Published online December 17, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 22 ***6627–6635***
doi:10.1093/nar/gkh1005

Identifying estrogen receptor α target genes using integrated computational genomics and chromatin immunoprecipitation microarray

Victor X. Jin, Yu-Wei Leu, Sandya Liyanarachchi, Hao Sun, Meiyun Fan[1], Kenneth P. Nephew[1], Tim H.-M. Huang and Ramana V. Davuluri*

Human Cancer Genetics Program, Department of Molecular Virology, Immunology, and Medical Genetics, The Ohio State University, Columbus, OH 43210, USA and [1]Medical Sciences Program, Indiana University School of Medicine, Bloomington, IN 47405, USA

Received September 24, 2004; Revised November 8, 2004; Accepted November 29, 2004

ABSTRACT

The estrogen receptor α (ERα) regulates gene expression by either direct binding to estrogen response elements or indirect tethering to other transcription factors on promoter targets. To identify these promoter sequences, we conducted a genome-wide screening with a novel microarray technique called ChIP-on-chip. A set of 70 candidate ERα loci were identified and the corresponding promoter sequences were analyzed by statistical pattern recognition and comparative genomics approaches. We found mouse counterparts for 63 of these loci and classified 42 (67%) as direct ERα targets using classification and regression tree (CART) statistical model, which involves position weight matrix and human-mouse sequence similarity scores as model parameters. The remaining genes were considered to be indirect targets. To validate this computational prediction, we conducted an additional ChIP-on-chip assay that identified acetylated chromatin components in active ERα promoters. Of the 27 loci upregulated in an ERα-positive breast cancer cell line, 20 having mouse counterparts were correctly predicted by CART. This integrated approach, therefore, sets a paradigm in which the iterative process of model refinement and experimental verification will continue until an accurate prediction of promoter target sequences is derived.

INTRODUCTION

Recent completion of human and mouse genome sequences and accumulation of an increasing number of gene annotations have made it possible for bioinformaticians to develop new approaches that help experimental researchers tackle biological problems. To fully understand the regulation of transcription by estrogen receptors (ER) α and β, members of the nuclear receptor superfamily, computational approaches capable of integrating vast amounts of complex genomic data are needed. ERs mediate estrogen signaling, primarily by 17β-estradiol, in various target tissues, including reproductive, bone, cardiovascular and the central nervous system (1). Estrogens and ERs also play important roles in breast cancer genesis and progression (2), and tumor ER status is a critical determinant in breast cancer patients to elucidate response to adjuvant treatment with endocrine agents (2). Thus, a better understanding of ERs may lead to advances in both normal physiology and disease states, which requires in-depth understanding of the spectrum of genes regulated by ERs in different tissues and cell types.

ERs function as ligand-inducible transcription factors (TFs) that either up- or down-regulate transcription of various target genes by binding to specific estrogen response elements (EREs) or interacting with other TFs, such as SP1, nuclear factor-κB or AP1 (3–6). Both processes result in recruitment of co-activators and components of RNA polymerase II that initiate gene transcription (3). The ERE consensus sequence, an inverted repeat of the sequence (GGTCA) separated by 3 bp, rarely occurs in nature (7); however, the imperfect ERE (GGTCANNNTNNCY) and ERE half-site (AGGTCA) are widely accepted as alternative binding sites (8–11). Binding to different ERE sequences alters the conformation of ER, allowing interaction with co-activators in a cell-type and DNA context-dependent manner (12–15). Although the interaction between ERs and EREs is under intense investigation (10,11), few studies have utilized the vast amount of genomic data available in the post-genome era. Thus, the development of a systematic computational approach not only contributes significantly to ongoing research in the characteristics of ER binding, but also allows for a better understanding of its functional connections in a cell.

Computational tools widely used to identify TF binding sites include MATCH (16) and MSCAN (17). However, due to lack of experimental verification, these tools are prone to false predictions. Consequently, there is a need to facilitate interactions between computational and experimental

*To whom correspondence should be addressed. Tel: +1 614 688 3088; Fax: +1 614 688 4006; Email: davuluri-1@medctr.osu.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

scientists who conduct integrated research for the identification of TF binding sites.

In this study, we combined a systematic computational approach and microarray-based ChIP-on-chip for the genome-wide identification of ERα target genes. Our computational approach entailed the development of a classification and regression tree (CART) model and the implementation of OMGProm (18), a comparative genomics database for the analysis of human/mouse orthologous promoters. A dataset containing 70 ERα candidate loci was created by the ChIP-on-chip screening of ~9000 putative GC-rich promoter sequences (19). A CART model was built to predict ERα promoter sequences and experimentally re-verified by ChIP-on-chip using a small ERα genomic microarray panel.

MATERIALS AND METHODS

Promoter sequence retrieval

The orthologous promoter sequences, corresponding attributes and annotation data were stored in a relational database, OMGProm, recently developed in our laboratory (18) (http://bioinformatics.med.ohio-state.edu/OMGProm). The OMGProm data were obtained via an efficient data mining pipeline, which collected experimentally substantiated full-length mRNA/5′-untraslated regions, first exons and promoters from GenBank, dbEST, RefSeq and Ensembl. A 5′ flanking region of 2 kb upstream to 1 kb downstream of the transcription start site (TSS) was designated as an ERα promoter sequence, since most of the experimentally known EREs are located within these regions (10).

Computational approaches to identify EREs and other binding sites

We used human-mouse orthologous promoters in OMGProm to map the 46 experimentally known EREs within 38 target genes. An ERE position weight matrix (ERE_PWM) was then constructed by using the TRANSFAC position weight matrix [TRANSFAC_PWM (20)] procedure, as shown in Supplementary Table 1. A computational program, ConScan, was developed in Perl and C languages to scan for conserved putative EREs in the ClustalW (21) sequence alignments of human-mouse orthologous promoters in the OMGProm database. Each predicted ERE had scores for five parameters: (i) human core score, (ii) human PWM score (iii) mouse core score, (iv) mouse PWM core score and (v) sequence similarity score of 13 bp ERE in the sequence alignment. The core score and PWM score, ranging from 0 to 1, reflect the closeness of predicted sites to the half-site and perfect ERE consensus sequences.

The other (non-ERE) TF binding sites were analyzed by the MATCH (16) program, using the PWMs from TRANSFAC database (20). For a given gene, we scanned both human and mouse orthologous promoters for all the 290 TF binding sites corresponding to known human TFs using 'minFN_good71.prf' profile (profile of cut-off values with minimum number of false-negative predictions) of MATCH. If the sequence similarity of the binding motif is ⩾60% in the ClustalW (21) sequence alignment, a predicted binding site is considered as conserved. Those binding sites that fall within the range (−220 to +220 bp) of a predicted ERE were further used in the CART analyses.

CART

CART (22) analysis was employed to develop a classification model for separating ERα targets from non-targets. The approach is an advanced data-mining tool for tree-structured non-parametric data analysis, based on binary recursive partitioning methodology (22). It partitions data into discrete classes using the value of a user-defined classification variable (e.g. target = 1 and non-target = 0) and computation-intensive searching and testing techniques to identify useful tree structures of data. CART selects the predictor variables in the data, depending on whether they provide a segregation of the data between different values of the classification variable. The 'Gini' method was selected as the splitting method for growing the tree, and the 10-fold cross validation method was used to obtain the minimal tree.

In our analysis, the CART procedure was divided into two phases. In the first phase, a CART model (model 1) was constructed to identify cut-off values for five parameters of ConScan program on two sets of promoters: ERα target promoters in ERTargetDB (see Results) and promoters of housekeeping genes (non-ERα targets). The determined cut-off values of parameters were used as predictor variables for model 1 to classify ERα targets from non-ERα targets. Only those promoters that had EREs predicted by this phase (in both targets and non-targets) were considered as a learning sample for the next phase of CART model (model 2).

In the second phase (model 2), we used ERE sequences predicted by ConScan and other TF binding sites predicted by MATCH as predictor variables. Each binding site was considered as a binary variable, such that it was either 1 or 0, depending on its presence within a −220 bp to +220 bp region of a predicted ERE (by model 1). All possible over-represented TF binding sites from the learning samples were first identified using model 2. Subsequently, these sites were used to construct a decision tree, with TF binding sites as the final categorical predictor variables for classifying ERα targets from non-ERα targets. Our analysis was performed on the commercially available CART software (Salford Systems, San Diego, CA).

ChIP-on-chip

The ERα-positive breast cancer cell line, MCF-7, was maintained in culture under conditions, as we have described previously (23). The cells (1×10^7 cells/well) were treated with 17β-estradiol (10 nM) for 24 h. The cells were then cross-linked with 1% formaldehyde, and cell nuclei were isolated as described previously (24). Isolated protein-DNA complexes were sonicated to generate smaller chromatin fragments (~500 bp). These chromatin fragments were then immunoprecipitated by an ERα antibody (Upstate) and treated with proteinase K (38 μg/ml) to reverse the cross-linked complexes. The purified genomic DNA fragments were labeled with the fluorescence dye Cy5 and then hybridized with microarray slides containing ~9000 GC-rich genomic sequences. These sequences had previously been shown to be preferentially located at the 5′ ends of genes (25). Standard hybridization and post-hybridization washing procedures, originally developed

by DeRisi, were followed (http://www.microarrays.org). Microarray slides were scanned with the GenePix 4000A scanner (Axon), and the acquired images were analyzed with the software GenePix Pro 4.0. The average ratio of signal intensities of known repeat sequences was used as a normalization factor. Normalized ratios (≥2) above pre-immune sera counterparts were considered to be ERα-positive loci. Individual sequences were determined by a standard sequencing method using flanking primers corresponding to vector sequences.

ERα target promoter microarray

A small panel of 70 ERα promoter sequences was spotted in triplicate on microarray slides. Repeat sequences were also arrayed as negative controls. A ChIP-on-chip assay was developed to determine the histone acetylation status of these 70 loci. The presence of acetylated histone 3 (AcH3) components in the promoter region is typically associated with actively transcribed genes (26). The AcH3 antibody (Upstate Catalog no.06-942) was used to immunoprecipitate chromatin from the ER+ (MCF-7) and ER− (MDA-MB-231) cells. The processed chromatin DNA was labeled with Cy5 dye and hybridized together with the total input (labeled with Cy3) onto a microarray slide. Normalized AcH3 Cy5/Cy3 ratios of these loci were calculated using GenePix Pro 4.0. A paired *t*-test was used to compare these ratios for MCF-7 and MDA-MB-231 cells.

RESULTS

Database of mammalian ERα promoters

In order to model the ERE and associated TF binding sites, we first constructed a database for mammalian ERα target promoters (ERTargetDB; http://bioinformatics.med.ohio-state.edu/ERTargetDB). ERTargetDB consists of 38 ERα targets for human, mouse and rat, with annotation of 46 EREs experimentally confirmed by individual laboratories. Other related TF binding sites and their TSSs are included. Of these 46 sequences, only 4 (9%) have perfect palindromic EREs and 37 (80%) have at least one perfect half-site, while the remaining sequences contain imperfect EREs (defined as mismatches to the consensus sequence). After we identified orthologous counterparts for all 38 genes (human for mouse/rat; mouse for human), we found that 41 (89%) of 46 EREs were conserved in human and mouse/rat (sequence similarity score >0.6 in ClustalW sequence alignments).

CART model for target classification

The 46 experimentally verified EREs and a set of 340 housekeeping gene promoters (i.e. non-targets) (27) were used to generate CART model 1. The cut-off values derived by model 1 for the five parameters of ConScan program were human core score >0.8, human PWM score >0.8, mouse core score >0.7, mouse PWM score >0.7 and sequence similarity score of the predicted ERE >0.6. After applying these cut-off values, 42 of 48 EREs, and 97 of 340 housekeeping genes, were predicted to have at least a putative ERE by the ConScan program. Of the 42 EREs, only 27 were unique, i.e. occurring in only one species. These were combined with the 97 housekeeping genes and used as the learning sample for CART model 2. The main purpose of model 2 was to discriminate ER target promoters from non-targets, although both promoters seemed to have contained at lease one ERE. A sequence length of 453 bp was trained by CART model 2 (−220 to +220 bp region surrounding the predicted ERE in both targets and housekeeping genes in the learning sample). TF binding sites, including EREs, were used as predictor variables in this model. The results showed that 32 TF binding sites were over-represented in the ERα target promoters, compared with those of the housekeeping gene promoters, with at least 20% of presence in the targets. A decision tree was constructed based on the 32 TF binding sites, and the 'Gini' method was used to obtain a minimal tree. Figure 1 illustrates a decision tree representation of CART model 2 for the minimal tree. The most discriminative feature distinguishing ERα targets from non-ERα targets for the learning sample of 124 genes was the presence of a MYOGENIN (AGCAGGTG) binding site within the 453 bp sequence around ERE. In the absence of MYOGENIN, the presence of DBP (TTTTGCT), AP1 (GCTGCGTCAGC) or GATA3 (TCCTATCGC) binding sites, but not USF2 (CAGGTG), indicated that the corresponding gene was an ERα target. Misclassification errors of the CART model by class can be found in Table 1. The classification accuracy of the minimal tree was 96%, as estimated by minimal cost of the tree. Furthermore, the model correctly predicted 96% of ERα targets and 45% of non-ERα targets.

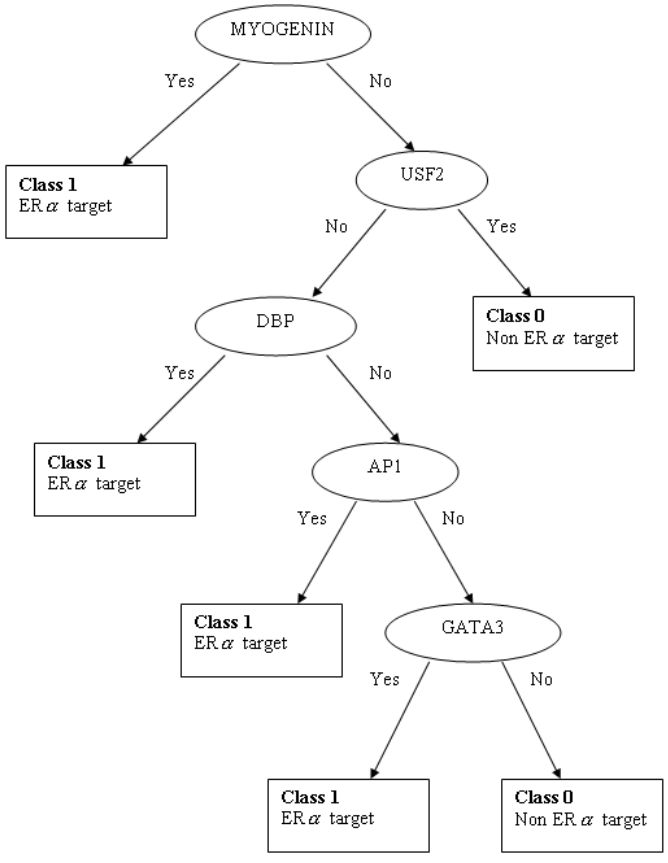

Figure 1. A CART model that discriminates ER target promoters from non-targets, using a learning sample of 27 ERα targets and 97 non-ERα targets.

Table 1. Misclassification estimates by class

Class	Learning sample Class size	Number misclassified	Percentage misclassified	Cost	Prediction success Total case	Percent correct
1 (ERα target)	27	1	3.7	0.04	27	96
0 (Non-ERα target)	97	54	55.67	0.33	97	44

The combinatorial theory of gene activation by transcription factors states that many transcription factors act together to mediate target gene activation. The close spacing of binding sites in a promoter sequence suggests that either similar proteins promote cooperative binding or dissimilar proteins provoke competition for binding sites (28). Accordingly, the identification of the putative conserved binding sites in the different ERα targets may help discover other proteins that are involved in the ERα signaling pathway. Therefore, we classified ERα targets into four modules of combinatorial control (Supplementary Figure 1), based on the discovery of over-represented TF binding sites identified by CART, Module 1: ERE+MYOGENIN (the presence of both ERE and MYOGENIN); Module 2: ERE+DBP−USF2 (the presence of ERE and DBP but the absence of USF2); Module 3: ERE+AP1−USF2 (the presence of ERE and AP1 but the absence of USF2); and Module 4: ERE+GATA3−USF2 (the presence of ERE and GATA3 but the absence of USF2). The list of target promoters that are predicted to have these different modules are presented in Supplementary Table 2.

ERα targets identified by ChIP-on-chip

A set of 70 putative ERα target candidate loci, identified by ChIP-on-chip, were mapped to the human genome by BLAT (29). Of the 70 loci, 63 corresponded to known genes with GenBank accession ID numbers (30) and seven were pseudo genes, as predicted by Genescan (31) and FirstEF (32). All 63 known genes have mouse counterparts, and their promoter sequences can be retrieved from the OMGProm database (Table 2). When the previously built CART model was applied to the 63 genes, 42 candidates fit the model and were thus considered to be ERα target genes (shown in row 2, Table 3). Among these 42 ERα target genes identified by CART, 12 (28.5%) were classified into Module 1, 17 (40.5%) were classified into Module 2, 6 (14.3%) were classified into Module 3 and 7 (16.7%) were classified into Module 4.

Comparison of CART model with other programs

The quality of our CART model was further assessed by testing it on three different datasets. Dataset 1 consisted of 46 experimentally verified EREs in 38 ERα target genes. Dataset 2 contained a set of 63 ERα candidate target genes, identified by ChIP-on-chip. Dataset 3 included 8124 promoters from the entire OMGProm database. In addition, these datasets were tested using other approaches, such as a perfect palindrome (GGTCAnnnTGACC) search, a perfect half-site (GGTCA) search, ERE_PWM and TRANSFAC_PWM. Results of these comparisons can be found in Table 3. In all three datasets, the TRANSFAC_PWM predicted the highest rate of putative EREs. The lowest prediction rate was for a perfect

Table 2. ERα target genes used on the Promoter Microarray[a]

Gene symbol	Unigene ID	ERE half-site[b]	ERE palindrome[c]	ER Binding site[d] Human	Mouse
BCAN	Hs.158244	+	−	+	+
*LOC91661**	Hs.190394	+	−	+	+
PEPP3	Hs.343666	+	−	+	+
*KIAA0182**	Hs.222171	+	−	+	+
EFNA5	Hs.37142	+	−	+	+
*ZNF600**	Hs.166312	+	−	+	+
TRIP10	Hs.445226	+	−	+	+
NOPE	Hs.20924	+	−	+	+
*CCNH**	Hs.514	+	−	+	+
LTA4H	Hs.81118	+	−	+	+
ADRBK2	Hs.445563	+	−	+	+
MOV10	Hs.512586	+	−	+	+
PGRMC1	Hs.90061	+	−	+	+
SDC3	Hs.158287	+	−	+	+
*ENSA**	Hs.511916	+	−	+	+
FLJ14768	Hs.129888	+	−	+	+
BAIAP1	Hs.169441	+	−	+	+
CGN	Hs.18376	+	−	+	+
*LOC84661**	Hs.402525	+	−	+	+
TP53	Hs.408312	+	−	+	+
DKFZP434A 1022	Hs.324335	+	−	+	+
PMPCA	Hs.75353	+	−	+	+
EIF3S8	Hs.192425	+	−	−	−
C6orf79	Hs.214043	+	−	−	−
*BRF1**	Hs.424484	+	−	+	+
*ZNF525**	Hs.352638	+	−	+	+
CNTNAP1	Hs.408730	+	−	+	+
MCM3	Hs.179565	+	−	+	+
RPS16	Hs.397609	+	−	+	+
ZNF566	Hs.528697	+	−	+	+
*ZNF217**	Hs.155040	+	−	+	+
*SFRS1**	Hs.68714	+	−	+	+
HSF2BP	Hs.406157	+	−	+	+
FLJ39739	Hs.523568	−	−	+	+
FAM11A	Hs.37106	+	−	+	+
C19orf7	Hs.119667	+	−	+	+
LOC169834	Hs.511892	+	−	−	−
OIP2	Hs.274170	+	−	+	+
ASB16	Hs.458471	+	−	+	+
*ZNF611**	Hs.446500	+	−	+	+
*PITX2**	Hs.92282	+	−	+	+
DGKI	Hs.242947	+	−	+	+
NMNAT2	Hs.158244	+	−	+	+
LOC152485	Hs.133916	+	−	−	−
*HOXC13**	Hs.118608	+	−	+	+
*HIST1H2BG**	Hs.68714	+	−	+	+
SIRT3	Hs.511950	+	−	+	+
MGA	Hs.435961	+	−	+	+
SCARB1	Hs.130981	+	−	+	+
NUP155	Hs.232255	+	−	−	−
*MLR2**	Hs.176120	+	−	+	+
*LOC400615**	Hs.405627	+	−	+	+
BRIP1	Hs.87507	+	−	+	+
DCC	Hs.172562	+	−	+	+
*CMAS**	Hs.311346	+	−	+	+
*IMAGE 3455200**	Hs.324844	+	−	+	+
KIAA0356	Hs.420584	+	−	−	−
COL1A2	Hs.232115	+	−	+	+
*CASP8AP2**	Hs.122843	+	−	+	+
RCP9	Hs.300684	+	−	−	−
TNPO2	Hs.278378	+	−	+	+
*D1S155E**	Hs.69855	+	−	+	+
LOC400713	Hs.528705	+	−	+	+
*Predicted gene**	NT_011109.956	+	−	+	N
Predicted gene	NT_004321.27	+	−	+	N
*Predicted gene**	NT_022517.1153	+	−	+	N

Table 2. *Continued*

Gene symbol	Unigene ID	ERE half-site[b]	ERE palindrome[c]	ER Binding site[d] Human	Mouse
*Predicted gene**	NT_008076.50	+	–	+	N
*Predicted gene**	NT_016354.964	+	–	+	N
*Predicted gene**	NT_010194.90	+	–	+	N
*Predicted gene**	NT_010194.30	+	–	+	N

[a]All promoter sequences are retrieved from 2 kb upstream to 1 kb downstream of the transcriptional start position. '+' indicates the existence of such element, '–' indicates no such element in this category, 'N' means no mouse counterparts for the human genes.
[b]The consensus of the half-site of ERE sequence (GGTCA).
[c]The consensus of the ERE palindrome (GGTCAnnnTGACC).
[d]The transcriptional factor of ER for both human and mouse was identified by our approach. Mouse orthologous counterpart was shown if available.

palindrome search, with only 0.1% conserved in orthologous pairs of 8124 promoters in OMGProm. Not surprisingly, at least one ERE half-site was found in the 5′ flanking region of almost all promoters ~3 kb uspstream of the TSSs. Although TRANSFAC_PWM accurately predicted ERα targets in Dataset 1 (100%) and Dataset 2 (97%), for Dataset 3, the model falsely predicted that 96% of the genes were ERα targets. Thus, TRANSFAC_PWM appears to suffer from an unacceptably high false-positive rate. Although our CART model correctly classified only 38 of 46 (83%) EREs in Dataset 1, corresponding to 35 of the experimentally confirmed ERα targets, the false-positive rate for the entire dataset was substantially reduced, with a predicative rate of 17%. We also used CART to classify the Dataset 3 into four modules and found that of 1696 promoters, 618 (36.4%) are in Module 1, 612 (36.1%) belong to Module 2, 161 (9.5%) are classified into Module 3 and 305 (18.0%) are in Module 4 (see Supplementary Table 2 for a list of classified genes).

Using ERα promoter microarrays to validate computational predictions

To validate the prediction results from CART, we performed another ChIP-on-chip experiment, this time using a small microarray panel of the aforementioned 70 ERα putative targets. The AcH3 antibody was used to assess binding of AcH3 components to these promoters and thus identify actively transcribed genes in ERα-positive MCF-7 and ERα-negative MDA-MB-231 cells. After treatment with E2, the number of 'active' promoters was greater ($P < 0.0001$) in MCF-7 (27/70 loci) compared with MDA-MB-231 cells (0/70 loci) Figure 2C and Table 2; an asterisk indicates active promoters). Our CART model correctly assigned 20 of these 27 loci to known genes, 2 loci to the same gene (MGC40455 and KIAA0182), and 6 loci to 7 pseudo genes. The 27 loci most likely represent genes up-regulated by ERα in MCF-7 cells, and the 43 'inactive' promoters presumably correspond to genes either down-regulated after E2 treatment or indirect target loci.

Identification of over-represented motifs prevalent in promoter regions of ERα targets

In order to identify novel motifs in ERα targets, we used the program MEME to search the promoter regions of −220 to +220 bp surrounding the ERE on the both strands for a dataset which consists of 35 experimentally confirmed ERα targets and 42 ERα target candidates from ChIP-on-chip. MEME uses position-dependent letter-probability matrices to represent motifs and describe the probability of each possible letter at each position in the pattern. After the 10 most significant motifs were found by this method, we then examined them by the TRANSFAC database and assigned the possible motifs corresponding to the binding site motifs in the database. A list of 10 consensus motifs, frequency of each motif occurring in the dataset, possible corresponding TFs are presented in Table 4. Of these motifs, Motif 6 was assigned to ERE half-site, Motif 2 was a Sp1 binding motif and Motif 5 was considered as a GC-rich box. We were also able to identify 2 novel motifs, Motifs 7 and 10, which did not match any known motif in the TRANSFAC database. It was not surprising to identify the ERE half-site and SP1 motifs, since most promoter sequences are GC-rich and include a consensus ERE half-site.

DISCUSSION

In this study, an integrated computational genomics approach was developed to identify ERα target promoters from ChIP-on-chip data. CART, a robust statistical method to select learning samples from two individual sets of loci, was used to differentiate ERα targets from non-ERα, housekeeping genes. The key phases of this approach were (i) the CART model construction phase, including a PWM built from a set of experimentally confirmed EREs; (ii) the model test phase, using a set of potential ERα targets from a genomic array; and (iii) the validation phase, using a small microarray panel of ERα targets to confirm the results in two different breast cancer cell lines, one positive for ERα signaling and the other negative. Although several studies have recently focused on either motif identification by ChIP-on-chip (33) or a computational approach to search consensus EREs (10), to our knowledge, this is the first study that combines a robust statistical model with a genomic microarray approach to systematically identify ERα target promoters.

The methodology of comparative genomics has recently been adopted by computational biologists to study transcriptional regulation of genes. However, the lack of publicly available, comprehensive databases for orthologous promoter sequences has hindered the application of comparative genomics to the transcriptional regulation field. Here, we have taken advantage of a unique orthologous database, OMGProm (18), and further tested it using CART. As indicated earlier, we defined a set of parameters for our ERE_PWM and then determined their cut-off values for CART model 1. Compared with the TRANSFAC_PWM, without using counterpart species, our ERE_PWM dramatically reduced the false-positive rate in the entire OMGProm database (from 96 to 29% in 8124 promoters). In addition, compared with ERE_PWM, our CART model further reduced false-positive rates from 96 to 21% for an entire OMGProm database. The results of the comparison between TRANSFAC_PWM and ERE_PWM further suggest that the source of the data used to build the PWM may play an important role in predictive rates. For example, binding sequences from more than four species (including chicken, human, mouse and rat) were used to construct

Table 3. Statistical summaries for three different datasets of ERα targets predicted from different approaches

Source of data	The number of EREs identified via				
	Perfect palindrome[a] (%)	Perfect half-site[b] (%)	TRANSFAC_PWM[c] (%)	ERE_PWM[d] (%)	CART[e] (%)
Literature search (46)	4 (9)	37 (80)	46 (100)	41 (89)	38 (83)
Conserved[f]	4 (9)	37 (80)			
ChIP-on-chip (63)	0 (0)	62 (98)	61 (97)	56 (89)	42 (67)
Conserved[f]	0 (0)	58 (92)			
OMGProm (8124)	19 (0.2)	8012 (99)	7777 (96)	2372 (29)	1696 (21)
Conserved[f]	8 (0.1)	7603 (94)			

[a]The consensus of the ERE half-site sequence (GGTCA).
[b]ERE palindrome consensus (GGTCAnnnTGACC).
[c]The putative ERE predicted using the PWM built from TRANSFAC database. The core score and PWM score cut-off values are default at >0.8 and 0.8, respectively.
[d]The putative ERE predicted using the PWM built from the experimentally verified ERE motifs. The cut-off values of human core score, human PWM score, mouse core score, mouse PWM score and sequence similarity score are >0.8, 0.8, 0.7, 0.7 and 0.6, respectively.
[e]The CART model developed in this study.
[f]The conservation of percent identity between orthologous pairs of human and mouse or human and rat is >60%.

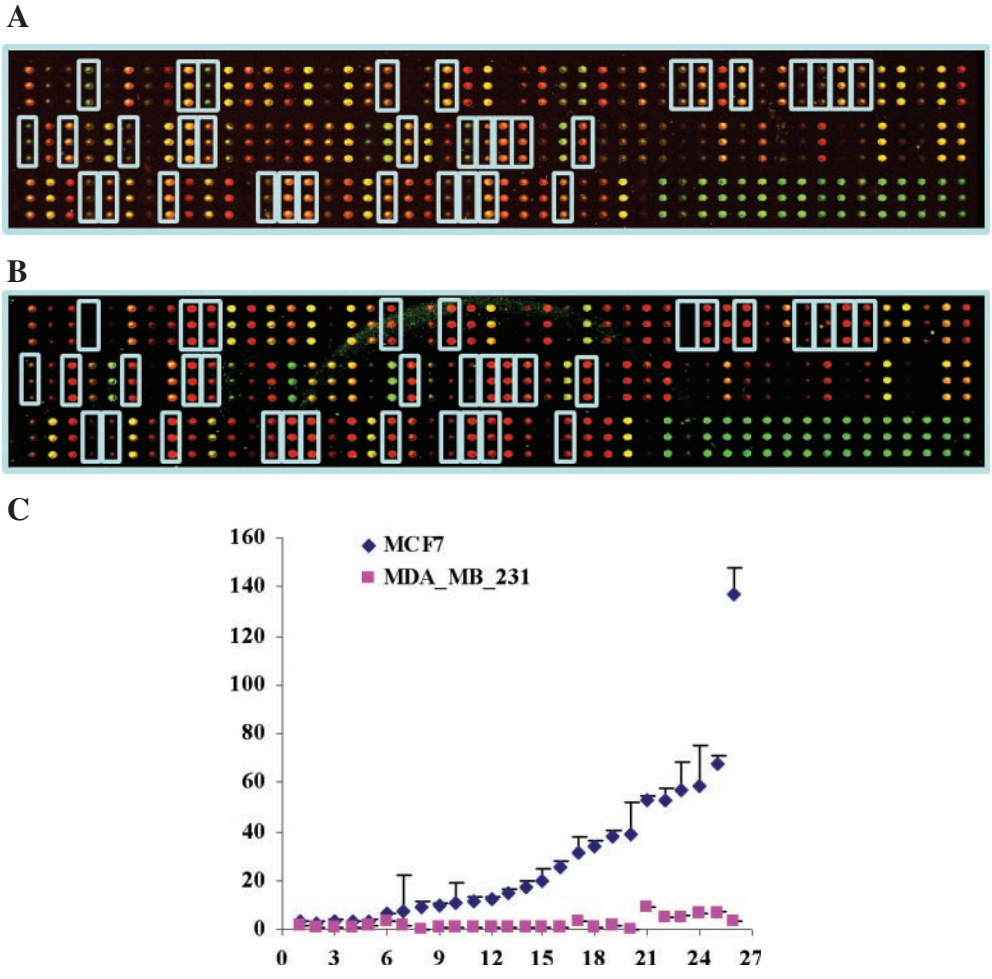

Figure 2. (**A**) AcH3 antibody (Upstate Cat. 06–942) was used to immunoprecipitate transcriptionally active chromatin components from the MDA-MB-231 (ER−) cell line. Each gene was printed three times on the array. (**B**) The same AcH3 antibody was used to immunoprecipitate the chromatin components from the MCF-7 (ER+) cell line. The loci that contain the active chromatin components in MCF-7, but not in the MDA-MB-231, cell are highlighted. (**C**) A plot of the Cy5/Cy3 ratios *versus* loci in both MCF7 and MDA-MB-231 cells. The ratios of MCF-7 (Cy5/Cy3)/MDA-MB-231(Cy5/Cy3) >15 suggest that the loci have active chromatin and are targets of ERα. The loci that contain the active chromatin components in MCF-7 but not MDA-MB-231 are highlighted.

TRANSFAC_PWM, whereas ERE_PWM was based on human, mouse and rat sequences only.

An advantage of CART is its ability to classify data with highly nonlinear structures, such as our ERα data. Although CART accurately predicted 96% of the experimentally confirmed ERα target genes, it performed poorly on the housekeeping gene data set. Housekeeping genes are constitutively expressed and play a role in most basic cellular functions.

Table 4. Over-represented motifs identified by MEME[a]

Motifs	Consensus	Frequency (%)	TF	E-value
1	GSGDGSMWGRGGSRGGRS	95	MZF1	2.2E−088
2	VSSYYYMCYCYCKCCHRS	100	SP1	6.9E−041
3	YCYRSSYYCCSGSYYC	100	AP2	5.3E−019
4	CCCYKCCYCHBSCYCSCB	100	GKLF	3.2E−089
5	CCCCGCCCCCCCCCCTCC	100	GC-signal	1.2E−030
6	KYCRCCAKGTTGGTCAGG	24	ERE half-site	1.9E−009
7	GRSCRSCDSCHBBYCC	61	novel	1.8E−002
8	SSSSGSGKYSGSKSSSGS	72	E2F	1.4E−016
9	KRSRGRGRSMSVGGCDG	66	ZF5	2.4E−007
10	YVMKKBBMCYSCYKYDSY	39	novel	2.6E−005

[a]The motifs identified by MEME program on a basis of a dataset which consists of 35 experimentally confirmed ERα targets and 42 ERα target candidates from ChIP-on-chip.

We and others routinely use housekeeping genes as internal controls in experimental assays. However, not all housekeeping genes have been confirmed experimentally as being non-ERα target genes, and any binding of ERα to these gene promoters could contribute to background noise and the misclassification rate observed in our model. To resolve this issue will require additional experimental investigation.

ChIP-on-chip provides strong *in vivo* evidence of direct binding of a specific protein complex to DNA (34). This technique differs markedly from the gene expression microarray, which has been used to investigate non-ERα mediated, E2-induced gene expression. For example, ER-independent regulation of fork genes by E2, through E2F, has been reported (35). After previously the analysis of ChIP-on-chip data in yeast, Kato *et al*. (36) proposed three types of interactions between a TF and its binding site on DNA: (i) direct binding (TF binds to specific binding motif); (ii) piggy-back binding (TF binds to another TF which is already bound to a specific motif); or (iii) cross-binding (TF binds its specific binding site on DNA but can interact with another TF bound to its specific motif). These three types of interactions have also been shown to apply to ERs, such as pS2 [directing binding (37), Interleukin-6 [piggy-back binding (38)] and Thymidylaste synthase [cross-binding (39)].

In the present study, we discovered four modules that can be used not only to classify ERα targets from non-ERα targets, but also to explore the possible combinatorial control of different TFs in different tissues. Based on many studies, there is slight doubt that module 3 (ERE+AP1) is one of the three most common indirect tethering models [(40) and references therein]; however, at this time, we lack the experimental results necessary to prove the other three modules proposed in this study. However, regarding the other modules, Orimo *et al*. (41) reported that vascular smooth muscle cells possess ERα and respond to estrogen, supporting the idea that estrogen may directly influence vascular cells through ERα. Speir *et al*. (42) also demonstrated the interaction and reciprocal interference of ER with p65, the NF-kappaB component, in ER-positive smooth muscle cells. Based on the discovery of those modules, we speculate that TFs may act together to build fully operational complexes on promoters, perhaps in a tissue specific manner. Unfortunately, due to the fact that the frequency of SP1 in both the training dataset and the control housekeeping gene dataset are very close (77.8 and 78.4%, respectively), our CART model failed to identify the combination of ERE and SP1. In contrast, the MEME was able to identify two novel motifs, ERE half-site and SP1 binding site. A comparison of the motifs identified by our model with the *ab initio* motifs discovery program, MEME, suggests that either method has its own advantages and disadvantages. MEME can detect novel motifs and our model is able to classify the targets by the motifs we identified. Combination of two methods will enable us to provide a powerful and functionally known set of motifs for the experimentally confirmed data and our ChIP-on-chip assay data.

Recently, Bourdeau *et al*. (10) identified 660 conserved EREs corresponding to 1% of total EREs in the flanking regions (−10 to +5 kb) of both the human and mouse genomes, by using a search criterion of ERE consensus palindrome sequence that allowed for a maximum of two mismatches. It was argued that a PWM-based approach would produce too many false predictions. Indeed, a search of the OMGProm database for conserved EREs based on TRANSFAC_PWM approach found 96% promoters as ER targets (Table 3), most of which probably are false predictions. In order to reduce the false-positive predictions and predict as many true ER targets as possible, we utilized the combinatorial association of ERE with other TF binding sites in the ER target promoters by CART analysis. Our CART model predicted 21% of OMGProm promoters as ERα target genes. The ERE_PWM we constructed is quite consistent with the matrices used in the study by Hansen and co-workers (11), who suggested that ER can cooperate with other TFs to bind to a half-site ERE or the consensus ERE can actually be more divergent and thus facilitate ER binding.

The important contributions of the present study to the study of transcriptional regulation networks are as follows: (i) we have developed a strategy or blueprint for the analysis of the regulation of transcription, not just a simple computational tool; (ii) the approach we have established and tested on the ChIP-on-chip microarray data, combined with the methodology of comparative genomics, provides deeper insight into the ChIP-on-chip data and contributes toward a better understanding of how the data from this new microarray technology can affect the accuracy of results; (iii) although our approach has been designed to analyze ERα target genes, it can easily be extended to the systematic study of other TFs. For instance, we have successfully used our CART model to analyze C/EBP, another important TF in breast cancer (data not shown). In conclusion, we believe that the modules discovered in this study and the ERTargetDB are novel integrated information resources for characterizing ER binding and studying transcriptional regulation of ER target genes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Twyla T. Pohar and Saranyan K. Palaniswamy for helpful discussions. V.X.J. is an Up On the Roof Postdoctoral Fellow at the Human Cancer Genetics

Program. The authors wish to thank Nicholas Berry for helping with data preparation. This work was supported in part by National Cancer Institute grant RO1 CA-69065 (T.H.-M.H.) and American Cancer Society Research Grant, Alaska Run for Women TBE 104125 (K.P.N.) and by funds from the Ohio State University Comprehensive Cancer Center—Arthur G. James Cancer Hospital and Richard J. Solove Research Institute (T.H.-M.H. and R.V.D.).

REFERENCES

1. Mangelsdorf,D.J., Thummel,C., Beato,M., Herrlich,P., Schutz,G., Umesono,K., Blumberg,B., Kastner,P., Mark,M., Chambon,P. and Evans,R.M. (1995) The nuclear receptor superfamily: the second decade. *Cell*, **83**, 835–839.
2. Kun,Y., How,L.C., Hoon,T.P., Bajic,V.B., Lam,T.S., Aggarwal,A., Sze,H.G., Bok,W.S., Yin,W.C. and Tan,P. (2003) Classifying the estrogen receptor status of breast cancers by expression profiles reveals a poor prognosis subpopulation exhibiting high expression of the ERBB2 receptor. *Hum. Mol. Genet*., **12**, 3245–3258.
3. Klinge,C.M. (2000) Estrogen receptor interaction with co-activators and co-repressors. *Steroids*, **65**, 227–251.
4. Abdelrahim,A., Samudio,I., Smith,R., III, Burghardt,R. and Safe,S. (2002) Small inhibitory RNA duplexes for Sp1 mRNA block basal and estrogen-induced gene expression and cell cycle progression in MCF-7 breast cancer cells. *J. Biol. Chem*., **277**, 28815–28822.
5. Gaub,M.P., Bellard,M., Scheuer,I., Chambon,P. and Sassone-Corsi,P. (1990) Activation of the ovalbumin gene by the estrogen receptor involves the fos-jun complex. *Cell*, **63**, 1267–1276.
6. Jakacka,M., Ito,M., Weiss,J., Chien,P.Y., Gehm,B.D. and Jameson,J.L. (2001) Estrogen receptor binding to DNA is not required for its activity through the nonclassical AP1 pathway. *J. Biol. Chem*., **276**, 13615–13621.
7. Klein-Hitpass,L., Schorpp,M., Wagner,U. and Ryffel,G.U. (1986) An estrogen-responsive element derived from the 5′ flanking region of Xenopus vitellogenin A2 gene functions in transfected human cells. *Cell*, **46**, 1053–1061.
8. Furlow,J.D., Murdoch,F.E. and Gorski,J. (1993) High affinity binding of the estrogen receptor to a DNA response element does not require homodimer formation or estrogen. *J. Biol. Chem*., **268**, 12519–12525.
9. Kim,J., Petz,L.N., Ziegler,Y.S., Wood,J.R., Potthoff,S.J. and Nardulli,A.M. (2000) Regulation of the estrogen-responsive pS2 gene in MCF-2 human cancer cells. *J. Steroid Biochem. Mol. Biol*., **74**, 157–168.
10. Bourdeau,V., Deschenes,J., Metivier,R., Nagai,Y., Nguyen,D., Bretschneider,N., Gannon,F., White,J.H. and Mader,S. (2004) Genome-wide identification of high-affinity estrogen response elements in human and mouse. *Mol. Endocrinol*., **18**, 1411–1427.
11. O'Lone,R., Frith,M.C., Karlsson,E.K. and Hansen,U. (2004) Genomic targets of nuclear estrogen receptors. *Mol. Endocrinol*., **18**, 1859–1875.
12. Wood,J.R., Greene,G.L. and Nardulli,A.M. (1998) Estrogen response elements function as allosteric modulators of estrogen receptor conformation. *Mol. Cell. Biol*., **18**, 1927–1934.
13. Wood,J.R., Likhite,V.S., Loven,M.A. and Nardulli,A.M. (2001) Allosteric modulation of estrogen receptor conformation by different estrogen response elements. *Mol. Endocrinol*., **15**, 1114–1126.
14. Loven,M.A., Likhite,V.S., Choi,I. and Nardulli,A.M. (2001) Estrogen response elements alter coactivator recruitment through allosteric modulation of estrogen receptor beta conformation. *J. Biol. Chem*., **276**, 45282–45288.
15. Schultz,J.R., Loven,M.A., Melvin,V.M., Edwards,D.P. and Nardulli,A.M. (2002) Differential modulation of DNA conformation by estrogen receptors alpha and beta. *J. Biol. Chem*., **277**, 8702–8707.
16. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.
17. Alkema,W.B., Johansson,O., Lagergren,J. and Wasserman,W.W. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res*., **32**, W195–W198.
18. Palaniswamy,S.K., Jin,V.X, Sun,H. and Davuluri,R.V. (2004) OMGProm: A Database of orthologous mammalian gene promoters. *Bioinformatics*, in press.
19. Leu,Y.-W., Yan,P.S., Fan,M., Jin,V.X., Liu,J.C., Curran,E.M., Welshons,W.V., Wei,S.H., Davuluri,R.V., Plass,C., Nephew,K.P. and Huang,T.H.-M. (2004) Loss of estrogen receptor signaling triggers epigenetic silencing of downstream targets in breast cancer. *Cancer Res*., **64**, 8184–8192.
20. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC®: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
21. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
22. Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) Classification and regression trees. Chapman & Hall, New York, NY.
23. Fan,M., Bigsby,R.M. and Nephew,K.P. (2003) The NEDD8 pathway is required for proteasome-mediated degradation of human estrogen receptor (ER)-alpha and essential for the antiproliferative activity of ICI 182,780 in ERalpha-positive breast cancer cells. *Mol. Endocrinol*., **17**, 356–365.
24. Yan,P.S., Chen,C.-M., Shi,H., Rahmatpanah,F., Wei,S.H., Caldwell,C.W. and Huang,T.H.-M. (2001) Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res*., **61**, 8375–8380.
25. Cross,S.H., Charlton,J.A., Nan,X. and Bird,A.P. (1994) Purification of CpG islands using a methylated DNA binding column. *Nature Genet*., **33**, 61–65.
26. Li,T., Vu,T.H., Ulaner,G.A., Yang,Y., Hu,J.F. and Hoffman,A.R. (2004) Activating and silencing histone modifications form independent allelic switch regions in the imprinted Gnas gene. *Hum. Mol. Genet*., **13**, 741–750.
27. Hsiao,L.-L., Dangond,F., Yoshida,T., Hong,R., Jensen,R.V., Misra,J., Dillon,W., Lee,K.F., Clark,K.E., Haverty,P. *et al*. (2001) A compendium of gene expression in normal human tissues reveals tissue-selective genes and distinct expreession patterns of housekeeping genes. *Physiol. Genomics*, **7**, 97–104.
28. Ringrose,L., Rehmsmeier,M., Dura,J.-M. and Paro,R. (2003) Genome-wide prediction of polycomb/trithorax response elements in *Drosophila melanogaster*. *Dev. Cell*, **5**, 759–771.
29. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
30. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res*., **31**, 23–27.
31. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78–94.
32. Davuluri,V.R., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet*., **29**, 412–417.
33. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipation microarray experiments. *Nat. Biotechnol*., **20**, 835–839.
34. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
35. Lobenhofer,E.K., Bennett,L., Cable,P.L., Li,L., Bushel,P.R. and Afshari,C.A. (2002) Regulation of DNA replication fork genes by 17beta-estradiol. *Mol. Endocrinol*., **16**, 1215–1229.
36. Kato,M., Hata,N., Banerjee,N., Futcher,B. and Zhang,M.Q. (2004) Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol*., **5**, R56.
37. Kim,J., Petz,L.N., Ziegler,Y.S., Wood,J.R., Potthoff,S.J. and Nardulli,A.M. (2000) Regulation of the estrogen-responsive pS2 gene in MCF-7 human breast cancer cells. *J. Steroid Biochem. Mol. Biol*., **74**, 157–168.
38. Galien,R. and Garcia,T. (1997) Estrogen receptor impairs interleukin-6 expression by preventing protein binding on the NF-kappaB site. *Nucleic Acids Res*., **25**, 2424–2429.

39. Xie,W., Duan,R., Chen,I., Samudio,I. and Safe,S. (2000) Transcriptional activation of thymidylate synthase by 17beta-estradiol in MCF-7 human breast cancer cells. *Endocrinology*, **141**, 2439–2449.
40. Kushner,P.J., Agard,D.A., Greene,G.L., Scanlan,T.S., Shiau,A.K., Uht,R.M. and Webb,P. (2000) Estrogen receptor pathways to AP-1. *J. Steroid Biochem. Mol. Biol.*, **74**, 311–317.
41. Orimo,A., Inoue,S., Ikegami,A., Hosoi,T., Akishita,M., Ouchi,Y., Muramatsu,M. and Orimo,H. (1993) Vascular smooth muscle cells as target for estrogen. *Biochem. Biophys. Res. Commun.*, **195**, 730–736.
42. Speir,E., Yu,Z.X., Takeda,K., Ferrans,V.J. and Cannon,R.O.III. (2000) Competition for p300 regulates transcription by estrogen receptors and nuclear factor-kappaB in human coronary smooth muscle cells. *Circ. Res.*, **87**, 1006–1111.