HRTBLDb: an informative data resource for hormone receptors target binding loci

Brian A. Kennedy1, Wenqing Gao2, Tim H.-M. Huang3 and Victor X. Jin1,*

1Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, 2Department of Pharmaceutical Sciences, SUNY at Buffalo, Buffalo, NY 14260, USA and 3Human Cancer Genetics Program, The Ohio State University, Columbus, OH 43210

Received June 24, 2009; Revised August 17, 2009; Accepted August 19, 2009

ABSTRACT

Three hormone receptors, the estrogen receptor (ER), the androgen receptor (AR) and glucocorticoid receptor (GR) play an important role in regulating the cellular differentiation tissue development of skin, bone, the brain and the endocrine system; therefore, there is a strong scientific need to identify and characterize hormone receptor transcriptional regulation. Given that the vast amount of regulatory data for hormone being produced by ChIP-based high-throughput experiments is widely scattered in disparate, poorly cross-indexed data stores, a flexible platform for organizing and relating these data would provide significant value. We created a data management system called the Hormone Receptor Target Binding Loci, HRTBLDb (http://motif.bmi.ohio-state.edu/hrtbldb), to address this problem. This database contains hormone receptor binding regions (binding loci) from in vivo ChIP-based high-throughput experiments as well as in silico, computationally predicted, binding motifs and cis-regulatory modules for the co-occurring transcription factor binding motifs, which are within a binding locus. It also contains individual binding sites whose regulatory action has been verified by in vitro experiments. The current version contains 44 673 binding elements with 114 hormone response elements which are verified by in vitro experiments; 75 binding motifs which occur with a hormone response element and whose co-regulatory action is verified by in vitro experiments; 18 472 binding loci from in vivo experiments; and 26 012 computationally predicted binding motifs.

INTRODUCTION

Hormone receptors (HRs) are steroid hormone activated transcription factors which are a part of the nuclear receptor super-family consisting of >60 members. Three hormone receptors, the estrogen receptor (ER), the androgen receptor (AR) and glucocorticoid receptor (GR) play an important role in regulating the cellular differentiation tissue development of skin, bone, the brain and the endocrine system (1). Many studies (2–4) have shown that all three of HRs regulates gene expression for many significant phenotypes. Dysregulation of any of these hormone signaling pathways is linked to many different types of diseases. For example, 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 (5). Dysregulation in the production of androgens can affect different organ systems, in which it caused the diseases such as androgen insensitivity syndromes, prostate cancer, hepatocellular carcinomas, acne and male pattern alopecia (6,7). GR activation is often associated with inducing apoptosis in lymphocytes and lymphomas, and paradoxically inhibiting the apoptotic response to several cell stressors including growth factor withdrawal (8) in breast cancer cell lines.

HRs modulate the transcription of target genes through hormone activation, and they do so either through (i) direct binding, in which HRs directly bind to hormone response elements; (ii) indirect binding, where HRs interact with other transcription factors, which instead bind to the DNA themselves; or (iii) co-occurrent binding, where HRs bind with lower affinity to DNA binding sequences and interact with other TFs through co-regulators. These modes of binding are shown at http://motif.bmi.ohio-state.edu/hrtbldb. ARs bind to the androgen response element 5′-GGAACAnnnTGTTTCT-3′ (9), ERs bind with the highest affinity to the estrogen

*To whom correspondence should be addressed. Tel: +614 292 4778; Fax: +614 688 6600; Email: victor.jin@osumc.edu

© The Author(s) 2009. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
response element consensus sequence 5\'-AGGTCAnnnTGACCT-3\' (10), and GRs bind to the glucocorticoid response element 5\'-GTTACAnnnTGTTCT-3\' (11).

Since hormone receptors are key regulators of growth, differentiation, and metabolism in multiple organs, the characterization of the transcriptional regulation of their target genes is quite important. Several new high-throughput technologies such as ChIP-chip (12,13), ChIP-seq (14,15) and ChIP-PET (16) are producing a large amount of whole-genome-wide hormone receptor regulatory data, in many different cell types and tissues [see Supplementary Table S1 for all literature as well as (17)]. Integrating these data into a single resource will allow expedited cross-reference, literature discovery and aid in the validation of novel data. To that end have created a database called the Hormone Receptor Target Binding Loci, HRTBLDb. This database mainly contains in vivo binding loci identified from ChIP-based high-throughput technologies and computationally predicted binding motifs and the cis-regulatory modules for the binding loci. It also collected in vitro experimentally defined binding elements. The data visualization focuses on a whole-genome-wide scale rather than traditional promoter regions and is accessed through simple yet expressive data filtering.

DATABASE CONTENT

In the current version of our HRTBLDb database, it contains two types of data, (i) in vitro binding elements, which were experimentally verified in vitro to bind one of the three hormone receptors using either EMSA or ChIP assays; (ii) in vivo binding loci identified from the ChIP-based high-throughput data using our peak detecting programs fdrPeak (18) for the ChIP-chip data and BELT (X. Lan et al., manuscript in preparation) for the ChIP-seq data, and binding motifs underneath of the binding loci using our ChIPMotifs (19) as well as cis-regulatory modules identified by our ChIPModules program (20). The database contains the nucleotide sequence for every indexed binding site. For experimentally verified binding sites, the sequence and coordinates in the paper are used when possible. In papers for which the sequence is not given, the coordinates given in the paper were used to obtain the sequence from the genome specified in the paper. For predicted binding sites, the sequence was taken from the genome used in the database for that species by the predictive algorithm used, after the coordinates from the paper were converted from the genome used in the paper to the version used in the database. For simplicity, the database only contains the annotation for one genome assembly per species. The human assembly used is HG18, and data was taken from papers in the HG18 and HG17 assemblies: HG17 data was converted to HG18 using UCSC's liftOver program. Mouse data is in the MM8 assembly, and rat data is in the RN4 assembly. The HRTBLDb contains AR, ER and GR in vitro data for all three species; however, all of the in vivo data is human. The total number of different data and sources by hormone receptor is shown in Figure 1.

DATABASE IMPLEMENTATION

The database is designed to have a close mapping to the real objects that it represents (Supplementary Data, and the web site http://motif.bmi.ohio-state.edu/hrtbldbabout#schema). One benefit of this is that entries in the database express the natural relationships between their real world counterparts. This was done to make complex queries to the database simpler and more intuitive. Since the database stores in vivo ChIP-based data, with computationally predicted binding motifs, and in vitro experimentally verified individual binding motifs, these data were split into two separate entities, binding loci and binding motifs. ChIP-based experiments produce many larger regions on the chromosome without fine interior details, aside from predicted binding sites; therefore, these regions are stored in a separate table, Loci. Both the predicted binding motifs within binding loci and the experimentally verified binding motifs are stored in the Binding Sites table. Each locus and binding site is also associated with an entry in the Experiments table which contains the meta-data common to all data from a given resource, reference information for the scientific paper, cell line, hormone receptor, etc. The database supports the separation of all data by species; however, for speed and efficiency, no meta-data on the species is stored in the database.

Binding loci have a one-to-many relationship with binding sites: i.e. a locus contains multiple binding sites, but no binding site is a part of more than one locus, if any. Overlapping loci from different experiments may share common predicted binding sites; however, these are represented as separate entities in the database. Each binding site is a collection of information about a single predicted or experimentally verified transcription factor binding site. They contain the coordinates, name and the exact sequence of that individual binding site. Loci and Binding are associated with a single ‘nearest gene’, which is the gene whose 5\'- or 3\'-end is nearest to the mid-point of that loci. Based on its proximity to this gene they are labeled with a region, which is a short description of its location relative to the gene, such as ‘within gene’, ‘5\'-promoter’, ‘3\'-distal’, etc. For the details of the distances from the gene that each region represents and how they are calculated, see Supplementary Data and our web site http://motif.bmi.ohio-state.edu/hrtbldbabout#regions. The publications which were used as data sources are recorded in the experiments table; however, there is a one-to-many mapping between source publications and experiment entries. Each experiment entry represents a combination of a publication, a species, a cell line and an experimental method. If one publication has datasets for which these characteristics differ, there will be one experiment entry for each dataset.

This database is implemented in MySQL using the MyISAM engine. All of the data are available for download as either a MySQL dump or as collection of tab separated text files, along with the database schema and the queries used to implement the schema, at the web site http://motif.bmi.ohio-state.edu/hrtbldb/downloads.
WEB INTERFACE AND DATA VISUALIZATION

We have developed a data visualization tool, the Portable Epigenetic Regulator Framework (PEGRF, Figure 2); it is a comprehensive, highly modular application for the storage of genomic and epigenomic data, data search and graphical display, all over the internet, through a web browser. The graphical rendering is done using the PHP light-weight regulator visualization tool (PLRVT), a module to PEGRF which we developed in tandem. Both are implemented entirely in the PHP scripting language and run on a web server. The primary benefits of this approach are that the application is both operating system and database agnostic. This results in PEGRF being highly portable across a wide variety of systems, and easily integrated into the existing web presence for many organizations (Supplementary Data).

There are multiple methods of searching through the data in the HRTBLDb. The simplest, ‘Basic Search’, is similar to the search engine of UCSC genome browser in that a location on the genome in the format chrX:0000–9999 is entered to view all data in the database within that region. Basic search also supports the display of all data near the transcription start site of a gene searched for by its gene symbol, RefSeq ID or its GenBank Gene ID. For example, a search results for the region of chr19:56047091–56052091 shows KLK3 (PSA) gene has both in vitro validated androgen response element and the predicted androgen response element from in vivo ChIP-chip AR-binding loci in the same location. The more complex search, ‘Advanced Search’, filters the data in the database by various criteria. The data elements from it are described in detail in a source specific table below the visualization in the order they are displayed. The data visualization also provides a way to visually browse the genome for target data by moving the region being displayed up or down the chromosome and zooming in and out to see more or less data (Figure 3).

DISCUSSION

The HRTBLDb is a comprehensive data management system for the aggregation, discovery and analysis of hormone receptor binding sites in the mammalian genomes including human, mouse and rat. Given that its primary purpose is to allow other researchers to build upon the information generated by high-throughput experiments, the current version contains 44,673 binding
elements with 114 hormone response elements which are verified in in vitro experiments; 75 binding motifs which occur with a hormone response element and whose co-regulatory action is verified in in vitro experiments; 18,472 binding loci from in vivo experiments; and 26,012 computationally predicted binding motifs. The first of several distinguishing features is that HRTBLDb is focused on a whole-genome-wide scale of HR regulatory information and classifies the binding sites based on their proximity to genes from RefGene databases, such as HG18 RefSeq, MM8 RefSeq and RN4 RefSeq (http://genome.ucsc.edu/). The importance of this feature is that it accommodates new conclusions from the ENCODE consortium (21) that transcription factor binding sites may be located anywhere in the human genome (See the Supplementary Data and the web site http://motif.bmi.ohio-state.edu/hrtbldbabout #regions for the classification scheme used). The second feature is that we not only collected the binding loci identified from in vivo high-throughput data, but also computationally predicted binding motifs which the TFs could possibly bind to and cis-regulatory modules for co-occurrent binding with other TFs. The third feature is that, to adequately verify the regulatory mechanisms at work on several genes in our database, we included direct and indirect binding patterns from in vitro EMSA and ChIP assay experiments. These patterns show regulatory action along with the reported motifs where available and publication references. For example, cathepsin D (CTSD) is reported to be involved with a direct binding estrogen response element in (22) and an endrogen response element and SP1 indirect binding in (23); therefore, this data was added to our database for comparison with in vivo binding data. In summary, the abundance of data in our HRTBLDb will hopefully contribute toward the progression of the ongoing investigation in to HR gene regulation and its mechanisms.

Although we have migrated all the data from our previous ERTargetDB (24) into our new database, there are significant differences between two databases: (i) HRTBLDb focuses on whole-genome-wide binding sites and visualization; whereas, ERTargetDB only contained promoter regions; (ii) The new database contains data for three major hormone receptors (ER, androgen receptor and GR). It contains binding loci from both in vivo ChIP-based high-throughput data and in vitro experimentally validated binding motifs from EMSA and ChIP assay; (iii) it will display all possible binding sites, different types of experiment, HRs and cell lines, in the queried region; (iv) the visualization is done by a modular binding element visualization tool we developed in tandem, the PHP light-weight regulator visualization tool (PLRVT) and ERTargetDB uses the GDVTK tool written in Java; and (v) the internal database schema are different as well in that the HRTBLDb is designed to be generic and store many different forms of DNA-binding elements for present and future research. We should also point out that our HRTBLDb is different from other similar publicly available databases, such as MRbase and NuReBase (25), which both contain the chemical and biological properties of hormones and their receptors, focus on all properties of nuclear receptors and their data are manually collected, hence these databases include a limited amount of information, but our database focuses on a subset of hormone receptors and their targets’ information and uses computational processing to predict additional binding motifs and transcriptional regulatory modules. This information can provide direction for future research which would likely be fruitful. Our software used to make these binding motif predictions are published on genome research, ChIPMotifs (19) for identified binding motifs as well as ChIPModules (20) for identified cis-regulatory modules. Ours is different from ERGDB (26) and KBERG (27) which were...
developed by the same group. Their databases only focus on ER responsive genes, and functional classifications, gene expression data. We think ChIP-based data with direct targets, can help determine transcriptional regulatory networks, therefore we created the HRTBLDb.

As part of the Integrative Cancer Biology Program of the National Cancer Institute (28), we have developed a ChIP-seq-based computational method to characterize the epigenetic mechanisms involved in the target genes of the ER pathway (J. Wu et al., manuscript in preparation). The generated ER target-binding loci have been deposited into this database. We will add epigenetic regulatory data such as DNA methylation and histone modification patterns across the loci for these HRs to the database in the future. Other hormone receptors data such as the progesterone receptor are also planned to be added to our database in the near future.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
The authors thank the members of their teams for discussion and comments.

FUNDING
National Institute of Health (grants U54CA113001 and R01CA069065). Funding for open access charge: Department of Biomedical Informatics, College of Medicine, The Ohio State University.

Conflict of interest statement. None declared.

REFERENCES
1. Parker,M.G. (1991) Nuclear Hormone Receptors. Academic Press, San Diego, CA.
2. Conzen,S.D. (2008) Nuclear receptors and breast cancer. Mol. Endocrinol., 22, 2215–2228.
3. Heemers,H.V.A.D.J.T. (2007) Androgen Receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex. Endcr. Rev., 28, 778-808.
4. Green,K.A. and Carroll,J.S. (2007) Oestrogen-receptor-mediated transcription and influence of co-factors and chromatin state. *Nat. Rev., Cancer*, 7, 713–722.
5. Pearce,S.T. and Jordan,V.C. (2004) The biological role of estrogen receptors alpha and beta in cancer. *Crit. Rev. Oncol. Hematol.*, 50, 3–22.
6. Quigley,C.A., De Bellis,A., Marschke,K.B., el-Awady,M.K., Wilson,E.M. and French,F.S. (1995) Androgen receptor defects: historical, clinical, and molecular perspectives. *Endocr. Rev.*, 16, 271–321.
7. Brinkmann,A.O. (2001) Molecular basis of androgen insensitivity. *Mol. Cell Endocrinol.*, 179, 105–109.
8. Moran,T.J., Gray,S., Mikosz,C.A. and Conzen,S.D. (2000) The glucocorticoid receptor mediates a survival signal in human mammary epithelial cells. *Cancer Res.*, 60, 867–872.
9. Read,J.T., Rahmani,M., Boroomand,S., Allahverdian,S., McManus,B.M. and Rennie,P.S. (2007) Androgen receptor regulation of the versican gene through an androgen response element in the proximal promoter. *J. Biol. Chem.*, 282, 31954–31963.
10. Chusacultanachai,S., Glenn,K.A., Rodriguez,A.O., Read,E.K., Gardner,J.F., Katzenellenbogen,B.S. and Shapiro,D.J. (1999) Analysis of estrogen response element binding by genetically selected steroid receptor DNA binding domain mutants exhibiting altered specificity and enhanced affinity. *J. Biol. Chem.*, 274, 23591–23598.
11. Malkoski,S.P. and Dorin,R.I. (1999) Composite glucocorticoid regulation at a functionally defined negative glucocorticoid response element of the human corticotropic-releasing hormone gene. *Mol. Endocrinol.*, 13, 1629–1644.
12. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. et al. (2000) Genome-wide location and function of DNA binding proteins. *Science*, 290, 2306–2309.
13. Weinmann,A.S. and Farnham,P.J. (2002) Identification of unknown target genes of human transcription factors through the use of chromatin immunoprecipitation. *Methods*, 26, 37–47.
14. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4, 651–657.
15. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, 129, 823–837.
16. Loh,Y.-H., Wu,Q., Chew,J.-L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. et al. (2006) The Otx1 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, 38, 431–440.
17. Deblois,G. and Giguère,V. (2008) Nuclear receptor location analyses in mammalian genomes: from gene regulation to regulatory networks. *Mol. Endocrinol.*, 22, 1999–2011.
18. Rabinovich,A., Jin,V.X., Rabinovich,R., Xu,X. and Farnham,P. (2008) E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res.*, 18, 1763–1777.
19. Jin,V.X., O’Gee,H., Iyengar,S., Green,R. and Rabinovich,P.J. (2007) Identification of cis-regulatory modules for Oct4 using de novo motif discovery and integrated computational genomics approaches. *Genome Res.*, 17, 807–817.
20. Jin,V.X., Rabinovich,A., Squazzo,S.L., Green,R. and Farnham,P.J. (2006) A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data—a case study using E2F1. *Genome Res.*, 16, 1585–1595.
21. ENCODE.P.C. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, 799–816.
22. Augereau,P., Miralles,F., Cavaillès,V., Gaudelet,C., Parker,M. and Rochefort,H. (1994) Characterization of the proximal estrogen-responsive element of human cathepsin D gene. *Mol. Endocrinol.*, 8, 693–703.
23. Krishnan,V., Wang,X. and Safe,S. (1994) Estrogen receptor-Sp1 complexes mediate estrogen-induced cathepsin D gene expression in MCF-7 human breast cancer cells. *J. Biol. Chem.*, 269, 15912–15917.
24. Jin,V.X., Sun,H., Pohar,T.T., Liyanarachchi,S., Palaniswamy,S.K., Huang,T.H. and Davuluri,R.V. (2005) ERTargetDB: an integral hormone receptor functional genomics information resource of transcription regulation of estrogen receptor target genes. *J. Mol. Endocrinol.*, 35, 225–230.
25. Ruau,D., Duarte,J., Ourgidal,T., Perriere,G., Laude,V. and Robinson-Rechavi,M. (2004) Update of NURBASE: nuclear hormone receptor functional genomics. *Nucleic Acids Res.*, 32, D165–D167.
26. Jiang,V., Han,H. and Baj,V. (2004) ERGDB: estrogen responsive genes database. *Nucleic Acids Res.*, 32, D533–D536.
27. Tang,S., Zhang,Z., Tan,S.L., Tang,M.-H.E., Kumar,A.P., Ramadoss,S.K. and Bajic,V.B. (2007) KBERG: knowledgebase for estrogen responsive genes. *Nucleic Acids Res.*, 35, D732–D736.
28. von Eschenbach,A.C. (2004) A vision for the national cancer program in the United States. *Nat. Rev. Cancer*, 4, 820–828.