ERTargetDB: an integral information resource of transcription regulation of estrogen receptor target genes

V X Jin*, H Sun*, T T Pohar, S Liyanarachchi, S K Palaniswamy, T H-M Huang and R V Davuluri

Human Cancer Genetics Program, Comprehensive Cancer Center, Department of Molecular Virology, Immunology and Medical Genetics, Ohio State University, 420 W 12th Avenue, TMRF 524A, Columbus, Ohio 43210, USA

(Requests for offprints should be addressed to R V Davuluri; Email: davuluri-1@medctr.osu.edu)

*(V X Jin and H Sun contributed equally to the work)

Abstract

The estrogen receptor (ER) plays an important role in several physiologic functions of both the reproductive and non-reproductive systems. Malignancies of the ER have been associated with the development of cancers, including those of the prostate and breast. Hence it has become of significant importance to characterize the transcriptional regulation of ER target genes. We have created ERTargetDB in order to integrate the previously published ER target gene information that is available in various publications and databases. This information resource provides researchers with an easy access to ER target genes and the regulatory mechanisms in the corresponding promoters. The current version contains 40 genes with experimentally verified estrogen response elements (EREs), 32 experimentally verified ERE tethering sites, 40 genes identified by the chromatin immunoprecipitation microarray, 381 genes from gene expression microarray and 2948 genes from computational prediction. ERTargetDB provides an integral information resource for direct target genes of ERs for the endocrinology research community. It should prove useful in the investigation of gene regulation and aid the development of computational tools for the prediction of ER target genes.

Journal of Molecular Endocrinology (2005) 35, 225–230

Introduction

The estrogen receptors (ER) α and β are steroid hormone receptors that play important roles in the normal development of various organs, such as the brain, heart, bone, breast, uterus and prostate. Malignancy of the ER has been associated with endocrine diseases such as breast and prostate cancers, and the action of ERs is linked to the growth promotion and invasion of breast cancer cells (Pearce & Jordan 2004). ERs modulate the transcription of target genes when bound to ligand, and mediate target gene expression either through the classical pathway, in which ERs directly bind to estrogen response elements (EREs), or through the ERE-independent pathway, also known as indirect tethering, where ERs interact with other transcription factors such as activator protein-1 (AP-1), Sp1 and nerve factor-κβ (NF-κβ), which bind to the corresponding binding sites. The classical pathway is also denoted as the ER/ERE pathway, while the ERE-independent pathway is referred to as the ERE/X pathway, where X can be AP-1 (Cheung et al. 2005), Sp1 (Schultz et al. 2005) or NF-κβ (Qin et al. 2002). ERs bind with the highest affinity to the ERE consensus sequence, an inverted repeat of PuGGTC separated by three base pairs. However, very few mammalian promoters contain perfect ERE consensus sequences (Jin et al. 2004). The near-consensus EREs with one or two mismatches were shown to bind to ERs in vitro (Bourdeau et al. 2004) and have been widely accepted as alternative binding sites (Gruber et al. 2004).

Since ERs are key regulators of growth, differentiation and metabolism in multiple organs, the characterization of transcriptional regulation of their target genes is of significant importance. The complex biological effects mediated by ERα and ERβ involve cross-talk between many signaling pathways. Although the biomedical research community has produced a vast amount of regulatory information for ERs, it is difficult to obtain relevant data easily, as they are often dispersed throughout different places. In order to integrate the available information on ER target genes and improve upon the existing tools, we have created ERTargetDB. This database contains a combination of both experimentally defined and computationally predicted data, with a primary focus on the promoter architecture of ER direct target genes.
Construction and content

A portion of the direct ER target genes, along with the pathway information, has been collated manually from published literature. The data under the ER/ERE pathway were included in our database only if the EREs were located in the corresponding gene promoter sequences. Totally 48 of 52 experimentally verified ERE-binding sites from human, mouse and rat were found in the corresponding promoter sequences, while the remaining four were not located on either strand, which might be due to incorrect information reported or gaps in the corresponding genome (Supplementary Table 1 available at http://bioinformatics.med.ohio-state.edu/ERTargetDB/SupplementaryInfo.pdf). For example, the reported ERE motifs GGTCATGGTGACC in the human estrogen-responsive gene (elp) that encodes a RING finger protein (Inoue et al. 1993) and GGACACCATCTGTC in the rat luteinizing hormone β gene (Lhb) (Shupnik et al. 1989) were not found in the corresponding 7 kb promoter region. The 32 ER target genes that belong to the ER/X pathway of human, mouse and rat are reported to use the ‘tethering,’ mechanism with AP-1, Sp1 or NF-κB by individual laboratories (Supplementary Table 2 available at http://bioinformatics.med.ohio-state.edu/ERTargetDB/SupplementaryInfo.pdf). In order to ensure inclusion of recently published data, we will continue to perform a comprehensive literature review on a periodic basis.

We reliably recognized 40 target genes via our Chromatin immunoprecipitation microarray (ChIP)-on-chip assay (Leu et al. 2004) and 381 target genes that were experimentally proven to be up- or down-regulated by estrogen via gene expression microarray; the conditions of this experiment are provided as a user’s query option. We subsequently identified putative EREs by our integrated bioinformatics approach (Jin et al. 2004). The data predicted from our program ERTarget is performed through a genome-wide dataset housed in OMGProm (Palaniswamy et al. 2005), which is publicly available at http://bioinformatics.med.ohio-state.edu/OMGProm. Since most ER target genes are associated with CpG islands, all of the promoters in the database are classified into CpG-related or non-CpG-related categories. The current version of ERTargetDB contains information about three organisms: human, mouse and rat. A summary of the database is provided in Table 1.

Interface and visualization

ERTargetDB uses a native XML database, mobius (Langella et al. 2004) http://projectmobius.osu.edu, as the backend database server to store the XML files, presented by promoter annotation markup language (PAML) (Sun & Davuluri 2004a), which was created as an XML tool to facilitate the data integration process and present the ER gene annotation information for visualization. The elements defined in PAML provide a structured document representation of ER target gene annotation data that can be exchanged over the internet as universally understandable, self-describing documents. The main advantage of PAML is that data could be integrated in context within a single document, thereby making data immediately meaningful to both biologists and the computer. The elements defined in PAML with the support of the genome data visualization toolkit (Sun & Davuluri 2004b) provide a flexible way to generate different forms of images in order to fit different levels of visualization of ER target genes and their promoter annotations. The detailed information and XML DTD (Document Type Definitions) of PAML can be obtained from the website http://bioinformatics.med.ohio-state.edu/PAML.

JBoss (http://www.jboss.org) is used as the HTTP application server. It runs on Red Hat Linux Enterprise edition 9.0 (Redhat, Raleigh, NC, USA). All of the analysis for ERTargetDB was written in Perl.

Users can access the data from ERTargetDB in several ways. The search engine provides an easy way

<table>
<thead>
<tr>
<th>Table 1 A summary of current entries in ERTargetDB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
</tr>
<tr>
<td>Species</td>
</tr>
<tr>
<td>Human</td>
</tr>
<tr>
<td>Mouse</td>
</tr>
<tr>
<td>Rat</td>
</tr>
</tbody>
</table>

*a*Data collected from experimentally verified ERE-binding sites by an individual laboratory.

*b*Data collected from experimentally verified indirect tethering binding sites by an individual laboratory.

*c*Data identified from ChIP-on-chip technology developed by our laboratory.

*d*Data identified from gene expression microarray experiments.

*e*Data source from the computational prediction developed by our laboratory.
for users to locate a gene of interest by searching with gene symbol, gene identification (ID), UniGene ID or GenBank accession ID. Users may also obtain a list of ER target genes with composite queries such as species, experimental type, cell lines, gene regulatory pathways and transcription factors involved in the ER target gene regulatory pathway.

Each search result specifies whether the gene is a target of ERα or ERβ, species, the source, which includes literature, ChIP-on-chip, microarray or predicted. We also provide orthologous gene information and, if available, detailed gene expression information for the genes.

Regulatory information at the transcriptional level in textual and graphical forms for each gene is also provided as a useful feature for the user. The visual module provides depiction of the promoter with corresponding Transcription Start Site (TSS), ERE elements, other Transcription Factors (TFs) relative to the TSS and CpG island information. The textual data provide information in static form, which includes TF name, binding site positions and sequence and respective binding site reference, with links to PUBMED. In addition, each target gene is presented with the corresponding descriptive functional role in cell types or tissues (Fig. 1, snapshot).

### Discussion

ERTargetDB contains promoter annotations, including combinatorial associated cis-regulatory elements, within 220 bases around ERE determined by a non-parametric decision tree model that discriminates ER targets from non-targets, implemented by our recently developed program, ERTarget (Jin et al. 2004), and CpG island information for each promoter, which is provided as an additional track. The database specifies the transcriptional pathway for each target gene, and provides the promoter sequence of a maximum length of 3 kb for each gene. Furthermore, ERTargetDB also includes target gene information obtained from high-throughput experiments, such as those generated from gene expression and ChIP-on-chip microarrays. The target genes that were identified using gene expression microarrays (Frasor et al. 2003) usually contain both direct and downstream indirect target genes. Hence, we only include genes that contain putative EREs, predicted by the ERTarget program, in the corresponding promoters. The target genes identified through the ChIP-on-chip technology (Ren et al. 2000) are more reliable and mostly contain the direct target genes. The current version of ERTargetDB contains 40 genes of the ER/ERE pathway with experimentally verified ERE direct binding sites and 32 genes of the ER/X pathway with experimentally verified ERE tethering sites (AP-1, Sp1 or NF-κB), 40 genes identified via ChIP-on-chip, 381 genes from gene expression microarray (Table 2) and 2948 genes from computational prediction. To our knowledge, this is the only publicly available database that focuses on ER target genes, with the exception of the estrogen responsive genes database (ERGDB) (Tang et al. 2004).

There are several distinguishing features of ERTargetDB that improve upon the current publicly available database, ERGDB. ERTargetDB is exclusively focused on ER direct target genes which are

### Table 2 Current statistics of the ER target genes computationally predicted by scanning the promoters of genes identified in ChIP-on-chip and Microarray experiments

<table>
<thead>
<tr>
<th>Data source</th>
<th>Publication</th>
<th>Cell line</th>
<th>No. of target genes identified from the analysis of array data</th>
<th>No. of target genes predicted by our integrated computational approach (Jin et al. 2004) and included in ERTargetDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-on-chip</td>
<td>Leu et al. (2004)</td>
<td>Human MCF-7</td>
<td>70</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Jelinsky et al. (2003)</td>
<td>Mouse kidney</td>
<td>29</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Islamov et al. (2002)</td>
<td>Mouse lumbar spinal cord</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Charpentier et al. (2000)</td>
<td>Human MCF-7</td>
<td>41</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Frasor et al. (2003)</td>
<td>Human MCF-7</td>
<td>156</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>Inoue et al. (2002)</td>
<td>Human MCF-7</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Microarray</td>
<td>Soulez &amp; Parker (2001)</td>
<td>Human ZR75-1</td>
<td>48</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Kian Tee et al. (2004)</td>
<td>Human U2OS osteosarcoma</td>
<td>34</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Monroe et al. (2003)</td>
<td>Human U2OS osteosarcoma</td>
<td>129</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Lin et al. (2004)</td>
<td>Human T-47D</td>
<td>89</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Stossi et al. (2004)</td>
<td>Human U2OS osteosarcoma</td>
<td>52</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>DeNardo et al. (2005)</td>
<td>Human MCF-7</td>
<td>19</td>
<td>15</td>
</tr>
</tbody>
</table>
experimentally confirmed, with a reported motif of ERE and relative position to the TSS, confirmed by electrophoretic mobility shift assay or ChIP assay. The microarray data are considered evidence for ER direct target genes only if putative EREs that are conserved between human and mouse are present in the corresponding promoters. Furthermore, we specify the pathway for each target gene, i.e., classical or ERE independent, to adequately describe the regulatory mechanism. For target genes that are reported to have two different pathways by separate laboratories, we include both pathways in our database with reference to Figure 1.

**Figure 1** Screenshot of two promoter records for human gene BCL-2.
the publications. For example, cathepsin D is reported to be involved with a classical pathway ERE/ER in Augereau et al. (1994) and an ERE/Sp1 pathway in Krishnan et al. (1994). Accordingly, we have included both pathways with corresponding references. In addition, ERTargetDB includes significantly more genes than does ERGDB (Fig. 2), likely due to inclusion of different types of sources for datasets. Only 243 genes overlapped between both databases due to the different focus of each database. ERTargetDB only includes direct ER target genes with putative EREs, which are predicted by our in-house program ERTarget (Jin et al. 2004). Unlike ERTargetDB, ERGDB also includes downstream regulated genes, therefore many of the genes fail to contain a putative ERE. An additional, distinct feature of ERTargetDB is the specification of detailed promoter information and with annotation of, not only the putative ERE, but also other important transcriptional regulation elements around the ERE region such as AP-1, TATA and so on. This information would greatly assist experimental biologists with the design of primers in an easier and more efficient fashion. In summary, the host of data that are provided by ERTargetDB will, hopefully, contribute toward the progression of the ongoing investigation of ER and the implications of ER aberrances.

As part of the Integrative Cancer Biology Program of the National Cancer Institute (von Eschenbach 2004), we are developing an integrative computational and high-throughput experimental approach to characterize the epigenetic mechanisms involved in the target genes of the ER pathway. ERTargetDB integrates information from the ongoing ChIP-on-chip experiments in our laboratory (Jin et al. 2004, Leu et al. 2004) and promoter sequence conservation from the OMGProm database (Palaniswamy et al. 2005). With this approach, we are able to identify a set of very reliable ERα targets by both ChIP-on-chip and systematic computational prediction. The identified putative EREs by our computational approach were further verified by ERα promoter microarray (Jin et al. 2004). We will add methylation and acetylation patterns across the target gene promoters in future versions of the database.

### Available

ERTargetDB may be accessed from http://bioinformatics.med.ohio-state.edu/ERTargetDB.

### Supplementary information

Additional information is available at http://bioinformatics.med.ohio-state.edu / ERTargetDB / SupplementaryInfo.pdf.

### Acknowledgements

We are grateful to Greg A C Singer and Francisco J Agosto-Pérez for their helpful discussion. This work was supported in part by National Cancer Institute grant P50 CA-11300 (T H-M H and R V D) and by funds from the Ohio State University Comprehensive Cancer Center (OSU-CCC)–Arthur G James Cancer Hospital and Richard J Solove Research Institute (T H-M H and R V D). V X J was supported in part by an Up on the Roof Postdoctoral Fellowship of the OSU-CCC. We declare that there is no conflict of interest that would prejudice the impartiality of this scientific work.

### References


Sun H & Davuluri RV 2004 Gene regulatory region annotation and visualization – Java and XML based software tools. In Systems Biology: Genome Approaches to Transcriptional Regulation, p 100, 4–7 March Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA.


Received 6 June 2005
Accepted 1 July 2005
Made available online as an Accepted Preprint 21 July 2005