D188–D194 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 27 November 2012*
doi:10.1093/nar/gks1165

CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization

Jesse D. Ziebarth[1,2], Anindya Bhattacharya[1,2] and Yan Cui[1,2,*]

[1]Department of Microbiology, Immunology and Biochemistry and [2]Center for Integrative and Translational Genomics, University of Tennessee Health Science Center, Memphis, TN 38163, USA

Received September 14, 2012; Revised October 26, 2012; Accepted October 28, 2012

ABSTRACT

CTCF is a highly conserved transcriptional regulator protein that performs diverse functions such as regulating gene expression and organizing the 3D structure of the genome. Here, we describe recent updates to a database of CTCF-binding sites, CTCFBSDB (http://insulatordb.uthsc.edu/), which now contains almost 15 million CTCF-binding sequences in 10 species. Since the original publication of the database, studies of the 3D structure of the genome, such as those provided by Hi-C experiments, have suggested that CTCF plays an important role in mediating intra- and inter-chromosomal interactions. To reflect this important progress, we have integrated CTCF-binding sites with genomic topological domains defined using Hi-C data. Additionally, the updated database includes new features enabled by new CTCF-binding site data, including binding site occupancy and the ability to visualize overlapping CTCF-binding sites determined in separate experiments.

INTRODUCTION

The CCCTC-binding factor, CTCF, is a ubiquitously expressed transcriptional regulator protein that is highly conserved from fly to man (1,2). It was first identified as a transcriptional repressor of the MYC oncogene (3,4) and, subsequently, has been shown to be involved in an extraordinarily diverse set of regulatory functions including transcriptional activation, imprinting, X-chromosome activation and acting as an enhancer-blocking and/or barrier insulator-binding protein (2). A few years ago, several groups attempted to better characterize CTCF function by identifying human and mouse CTCF-binding sites genome wide using both experimental and computational methods (5–8). These studies focused on CTCF's role as an insulator-binding protein, finding that CTCF-binding sites were detected between active and silent chromatin domains (7) and that the expression of neighboring genes separated by predicted CTCF-binding sites is less correlated than random sets of neighboring genes (6). Additionally, these datasets of CTCF-binding sites were used to establish that, while many functional CTCF-binding sites do not match a consensus motif (9), there is a CTCF-binding site motif that is highly conserved in vertebrates (5). Initial consensus CTCF-binding site motifs were then used to computationally predict CTCF-binding sites (5,6). Within this context, we introduced the first public database of CTCF-binding sites, CTCFBSDB, in 2007 (10). The initial version of CTCFBSDB contained 34 420 experimental and 18 905 predicted CTCF-binding sequences and integrated these sites with functional annotations and gene expression profiles to examine how the binding sites may provide insulator function.

Since the introduction of CTCFBSDB, there have been many significant developments in understanding the role of CTCF. To a large extent, these developments have focused on how CTCF functions as the 'master weaver' of the genome by establishing the long-range intra- and inter-chromosomal contacts between chromatin fibers that organize the genome in three dimensions (2,9). In addition to CTCF being responsible for long-range interactions at specific loci such as β-globin, *H19* ICR and MHC-II (2), CTCF-binding sites have been connected to several key observations from Hi-C experiments that provide genome-wide 3D maps of chromatin interactions (11,12). Specifically, CTCF-binding sites were found to be significantly overrepresented both on Hi-C fragments that had a large number of long-range interactions (13) and at the boundaries of the topological domains that spatially organize the genome (12). In parallel with this changing understanding of the importance of CTCF, there has been remarkable growth in the number of experimentally identified CTCF-binding sites. These new binding sites have been used to investigate the mechanism through which CTCF binds to DNA sequences, resulting in the identification of multi-part sequence motifs that bind to

[*]To whom correspondence should be addressed. Tel: +1 901 448 3240; Fax: +1 901 448 7360; Email: ycui2@uthsc.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

CTCF (14–16) and the suggestion that the degree of occupancy at a binding site may be related to the binding type and function at the site (17). In this article, we discuss improvements to the CTCFBSDB that reflect this recent progress in understanding the function of CTCF.

NEW FEATURES

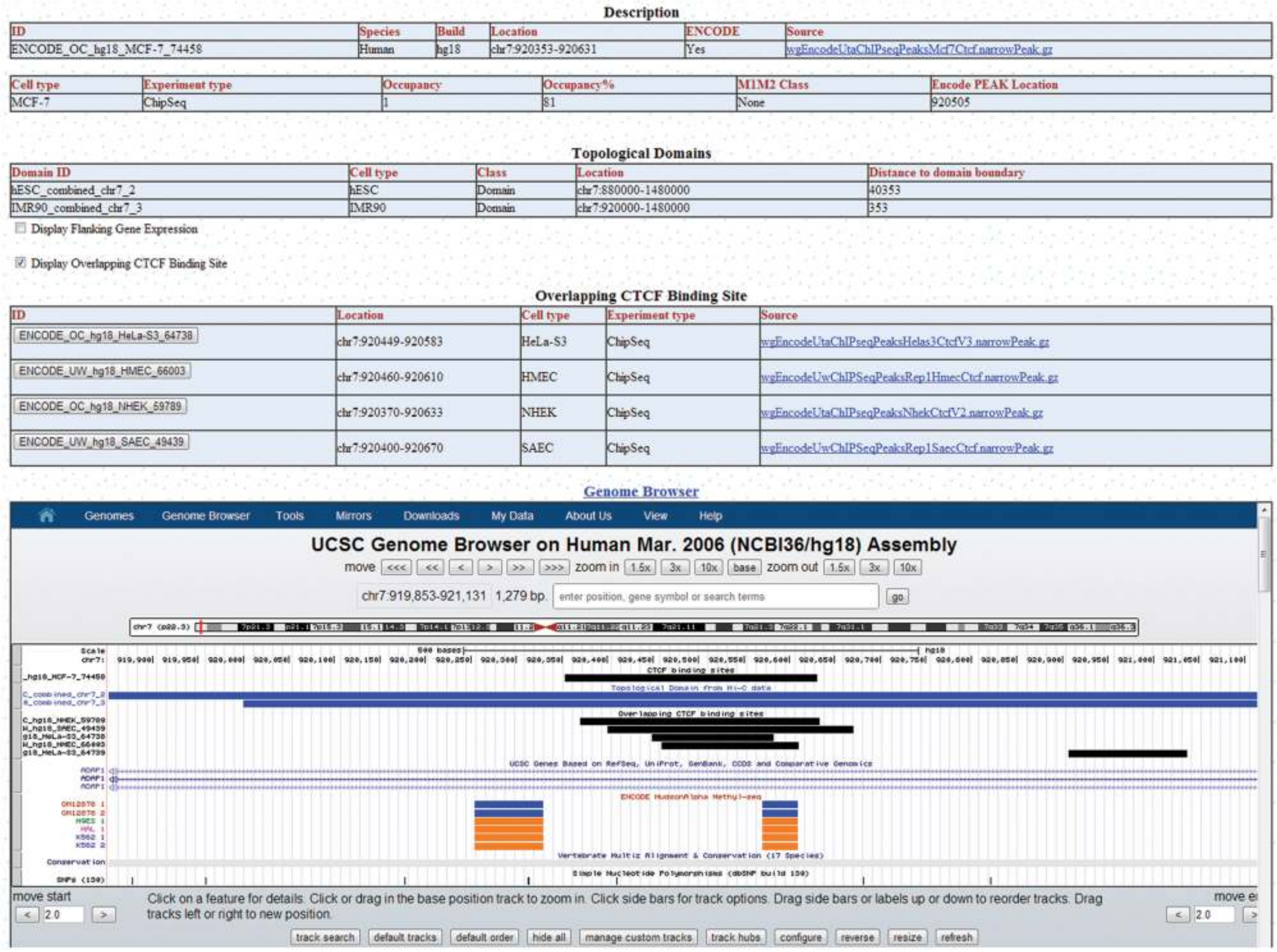
In addition to the significant expansion in the number of binding sequences available in the database which will be discussed in the next section, we have modified the presentation of binding sites in the CTCFBSDB (Figure 1) to include several new features:

(i) Inclusion of genomic topological domains defined using Hi-C data: the boundaries of these domains are enriched for CTCF-binding sites (12). We calculated the distance from each CTCF-binding sequence to the nearest domain boundary to help identify binding sites that may function to organize these domains. We also allow users to browse the topological domains to identify CTCF-binding sites at the boundaries of specific domains.

(ii) Identification of CTCF-binding sequences that overlap a given CTCF-binding sequence: the database now contains CTCF-binding sites identified in many tissues and cell types in mice and humans, making it possible to investigate if CTCF binding is specific to a particular cell type or conserved and, potentially, help limit the location of a binding site to a more narrow range.

(iii) Inclusion of occupancy data: we display the occupancy of the CTCF-binding site, when available. CTCF-binding site occupancy has been used to investigate both the potential for buffering of polymorphisms within binding sites (18) and how the CTCF-binding motif changes depending on the occupancy (17).

(iv) Classification of motif match type: recent analysis of the conservation of CTCF-binding sites across vertebrates has found that CTCF binding at many sites



Description

ID	Species	Build	Location	ENCODE	Source
ENCODE_OC_hg18_MCF-7_74458	Human	hg18	chr7:920353-920631	Yes	wgEncodeUtaChIPseqPeaksMcf7Ctcf.narrowPeak.gz

Cell type	Experiment type	Occupancy	Occupancy%	M1M2 Class	Encode PEAK Location
MCF-7	ChipSeq	1	81	None	920505

Topological Domains

Domain ID	Cell type	Class	Location	Distance to domain boundary
hESC_combined_chr7_2	hESC	Domain	chr7:880000-1480000	40353
IMR90_combined_chr7_3	IMR90	Domain	chr7:920000-1480000	353

☐ Display Flanking Gene Expression

☑ Display Overlapping CTCF Binding Site

Overlapping CTCF Binding Site

ID	Location	Cell type	Experiment type	Source
ENCODE_OC_hg18_HeLa-S3_64738	chr7:920449-920583	HeLa-S3	ChipSeq	wgEncodeUtaChIPseqPeaksHelas3CtcfV3.narrowPeak.gz
ENCODE_UW_hg18_HMEC_66003	chr7:920460-920610	HMEC	ChipSeq	wgEncodeUwChIPSeqPeaksRep1HmecCtcf.narrowPeak.gz
ENCODE_OC_hg18_NHEK_59789	chr7:920370-920633	NHEK	ChipSeq	wgEncodeUtaChIPseqPeaksNhekCtcfV2.narrowPeak.gz
ENCODE_UW_hg18_SAEC_49439	chr7:920400-920670	SAEC	ChipSeq	wgEncodeUwChIPSeqPeaksRep1SaecCtcf.narrowPeak.gz

Genome Browser

UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

chr7:919,853-921,131 1,279 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Figure 1. Screenshot of an example webpage for a CTCF-binding sequence (ENCODE_OC_hg18_MCF-7_744758) in CTCFBSDB 2.0. The database provides a description of the binding site, where the binding sequence is located within topological domains, and a Genome Browser viewer showing the genomic context of the binding site. Users also have the option to display the expression of genes flanking the binding site and CTCF-binding sequences that overlap the sequence. This CTCF-binding sequence, which was identified in MCF-7 cells, overlaps binding sequences that were identified in four other cell types.

can be understood in terms of a two-part motif in which each part interacts with distinct CTCF zinc fingers (16). We classify the CTCF-binding sequences based on how they match these motifs, allowing users to investigate the types of interactions that take place in the binding event.

(v) Integration with Genome Browser: the sequence context of each binding site in CTCFBSDB, including polymorphisms and DNA-methylation sites within the binding sequences, can be visualized in a Genome Browser viewer (19). Overlapping CTCF-binding sequences and topological domains are also displayed, facilitating the use of these new features.

DATABASE CONTENT

Sources of CTCF-binding sites

We expanded the CTCFBSDB using data from a variety of sources containing CTCF-binding sites determined using genome-wide experimental methods, and CTCFBSDB now contains 14 735 367 experimentally determined CTCF-binding sequences, including 13 760 124 human-binding sequences and 821 858 mouse-binding sequences. For human and mouse, the database contains CTCF-binding sequences identified in many cell types and experiments. Therefore, these sequences may include binding sites that have been repeatedly found in different cell types and experiments. We grouped overlapping binding sequences into CTCF-binding sequence clusters and identified 433 747 and 149 141 clusters in human (hg19) and mouse (mm9), respectively.

The sources of binding sites collected in the database include six published articles that utilized ChIP-Seq (16,20–24) and two articles using new ChIP-exo (15) and ChIA-PET (25) methods that have provided tens of thousands of CTCF-binding sequences in each of seven species (human, macaque, mouse, rat, dog, opossum and chicken). Additionally, we collected 145 human and 18 mouse CTCF-binding site datasets identified by the ENCODE project (26,27) that were publicly available as of 30 June 2012. Each CTCF-binding sequence in the database is identified by a prefix containing information about the data source appended to a number, creating a unique identifier for each binding sequence. For binding site datasets from ENCODE, the cell type and experimental treatment, if specified, were added to the end of the identifier prefix. A table containing a complete listing of the sources of the data in CTCFBSDB and the binding sequence identifier prefixes is provided on the database website ‘Help’ page.

CTCF-binding sites at topological domains boundaries

As technological advancements have enabled the study of how the genome is packaged into the nuclei of eukaryotes, they have consistently confirmed that there are strong links between the spatial organization of the genome and biological function (2,9). One of the most significant new experimental techniques that investigate the 3D structure of the genome is Hi-C, which was first applied to create a genome-wide map of chromatin interactions in a human lymphoblastoid cell line (11), and more recently, has been used to study mouse and human embryonic stem (ES) cells, mouse cortex and human IMR90 fibroblasts (12). A primary result of this later study is that the genome is organized into megabased-sized topological domains that occur throughout the genome and are conserved across different cell types and between mouse and human. Local chromatin interactions within a topological domain are common, while interactions between different domains or with boundary regions that separate domains are comparatively rare. While only 15% of CTCF-binding sites were located within boundary regions, there was a significant enrichment of CTCF-binding sites at domain boundaries (12), adding to the evidence that CTCF plays an important role in higher order genome organization.

To integrate CTCF-binding sites with topological domains, we downloaded 7947 human (hg18) and 8937 mouse (mm9) topological domains from the project website (http://chromosome.sdsc.edu/mouse/hi-c/download.html) of recent Hi-C experiments. The topological domains included in CTCFBSDB were determined for a bin size of 40 kb combined across multiple replicates of Hi-C interactions determined using the HindIII or NcolI restriction enzyme for human and mouse ES cells, mouse cortex and human IMR90 fibroblasts. We then determined if each CTCF-binding sequence in the hg18 or mm9 genomes was located within a topological domain or within a boundary region between domains and calculated the distance, in bp, between the edges of the binding sequence and the topological region.

Binding site motif classification

Due to the diversity in CTCF function, it has been suggested that different functions may be conferred by different CTCF-DNA-binding modes, potentially involving different combinations of interactions with the 11 zinc fingers that compose CTCF’s DNA-binding domain (1,2). Using genome-wide CTCF-binding sites determined in six mammalian species, Schmidt *et al.* (16) recently investigated this possibility by examining the binding site sequences and, agreeing with previous observations (14,15), delineated a multi-part CTCF-binding motif. They observed that, for the majority of CTCF–DNA-binding events, the N-terminal zinc fingers interact with a 14-bp long M1 motif. Additionally, in a subset of binding events, the C-terminal fingers interact with a shorter M2 interaction, creating a 34-bp-long M1+M2 motif. In the most common arrangement of sites containing M1+M2 motifs, the half-site distance between M1 and M2 was 21 or 22 bp. In order to classify the CTCF-binding sites based on the type of binding event, each binding sequence was scanned for matches to the M1 and M2 CTCF-binding motifs described by Schmidt *et al.* (16) and provided at http://www.ebi.ac.uk/∼schwalie/CTCFCell2012/ using the nmscan module of NestedMica (28) with a cutoff of −15. They were then classified as None (no M1 motif matches), M1 (the sequence matches the M1 motif), M1M2 (the sequence contains matches to the M1 and M2 motifs separated by a half-site distance of 12–42 bp) and M1M2_21_22 (the sequence contains matches to the M1 and M2 motifs that were separated by a half-site distance of 21 or 22 bp). Additionally, we included the position weight matrices of the M1 and M2 motifs in the CTCFBSDB Prediction Tool, which has been described previously (10), allowing users to scan query sequences for CTCF-binding site motifs.

Flanking gene expression

To investigate the potential for CTCF-binding sites to function as insulators, CTCFBSDB includes a comparison of the expression of the genes flanking each CTCF-binding site (Figure 2). In the original version of the database, this comparison was a heatmap image comparing microarray-based gene expression profiles from 61 mouse and 79 human tissues (29). We have maintained these microarray gene expression heatmaps in the updated version of the database, but present an additional figure containing RNA-Seq gene expression profiles determined in 10 human tissues (30). As the RNA-Seq data contain only 10 tissues, we display a column chart comparing the number of normalized Reads Per Kilobase of exon per Million mapped reads for the flanking genes of each CTCF-binding site. In CTCFBSDB 2.0, the microarray expression profiles are rendered using the BioHeatmap Javascript library (http://code.google.com/p/systemsbiology-visualizations/), whereas the RNA-Seq column charts use Google Visualization APIs (https://developers.google.com/chart/).

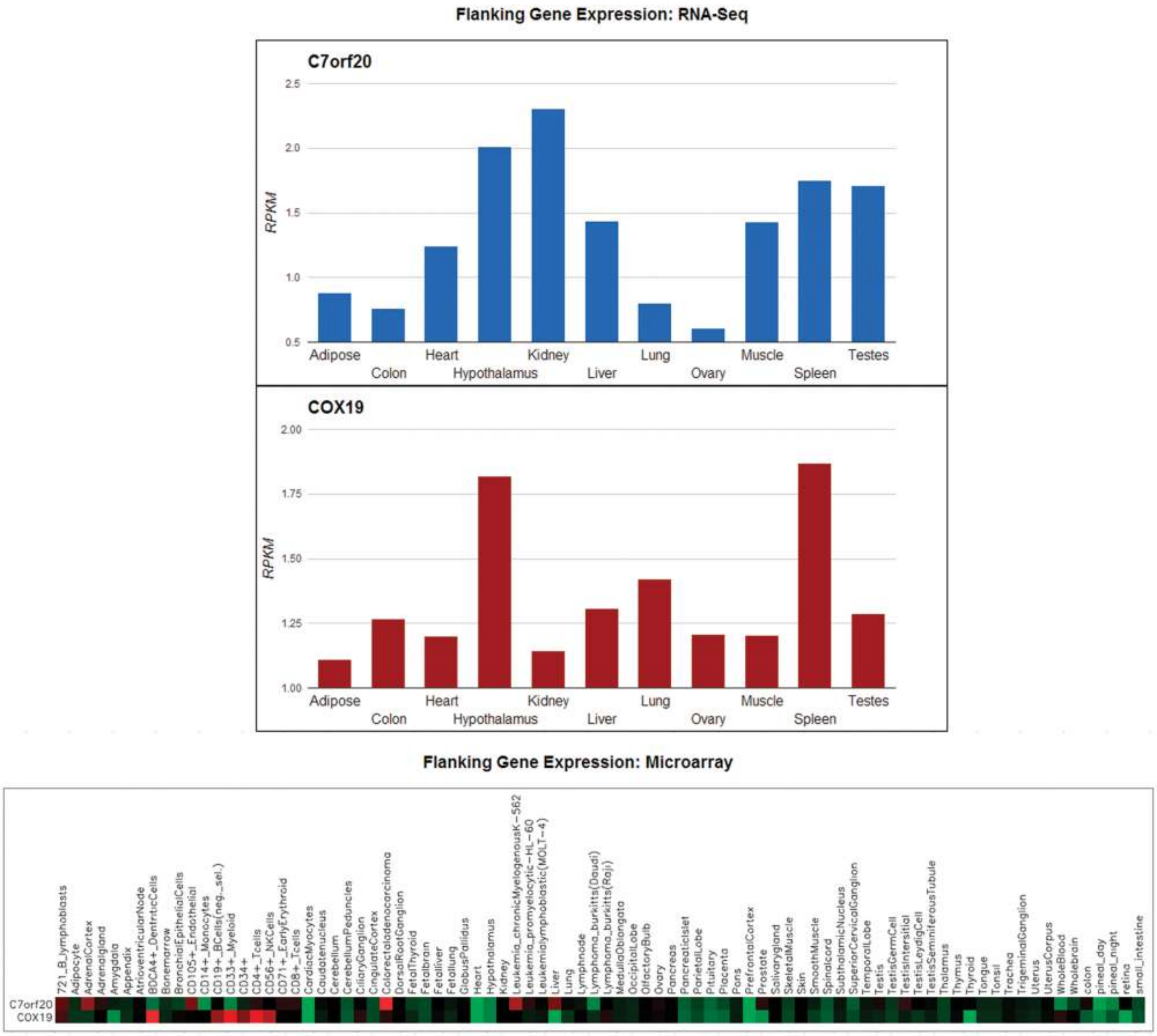


Figure 2. Gene expression profiles for genes flanking a CTCF-binding site (ENCODE_OC_hg18_MCF-7_744758). CTCFBSDB provides images comparing expression profiles identified using both RNA-Seq (top) and microarrays (bottom) for genes flanking the CTCF-binding site.

Table 1. Description of fields used to annotated CTCF-binding sites

Field name	Description
ID	Unique database identifier for the binding sequence
Species and build	The species and genomic build in which the binding site was determined
Location	Genomic location of the binding sequence
ENCODE	Whether or not the site was determined in an ENCODE dataset
Source	PubmedID or ENCODE accession number containing the binding site
Cell and experiment type	Experimental conditions in which the site was identified
Occupancy	Numerical value of the occupancy of the binding site reported in the original source
Occupancy%	Percentile of Occupancy within sites of the source dataset
M1M2 Class	Binding site motif class
ENCODE Peak location	Location of the ChIP-Seq peak of the binding site for ENCODE datasets

Additional CTCF-binding site annotation

In addition to identifying the topological domain location and binding motif type, each CTCF-binding sequence in the database is annotated with descriptions of the binding site and the experiment in which the site was identified (Table 1). Of particular interest among these annotations are two fields that show the occupancy of the binding site that was determined in the experiment. The first of these, ‘Occupancy’, provides a numeric value (i.e. read count for ChIP-Seq experiments or signal strength for ENCODE data) indicating the extent to which the binding site was occupied in the experiment, if available. As the values in the ‘Occupancy’ field had different scales for different experiments, we calculated the percentile of the occupancy value for each binding site within its dataset to allow for comparisons across experiments in the ‘Occupancy%’ field. A value of 99 in this field indicates that the binding site was in the top 1% of high occupancies within the dataset.

DATABASE USE AND ACCESS

Users can access CTCFDB through a variety of browse and search options. The contents of the database can be browsed through three tables containing experimentally identified CTCF-binding sequences, topological domains and computationally predicted CTCF-binding sites, respectively. The browseable experimental binding sequence table contains the unique CTCFBSDB identifier, which links to a page containing the full database record of the binding sequence, and a brief description of the binding site. This table can be filtered by species, cell type and chromosome, allowing users to quickly view relevant binding sites. The topological domain table can be filtered by species, cell type and chromosome and displays a unique database identifier for the topological domain or boundary and location of the domain. Clicking on the domain identifier presents a list of all CTCF-binding sequences located within the domain sorted by chromosome location. The predicted CTCF-binding site table remains unchanged from the first release of the database.

CTCFBSDB contains two options for searching the database. First, users can search for all binding sites in a species within a genomic range. Optionally, the search results can be filtered to present only binding sequences from a single data source or, due to the large percentage of database records that were collected from ENCODE project data causing ENCODE-binding sites to sometimes overwhelm search results, the search can be filtered to include all binding sequences, include only those binding sequences identified in the ENCODE project, or exclude ENCODE-binding sequences. Second, for quick access to previously investigated binding sequences, a keyword search can be used to search for a particular CTCFBSDB identifier.

Each experimental binding site in CTCFBSDB is presented on a webpage (Figure 1) that contains the following five sections: (i) Description: a table presenting a description of binding site, including the annotation information presented in Table 1; (ii) Topological Domains: a table presenting the domain identifier, type, location and distance from the binding sequence to the nearest edge of the domain boundary for the topological domains in which the binding site is located; (iii) Flanking Gene Expression: figures (Figure 2) comparing RNA-Seq and microarray expression profiles of the genes flanking the binding site; (iv) Overlapping CTCF-Binding Sites: a table containing CTCF-binding site sequences that overlap this sequence and (v) Genome Browser: a Genome Browser viewer (19) that displays the genomic context of the binding site, including UCSC genes, SNPs and custom tracks for the binding site and topological domains. Additionally, as methylation at CTCF-binding sites has been shown to impact CTCF binding (31,32), we display methylation tracks provided by the ENCODE project for human genome (the ENC DNA Methyl track for hg19 and the HAIB Methyl-seq and HAIB Methyl27 tracks for hg18) in the Genome Browser viewer, allowing users to quickly identify methylation sites within CTCF-binding sequences. By default, the flanking gene expression figure and overlapping binding site table are hidden, but can quickly be displayed by selecting a clearly labeled box. Displaying the overlapping CTCF-binding site table automatically adds a custom track to the Genome Browser containing these sites, allowing for visualization of the extent to which the binding site sequence overlaps other CTCF-binding sequences identified in other cell types or experiments.

DISCUSSION AND FUTURE DIRECTIONS

Updates made in version 2.0 of the CTCFBSBD reflect significant advances in both the number of known

CTCF-binding sites and the function of CTCF. In addition to a 250-fold increase in the binding site sequences included in CTCFBSDB, the database now integrates new data describing the details of the binding site (i.e. binding site occupancy, motif match type and location within topological domain), which can potentially be used to investigate the function of specific binding sites. With the large number of experiments that have determined CTCF-binding sites, it is likely that the majority of binding sites in the mouse and human genomes have already been identified. A next step in understanding the function of CTCF is determining if and how specific features of these binding sites allow CTCF to perform its diverse functions. The CTCFBSDB has the potential to be particularly useful to this effort, as it may not require the identification of new binding sites, but, instead, can be based on analysis of known binding sites. For example, data contained in the database can be used to compare binding sites located at the boundaries of topological domains with those in the domain centers to determine the characteristics that distinguish these types of binding sites.

In the future, the utility of the CTCFBSDB can be improved in several ways. The results of Hi-C and similar experiments will continue to increase understanding of the 3D structure and the role that CTCF plays in organizing this structure. It is likely that some data generated by these studies can be integrated with the CTCFBSDB, similar to how we have included the locations of topological domains. Additionally, while CTCF has been shown to interact with several other proteins (9), such as cohesin (33–35), the interactions between CTCF and these other proteins are not completely understood. As more binding sites of cohesin or other proteins that interact with CTCF are identified, these binding sites can be integrated into the CTCFBSDB, adding new data that can be used to determine the function of a binding site.

FUNDING

The UT Center for Integrative and Translational Genomics; Department of Defense [W81XHW-05-01-0227]. Funding for open access charge: Department of Defense [W81XHW-05-01-0227].

Conflict of interest statement. None declared.

REFERENCES

1. Ohlsson,R., Renkawitz,R. and Lobanenkov,V. (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.*, **17**, 520–527.
2. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
3. Lobanenkov,V.V., Nicolas,R.H., Adler,V.V., Paterson,H., Klenova,E.M., Polotskaja,A.V. and Goodwin,G.H. (1990) A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5′-flanking sequence of the chicken c-myc gene. *Oncogene*, **5**, 1743–1753.
4. Filippova,G.N., Fagerlie,S., Klenova,E.M., Myers,C., Dehner,Y., Goodwin,G., Neiman,P.E., Collins,S.J. and Lobanenkov,V.V. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.*, **16**, 2802–2813.
5. Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenkov,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.
6. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.
7. 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.
8. Mukhopadhyay,R., Yu,W., Whitehead,J., Xu,J., Lezcano,M., Pack,S., Kanduri,C., Kanduri,M., Ginjala,V., Vostrov,A. *et al.* (2004) The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res.*, **14**, 1594–1602.
9. Ohlsson,R., Lobanenkov,V. and Klenova,E. (2010) Does CTCF mediate between nuclear organization and gene expression? *Bioessays*, **32**, 37–50.
10. Bao,L., Zhou,M. and Cui,Y. (2008) CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res.*, **36**, D83–D87.
11. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
12. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
13. Botta,M., Haider,S., Leung,I.X., Lio,P. and Mozziconacci,J. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol.*, **6**, 426.
14. Boyle,A.P., Song,L., Lee,B.K., London,D., Keefe,D., Birney,E., Iyer,V.R., Crawford,G.E. and Furey,T.S. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, **21**, 456–464.
15. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
16. Schmidt,D., Schwalie,P.C., Wilson,M.D., Ballester,B., Goncalves,A., Kutter,C., Brown,G.D., Marshall,A., Flicek,P. and Odom,D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.
17. Essien,K., Vigneau,S., Apreleva,S., Singh,L.N., Bartolomei,M.S. and Hannenhalli,S. (2009) CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol.*, **10**, R131.
18. Maurano,M.T., Wang,H., Kutyavin,T. and Stamatoyannopoulos,J.A. (2012) Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.*, **8**, e1002599.
19. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
20. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
21. Cuddapah,S., Jothi,R., Schones,D.E., Roh,T.Y., Cui,K. and Zhao,K. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.
22. Martin,D., Pantoja,C., Fernandez Minan,A., Valdes-Quezada,C., Molto,E., Matesanz,F., Bogdanovic,O., de la Calle-Mustienes,E., Dominguez,O., Taher,L. *et al.* (2011) Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the

identification of disease-associated genes. *Nat. Struct. Mol. Biol.*, **18**, 708–714.

23. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
24. Kunarso,G., Chia,N.Y., Jeyakani,J., Hwang,C., Lu,X., Chan,Y.S., Ng,H.H. and Bourque,G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, **42**, 631–634.
25. Handoko,L., Xu,H., Li,G., Ngan,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
26. The ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.
27. The ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
28. Down,T.A. and Hubbard,T.J. (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res.*, **33**, 1445–1453.
29. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
30. Krupp,M., Marquardt,J.U., Sahin,U., Galle,P.R., Castle,J. and Teufel,A. (2012) RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, **28**, 1184–1185.
31. Bell,A.C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.
32. Shukla,S., Kavak,E., Gregory,M., Imashimizu,M., Shutinoski,B., Kashlev,M., Oberdoerffer,P., Sandberg,R. and Oberdoerffer,S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, **479**, 74–79.
33. Wendt,K.S., Yoshida,K., Itoh,T., Bando,M., Koch,B., Schirghuber,E., Tsutsumi,S., Nagae,G., Ishihara,K., Mishiro,T. *et al.* (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, **451**, 796–801.
34. Rubio,E.D., Reiss,D.J., Welcsh,P.L., Disteche,C.M., Filippova,G.N., Baliga,N.S., Aebersold,R., Ranish,J.A. and Krumm,A. (2008) CTCF physically links cohesin to chromatin. *Proc. Natl Acad. Sci. USA*, **105**, 8309–8314.
35. Hou,C., Dale,R. and Dean,A. (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc. Natl Acad. Sci. USA*, **107**, 3651–3656.