*W524–W528* *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
*doi:10.1093/nar/gkl041*

# TFBScluster web server for the identification of mammalian composite regulatory elements

**Ian J. Donaldson and Berthold Göttgens***

Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge, CB2 2XY, UK

Received January 10, 2006; Revised and Accepted January 24, 2006

## ABSTRACT

**Identification of transcriptional regulatory elements represents a critical step in our ability to reconstruct transcriptional regulatory networks from gene expression profiling datasets. To facilitate computational identification of candidate gene regulatory elements from whole genome sequences, we have developed the TFBScluster web server that integrates several tools for the genome-wide identification and subsequent characterization of transcription factor binding site clusters that are conserved in multiple mammalian species. Either the human or mouse genomes can be used as the reference sequence with direct links from the search results to the ENSEMBL and UCSC genome browsers. Moreover, TFBScluster provides seamless integration of transcription factor binding site searches with genome annotation and gene expression profiling data, to allow prioritising computational predictions for subsequent experimental validation. TFBScluster is publicly available at http://hscl.cimr.cam.ac.uk/TFBScluster_genome_portal.html.**

## INTRODUCTION

One of the great challenges of the post genomic era is to integrate the various 'omics' approaches with the accumulated data from traditional hypothesis driven research to make a 'systems level' understanding of biological processes a feasible goal. Systems biology 'parts lists' (e.g. whole genome sequences and gene expression profiles) are increasingly generated for higher eukaryotes, including human and vertebrate model systems. Based on recent studies in sea urchins, flies and worms (1–6), analysis of transcriptional gene regulatory networks holds particular promise for applying systems biology to higher eukaryotes. One particularly attractive feature of transcriptional regulatory networks lies in the fact that computational analysis of transcriptional regulatory networks should be facilitated since major components of the 'parts lists' will be intimately connected through a regulatory code present in the genomic sequence that determines where, when and at what level each gene will be expressed. Nevertheless, the higher genomic and biological complexities of vertebrates currently pose major challenges to directly transferring approaches developed in lower eukaryotes.

Identification and subsequent characterization of gene regulatory elements will be a key step in assembling transcriptional regulatory networks from gene expression profiling data, with the ultimate goal of unravelling the regulatory codes that govern gene expression in various cell types. The resulting models of transcriptional regulatory networks will have the ability to elucidate system properties, which may not be apparent when studying individual components. Moreover, models can predict the outcome of perturbations, experimental or pathological, and are therefore likely to play a key role in future drug discovery. The network nodes of transcriptional regulatory networks consist of gene regulatory elements; these are stretches of DNA sequence containing multiple transcription factor binding sites (TFBSs) that act combinatorially on the transcriptional regulation of a given gene. Therefore, genome-wide analysis of regulatory elements is ultimately required to construct transcriptional regulatory networks. However, even more so than in simpler genomes such as fly, yeast or worm, the complexity of mammalian genomes makes computational identification of functional regulatory elements a formidable task. The DNA sequence motifs recognized by transcription factors are mostly short and degenerate, and regulatory elements can be located many kilobases away from the proximal promoter of a gene in distal 5′ and 3′ enhancers or in introns (7).

To facilitate the identification of mammalian gene regulatory elements, we developed the TFBScluster tool for the genome-wide discovery of candidate regulatory elements and the genes under their control. The original version of TFBScluster was designed specifically for the characterization of gene regulatory elements active in human blood stem cells (8). In a proof of principle study, TFBScluster allowed

*To whom correspondence should be addressed. Tel: +44 1223 336829; Fax: +44 1223 762670; Email: bg200@cam.ac.uk

us to identify candidate gene regulatory sequences with predicted biological activity, confirmed using transgenic mouse assays (9).

The newly updated version of the TFBScluster web server has many new features that make it a widely applicable resource for studying mammalian gene regulatory networks. For example, we have now implemented mouse and human versions to streamline analysis for researchers using murine or human systems as their principal experimental setting. Moreover, integration within TFBScluster of genomic sequence searches with gene expression profiling datasets provides the ability to datamine across diverse yet complementary datasets, in order to increase the likely accuracy of computational predictions. We have implemented a much wider range of TFBSs, including IPUAC consensus sequences curated from the literature, published datasets and a large set of conserved positional weight matrices. Our emphasis has moved from the use of multiple alignments to the initial use of pair-wise alignments, such as human–mouse. Conserved TFBSs can now be further filtered to retain only those sites that are also conserved between human/mouse and other species including dog and opossum.

TFBScluster significantly differs from other tools developed for the analysis of lower invertebrate species (3,10–12), tools that use limited datasets (13–16) or are restricted to sequence flanking predicted transcription start-sites (17–24). TFBScluster shares the concept of identifying binding site clusters with the SynoR program (25). However, TFBScluster utilizes three datasets of IUPAC consensus sequences in addition to TRANSFAC matrices, and provides links to mainstream genome browsers thereby providing a variety of external information.

## METHODOLOGY OF TFBScluster

TFBScluster utilizes three principle methods to help distinguish functional binding sites (true positives) from the background noise of non-functional sites (false positives): (i) queries focus upon TFBSs that form part of binding site clusters, (ii) TFBScluster only considers sites that are conserved between two or more species and (iii) searches can be restricted to areas of the genome thought to be involved in regulating gene expression. TFBScluster can subject candidate binding site clusters to a series of filters based upon associated gene expression, and by their location in areas of regulatory potential or active promoters.

### Motif datasets

Genome-wide positions for three sets of IUPAC code defined TFBS consensus sequences and one set of positional weight matrices have been determined. The first IUPAC code set consists of 41 consensus sequences curated from the literature and the databases TRANSFAC (26) and JASPAR (27). Background information and references for each IUPAC consensus sequence can be found at (http://hscl.cimr.cam.ac.uk/TFBScluster_genome_35_filters_background.html). Five datafiles containing genome-wide collections of matching conserved sites with increasing levels of sequence conservation have been generated for all 41 IUPAC consensus sequences. The first datafile contains 'non-exact' matches to the core sequence; both sequences match the IUPAC consensus, but degenerate IUPAC codes are allowed to differ between the two sequences. The second datafile contains 'exact' matches, where degenerate IUPAC codes must be identical between the two sequences, thus requiring degenerate consensus sequences to be aligned in regions of high sequence identity. The last three datafiles also require an exact match and extend the overall length of sequence identity. To achieve this, the IUPAC code 'N' (any nucleotide) is added to both ends of the consensus, resulting in three files with 2, 4 and 6 conserved nucleotides flanking the core sequence. Functional binding sites are likely to be located in highly conserved sequence regions. Therefore, our method of increasing the degree of conservation in our TFBS datafiles aims to enrich functional binding sites whilst decreasing the number of false positive sites. The identification of completely or near completely conserved binding sites is a method that has been successfully used in a variety of other approaches (28).

The second set of IUPAC codes was taken from a recently published study (29), which identified common regulatory motifs conserved in human, dog, mouse and rat genomes. We have now taken the top 50 IUPAC consensus sequences from this study and determined their positions in whole genome alignments (human–mouse, human–dog, human–opossum, mouse–human, mouse–dog and mouse–opossum). The third set of IUPAC codes was taken from a similar study (30) that identified a 'dictionary' of conserved consensus sequences in the promoter regions of orthologous human and mouse genes. Again, we have determined the genome-wide positions of the non-degenerate IUPAC consensus sequences detailed in this second study, which differentiated between those located in CpG rich regions (35 consensus sequences) and those in non-CpG rich regions (19 consensus sequences). Finally for the human version of TFBScluster, we have incorporated the genome-wide positions of TFBSs matching 410 positional weight matrices conserved in human, mouse and rat whole genome alignments. We obtained these from the UCSC genome browser (http://genome.ucsc.edu/). The positional weight matrices originate from the TRANSFAC database v8.3.

### Whole genome alignments

Human and mouse versions of TFBScluster have been implemented to serve these two large research communities. Both versions of TFBScluster incorporate the genome-wide positions of TFBS consensus sequences conserved in a series of pair-wise genome alignments, where either the human or mouse genome is the reference sequence. The pair-wise genome comparisons were downloaded from Genome Bioinformatics at the UCSC (http://genome.ucsc.edu/downloads.html). For the human centric version (NCBI v35/hg17) TFBSs conserved in the mouse genome (NCBI 33/mm5) represent the default level of conservation. Positions of TFBSs are also catalogued in human–dog (canFam1) and human–opossum (monDom1) alignments. Human–mouse conserved sites have then been filtered to retain only those that are also present in the latter two pair-wise alignments. For the mouse version (NCBI 34/mm6) the default level of conservation uses sites conserved in the human (NCBI 35/hg17)

genome and filtered sets have been produced using mouse–dog (canFam1) and mouse–opossum (monDom1) alignments.

### TFBScluster input

The first screen of TFBScluster requires the user to choose the reference genome (human or mouse) to determine the relevant genome annotation and external data used for the subsequent analysis of candidate regulatory regions. The next screen specifies the number of different consensus sequences that can be present in a cluster of TFBSs. The choice is also made as to which dataset of TFBSs should be used, either in-house consensus sequences, conserved regulatory motifs from two other studies (29,30) or TRANSFAC v8.3 matrices. The next screen contains fields to specify the main parameters of candidate regulatory clusters, such as the number of motifs within a user defined window of sequence. To ensure that a minimum set of TFBSs are free to bind their corresponding proteins simultaneously, an option is included to discard clusters that do not contain the required number of sites when overlapping TFBSs are discounted. By default TFBScluster will use datafiles of binding sites that are conserved between human–mouse or mouse–human genomes. The degree of conservation and therefore the stringency of the search can be increased by selecting sites that are also conserved in dog or opossum.

In the human version of TFBScluster we have implemented additional approaches to increase the likelihood that TFBSs represent functional sites. Regulatory potential scores have been shown by others (31,32) to provide significant enrichment of regulatory sequences. A filter has therefore been implemented that can restrict TFBScluster to sites located in areas of regulatory potential with scores greater than zero (based on the threshold suggested by UCSC http://hgdownload.cse.ucsc.edu/goldenPath/hg17/regPotential/). It is also possible to only consider those motifs that are present within experimentally determined active promoters, using the fibroblast cell line IMR90. This dataset is based on a recent study that identified active promoters using a microarray based chromatin immunopreciptation method to detect all RNA polymerase II preinitiation complexes assembled on DNA throughout the human genome (33).

### TFBScluster output

The 'short' analysis returns the chromosomal start and end positions of all candidate clusters. This method is useful to gauge how many candidate regulatory clusters are found using a given set of search parameters. Moreover, as the data is returned in the general feature format (GFF; http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml), results files can be readily exported to other genome analysis programs. Users are advised to avoid 'long' searches using parameters that generate more than 5000 candidate clusters, due to the length of time it will take to complete.

TFBScluster can provide a more comprehensive analysis of candidate clusters by selecting the 'long' analysis. Using the Ensembl API (34), this option will identify all genes that are localized to within 100 kb of a cluster. A single gene will be associated with a cluster if the cluster is located in an intron, otherwise up to two genes will be associated with a cluster by selecting the closest gene 5′ and 3′ to the cluster. In TFBScluster a gene is defined as the footprint of all Ensembl gene transcripts on the genome sequence. All candidate clusters located in an exon are excluded from further analysis. Clusters can be restricted to genes that are annotated with EntrezGene (35) and/or UniProt-SwissProt (36,37) identifiers. The choice of SwissProt identifiers ensures that only those genes with experimentally characterized proteins are detailed in the results.

TFBScluster has the ability to retain or exclude cluster candidates based on the presence of user supplied IUPAC consensus sequences. This facility can be useful, e.g. to reject clusters containing a consensus that is known not to be associated with the others. It can also be used to determine space and order constraints within the cluster by submitting consensus sequence patterns. For example, GATA*5–8*GATA will look for two Gata sites within 5 to 8 nt on the same strand.

The 'long' analysis results page provides detailed information for each candidate cluster together with information about the genes that are potentially under their control. Access is also provided to other files containing subsets of information from the main results file, such as a list of Swissprot identifiers that can be used in external analyses (see http://hscl.cimr.cam.ac.uk/TFBScluster_examples/TFBScluster_ex_file.html). Links are provided to both UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway) and Ensembl (http://www.ensembl.org/index.html) genome browsers; they both indicate the location of a cluster (with 300 flanking nucleotides) in relation to the surrounding chromosome annotation. Information is therefore provided to streamline the experimental verification of interesting candidates; 300 nt of sequence flanking the cluster is shown for the reference genome, as well as an alignment of the predicted cluster for the default genomes.

The selection of clusters can be further constrained by only considering those that have been localized to genes that are expressed in a tissue specific manner. To achieve this, both the human and mouse versions of TFBScluster make use of information from the Gene Expression Atlas 2 (38). The Gene Expression Atlas 2 contains global gene expression profiles for 79 different human tissues and 61 different mouse tissues, together with extensive statistical analysis. For example, a TFBScluster analysis for muscle-specific binding site clusters can be focussed on those genes that show differential expression in muscle. To achieve this, TFBScluster would only retain those clusters that localize to genes with a specified fold over median expression in muscle when compared to the other 78 tissues. This important feature therefore allows the user to connect candidate cluster regions to a particular biological context in human or mouse.

## SOFTWARE AND ACCESS

TFBScluster.pl and supporting programs are written in PERL and are accessible via a PERL CGI interface on a web server, hosted by the University of Cambridge. The PERL scripts are available on request. The run time of a submitted job is typically less than 15 min using the 'short' analysis. For 'long' analyses, run time is dependent on the number of candidate clusters. Although users are initially restricted to the TFBS consensus sequences already present in TFBScluster, we

welcome any requests for new motifs to be added. The default method of obtaining results files is via email. However, if there is a desire for anonymity then an alternative method allows the user to manually check whether a job has finished by following a web page link that is provided when the job is submitted (without the need to enter an email address). If the link is followed before the job is finished, a message to that effect is displayed. This page will refresh every 30 s until the job is completed. Either the results will be displayed or a message stating that no clusters were found. Throughout, all user input screens are designed to be simple and used in a step-wise manner; appropriate default values have been pre-entered as good starting points to run the analysis.

## CONCLUSIONS

Modern biology is increasingly dependent on computational infrastructure to integrate diverse datasets from multiple different organisms and of multiple different types. TFBScluster is designed to utilize diverse information (e.g. genome annotation, comparative genomics and expression profiling) to permit the formulation of novel experimentally testable propositions about the likely biological function of gene regulatory elements. Understanding the function of transcriptional regulatory elements represents a critical step in our ability to reconstruct transcriptional regulatory networks. TFBScluster is therefore potentially widely applicable for future studies of human and murine developmental systems and disease models. Moreover, TFBScluster could be readily adaptable for other genomes annotated in Ensembl.

## ACKNOWLEDGEMENTS

Work in the authors' laboratory is funded by the Cambridge MIT Institute, an SUR grant from IBM and the Leukaemia Research Fund. Funding to pay the Open Access publication charges for this article was provided by Cambridge MIT Institute.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al*. (2002) A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Dev. Biol*., **246**, 162–190.
2. Markstein,M. and Levine,M. (2002) Decoding cis-regulatory DNAs in the *Drosophila* genome. *Curr. Opin. Genet. Dev*., **12**, 601–606.
3. Markstein,M., Markstein,P., Markstein,V. and Levine,M.S. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA*, **99**, 763–768.
4. Senger,K., Armstrong,G.W., Rowell,W.J., Kwan,J.M., Markstein,M. and Levine,M. (2004) Immunity regulatory DNAs share common organizational features in *Drosophila*. *Mol. Cell*, **13**, 19–32.
5. Stathopoulos,A., Van Drenth,M., Erives,A., Markstein,M. and Levine,M. (2002) Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell*, **111**, 687–701.
6. Wenick,A.S. and Hobert,O. (2004) Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in *C.elegans*. *Dev. Cell*, **6**, 757–770.
7. Nobrega,M.A., Ovcharenko,I., Afzal,V. and Rubin,E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
8. Donaldson,I.J., Chapman,M. and Gottgens,B. (2005) TFBScluster: a resource for the characterization of transcriptional regulatory networks. *Bioinformatics*, **21**, 3058–3059.
9. Donaldson,I.J., Chapman,M., Kinston,S., Landry,J.R., Knezevic,K., Piltz,S., Buckley,N., Green,A.R. and Gottgens,B. (2005) Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum. Mol. Genet*., **14**, 595–601.
10. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
11. Grad,Y.H., Roth,F.P., Halfon,M.S. and Church,G.M. (2004) Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D.pseudoobscura*. *Bioinformatics*, **20**, 2738–2750.
12. Sosinsky,A., Bonin,C.P., Mann,R.S. and Honig,B. (2003) Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors. *Nucleic Acids Res*., **31**, 3589–3592.
13. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*., **31**, 3666–3668.
14. Johansson,O., Alkema,W., Wasserman,W.W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19**, i169–i176.
15. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*., **32**, W217–W221.
16. Yu,H., Yoo,A.S. and Greenwald,I. (2004) Cluster analyzer for transcription sites (CATS): a C++-based program for identifying clustered transcription factor binding sites. *Bioinformatics*, **20**, 1198–1200.
17. Aerts,S., Van Loo,P., Moreau,Y. and De Moor,B. (2004) A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, **20**, 1974–1976.
18. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of cis -regulatory modules. *Bioinformatics*, **19**, II5–II14.
19. Kankainen,M. and Holm,L. (2004) POBO, transcription factor binding site verification with bootstrapping. *Nucleic Acids Res*., **32**, W222–W229.
20. Karanam,S. and Moreno,C.S. (2004) CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. *Nucleic Acids Res*., **32**, W475–W484.
21. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*., 127–138.
22. Sharan,R., Ben-Hur,A., Loots,G.G. and Ovcharenko,I. (2004) CREME: cis-regulatory module explorer for the human genome. *Nucleic Acids Res*., **32**, W253–W256.
23. Sharan,R., Ovcharenko,I., Ben-Hur,A. and Karp,R.M. (2003) CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, **19**, i283–i291.
24. Vega,V.B., Bangarusamy,D.K., Miller,L.D., Liu,E.T. and Lin,C.Y. (2004) BEARR: batch extraction and analysis of cis-regulatory regions. *Nucleic Acids Res*., **32**, W257–W260.
25. Ovcharenko,I. and Nobrega,M.A. (2005) Identifying synonymous regulatory elements in vertebrate genomes. *Nucleic Acids Res*., **33**, W403–W407.
26. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res*., **27**, 318–322.
27. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.
28. Prakash,A. and Tompa,M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol*., **23**, 1249–1256.
29. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′-UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
30. Ettwiller,L., Paten,B., Souren,M., Loosli,F., Wittbrodt,J. and Birney,E. (2005) The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol*., **6**, R104.

31. King,D.C., Taylor,J., Elnitski,L., Chiaromonte,F., Miller,W. and Hardison,R.C. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res*., **15**, 1051–1060.
32. Kolbe,D., Taylor,J., Elnitski,L., Eswara,P., Li,J., Miller,W., Hardison,R. and Chiaromonte,F. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse and rat. *Genome Res*., **14**, 700–707.
33. Kim,T.H., Barrera,L.O., Zheng,M., Qu,C., Singer,M.A., Richmond,T.A., Wu,Y., Green,R.D. and Ren,B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.
34. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl core software libraries. *Genome Res*., **14**, 929–933.
35. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
36. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The universal protein resource (UniProt). *Nucleic. Acids Res*., **33**, D154–D159.
37. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
38. 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.