpm for 2 minutes. The supernatant of each sample was transferred into new 2ml screw cap tube. ~ 250µl elute was recovered for all samples. 230µl of TE buffer (for dissolving DNA) was added into the beads. The beads were re-suspended and spun down. Another ~ 230µl elute was recovered and pooled together to yield ~ 480µl elute.

20µl of pronase (20 mg/ml) was added to each sample to yield a total of 500µl. Incubation was done at 42°C water bath for 2 hours as protease treatment. The nuclear lysate and the supernatant for input DNA (in day 3 above) was thawed. The nuclear lysate was diluted to ~ 500µl by TE buffer (for dissolving DNA). The same protease treatment was done for all the samples.

The CHIP samples, nuclear lysate and the total input samples were de-crosslinked together at 65°C overnight. (All the caps of the 2ml screw cap tubes were wrapped and sealed with parafilm before incubation to prevent evaporation of the volume.)

Day 5: DNA Extraction and Precipitation

~ 500 μ l (1 vol) of phenol /chloroform /isoamyl alcohol (25: 24:1) was added to each sample, mixed well and spun at maximum speed for 10 minutes. The upper aqueous phase was saved into new tubes. ~500 μ l (1 vol) of chloroform was added to each sample, mixed well and spun at maximum speed for 10 minutes. The upper aqueous phase was saved into new tubes again. 1 μ l of glycogen (20 μ g/ μ l), 45 μ l (0.1 vol) of 4M LiCl, and 500 μ l (1 vol) of 100% isopropanol or 1250 μ l (2 – 3 vol) of 100% ethanol were added, mixed well and incubated at -80°C for 1.5 hours for precipitation of DNA.

The DNA was spun at a maximum for 30 minutes at 4°C. The supernatant was discarded and the DNA pellet was rinsed with 1ml 95% ethanol, spun again at maximum speed for 15 minutes at 4°C. The supernatant was discarded again and the DNA pellet was air-dried. Next, 15 μ l of H₂O was added into each of the CHIP samples to dissolve DNA and the CHIP DNA was stored at -20°C. 50 μ l of H₂O was then added into each of the total input samples and nuclear lysate samples to dissolve DNA, followed by the addition of 1 μ l of RNase (10 mg/ml) into each input sample. Incubation was done at 37°C for 1 hour and the input DNA was stored at -20°C.

WGA Amplification**Library Preparation**

2 μ l of 1X Library Preparation Buffer was added to each sample, followed by 1 μ l of Library Stabilization Solution. The mixture was vortex thoroughly, consolidated by centrifugation and placed in thermal cycler at 95°C for 2 minutes. The sample was cooled on ice, consolidated by centrifugation and replaced on ice. Thereafter, 1 μ l of Library Preparation Enzyme was added to the sample and it was vortex thoroughly and before centrifugation was done briefly. Next, the sample was placed in a thermal

cycler and incubated as follows: 16°C for 20 minutes, 24°C for 20 minutes, 37°C for 20 minutes, 75°C for 5 minutes and lastly, held at 4°C. The sample was then removed from the thermal cycler, centrifuged briefly and labeled as X. Samples may be amplified immediately or stored at -20 °C for three days. In the process of amplification, a master mix was prepared by adding the following reagents to the 15µl of X prepared earlier: 7.5µl of 10X Amplification Master Mix, 47.5 µl of Nuclease-Free Water and 5µl (12.5 units) of Jumpstart Taq DNA Polymerase. The master mix was vortex thoroughly and centrifuged briefly before themocycling began.

The following profile has been optimized for a PE 9700 or equivalent thermocycler:

Initial Denaturation took place at 95°C for 3 minutes, 14 cycles were performed as follows: denaturation at 94°C for 15 seconds followed by extension at 65°C for 5 minutes. After thermocycling was completed, the reactions were maintained at 4°C or stored at -20°C until they were ready for analysis or purification. The stability of WGA DNA was equivalent to genomic DNA stored under the same conditions.

Nimblegen ChIP-on-chip protocols

Sample Labeling

Cy3 and Cy5 dye-labeled 9mers were diluted to 1 O.D./42µl Random 9mer Buffer (8.6ml of VWR deionised water, 1.25ml of 1M Tris-HCl, 125µl of 1M of MgCl₂, 17.5µl of β-Mercaptoethanol) . The random 9mer buffer was aliquot to 40µl individual reaction volumes in 0.2ml thin-walled PCR tubes and stored at -20°C. 1µg

of WGA amplified ChIP Sample (1 μ g), 40 μ l of Cy5-9mer Primers and VWR Water was added to make up a total volume of 80 μ l of ChIP sample. 1 μ g of WGA input DNA (1 μ g), 40 μ l of Cy3-9mer Primers and VWR Water was added to make up a total volume of 80 μ l of reference input DNA. Samples were denatured by heating in thermocycler at 98 $^{\circ}$ C for 10 minutes and quick chilled in ice water bath. 10 μ l of 50X dNTP mix, 8 μ l of VWP deionised water and 2 μ l of Klenow (50 U/ml) were added to both the denatured ChIP samples and input DNA samples. ChIP samples and input DNA samples were mixed well by pipetting 10 times. The contents were forced to the bottom of the tube by spinning down at low RPM. ChIP samples and input DNA samples were incubated at 37 $^{\circ}$ C for 2 hours in a thermocycler protected from light. After 2 hours, 10 μ l of Stop Solution (0.5M EDTA) was added to stop the solution. The samples were each transferred to a 1.5ml tube and precipitated by adding 11.5 μ l of 5M NaCl and 110 μ l of isopropanol to each tube, each with a cumulative total volume of 231.5 μ l. The samples were vortex and incubated for 10 minutes at room temperature in the dark. Centrifugation was done for 10 minutes at maximum speed. The supernatants were then removed using a pipette. The pellets were rinsed with 500 μ l of 80% ice-cold ethanol and dislodged from the tube wall. Centrifugation was carried out again at maximum speed for 2 minutes and the supernatants were removed with a pipette. The supernatants were then speed-vac on low heat and protected from light for 5 minutes until they became dry. The labeled samples were stored at -20 $^{\circ}$ C and protected from light. The dried pellets were then rehydrated in 25 μ l VWR deionised water, vortex for 30 seconds and quick spun to collect contents at the bottom of each tube. The process of vortexing for 30 seconds and quick spinning to

collect contents at the bottom of each tube was done repeatedly until complete rehydration was achieved.

The A_{260} in each sample was measured. It was important to consume as little sample as possible when performing this measurement. NimbleGen recommended the use of a NanoDrop spectrophotometer (www.nanodrop.com) that allows the measurement of a 1 μ l sample without dilution and the typical yields range from 10 μ g to 30 μ g per reaction.

Hybridization of Cy-labeled ChIP Samples

The MAUI hybridization unit was set to 42°C and time was allowed for the temperature to stabilise. Based on the A_{260} measurement performed in the sample labeling step, 12 μ g of each of the ChIP and input DNA samples were combined into a single 1.5ml microcentrifuge tube. The combined contents were dried in a Speed-Vac on low heat and then resuspended in 3.5 μ l VWR water and vortex to completely dissolve the sample. The tube was spun down briefly to collect the contents in the bottom. Using the NimbleGen Array Reuse Kit, 3.5 μ l of Cy labeled test & reference samples, 31.5 μ l of 2X Hybridization Buffer, 9 μ l of Hybridization Component A, 0.5 μ l of 50nM Cy3 CPK6 50mer Oligo and 0.5 μ l of 100nM Cy5 CPK6 50mer Oligo were added to the resuspended sample, making up a total volume of 45 μ l. The tube was mixed briefly and spun down to collect the contents in the bottom and placed at 95°C for 5 minutes. Thereafter, the tube was immediately transferred to the MAUI 42°C sample block and was held at this temperature until sample loading is ready to be done. The MAUI Hybridization Chamber was placed on the array using the provided assembly/disassembly jig and the MAUI setup instructions were followed carefully. The sample was loaded using the pipet supplied with the MAUI Station.

During loading, a small amount (3-7 μ l) of the sample may flow out of the outlet port. It was important to ensure that there were no bubbles in the chamber by careful loading and remove any air in the pipet tip. If bubbles was present, very gently massaged the bubbles to either of the ends, away from the center of the array. Avoid applying too much pressure since this would force liquid out of the ports. The loaded array was placed into one of the four MAUI bays and allowed to equilibrate for 30 seconds. Any sample leakage at the ports was wiped off with a Kim-Wipe, and MAUI stickers were adhered to both ports. The bay clamp was closed and mix mode B was selected. The mix button was held down to start mixing. The mixing was in progress before the cover was closed. The sample was left to be hybridised overnight (16-20 hours).

Washing of Arrays

Before the array was removed from the MAUI Hybridization Station, the following solutions were prepared: two 250 ml dishes of Wash I (225ml of VWR water, 25ml of 10X Wash Buffer I and 25 μ l of 1M DDT), one 250ml of Wash II (225ml of VWR water, 25ml of 10X Wash Buffer II and 25 μ l of 1M DDT) and one 250ml of Wash III (225ml of VWR water, 25 μ l of 10X Wash Buffer III and 25 μ l of 1M DDT). One of the dishes of Wash I had to be shallow and wide enough to accommodate the array and mixer loaded in the MAUI assembly/disassembly jig. The other dish of Wash I, dish of Wash II and dish of Wash III solutions were placed in 300ml wash tanks. The chip from MAUI Hybridization Station was removed, loaded back into the MAUI assembly/disassembly jig and immersed in the shallow dish of 250ml Wash I. While the chip was submerged, the lid of the dish was carefully peeled off. The chip was then gently agitated in Wash 1 for 10-15 seconds. The slide was transferred into a slide rack in the second dish of Wash I and incubated for 2 minutes

with agitation. It was then transferred to the dish of Wash II and incubated for 1 minute with agitation. The dish was rocked to move the wash over the tops of the arrays. Thereafter, it was transferred to the dish of Wash III and incubated for 15 seconds with agitation. The array was removed and spun dry in an array-drying unit (e.g. the NimbleGen Array-Go-Round) for 1 minute. The dried array was then stored in a dark desiccator and the protocol for NimbleGen Two-Color Scanning of NimbleGen Arrays was followed immediately.

Two-Colour Array Scanning

The hybridized NimbleGen two-colour arrays were kept in a dark desiccator before they were scanned. The slide was placed in the slide carriage so that the array was faced down and the barcode end was closest to me. When the slide was lying flat on the right side of the carriage and held firmly, the scanning could then begin. While scanning, the zoom was set to view the whole image, the brightness and contrast of the displayed image was adjusted to eliminate visible saturation. The PMT setting was adjusted as appropriate so that the array features should be mostly yellow. It was then zoomed into a region as large as possible to get an accurate reading for the whole array. The histogram was calculated from all regions bounded by the current view, but ratios were selected for array areas only. The global intensity of the features was checked. It was made sure that the 532 and 635 wavelength boxes were checked so both wavelength histograms were displayed. Under Y-Axis, the Log Axis box was also checked. The red or green PMT setting was adjusted as appropriate so that the red and green curves were superimposed or as close as possible to each other. When the PMT setting was properly adjusted, the scan was restarted and allowed to run completely at the new settings. The scan was completed when the histogram was no longer changing. The images for the 532nm and 635nm wavelengths were saved as

single image .tif files as both images were needed to calculate test vs. reference ratios. After scanning, the slide was removed from the scanner and stored in a dark desiccator in the event that rescanning was necessary.

Process NimbleChip Microarray Slides to Remove Sample (Strip array)

All the steps to strip samples from the NimbleChip microarray slides were performed without interruption.

The following is the preparation of processing stations. A water bath was filled and equilibrated to 45°C. One to four slide processing containers were filled with 25ml of Array Reuse Solution. The slide processing containers were placed in a conical tube rack. The conical tube rack was submerged in the water bath and incubated for 30 minutes at 45°C to warm the Array Solution to 45°C. Each of the two processing tanks was filled with 200ml of purified water at room temperature and labeled as “processing tank 1” and “processing tank 2” respectively.

Next, each slide was placed in the array processing rack. The array processing tank was submerged in processing tank 1 containing 200ml of purified water and incubated for 5 minutes at room temperature. The array processing rack was agitated vigorously for the first minute of incubation. The array processing rack was lifted from processing tank 1 and water was allowed to drain off the rack for 5 seconds. Each slide was transferred to a slide processing container in the conical tube rack submerged in the water bath. The conical tube rack was incubated for 2 hours at 45°C before it was agitated gently for the first minute of incubation. The array processing rack was placed in processing tank 2 containing 200ml of purified water. Immediately, the slides were transferred from the slide processing containers to the array processing rack in processing rack 2 and incubated for 5 minutes at room temperature. The array

processing rack was agitated vigorously for the first minute of incubation. During the 5-minute incubation period, water from processing tank 1 was discarded and the tank was rinsed with clean purified water. The tank was then filled again with 200ml of clean purified water at room temperature. The array processing rack was transferred to processing tank 1 containing purified water and tank 1 was incubated for 5 minutes at room temperature. The array processing rack was agitated vigorously for the first minute of incubation.

Thereafter, it was the drying of slides one at a time, step-by-step until all the slides were dried. To dry the slides, the slide was removed from the array processing rack and spun dry in the ArrayIt Microarray High-Speed Centrifuge for 1 minute. The slide was removed and the edges were blotted to remove any residual moisture. The slide was stored in the original case in a dark desiccator at room temperature until ready for use.

Decision tree

Decision tree (Quinlan, J. R. 1993) is utilized to define conditions for the classification of transcription factor binding or non-binding events. Decision tree induction is one of the data mining methods to discover relationships from a set of data and present the data structure in the form of tree for easy visualization. (Quinlan, J. R. 1993) The constructed decision tree is subsequently used for predicting the outcome, which can be either categorical or numerical. All nodes of the tree except leaf nodes have splits, evaluating the values on the data attributes. Each leaf node carries a class label. As such, Decision rules can be inferred from the tree by linking the conditions along the paths from the top node to each leaf node.

Motif Analysis

Weeder program was used for the de novo prediction of the motif (Pavesi, Mereghetti et al. 2004). Subsequently, the obtained motif was drawn using Weblogo (<http://weblogo.berkeley.edu/>).

7.2 Materials and Methods for Affymetrix Array

RNA isolation

Using the protocol – Column RNA isolation.pdf , RNA was extracted from MCF-7 breast tumor cells subjected to different treatment conditions: Vehicle (DMSO), Estradiol (E2), SERM and SERM + E2. Concentration of E2 is 10nM while that of SERMs was 1 μ M. MCF-7 breast tumor cells were serum starved for 3 days, followed by treatments for 0, 3, 6, 9, 12, 24 and 48 hours. SERMs used in the studies included tamoxifen, raloxifene and ICI 182,780. After extraction of RNA, the RNA was hybridized to Affymetrix U133 plus human array following the manufacturer's protocol. (expression_analysis_technical_manual.pdf). The steps are summarized as below:

Affymetrix protocols

One-Cycle cDNA Synthesis

First-strand cDNA Synthesis: 5 μ g of RNA, 2 μ l of the appropriately diluted poly-A RNA controls, 2 μ l of 50 μ M T7-Oligo(dT) Primer were mixed in a 0.2ml PCR tube. RNase-free Water was added to a final volume of 12 μ l. The tube was gently flicked a few times to mix and then centrifuged briefly (~ 5 seconds) to collect the reaction at the bottom of the tube. The reaction was incubated for 10 minutes at 70°C. The

sample was cooled at 4°C for at least 2 minutes. The tube was centrifuged briefly (~ 5 seconds) to collect the sample at the bottom of the tube. In a separate tube, First-Strand Master Mix for all the RNA samples was prepared. The First-Strand Master Mix (4µl of 5X 1st Strand Reaction Mix, 2µl of 0.1M DDT, 1µl of 10mM dNTP) was mixed well by flicking the tube a few times. The tube was centrifuged briefly (~ 5 seconds) to collect the master mix at the bottom of the tube. 7µl of First-Strand Master Mix was transferred to each RNA/T7-Oligo(dT) Primer Mix for a final volume of 19µl. It was mixed thoroughly by flicking the tube a few times. The tube was centrifuged briefly (~ 5 seconds) to collect the reaction at the bottom of the tube. The tubes were immediately placed at 42°C and incubated for 2 minutes at 42°C. 1µl of SuperScript II was added to each RNA sample for a final volume of 20µl. It was mixed thoroughly by flicking the tube a few times. The tube was centrifuged briefly (~ 5 seconds) to collect the reaction at the bottom of the tube. The tubes were immediately placed at 42°C and incubated for 1 hour at 42°C. The sample was then cooled for at least 2 minutes at 4°C. After incubation at 4°C, the tube was centrifuged briefly (~ 5 seconds) to collect the reaction at the bottom of the tube and the next step was performed immediately.

Second-strand cDNA Synthesis: In a separate tube, Second-Strand Master Mix for all the samples was prepared. The Second-Strand Master Mix (91µl of RNase-free Water, 30µl of 5X 2nd Strand Reaction Mix, 3µl of 10mM dNTP, 1µl of *E. coli* DNA ligase, 4µl of *E. coli* DNA polymerase I, 1µl of RNase) was mixed well by flicking the tube a few times. The tube was centrifuged briefly (~ 5 seconds) to collect the solution at the bottom of the tube. 130µl of Second-Strand Master Mix was added to each first-strand synthesis sample for a total volume of 150µl. The tube was gently flicked a few times to mix and then centrifuged briefly (~ 5 seconds) to collect the

reaction at the bottom of the tube. It was then incubated for 2 hours at 16°C. 2µl of T4 DNA Polymerase was added to each sample and incubated for 5 minutes at 16°C. After incubation with T4 DNA Polymerase, 10µl of 0.5M EDTA was added and steps were carried out to clean up the double-stranded cDNA.

Cleanup of Double-Stranded cDNA

600µl of cDNA Binding Buffer was added to the double-stranded cDNA synthesis preparation and mixed by vortexing for 3 seconds to obtain a yellow mixture. 500µl of the sample was applied to the cDNA Cleanup Spin Column sitting in a 2ml Collection Tube and centrifuged for 1 minute at $\geq 10,000$ rpm. Flow-through was discarded. The spin column was reloaded with the remaining mixture and centrifuged as above. Flow-through and Collection Tube were discarded. The spin column was transferred into a new 2ml Collection Tube. 750µl of the cDNA Wash Buffer was pipetted onto the spin column. It was centrifuged for 1 minute at $\geq 10,000$ rpm. Flow through was discarded. The cap of the spin column was opened and centrifugation was carried out for 5 minutes at maximum speed ($\leq 25,000 \times g$). Flow-through and Collection Tube were discarded. The spin column was transferred into a 1.5ml Collection Tube. 14µl of the cDNA Elution Buffer was pipetted directly onto the spin column membrane, incubated for 1 minute at room temperature and centrifuged for 1 minute at maximum speed ($\leq 25,000 \times g$) to elute. Next, synthesis of Biotin-Labeled cRNA was carried out.

Synthesis of Biotin-Labeled cRNA for One-Cycle Target Labeling Assays

~ 12µl of template cDNA was transferred to RNase-free microfuge tubes and 4µl of 10X IVT Labeling Buffer, 12µl of IVT Labeling NTP Mix, 4µl of IVT Labeling Enzyme Mix were added. RNase-free Water was added to give a final volume of 40µl. The reagents were mixed and the mixture was collected at the bottom of the tube by

brief (~ 5 seconds) microcentrifugation. The tube was incubated in oven incubator at 37°C for 16 hours. The next step was the cleanup of biotin-labeled cRNA.

Cleanup of Biotin-Labeled cRNA

60µl of RNase-free Water was added to the IVT reaction and mixed by vortexing for 3 seconds. 350µl of IVT cRNA Binding Buffer was then added to the sample and mixed by vortexing for 3 seconds. 250µl of ethanol (96-100%) was added to the mixture and mixed well by pipetting. 700µl of sample was applied to the IVT cRNA Cleanup Spin Column sitting in a 2ml Collection Tube and it was centrifuged for 15 seconds at $\geq 10,000$ rpm. Flow-through and Collection Tube were discarded. The spin column was transferred into a new 2ml Collection Tube. 500µl of IVT cRNA Wash Buffer was pipette onto the spin column and centrifuged for 15 seconds at $\geq 10,000$ rpm to wash. Flow-through was discarded. 500µl of 80% (v/v) ethanol was then pipette onto the spin column and centrifuged for 15 seconds at $\geq 10,000$ rpm and flow-through was discarded. The cap of the spin column was opened and centrifuged for 5 minutes at maximum speed of ($\leq 25,000 \times g$). Flow-through and Collection Tube were discarded. Columns were placed into the centrifuge using every second bucket. Caps were positioned over the adjoining bucket so that they were oriented in the opposite direction to the rotation to avoid damage of the caps. The spin column was transferred into a new 1.5ml Collection Tube. 11µl of RNase-free Water was pipetted onto the spin column membrane. It was ensured that water was dispensed directly onto the membrane. Centrifugation was done for 1 minute at maximum speed of ($\leq 25,000 \times g$) to elute the cRNA.

Fragmenting the cRNA for Target Preparation

The Fragmentation Buffer has been optimised to break down full-length cRNA to 35 to 200 base fragments by metal-induced hydrolysis. 20µg of cRNA, 8µl of 5X

Fragmentation Buffer were mixed and RNase-free Water was added to give a total volume of 40 μ l. The reaction mixture was incubated at 94°C for 35 minutes and then put on ice. An aliquot of it was used for analysis on the Bioanalyser. Undiluted, fragmented sample of cRNA was stored at -20°C until ready to perform the hybridization.

Eukaryotic Target Hybridisation

15 μ g of Fragmented and Labeled cRNA, 5 μ l of 3nM Control Oligonucleotide B2, 15 μ l of 20X Eukaryotic Hybridisation Controls (*bioB*, *bioC*, *bioD*, *cre*), 150 μ l of 2X Hybridisation Mix, 30 μ l of DMSO were mixed and Nuclease-free Water was added to give a final volume of 300 μ l. Probe array was equilibrated to room temperature immediately before use. The hybridization cocktail was heated to 99°C for 5 minutes in a heat block. Meanwhile, the array was wet with 200 μ l of Pre-Hybridisation Mix by filling it through one of the septa. The probe array was then incubated at 45°C for 10 minutes with rotation. The hybridization cocktail that has been heated at 99°C, was transferred to a 45°C heat block for 5 minutes before spinning it at maximum speed in a microcentrifuge for 5 minutes to collect any insoluble material from the hybridization mixture. The array was removed from the hybridization oven and was vented with a clean pipette tip. The Pre-Hybridisation Mix was extracted from the array with a micropipettor. The array was refilled with 200 μ l of the clarified hybridization cocktail, avoiding any insoluble matter at the bottom of the tube. The probe array was then placed into the hybridization oven set to 45°C. To avoid stress to the motor, probe arrays were loaded in a balanced configuration around the axis and rotated at 60 rpm. Hybridisation was carried out for 16 hours.

Eukaryotic Arrays Washing, Staining and Scanning

A. Preparing the stain reagents (the day before)

The bottles of reagents were mixed. 600µl Stain cocktail 1 was aliquot into 1.5ml tubes (x 2): POSITION 1 & 3. 600µl Stain cocktail 2 was aliquot into a 1.5ml tube: POSITION 2. 800µl Array holding buffer was aliquot into another 1.5ml tube. The tubes were spun to remove bubbles. The rack was wrapped in aluminum foil and kept at 4°C. Wash A and Wash B were filtered and kept at 4°C (1 set for every 4 chips).

B. Probe array wash

Hybridization cocktail was removed with a pipettor and stored in a tube at -80°C. The probe array was refilled with 300µl of Array holding buffer.

C. Priming the fluidics station

In the Fluidics window, the Fluidic station(s) to be used, were selected. The Fluidics station (left side of the machine) was turned on. The respective media bottles were filled with the correct solutions (MiliQ water, Wash A & Wash B) and it was ensured that there are fresh tubes in every station. The lever was pulled downwards towards the tubes. Protocol: “Prime 450” was selected for “all modules”.

E1. Probe array stain

After priming was completed, the tubes were removed. Stain cocktail 1 was loaded onto positions 1 & 3, Stain cocktail 2 onto position 2. In Fluidics window, “Experiment name”, “Probe array type” & Protocol = EukGE_WS2v5_450 for individual module were selected. The array probe was put into the respective cartridge holder with the label side facing out. “Run” was clicked for individual module to begin wash & stain. (~1.5hrs)

E2. Probe array wash

When the protocol was completed, the probe arrays were removed and any large bubbles or air pockets were checked. (If bubbles were present, the Array holding

buffer was removed using a pipettor and a fresh buffer was loaded). The tubes were removed and replaced with fresh tubes. The lever was pulled down to end the protocol.

E3. Shutdown

The new tubes were replaced again. Wash A & B were removed and the tubings were placed into the MiliQ bottle, before topping the bottle up with water. In Fluidics window, Protocol = Shutdown_450 for “all modules” was selected and run for 3 times before removing the waste solutions.

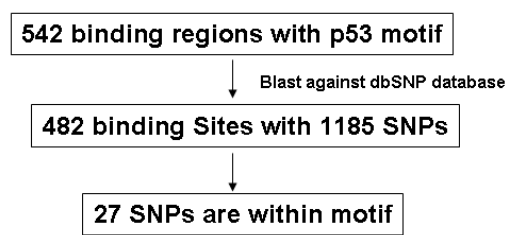
F. Probe array scan

The scanner was warmed up for 15 minutes. The 1st cartridge was inserted into the red slot of the scanner, the next 1 was in front of it (away from user). The Start Scan icon was selected. It took ~7 minutes per scan.

G. Analysis

The image file was opened (D drive>program files>Affymetrix>genechip>affy data>data). Using the Grid, the runway lights was checked (=oligoB2 controls) in all 4 corners and for any artifacts. Analysis was selected, followed by File→ Report to create a report file.

7.3 Materials and Methods for Functional Studies



The studies began with SNPs detection in p53 binding sites. Region of interests were found in 542 loci with high confidence of p53 interaction. Blasting against dbSNP databases, 482 binding sites were found to contain 1185 SNPS. Among these,

27 SNPs are within p53 motif. The polymorphism studies are performed in lymphoblastoid cell lines(LCLs) as they are more readily available and from greater sources compared to breast cancer cell lines. Initially, LCLs were tested for p53 response. This was followed by ChIP analysis and transcriptional activity analysis.

Samples

The current study included the clinical samples from Sweden and Finland. All the Swedish cases were randomly selected from a population-based Swedish cohort that included all Swedish-born breast and endometrial cancer patients between 50 and 74 years of age and resident in Sweden between October 1993 and March 1995. A similar number of age-matched controls were randomly selected from the Swedish Registry of Total Population. Briefly, after informed consent, 1596 breast cancer patients, 719 endometrial cancer patients and 1730 healthy volunteers participated into this study by providing either whole blood or non-malignant paraffin-embedded tissues for DNA analysis. From whole blood samples, DNAs were extracted by using the QIAamp DNA Blood Maxi Kit (Qiagen) according to the manufacturer's instruction. From non-malignant paraffin-embedded tissues, DNA was extracted using a standard phenol/chloroform/isoamyl alcohol protocol.

The Finnish breast cancer cases consist of two series of unselected breast cancer patients and additional familial cases ascertained at the Helsinki University Central Hospital. The first unselected series of 884 breast cancer patients studied were collected at the Department of Oncology, Helsinki University Central Hospital in 1997-1998 and 2000 and cover 79% of all consecutive, newly diagnosed breast cancer cases during the collection periods. 876 patients (99%) from this series were successfully genotyped in this study.

The second unselected series, containing 986 consecutive newly diagnosed breast cancer patients, were collected at the Helsinki University Central Hospital 2001 – 2004 and covered 87% of all such patients treated at the Department of Surgery during the collection period. Of this series, 979 patients (99%) were successfully genotyped.

The series of 538 additional familial breast cancer cases in this study have been collected at the Helsinki University Central Hospital as described (Eerola, Blomqvist et al. 2000). The genotyped series included 295 patients with strong family history, defined as three or more breast or ovarian cancer cases in the first or second degree family members including the index case. These families were screened negative for BRCA1/2 mutations as previously described in detail (Vehmanen, Friedman et al. 1997; Vahteristo, Eerola et al. 2001; Vahteristo, Bartkova et al. 2002). The remaining 243 genotyped familial cases had a single affected first degree family member; for 213 of these cases, the Finnish BRCA1/2 founder mutations have been excluded as described (Vahteristo, Eerola et al. 2001; Vahteristo, Bartkova et al. 2002). All the cancer diagnoses have been verified through the Finnish Cancer Registry and hospital records. Allele and genotype frequencies in the normal population were determined in 1256 healthy female population controls collected from the same geographical region.

This study was approved by the Institutional Review Boards in Sweden and Finland and the National University of Singapore.

SNP Genotyping

Genotyping analysis of SNPs was performed by using the MALDI-TOF mass spectrometry-based MassARRAY™ system from the Sequenom (San Diego, CA, US) (Swedish samples) as well as the TaqMan assays from the AppliedBiosystems (ABI) (Foster City, CA, US) (Finnish samples). All genotyping plates included positive and negative controls, DNA samples were randomly assigned to the plates, and all genotyping results were generated and checked by laboratory staff unaware of case-control status.

Lymphoblastoid cell lines and culture

All lymphoblastoid cell lines (LCLs) used in this study were obtained from the Coriell depository (<http://www.coriell.org/>). Cells were cultured in RPMI medium supplemented with 20% fetal bovine serum. For ChIP, real-time qPCR and western blot analyses, cells were treated with 5FU at the concentration of 375µM for various hours. All the drug treatments were done during the log phase of cell growth (about 1 to 1.5 millions of cells per ml). Cells were harvested after culture with or without drug treatment(s) and stored at -80°C. 5FU was obtained from the Sigma.

ChIP Analysis: ChIP assays were performed in LCLs using the protocol described previously (Weinmann and Farnham 2002; Wells and Farnham 2002). For all ChIP analyses, the DO1 monoclonal antibody for p53 (Santa Cruz Biotechnology, **Santa** Cruz, CA) was used for immunoprecipitation, and real-time quantitative PCR analyses were performed using the PRISM 7900 Sequence Detection System and the SYBR protocol as described (Ng et al 2003). The real-time PCR analysis was performed using the following primers: CCATCCTGCCTGAGCATGTCTGAAC (forward) and CCGGCTTTGCCAGACAATTGG (reverse) (For PRKAG2);

CAGGCTGTGGCTCTGATTGGCTTTC (forward) and
GCTGGCAGATCACATACCCTGTTTCAGAGTA (reverse) (For p21);
ACCCACACTGTGCCCATCTACGAG (forward) and
TCTCCTTAATGTCACGCACGATTTCC (reverse) (For Actin). The primers were designed using Vector NTI. Relative occupancy was calculated by determining the immunoprecipitation efficiency (ratios of the amount of immunoprecipitated DNA over that of the input sample) and normalized to the level observed at a control region, which was defined as 1.0. The control region was a distal site around the binding site for Actin and not enriched by the immunoprecipitation. Each real-time quantitative PCR analysis was done in triplicate.

Allele Enrichment Analysis of ChIP pull-down DNAs by real-time PCR

The allele enrichment analysis of the ChIP input and pull-down DNAs from heterozygous cell lines was performed by real-time quantitative PCR using a made-to-order TaqMan SNP assay for rs1804674 from the ABI. The quality of the TaqMan SNP assay was first verified by genotyping 30 CEPH DNA samples, and all the genotype results are consistent with the ones from the HapMap project (data not shown). For real-time PCR analysis, the Ct value difference (ΔCt) between G and T alleles of a ChIP pull-down DNA was normalized by the ΔCt value of the corresponding input DNA (reflecting the equal numbers of G and T alleles in normal genomic DNAs from the heterozygous cell lines). The normalized ΔCt value ($\Delta\Delta\text{Ct}$) was then used to calculate the enrichments (Fold Change using the formula of $2^{\Delta\Delta\text{Ct}}$) of the wild-type G allele over the mutant T allele in the ChIP pull-down DNA. All the real-time PCR analyses were done in triplicate.

Expression Analysis by Real-time PCR

Total RNAs were extracted from cells (with or without 5FU treatment) using the RNeasy Kit from the Qiagen (with DNase digestion step). 200 ng RNA was then reverse transcribed into 20 μ l cDNA using the SuperScript kit from the Invitrogen (CA, USA), and real-time PCR analysis was subsequently performed by using 2 μ l cDNA as template. All the real-time PCR analyses were done in the ABI Prism 7700 sequence detection system by using the TaqMan assays from the ABI. For PRKAG2, assay-by-demand assay was developed by using the Primer Express software from the ABI:

GTTTCCCCTGGAATCCTATAAGC (Forward),
CGAGGCATAGATGCGATTCTC (reverse) and *CGAGCCTGAACGGT* (probe).

For normalization, a ready-to-use TaqMan probe for the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene was analyzed as endogenous control. Each real-time PCR analysis was done in triplicate.

All the Ct values from the real-time PCR analyses were analyzed by using the comparative Ct method provided by the manufacturer (ABI). Briefly, the Ct values from the PRKAG2 analysis were first normalized by the Ct values of the endogenous control, GHAP. The normalized Ct (Δ Ct) values were then used to calculate the Ct value difference ($\Delta\Delta$ Ct) between 10h treatment and the baseline. Fold change in the expression of PRKAG2 between the baseline and the 10h treatment of 5FU was calculated by using the formula of $2^{\Delta\Delta\text{Ct}}$.

Promoter Assay Analysis

A 226 bp region encompassing the intronic p53 binding site within *PRKAG2* was amplified using hotstart PCR with forward primer 5'-

TAGGAGACCTGGGGGACTTT-3' and reverse primer 5'-CAGGCATCTCGAAGAGATCA -3' and 50 ng of genomic DNAs isolated from the individuals carrying either the wild-type (WT) G or mutant (MUT) A allele. The PCR conditions were; 94°C for 15 mins, followed by 35 cycles of denaturation at 94°C for 45s, annealing 55°C for 45s, and extension at 72°C for 45s. The resultant PCR products of 226 bp were purified from agarose gels and cloned using TOPO-TA cloning system (Invitrogen, Calsbad, CA). The genotypes of the cloned DNA fragments were confirmed by DNA sequencing. Subsequently, the DNA fragments were subcloned into the upstream of TATA-luciferase (fire-fly) containing pGL4 vector (Promega) using Kpn I and Xho I restriction enzymes (New England Biolabs).

Reporter assay analysis was performed by using both HCT116 wild type and null for p53 cells (provided by Dr Bert Vogelstein's lab at the Johns Hopkins School of Medicine) that were maintained in DMEM containing 10% fetal bovine serum. 5×10^4 cells were plated in triplicate in 24-well plates and transfected next day with 250 ng of either parent TATA-luc, WT-TATA-luc or MUT-TATA-luc plasmid DNAs under serum free conditions using 1 μ g per well of Lipofectamine 2000 (Invitrogen, Calsbad, CA). 2.5 ng of pRL-CMV vector containing renilla luciferase was co-transfected in each well to normalize transfection efficiency across wells. After 8 hours the cells were recovered for 3 hours in serum containing medium, following which the cells were treated for 12 hours with 375 μ M 5-Fluorouracil or DMSO. The cells were lysed in passive lysis buffer and promoter assays were carried out as per manufacturer's instructions using Promega Dual-luciferase assay system. The values obtained for each construct were normalized as fold change to that of the activity of parental TATA-luc vector in HCT116 WT cells (designated as 1).

Extraction of proteins using modified radioimmunoprecipitation (RIPA) lysis buffer

100ml of modified RIPA buffer was prepared as follows. 790mg of Tris base was added to 75ml of distilled H₂O. 90mg of NaCl was added and the solution was stirred until all solids were dissolved. The pH was adjusted to 7.4 by adding HCl. 10ml of 10% NP-40 (stored at room temperature) was added to the solution. 2.5ml of 10% Na-deoxycholate (stored at room temperature) was added and the solution was stirred until it was clear. 1ml of 100mM EDTA was then added to the solution and the volume of the solution was adjusted to 100ml using a graduated cylinder. *RIPA buffer stock was prepared and stored in the fridge at 4°C. The remaining protease and phosphatase inhibitors were added to the solution on the same day the assay was run.

Protocol for Protein Extraction with 1x Modified RIPA Buffer

After cells were harvested by trypsinisation or scraping, the cells were washed twice with ice-cold 1x PBS. Pellet cells were spun at 1200 rpm at 4°C for 4 minutes. The cell pellet was resuspended in 5 volume of 1x Modified RIPA Buffer. Cells were lysed on ice for 30 to 60 minutes where vortex was done for 15 seconds in every 10 minutes. Samples were centrifuged at 16000xg at 4°C for 15 minutes. The supernatant containing soluble proteins, was transferred to new tubes.

Preparation of stock solutions

10% NP-40: 100% NP-40 stock (stored at room temperature) was melted by heating. 10ml of 100% NP-40 solution was dissolved in 100ml of distilled H₂O, mixed well and stored at room temperature.

10% Na-deoxycholate: 10g of Na-deoxycholate salt was dissolved in 100ml of distilled H₂O, mixed by stirring and stored at room temperature, protected from light.

200mM PMSF: 3.48g of Phenylmethylsulfonyl fluoride (PMSF) salt was dissolved in 100ml of isopropanol, mixed with stirring and heating and stored at room temperature.

200mM NaF: 0.4g of sodium fluoride (NaF) was dissolved in 50ml of distilled H₂O, mixed well and stored at room temperature.

Activation of Sodium Orthovanadate

Sodium orthovanadate was activated for maximal inhibition of protein phosphotyrosyl-phosphatases. A 200mM solution of sodium orthovanadate (Na₃VO₄) was prepared by dissolving 1.8g of Na₃VO₄ salt in 50ml of distilled water. The pH was adjusted to 10.0 by using either 1M NaOH or 1M HCl. The yellow solution was boiled until it turned colourless. It was then cooled to room temperature. The pH was re-adjusted to 10.0 and the boiling and cooling of the solution to room temperature was repeated until the solution remained colourless and the pH stabilized at 10.0. The activated sodium orthovanadate was stored as aliquots at -20°C.

Protein Quantitation

1x Bradford's reagent was prepared from 5x stock solution. 1mg/ml Bovine Serum Albumin(BSA) was also prepared. Each well was loaded with 200µl of 1x Bradford's reagent. For standard curve, the lanes for the amount of protein in µg were Blank, 0.5, 1, 2, 4, 6, 8 and 10. (i.e, 0.5µl was loaded for 0.5µg or 8µl was loaded for 8µg, etc.) Next, the protein samples were loaded and the machine was used to quantitate the proteins. Finally, the standard curve was plotted and the equation obtained was used to calculate the unknown protein concentration.

SDS-PAGE

Assembly: Plates/ spacers were washed with 70% ethanol. The plates were slid into casting frame, keeping the short plate facing the front of the frame. Pressure cams were engaged to secure glass plates. The casting frame and plates were secured in the casting stand.

Pouring Resolving Gel: APS and TEMED were added just before casting. The resolving gel was poured between plates to a level ~1cm below the bottom of wells. The top level was made flat with water-saturated butanol and the gel was allowed to set for ~1 hour before the water-saturated butanol was rinsed off with milliQ water.

Pouring Stacking Gel: APS and TEMED were added just before casting and it was poured between plates until they were full. The plates were then inserted in a comb carefully, ensuring that no air bubbles were trapped. ~1 hour was allowed for gel to polymerize.

Loading Samples and Gel Electrophoresis: Casting plates were unclipped from stand and clipped to the U-shaped gaskets, making sure that the short plate faced inward toward the notches of the U-shaped gaskets. The whole thing was lowered into the tank and the inner chamber was filled until full with running buffer. The samples were heated at 95°C for 5 minutes in 1x sample loading buffer. Samples (usually 10µg) were loaded into wells. 5µl of protein molecular weight marker was also loaded into 1 lane. The rest of buffer was poured into the lower buffer chamber, making sure the buffer was 1cm above bottom of plates. For 1 run, 500ml 1x running buffer was prepared and the power was connected to run gel at constant voltage of 100V for 1.5 hours.

Western Blot Analysis

Total protein was extracted from cells using the Modified RIPA buffer. The Micro BCA Protein Assay Reagent Kit (Pierce, Rockford, IL, U.S.A) was used to quantify protein concentration. Western blot was performed using 40µg of protein using the established protocol and the following antibodies: 1) antibody for actin (control, 1:5000 dilution), 2) p53(DO-1) sc-126 (Santa Cruz Biotechnology, 1:1000 dilution); 3) AMPK α , Phospho-AMPK α (Thr172) antibodies for both the total- and phosphor-AMPK proteins (Cell Signaling technology, 1:1000 dilution), and 4) AMPK γ 2 antibody (Cell Signaling technology, 1:1000 dilution) .

Preparation for blotting: The gels after electrophoresis were equilibrated in transfer buffer for 15 minutes at room temperature. The membranes were cut to 8.8mm by 6.2mm and wet in 100% methanol for 15 seconds, followed by ultrapure water for 2 minutes and lastly soaked in transfer water for 15-30 minutes. Two thick and two thin filter papers were also cut to 8.8mm by 6.2mm and were soaked in transfer buffer.

Assemble the transfer stack: The safety cover and the stainless steel cathode assembly were removed. 1 thick and 1 thin pre-soaked filter papers were placed onto the platinum anode and a 50ml tube was rolled over the surface of the papers to exclude all air bubbles. Pre-wetted Immobilon-P membrane (transfer membrane) was placed on top of the filter papers followed by equilibrated gel on top of the transfer membrane. 1 thick and 1 thin filter papers were then placed on top of the equilibrated gel. Finally, the cathode assembly and the safety cover were placed back. The power was turned on and the gel was transferred at 15V for 1 hour. After transferring, the blotted membrane was dried by soaking it in 100% methanol for 10 seconds, followed by drying on top of a piece of filter paper.

Immunodetection: (1) Blocking- Non-specific binding sites were blocked by immersing the membrane in 5% non-fat dried milk, 0.1%(v/v) Tween 20 in PBS for 1 hour at room temperature on an orbital shaker. (2) Washing – The membrane was rinsed with washing buffer (PBS-T) for 2 minutes and repeated once. (3) Binding of Primary Antibody (Ab)- The primary antibody (anti-goat antibody) was diluted in blocking solution, i.e. 2.5 μ l Ab in 2ml blocking solution. The membrane was incubated in diluted primary antibody on an orbital shaker at room temperature for 1 hour. (4) Washing – The membrane was briefly rinsed with 2 changes of washing buffer. It was then soaked in washing buffer in a rotary shaker for 15 minutes. The membrane was then rinsed with wash buffer in rotary shaker for 5 minutes and was repeated twice. (5) Binding of Secondary Antibody- The secondary antibody (anti-goat antibody) was diluted in blocking solution, i.e. 1.5 μ l Ab in 2ml blocking solution. The membrane was incubated in diluted secondary antibody on an orbital shaker at room temperature for 1 hour. (6) Washing – The membrane was briefly rinsed with 2 changes of washing buffer. It was then soaked in washing buffer in a rotary shaker for 15 minutes. The membrane was then rinsed with wash buffer in rotary shaker for 5 minutes and was repeated twice. (7) Detection – 2ml of solution 1 and 50 μ l of solution 2 were mixed to get the detection reagent ready. Excess wash buffer was drained from the washed membrane and the membrane with the protein side faced up was placed on a large glass panel. The protein side of the membrane was covered with the detection reagent and incubated for exactly 1 minute. Excess detection reagent was drained off by holding the membrane vertically and letting its edge to touch a tissue paper. The membrane with protein side faced down was then gently placed onto a transparency and covered by another transparency. Any air pocket was gently smoothed out. The blots with protein side up was placed in the film cassette. Lights

were switched off and a sheet of auto radiography film such as Hyperfilm ECL was carefully placed on top of the membranes. The cassette was closed and exposed for 8 seconds with the exposure duration adjusted accordingly. The film was removed and developed using the machine.

Statistical Analysis

Hardy-Weinberg Equilibrium (HWE) test was performed in the Finnish and Swedish control samples separately, and no evidence for deviation from HWE was found. Association analysis was performed using the χ^2 test under a recessive model of inheritance. For the joint association analyses of the combined Swedish-Finnish breast cancer sample and the combined breast-endometrial cancer sample, the Mantel-Haenszel method for meta-analysis was used by assuming fixed effect. For the joint analysis of the breast-endometrial sample, the Swedish cases were defined as having either breast or endometrial cancer. All statistical analyses were performed by using the StataSE8 system.

Appendices

Appendix A: Enrichment Results for 24 validated ERE binding sites

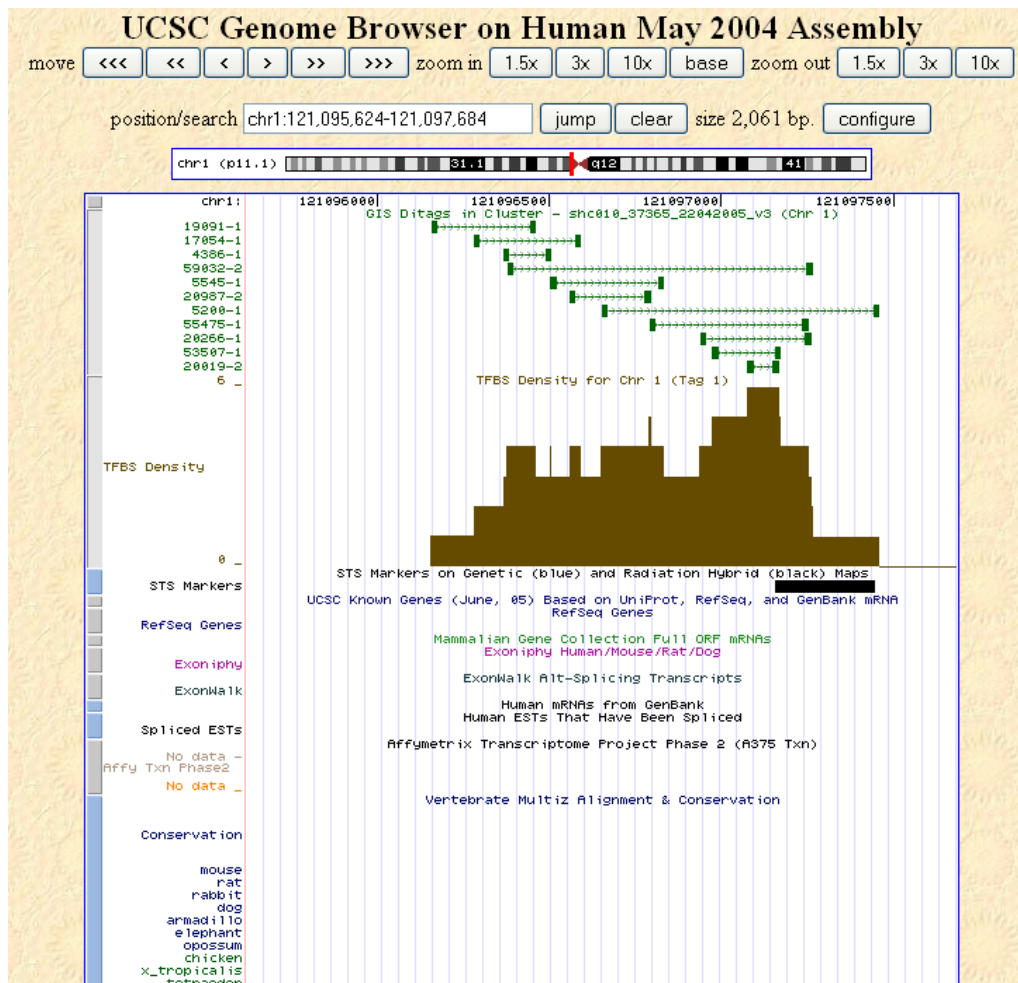
Acc	Gene Name	Chr.	Sequence	(ERα/GST) _{+E2}	(ERα/GST) _{-E2}	(ERα/GST) _{RA+E2}	(ERα/GST) _{+RA}	(ERα/GST) _{TA+E2}	(ERα/GST) _{+TA}	(ERα/GST) _{ICI+E2}	(ERα/GST) _{+ICI}
NM_003489	NRIP1	chr21	GGTC aa gTG CC	16 ± 1	4 ± 2	7 ± 1	12 ± 1	10 ± 2	9 ± 5	6 ± 5	4 ± 1
NM_001657	AREG	chr4	GGAC agg TG TCC	2 ± 1	1 ± 0	2 ± 1	1 ± 0	2 ± 1	1 ± 0	2 ± 0	4 ± 1
NM_003225	TFF1	chr21	GGTC acgg TG CC	103 ± 36	8 ± 4	38 ± 10	22 ± 12	13 ± 6	19 ± 5	60 ± 42	24 ± 12
NM_004878	PTGES	chr9	GGAC agcc TG CC	29 ± 5	4 ± 2	17 ± 7	9 ± 8	20 ± 10	16 ± 5	13 ± 6	5 ± 0
NM_001089	ABCA3-F1	chr16	GGTC acgg TG TTC	135 ± 8	17 ± 6	44 ± 18	34 ± 12	34 ± 19	32 ± 10	60 ± 21	24 ± 4
NM_017770	ELOVL2-F3	chr6	GGTC atct TG ATG	22 ± 2	10 ± 4	19 ± 14	8 ± 6	15 ± 15	20 ± 13	16 ± 0	7 ± 1
NM_148903	GREB1-F1	chr2	AGTCA gtgT CA CC	171 ± 42	39 ± 0	45 ± 8	49 ± 6	38 ± 8	26 ± 16	67 ± 38	34 ± 3
NM_148903	GREB1-F2	chr2	GGTC atct TG CC	194 ± 28	95 ± 30	106 ± 44	75 ± 12	100 ± 41	137 ± 56	184 ± 67	48 ± 3
NM_000029	AGT	chr1	GG CA tctgTG CC	2 ± 1	1 ± 0	1 ± 0	1 ± 1	2 ± 0	2 ± 1	1 ± 0	1 ± 1
NM_000633	BCL2	chr18	GGTC Gcca GG CC	1 ± 0	1 ± 0	1 ± 1	1 ± 0	1 ± 0	1 ± 0	2 ± 1	1 ± 0
NM_001909	CTSD	chr11	GGCC Gggc TG CC	5 ± 3	2 ± 0	3 ± 1	1 ± 1	3 ± 2	1 ± 0	2 ± 1	1 ± 1
NM_000064	C3	chr19	GGT GG cccTG CC	5 ± 2	1 ± 0	2 ± 1	3 ± 2	4 ± 2	2 ± 1	3 ± 0	2 ± 0
NM_002343	LTF	chr3	GGTC aggc GA TCC	3 ± 1	1 ± 0	2 ± 1	1 ± 1	2 ± 1	1 ± 0	3 ± 2	3 ± 1
NM_005082	EFP	chr17	GGTC atgg TG CC	46 ± 10	6 ± 1	13 ± 4	12 ± 6	16 ± 5	15 ± 0	13 ± 1	5 ± 1
NM_000926	PGR-F2	chr11	GCA GG agc TG CC	2 ± 0	1 ± 0	1 ± 0	1 ± 1	1 ± 0	1 ± 1	1 ± 0	1 ± 0
NM_000926	PGR-F3	chr11	GGTC acca G CTCT	3 ± 1	1 ± 0	1 ± 1	1 ± 1	2 ± 0	1 ± 0	1 ± 0	2 ± 0
NM_003376	VEGF	chr6	AATCA gacTG ACT	1 ± 0	1 ± 0	1 ± 0	1 ± 1	1 ± 0	1 ± 0	1 ± 0	0 ± 0
NM_000674	ADORA1	chr1	GGT T agggTG CC	164 ± 54	26 ± 12	47 ± 10	45 ± 10	86 ± 46	53 ± 4	14 ± 0	23 ± 5
NM_001227	CASP7-F1	chr10	GGTC aggg TG AAC	49 ± 1	18 ± 4	24 ± 4	27 ± 11	40 ± 19	43 ± 7	23 ± 9	11 ± 1
NM_001227	CASP7-F7	chr10	GGTC aggg TG AAC	143 ± 71	25 ± 6	33 ± 3	37 ± 4	82 ± 39	48 ± 22	27 ± 9	15 ± 2
NM_148903	GREB1-F5	chr2	GGTC aaa TG CC	26 ± 7	5 ± 0	13 ± 2	11 ± 4	44 ± 29	31 ± 21	12 ± 4	3 ± 0
NM_148903	GREB1-F6	chr2	GGTC atca TG CC	31 ± 2	6 ± 0	12 ± 1	10 ± 4	16 ± 2	13 ± 0	17 ± 5	4 ± 1
NM_001552	IGFBP4-F3	chr17	GGTC attg TG CA	4 ± 1	1 ± 0	2 ± 1	2 ± 1	3 ± 1	2 ± 1	2 ± 1	3 ± 2
NM_002346	LY6E	chr6	GGAC aga TG CC	75 ± 22	8 ± 5	32 ± 3	40 ± 8	24 ± 6	47 ± 20	39 ± 9	12 ± 4

Appendix B: Enrichment Results for 27 validated ERE non-binding Sites

Acc	Gene Name	Chr.	Sequence	(ERα/GST) _{+E2}	(ERα/GST) _{-E2}	(ERα/GST) _{RA+E2}	(ERα/GST) _{-RA}	(ERα/GST) _{TA+E2}	(ERα/GST) _{-TA}	(ERα/GST) _{ICI+E2}	(ERα/GST) _{-ICI}
NM_000427	LOR	chr1	GGTC caaa GGACC	1±0	1±0	2±2	1±0	1±0	1±0	1±0	1±0
NM_000602	SERPINE1-F2	chr7	GGCA agc TGCC	1±0	1±1	1±2	1±0	1±0	1±0	1±0	1±0
NM_000602	SERPINE1-F3	chr7	TGTC aga AGACC	2±0	2±0	1±0	1±0	1±0	1±0	1±0	1±0
NM_000609	CXCL12-F1 or SDF1-F1	chr10	GGTC cagc TGCC	2±0	2±0	1±0	1±0	1±0	1±1	1±0	1±0
NM_000609	CXCL12-F2 or SDF1-F2	chr10	TGTC aaa TGGCC	1±0	1±0	1±0	1±0	1±0	1±0	1±0	1±0
NM_000712	BLVRA-F1	chr7	AGTC acc TTACC	1±0	1±0	1±0	0±0	1±0	1±0	1±0	1±0
NM_000712	BLVRA-F4	chr7	GGTC actc TGGCT	1±0	1±0	2±0	2±0	2±0	2±0	2±0	1±0
NM_000926	PGR-F1	chr11	AGTC atgt TG&CA	1±0	0±0	2±0	1±0	2±0	2±0	2±0	1±0
NM_001089	ABCA3-F3	chr16	GGTC ttt TTACC	4±1	2±0	2±1	2±2	2±1	2±2	2±0	1±0
NM_001116	ADCY9-F1	chr16	GGTC agge TGGTC	0±0	1±0	2±1	1±1	1±0	1±0	1±0	1±0
NM_001116	ADCY9-F2	chr16	GGTC aaa TGTCC	1±0	1±0	2±0	1±0	1±0	1±0	1±0	1±0
NM_001497	B4GALT1-F1	chr9	GATC gaa GGACC	1±0	2±0	2±1	1±0	1±0	1±0	1±0	1±0
NM_001552	IGFBP4-F1	chr17	GATC actg TAACC	1±0	1±0	1±0	0±0	1±0	1±0	1±0	1±0
NM_003646	DGKZ-F1	chr11	GGCC atgc TGGCC	1±0	1±0	2±0	1±0	1±0	1±0	1±0	1±0
NM_004354	CCNG2	chr4	GGCA act TGATC	1±0	1±0	2±0	1±0	1±0	1±0	1±0	1±0
NM_005067	SIAH2	chr3	GCTC atag TGCC	2±1	1±0	2±2	1±0	1±0	1±0	1±0	1±0
NM_006472	TXNIP	chr1	GGTC agtg GGATC	1±0	1±0	3±1	2±0	2±0	2±0	2±0	2±0
NM_014583	LMCD1 F2	chr3	GGCC tgc aTGACC	5±1	2±0	3±2	3±0	2±1	2±1	2±0	2±0
NM_020120	UGCGL1-F1	chr2	TGTC aaa TGTCC	1±0	1±0	1±0	1±0	1±0	1±0	1±0	1±0
NM_020120	UGCGL1-F2	chr2	TGTC acat TGAGC	1±0	1±0	1±0	1±0	1±0	1±0	1±0	1±0
NM_022365	DNAJC1	chr10	GITC act TGTCC	1±0	1±0	1±2	1±0	1±0	1±0	1±0	1±0
NM_024524	AFUR51 FLJ20986	chr3	GGTC atta ATACC	1±0	1±0	1±0	1±0	1±0	1±0	1±0	1±0
NM_024817	FLJ13710-F1	chr15	AGTC attg TTACC	1±0	1±0	1±1	0±0	1±0	1±0	1±0	1±0
NM_024817	FLJ13710-F2	chr15	GCTC actt TGTCC	1±0	1±0	1±0	0±0	1±0	1±0	1±0	1±0
NM_024817	FLJ13710-F3	chr15	GGTC aatg TG C&C	1±0	1±0	2±1	1±0	2±0	1±1	1±0	2±0
NM_032219	FLJ22269	chr4	GGC agag TG A&T	1±0	1±0	2±0	1±0	2±0	2±0	2±0	2±0
NM_148903	GREB1-F4	chr2	TGTC aate TGTCC	3±1	2±0	2±0	2±1	2±0	2±1	2±0	3±1

Appendix of Technology platforms

ChIP-PET platform: ChIP-PET data were generated in the steps below. Chromatin Immunoprecipitation (ChIP) assay with E2 treatment was performed first. Cloning was performed through a DNA tag sequencing and mapping strategy called gene identification signature (GIS) analysis, in which 5' and 3' signatures of the ChIP DNA fragments were extracted into paired-end ditags (PETs), which are subsequently concatenated for efficient sequencing and mapping to human genome (Buckland, Hoogendoorn et al. 2005).



This is an example of a cluster whose PET is 11 but moPET is 6.

Bibliography

- Acconcia, F., C. J. Barnes, et al. (2006). "Estrogen and tamoxifen induce cytoskeletal remodeling and migration in endometrial cancer cells." Endocrinology **147**(3): 1203-12.
- Ali, S. and R. C. Coombes (2000). "Estrogen receptor alpha in human breast cancer: occurrence and significance." J Mammary Gland Biol Neoplasia **5**(3): 271-81.
- Bensaad, K. and K. H. Vousden (2007). "p53: new roles in metabolism." Trends Cell Biol **17**(6): 286-91.
- Brown, A. M., J. M. Jeltsch, et al. (1984). "Activation of pS2 gene transcription is a primary response to estrogen in the human breast cancer cell line MCF-7." Proc Natl Acad Sci U S A **81**(20): 6344-8.
- Bu, P., Y. A. Evrard, et al. (2007). "Loss of Gcn5 acetyltransferase activity leads to neural tube closure defects and exencephaly in mouse embryos." Mol Cell Biol **27**(9): 3405-16.
- Buckland, P. R., B. Hoogendoorn, et al. (2005). "Strong bias in the location of functional promoter polymorphisms." Hum Mutat **26**(3): 214-23.
- Carroll, J. S., X. S. Liu, et al. (2005). "Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1." Cell **122**(1): 33-43.
- Carroll, J. S., C. A. Meyer, et al. (2006). "Genome-wide analysis of estrogen receptor binding sites." Nat Genet **38**(11): 1289-97.
- Cawley, S., S. Bekiranov, 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**(4): 499-509.
- Cheong, J. Y., S. W. Cho, et al. (2006). "Genetic Polymorphism of Interferon-gamma, Interferon-gamma Receptor, and Interferon Regulatory Factor-1 Genes in Patients with Hepatitis B Virus Infection." Biochem Genet.
- Clark, M. L., J. B. Burch, et al. (2007). "Biomonitoring of Estrogen and Melatonin Metabolites Among Women Residing Near Radio and Television Broadcasting Transmitters." J Occup Environ Med **49**(10): 1149-1156.
- Cohen, A., M. Shmoish, et al. (2008). "Alterations in micro-ribonucleic acid expression profiles reveal a novel pathway for estrogen regulation." Endocrinology **149**(4): 1687-96.
- Crute, B. E., K. Seefeld, et al. (1998). "Functional domains of the alpha catalytic subunit of the AMP-activated protein kinase." J Biol Chem **273**(52): 35347-54.
- Dion, V. and B. Coulombe (2003). "Interactions of a DNA-bound transcriptional activator with the TBP-TFIIA-TFIIB-promoter quaternary complex." J Biol Chem **278**(13): 11495-501.
- Durinovic-Bello, I., E. Jelinek, et al. (2005). "Class III alleles at the insulin VNTR polymorphism are associated with regulatory T-cell responses to proinsulin epitopes in HLA-DR4, DQ8 individuals." Diabetes **54 Suppl 2**: S18-24.
- Easton, D. F., K. A. Pooley, et al. (2007). "Genome-wide association study identifies novel breast cancer susceptibility loci." Nature **447**(7148): 1087-93.
- Eckert, R. L., A. Mullick, et al. (1984). "Estrogen receptor synthesis and turnover in MCF-7 breast cancer cells measured by a density shift technique." Endocrinology **114**(2): 629-37.

- Eerola, H., C. Blomqvist, et al. (2000). "Familial breast cancer in southern Finland: how prevalent are breast cancer families and can we trust the family history reported by patients?" Eur J Cancer **36**(9): 1143-8.
- Fitzpatrick, S. L., T. J. Berrodin, et al. (1999). "Effect of estrogen agonists and antagonists on induction of progesterone receptor in a rat hypothalamic cell line." Endocrinology **140**(9): 3928-37.
- Green, S., V. Kumar, et al. (1988). "The N-terminal DNA-binding 'zinc finger' of the oestrogen and glucocorticoid receptors determines target gene specificity." Embo J **7**(10): 3037-44.
- Greene, G. L., P. Gilna, et al. (1986). "Sequence and expression of human estrogen receptor complementary DNA." Science **231**(4742): 1150-4.
- Guo, Y. and D. C. Jamison (2005). "The Distribution of SNPs in Human Gene Regulatory Regions." BMC Genomics **6**(1): 140.
- Heintzman, N. D., R. K. Stuart, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet **39**(3): 311-8.
- Horak, C. E., M. C. Mahajan, et al. (2002). "GATA-1 binding sites mapped in the beta-globin locus by using mammalian ChIP-chip analysis." Proc Natl Acad Sci U S A **99**(5): 2924-9.
- Howell, A., C. K. Osborne, et al. (2000). "ICI 182,780 (Faslodex): development of a novel, "pure" antiestrogen." Cancer **89**(4): 817-25.
- Ihaka R. & Gentleman R. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics **5**: 299-314
- Imhof, M. O. and D. P. McDonnell (1996). "Yeast RSP5 and its human homolog hRPF1 potentiate hormone-dependent activation of transcription by human progesterone and glucocorticoid receptors." Mol Cell Biol **16**(6): 2594-605.
- Impey, S., S. R. McCorkle, et al. (2004). "Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions." Cell **119**(7): 1041-54.
- Irizarry, R. A., S. L. Ooi, et al. (2003). "Use of mixture models in a microarray-based screening procedure for detecting differentially represented yeast mutants." Stat Appl Genet Mol Biol **2**: Article1.
- Jones, R. G., D. R. Plas, et al. (2005). "AMP-activated protein kinase induces a p53-dependent metabolic checkpoint." Mol Cell **18**(3): 283-93.
- Kaesler, M. D. and R. D. Iggo (2002). "Chromatin immunoprecipitation analysis fails to support the latency model for regulation of p53 DNA binding activity in vivo." Proc Natl Acad Sci U S A **99**(1): 95-100.
- Klinge, C. M. (1999). "Role of estrogen receptor ligand and estrogen response element sequence on interaction with chicken ovalbumin upstream promoter transcription factor (COUP-TF)." J Steroid Biochem Mol Biol **71**(1-2): 1-19.
- Konev, A. Y., M. Tribus, et al. (2007). "CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin in vivo." Science **317**(5841): 1087-90.
- Kuiper, G. G., E. Enmark, et al. (1996). "Cloning of a novel receptor expressed in rat prostate and ovary." Proc Natl Acad Sci U S A **93**(12): 5925-30.
- Kushner, P. J., D. A. Agard, et al. (2000). "Estrogen receptor pathways to AP-1." J Steroid Biochem Mol Biol **74**(5): 311-7.
- Lee, C. K., Y. Shibata, et al. (2004). "Evidence for nucleosome depletion at active regulatory regions genome-wide." Nat Genet **36**(8): 900-5.

- Lim, L. P., N. C. Lau, et al. (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." *Nature* **433**(7027): 769-73.
- Lin, C. Y., V. B. Vega, et al. (2007). "Whole-genome cartography of estrogen receptor alpha binding sites." *PLoS Genet* **3**(6): e87.
- Lindberg, M. K., S. Moverare, et al. (2003). "Estrogen receptor (ER)-beta reduces ERalpha-regulated gene transcription, supporting a "ying yang" relationship between ERalpha and ERbeta in mice." *Mol Endocrinol* **17**(2): 203-8.
- Lupien, M., J. Eeckhoute, et al. (2008). "FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription." *Cell* **132**(6): 958-70.
- Macaluso, M., M. Montanari, et al. (2006). "Nuclear and cytoplasmic interaction of pRb2/p130 and ER-beta in MCF-7 breast cancer cells." *Ann Oncol* **17 Suppl 7**: vii27-9.
- Matthews, J., B. Wihlen, et al. (2006). "Estrogen receptor (ER) beta modulates ERalpha-mediated transcriptional activation by altering the recruitment of c-Fos and c-Jun to estrogen-responsive promoters." *Mol Endocrinol* **20**(3): 534-43.
- Mizuguchi, G., X. Shen, et al. (2004). "ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex." *Science* **303**(5656): 343-8.
- Monsma, F. J., Jr., B. S. Katzenellenbogen, et al. (1984). "Characterization of the estrogen receptor and its dynamics in MCF-7 human breast cancer cells using a covalently attaching antiestrogen." *Endocrinology* **115**(1): 143-53.
- Mottagui-Tabar, S., M. A. Faghihi, et al. (2005). "Identification of functional SNPs in the 5-prime flanking sequences of human genes." *BMC Genomics* **6**(1): 18.
- Murphy, L. C., G. E. Weitsman, et al. (2006). "Potential role of estrogen receptor alpha (ERalpha) phosphorylated at Serine118 in human breast cancer in vivo." *J Steroid Biochem Mol Biol* **102**(1-5): 139-46.
- Nardulli, A. M. and B. S. Katzenellenbogen (1986). "Dynamics of estrogen receptor turnover in uterine cells in vitro and in uteri in vivo." *Endocrinology* **119**(5): 2038-46.
- Nuckel, H., U. H. Frey, et al. (2006). "Association of a novel regulatory polymorphism (-938C>A) in the BCL2 gene promoter with disease progression and survival in chronic lymphocytic leukemia." *Blood*.
- Olivier, M., D. E. Goldgar, et al. (2003). "Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype." *Cancer Res* **63**(20): 6643-50.
- Orlando, V., H. Strutt, et al. (1997). "Analysis of chromatin structure by in vivo formaldehyde cross-linking." *Methods* **11**(2): 205-14.
- Pavesi, G., P. Mereghetti, et al. (2004). "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes." *Nucleic Acids Res* **32**(Web Server issue): W199-203.
- Philips, A., D. Chalbos, et al. (1993). "Estradiol increases and anti-estrogens antagonize the growth factor-induced activator protein-1 activity in MCF7 breast cancer cells without affecting c-fos and c-jun synthesis." *J Biol Chem* **268**(19): 14103-8.
- Pike, A. C., A. M. Brzozowski, et al. (2001). "Structural insights into the mode of action of a pure antiestrogen." *Structure* **9**(2): 145-53.
- Plohl, M., A. Luchetti, et al. (2008). "Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin." *Gene* **409**(1-2): 72-82.
- Porter, W., B. Saville, et al. (1997). "Functional synergy between the transcription factor Sp1 and the estrogen receptor." *Mol Endocrinol* **11**(11): 1569-80.

- Shaw, R. J. (2006). "Glucose metabolism and cancer." Curr Opin Cell Biol **18**(6): 598-608.
- Shou, J., S. Massarweh, et al. (2004). "Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/neu cross-talk in ER/HER2-positive breast cancer." J Natl Cancer Inst **96**(12): 926-35.
- Sledge, G. W., Jr. and K. D. Miller (2003). "Exploiting the hallmarks of cancer: the future conquest of breast cancer." Eur J Cancer **39**(12): 1668-75.
- Vahteristo, P., J. Bartkova, et al. (2002). "A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer." Am J Hum Genet **71**(2): 432-8.
- Vahteristo, P., H. Eerola, et al. (2001). "A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families." Br J Cancer **84**(5): 704-8.
- Vega, V. B., C. Y. Lin, et al. (2006). "Multiplatform genome-wide identification and modeling of functional human estrogen receptor binding sites." Genome Biol **7**(9): R82.
- Vehmanen, P., L. S. Friedman, et al. (1997). "Low proportion of BRCA1 and BRCA2 mutations in Finnish breast cancer families: evidence for additional susceptibility genes." Hum Mol Genet **6**(13): 2309-15.
- Vousden, K. H. and X. Lu (2002). "Live or let die: the cell's response to p53." Nat Rev Cancer **2**(8): 594-604.
- Walsh, T., S. Casadei, et al. (2006). "Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer." JAMA **295**(12): 1379-88.
- Wei, C. L., Q. Wu, et al. (2006). "A global map of p53 transcription-factor binding sites in the human genome." Cell **124**(1): 207-19.
- Weinmann, A. S. and P. J. Farnham (2002). "Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation." Methods **26**(1): 37-47.
- Wells, J. and P. J. Farnham (2002). "Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation." Methods **26**(1): 48-56.
- Wolczynski, S., A. Surazynski, et al. (2001). "Estrogenic and antiestrogenic effects of raloxifene on collagen metabolism in breast cancer MCF-7 cells." Gynecol Endocrinol **15**(3): 225-33.
- Wu, Y. L., X. Yang, et al. (2005). "Structural basis for an unexpected mode of SERM-mediated ER antagonism." Mol Cell **18**(4): 413-24.