Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences

Y. V. Sun1,3, D. R. Boverhof1,3,4, L. D. Burgoon1,2,4, M. R. Fielden1,3,4 and T. R. Zacharewski1,3,4,*

1Department of Biochemistry and Molecular Biology, 2Department of Pharmacology and Toxicology, 3National Food Safety and Toxicology Center and 4Center for Integrative Toxicology, Michigan State University, East Lansing, MI 48824, USA

Received May 18, 2004; Revised June 24, 2004; Accepted August 3, 2004

ABSTRACT

Comparative approaches were used to identify human, mouse and rat dioxin response elements (DREs) in genomic sequences unambiguously assigned to a nucleotide RefSeq accession number. A total of 13 bona fide DREs, all including the substitution intolerant core sequence (GCGTG) and adjacent variable sequences, were used to establish a position weight matrix and a matrix similarity (MS) score threshold to rank identified DREs. DREs with MS scores above the threshold were disproportionately distributed in close proximity to the transcription start site in all three species. Gene expression assays in hepatic mouse tissue confirmed the responsiveness of 192 genes possessing a putative DRE. Previously identified functional DREs in well-characterized AhR-regulated genes including Cyp1a1 and Cyp1b1 were corroborated. Putative DREs were identified in 48 out of 2437 human–mouse–rat orthologous genes between −1500 and the transcriptional start site, of which 19 of these genes possessed positionally conserved DREs as determined by multiple sequence alignment. Seven of these nineteen genes exhibited 2,3,7,8-tetrachlorodibenzo-p-dioxin-mediated regulation, although there were significant discrepancies between in vivo and in vitro results. Interestingly, of the mouse–rat orthologous genes with a DRE between −1500 and +1500, only 37% had an equivalent human ortholog. These results suggest that AhR-mediated gene expression may not be well conserved across species, which could have significant implications in human risk assessment.

INTRODUCTION

Novel computational approaches are being developed to annotate genome sequence data in order to predict gene structure, function and higher order biological control processes, such as gene regulation (1). Deciphering gene regulation at the transcriptional level through the computational identification of response elements is a valuable complement to empirical approaches seeking to develop biochemical networks (2–5). Several software tools have been developed to assist with the identification of potential regulatory elements and to predict their functionality using sequence, structure, context and comparative-based methods (1,6). Sequence-based searches are the most common means of identifying putative regulatory elements in silico due to the availability of software (1) and databases of transcription factor binding sites (7). This approach can complement experimental genome-wide analysis of gene expression and can be used to predict gene regulatory networks (8–10).

The availability of the human, mouse and rat genomes has provided unprecedented opportunities in understanding mammalian evolution, elucidating the etiology of human disease and facilitating drug development. Comparative analysis of genomic sequence data is a powerful tool for identifying functional non-coding sequences, such as gene regulatory elements, which tend to be conserved through evolution for common responses (11–14). Putative functional regulatory elements can then be identified by searching for conserved DNA sequence motifs in orthologous genes across multiple species, such as human, mouse and rat. However, the value of computational approaches has been questioned due to supposed high false-positive rates, and therefore must be verified using empirical approaches such as genome-wide gene expression and chromatin immunoprecipitation assays.

2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) is a persistent environmental contaminant that elicits a broad spectrum of aryl hydrocarbon receptor (AhR)-mediated biochemical and toxic effects. These effects are well correlated with the ability of TCDD and related compounds to bind to the AhR (15) and are manifested by inappropriate AhR-mediated regulation of gene expression. This is supported by studies demonstrating that mice with low-affinity AhR alleles are less susceptible to the effects of TCDD (16), and AhR-null mice are resistant to the prototypical toxicities elicited by TCDD and related ligands (17,18). The AhR is a basic helix–loop–helix PAS (bHLH–PAS) protein that acts as a ligand-dependent DNA-binding transcription factor (19,20).
Binding of ligand to the AhR triggers nuclear translocation and subsequent dissociation of the AhR from cytosolic components. Once in the nucleus, the AhR heterodimerizes with its bHLH–PAS binding partner, the AhR nuclear translocator (ARNT). This heterodimer binds to specific genomic sequences of responsive genes to modulate their gene expression. These specific binding sequences contain the substitution intolerant 5′-GCGTG-3′ core sequence and are referred to as the dioxin response elements (DREs). Ultra violet cross-linking (21) and site selection experiments (22) indicate that the AhR occupies the 5′-TGTC half-site, while ARNT contacts the GTG-3′ half-site. However, strong evidence indicates that the 5′- and 3′-flanking nucleotides play an important role in modulating DNA-binding affinity and enhancer function (23–25).

TCDD and related compounds, including polychlorinated biphenyls and polyaromatic hydrocarbons, alter the expression of various genes involved in metabolism and detoxification (26). However, dysregulation of classical drug-metabolizing enzymes alone fails to adequately explain the tissue-, sex- and species-specific toxicity of AhR ligands (27). Consequently, the genome-wide identification of additional targets is required in order to investigate the pathological and physiological role of the AhR, and to elucidate the mechanisms of toxicity of TCDD and related compounds.

We have taken a comparative computational scanning approach to identify putative DREs in the genomic sequences of human, mouse and rat target genes. The AhR regulon provides an ideal model as most, if not all, TCDD-elicited effects are mediated through the interaction between the AhR complex with the DRE core sequence. A total of 13 bona fide DREs were used to establish a position weight matrix (PWM) and a matrix similarity (MS) score threshold to prioritize computationally identified DREs. Comparative analysis and complementary in vitro and in vivo gene expression studies validate the approach and highlight challenges in verifying the genome-wide functionality of computationally identified putative response elements.

**MATERIALS AND METHODS**

**Computational scanning for DREs**

Unambiguous genomic sequence (−5000 to +2000 bp) for 17,882 human (hg15), 11,697 mouse (mm3) and 3,896 rat (rn2) genes corresponding to RefSeq accession numbers was extracted from the UCSC Genome Browser (http://genomic.ucsc.edu). Genomic sequences were scanned for exact matches to the DRE core sequence, GCGTG, on both positive and negative strands. For each match, the extended 19 bp sequence was used to calculate an MS score (28), and compared to an MS score threshold of 0.85 which was based on the lowest MS score of 13 bona fide DREs (Table 1). A Java application was developed to implement the search algorithm and to calculate an MS score (available upon request). DRE frequency and location were subsequently mapped for each gene in the region of −5000 to +2000 bp in 500 bp increments. In order to investigate the change occurrence of the DRE core sequence, a set of 10,000 DNA sequences, with each sequence having a length of 5000 bp, was compiled using a Java application that randomly selected A, C, G or T to ensure independent and identical distributions for the nucleotides. This DNA sequence set was then analyzed as described above. The Wilcoxon’s rank–sum test was used to compare the DRE distributions on a per species basis to the uniform, chance distribution observed with the set of 10,000 random DNA sequences. Comparison of the number of expected (due to chance) and observed DREs per species was performed using the chi-square test. Both statistical tests were performed using R v1.8.1 (http://www.r-project.org).

**Identification of DREs in orthologous genes**

Orthologous genes were retrieved from NCBI Homologene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=homologene, Build 31). The multiple sequence alignment tool, ClustalW (http://www.ebi.ac.uk/clustalw/), was used to identify consensus regions for all genes possessing DREs between −1500 and the transcriptional start site (TSS) with an MS score above the threshold. Positively conserved DREs were then identified by searching DREs located within conserved regions. Using indices of putative DREs, we further mapped first level conserved DREs in Cyp1al to find out whether they were located within conserved regions. For Cyp1al, a poorly annotated TSS required upstream regulatory sequences associated with mRNA RefSeq NM_000499 (Human CYPIA1), NM_009992 (Mouse Cyp1al) and NM_012540 (Rat Cyp1al) to be downloaded from the UCSC Genome Browser and manually curated for analysis.

**Animal and cell treatment**

Female C57BL/6 mice, ovariectomized by the vendor on postnatal day 20 and all having body weights within 10% of the average body weight, were obtained from Charles River Laboratories on postnatal day 26 (Raleigh, NC). Immature ovariectomized C67BL/6 mice provided a well-controlled responsive model to examine temporal and dose-dependent changes in gene expression that is not confounded by circulating endogenous estrogens as a result of the onset of sexual maturation and the estrus cycle. Mice were housed

---

Table 1. Sequence of bona fide DREs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Site</th>
<th>DRE sequence 5′-3′</th>
<th>MS score</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCyp1al</td>
<td>a</td>
<td>caagctcGCGTGgaagaag</td>
<td>0.94</td>
<td>(24)</td>
</tr>
<tr>
<td>s</td>
<td>ccctgtGCGTGcgaaga</td>
<td>0.95</td>
<td>(24)</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>cggattGCGTGcgaagaag</td>
<td>0.98</td>
<td>(24)</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>caagctcGCGTGcagcac</td>
<td>0.91</td>
<td>(24)</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>cggattGCGTGcgaagaag</td>
<td>0.97</td>
<td>(24)</td>
<td></td>
</tr>
<tr>
<td>mCyp1b1</td>
<td>XRE1</td>
<td>ccctgtGCGTGcgaagaag</td>
<td>0.96</td>
<td>(38)</td>
</tr>
<tr>
<td>c</td>
<td>ccctgtGCGTGcgaagaag</td>
<td>0.95</td>
<td>(55)</td>
<td></td>
</tr>
<tr>
<td>rCyp1al</td>
<td>XRE2</td>
<td>ggattGCGTGcagcac</td>
<td>0.88</td>
<td>(55)</td>
</tr>
<tr>
<td>t</td>
<td>caaattGCGTGcagcac</td>
<td>0.85</td>
<td>(62)</td>
<td></td>
</tr>
<tr>
<td>rAdh3</td>
<td>tccctgtGCGTGcagcac</td>
<td>0.95</td>
<td>(63)</td>
<td></td>
</tr>
<tr>
<td>rSodl</td>
<td>gaggctGCGTGcagcac</td>
<td>0.89</td>
<td>(64)</td>
<td></td>
</tr>
<tr>
<td>rGstya</td>
<td>gcagttGCGTGcagcac</td>
<td>0.89</td>
<td>(48)</td>
<td></td>
</tr>
<tr>
<td>rUGT1A1</td>
<td>agaagttGCGTGcagcac</td>
<td>0.92</td>
<td>(65)</td>
<td></td>
</tr>
</tbody>
</table>

*Abbreviations: mCyp1al, mouse cytochrome P450A1; mCyp1b1, mouse cytochrome P450B1; rCyp1al, rat Cyp1al; tAdh3, rat aldehyde dehydrogenase-3; rNqr1, rat NADPH:quinone oxidoreductase; rSodl, rat Cu/Zn superoxide dismutase; rGstya, rat glutathione S-transferase Ya; and rUGT1a1, rat UDP-glucuronosyltransferase 1A1.*
in polycarbonate cages containing cellulose fiber chips (Aspen Chip Laboratory Bedding, Northeastern Products, Warrenberg, NY) in a 23°C high-efficiency particulate air (HEPA)-filtered environment with 30–40% humidity and a 12 h light/dark cycle (07.00–19.00 h). Animals were allowed free access to deionized water and Harlan Teklad 22/5 Rodent Diet 8640 (Madison, WI) and acclimatized for four days prior to dosing. Animals were treated by gavage with 0.1 ml of sesame oil for a nominal dose of 0 (vehicle control) or 10 µg/kg body weight of TCDD (provided by S. Safe, Texas A&M University, College Station, TX) for the quantitative real-time PCR (QRTPCR) time course study. Four animals were treated per dose and time point, and groups for each dose and time point were housed in separate cages. Treated mice and their time-matched vehicle controls were sacrificed 12, 24 and 72 h after dosing by gavage for QRTPCR analysis of specific genes. Animals for the dose–response microarray analysis were sacrificed 24 h following treatment for RNA isolation and a section of the liver was snap frozen in liquid nitrogen and stored at −80°C until further use. All procedures were performed with the approval of the Michigan State University All-University Committee on Animal Use and Care.

Mouse Hepa1c1c7 cells (obtained from Dr O. Hankinson, UCLA) were seeded (1 × 10⁶ cells) onto 150 mm culture dishes (Corning, Acton, MA) in triplicates in phenol red-free DMEM (Invitrogen, Carlsbad, CA) supplemented with 5% fetal bovine serum (SeroLogicals Inc., Norcross, GA), 100 U/ml of penicillin/streptomycin (Invitrogen), 2.5 µg/ml of amphotericin B (Invitrogen) and 100 µg/ml of gentamicin reagent solution (Invitrogen) at 37°C in a 5% CO₂ humidified environment. The cells were treated with dimethyl sulfoxide (vehicle control) or TCDD (10 nM) for 2, 4, 6, 8 and 12 h, and harvested by scraping in Trizol (Invitrogen).

RNA isolation
Liver samples (~70 mg) were transferred to 1.0 ml of Trizol in a 2.0 ml microfuge tube and homogenized using a Mixer Mill 300 tissue homogenizer (Retsch, Germany). Total RNA was isolated from homogenized tissue samples and cell culture samples according to the manufacturer’s protocol with an additional phenol–chloroform extraction. Isolated RNA was resuspended in RNA storage solution (Ambion Inc., Austin, TX), quantified (A₂₆₀) and assessed for purity by determining the A₂₆₀/A₂₈₀ ratio and by visual inspection of 1.0 µg on a denaturing gel.

cDNA microarray analysis
Detailed protocols for microarray construction, labeling of the cDNA probe, sample hybridization and slide washing can be found at http://dbzach.fst.msu.edu/interfaces/microarray.html. Briefly, PCR-amplified DNA was robotically arrayed in duplicate onto epoxy-coated glass slides (Quantifoil, Germany) using an Omniprobe arrayer (GeneMachines, San Carlos, CA) equipped with 16 (4 × 4) Chipmaker 2 pins (Telechem) at the Genomics Technology Support Facility (http://www.genomics.msu.edu) at Michigan State University. Total RNA from TCDD-treated liver samples was compared to total RNA from time-matched control (vehicle-treated samples) using the ‘spoke’ method which involves comparing treated samples to a common vehicle control with appropriate dye swaps as described previously (29). Four replicates were performed with each hepatic sample representing a single animal. Dye-swap hybridizations were conducted to account for dye biases for a total of eight hybridizations per treatment. The animal was considered as the experimental unit and there was no pooling of samples. Total RNA (25 µg) was reverse transcribed in the presence of Cy3- or Cy5-dUTP to create fluor-labeled cDNA which was purified using a Qiagen PCR purification kit (Qiagen, Valencia, CA). The Cy3 and Cy5 samples were mixed, vacuum dried and resuspended in 32 µl of hybridization buffer (40% formamide, 4× SSC and 1% SDS) with 20 µg poly(dA) and 20 µg of mouse Cot-1 DNA (Invitrogen) as competitor. This probe mixture was heated at 95°C for 5 min and was then hybridized on an array under a 22 × 40 mm² coverslip (Corning Inc., Corning NY) in a light protected and humidified hybridization chamber (Corning Inc.). Samples were hybridized for 18–24 h at 42°C in a water bath. Slides were then washed, dried by centrifugation and scanned at 635 nm (Cy5) and 535 nm (Cy3) on an Affymetrix 428 Array Scanner (Santa Clara, CA). Images were analyzed for feature and background intensities using GenePix Pro 5.0 (Axon Instruments, Union City, CA).

cDNA microarray data were normalized using a semiparametric method that combines both parametric and non-parametric methods (J. E. Eckel, C. Gennings, T. M. Therneau, D. R. Boverhof, L. D. Burgoon and T. Zacharewski, submitted for publication). A model-based t-test was used to calculate t-scores using the General Linear Mixed Model: yicad = μi + Ci + Ai + Dij + γicad where γicad is the normalized feature intensity for the i-th cDNA, μ is the global mean for the i-th gene, Ci is the fixed effect of dye for the i-th gene and the c-th dye (Cy3 or Cy5), Ai is the random main effect for the α-th microarray (1, ..., A), Dij is the fixed effect for the d-th dose (0, ..., D) and γicad is the global error term with normal mean and SD. The model-based t-test was calculated per cDNA and per dose-group, comparing each dose individually to the expression in vehicle-treated samples for that cDNA.

An empirical Bayes method was used to identify active genes based on the model-based t-scores (30). The empirical Bayes was used to filter genes for activity based on the P1(r)-value, the posterior probability of activity, with a P1(r) cutoff of 0.99.

QRTPCR analysis
QRTPCR was performed on an Applied Biosystems Prism 7000 Sequence Detection System (Applied Biosystems, Foster City, CA) using the SYBR Green PCR Core Reagents (Applied Biosystems) as described previously (29,31). Total RNA (1 µg) was used as template for a reverse transcriptase reaction in 20 µl of 1× First-Strand Synthesis buffer (Life Technologies) containing 1 µg of oligo (dT₁₈A/C/GN), 0.2 mM of dNTPs, 10 mM of DTT and 200 U of Superscript II reverse transcriptase (Life Technologies). The reaction mixture was incubated at 42°C for 60 min and was stopped by incubation at 75°C for 15 min. Amplification of cDNA was performed in a single MicroAmp Optical 96-well Reaction
Plate (Applied Biosystems). Primer pairs for each gene were designed using Primer3 (32). Gene names, accession numbers, the forward and reverse primer sequences, amplicon size, and optimal primer and Mg²⁺ concentrations are listed in Table S5. Each plate contains standards of purified PCR products with known template concentration covering at least six orders of magnitude in order to interpolate relative template concentration of the experimental samples from standard curves of log copy number versus threshold cycle (C_t). No template controls (NTC) were also included on each plate. The relative level of mRNA was standardized to the housekeeping genes β-actin in order to control the differences in RNA loading, quality and cDNA synthesis.

The relative mRNA copy number was transformed using the natural logarithm for statistical analysis, and normality assured by normal probability–probability plots. Transformed copy numbers were analyzed on a gene-by-gene basis across time, between time-matched TCDD and vehicle treatments using a General Linear Model (GLM) in SAS v8.02. Contrasts between time-matched treated and vehicle samples were performed using a GLM model-based t-test.

RESULTS
Establishing a DRE MS score threshold
Prior to computationally scanning for putative DREs within genomic sequence data, an MS score threshold was established that could subsequently be used to assess the putative functionality of a motif when compared to bona fide DREs empirically determined to be functional. A 19 bp PWM, which included the intolerant 5 bp core DRE sequence (GGCTG) and the adjacent variant 7 bp flanking sequences, was developed using 13 reported functional DREs (Table 1). The PWM considers most of the provided sequence information to represent the nucleotide frequency at each position (Figure 1). In addition, it has the advantage of increased specificity in distinguishing relevant binding sequences due to the ability to assign an MS score to any oligonucleotide sequence of equal length to the matrix (28). DREs with MS scores above a selected threshold are expected to have a greater probability of possessing a measurable binding affinity and presumable biological relevance. MS scores for the 13 functional DREs are listed in Table 1 with the rat aldehyde dehydrogenase-3 DRE exhibiting the lowest MS score of 0.85, which was adopted as the threshold cutoff to initially assign tentative functionality to putative DREs that were identified by computational scanning.

Defining the genomic search region
Previous studies fail to define a region to restrict the computational search for putative DREs. We took advantage of the availability of genomic sequence data to examine the distribution of DREs in 17 882 human, 11 697 mouse and 3896 rat genes. Occurrence of the DRE core sequence was scanned within 5000 bp upstream and 2000 bp downstream of the TSS for all available genes from the human (hg15), mouse (mm3) and rat (rm2) genomes. In total, 12 420 human, 7835 mouse and 3043 rat DRE core sequences were identified with an MS score greater than the threshold (0.85), which corresponds to the presence of at least one putative DRE in 8290 human, 5238 mouse and 1837 rat genes (Figure 2). The distributions of putative functional DREs in the human, mouse and rat genomes were higher in regions proximal to the TSS (Figure 3A–C). Supplementary Table S1a–c provides MS scores for each computationally identified DRE as well as the associated gene name, LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink), RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/), location and sequence.

A Wilcoxon’s rank-sum test confirmed (P < 0.0001) an even DRE core sequence distribution pattern (Figure 3D) within 10 000 random DNA sequences, each consisting of 5000 bp, and that the DRE core sequence distribution in human, mouse and rat genomes differed significantly (P < 0.0001) from the uniform, random distribution (Figure 4). Moreover, the occurrence of the DRE core sequences in areas distal to the TSS is significantly (P < 0.0001) less than what would occur due to chance, and overall, the number of expected DRE core sequences is significantly different from the number of observed DREs for all species (P < 0.0001).

Further assessment of putative DREs
Within −1500 to +1500 bp, 5763 human, 3539 mouse and 1190 rat genes possessed at least one putative DRE. Genes annotated with the same TSS and first codon were removed, leaving 5489, 3182 and 1017 human, mouse and rat genes, respectively (Figure 2). Poor annotation of the TSS affected ~5, 10 and 15% of the human, mouse and rat genes, respectively. This has been attributed to inaccurate mapping of the TSS as a result of incomplete sequencing information for the 5’-untranslated region since expressed sequence tag sequencing is biased to the 3’ end (33).

The number of DREs within the −1500 to +1500 genomic region varied considerably between genes. We identified 1368 human, 736 mouse and 262 rat genes that contained multiple putative DREs within −1500 to +1500 bp of the TSS (Supplementary Table S2a–c). For example, the human vesicle-associated membrane protein (VAMP)-associated protein B and C gene (NM_004738) possessed 9 putative DREs, the mouse barttin gene (NM_080458) possessed 11 putative DREs and the rat insulin receptor substrate 3 gene (NM_032074) had 9 putative DREs. These genes are good candidates for AhR-mediated regulation since several of the DREs within the genomic sequence had MS scores above the threshold. Furthermore, tandem DREs are known to have a cooperative effect on gene expression (34,35), and therefore may indicate genes with a higher probability of responsiveness. Nonetheless, gene expression can be dependent on the species, sex, developmental stage, tissue, cell type and promoter context. Therefore, verification will be highly dependent on the selection of a suitable model maintained under the appropriate conditions. This may be a significant factor contributing to the high false-positive rate for identifying functional transcriptional response elements.

Microarray analysis of putative DRE containing genes
Mouse cDNA microarrays containing 13 362 features representing 8891 unique LocusLink IDs were used to identify AhR-regulated genes in C57BL/6 liver samples following treatment by gavage with 100 and 300 μg/kg body weight of TCDD.
for 24 h as part of a more comprehensive study examining
temporal- and dose-dependent changes in gene expression
(D.R. Boverhof, L.D. Burgoon, C. Tashiro, B. Chittim,
J.R. Harkema and T.R. Zacharewski, manuscript in prepara-
tion). Of the genes represented, 6061 had available genomic
sequence and a TSS annotated separately from the first codon
(Figure 5). Putative DREs were computationally identified
within the −1500 to +1500 bp genomic sequence for 1856
(21%) of these genes.

The hepatic expression level of 943 features representing
739 unique LocusLink IDs were found to be responsive
\[ P1(t) > 0.99 \] following 24 h treatment with TCDD, the
prototypical AhR ligand (Supplemental Table S3). Genomic sequence was available for 585 responsive genes with well-annotated TSSs. A total of 192 genes exhibited significant $[P1(t) > 0.99]$ induction or repression and had putative DREs within $-1500$ and $+1500$ bp with 81 of these responsive genes exhibiting a significant change in expression of at least 1.5-fold. If limited to $-1500$ to the TSS, the region where all known functional DREs have been reported, 118 responsive genes possessed a putative DRE. It is important to note that preliminary time course studies indicate that 24 h may not be the optimal time to capture TCDD-elicited expression changes for all responsive genes (data not shown). It is expected that as other models and time points are examined, the number of responsive genes will increase, although some responses may be due to secondary effects.

No evidence of overt toxicity was observed for 24 h at any of the doses examined. However, histopathological examination using hematoxylin and eosin staining indicated some evidence of very mild hydropic degeneration at 100 and 300 $\mu$g/kg TCDD. This likely contributes to the high-false-positive rate for TCDD responsive genes that did not possess a putative DRE as identified by computational scanning.

Comparative analysis and verification of positionally-conserved DREs in orthologous genes

The identification of evolutionary conserved regions has proven to be a powerful strategy to identify putative functional elements within genomic sequence (11–14). To increase the probability of computationally identifying functional response elements, the $-1500$ to TSS genomic region of orthologous human, mouse and rat genes were comparatively scanned for the presence of putative DREs. Coding regions (i.e. TSS to $+1500$ bp) were not included in this analysis as no previously characterized DREs have been located within these regions, and homology biases inherent in coding regions are likely to increase the false-positive rate. Orthologous human, mouse and rat genes were obtained from HomoloGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=homologene, Build 31) based on a reciprocal best match. Among 3087 human, 1745 mouse and 554 rat DRE-containing genes within $-1500$ bp and the TSS, 365 human–mouse, 140 human–rat and 133 mouse–rat orthologous genes containing DREs were identified (Figure 6 and Supplementary Table S4a–c). In addition, 48 human–mouse–rat orthologous genes were identified that contain a putative DRE in the $-1500$ to TSS genomic region (Supplementary Table S4d).

ClustalW, a multiple sequence alignment tool, was then used to identify 19 human, mouse and rat orthologous genes containing positionally conserved DREs in the $-1500$ to TSS genomic region (Table 2). Another 21 mouse–rat orthologous genes were identified to contain positionally conserved DREs but were not conserved in human. Surprisingly, no additional positionally conserved DREs were identified between human–mouse and human–rat orthologs.

AhR-mediated regulation for the 19 orthologous genes containing positionally conserved DREs was examined using QRTPCR in C57BL/6 mice liver and mouse Hepa1c1c7...
hepatoma cell models following TCDD treatment. Six of the nineteen genes exhibited *in vivo* hepatic induction with marked differences in kinetics and levels of expression (Table 3). For example, *Cyp1a1* and *Cyp1b1*, two well-characterized responsive genes, were significantly induced by TCDD at 12 h while reaching 1420- and 82-fold induction, respectively, at 72 h in the mouse liver. Four other genes, *Ugdh, Sta2, Znf148* and *Rab10*, were also significantly induced. In contrast, only two of the six *in vivo* responsive genes, *Cyp1a1* and *Cyp1b1*, were significantly induced in Hepa1c1c7 cells. Moreover, *Khsnpd1* repression (2.6-fold at 2 h) was specific to Hepa1c1c7 cells. Gene expression levels for the other 12 genes were not significantly affected by TCDD in either model.

Computational scanning of the human, mouse and rat *Cyp1b1* gene, identified 7, 10 and 10 DREs, respectively, between −1500 and the TSS region (Table 4). ClustalW alignment identified four positionally conserved DREs in highly homologous regions between −1100 and −800 bp relative to the TSS (Figure 7). Three of the four positionally conserved DREs (−1026, −855 and −835 for human; −1033, −891 and −872 for mouse; −1001, −859 and −840 for rat) had MS scores above the 0.85 threshold. The four positionally conserved DREs have an average MS score of 0.91, well above the average MS score of 0.80 for the 15 non-conserved identified DREs. Gel-mobility shift assays have demonstrated that the AhR/ARNT heterodimers bind to DREs at positions −1026, −855 and −835 in the human *CYP1B1* gene (36). Moreover, the positionally conserved DREs located at −872 and −891 in the mouse *Cyp1b1* promoter have also proved functional (37,38). Extrapolation of these results also facilitates the ranking and prioritization of putative DREs in the rat promoter for functional analysis.

**DISCUSSION**

Following completion of the human, mouse and rat genomes, attention has focused on identifying and elucidating the precise location of all sequence-based functional elements. While significant progress has been made identifying protein-coding sequences, most other sequence-based functional elements within the genome, including transcriptional regulatory elements, remain largely uncharacterized. The AhR, a phylogenetically ancient protein with homologs in nearly all living vertebrates including early chordates such as the sea...
database (Build 31).

indicate the number of orthologous genes as defined by the Homologene TSS and an MS score greater than 0.8548 were identified. Overlapped regions of cis-regulatory elements, particularly in vertebrate organisms such as human and rodents, is challenging. Within well-characterized genomes (40).

Figure 5. cDNA Microarray analysis of TCDD responsive genes. Immature, female ovariectomized C57BL/6 mice were treated by gavage with 100 and 300 µg/kg body weight of TCDD or vehicle for 24 h. Total RNA was extracted from hepatic tissue and reverse transcribed in the presence of Cy3- or Cy5-dUTP to create fluor-labeled cDNA as described previously (29). The probe mixture was hybridized to a 13,362 mouse cDNA array representing 8891 unique LocusLink IDs. Four independent replicates were performed and dye-swap hybridizations for a total of eight hybridizations per dose of TCDD. The animal was the experimental unit and there was no pooling of samples. Raw intensity data were normalized using a semiparametric method that combines both parametric and non-parametric methods (J. E. Eckel, C. Gennings, T. M. Therneau, D. R. Boverhof, L. D. Burgoon and T. Zacharewski, submitted for publication), and significant changes in gene expression were identified using an empirical Bayes method (30). Only 192 of the 1856 well-annotated genes with putative DREs between −1500 to +1500 bp exhibited a significant change in expression.

Figure 6. Venn diagram of human, mouse and rat DRE-containing genes. Well-annotated human, mouse and rat genes with DREs between −1500 bp and the TSS and an MS score greater than 0.8548 were identified. Overlapped regions indicate the number of orthologous genes as defined by the HomoloGene database (Build 31).
### Table 2. Orthologous human, mouse and rat genes containing positionally conserved DREs

<table>
<thead>
<tr>
<th>Human RefSeq</th>
<th>Locuslink</th>
<th>Gene abbreviation</th>
<th>Mouse RefSeq</th>
<th>Locuslink</th>
<th>Gene abbreviation</th>
<th>Rat RefSeq</th>
<th>Locuslink</th>
<th>Gene abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_000475</td>
<td>190</td>
<td>NR0B1</td>
<td>NM_007430</td>
<td>11614</td>
<td>Nr0B1</td>
<td>NM_053317</td>
<td>58850</td>
<td>Nr0B1</td>
</tr>
<tr>
<td>NM_001609</td>
<td>36</td>
<td>ACADS</td>
<td>NM_025826</td>
<td>66885</td>
<td>Acads</td>
<td>NM_013084</td>
<td>25618</td>
<td>Acads</td>
</tr>
<tr>
<td>NM_002156</td>
<td>3329</td>
<td>HSPD1</td>
<td>NM_010477</td>
<td>55110</td>
<td>Hspd1</td>
<td>NM_022299</td>
<td>63868</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_003359</td>
<td>7358</td>
<td>UGDP</td>
<td>NM_009466</td>
<td>22235</td>
<td>Ugd</td>
<td>NM_031325</td>
<td>83472</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_004501</td>
<td>3192</td>
<td>HNRPU</td>
<td>NM_0805</td>
<td>51810</td>
<td>Hnrpu</td>
<td>NM_057139</td>
<td>117280</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_005389</td>
<td>5110</td>
<td>PCTM1</td>
<td>NM_008786</td>
<td>18537</td>
<td>Pctm1</td>
<td>NM_013073</td>
<td>25604</td>
<td>Pctm1</td>
</tr>
<tr>
<td>NM_005662</td>
<td>7419</td>
<td>VDAC3</td>
<td>NM_011696</td>
<td>22335</td>
<td>Vdac3</td>
<td>NM_031355</td>
<td>83532</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_006569</td>
<td>10657</td>
<td>KHDRBS1</td>
<td>NM_011317</td>
<td>20218</td>
<td>Khdrbs1</td>
<td>NM_130405</td>
<td>117268</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_012479</td>
<td>7532</td>
<td>YWHAG</td>
<td>NM_018871</td>
<td>22628</td>
<td>Yw Hag</td>
<td>NM_019376</td>
<td>56010</td>
<td>Yw Hag</td>
</tr>
<tr>
<td>NM_021964</td>
<td>7707</td>
<td>ZNF148</td>
<td>NM_011749</td>
<td>22661</td>
<td>Zfp148</td>
<td>NM_031615</td>
<td>58820</td>
<td>Zfp148</td>
</tr>
<tr>
<td>NM_053024</td>
<td>5217</td>
<td>P62</td>
<td>NM_019410</td>
<td>18645</td>
<td>P62</td>
<td>NM_030873</td>
<td>81531</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_145805</td>
<td>64843</td>
<td>ISL2</td>
<td>NM_027397</td>
<td>104360</td>
<td>Isl2</td>
<td>NM_020471</td>
<td>57233</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_007222</td>
<td>781</td>
<td>CACNA2D1</td>
<td>NM_009784</td>
<td>12293</td>
<td>Cacna2d1</td>
<td>NM_012919</td>
<td>25399</td>
<td>Cacna2d1</td>
</tr>
<tr>
<td>NM_003040</td>
<td>6522</td>
<td>SLC4A2</td>
<td>NM_009207</td>
<td>20535</td>
<td>Sla4a2</td>
<td>NM_017048</td>
<td>24780</td>
<td>Sla4a2</td>
</tr>
<tr>
<td>NM_003129</td>
<td>6713</td>
<td>SQLE</td>
<td>NM_009270</td>
<td>20775</td>
<td>Sqcle</td>
<td>NM_017136</td>
<td>29230</td>
<td>Sqcle</td>
</tr>
<tr>
<td>NM_016131</td>
<td>10890</td>
<td>RAB10</td>
<td>NM_016676</td>
<td>19325</td>
<td>Rab10</td>
<td>NM_017359</td>
<td>50993</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_000104</td>
<td>1543</td>
<td>CYP1B1</td>
<td>NM_009994</td>
<td>13078</td>
<td>Cyp1b1</td>
<td>NM_012940</td>
<td>25426</td>
<td>Cyp1b1</td>
</tr>
<tr>
<td>NM_003714</td>
<td>8614</td>
<td>STC2</td>
<td>NM_011491</td>
<td>20856</td>
<td>Stc2</td>
<td>NM_022280</td>
<td>63878</td>
<td>N/A</td>
</tr>
<tr>
<td>NM_004099</td>
<td>1543</td>
<td>CYP1A1</td>
<td>NM_009992</td>
<td>13076</td>
<td>Cyp1a1</td>
<td>NM_012540</td>
<td>24296</td>
<td>Cyp1a1</td>
</tr>
</tbody>
</table>

N/A, gene abbreviation not available.

### Table 3. In vivo and in vitro qRT-PCR examination of orthologous genes with positionally conserved DREs

<table>
<thead>
<tr>
<th>Mn RefSeq</th>
<th>Locuslink</th>
<th>Gene</th>
<th>Maximum fold change (time point)</th>
<th>P-value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_004666</td>
<td>22235</td>
<td>Ugd1</td>
<td>+4.0 (12 h)</td>
<td>2.9E–06</td>
</tr>
<tr>
<td>NM_009992</td>
<td>13076</td>
<td>Cyp1a1</td>
<td>+1420 (72 h)</td>
<td>2.89E–12</td>
</tr>
<tr>
<td>NM_009994</td>
<td>13078</td>
<td>Cyp1b1</td>
<td>+86 (72 h)</td>
<td>1.7 (4 h)</td>
</tr>
<tr>
<td>NM_011317</td>
<td>20218</td>
<td>Kspag1</td>
<td>No change</td>
<td>2.6 (2 h)</td>
</tr>
<tr>
<td>NM_011491</td>
<td>20856</td>
<td>Stc2</td>
<td>+2.2 (4 h)</td>
<td>0.009</td>
</tr>
<tr>
<td>NM_011749</td>
<td>22661</td>
<td>Znf148</td>
<td>+1.9 (24 h)</td>
<td>0.105</td>
</tr>
<tr>
<td>NM_016676</td>
<td>19325</td>
<td>Rab10</td>
<td>−2.1 (72 h)</td>
<td>0.068</td>
</tr>
</tbody>
</table>

<sup>a</sup>P-values not reported for genes with no change in expression following treatment with TCDD.

### Table 4. Orthologous Cyp1b1 DRE locations, sequences and MS scores

<table>
<thead>
<tr>
<th>DRE</th>
<th>Human Location&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Sequence&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Mouse Location</th>
<th>Sequence</th>
<th>Rat Location</th>
<th>Sequence</th>
<th>MS score&lt;sup&gt;c&lt;/sup&gt;</th>
<th>MS score&lt;sup&gt;c&lt;/sup&gt;</th>
<th>MS score&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conserved</td>
<td>−1026</td>
<td>cagctgCGTGGcgacgcg</td>
<td>0.93</td>
<td>−1033</td>
<td>cagctgCGTGGcgacgcg</td>
<td>0.91</td>
<td>−1001</td>
<td>cagctgCGTGGcgacgcg</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>−942</td>
<td>cgcctcCGTGGcaggtt</td>
<td>0.84</td>
<td>−978</td>
<td>cgcctcCGTGGcaggtt</td>
<td>0.83</td>
<td>−946</td>
<td>cgcctcCGTGGcaggtt</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>−855</td>
<td>ggccttCGTGGgacgcg</td>
<td>0.92</td>
<td>−891</td>
<td>ggccttCGTGGgacgcg</td>
<td>0.92</td>
<td>−859</td>
<td>ggccttCGTGGgacgcg</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>−836</td>
<td>ccccttCGTGGgacgcg</td>
<td>0.95</td>
<td>−872</td>
<td>ccccttCGTGGgacgcg</td>
<td>0.96</td>
<td>−840</td>
<td>ccccttCGTGGgacgcg</td>
<td>0.95</td>
</tr>
<tr>
<td>Non-conserved</td>
<td>−1494</td>
<td>agagttGCGTGTGacatac</td>
<td>0.87</td>
<td>−712</td>
<td>agagttGCGTGTGacatac</td>
<td>0.83</td>
<td>−677</td>
<td>agagttGCGTGTGacatac</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>−991</td>
<td>acccttCGTGGgttcgcg</td>
<td>0.82</td>
<td>−709</td>
<td>acccttCGTGGgttcgcg</td>
<td>0.81</td>
<td>−590</td>
<td>acccttCGTGGgttcgcg</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>−265</td>
<td>gttcagCGTGGttcgccg</td>
<td>0.73</td>
<td>−625</td>
<td>gttcagCGTGGttcgccg</td>
<td>0.75</td>
<td>−351</td>
<td>gttcagCGTGGttcgccg</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>−387</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.77</td>
<td>−397</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.77</td>
<td>−293</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>−322</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.76</td>
<td>−322</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.76</td>
<td>−293</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>−94</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.83</td>
<td>−94</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.83</td>
<td>−18</td>
<td>gggaggCGTGGgctcgccg</td>
<td>0.83</td>
</tr>
</tbody>
</table>

<sup>a</sup>Location of DREs was relative from TSS. The negative numbers indicate the upstream region of TSS.

<sup>b</sup>The DRE sequences included the core region and flanking sequences comparable to the DRE PWM regardless of the strand located.

−1500 and +1500 bp (Figure 4). Microarray analysis using hepatic tissue from mice treated with TCDD for 24 h identified 739 genes that exhibited a significant change in expression, with 192 of these genes possessing at least one DRE between −1500 and +1500 with a MS score above the threshold. Collectively, these results provide compelling indirect evidence for the presence of a functional DRE, and support hypotheses regarding the elucidation of the DRE(s)
responsible for mediating the response when considered in combination with MS score data. Although only $\approx 10\%$ of the genes possessing DREs within $-1500$ to $+1500$ bp of the TSS that were represented on the array exhibited a significant change in mRNA transcript levels (either induction or repression) following exposure to TCDD, this percentage will increase as other models and additional time points are examined. However, given the exposure time, some of the responses may be due to secondary effects, and therefore must be verified by further empirical studies. In addition, some responsive genes may have functional DREs that are below the MS score threshold. Nevertheless, gene expression and computational scanning data have provided valuable supporting data for subsequent verification by chromatin immunoprecipitation procedures including genome-wide approaches such as ChIP (51,52).

Comparative human, mouse and rat genome analysis was pursued using ClustalW, a multiple sequence-alignment tool, in order to further enhance the probability of identifying functional DREs by computational scanning. Sequence alignment of homologous regions has proven to be a superior approach to identify functional motifs when compared to simple alignment approaches. *Cis*-regulatory elements often cannot be simply aligned in promoter sequences in order to identify putative functional motifs. Empirical studies have shown that the distance of some functionally conserved response elements from the TSS varies significantly due to gap region length (53,54). For example, comparative analysis of the human *CYP1B1* from $-1500$ bp to the TSS identified a total of seven DREs. Four of these are positionally conserved in the mouse and rat, and three have been shown to be functional in gel-mobility shift assays (36). Other studies have also verified the functionality of positionally conserved *Cyp1b1* and *Cyp1a1* DREs in the mouse (24,37,38,55–57).

Examination of the 48 human–mouse–rat orthologs (Figure 6), in addition to the manually annotated *Cyp1a1* gene, resulted in the identification of 19 genes with positionally conserved DREs. Only two of these nineteen genes, *Cyp1a1* and *Cyp1b1*, have been reported to be AhR regulated. Seven of these nineteen genes were found to be TCDD responsive with six (i.e. Ugdh, Cyp1a1, Cyp1b1, Stc2, Znf48, Rab10) exhibiting a change in mouse liver and three in mouse Hepa1c1c7 cells (i.e. Cyp1a1, Cyp1b1, Khsnpd1). These results (Table 3) illustrate the importance of selecting an appropriate model when verifying putative response elements, and support our earlier argument that additional putative DREs will likely be verified as other models and time points are investigated. For example, maximum induction of *Cyp1b1* mRNA was 87-fold in mouse liver but only 1.7-fold in Hepa1c1c7 cells while 18-fold induction was reported in Hepa1c1c7 cells using different RT–PCR primers and conditions (58). These studies also demonstrate that different probes for different regions of *Cyp1b1* can significantly affect gene expression results which may be a contributing factor in the discrepancy between the reported *Cyp1b1* microarray and QRTPCR results in this study. Although the clone representing *Cyp1b1* was sequenced verified in-house, and our records indicate that it amplified well, basic local alignment sequence tool (BLAST) analysis indicates that regions of the *Cyp1b1* are homologous to other mouse BAC clones. In addition, examination of *Cyp1b1* mRNA expression in other ongoing microarray studies using the same representative clone also

Figure 7. Positionally conserved DREs in human, mouse and rat *Cyp1b1* genomic sequence. Alignment of orthologous human, mouse and rat *Cyp1b1* genomic sequences extracted from the UCSC Genome Browser. Four positionally conserved DREs [human *CYP1B1* (NM_000104), $-1123$ to $-787$; mouse *Cyp1b1* (NM_009994), $-1133$ to $-823$; and rat *Cyp1b1* (NM_012940), $-1101$ to $-791$], shaded in gray, were identified. Three of these positionally conserved DREs have been shown to be functional in gel–mobility shift assays (36).
provided equivocal results (D.R. Boverhof, L.D. Burgoon, C. Tashiro, B. Chittock, J.R. Harke and T.R. Zacharewski, manuscript in preparation). Collectively, this suggests that although the clone is representative of Cyp1b1, it probes a poor region of the transcript which may be a contributing factor in the discrepancy between the Cyp1b1 microarray and RT-PCR results.

Gene product functions associated with these orthologs are also consistent with biological activities elicited by TCDD and related compounds that are conserved between species. Among the 48 genes, 7 could be classified as being involved in oxidative stress, hypoxia and detoxification, 5 are associated with calcium homeostasis, and 5 are localized to the endoplasmic reticulum. Associations between oxidative stress, hypoxia and detoxification (59), calcium homeostasis (60,61) and TCDD exposure are well-established effects.

In addition to examining similarities, the availability of the human, mouse and rat genomes also facilitates identifying potential differences in DRE distributions that may reflect species-specific mechanisms of toxicity and interactions with other signaling pathways that could affect AhR-mediated gene expression. On a genome-wide scale, only 37% (49 out of 134) of the manually annotated Cyp1A1 is included) of mouse–rat orthologs with a DRE between −1500 and +1500 had an equivalent human ortholog. About 82% (40 out of 49) of the mouse–rat orthologs shared positionally conserved DREs. In contrast, only 39% (19 out of 49) of the human orthologs with positionally conserved DREs had a rodent counterpart with a positionally conserved DRE. Although biases may have been introduced through the pre-dominant use of bona fide rodent DREs in the development of the PWM, these statistics fuel the debate regarding the suitability of rodent models to assess the potential human health risks associated with exposure to AhR ligands. However, TCDD and related compounds are known to elicit species-specific effects and therefore, not all regulatory elements will be identified using comparative approaches. Nevertheless, these searches rank and prioritize the most promising DREs for further investigation which will facilitate the development of AhR regulons for physiological and toxic responses when integrated with other genome-wide technologies.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Jeremy Burt, Cora Fong, Josh Kwékel and Ed Dere for their constructive comments. D.R.B. is supported by a Michigan Agricultural Experiment Station Competitive Grant. T.R.Z. is partially supported by the Michigan Agricultural Experiment Station. This work was funded by NIH grant ES012245.

REFERENCES


