Identification of Potential E2F Target Genes Through cis-regulatory Modules Derived From Chromatin Immunoprecipitation Microarray Data

Wen-Shyong Tzou1,2*

1Institute of Bioscience and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan
2Center for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, Taiwan

Received: August 17, 2010 Revised: September 4, 2010 Accepted: September 7, 2010

In the postgenome era, employment of high-throughput data via the integrated use of resources from various domains will lead to the generation of new knowledge and testable hypothesis. In this bioinformatic research, we utilized published chromatin immunoprecipitation microarray (ChIP-chip) results for the promoter binding by E2F1 or E2F4 proteins observed in the primary human WI-38 cells to infer the potential transcription factor binding sites (TFBS). We have compiled “gene vs. motif” and “motif vs. gene” tables from more than 2,700,000 computational predicted transcriptional regulatory motifs representing the regulatory potential for 230 transcription factors families within the proximal promoter sequence (1,200 nucleotides) of human genome. From this approach, for the first time, the transcription of 23 genes is predicted to be under the control of a cis-regulatory module containing four TFBS motifs (CREB, E2F, NF-Y and Nrf-1).

Key Words: ChIP-chip; cis-regulatory module; E2F; transcription factor binding sites

Introduction

Several tools had been developed to search for the overrepresented motifs and cis-regulatory elements and modules from the promoter sequences of a gene set; however, it is also interesting to search for the genes of which the promoters contained a precomputed cis-regulatory module.1,2 We have compiled a “gene vs. motif” table using a scoring system adopted previously.3

Materials and Methods

Briefly, more than 2,700,000 computational predicted transcriptional regulatory motifs representing the regulatory potential for 230 transcription factors families (TRANSFAC 10.2) within the proximal promoter sequence (1,200 nucleotides) of human genome (NCBI36, 22,254 genes) are prepared. The nucleotide sequences were evaluated for its similarity to each transcription factor binding motif (position weight matrix) using a log-likelihood ratio score with a sixth-order Markov background model. A statistical method is employed to identify the significantly overrepresented transcription factor binding site motifs in the promoter sequences of a gene set and the pairs of transcription factor binding site motifs showing a significant co-occurrence rate.3 This method can suffice to answer the question of what are overrepresented transcription factor binding site motifs of the promoter sequences of a given...
Table 1  The predicted 27 genes of which the gene transcription is under the control of a cis-regulatory module containing four TFBS motifs (CREB, E2F, NF-Y and Nrf-1)\textsuperscript{4,11–25}

<table>
<thead>
<tr>
<th>ENSG ID</th>
<th>p</th>
<th>Gene ID</th>
<th>Official symbol</th>
<th>Official full name</th>
<th>Pubmed</th>
<th>Description</th>
<th>Evidence level (1–3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000205659</td>
<td>$7.4 \times 10^{-9}$</td>
<td>91750</td>
<td>LIN52</td>
<td>lin-52 homolog (C. elegans)</td>
<td>12</td>
<td>Fumarylacetoacetate hydrolase is responsible for the conversion of fumarylacetoacetate to fumarate and acetoacetate. The accumulation of fumarylacetoacetate induces spindle disturbances and segregational defects, leading to mitotic abnormality and genomic instability.</td>
<td>1</td>
</tr>
<tr>
<td>ENSG00000180185</td>
<td>$9.6 \times 10^{-9}$</td>
<td>81889</td>
<td>FAHD1</td>
<td>Fumarylacetoacetate 1 hydrolase domain containing</td>
<td>12</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000184207</td>
<td>$2.0 \times 10^{-8}$</td>
<td>283871</td>
<td>PGP</td>
<td>Phosphoglycolate phosphatase protein phosphatase 1, regulatory (inhibitor) subunit 10</td>
<td>13</td>
<td>PNUTS is a nuclear regulatory subunit of protein phosphatase 1. Targeting of PNUTS to reforming nuclei in telophase is a key signaling event promoting chromatin decondensation.</td>
<td>1</td>
</tr>
<tr>
<td>ENSG00000090520</td>
<td>$6.8 \times 10^{-8}$</td>
<td>51726</td>
<td>DNAJB11</td>
<td>DnaJ (Hsp40) homolog, subfamily B, member 11</td>
<td>15</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>ENSG00000165280</td>
<td>$7.1 \times 10^{-8}$</td>
<td>2189</td>
<td>FANCG, XRCC9</td>
<td>Fanconi anemia, complementation group G</td>
<td>14</td>
<td>XRCC9 gene product can correct the spontaneous chromosomal aberrations, a candidate of tumor suppressor gene that may have a postreplication repair or a cell cycle checkpoint function.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000090316</td>
<td>$7.2 \times 10^{-8}$</td>
<td>10296</td>
<td>MAEA, EMP</td>
<td>Macrophage erythroblast attacher</td>
<td>15</td>
<td>Emp migrates from intranuclear sites in interphase to regions associated with mitotic spindle poles and contractile ring during cell division.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000204568</td>
<td>$7.6 \times 10^{-8}$</td>
<td>28973</td>
<td>MRPS18B, MRP-518-2</td>
<td>Mitochondrial ribosomal protein S18B</td>
<td>16</td>
<td>EBV-encoded nuclear antigen 6 (EBNA-6) binds to MRPS18B and targets it to the nucleus. The binding will raise the level of free E2F1 and promote the S-phase entry.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000170315</td>
<td>$9.1 \times 10^{-8}$</td>
<td>7314</td>
<td>UBB</td>
<td>Ubiquitin B</td>
<td>17</td>
<td>Ubiquitin is involved in the regulation of onset of mitosis of cell cycle through the ubiquitin-protein ligase anaphase-promoting complex (APC) and Skp1-Rbx1-Cul1-F-box (SCF).</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000204308</td>
<td>$9.2 \times 10^{-8}$</td>
<td>6048</td>
<td>RNF5</td>
<td>Ring finger protein 5</td>
<td>18</td>
<td>RNF5 may regulate the cytokinesis by targeting ubiquitination and altered localization of paxillin, a focal adhesion-associated protein.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000180879</td>
<td>$1.1 \times 10^{-7}$</td>
<td>6748</td>
<td>SSR4, TRAPD</td>
<td>Signal sequence receptor, delta (translocon-associated protein delta)</td>
<td>18</td>
<td>TRAPD is upregulated in ultraviolet-induced melanoma tumors of advanced stages.</td>
<td>2</td>
</tr>
<tr>
<td>ENSG00000185359</td>
<td>$1.2 \times 10^{-7}$</td>
<td>9146</td>
<td>HGS, HRS</td>
<td>Hepatocyte growth factor-regulated tyrosine kinase substrate</td>
<td>19</td>
<td>Hrs associates with signal-transducing adaptor molecule (STAM) to suppress DNA synthesis or neurofibromin 2 (merlin) to inhibit cell growth and motility.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000115687</td>
<td>$1.3 \times 10^{-7}$</td>
<td>23178</td>
<td>PASK</td>
<td>PAS domain containing serine/threonine kinase</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

(Contd)
<table>
<thead>
<tr>
<th>ENSG ID</th>
<th>$p$</th>
<th>Gene ID</th>
<th>Official symbol</th>
<th>Official full name</th>
<th>Pubmed</th>
<th>Description</th>
<th>Evidence level (1−3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000125398</td>
<td>$1.3 \times 10^{-7}$</td>
<td>6662</td>
<td>SOX9</td>
<td>SRY (sex determining region Y)-box 9</td>
<td>20</td>
<td>SOX9 regulates the expression of p21&lt;sup&gt;cip1&lt;/sup&gt;, ERK1 and N-cadherin, altering the rate of cell cycle progression.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000094916</td>
<td>$1.9 \times 10^{-7}$</td>
<td>23468</td>
<td>CBX5, HP1, HP1A</td>
<td>Chromobox homolog 5 (HP1 alpha homolog, Drosophila)</td>
<td>4,21</td>
<td>HP1, a member of the heterochromatin protein family, is associated with centromere through the interaction with kinetochore. HP1 is involved in the segregation of sister kinetochores and rescues of the perturbed heterochromatin during mitosis.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000159199</td>
<td>$2.0 \times 10^{-7}$</td>
<td>516</td>
<td>ATP5G1</td>
<td>ATP synthase, H&lt;sup&gt;+&lt;/sup&gt;-transporting, mitochondrial F&lt;sub&gt;0&lt;/sub&gt; complex, subunit C1 (subunit 9)</td>
<td>22</td>
<td>hnRNPLL can regulate the alternative splicing of several genes including CD45, a transmembrane phosphatase marking the transition from naïve to activated T cells. hnRNPLL may regulate the cell differentiation and proliferation of hematopoietic cell.</td>
<td>1</td>
</tr>
<tr>
<td>ENSG00000185065</td>
<td>$3.3 \times 10^{-7}$</td>
<td>128977</td>
<td>C22orf39</td>
<td>Chromosome 22 open reading frame 39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000104824</td>
<td>$3.3 \times 10^{-7}$</td>
<td>3191</td>
<td>HNRNPL</td>
<td>hnRNPLL</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000077458</td>
<td>$3.3 \times 10^{-7}$</td>
<td>143684</td>
<td>FAM76B</td>
<td>Family with sequence similarity 76, member B</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000158373</td>
<td>$3.3 \times 10^{-7}$</td>
<td>3017</td>
<td>HIST1H2BD</td>
<td>Histone cluster 1, H2bd</td>
<td>4,10</td>
<td>Histone is required for assembly of DNA into nucleosome. E2F activates the gene expression of H2A in early S phase.</td>
<td>2</td>
</tr>
<tr>
<td>ENSG00000164040</td>
<td>$3.6 \times 10^{-7}$</td>
<td>10424</td>
<td>PGRMC2</td>
<td>Progesterone receptor membrane component 2</td>
<td>23</td>
<td>High frequency of copy number loss in PGRMC2 is associated with carcinogenesis of endocervical adenocarcinoma of uterus and metastasis.</td>
<td></td>
</tr>
<tr>
<td>ENSG00000197903</td>
<td>$3.7 \times 10^{-7}$</td>
<td>85236</td>
<td>HIST1H2BK</td>
<td>Histone cluster 1, H2bk</td>
<td>11</td>
<td>Histone is required for assembly of DNA into nucleosome. E2F activates the gene expression of H2A in early S phase.</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000181218</td>
<td>$4.1 \times 10^{-7}$</td>
<td>92815</td>
<td>HIST3H2A</td>
<td>Histone cluster 3, H2a</td>
<td></td>
<td>Vaccinia-related kinase 3 (VRK3) is a negative feedback regulator of ERK activity by binding to mitogen-activated protein kinase phosphatases (MKPs).</td>
<td>3</td>
</tr>
<tr>
<td>ENSG00000155287</td>
<td>$4.3 \times 10^{-7}$</td>
<td>81894</td>
<td>SLC25A28</td>
<td>Solute carrier family 25, member 28</td>
<td>25</td>
<td>SLTM modulates the number of pre-G1 phase cells, inhibits gene expression and induces apoptosis.</td>
<td></td>
</tr>
<tr>
<td>ENSG00000105053</td>
<td>$4.7 \times 10^{-7}$</td>
<td>51231</td>
<td>VRK3</td>
<td>Vaccinia related kinase 3</td>
<td>10,24</td>
<td>Vaccinia-related kinase 3 (VRK3) is a negative feedback regulator of ERK activity by binding to mitogen-activated protein kinase phosphatases (MKPs).</td>
<td></td>
</tr>
<tr>
<td>ENSG00000137776</td>
<td>$4.9 \times 10^{-7}$</td>
<td>79811</td>
<td>SLTM</td>
<td>SAFB-like, transcription modulator</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TFBS: transcription factor binding site; PNUTS: protein phosphatase 1 nuclear targeting subunit; hnRNPLL: heterogeneous ribonucleoprotein L-like; SLTM: SAF (scaffold attachment factor)-like transcription modulator.
gene sets. Importantly, we move further by addressing the question of what are the genes which the promoter sequence contains the identified over-represented transcription factor binding sites. For achieving this, we prepared the “motif vs. gene” table, demonstrating the potential gene targets of each transcription factor binding motif.

Results

A total of 115 genes bound by either E2F1 or E2F4 observed in the primary human WI-38 cells through ChIP-chip (chromatin immunoprecipitation micro-array) experiment were used as an example. First, we used the algorithm mentioned above (“gene vs. motif” table) to reveal the overrepresented transcription factor binding site motifs in the proximal promoter sequence of the gene sets (115 genes). Four motifs forming a cis-regulatory module (CREB, E2F, NF-Y and Nrf-1) are found to be enriched ($p \leq 0.001$ for the individual motif; $p \leq 0.03$ for the co-occurring motifs). CREB, E2F, NF-Y and Nrf-1 had been observed to be involved in the regulation of cell cycle and cell proliferation and our computation result is similar to those of the previous study.3

We next sought to address the question of what are the genes of which the promoter sequences contain the identified transcription factor binding site motifs, CREB, E2F, NF-Y and Nrf-1. We employed the “motif vs. gene” table and revealed 27 overrepresented genes ($p \leq 5 \times 10^{-7}$, Table 1), of which the promoter sequences contain the identified transcription factor binding site motifs, CREB, E2F, NF-Y and Nrf-1. It is interesting to note that only one gene (ENSG00000094916) appeared in the original gene sets (115 genes). In order to confirm if the promoter sequences of these 27 genes still harbored the identified transcription factor binding site motifs (CREB, E2F, NF-Y and Nrf-1), we investigated the promoter sequence of 27 genes. We found the same four transcription factor binding site motifs ($p \leq 0.001$) are found enriched in the promoter sequence of 27 genes. Interestingly, if we only focused on the regions in the conserved promoter sequences obtained by comparing human and mouse genomes (ECRbase), the four transcription factor binding site motifs (CREB, E2F, NF-Y and Nrf-1) are also enriched in the promoter sequence of 27 genes.

In order to discover the possible link between transcription factor E2F and 27 genes, we searched in the National Center for Biotechnological Information PubMed (http://www.ncbi.nlm.nih.gov/PubMed) for publications that demonstrated the involvement of the gene product in cell cycle control that is closely related to the function of E2F family. The search results can be categorized into three groups running from 1 to 3 with increasing confidence (Table 1). Fourteen out of the 27 (52%) genes fall into Category 3 with experimental evidence demonstrating the involvement of the gene product or its interacting partner on the cell cycle control. In this category, the direct interaction of the distal promoter sequence of four genes with E2F has been observed (ENSG00000094916, ENSG00000197903, ENSG00000181218, ENSG00000105053) (Table 1). Two out of the 27 genes (category 2, 7%) were found to be involved in the maintenance of genome stability during cell cycle. We cannot find the information for the relationship between the function of eleven genes (category 1, 41%) and E2F.

Discussion

E2F has the most abundant binding sites (54 TFBS) in the promoter regions of the 27 identified genes, while CREB and NF-Y have less binding sites (Table 2). TFBS in the region 300 bps upstream the transcription start sites occupies 50% of the whole TFBS (Table 2), indicating the possible distance effect.

Analyzing the spatial arrangement prompts four most abundant linked TFBS (the direction is from upstream to downstream): E2F-NF-Y (11 genes), NF-Y-E2F (11 genes), CREB-E2F (10 genes) and CREB-NF-Y (10 genes). However, in this study, we cannot predict that those two transcription factors interact with each other directly or indirectly.

Table 2: Contribution and localization of four transcription factor binding site (TFBS) in the promoter regions of 27 identified genes

<table>
<thead>
<tr>
<th>CREB</th>
<th>E2F</th>
<th>NF-Y</th>
<th>NRF-1</th>
<th>No. of TFBS per region</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1200 to -900bp</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>-900 to -600bp</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>-600 to -300bp</td>
<td>10</td>
<td>19</td>
<td>18</td>
<td>54</td>
</tr>
<tr>
<td>-300 to -1bp</td>
<td>23</td>
<td>21</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>No. of TFBS for each TF</td>
<td>41</td>
<td>54</td>
<td>47</td>
<td>20</td>
</tr>
</tbody>
</table>

CREB: cAMP response element-binding; NF-Y: nuclear factor Y; NRF-1: nuclear respiratory factor 1; bp: base pair.
This observation warrants further experimental design for testing if the expression of these 23 genes is controlled by CREB, E2F, NF-Y and Nrf-1.

References