Supplementary information for

Integrative model of genomic factors for determining binding site selection by estrogen receptor α

Roy Joseph*, Yuriy L. Orlov*, Mikael Huss*, Wenjie Sun, Say Li Kong, Leena Ukil, You Fu Pan, Guoliang Li, Michael Lim, Jane S. Thomsen, Yijun Ruan, Neil D. Clarke, Shyam Prabhakar, Edwin Cheung, Edison T. Liu[1]

Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672.

*These authors contributed equally to this work

[1]Corresponding author: Edison T. Liu, Genome Institute of Singapore, 60 Biopolis, Singapore 138672, T. 65 6808 8038; F. 65 6808 9051; Email: liue@gis.a-star.edu.sg

This supplementary file includes:

Supplementary Methods and Notes

Supplementary Figures 1 to 17

Supplementary Tables I to XIII

Supplementary References

Supplementary Methods and Notes

Library construction, sequencing of ChIP/FAIRE enriched DNA samples and data analysis

All DNA samples were processed as per the Illumina Solexa ChIP-seq sample processing methods. 10 ng of ChIP DNA was end polished with T4 DNA polymerase and kinase. An A base was added to the polished DNA fragments followed by the Qiaquick column (Qiagen) clean up. Solexa adaptors were ligated to the ChIP DNA fragments and enriched by 15 cycles of PCR amplification. 200-300 bp size fractions were selectively isolated from the 1% agarose gel and eluted by Qiagen gel extraction kit. The extracted DNA was quantified by Agilent Bioanalyzer and subjected to Solexa sequencing according to the manufacturer's instruction. The processed ChIP or FAIRE-enriched DNA fragments were then used for Illumina single read sequencing analysis. We used ELAND program provided with the 1G analyzer software package and in-house computational tools for mapping the sequence tags to the reference genome hg18 and clustering short sequences. In order to avoid potential PCR amplification bias, tags that shared the same mapping location on the same strand were removed. The uniquely-mapped reads with at most 2-mismatches were kept for further processing.

The oriented 25-36 bp DNA reads were extended to 200 bp regions to count clusters of overlapping sequences. The enrichment peaks for corresponding libraries were identified using ChIP-seq peak calling algorithm as previously described (Chen *et al*, 2008; De Santa *et al*, 2009). The identified peaks were filtered in three steps. First, an estimated peak intensity threshold based on a random distribution of tags over the genome was used to remove random low-intensity peaks. The FDR for a library was determined by a Monte Carlo simulation, in which extended (to 3' direction) 200 bp fragments randomly extracted from the genome were used to estimate the numbers of random peaks with different intensity values. The number of random fragments was equal to the sequencing depth of each library under analysis. The minimum intensity that satisfies the 0.001 criterion is selected as the lower cut-off for calling confident peaks. Then, we further filtered the peaks based on

the 5 fold-change of peak intensity against local count of sequence reads from the input DNA control sequencing library. To remove bias due to duplications in MCF-7, we used published data on copy number variations (Shadeo and Lam, 2006) and kept only ChIP-seq peaks from non-amplified regions, as defined by array CGH experiments in the cited paper. These three steps resulted in 16,043 ERα binding sites in MCF-7 from non-amplified regions for the downstream analysis. The processing of histone modification ChIP-seq data followed same method, but without restriction to non-amplified regions. For cross-comparison of chromatin modified regions from different ChIP-seq libraries we used downsampling to smallest available library size (i.e.7 millions tags for ChIP-seq libraries from MCF-7 cells, see Supplementary Table I) using random number generator to removed excess tags from counting).

The analysis of the ER binding as well as histone modifications in MCF-7 cells was done by ChIP-seq data obtained either with E2 stimulation or without stimulation using vehicle as a control. We noticed that E2 induction caused several fold increase in ERα binding affinity (Supplementary Figure 4). By comparing the numbers of unique ER ChIP-seq peaks obtained by E2 and vehicle treatments, we found that the number of sites in vehicle condition is an order of magnitude less than the number of ER sites following E2 administration (1,110 vs. 16,043) and the majority of ER sites at vehicle condition are presented at E2-defined ER ChIP-seq sites [only 59 ER binding sites are vehicle-specific (in non-amplified regions of MCF-7 cells)].

Comparative analysis of ERα binding sites

To understand whether the ER binding sites identified by ChIP-seq were valid, we first compared the ChIP-seq data with previously published data sets. Previously 3,665 sites were identified by ChIP-on-chip experiments (Carroll *et al*, 2006) and 1,226 ER binding sites were identified by ChIP-PET (Lin *et al*, 2007) (non-redundant number of published 1,234 sites without chromosomes M and Y) in MCF-7 cells. We extended ±200 bp from the mid-point of each binding

site in this study to overlap with the published binding sites. Supplementary Figure 1 shows the overlap between ER binding sites defined by ChIP-seq method (this study), and other published genome wide ER binding site data sets in MCF-7 cells. Our ChIP-seq data contain 68.6% (0.686=841/1226) and 86% (0.86=3152/3665) of binding sites, respectively, of these high-confidence datasets published previously. Regarding extended ER dataset containing 8,525 sites defined by the same ChIP-on-chip technology in MCF-7 cells (Hurtado *et al*, 2008) our ChIP-seq data set contains 74% (6,321/8,525) of binding sites. The high percent of overlapped binding sites indicates the high sensitivity of the Solexa based platform. Additionally we checked overlap with recently published ER binding sites defined in MCF-7 cells (Welboren *et al*, 2009) by using the same ChIP-seq technology. We found that the overlap between two ChIP-seq datasets in MCF-7 was 62% (6,261/10,191 sites).

Genomic distribution of ERα binding sites

When regions of gene amplification are accounted for, the frequency of binding clusters per chromosome generally corresponds to the size and gene density of the chromosome, and ER does not appear to localize to specific chromosomal regions within the genome. We analyzed the location of all the 16,043 ERBS with respect to the known RefSeq genes. We grouped these ERBS based on their genomic locations [promoter: (-5 Kb to +1 Kb of the TSS); intragenic: (+1 Kb from TSS to the 3' end); 3' end: (from 3'end to 5 Kb downstream); 5' distal: (-100 Kb to -5 Kb of the TSS); 3' distal: (+5 Kb to +100 Kb of the 3' end) and gene desert: all the rest]. The pie chart (Supplementary Figure 2A) shows the ERBS distribution relative to the nearest gene border of RefSeq annotation track. We found that only 9% of ERBS are actually located in the proximal promoter regions with respect to the known RefSeq gene borders. The largest fraction (40%) of binding sites lie in intragenic regions of transcripts and are generally localized to introns, whereas 17% and 14% of sites are present in distal regions (from 5' and 3' ends respectively), and 4% sites are present in the

vicinity of 3' polyadenylation sites. Based on our definition of the location of binding sites, 16% of the ER binding regions are located in gene deserts. Our findings suggest that DNA-bound ER can interact with the transcriptional machinery through both proximal- and distal-acting mechanisms, and these interactions are not likely to be limited by ERBS orientation (5' or 3') relative to the TSS. The enrichment of ERBS for both 5' and 3'-proximal regions of genes (Figure 2B and Supplementary Table V) is significant. Furthermore, our data is consistent with earlier studies (Carroll *et al*, 2006; Lin *et al*, 2007) which suggested functional ERBS were rarely present in exons and when in exons were in untranslated regions.

Enrichment of chromatin marks around ERα binding sites

The 16,043 ER binding sites (ERBS) defined by ChIP-seq sequence tag occupancy were ranked by descending order of their induction and subsequently stratified into quartiles (Q1 to Q4); where quartile 1 (Q1) contains top 25% strongest induced ER binding sites (i. e. strongest ER occupancy), quartile 2 (Q2) contains next 25%, and so on for quartiles 3 and 4 where quartile 4 (Q4) represents the weakest induced binding sites. The association of chromatin marks to ERBS was estimated by counting of ChIP-seq tags from histone modifications and FAIRE in proximity to the ERα binding peak at various intervals from the centre of binding sites. We analyzed chromatin activation and repressive mark profiles of ER binding sites at both E2 stimulated and non-stimulated conditions. First, the tags from each library were mapped relative to ERBS and then used to calculate the average enrichment in intervals +/- 2 Kb of the centre of the binding sites. For comparison, we also performed same analysis for unbound EREs. From the tag profile of various chromatin marks, we found an open chromatin conformation within 1 Kb region around ERBS with either E2 stimulated [main text, Figure 1 (Q1-Q4)] or both stimulated and non-stimulated states (main text, Figure 2B; where only Q1 binding sites are presented to clearly show the ligand effect) with significant enrichment for histone activation marks and RNA Pol II. Interestingly enough, we

also observed a gradient for this open chromatin conformation and active histone modifications enrichment, for each quartile of ER binding sites, with quartile 1 (highest ER occupancy) binding sites showing the maximum openness and maximum enrichment for active histone marks and RNA Pol II, and quartile 4 (least ER occupancy) binding sites showing the least (Figure 1, main text). On the other hand, the unbound sites (EREs with no ER binding) lacked this open chromatin conformation or active histone mark enrichment, either in the E2 induced (Figure 1) or non-induced state (data not shown). In terms of repressive chromatin marks (H3K9me3 and H3K27me3), our analyses showed that these signals were very low for both bound and unbound sites with no significant difference between the E2 stimulated or non-stimulated states. We also analyzed whether the characteristics of the ERα binding sites changed upon ligand induction of ER and found that some of the signatures were increased (H3K9ac, H3K14ac and RNA Pol II), or decreased (H3K4me1) or not changed significantly (FAIRE) after E2 stimulation (see result section for more details). Secondly, in order to see the correlation of various chromatin marks with bound ER sites, we also analyzed the association between chromatin marks with tag counts at bound versus unbound ER sites before E2 induction (Supplementary Figure 5). Supplementary Figure 5 represents the fraction of ER sites that are enriched by chromatin marks over background sequencing reads in ±250bp around the centre of ER ChIP-seq binding sites. We expected potential problems when making conclusions of the comparative data in order to explain population trend of ER bound sites with various chromatin marks of different sequencing depth for corresponding libraries. Therefore, whenever necessary, all ChIP-seq libraries were down sampled to the same size (7M tags; i. e. to the smallest available library in the dataset), after random removal of excess sequence tags. We also analyzed the trends in chromatin marks occupancy by different quartiles of ERBS, both at E2 stimulated and at non-stimulated conditions (Supplementary Figure 7). For all activation chromatin marks as well as RNA Pol II and FAIRE, the Q1 ERBS have higher chromatin enrichment than following Q2 sites, Q2 sites have higher enrichment than Q3 sites and Q3 sites have higher

enrichment than Q4 binding sites. At the same time repressive chromatin marks have no difference across different quartiles and treatment conditions (not significant by Mann-Whitney test *P*-value >0.01) (Supplementary Figure 7). Finally, we analyzed the positional biasness of various chromatin signatures around ERBS with respect to the genomic locations. It is well known that certain histone marks are enriched at the gene boundaries, especially the promoter regions. In order to see whether the enrichment of histone activation marks at the ERBS is due to the positional bias, we analyzed the tag density profile for each chromatin mark at the ERBS from different genomic locations relative to RefSeq genes. For this analysis, we classified the 16K binding sites into promoter (5Kb), intragenic and distal (5-100Kb upstream TSS) binding sites and looked at the average tag counts in intervals of ±2 Kb of the center of the binding sites. We found that except two histone modifications (H3K4me3 and H3K9ac), the ChIP-seq enrichment of the remaining activation marks, especially H3K4me1, and the other binding site signatures (RNA Pol II and FAIRE) were significantly associated with ERα binding sites regardless of their location relative to gene boundaries. H3K4me3 and H3K9ac marks were more enriched at the promoter specific ER binding sites (Figure 2C, main text).

Motif analysis at the ERα binding sites

We initially analyzed the presence of the Estrogen Response Element (ERE) motif in the ERα binding sites identified in this study, according to the method in our previous publication (Vega *et al*, 2006). Briefly, we looked for the presence of the core consensus ERE motif (GGTCA-nnn-TGACC), allowing for a maximum of 2 mismatches. ERE motif cores were enriched at the center of the ChIP-seq identified ERα binding sites.

To estimate potential ER binding landscape in the human genome we used an extended ERE motif identified in our previous study (Vega *et al*, 2006) as well as standard TRANSFAC PWM (ER_Q6). In hg18 (without Y chromosome absent in female breast cancer samples) we found 32,614 and 283,999 potential binding sites correspondingly. Next, the coordinates of PWM-defined sites

were filtered by presence of experimentally defined ERα binding sites in +/-500bp interval as detected by any published binding site studies: ChIP-PET by (Lin *et al*, 2007), ChIP-seq by (Welboren *et al*, 2009), ChIP-on-chip (Hurtado *et al*, 2008) and ChIA-PET by (Fullwood *et al*, 2009). Resulting negative binding sets of computationally predicted but experimentally unbound ERE as well as set of 820,000 random non-promoter locations in the genome were used for ROC-AUC calculations and motif updates.

In parallel with the elementary consensus-and-mismatch motif and PWM analysis described above, we also assessed ER binding affinity using the TherMoS (*Ther*modynamic *Mo*deling of chip-*S*eq) algorithm. When trained on the ER ChIP-seq profile in 1Kb regions centered on the 16,043 binding peaks from non-amplified genomic regions in the MCF-7 genome, the algorithm identified the palindromic ER motif illustrated in Figure 3A (main text). In generating the figure, the PSEM (Position Specific Energy Matrix) was first converted to a traditional position-specific frequency matrix using an exponential transformation (Stormo, 2000). The motif shown in the Figure 3A (main text) is perfectly palindromic, since palindromicity was imposed as a constraint while running the algorithm on ER binding data. The ER PSEM and binding affinity scores are available at http://www.gis.a-star.edu.sg/~liue/sup/.

Based on the palindromic motif defined by free energy-based binding model, we examined the 16,043 ER binding regions for evidence of various subpopulations of binding sites, such as ER full-sites, half-sites and non-ERE sites. It is not obvious how one would classify binding sites as half-sites or full palindromic sites based only on scores for the 17-mer palindromic motif. We therefore decomposed the G-score of 17-mers into their left and right half-site components G_L and G_R. In this scoring scheme, genuine palindromic binding sites would have high affinity at both the left and right half-sites, i.e. both G_L and G_R would be low. On the other hand, half-ERE binding sites would score well only at one half of the 17-mer, and poorly at the other half (for example, G_L would

be low and G_R would be high). With this decomposition, it is therefore possible to separate half-site enrichment from full-site enrichment at ER ChIP-seq peaks, and therefore to assess the relative contribution of half-sites and full-sites to ER binding.

In order to quantify ER motif enrichment in the two-dimensional space of left and right half-site binding free energy scores, we partitioned the two-dimensional G_L-G_R space into square bins, and counted the number of 17-mers assigned to each bin. The "foreground" tally was based on 17-mers that lay in 100 bp regions centered at ChIP peaks. In order to estimate the "background" frequency of random 17-mers, we scanned 4.1 million random 100-bp regions with no evidence of ER binding. Figure 3B (main text) illustrates the enrichment of 17-mer scores in the 16,043 binding regions relative to background, plotted in G_L-G_R space.

It is evident from Figure 3B that the binding energy of high-affinity 17-mers is more or less randomly distributed among the left and right half-sites. This can be inferred from the fact that, along any given antidiagonal slice near the bottom left corner of the plot (for example, $G_L + G_R = 3$), the degree of motif enrichment is more or less uniform. However, at lower levels of 17-mer affinity (say, $G_L + G_R = 10$), it is clear that motif enrichment is most pronounced close to the X and Y axes, indicating that the affinity of these 17-mers derives mostly from one of the two half-sites. In order to quantify the affinity or occupancy level at which binding n-mers transition from unbiased full-site binding to predominantly half-site binding, we computed a motif asymmetry score given by the equation:

$$A = \sum_{i=0}^{N} \cos[2\pi \frac{i}{N}] E(i)$$

where *E(i)* represents the motif enrichment in the *i*-th bin that lies on along any particular anti-diagonal $G_L + G_R = C$ (Figure 3B, main text). As defined here, motif asymmetry $A = 0$ when motif enrichment is constant along the anti-diagonal, and $A > 0$ when enrichment is greater near the axes (half-site zones). We determined that motif asymmetry was relatively low above an occupancy of

0.05, which we took to be our threshold for a full-site (Figure 3C; main text). This threshold of 0.05 can be interpreted as a strict cut off, such that any site that has an occupancy greater than 0.05 is a definite full-site. From here on, we will use the term "full-site" to refer to definite full-sites with predicted occupancy greater than 0.05. We defined intermediate sites as 17-mers that could possibly be considered full-sites, but with slightly lower binding affinity. To determine the threshold for intermediate sites, we identified the point on the diagonal line (G_L=G_R) of the enrichment plot at which enrichment in ER binding regions dropped to less than twofold. This yielded an occupancy threshold of 0.02 for intermediate sites (Figure 3D, main text). Thus, intermediate sites are defined by $0.02 < \text{occupancy} \leq 0.05$. A similar two -fold enrichment criterion was used to determine the G-score threshold for definite half-sites (Figure 3E; main text). Intermediate half-sites were defined as 17-mers that did not qualify in any of the previous categories, but had a moderate left or right half-site score (Figure 3B, area IV). "No ERE" was used to describe 17-mers that did not fit any of these descriptions (area V).

We sought to identify transcription factors that might modulate ER binding to the MCF-7 genome by performing de novo motif detection in 100-bp regions centred on ER ChIP-seq peaks, using the MDscan program (Bailey *et al*, 2009). In order to identify motifs besides the palindromic ERE motif and its half-site variant, all definite ER half-sites were masked in the binding sequences supplied to MDscan. This screen identified CACD (similar to SP1), AP1, Forkhead and AP2 motifs as potential cooccupants at ERE half-sites and "no-ERE" binding regions (Supplementary Figure 8). In order to independently estimate the enrichment of these co-motifs in the various subsets of ER binding sequence, we replaced the motifs defined by MDscan with their TRANSFAC equivalents (Symbol "V$" standing for vertebrate PWM ID is omitted): CACD/SP1 by CACD_01, AP1 by AP1_C, Forkhead by HNF3ALPHA_Q6 and AP2 by AP2ALPHA_03. Score thresholds for these four TRANSFAC motif models were set so as to maximize enrichment of motif matches in the 16,043 ERBS (100 bp binding regions) relative to randomly chosen genomic regions. With these

threshold definitions, we found following trend: as the quality of the ERE motif deteriorated from full sites to intermediate half-sites, the likelihood of co-motif occurrence in the ER binding region increased for all four co-motifs (Supplementary Figure 9, Supplementary Table IV). Binding regions with no ERE showed no clear trend, perhaps due to the fact that they were too few in number to characterize statistically. The inverse relationship we observed between ER motif quality and co-motif occurrence is consistent with a model wherein low-affinity EREs are more likely to require assistance from other transcription factors in recruiting ER. This assistance could take the form of direct protein-protein interactions, as has been suggested for AP-1 and ER (Safe and Kim, 2008), or an indirect cooperative or chromatin-modifying role as suggested for FOXA1 (Carroll *et al*, 2006).

In order to refine our understanding of the role of co-motifs in recruiting ER to DNA, we analyzed the location of the co-motif relative to the position of the ChIP-seq peak (data not shown). Although the AP1, FOXA1 and AP2-alpha motifs showed a positional preference for being close to the centre of the binding region, the CACD/SP1 motif was uniformly distributed in the 1Kb extended binding regions. Thus, although we found no evidence of widespread ER recruitment by CACD/SP1, our results are consistent with earlier reports of the role of AP1 and FOXA1 in recruiting ER to DNA, and also confirms a novel enabling or cooperative role for AP2 in ER recruitment (Cheung, E. *et al*., manuscript in preparation).

De novo **motif discovery for defined ER binding site categories**

MEME (Bailey *et al*, 2009) was used to scan 100 bp sequences around the 16,043 ERBS divided into five categories (definite full, intermediate full, definite half, intermediate half and no-ERE sites) according to the classification described in the main text. For the full and intermediate full sites, the search was constrained to 13 bp motifs, and for the other, to 6 bp motifs. In some cases, the motif search was initialized from the consensus sequence or constrained to find a

palindrome; these cases are indicated in the caption to Supplementary Figure 10C. Both strands of the sequences were searched and zero or one occurrence of the motif per sequence was assumed.

Genome-scale cross-validation of ERα binding site prediction

In order to further demonstrate the efficacy of our 4-parameter predictive model (as explained in the main text) in a cell line specific manner, we chose another ERα positive breast cancer cell line, the T47D and monitored the ERα binding sites as well as selected most predictive chromatin marks (FAIRE, FOXA1 and H3K4me1) by using same experimental and computational techniques that were used for MCF-7 cells. We then compared the ERα binding sites from two cell lines and defined a set of binding sites either common or unique for each of these cell lines. We selected 73 binding sites from each category and experimentally validated these binding sites for their ERα occupancy using qPCR experiments (Supplementary Figure 14). The functional importance of ERα cell-specific recruitment raises the question as to how ERα is able to bind to distinct regions within the genome of the MCF-7 and T47D cells. Accordingly, we considered the possibility that the sequence recognized by ERα or the chromatin signatures at these sites could be different between the two cell lines. We therefore compared ERα motifs as well as the ChIP-seq tag counts of chromatin marks enriched within the common ERα recruitment sites, as well as those unique to each cell line (Supplementary Figures 14-16). We found that ER binding sites common to both cell lines had the highest levels of each of TherMoS affinity score, H3K4me1, FOXA1 occupancy, and FAIRE, the 4 predictive parameters. Importantly, we noticed that, despite similar low ER affinity scores, all other three parameters (H3K4me1, FAIRE, and FOXA1 occupancy) were significantly enriched in the MCF-7 unique sites as compared to the T47D unique sites in MCF-7 cells. On the other hand, in T47D cells we noticed higher enrichment of FAIRE and FOXA1 marks in proximity of T47D unique ER binding sites, as compared to MCF-7 unique sites. Indeed these data suggested that ERα translates an epigenetic signature into functional cell type-specific enhancers leading to the establishment of cell

type-specific transcriptional programs. We then took a logistic regression model (with TherMos ER affinity score, FOXA1, H3K4me1 and FAIRE as features) which had been fitted to all distal ER binding sites in MCF7 and applied it directly to a dataset containing the same features measured in T47D for all T47D ChIP-seq ER peaks and the 820,000 random regions. The ROC-AUC for this prediction task was 0.86. We also attempted to discriminate between TherMoS-predicted EREs which were bound in T47D vs those that were not bound in T47D using the three features H3K4me1, TherMoS affinity score and FOXA1. Again, the model was trained on the equivalent data set in MCF7, and the resulting logistic regression model was evaluated on the T47D data. The resulting ROC-AUC for this task was 0.93. Therefore, the FOXA1, FAIRE and H3K4me1 marks along with the ER binding sites in T47D cells, provided us an additional opportunity to demonstrate and test the 4-parameter predictive model for T47D cells in terms of discriminating the bound sites from random sites and from non-bound EREs. Our data also suggested the possibility that the cell type-specific recruitment of ERα to the chromatin is linked to through specific collaborations with chromatin marks in two different breast cancer cell lines.

Logistic regression modeling of ER binding

Task 1. Distinguishing ER bound sites from random genomic background.

14,338 non-TSS-proximal ER ChIP-seq determined binding sites were used as the positive set and 820,000 regions that neither overlapped with ER ChIP-seq peaks nor were near transcription start sites were used as the negative set. Since the free-energy based ER affinity score, which is in itself derived from the ChIP-seq-defined ER bound sites, was used as one of the features predicting ER binding here, it was impermissible to score a site using a PSEM that had been fitted using a training set that included that site. Therefore, we divided the 14,338 distal sites into five approximately equal sets, which were scored separately using an ER binding PSEM fitted on the remaining 4/5 of the sites. Similarly, the 820,000 random distal sites were divided into five equally

large sets and scored using the five different PSEMs. Accordingly, each logistic regression model was built five times, each time with a different positive (ER bound) and negative set (random) of regions. In each of the five rounds, 70% of the data were randomly selected to fit the logistic regression model, and the remaining 30% was used to assess the performance of the model and produce ROC and precision-recall curves. All possible feature combinations were tested. The performance of the best N-variable models (where N=1,2,3,…12) across five cross-validation runs are shown in Supplementary Table VIII.

Task 2. Distinguishing ER bound proximal promoters from non-ER-bound proximal promoters

We extracted 500-bp regions just upstream of each RefSeq gene [-1 to -500 bp relative to the TSS] and summed the tag counts for each library for these regions as well as scoring them using the ER PSEM. The promoters where an ER ChIP-seq peak had been called (after ligand induction) were used as the positive set, and the rest was used as a negative set. 5-fold cross-validation was used to assess performance. All possible feature combinations were tested. The best ROC-AUC was obtained for a model using all features, which meant that overfitting was not a serious problem in this particular case (Supplementary Table IX).

Task 3. Distinguishing ER bound predicted EREs from non ER bound predicted EREs

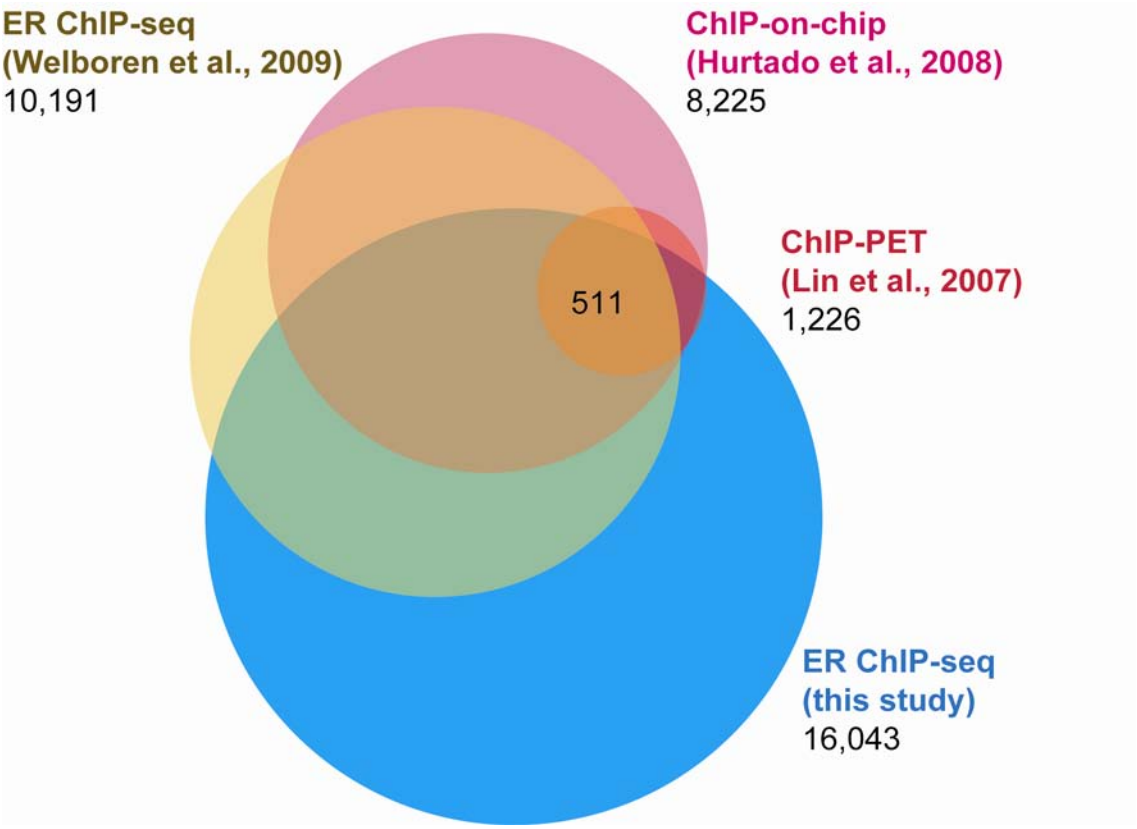
Here, the 6,900+ predicted EREs that were close to ChIP-seq peaks (EREs within 200 bp of a ChIP-seq peak) became the positive set, and the remaining sites became the negative set after additional removal of possibly bound ERα sites as judged from comparisons with previously published results (ChIP-PET, ChIP-chip, ChIA-PET). Furthermore, the sites were divided into TSS-proximal and distal sites and classification models were built separately for these two sets. ROC-AUC scores were obtained by testing on a randomly selected hold-out set of 30% of the original examples after fitting the logistic regression model on the other 70%. All possible parameter

combinations were tested (but the ER affinity score was excluded from the set of features, because the positive and negative sets had been defined based on ER binding). Refer to Supplementary Tables X and XI for the performance of the best N-parameter combinations.

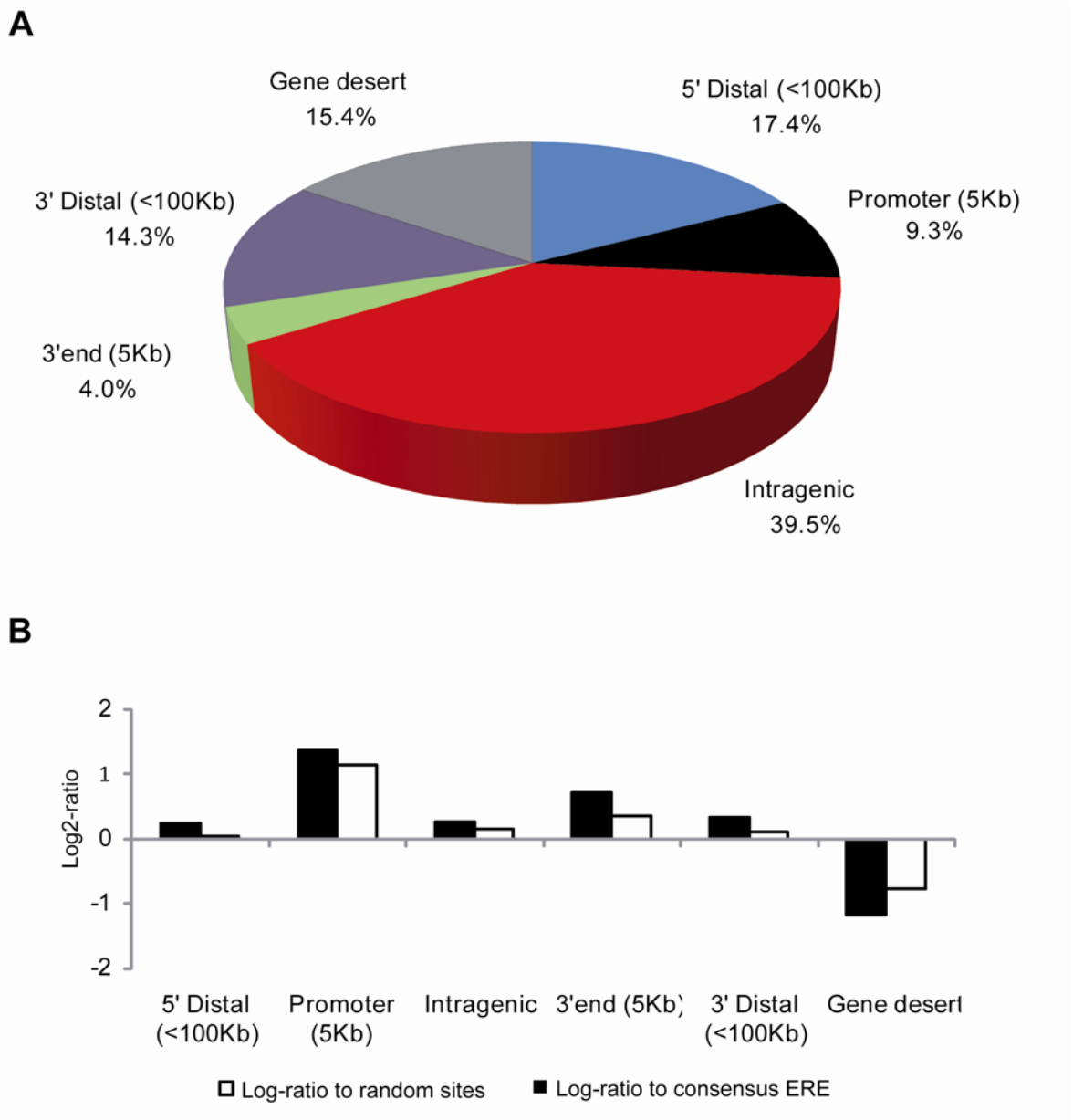
Table XII shows comparison of ER binding prediction with general binding score from Ernst *et al*. (2010) on same sets of predicted EREs.

Table XIII contains the PSEM (Position Specific Energy Matrix) of ERE used in affinity scores.

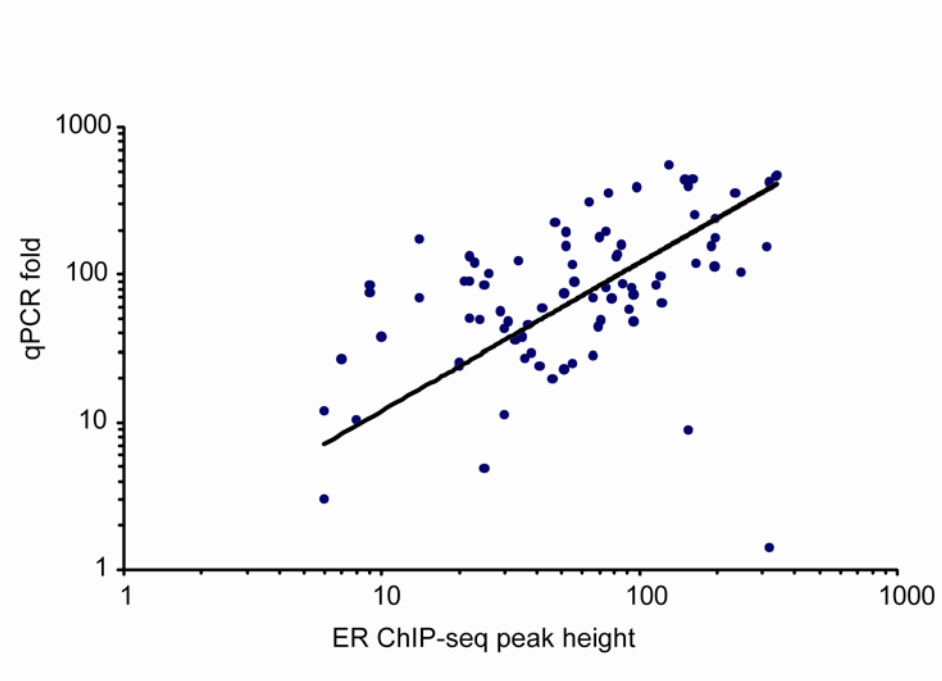
Supplementary Figures



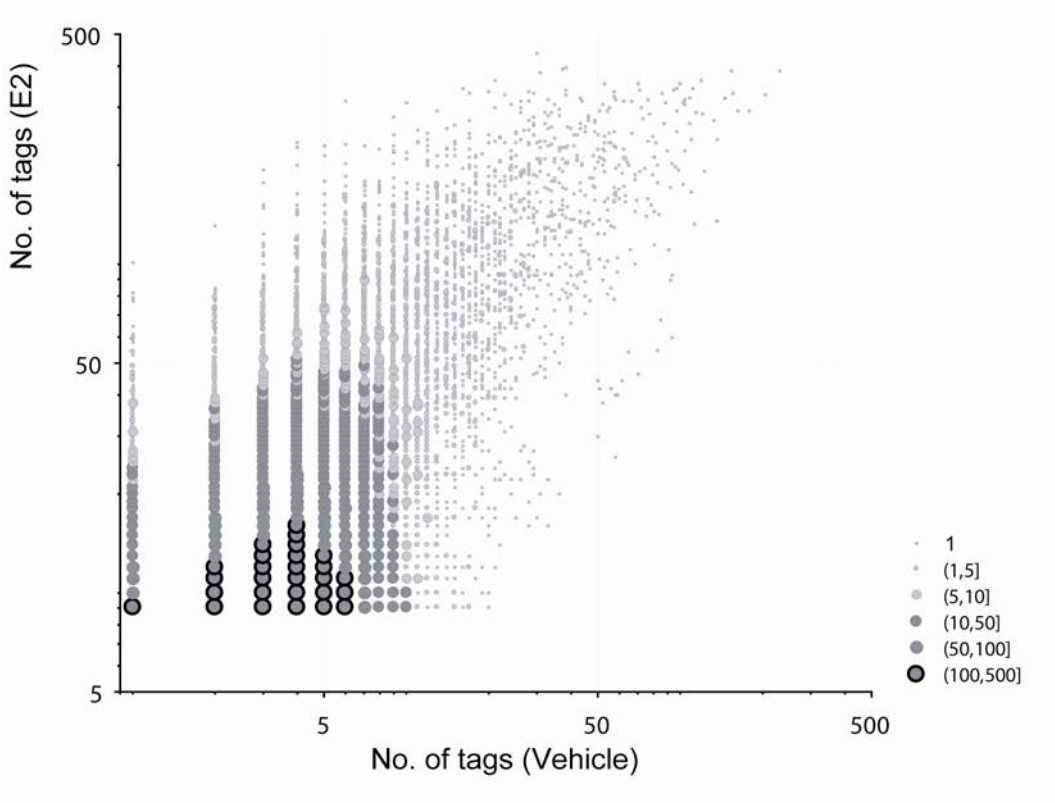
Supplementary Figure 1 Overlap of ChIP-seq ER binding sites with data from previous published studies. Venn diagram indicating the overlap between ER binding sites defined in MCF-7 cells by ChIP-seq in this study, 1,226 ChIP-PET identified sites (Lin *et al*, 2007), 8,225 binding sites from ChIP-on-chip experiments (Hurtado *et al*, 2008) (an updated version of previous ChIP-chip identified 3,665 sites; Carroll *et al*, 2006) and 10,191 sites identified by ChIP-seq in a recent study (Welboren *et al*, 2009). There is an overlap of 69%, 74% and 62%, respectively, between previously published data and the data presented here (see Supplementary Table II for details).



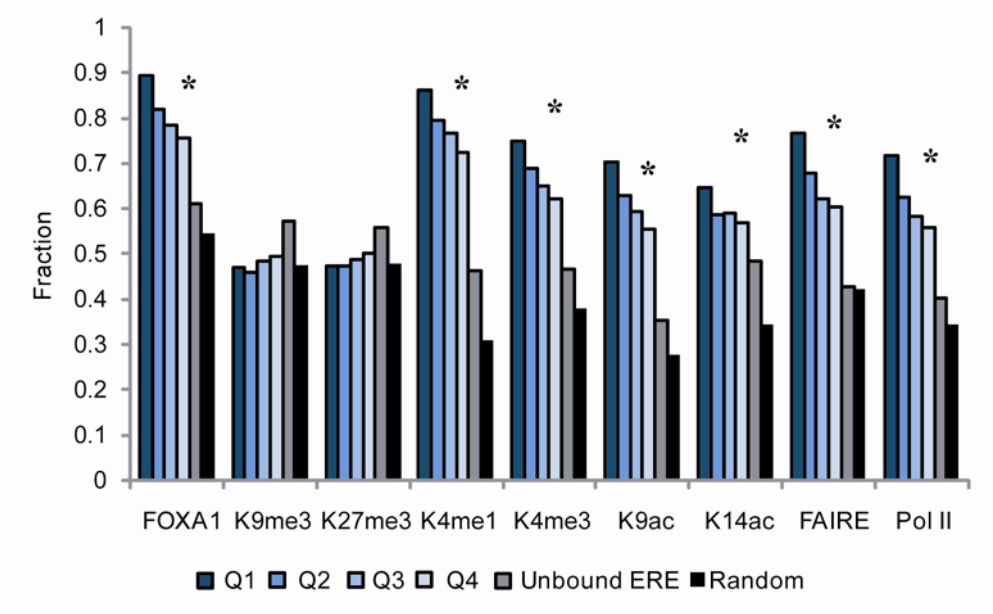
Supplementary Figure 2 Distribution of ER binding sites in the MCF-7 genome. **(A)** ER binding site distribution in the genome relative to the nearest RefSeq gene. The 16,043 ERBS were grouped based on their genomic locations as per the following definitions [promoter: (-5 Kb to +1 Kb of the TSS); intragenic: (+1 Kb from TSS to the 3' end); 3' end: (from 3'end to 5Kb downstream); 5' distal: (-100 Kb to -5 Kb of the TSS); 3' distal: (+5 Kb to +100 Kb of the 3' end) and gene desert: (all the rest)]. **(B)** Comparison of ChIP-seq ERBS distribution in genomic locations (same as in Panel **A**) to the distributions of random coordinates and consensus ERE in human genome. Random sites (computer generated coordinates) and ERE predicted by consensus (perfect 13 bp ERE consensus with no more than 1 mismatch, but not detected by ChIP-seq, 56,746 sites in total) were used for comparison. Ratio of fraction of ChIP-seq ER sites to random sites (black bars) and ratio of same fraction to consensus ERE sites regardless of binding (white bars) are in log2 scale. Although major fractions of ChIP-seq sites are within gene borders or distal regions, the most enriched fraction are promoters and 3'end regions of RefSeq genes. Fraction of ChIP-seq ERBS in gene deserts are depleted even relative to consensus ERE sites indicating the gene-centric nature of ChIP-seq defined sites. Exact *P*-values for this analysis are shown in Supplementary Table V.



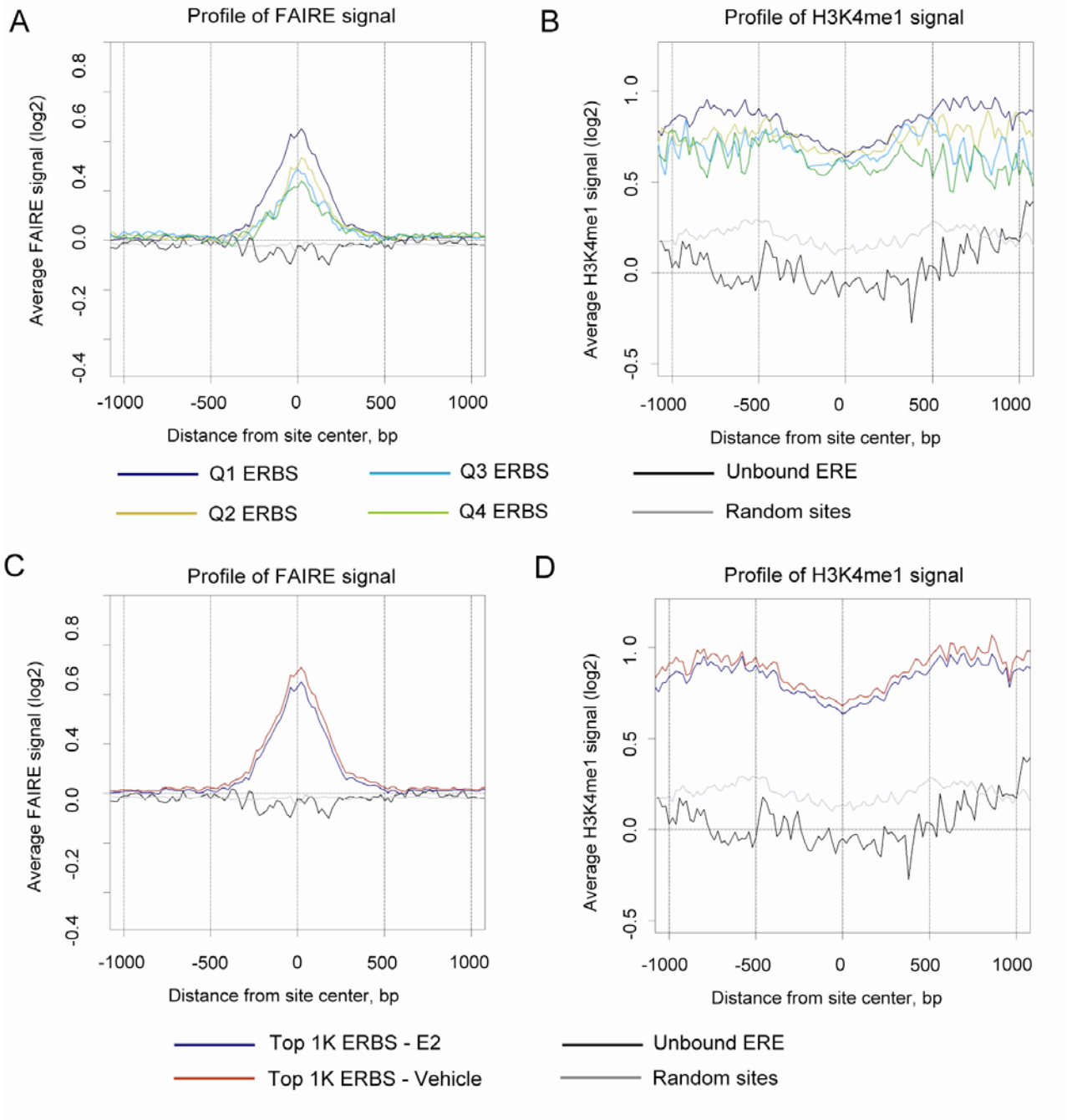
Supplementary Figure 3 Correlation between qPCR fold changes of 81 validated ER binding sites (Lin *et al*, 2007) and ChIP-seq peak height after E2 activation (log-log scale). We observed a linear correlation value of CC=0.56 (*P*=5.0E-8), and a Kendall tau rank correlation of τ=0.375 (*P*= 7.17E-07).



Supplementary Figure 4 Correlation between ER ChIP-seq peaks at E2 vs vehicle treated states in MCF-7 cells. ChIP-seq ERBS selected separately at E2 and vehicle treatment conditions were compared by (unnormalized) tag count. ChIP-seq peak height is on average much higher for ERBS following E2 treatment. Linear correlation coefficient r=0.662 (P < 2.2e-16), rank Kendall correlation τ=0.328 (P< 2.2e-16).

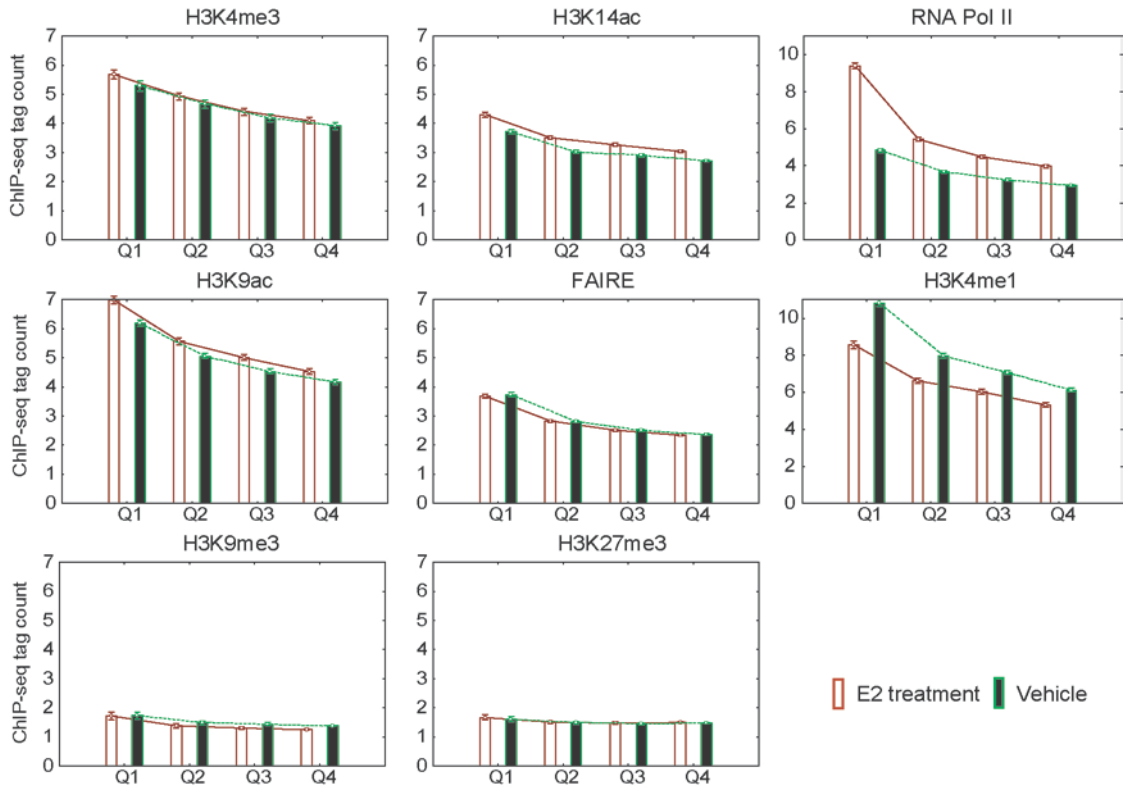


Supplementary Figure 5 Association between chromatin marks with tag counts at bound versus unbound ER sites before E2 induction. The bound ER sites are stratified into quartiles; quartile 1 (Q1) contains the 25% of ER sites that had the highest ChIP-seq tag counts, quartile 2(Q2) contains second 25% and so on. The bar graph represent the fraction of ER sites that are enriched by chromatin marks over background sequencing in +/-250bp of ER ChIP-seq center (at least 2 fold change), thus belong to histone modification regions. All ChIP-seq libraries were down sampled to the same size (7M unique tags) in order to avoid bias by sequencing depth. Note that association of all activation marks and RNA Pol II is statistically different (indicated by *) for four quartiles of ERBS compared to unbound or random sites. The difference between ChIP-seq chromatin mark occupancy at ERBS and random loci is statistically significant for activation marks (P<1E-11) and not significant for repressive marks. Similar results were observed for chromatin marks in ChIP-seq ERBS versus binding sites computationally predicted in human genome by PWM ERBS ('Unbound ERE').



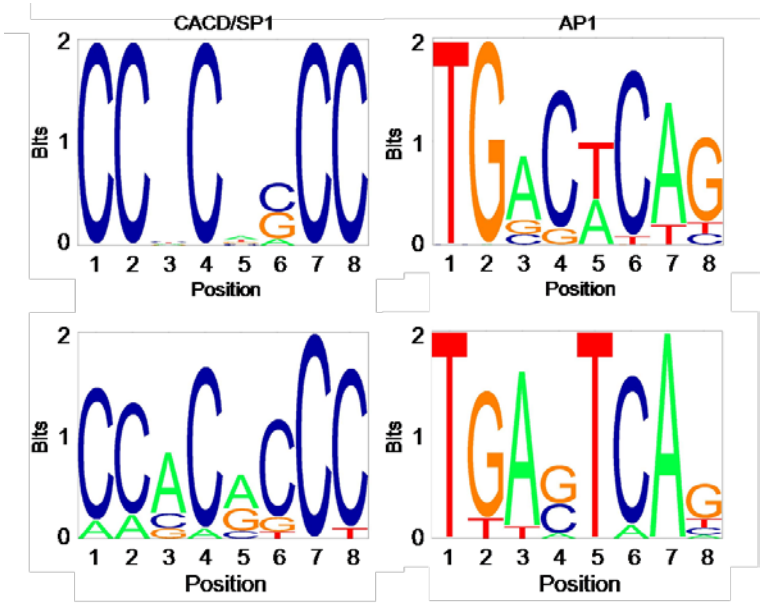
Supplementary Figure 6 ChIP-chip microarray experiments for validation of ChIP-seq results. Validation of ChIP-seq profiles for FAIRE and H3K4me1 around ER-bound sites using a custom made NimbleGen array containing for 40,000 ER binding sites (all validated binding sites from two published genome-wide ERα mapping studies (Carroll *et al*, 2006; Lin *et al*, 2007) and most of the computationally predicted high affinity binding sites using the hERE algorithm (Vega *et al*, 2006) as well as 10,000 randomly selected binding sites). The array was covered with about 385,000 isothermal probes and each binding site region was tiled with at least 6 probes. The FAIRE and H3K4me1 ChIP DNA samples were labelled and hybridized, following the array manufacturer's protocol, to the arrays with three biological replicates each for E2 or vehicle treated cells (Liu, E.T. *et al*., manuscript in preparation). The average FAIRE and H3K4Me1 signals were computed the average signals for all probes at the center of ER binding site and at every 20 bp on both sides until

1Kb away and comparing these signals with the 'random' group binding sites. The 'Unbound' group is defined as the list of binding sites containing predicted ERE from ERE motif-finding program that do not match to any binding site from ChIP-PET (Lin *et al*,2007) or ChIP-chip (Carroll *et al*, 2006) datasets. (**A**) and (**B**) Average ChIP-chip FAIRE and H3K4me1signal for the quartile 1-4 ChIP-seq ERα binding sites after E2 induction. (**C**) & (**D**), Average FAIRE and H3K4me1signal for the 1,000 strongest induced ER binding sites for profiling the signal between E2 stimulated or non-stimulated samples.

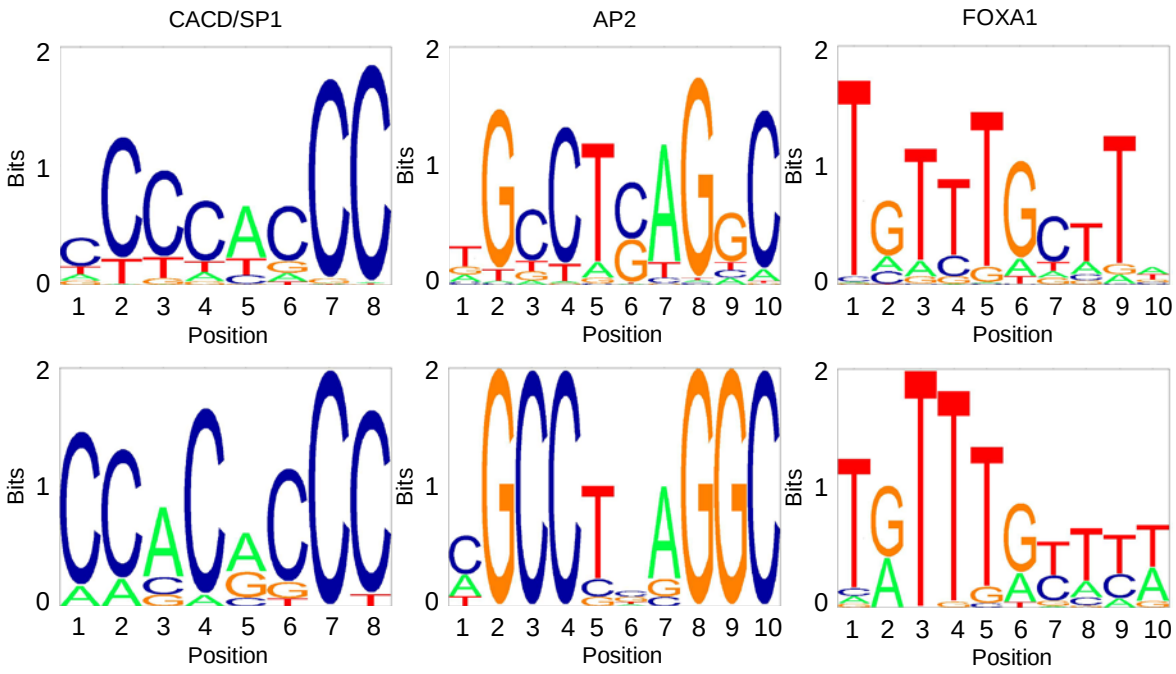


Supplementary Figure 7 Enrichment of chromatin signals at ERBS in MCF-7 cells by quartiles ranked by ER binding. ChIP-seq identified ER binding sites from MCF-7 cells were ranked by quartiles 1, 2, 3 and 4 based on the ER ChIP-seq binding occupancy were analyzed for average enrichment of chromatin marks in +/-250bp from the centre of binding sites. Enrichment of chromatin marks at E2 treatment is shown by brown color outlined empty bars and vehicle treatment is shown by green color outlined filled bars. Columns show average count of chromatin marks, whiskers indicate standard error; Y axis scale is common for E2 treated and non-treated conditions. ChIP-seq libraries were downsampled to 7M tags to make plots comparable. Trend in average chromatin mark enrichment from strongest Q1 binding sites to weakest Q4 binding sites is significant for all activation chromatin marks, FAIRE and RNA Pol II. At the same time repressive chromatin marks H3K27me3 and H3K9me3 have low signal for all quartiles and trend between quartiles as well as between treatments is not different.

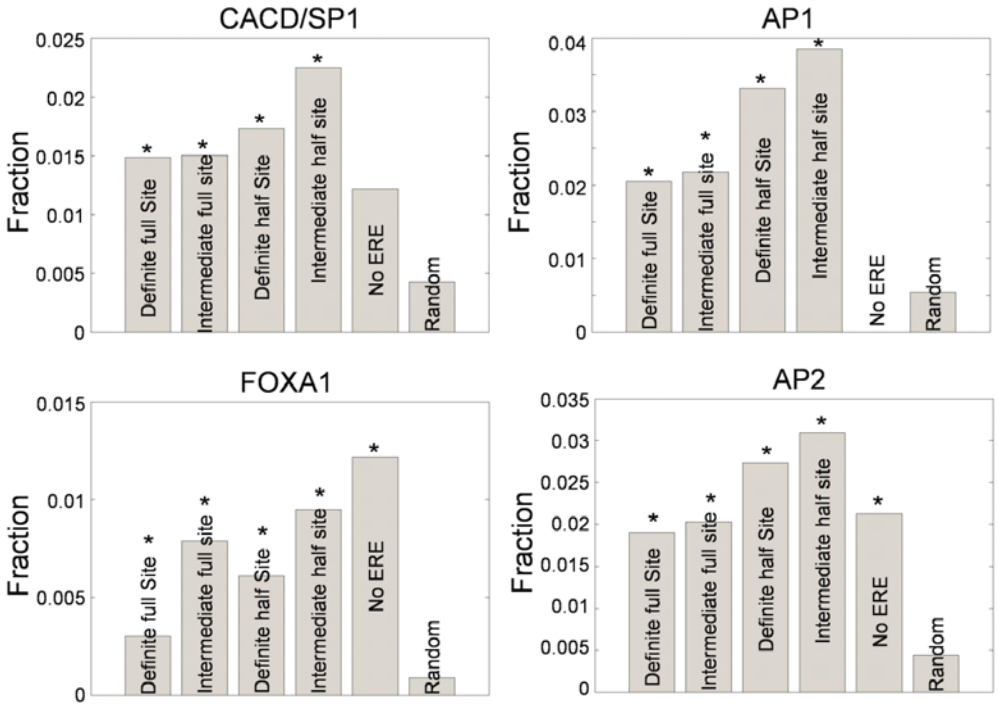
A



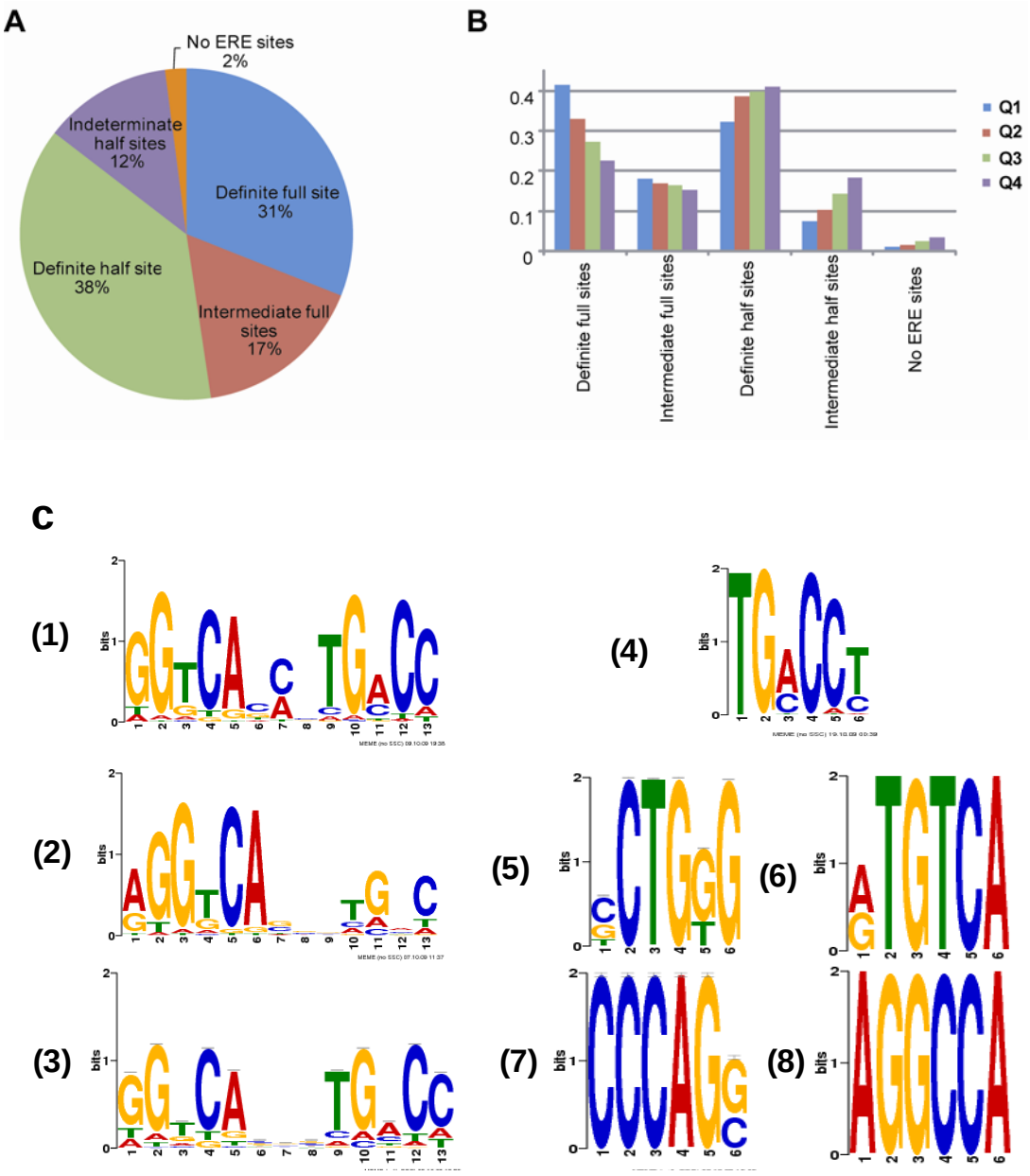
B



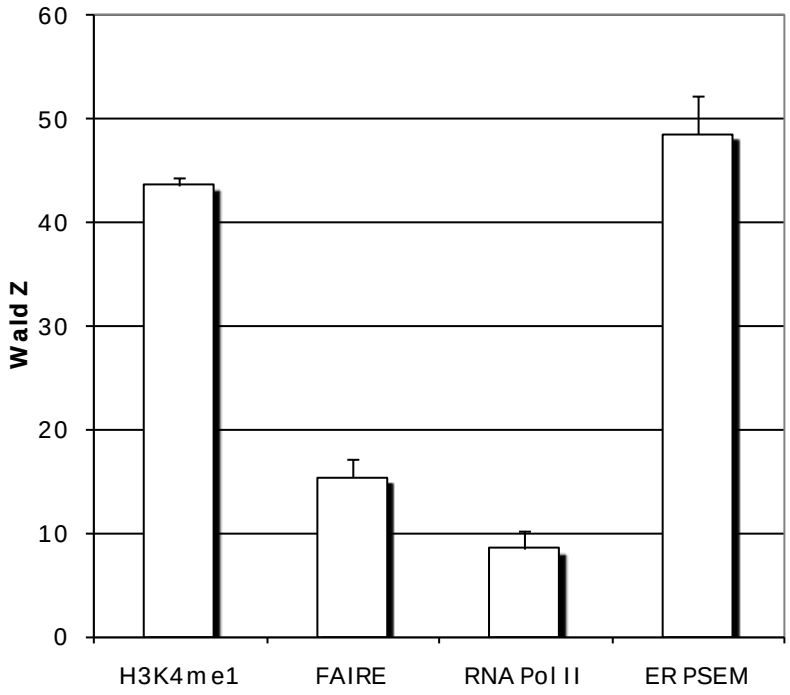
Supplementary Figure 8 Co-motif analysis at ER binding sites. ER co-motifs identified by MDscan in the vicinity of **(A)** definite ER half sites or **(B)** noERE binding regions. In both subfigures, the top panel displays MDscan motifs and the bottom panel illustrates the matching TRANSFAC motifs, for comparison.



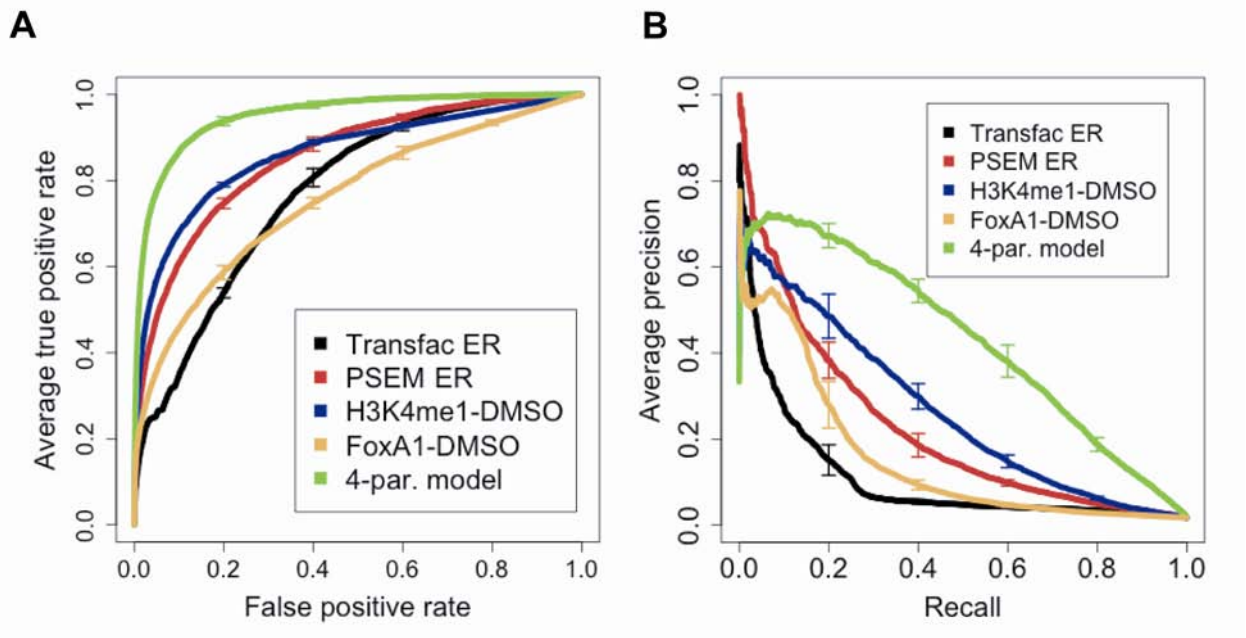
Supplementary Figure 9 *De novo* comotif analysis at ER binding sites. Frequency of occurrence of TRANSFAC (Matys *et al*, 2006), CACD, AP1, Forkhead (FOXA1) and AP2-alpha motifs in the five categories of ER binding region (definite full, intermediate full, definite half, intermediate half and no ERE sites), compared to random noncoding genomic regions. "*" indicates statistically significant deviation (P<0.05) from the motif frequency in random regions. Exact P-values for this analysis are indicated in Supplementary Table IV.



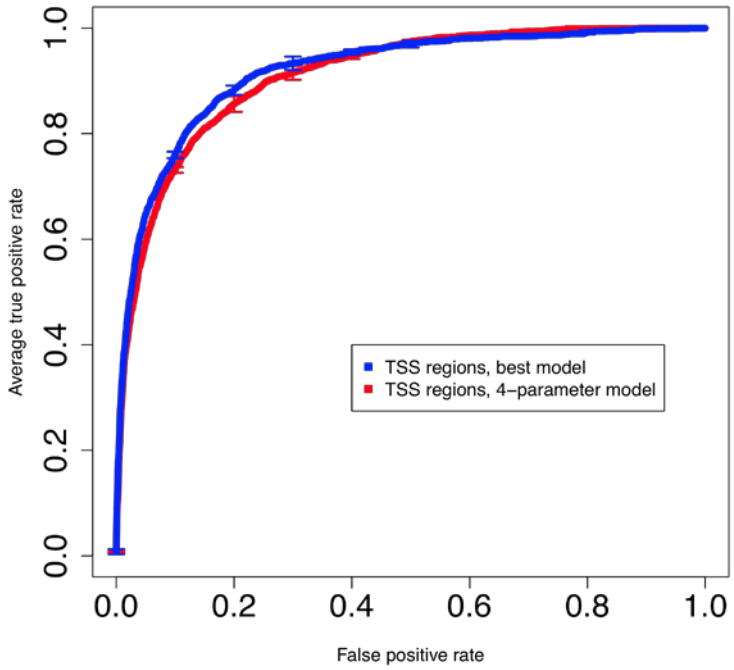
Supplementary Figure 10 TherMoS-defined ER binding site categories (definite full sites, intermediate full sites, definite half sites, intermediate half sites, and no ERE sites) and motif analysis. **(A)** Classification of ER ChIP-seq peaks into five subcategories, based on declining free energy scores of the underlying sequence motifs **(B)** Distribution of full, half or no EREs in 16K ChIP-seq ER binding sites by quartiles based on the PSEM. **(C)** Motifs found *de novo* using MEME with 100 bp sequences around ER ChIP-seq peaks in the five TherMos defined sub categories. For full and intermediate full sites, MEME was constrained to finding 13 bp motifs; for the other categories it was constrained to find 6bp motifs. 1. Best MEME full site motif. 2. Best MEME intermediate full site motif. 3. Best MEME intermediate full site motif constrained to finding palindromes. 4. Best MEME half site motif. 5. Best MEME intermediate half site motif. 6. Best MEME intermediate half site motif when initialized from the consensus sequence AGGTCA. 7. Best MEME no-ERE motif. 8. Best MEME no-ERE motif when initialized from consensus sequence AGGTCA (showing no recognizable ERE motif).



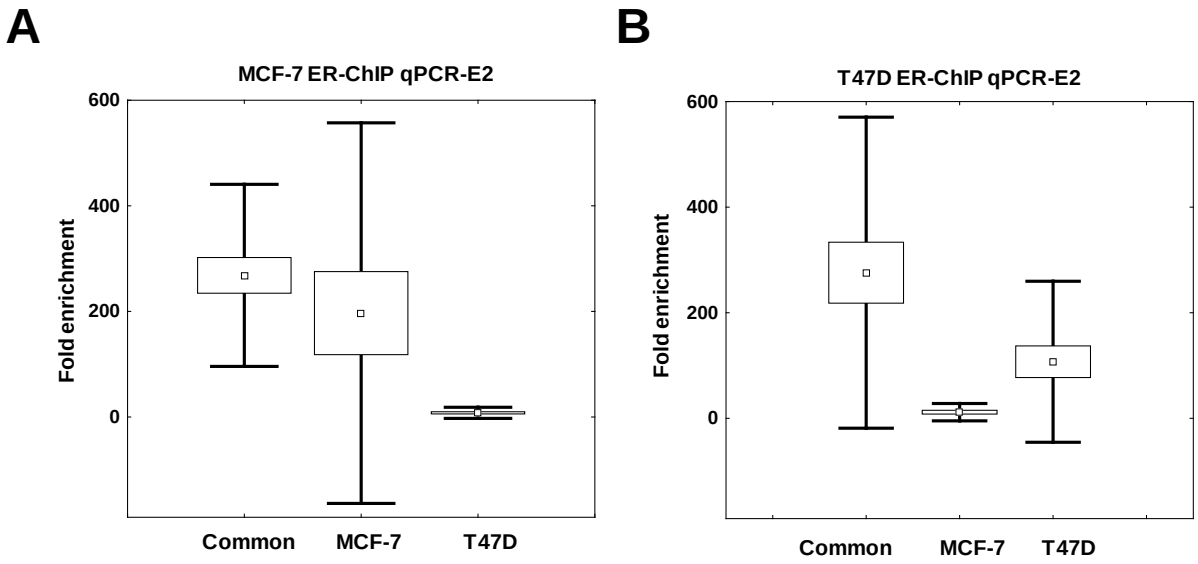
Supplementary Figure 11 The relative contributions of the ER binding site predictors quantified through the Wald Z scores using the fitted regression model (in MCF-7 cells).



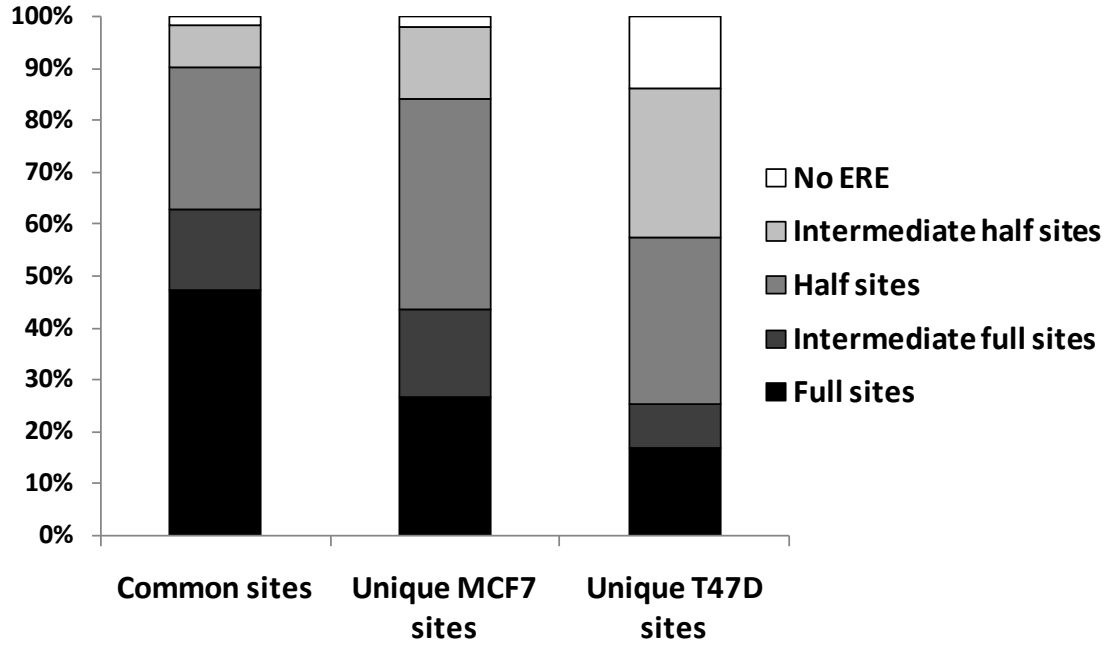
Supplementary Figure 12 Predictive factors of ER binding. MCF-7 ChIP-seq libraries were downsampled to 7 million reads. ROC and precision-recall curves for TherMoS PSEM ER, TRANSFAC PWM ER, H3K4me1 tag counts, and 4-parameter logistic regression models with PSEM score, H3K4me1, FOXA1 and FAIRE tag counts as predictors. The values are averages of five runs and bars show standard deviations. (**A**) ROC curves and (**B**) Precision-recall curves. Note that, when using ChIP-seq data normalized for 7 million reads for each of the parameters, the results showed the same high ROC AUC (0.952 for a model using the same four parameters) with the results using the entire tag data). See (Figure 4, main text).



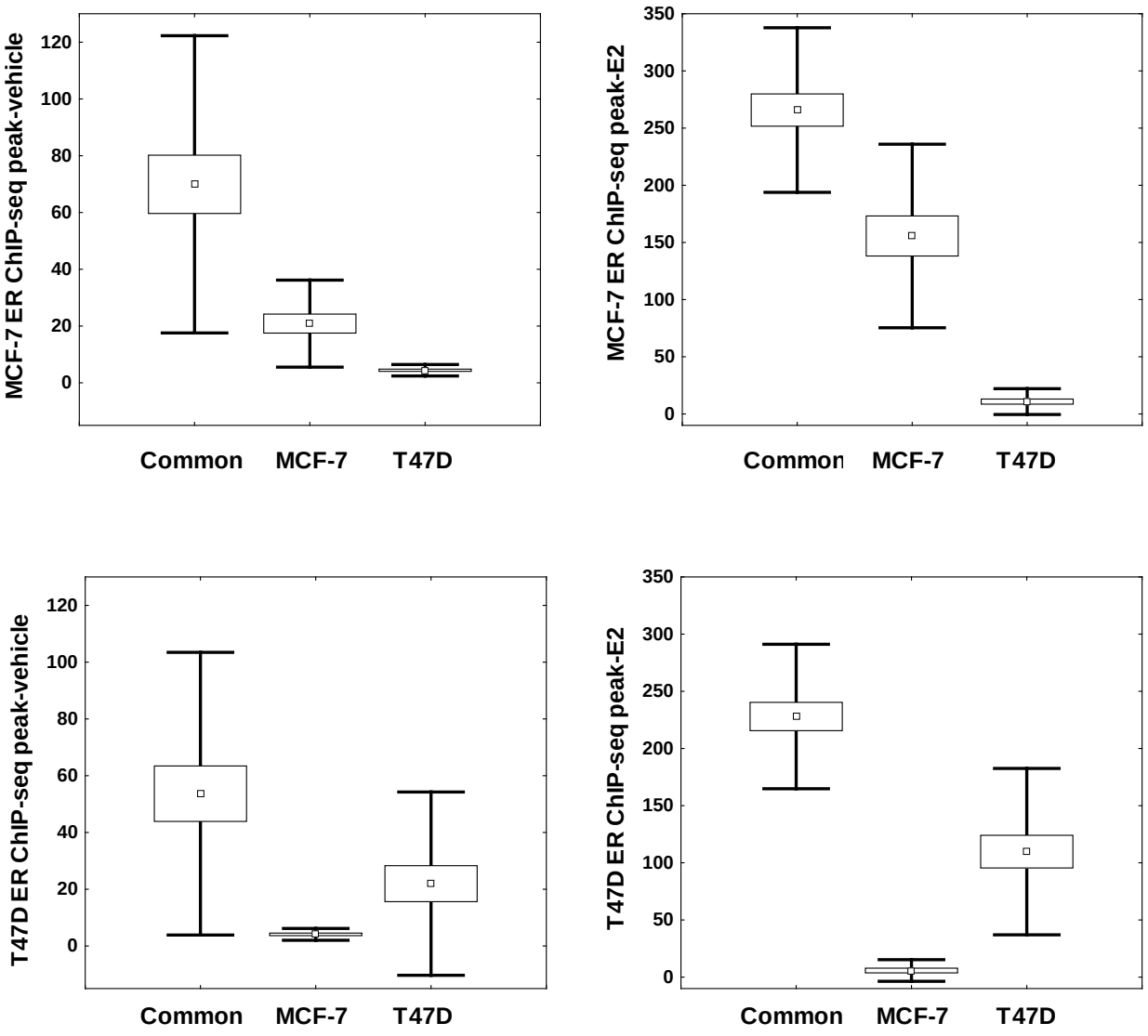
Supplementary Figure 13 Discriminating between TSS regions ([-500..-1 bp relative to RefSeq TSS]) that will be bound by ER after E2 treatment and between those that won't. Proximal promoter regions ([-500..-1 bp] intervals relative to RefSeq TSS]) were classified as ER bound and unbound by presence of ChIP-seq defined ER binding site (after E2 treatment in MCF-7 cells). A model using the 4-parameter model consisting of the TherMoS ER affinity score, H3K4me1, FOXA1 and FAIRE achieves a ROC-AUC of 0.915 (estimated by 5-fold cross-validation) while the best model, which uses all features, achieves a ROC-AUC of 0.9225 (estimated in the same way).



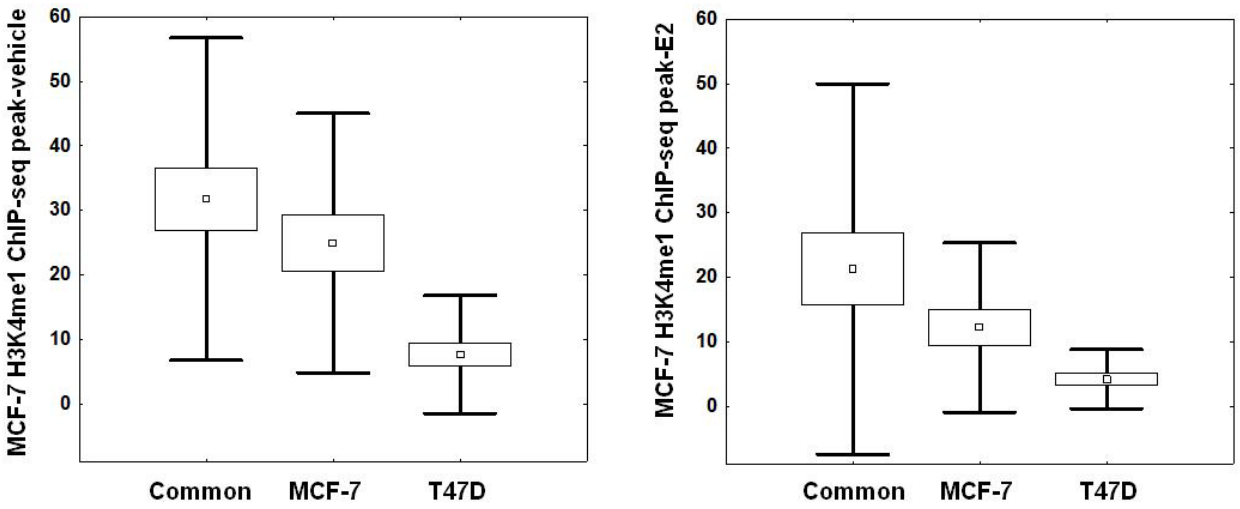
Supplementary Figure 14 qPCR validation of selected ER binding sites for their ER occupancy in (A) MCF-7 cells and (B) T47D cells. Specificity of the sites unique for cell lines was confirmed for binding in MCF-7 and T47D cells correspondingly, while common sites have higher binding affinity in both cell lines. ChIP enriched DNA (after E2) was used for syber green based qPCR experiments and fold enrichment was assessed. We used nearly 73 sites in total; 26 sites that are unique for MCF-7 cells, 26 sites that are unique for T47D and 21 sites that are common for both cell lines. ChIP followed by qPCR results are shown in (**A**) MCF-7 and (**B**) T47D cells.



Supplementary Figure 15 Distribution of different categories of binding sites in MCF-7 and T47D cells. Fraction of different ERα binding sites (percentage scale) in common or unique sites for MCF-7 and T47D cell lines based on the ChIP-seq identified ERα binding sites from MCF-7 (16,043) and T47D (5,421) cells. Using binding sites from non-amplified regions we compared here 3,335 ER binding sites common for both cell lines, 12,707 MCF-7-specific and 1,685 T47D-specific sites. Note that, fraction of full and intermediate full ERE sites is significantly higher in common sites. Also fraction of MCF-7 unique full sites is higher than in T47D unique ERα ChIP-seq sites.



Supplementary Figure 16 Average ER ChIP-seq tag count for ER binding sites either at all sites in common between the MCF-7 and T47D cell lines or sites specific for each cell line. Tag counts are shown for MCF-7 (upper panel) and T47D (lower panel) ChIP-seq libraries either before or after E2 activation. Common and MCF-7 specific binding sites have higher signal for ER binding in MCF-7 cell ChIP-seq data than T47D specific binding sites and common and T47D specific binding sites have higher signal for ER binding in T47D ChIP-seq data than MCF-7 specific binding sites (differences are statistically significant, $P<0.05$ by Mann-Whitney U Test). In both cases, the ER binding sites in common between the cell lines have greater ER occupancy than the cell line specific binding sites.



Supplementary Figure 17 Average ChIP-seq tag count found in MCF-7 for H3K4me1 mark at either all common or all specific sites for MCF-7 and T47D cell lines. Tag counts are shown for MCF-7 ChIP-seq library either before (left panel) or after E2 activation (right panel). Common and MCF-7 specific binding sites have higher signal for H3K4me1 mark in MCF-7 cells than in T47D specific binding sites (differences are statistically significant, $P<0.05$ by Mann-Whitney U Test).

Supplementary Tables

Supplementary Table I Total tag numbers and antibody used for each ChIP sample preparation. FAIRE is not a ChIP sample, but an enriched genomic DNA sample from open chromatin region by a phenol-chloroform extraction method as previously explained (Giresi *et al*, 2007).

Transcription factor/ Chromatin modification	**Cell line**	**Treatment**	**Antibody**	**Total library size (Number of unique tags, in thousands)**
ER α	MCF-7	E2	Cat# HC-20, Santa Cruz	7,009
		Vehicle		12,658
	T47D	E2		7,640
		Vehicle		11,724
FAIRE	MCF-7	E2	--	12,612
		Vehicle		12,287
	T47D	E2		20,258
		Vehicle		14,860
RNA Pol II	MCF-7	E2	Cat# ab5408, Abcam	7,556
		Vehicle		9,556
H3K9me3	MCF-7	E2	Cat# ab8898, Abcam	13,789
		Vehicle		14,846
H3K27me3	MCF-7	E2	Cat# 07-449, Upstate Biotechnology Inc.	17,253
		Vehicle		14,686
H3K4me1	MCF-7	E2	Cat# ab8895, Abcam	7,705
		Vehicle		10,171
	T47D	E2		18,377
		Vehicle		17,067
H3K4me3	MCF-7	E2	Cat# ab8580, Abcam	16,962
		Vehicle		14,162
H3K9ac	MCF-7	E2	Cat# 07-352, Upstate Biotechnology Inc.	8,527
		Vehicle		7,600
H3K14ac	MCF-7	E2	Cat# 07-353, Upstate Biotechnology Inc.	11,174
		Vehicle		9,276
FOXA1	MCF-7	E2	Cat# AB4124, Chemicon	13,182
		Vehicle		17,932
	T47D	E2		6,764
		Vehicle		14,860
c-Fos	MCF-7	E2	Cat# sc-7202, Santa Cruz	18,222
		Vehicle		15,261
c-Jun	MCF-7	E2	Cat# sc-45, Santa Cruz	15,696
		Vehicle		14,632

Supplementary Table II Overlap of ER ChIP-seq binding sites with published datasets.

	ER-ChIP-seq (this study)	ChIP-PET (Lin *et al*, 2007)	ChIP-on-chip (Carroll *et al*, 2006)	ER-ChIP-seq (Welboren *et al*, 2009)
Total no. of sites	16,043	1,226	3,665	10,191
ER-ChIP-seq (this study)	-	841	3,152	5,823
ChIP-PET (Lin *et al*, 2007)	838	-	610	814
ChIP-on-chip (Carroll *et al*, 2006)	2,977	608	-	2,041
ER-ChIP-seq (Welboren *et al*, 2009)	6,291	955	2,590	-

***** the overlaps were counted by 200 bp distance between sites from the datasets under comparison. Matrix of overlaps may be non-symmetrical due to the fact that one site can overlap two other sites from other dataset.

Supplementary Table III Overlap of ER ChIP-seq binding sites by quartiles.

	ChIP-PET (Lin *et al*, 2007)	ChIP-chip (Carroll *et al*, 2006)	ChIP-seq (Welboren *et al*, 2009)	All data overlap
Q1	626	1,945	2,749	407
Q2	129	727	1,497	32
Q3	56	319	969	6
Q4	30	161	608	1
Total No. of overlapping sites	841	3,152	5,823	446
Total No. of sites	1,226	3,665	10,191	-

Supplementary Table IV *P*-values of difference of frequency of occurrence of SP1, AP1, FoxA1 and AP2-alpha motifs between the five categories of ER binding region and random noncoding genomic regions (Fisher's exact test).

TF	Definite full site	Intermediate full site	Definite half site	Intermediate half site	No ERE
SP1	5.13E-19	3.06E-11	3.29E-31	1.46E-18	5.34E-02
AP1	6.88E-28	6.99E-18	1.53E-85	9.72E-39	4.33E-01
FOXA1	7.83E-05	2.50E-13	1.79E-18	1.36E-13	2.52E-04
AP2-alphaA	3.85E-30	3.93E-19	1.37E-71	3.56E-31	7.79E-04

Supplementary Table V Distribution of ER sites with respect to gene borders. *P*-values of the statistical difference between ChIP-seq sites and random set of loci for each category; p-value of the difference between ER ChIP-seq sites and unbound ERE sites (predicted by PWM) by Fisher's exact test (see Supplementary Figure 2).

	5' Distal	Promoter	Intra-genic	3' end	3' Distal	Gene desert
ChIP-seq ER sites	2792	1485	6344	645	2299	2478
Random genome loci	1473	359	3275	246	1141	3506
Consensus ERE motif (GGTCAnnnTGACC) sites*	9675	2362	20297	1797	7603	15012
ChIP-seq ER vs Random: Fisher's exact test *P*-value of difference (2-sided)	1.32E-08	1.84E-73	1.56E-28	6.8E-12	9.33E-12	1.23E-286
ChIP-seq ER vs Consensus ERE motif: Fisher's exact test *P*-value (2-sided)	0.296	4.42E-126	2.62E-18	2.05E-07	0.00248	4.29E-196

* 1 mismatch to consensus ERE motif was allowed. Motif defined ERE sites that overlapped with ChIP-seq defined ER sites were excluded. In total 56,746 motif sites in human genome (hg18) were tested.

Supplementary Table VI Statistical significance of difference between histone modification marks enrichment on cell line specific ER binding sites in MCF-7 and T47D cells (ERBS are MCF-7, T47D specific and common between cell lines). Z statistics and corresponding *P*-values are given by Mann-Whitney U test (see Figure 5, main text).

	Z statistics (*P*-value) of difference in chromatin marks for ERBS		
	Common vs MCF-7	**MCF-7 vs T47D**	**Common vs T47D**
ER affinity score	25.88(<1E-16)	17.61(<1E-16)	28.71(<1E-16)
ChIP-seq in MCF-7 cells			
FOXA1 (Vehicle)	0.54 (>0.1)	18.07 (<1E-16)	15.35 (<1E16)
H3K4me1 (Vehicle)	26.46 (>0.1)	25.74007 (<1E16)	23.08011(<1E16)
FAIRE (Vehicle)	0.148(>0.1)	24.06 (<1E16)	20.75 (<1E16)
ChIP-seq in T47D cells			
FOXA1 (Vehicle)	0.347 (>0.1)	-7.53 (4.95E-14)	-6.77 (1.29E-11)
H3K4me1 (Vehicle)	0.739 (>0.1)	26.45 (<1E-16)	22.23 (<1E-16)
FAIRE (Vehicle)	0.238 (>0.1)	-12.34 (<1E-16)	-10.56 (4.42E-26)

Supplementary Table VII Rank correlations between ER ChIP-seq peak height and count of chromatin modification tags depending on library size: downsampling to 1 and 5 millions uniquely mapped tags. Statistically significant values are given in bold.

Chromatin mark (vehicle)	Rank correlations (Kendall tau)			*P*-value		
	1M tags	5M tags	all tags	1M tags	5M tags	all tags
H3K4me1	**0.1308**	**0.1416**	**0.1512**	**2.54E-136**	**2.12E-159**	**1.49E-181**
RNA Pol II	**0.095**	**0.1348**	**0.1452**	**8.26E-73**	**1.17E-144**	**1.62E-167**
H3K9ac	**0.0921**	**0.1196**	**0.1352**	**1.71E-68**	**3.20E-114**	**1.80E-145**
FAIRE	**0.0786**	**0.1163**	**0.1469**	**2.02E-50**	**4.16E-108**	**2.14E-171**
H3K4me3	**0.0639**	**0.093**	**0.1187**	**7.16E-34**	**6.90E-70**	**1.49E-112**
H3K14ac	**0.0389**	**0.0655**	**0.0762**	**1.49E-13**	**1.50E-35**	**5.05E-07**
Input DNA	0.002	-0.0001	-0.0010	7.08E-01	9.82E-01	8.46E-01
H3K9me3	-0.0052	-0.0185	-0.0264	3.26E-01	**4.46E-04**	**5.05E-07**
H3K27me3	-0.0163	-0.0287	-0.0323	1.94E-03	**4.76E-08**	**8.89E-10**
FOXA1	**0.098**	**0.1565**	**0.1872**	**2.44E-77**	**3.75E-194**	**5.07E-277**
AP1 (cFos+cJun)	**0.0218**	**0.0375**	**0.0606**	**3.35E-05**	**1.69E-07**	**1.20E-30**

Supplementary Table VIII Best N-feature models for discrimination task 1 (bound vs. random distal genomic regions) using five-fold validation. “ER binding” stays for free energy based ER affinity score.

N	Most frequent top-scoring feature combination	Mean ROC-AUC
1	* H3K4me1	0.870
2	* H3K4me1+ER binding	0.929
3	* H3K4me1+ER binding+FOXA1	0.949
4	* H3K4me1+ER binding+FOXA1+FAIRE	0.952
5	H3K4me1+ER binding+FOXA1+FAIRE+RNA Pol II	0.953
6	H3K4me1+ER binding+FOXA1+FAIRE+RNA Pol II+H3K9ac	0.953
7	H3K4me1+ER binding+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3	0.953
8	H3K4me1+ER binding+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K27me3	0.950
9	H3K4me1+ER binding+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K9me3+ H3K14ac	0.948
10	H3K4me1+ER binding+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K9me3+ H3K14ac+H3K27me3	0.946
11	H3K4me1+ER binding+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K9me3+ H3K14ac+H3K27me3+cJun	0.945

Note: * *Combination appeared 5 times of 5 fold validations.*

Supplementary Table IX Best N-feature models for discrimination task 2 (bound vs. non-bound TSS-proximal regions) assessed as in Supplementary Table VIII.

N	Most frequent top-scoring feature combination	Mean ROC-AUC
1	ER binding	0.833
2	ER binding+H3K4me1	0.900
3	ER binding+H3K4me1+FOXA1	0.912
4	ER binding+H3K4me1+FOXA1+RNA Pol II	0.916
5	ER binding+H3K4me1+FOXA1+cFos+FAIRE	0.918
6	* *multiple combinations of 6 features from the 12*	0.920
7	* *multiple combinations of 7 features from the 12*	0.921
8	* *multiple combinations of 8 features from the 12*	0.922
9	* *multiple combinations of 9 features from the 12*	0.922
10	* *multiple combinations of 10 features from the 12*	0.922
11	* *multiple combinations of 11 features from the 12*	0.923

Note: * *No single combination of these N features of the 12 was noted to be the predominant combination.*

Supplementary Table X Best N-feature models for discrimination task 3 (bound vs. non-bound predicted EREs; distal).

N	Best feature combination	Mean ROC-AUC
1	H3K4me1	0.841
2	H3K4me1+FOXA1	0.870
3	H3K4me1+FOXA1+H3K9ac	0.880
4	H3K4me1+FOXA1+H3K4me3+FAIRE	0.882
5	H3K4me1+FOXA1+H3K9ac+FAIRE+RNA Pol II	0.883
6	H3K4me1+FOXA1+H3K9ac+FAIRE+RNA Pol II+H3K4me3	0.880
7	H3K4me1+FOXA1+H3K9ac+cFos+cJun+H3K9me3+ H3K27me3	0.872
8	H3K4me1+FOXA1+H3K9ac+FAIRE+cJun+H3K9me3+ H3K27me3+H3K14ac	0.863
9	H3K4me1+FOXA1+H3K9ac+FAIRE+H3K4me3+cJun+ H3K9me3+H3K27me3+H3K14ac	0.864
10	H3K4me1+FOXA1+H3K9ac+FAIRE+RNA Pol II+ H3K4me3+cFos+ H3K9me3+H3K27me3+H3K14ac	0.866

Supplementary Table XI Best N-feature models for discrimination task 3 (bound vs. non-bound predicted EREs; TSS-proximal)

N	Best feature combination	Mean ROC-AUC
1	FOXA1	0.699
2	FOXA1+H3K4me1	0.749
3	FOXA1+H3K9me3+H3K14ac	0.790
4	FOXA1+H3K14ac+cFos+cJun	0.798
5	H3K4me1+H3K9me3+H3K9ac+FAIRE+RNA Pol II	0.800
6	FOXA1+H3K4me1+cJun+H3K9ac+H3K4me3+H3K27me3	0.805
7	FOXA1+H3K4me1+H3K9me3+H3K14ac+FAIRE+H3K27me3+ H3K4me3	0.824
8	FOXA1+H3K4me1+cJun+H3K9me3+H3K9ac+H3K4me3+ FAIRE+H3K27me3	0.806
9	FOXA1+H3K4me1+cFos+cJun+H3K9me3+H3K9ac+FAIRE+ H3K4me3+H3K27me3	0.829
10	FOXA1+H3K4me1+H3K14ac+cFos+cJun+H3K9me3+H3K9ac+ FAIRE+RNA Pol II+H3K27me3	0.796

Supplementary Table XII Comparison of ER binding prediction for sites with ER TRANSFAC motif using GBP (general binding profile) from Ernst et al. (2010) and cell-specific chromatin marks in MCF-7 and T47D cells.

ER sites	GBP (Ernst *et al*, 2010)	H3K4me1	H3K4me1+ FOXA1	H3K4me1+ FAIRE
MCF-7 cells				
All in MCF-7	0.788	0.848	0.861	0.866
Distal	0.783	0.847	0.861	0.868
TSS-proximal	0.786	0.788	0.812	0.811
T47D				
All in T47D	0.804	0.809	0.879	0.841
Distal	0.802	0.812	0.879	0.847
TSS-proximal	0.789	0.733	0.860	0.763

Supplementary Table XIII PSEM (Position Specific Energy Matrix) of ERE (Scale factor τ = 2.39E-07). 17-mer contains left (L) and right (R) half-ERE sites.

		A	T	G	C
1	L1	0.6791	0.6132	0.8886	0
2	L2	0	1.5384	0.2713	0.8043
3	L3	1.3348	1.2546	0	3.246
4	L4	1.4382	1.6328	0	3.439
5	L5	1.2157	0	0.6739	1.9512
6	L6	2.9798	1.4546	1.8694	0
7	L7	0	1.8863	0.7029	2.2217
8		0	0	0	0
9		0	0	0	0
10		0	0	0	0
11	R1	1.8863	0	2.2217	0.7029
12	R2	1.4546	2.9798	0	1.8694
13	R3	0	1.2157	1.9512	0.6739
14	R4	1.6328	1.4382	3.439	0
15	R5	1.2546	1.3348	3.246	0
16	R6	1.5384	0	0.8043	0.2713
17	R7	0.6132	0.6791	0	0.8886

Supplementary references

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37:** W202-208

Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, Wang Q, Bekiranov S, Sementchenko V, Fox EA, Silver PA, Gingeras TR, Liu XS, Brown M (2006) Genome-wide analysis of estrogen receptor binding sites. *Nature Genetics* **38:** 1289-1297

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK *et al*. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:**1106-1117

De Santa F, Narang V, Yap ZH, Tusi BK, Burgold T, Austenaa L, Bucci G, Caganova M, Notarbartolo S, Casola S, Testa G, Sung WK, Wei CL, Natoli G. (2009) Jmjd3 contributes to the control of gene expression in LPS-activated macrophages. *EMBO J.* **28**: 3341-52.

Ernst J, Plasterer HL, Simon I, Bar-Joseph Z (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res* **20**: 526-536

Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by Matrix REDUCE. *Bioinformatics* **22:** e141-149

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EGY, Huang PYH, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KDSA *et al*. (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462:** 58-64

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17:** 877-885

Hurtado A, Holmes KA, Geistlinger TR, Hutcheson IR, Nicholson RI, Brown M, Jiang J, Howat WJ, Ali S, Carroll JS (2008) Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen. *Nature* **456:** 663-666

Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F, Yeo A, George J, Kuznetsov VA, Lee YK, Charn TH, Palanisamy N, Miller LD, Cheung E, Katzenellenbogen BS, Ruan Y *et al*. (2007) Whole genome cartography of estrogen receptor α binding sites. *PLOS Genet* **3:** e87

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B,Saxel H, Kel AE, Wingender E (2006) TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34:** D108-110

Roider HG, Kanhere A, Manke T, Vingron M (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* **23:**134-141

Safe S, Kim K (2008) Non-classical genomic estrogen receptor (ER)/specificity protein and ER/activating protein-1 signaling pathways. *J Mol Endocrinol* **41:** 263–275

Shadeo A, Lam WL (2006) Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res* **8:** R9

Stormo GD. (2000) DNA binding sites: representation and discovery. *Bioinformatics* **16:** 16-23.

Vega VB, Lin CY, Lai KS, Kong SL, Xie M, Su X, Teh HF, Thomsen JS, Yeo AL, Sung WK, Bourque G, Liu ET (2006) Multiplatform genome-wide identification and modeling of functional human estrogen receptor binding sites. *Genome Biol* **7:** R82

Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG (2009) ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *The EMBO J* **28:** 1418-1428