*Published online 2 November 2007* *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D83–D87***
*doi:10.1093/nar/gkm875*

# CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators

**Lei Bao, Mi Zhou and Yan Cui***

Department of Molecular Sciences, Center of Genomics and Bioinformatics, University of Tennessee Health Science Center, 858 Madison Avenue, Memphis, TN 38163, USA

Received August 9, 2007; Revised September 30, 2007; Accepted October 1, 2007



## ABSTRACT

**Recent studies on transcriptional control of gene expression have pinpointed the importance of long-range interactions and three-dimensional organization of chromatins within the nucleus. Distal regulatory elements such as enhancers may activate transcription over long distances; hence, their action must be restricted within appropriate boundaries to prevent illegitimate activation of non-target genes. Insulators are DNA elements with enhancer-blocking and/or chromatin-bordering functions. In vertebrates, the versatile transcription regulator CCCTC-binding factor (CTCF) is the only identified *trans*-acting factor that confers enhancer-blocking insulator activity. CTCF-binding sites were found to be commonly distributed along the vertebrate genomes. We have constructed a CTCF-binding site database (CTCFBSDB) to characterize experimentally identified and computationally predicted CTCF-binding sties. Biological knowledge and data from multiple resources have been integrated into the database, including sequence data, genetic polymorphisms, function annotations, histone methylation profiles, gene expression profiles and comparative genomic information. A web-based user interface was implemented for data retrieval, analysis and visualization. *In silico* prediction of CTCF-binding motifs is provided to facilitate the identification of candidate insulators in the query sequences submitted by users. The database can be accessed at http://insulatordb.utmem.edu/**

## INTRODUCTION

CCCTC-binding factor (CTCF) is a versatile transcription regulator that is evolutionarily conserved from fruit fly to human (1). CTCF binds to different DNA sequences by combinatorial use of 11-zinc fingers and plays a key role in many chromatin insulation events [reviewed in (2)]. In eukaryotic genomes, chromatins are organized into distinct domains. The chromatin domain architecture is critical for transcription control. Insulators are the key DNA sequence elements that establish and maintain such domain boundaries (2–12). They represent a class of diverged DNA sequences capable of shielding genes from inappropriate *cis*-regulatory signals from the genomic neighborhood. There are two types of insulators—enhancer-blocking insulators that block enhancer–promoter communication and barrier insulators that protect against heterochromatin-mediated silencing (13). Many recent studies have been devoted to the identification and characterization of insulators. CTCF-binding site is of particular interest because CTCF is the only protein identified so far in vertebrate that binds to enhancer-blocking insulators and shows enhancer-blocking activity. Recent studies also linked the CTCF-binding site to epigenetic processes, such as imprinting (14–16), X-chromosome inactivation (17,18) and interchromosomal colocalization (19). Despite their obvious importance, to our knowledge, there is no public database categorizing this type of regulatory elements. In addition to dozens of well-characterized CTCF-binding sites with validated insulation functions that are scattered in the biomedical literature, several recent high-throughput ChIP-chip analyses and comparative genomic studies (20–23) have identified tens of thousands of potential CTCF-binding sites in human and mouse genomes. Here we report our effort in creating a CTCF-binding site database, a collection of experimentally identified and computationally predicted CTCF-binding sites. Biological knowledge and data from multiple resources were integrated to annotate the CTCF-binding sites. The database is designed to facilitate the studies on insulators and their roles in regulating gene expression and demarcating functional genomic domains.

*To whom correspondence should be addressed. Tel: +1 9014483240; Fax: +1 9014487360; Email: ycui2@utmem.edu

## DATA SOURCES AND PROCESSING

### Data sources

Experimentally identified and computationally predicted CTCF-binding sites are processed separately. First, 34 417 experimentally identified CTCF-binding sites are collected from four sources: (i) 110 manually curated CTCF-binding sites from biomedical literature, denoted by identifiers starting with 'INSUL_MAN', (ii) 244 mouse CTCF-binding sites identified by Ohlsson and coworkers using ChIP-chip assay (21), denoted by identifiers starting with 'INSUL_OHL', (iii) 13 801 human CTCF-binding sites identified by Ren and coworkers using ChIP-chip assay (20), denoted by identifiers starting with 'INSUL_REN' and (iv) 20 262 human CTCF-binding sites identified by Zhao and coworkers using massive direct sequencing of ChIP DNA (23), denoted by identifiers starting with 'INSUL_ZHAO'. Second, we collected the conserved CTCF-specific sequence motifs (~20 bp) in the human and mouse genomes that were predicted using motif scan (20,22). We excluded those (~40%) overlapped with any of the experimentally determined CTCF-binding sites. The resulting 18 905 entries include 7736 human and 5504 mouse CTCF-binding sites predicted in (20) and 5665 human CTCF-binding sites predicted in (22). The computationally predicted CTCF-binding sites have identifiers beginning with 'INSUL_PRE'.

### Annotation of the CTCF-binding sites

Table 1 shows the major data fields of the database. For the 110 manually curated CTCF-binding sites, we used a set of controlled vocabularies to describe their properties. The 'Validation Method' field specifies whether CTCF binding was validated by *in vitro* and/or *in vivo* assays and whether this CTCF-binding sequence showed enhancer-blocking function in transgenic experiment (24). The '*In situ* Function' field annotates the biological roles of a CTCFBS in its natural genomic context (enhancer-blocking, chromatin boundary, etc.). The 'Description' field contains other features of the CTCF-binding site (e.g. methylation-sensitivity of the CTCF binding). Genomic coordinates of the CTCF-binding sequences were determined using the BLAT alignment program (25). The assemblies of genomes used are hg18 for human, mm8 for mouse, rn3 for rat and galGal2 for chicken. CTCF-binding sequences without chromosome location information usually mean that they were probably mapped to unsequenced portions (e.g. heterochromatic regions) of the genome (21).

### Sequence features of the CTCF-binding sites

The sequences of CTCF-binding sites in the database vary from 20 bp to several hundred bp. There are two reasons for this length heterogeneity. First, different experimental methods may have different basepair resolutions for locating CTCF-binding sites. Second, different laboratories often have different research goals when they publish the original sequences. Some researchers may stop at a 500-bp region encapsulating the CTCF-binding sites while others may further narrow down to the sequences covered by the CTCF protein physically. Most CTCF-binding sites were found to share a 20-bp motif (20), which is highlighted using consecutive arrows. The direction of the arrows shows the genomic orientation of the motif. We also highlighted all the single nucleotide polymorphisms (SNP) in the dbSNP database (26) that are located in a CTCF-binding site using vertical indicators. Mutations that disrupt CTCF-binding sites may lead to abnormal gene expression and cause diseases. Indeed, a recent study showed that inherited mutations that abolish CTCF-binding sites in the human H19 differentially methylated region (DMR) can cause Beckwith–Wiedemann syndrome (27,28). Thus, the naturally occurring mutations in CTCF-binding sites may represent new types of genetic variations that underlie phenotypes including disease status. To get more information about any of the SNPs, the user can click the SNP indicator to browse the corresponding dbSNP webpage (26).

**Table 1.** Description of the fields

| Field name | Description |
|---|---|
| ID[a] | Unique identifier of an entry |
| Species[a] | Species name |
| Name | Name used by the authors in the original paper |
| Chromosome location | CTCF-binding site position |
| Orientation | Forward (+) or reverse (−) strand |
| 5′-Flanking gene | 5′-Flanking gene of the CTCF-binding site along the genome |
| 3′-Flanking gene | 3′-Flanking gene of the CTCF-binding site along the genome |
| Validation method[a] | The validation methods including *in vitro* binding, *in vivo* binding, enhancer-blocking assay and sequence analysis |
| *In situ* function | *In situ* function of the CTCF-binding site |
| Description | Other important features of the CTCF-binding site |
| Reference[a] | PubMed reference |
| Sequence[a] | DNA sequence of the CTCF-binding site |

[a]A mandatory field.

### Genomic context track

The genomic context of a CTCF-binding site provides clues for its *in situ* functions. The CTCF-binding site (red) and flanking genes within 100 kb distance are displayed using the UCSC genome browser (29) (Figure 1). Other CTCF-binding sites located in this genomic region are also displayed and different colors are used to distinguish the sources of the CTCF-binding sites: yellow for INSUL_MAN, blue for INSUL_OHL, green for INSUL_REN, cyan for INSUL_ZHAO and black for INSUL_PRE. An important function of CTCF-bound insulators is to demarcate transcriptionally active and silent chromatin domains, which are marked by distinct histone methylation patterns. A recent study provided high-resolution maps of histone methylations (chromatin domains) in the human genome (23). H3K4 trimethylation (H3K4me3) and H3K27 trimethylation

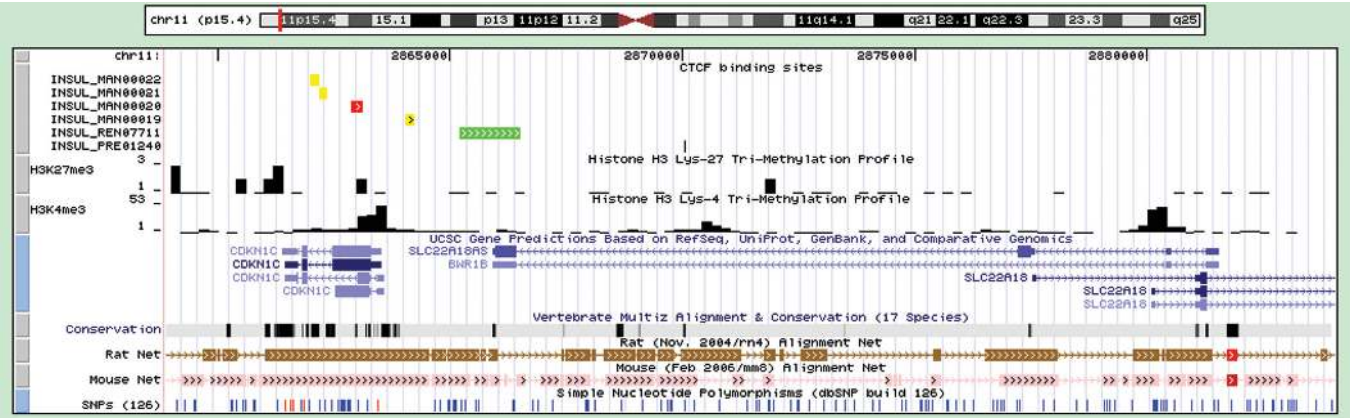


**Figure 1.** The genomic context of a few CTCF-binding sites. The CTCF-binding sites reside at the boundary between the two histone methylation domains (H3K4me3 and H3K27me3).

(H3K27me3) are a pair of 'Yin-Yang' modifications with high level of H3K4me3 and H3K27me3, representing gene activation and silencing, respectively (23). We integrated H3K4me3 and H3K27me3 maps with our genomic context track of CTCF-binding sites using the genome browser to facilitate the utilization of this valuable information.

### Flanking gene expression track

Another *in situ* function of insulators is to maintain independent expression patterns of neighboring genes. Suppose there is a tissue-specific enhancer that should control the transcription of one gene but not that of the other in a pair of neighboring genes. The CTCF-binding site located between the enhancer and the promoter of the second gene may function as enhancer-blocking insulator to protect against illegitimate transcriptional activation. In this scenario, the neighboring genes may have very different expression status in that tissue. We created a flanking gene expression track to compare the expression patterns of the genes flanking the CTCF-binding site. The data were obtained from The Genomics Institute of the Novartis Research Foundation (GNF) Gene Expression Atlas 2 (30), which contains genome-wide gene expression profiles of 61 mouse tissues and 79 human tissues. The raw data was log-transformed (base 2) and normalized to have a mean of 0 and SD of 1. The expression images were created using the Slcview software (http://slcview.stanford.edu), in which red indicates overexpression and green indicates underexpression. An example of gene expression track is shown in Figure 2.

### Mammalian orthologous region track

Comparative genomic studies on human, mouse and rat may provide insights into the evolution of CTCF-binding sites. To this end, we created a track of mammalian orthologous regions. For any of the three genomes, the regions containing CTCF-binding sites and flanking genes were used to query orthologous regions in the other two genomes from the UCSC precomputed block chains (31,32). Only the DNA blocks with the maximal alignment score against the query region were retained as orthologous regions. The aligned orthologous sequences in up to 16 vertebrate genomes can be displayed by clicking the 'view alignment' button (Figure 2).

### CTCF-binding site prediction

CTCF uses different combinations of its zinc fingers to recognize divergent DNA sequences. Recent studies have identified core motifs for CTCFBS sequences (20,22). The motifs are represented by position weight matrices (PWM). Altogether, four closely related PWM have been derived to accommodate the sequence divergences in CTCF-binding sites (20,22). The database provides a simple web tool to search for the core CTCF-binding motifs in a query sequence. It uses the STORM program (33) to scan for each of the four PWM in the query sequences and reports the best hits.

## UTILITY AND DISCUSSION

First, a web interface was developed for browsing the experimentally identified and computationally predicted CTCF-binding sites. Users can focus on entries of interest using four selection controls—Species, Validation Method, *In Situ* Function and Description. The *in situ* function of most known CTCF-binding sites is to act as boundary element. However, in some biological contexts, CTCF-binding sites may also function as elements for transcription activation/repression [reviewed in (34)]. Second, a text search interface was developed for querying the database. Users can search for CTCF-binding sites by element name or by the PubMed identifier of the original literature. A useful approach is to retrieve the CTCF-binding sites contiguous to a gene of interest by entering an official gene symbol or words used in the gene description. Third, the database provides sequence similarity search (35) for the comparison

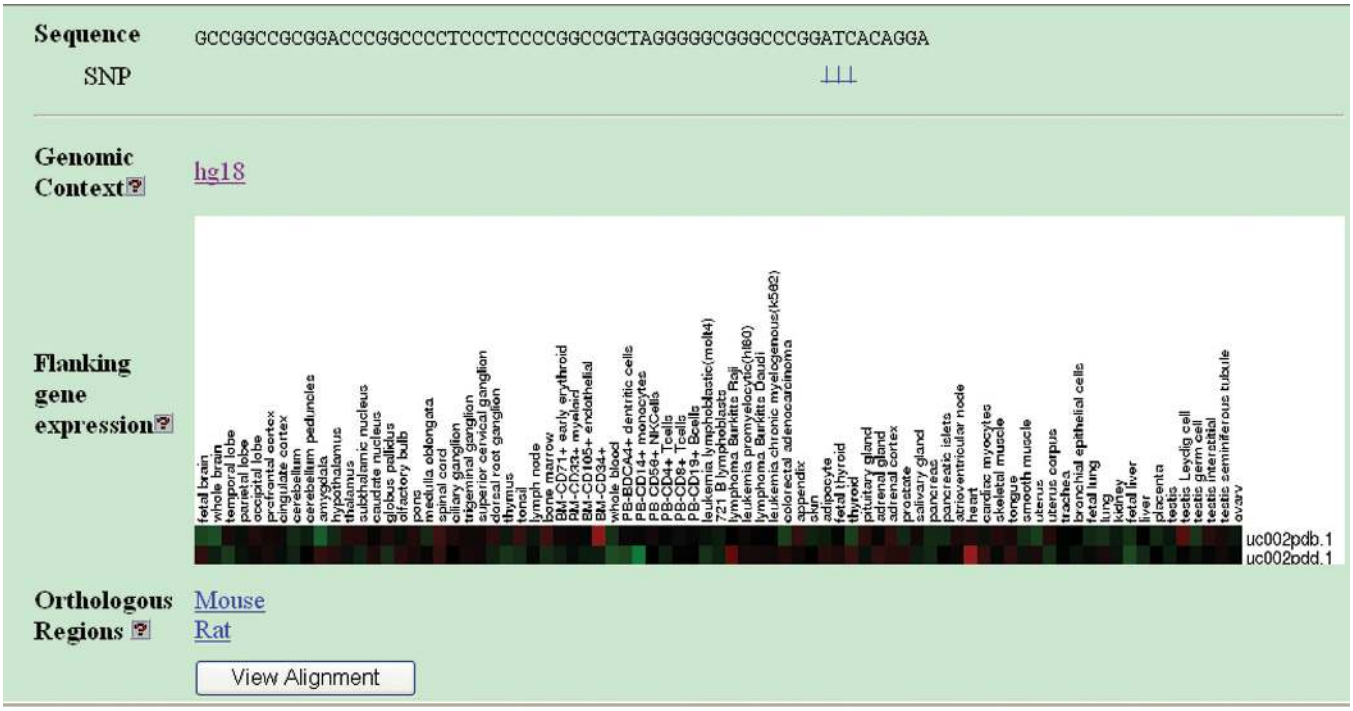


**Figure 2.** A CTCF-binding site (INSUL_MAN0004) webpage displays the flanking gene expression profiles and links to tracks of SNPs, genomic context and orthologous regions.

between query sequences and CTCF-binding sequences. Finally, an option of genomic range search is provided. Users can specify a genomic interval and retrieve all the CTCF-binding sites in the interval.

To maintain an up-to-date resource, we encourage researchers to submit newly identified CTCFBS sequences to the database. Data can be submitted directly through a web interface. The submissions will be manually checked before being added to the database.

The database is an integrative platform for storing, retrieving and characterizing vertebrate genomic insulators. We envision that with more and more experimentally validated CTCFBS sequences available in the database, a comprehensive analysis of these sequences may facilitate the extraction of meaningful sequence signals, uncover the functional basis of insulators, and ultimately enable the mapping of every distinct transcription domain along the genomes.

## ACKNOWLEDGEMENTS

We thank Dr Bing Ren for providing the 20-bp motif information for 13 801 experimentally identified CTCF-binding sites, Drs Bing Ren, Xiaohui Xie and Eric S. Lander for providing the predicted CTCF motifs and Dr Keji Zhao for providing the genomic coordinates of 20 262 CTCF-binding sites. Funding to pay the Open Access publication charges for this article was provided by The University of Tennessee Health Science Center.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Moon,H., Filippova,G., Loukinov,D., Pugacheva,E., Chen,Q., Smith,S.T., Munhall,A., Grewe,B., Bartkuhn,M. *et al.* (2005) CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Rep.*, **6**, 165–170.
2. West,A.G., Gaszner,M. and Felsenfeld,G. (2002) Insulators: many functions, many mechanisms. *Genes Dev.*, **16**, 271–288.
3. Bell,A.C., West,A.G. and Felsenfeld,G. (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic. *Science*, **291**, 447–450.
4. Gaszner,M. and Felsenfeld,G. (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.*, **7**, 703–713.
5. Kuhn,E.J. and Geyer,P.K. (2003) Genomic insulators: connecting properties to mechanism. *Curr. Opin. Cell Biol.*, **15**, 259–265.
6. Brasset,E. and Vaury,C. (2005) Insulators are fundamental components of the eukaryotic genomes. *Heredity*, **94**, 571–576.
7. Capelson,M. and Corces,V.G. (2004) Boundary elements and nuclear organization. *Biol. Cell*, **96**, 617–629.
8. Engel,N. and Bartolomei,M.S. (2003) Mechanisms of insulator function in gene regulation and genomic imprinting. *Int. Rev. Cytol.*, **232**, 89–127.
9. Fourel,G., Magdinier,F. and Gilson,E. (2004) Insulator dynamics and the setting of chromatin domains. *Bioessays*, **26**, 523–532.
10. Geyer,P.K. and Clark,I. (2002) Protecting against promiscuity: the regulatory role of insulators. *Cell Mol. Life Sci.*, **59**, 2112–2127.
11. Labrador,M. and Corces,V.G. (2002) Setting the boundaries of chromatin domains and nuclear organization. *Cell*, **111**, 151–154.
12. West,A.G. and Fraser,P. (2005) Remote control of gene transcription. *Hum. Mol. Genet.*, **14**, R101–R111.
13. Scott,K.C., Merrett,S.L. and Willard,H.F. (2006) A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. *Curr. Biol.*, **16**, 119–129.

14. Bell,A.C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.
15. Hark,A.T., Schoenherr,C.J., Katz,D.J., Ingram,R.S., Levorse,J.M. and Tilghman,S.M. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, **405**, 486–489.
16. Yu,W., Ginjala,V., Pant,V., Chernukhin,I., Whitehead,J., Docquier,F., Farrar,D., Tavoosidana,G., Mukhopadhyay,R. *et al.* (2004) Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin- insulation. *Nat. Genet.*, **36**, 1105–1110.
17. Chao,W., Huynh,K.D., Spencer,R.J., Davidow,L.S. and Lee,J.T. (2002) CTCF, a candidate trans-acting factor for X-inactivation choice. *Science*, **295**, 345–347.
18. Valley,C.M. and Willard,H.F. (2006) Genomic and epigenomic approaches to the study of X chromosome inactivation. *Curr. Opin. Genet. Dev.*, **16**, 240–245.
19. Ling,J.Q., Li,T., Hu,J.F., Vu,T.H., Chen,H.L., Qiu,X.W., Cherry,A.M. and Hoffman,A.R. (2006) CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science*, **312**, 269–272.
20. 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.
21. Mukhopadhyay,R., Yu,W., Whitehead,J., Xu,J., Lezcano,M., Pack,S., Kanduri,C., Kanduri,M., Ginjala,V. *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.
22. 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.
23. 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.
24. Chung,J.H., Whiteley,M. and Felsenfeld,G. (1993) A 5′ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, **74**, 505–514.
25. Kent,W.J. (2002) BLAT – The BLAST-Like Alignment Tool. *Genome Res.*, **12**, 656–664.
26. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
27. Sparago,A., Cerrato,F., Vernucci,M., Ferrero,G.B., Silengo,M.C. and Riccio,A. (2004) Microdeletions in the human H19 DMR result in loss of IGF2 imprinting and Beckwith-Wiedemann syndrome. *Nat. Genet.*, **36**, 958–960.
28. Prawitt,D., Enklaar,T., Gartner-Rupprecht,B., Spangenberg,C., Oswald,M., Lausch,E., Schmidtke,P., Reutzel,D., Fees,S. *et al.* (2005) Microdeletion of target sites for insulator protein CTCF in a chromosome 11p15 imprinting center in Beckwith-Wiedemann syndrome and Wilms' tumor. *Proc. Natl Acad. Sci. USA*, **102**, 4085–4090.
29. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
30. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
31. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
32. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad Sci. USA*, **100**, 11484–11489.
33. Schones,D.E., Smith,A.D. and Zhang,M.Q. (2007) Statistical significance of cis-regulatory modules. *BMC Bioinformatics*, **8**, 19.
34. 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.
35. Altschul,S.F., Gish,W., Miller,W., Meyers,E.W. and Lipman,D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403.