methods

Genome-wide identification of peroxisome proliferator response elements using integrated computational genomics[S]

Danielle G. Lemay and Daniel H. Hwang[1]

United States Department of Agriculture, Agricultural Research Service, Western Human Nutrition Research Center, and Department of Nutrition, University of California-Davis, Davis, CA 95616

Abstract Peroxisome proliferator-activated receptor (PPAR) agonists are currently used therapeutically in humans, even though many of their direct gene targets are unknown. Because PPARs can directly regulate gene expression through peroxisome proliferator response elements (PPREs), we pursued the computational prediction of PPREs on a genome-wide scale. Contrary to current hypotheses, PPREs are not isotype-specific, nor do flanking nucleotides confer additional information. However, a position weight matrix-based search for PPREs within upstream conserved elements yielded sufficient selectivity for a genome-wide search. Additionally, a novel motif occurring with greater prevalence than PPREs was revealed. Microarray and gene ontology analyses further validated our search technique and provided new functional clusters of genes that were not previously known to be directly regulated by PPARs (e.g., chromatin remodeling, DNA damage response, Wnt, and mitogen-activated protein kinase signaling). This first genome-wide library of high-confidence predicted PPAR target genes will be a valuable resource to PPAR biologists.—Lemay, D. G., and D. H. Hwang. **Genome-wide identification of peroxisome proliferation response elements using integrated computional genomics. *J. Lipid Res.* 2006.** 47: **1583–1587.**

Supplementary key words peroxisome proliferator-activated receptor • target genes • conserved elements • PACM

Peroxisome proliferator-activated receptors (PPARs) are a family of nuclear receptors that serve as cellular sensors of fatty acids and fatty acid derivatives and broadly regulate nutrient metabolism and energy homeostasis. Thus, PPARs are considered ideal targets for pharmaceutical intervention and are used therapeutically despite the fact that their mechanisms are incompletely understood. Many direct PPAR targets have been reported, but no comprehensive unbiased genome-wide search for PPAR target genes has been published.

The three PPAR isotypes, α, δ, and γ, are differentially expressed across tissue types and developmental stages. However, all three bind peroxisome proliferator response elements (PPREs) in regulatory regions of their target genes. In this study, all known PPREs are collected from the literature and basic assumptions about PPREs are investigated. The most selective detection technique, position weight matrix (PWM)-based search of PPREs within upstream conserved elements, is applied to the entire human genome to develop a library of PPAR target genes. This technique is further assessed by microarray and gene ontology (GO) analysis, yielding new insights in PPAR biology.

METHODS

DNA source

Sequences were downloaded using the University of California at Santa Cruz (UCSC) Table Browser (1) and the human May 2004 (2), mouse May 2004 (3), and rat June 2003 (4) genomic assemblies.

Collection of reported functional PPREs

Reported PPREs from 78 publications and collection methods are detailed in the Supplemental Section 1.

Detection of PPREs in DNA sequences

PWMs were generated using the CONSENSUS algorithm (5) on lists of reported PPREs. DNA sequences were scored against PWMs using the PATSER program (5). A DNA sequence whose

Manuscript received 11 November 2005 and in revised form 12 January 2006 and in re-revised form 30 March 2006.

Published, JLR Papers in Press, April 3, 2006.
DOI 10.1194/jlr.M500504-JLR200

Abbreviations: DR1, direct repeat with a 1 bp spacer; GO, gene ontology; GWM, generalized weight matrix; PACM, PPAR-associated conserved motif; PPAR, peroxisome proliferator-activated receptor; PPRE, peroxisome proliferator response element; PWM, position weight matrix; ROC, receiver operating characteristic; UCSC, University of California at Santa Cruz.

[1] To whom correspondence should be addressed.
e-mail: dhwang@whnrc.usda.gov

[S] The online version of this article (available at http://www.jlr.org) contains additional tables and references in 9 sections.

matrix score surpassed the cutoff value was a “detected” binding site. To evaluate PPREs for within-site correlations, the GMMPS program (6), which implements a generalized weight matrix (GWM) model, was used.

Evaluation of PPRE detection

Detection methods were evaluated using receiver operating characteristic (ROC) curves. Optimal discrimination techniques minimize the area under the curve. Each data point on the curve corresponds to a cutoff value: the false-negative coordinate is the fraction of reported PPREs that fall below the cutoff, and the false-positive coordinate is the fraction of random human promoter regions (5,000 bp) that contain a sequence that exceeds the cutoff. Although random promoter sequences may contain true binding sites, detection in such sequences is a direct measure of selectivity and a very useful benchmark when comparing detection methods.

Identification of overrepresented motifs

To determine the significance ($P \leq 0.05$) of motif occurrence between reported and random sets, the data were modeled using a binomial distribution.

Microarray data analyses

The National Center for Biotechnology Information’s PubMed and GEO databases were searched for PPAR microarray studies that published accession numbers of all regulated genes. Six were located (7–12). The two rat microarrays were excluded from tests involving conserved elements because such data were not available for the rat. See Supplemental Section 2 for further details.

GO analysis

The MAPPFinder tool (13) within GenMAPP version 2.0 was used to identify enriched GO terms with the Hs-Std_20041021.gdb and Mm-Std_20040824.gdb databases.

RESULTS

Development of a new PPRE matrix

Because the most promising of the TRANSFAC (14) matrices for PPRE detection were those based on reported functional sites (data not shown), the literature was searched for functional PPREs. In total, 73 DR1-like (direct repeat with a 1 bp spacer) PPREs (see Supplemental Table 1) were used to construct new PWMs (see Supplemental Section 3). The conservation of each nucleotide position in the core PPRE plus 5′ flanking nucleotides using WebLogo (15) is shown in **Fig. 1**.

To evaluate whether to include flanking nucleotides in the new matrix, matrices of widths 13–17 bp were generated using the reported PPREs. Matrices of widths >13 bp did not confer better discriminatory power (see Supplemental Figure 4A). Matrices of various widths constructed from sequences reported to bind PPARα alone or PPARγ alone were not better discriminators of PPREs overall (see Supplemental Figure 4B). Matrices constructed from PPREs reported to bind one of the isotypes were not better discriminators of PPREs of the same isotype than were matrices constructed from PPREs reported to bind the other isotype, even when flanking nucleotides were included (see Supplemental Figure 4C–D).

Higher order probability models were also evaluated. Use of a background model to reflect true GC content did not improve discrimination ability (data not shown). Furthermore, none of the nucleotide positions were cocorrelated under the GWM model, even when flanking nucleotides were included or PPREs were subgrouped by isotype.

Identification of PPREs and a novel motif in conserved elements

To improve selectivity, the search space was restricted to conserved elements within 5,000 bp upstream of reported human PPAR target genes. These elements, provided via the human Most Conserved track at UCSC, are conserved in space (neighboring nucleotides) and time (human, chimp, mouse, rat, dog, chicken, fugu, and zebrafish genomes), as identified by the phylogenetic hidden Markov model of Siepel et al. (16). Using our PWM (see Supplemental Section 3) to search these elements, PPREs were overrepresented among reported human genes compared with random genes ($P < 0.00001$). Furthermore, these PPREs occur with greater frequency than response elements of any PPAR or non-PPAR transcription factor in the TRANSFAC database using the MATCH program (17).

The upstream conserved elements were also searched for novel motifs using MEME (18). A motif of width 15 bp with the consensus TTCATTTGGACATTG was discovered. This motif, here named PACM, for PPAR-associated conserved motif (PWM in Supplemental Section 3), is more common than PPREs among these elements.

ROC curves for PPRE and PACM detection using our PWMs in upstream conserved elements are illustrated in **Fig. 2A**. With an average of only four conserved elements per gene, less than half of the reported human target genes have PPREs among their upstream conserved elements. Thus, this method cannot predict all direct PPAR targets. However, because the ROC curve is fairly sharp, a subset of targets can be predicted with high confidence. Of the reported genes that have any upstream conserved elements,

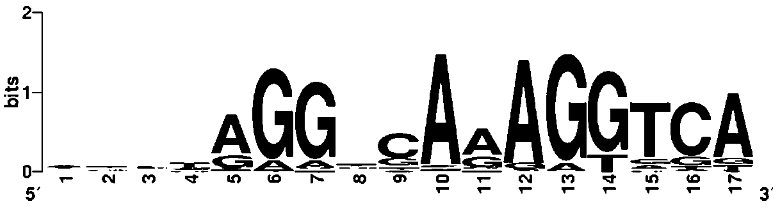

Fig. 1. Peroxisome proliferator response element (PPRE) sequence logo. The overall height of each column indicates conservation at that position in the alignment of 73 published DR1-like (direct repeat with a 1 bp spacer) PPREs, whereas the height of each letter within the column indicates the relative frequency of each nucleic acid at that position. Positions 5–7 represent the core DR1; positions 1–4 are 5′ flanking nucleotides.

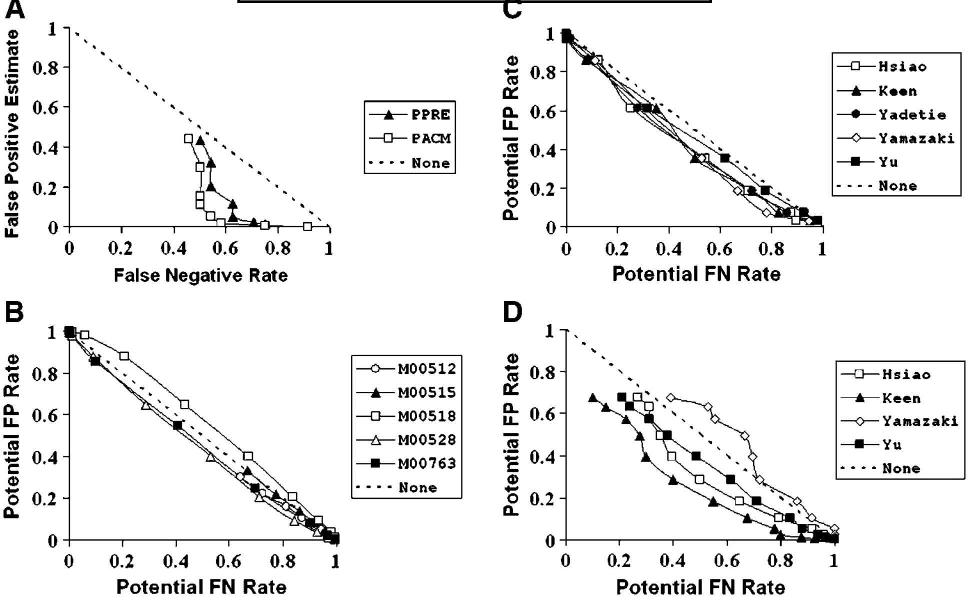

Fig. 2. PPRE detection by position weight matrix (PWM)-based search within upstream conserved elements is selective for a subset of the reported human peroxisome proliferator-activated receptor (PPAR) target genes and for genes upregulated in PPARγ microarray studies. A: PWM-based search within conserved elements 5,000 bp upstream of reported human PPAR target genes. B–D: TRANSFAC matrix-based search (B), PWM-based search (C), and PWM-based search (D) within conserved elements 5,000 bp upstream of upregulated genes. The false-negative (FN) rate or potential FN rate is the rate at which a PPRE was not detected, across a range of matrix score thresholds, upstream of the reported genes (A) or upregulated genes (B–D). The false-positive (FP) estimate or potential FP rate is the rate at which a PPRE was detected in conserved elements upstream of randomly selected genes (A) or nonregulated genes (B–D) at the same score thresholds. The dashed lines indicate the line of no discrimination.

70% have a PPRE or PACM at matrix scores for which detection in a random promoter is <10%.

Microarray data analyses

Although many regulated genes in PPAR microarrays may be indirect or secondary targets, at least some should be direct targets coordinated through a PPRE. Unfortunately, <5% of the reported PPAR targets were regulated in three of the four mouse microarrays. Nevertheless, the microarrays were used to assess PPRE detection methods, because one would expect some degree of PPAR target gene enrichment among regulated genes, however small.

Detection of upstream PPREs using TRANSFAC matrices does not distinguish upregulated from nonregulated genes (Fig. 2B). Using our PPRE matrix, the ROC curves fall beneath the nondiscrimination line, but the difference is underwhelming (Fig. 2C). However, when the search space is restricted to highly conserved elements, we see a distinct selectivity for upregulated genes in the three PPARγ microarrays but not in the PPARα microarray (Fig. 2D). None of the methods distinguish downregulated genes (data not shown).

Genome-wide prediction

We conducted a genome-wide search for PPREs among conserved elements in the 5,000 bp upstream of all human reference sequences. Of 24,033 genes, PPREs were detected upstream of 1,085 (cutoff score = 8). These genes and their PPREs are listed in Supplemental Section 5. The 1,207 genes with PACMs and the 172 genes with both PPREs and PACMs are also listed in Supplemental Sections 6 and 7.

GO analysis

Biological process GO terms that are statistically overrepresented (*Z* score ≥ 2) among the predicted PPAR target genes are given in **Table 1**. Only GO terms locally associated (nonnested) with three or more predicted genes were retained. Predicted PPAR target genes are sorted by these GO terms in Supplemental Section 8. GO analysis was also conducted on the reported PPAR target genes and regulated genes from all six microarrays. For each of these gene sets (reported, microarray, and predicted), the enriched GO terms were grouped into general categories to elucidate functional clusters (**Fig. 3**).

DISCUSSION

Prior researchers have highlighted the importance of the 5′ flanking nucleotides (19) and demonstrated isotype specificity (20, 21) using a small set of selected PPREs. Our study, based on a much larger set of reported PPREs, indicates that isotype-specific PWMs are not better predictors of same-isotype PPREs. When the matrix was extended to include flanking nucleotides, PPRE detection did not improve. Together, these results suggest that if there is any

TABLE 1. Biological processes of predicted peroxisome proliferator-activated receptor target genes

GO Identifier	GO Term	*Z* Score
45449	Regulation of transcription	6.098
6355	Regulation of transcription, DNA-dependent	5.94
6350	Transcription	5.85
6338	Chromatin remodeling	4.054
77	DNA damage response, cell cycle arrest	3.953
16568	Chromatin modification	3.786
7275	Development	3.72
8630	DNA damage response, induction of apoptosis	3.596
6631	Fatty acid metabolism	3.179
6366	Transcription from Pol II promoter	3.037
8152	Metabolism	3.012
6635	Fatty acid β-oxidation	2.984
7243	Protein kinase cascade	2.651
7223	Frizzled-2 signaling pathway	2.415
7001	Chromosome organization and biogenesis	2.412
8151	Cell growth and/or maintenance	2.405
9653	Morphogenesis	2.356
6629	Lipid metabolism	2.349
7517	Muscle development	2.345
6357	Regulation of transcription from Pol II promoter	2.299
9615	Response to virus	2.286
30154	Cell differentiation	2.042
16055	Wnt receptor signaling pathway	2.033
187	Activation of mitogen-activated protein kinase	2.016
188	Inactivation of mitogen-activated protein kinase	2.016

GO, gene ontology.

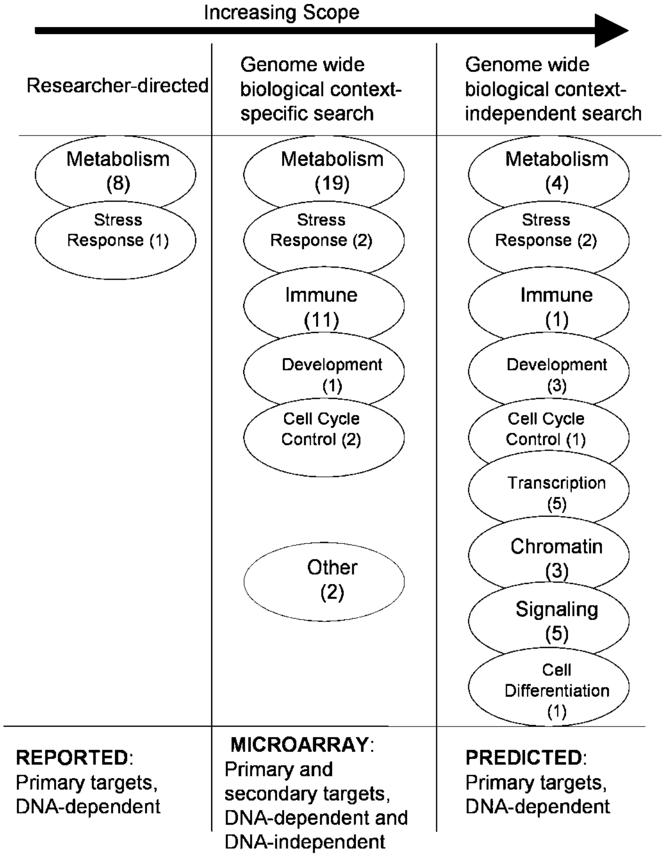

Fig. 3. Functional clusters of gene ontology (GO) terms among reported, microarray-regulated, and predicted PPAR target gene sets suggest new areas of research not yet explored by biological experiments. For comparison across gene sets, biological process GO terms that were overrepresented in each gene set were grouped together into functional clusters. Each oval represents a functional cluster, with the number of biological process GO terms within that cluster indicated in parentheses.

isotype specificity for PPREs beyond the generally higher binding affinity of PPARγ, it is not inherent to either the core DR1 or immediately flanking nucleotides.

Today, biologists commonly seek putative PPREs using the consensus or TRANSFAC matrix. Only 2 of the 73 reported DR1-like PPREs match an ideal DR1, and 8% have five or more mismatches from consensus. Unlike TRANSFAC matrix searches, our method detected PPREs with sufficient selectivity for a genome-wide search and preferentially selected upregulated genes in PPARγ microarray studies. Interestingly, the fact that downregulated genes were not favorably selected suggests that the primary mechanism by which PPARs suppress gene expression is not mediated through PPREs.

Although the PPAR targets predicted in our genome-wide search are not complete, they represent an important subset, namely those that are targets across many vertebrate species. A high false-negative rate (60%) was tolerated to minimize the false-positive rate (<10%). Nevertheless, the false-negative rate of our method is still an improvement over that of all microarrays analyzed (84–97%). Genes necessarily excluded from the predicted library include those with PPREs outside of upstream conserved elements and those whose PPREs are not DR1-like. Experimental verification of individual genes in future studies is necessary for unequivocal validation and to demonstrate the particular biological contexts (anatomical site, developmental stage, time of observation, stimulus, coregulatory molecules, DNA topology, etc.) in which they are regulated. Lastly, as the search space was restricted to conserved elements, there exists the possibility that this library is relevant only to vertebrate development. However, the economy of biology is such that developmental genes are often functional in the mature animal as well, albeit with a different function.

GO analysis supports the validity of the prediction method. First, both the predicted and microarray data sets contain the functional groups represented by the set of reported PPAR target genes. Second, GO terms represented by the predicted genes match areas of known PPAR biological function, such as DNA damage response, mitogen-activated protein kinase signaling, Wnt receptor signaling, cell differentiation, and muscle development, even though direct PPAR targets in these areas were previously unknown. Our study provides the first evidence that PPARs directly target such genes.

The GO analysis also suggests new mechanistic insights. The overwhelming number of immune-related GO terms among microarray-regulated genes, but not the reported or predicted direct targets, strongly suggests that immune function is primarily regulated by PPARs through indirect means. The enrichment of chromatin modification GO terms among the predicted set implies an exciting new hypothesis: PPARs directly target chromatin-remodeling genes. The fact that these genes were not regulated in the microarray studies may be time-dependent; the earliest observation point was at 24 h. Chromatin-remodeling genes may be targeted very early after stimulus and only transiently. Because PPARs can launch broad physiological changes, one might expect that a temporary increase in the quantity of chromatin-remodeling proteins would be necessary to implement such changes.

Surprisingly, a novel motif, PACM, was more prevalent than PPREs among conserved elements upstream of reported human PPAR target genes. This element may be bound by an unidentified transcription factor that coordinates with PPARs to regulate a subset of PPAR targets, especially considering that PPARγ itself contains both PPRE and PACM among its upstream conserved elements. Only 172 genes across the entire human genome contain both PPREs and PACMs in their upstream conserved elements. GO analysis of these genes (see Supplemental Section 9) suggests that their protein products are involved in lipid metabolism, perhaps in developing neurological tissue.

In summary, the major contributions of this study include the resolution of basic research questions about PPREs, a PPRE detection technique with demonstrated discriminatory ability, the identification of a novel *cis*-acting element through which PPAR-associated regulation is likely mediated, and new insights with respect to PPAR regulatory function. Additionally, the methodology, PWM- or GWM-based search within conserved elements as identified by a phylogenetic hidden Markov model and the monitoring of error rates using signal detection theory, can be applied to other *cis*-acting elements. Finally, the library of genes that contain high-confidence predicted PPREs should be a valuable resource for PPAR biologists.

The authors thank Ian Korf, Bruce German, and Charles Stephensen for helpful comments and Angie Hinrichs for UCSC technical support. This work was supported by grants from the National Institutes of Health (DK-064007), the U.S. Department of Agriculture (USDA; 01A095Rev), American Institute for Cancer Research, and program funds from the Western Human Nutrition Research Center (WHRNC)/Agricultural Research Service/USDA. D.G.L. was supported by a WHNRC fellowship.

REFERENCES

1. Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32:** D493–D496.
2. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature.* **409:** 860–921.
3. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature.* **420:** 520–562.
4. Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the brown Norway rat yields insights into mammalian evolution. *Nature.* **428:** 493–521.
5. Hertz, G. Z., and G. D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.* **15:** 563–577.
6. Zhou, Q., and J. S. Liu. 2004. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics.* **20:** 909–916.
7. Hsiao, A., D. S. Worrall, J. M. Olefsky, and S. Subramaniam. 2004. Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes. *Bioinformatics.* **20:** 3108–3127.
8. Keen, H. L., M. J. Ryan, A. Beyer, S. Mathur, T. E. Scheetz, B. D. Gackle, F. M. Faraci, T. L. Casavant, and C. D. Sigmund. 2004. Gene expression profiling of potential PPARgamma target genes in mouse aorta. *Physiol. Genomics.* **18:** 33–42.
9. Parton, L. E., F. Diraison, S. E. Neill, S. K. Ghosh, M. A. Rubino, J. E. Bisi, C. P. Briscoe, and G. A. Rutter. 2004. Impact of PPARgamma overexpression and activation on pancreatic islet gene expression profile analyzed with oligonucleotide microarrays. *Am. J. Physiol. Endocrinol. Metab.* **287:** E390–E404.
10. Yadetie, F., A. Laegreid, I. Bakke, W. Kusnierczyk, J. Komorowski, H. L. Waldum, and A. K. Sandvik. 2003. Liver gene expression in rats in response to the peroxisome proliferator-activated receptor-alpha agonist ciprofibrate. *Physiol. Genomics.* **15:** 9–19.
11. Yamazaki, K., J. Kuromitsu, and I. Tanaka. 2002. Microarray analysis of gene expression changes in mouse liver induced by peroxisome proliferator-activated receptor alpha agonists. *Biochem. Biophys. Res. Commun.* **290:** 1114–1122.
12. Yu, S., N. Viswakarma, S. K. Batra, M. Sambasiva Rao, and J. K. Reddy. 2004. Identification of promethin and PGLP as two novel up-regulated genes in PPARgamma1-induced adipogenic mouse liver. *Biochimie.* **86:** 743–761.
13. Doniger, S. W., N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, and B. R. Conklin. 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* **4:** R7.
14. Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24:** 238–241.
15. Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res.* **14:** 1188–1190.
16. Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
17. Kel, A. E., E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. 2003. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31:** 3576–3579.
18. Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2:** 28–36.
19. IJpenberg, A., E. Jeannin, W. Wahli, and B. Desvergne. 1997. Polarity and specific sequence requirements of peroxisome proliferator-activated receptor (PPAR)/retinoid X receptor heterodimer binding to DNA. A functional analysis of the malic enzyme gene PPAR response element. *J. Biol. Chem.* **272:** 20108–20117.
20. Juge-Aubry, C., A. Pernin, T. Favez, A. G. Burger, W. Wahli, C. A. Meier, and B. Desvergne. 1997. DNA binding properties of peroxisome proliferator-activated receptor subtypes on various natural peroxisome proliferator response elements. Importance of the 5′-flanking region. *J. Biol. Chem.* **272:** 25252–25259.
21. Kassam, A., J. Hunter, R. A. Rachubinski, and J. P. Capone. 1998. Subtype- and response element-dependent differences in transactivation by peroxisome proliferator-activated receptors alpha and gamma. *Mol. Cell. Endocrinol.* **141:** 153–162.