BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 2 2013, pages 295–297
doi:10.1093/bioinformatics/bts693

Databases and ontologies

Advance Access publication November 29, 2012

NURBS: a database of experimental and predicted nuclear receptor binding sites of mouse

Yaping Fang[1], Hui-Xin Liu[2], Ning Zhang[3], Grace L. Guo[4], Yu-Jui Yvonne Wan[2] and Jianwen Fang[1,*]

[1]Applied Bioinformatics Laboratory, University of Kansas, Lawrence, KS 66047, USA, [2]Department of Medical Pathology and Laboratory Medicine, University of California, Davis Health Systems, Sacramento, CA 95817, USA, [3]Department of Biomedical Engineering, Tianjin University, Tianjin 300072, China and [4]Department of Pharmacology and Toxicology, Rutgers University, Piscataway, NY 08854, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: Nuclear receptors (NRs) are a class of transcription factors playing important roles in various biological processes. An NR often impacts numerous genes and different NRs share overlapped target networks. To fulfil the need for a database incorporating binding sites of different NRs at various conditions for easy comparison and visualization to improve our understanding of NR binding mechanisms, we have developed NURBS, a database for experimental and predicted nuclear receptor binding sites of mouse (NURBS). NURBS currently contains binding sites across the whole-mouse genome of 8 NRs identified in 40 chromatin immunoprecipitation with massively parallel DNA sequencing experiments. All datasets are processed using a widely used procedure and same statistical criteria to ensure the binding sites derived from different datasets are comparable. NURBS also provides predicted binding sites using NR-HMM, a Hidden Markov Model (HMM) model.

Availability: The GBrowse-based user interface of NURBS is freely accessible at http://shark.abl.ku.edu/nurbs/. NR-HMM and all results can be downloaded for free at the website.

Contact: jwfang@ku.edu

Received on October 5, 2012; revised on November 15, 2012; accepted on November 24, 2012

1 INTRODUCTION

Nuclear receptors (NRs) are a class of ligand-regulated transcription factors involved in many important biological processes, such as development and metabolism, and implicated in various human diseases, including diabetes, cardiovascular disease and cancers (Overington *et al.*, 2006). A single NR often impacts numerous genes, and different NRs may compete for target sites, resulting in overlapped target gene networks (Cotnoir-White *et al.*, 2011). Therefore, it is important to analyse and compare the binding sites and target genes of various NRs and determine the cross-talk between NRs. The recent development of chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-Seq) has greatly advanced the NR binding site detection (Johnson *et al.*, 2007). Consequently, data of experimental NR binding sites have been accumulated rapidly. Although there is limited effort to collect and archive NR-related ChIP-Seq data (Kennedy *et al.*, 2010; Tang *et al.*, 2011), to the best of our knowledge, no comprehensive database of NR binding sites across whole genome has been developed for easy comparing binding sites for different NRs.

It is well known that the binding sites of various NRs share a similar sequence pattern that contains two half sites with a consensus sequence of RGKTCA or its reverse complementary with variable intervals from 0 to 8 base pairs (Sandelin and Wasserman, 2005). There are three forms of such a pattern: direct repeat, inverted repeat, everted repeat and their reverse complement. Several methods have been used to predict NR binding sites (Bulyk, 2003; Cartharius *et al.*, 2005; Denver and Williamson, 2009; Grau *et al.*, 2006; Sandelin *et al.*, 2004; Varga, 2010). One widely used method is NHR-Scan (Sandelin and Wasserman, 2005), a Hidden Markov Model (HMM) using 107 experimentally determined NRs to predict two half sites simultaneously. However, NHR-Scan is no longer actively maintained, and its web-based interface does not allow predicting binding sites at a large scale. To take advantage of recently discovered binding sites and also to provide a standalone application for genome-scale response element prediction to the community, we have developed NR-HMM, a virtual basic application implementing an HMM model based on the NHR-Scan algorithm. The new model was trained on 151 experimentally verified binding sites.

We present NURBS, a database of binding sites across the whole-mouse genome of 8 NRs identified in 40 ChIP-Seq along with predicted binding sites of mouse genome using NR-HMM.

2 IMPLEMENTATION

The ChIP-Seq data used in NURBS are collected from NCBI (Kodama *et al.*, 2012). We regularly search the NCBI database for each NR by its nomenclature committee name, abbreviation and common name. At the time the manuscript was written, NURBS had 40 ChIP-Seq datasets for 8 NRs. The information regarding these datasets and corresponding NRs is summarized in an online table available in the NURBS website.

To allow comparing multiple datasets, we perform genome alignment and peak annotation for all collected datasets using

*To whom correspondence should be addressed.

a well established procedure. In brief, all sequenced reads are aligned to mm9 mouse reference genome (http://genome.ucsc.edu/) (Dreszer *et al.*, 2012) using bowtie (version 0.12.7) (Langmead *et al.*, 2009). All peaks are detected using the model-based analysis for ChIP-Seq (MACS v.1.4.1) (Zhang *et al.*, 2008). The Poisson *p*-value cut-off for peak detection is set to 10^{-5}. Overlapped peaks are split using Mali Salmon's Peak Splitter program (http://www.ebi.ac.uk/bertone/software.html). The peak annotation is done using R package ChIPpeakAnno (Zhu *et al.*, 2010).

NR-HMM was built using NR binding sites collected in previous work and JASPAR CORE (http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl) (Sandelin and Wasserman, 2005). The sequences of the collected binding sites were manually inspected, and only those with two half sites were kept. The final dataset contained 151 NR binding sites. The NR-HMM model is built using the framework of NHR-Scan (Sandelin and Wasserman, 2005). The parameters of HMM were estimated using Baum–Welch algorithm. The transmission probability matrix was estimated based on the frequency of each state in the training data. Laplace's rule pseudo count was applied to estimate the emission probability matrix. The NR-HMM is then used to predict NR binding sites in mouse genome (mm9), available as a GBrowse track in NURBS.

The NURBS is implemented as a MySQL relational database and uses the Generic Genome Browser (GBrowse) (Donlin, 2009) as its web browser based interface. GBrowse is user-friendly and highly customizable. Links to the Mouse Genome Informatics database (MGI, http://www.informatics.jax.org/) are available in NURBS wherever possible. Because MGI also uses GBrowse as its user interface, all MGI annotation trackers can be easily downloaded and enabled in NURBS.

The main search page of NURBS is highly configurable by users. Detailed examples are provided in an online help file. In brief, users can click track buttons to select datasets for visualization and comparison. The coverage option allows users to choose whether to display the binding region intensities. NURBS supports searches by name and region. A search result page displays selected tracks of experimental data, predicted peaks by NR-HMM and annotations. A hypertext link is available for a peak, leading to additional information, such as sequence and genome location. Similarly, when an annotated feature such as a gene is visible in the result page, a hyperlink to the MGI database is provided so the users can access more information such as genetic map position, mammalian homology information, gene ontology and expression, as well as links to other public databases such as Ensembl genome browser, University of California at Santa Cruz (UCSC) genome browser and NCBI.

3 CONCLUSION

We have developed NURBS, a web-based database for experimental and predicted NR binding sites in mouse genome. It has a customizable and user-friendly interface easy for navigating, searching and comparing experimental and predicted NR binding sites for multiple NRs. All the data and the HMM model are freely available for download. Currently, we are incorporating transcriptome data in NURBS for more advanced studies and expanding the database to human NRs for cross-species comparison. We intend to make NURBS a database open to the community and encourage users to provide feedback and submit new data and references. We have been actively populating the database and plan to maintain regular updates for the years to come.

NURBS is distinct from other existing NR-related databases. It is developed to provide researchers a convenient way to compare experimental and predicted binding sites for various NRs, along with genomic annotations. Other existing NR databases are devoted to the sequences, structures and functions (Martinez *et al.*, 1998), mutations (Van Durme *et al.*, 2003; Vroling *et al.*, 2012) and phylogenies (Ruau *et al.*, 2004) of NRs, not their binding sites. Recently, Ochsner *et al.* (2012) set-up a database, Transcriptomine, which is focused on the expression and function of NRs-related genes. The well developed Cistrome (Tang *et al.*, 2011) provides information about a number of NRs and their co-factors and epigenomic information, and it is better suited for investigating individual NR, as it does not provide features for comparison between NRs. In addition, Cistrome dedicates genomic visualization to remote UCSC genome browser, whereas NURBS uses an integrated GBrowse, also used by MGI.

ACKNOWLEDGEMENT

The authors thank Dr Shan Gao for his assistance.

Funding: National Institutes of Health (DK092100 and CA53596 to Y.W. and DK090036 to G.G.).

Conflict of Interest: none declared.

REFERENCES

Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.

Cartharius,K. *et al.* (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.

Cotnoir-White,D. *et al.* (2011) Evolution of the repertoire of nuclear receptor binding sites in genomes. *Mol. Cell. Endocrinol.*, **334**, 76–82.

Denver,R.J. and Williamson,K.E. (2009) Identification of a thyroid hormone response element in the mouse Kruppel-like factor 9 gene to explain its postnatal expression in the brain. *Endocrinology*, **150**, 3935–3943.

Donlin,M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, **28**, 9.9.1–9.9.25.

Dreszer,T.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.

Grau,J. *et al.* (2006) VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees. *Nucleic Acids Res.*, **34**, W529–W533.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kennedy,B.A. *et al.* (2010) HRTBLDb: an informative data resource for hormone receptors target binding loci. *Nucleic Acids Res.*, **38**, D676–D681.

Kodama,Y. *et al.* (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Martinez,E. *et al.* (1998) The Nuclear Receptor Resource: a growing family. *Nucleic Acids Res.*, **26**, 239–241.

Ochsner,S.A. *et al.* (2012) Transcriptomine, a web resource for nuclear receptor signaling transcriptomes. *Physiol. Genomics*, **44**, 853–863.

Overington,J.P. *et al.* (2006) Opinion—how many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.

Ruau,D. *et al.* (2004) Update of NUREBASE: nuclear hormone receptor functional genomics. *Nucleic Acids Res.*, **32**, D165–D167.
Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
Sandelin,A. and Wasserman,W.W. (2005) Prediction of nuclear hormone receptor response elements. *Mol. Endocrinol.*, **19**, 595–606.
Tang,Q. *et al.* (2011) A comprehensive view of nuclear receptor cancer cistromes. *Cancer Res.*, **71**, 6940–6947.
Van Durme,J.J. *et al.* (2003) NRMD: Nuclear Receptor Mutation Database. *Nucleic Acids Res.*, **31**, 331–333.
Varga,G. (2010) Target gene identification via nuclear receptor binding site prediction. *Methods Mol. Biol.*, **674**, 241–249.
Vroling,B. *et al.* (2012) NucleaRDB: information system for nuclear receptors. *Nucleic Acids Res.*, **40**, D377–D380.
Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
Zhu,L.J. *et al.* (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, **11**, 237.