BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2519–2520
doi:10.1093/bioinformatics/btt432

Databases and ontologies

Advance Access publication August 9, 2013

TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors

Konika Chawla[1,†], Sushil Tripathi[2,†], Liv Thommesen[2,3], Astrid Lægreid[2] and Martin Kuiper[1,*]

[1]Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway, [2]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway and [3]Department of Technology, Sør-Trøndelag, University College, N-7004 Trondheim, Norway

Associate Editor: Janet Kelso

ABSTRACT

Summary: Gene regulatory network assembly and analysis requires high-quality knowledge sources that cover functional aspects of the various components of the gene regulatory machinery. A multiplicity of resources exists with information about mammalian transcription factors (TFs); yet, only few of these provide sufficiently accurate classifications of the functional roles of individual TFs, or standardized evidence that would justify the information on which these functional classifications are based. We compiled the list of all putative TFs from nine different resources, ignored factors such as general TFs, mediator complexes and chromatin modifiers, and for the remaining factors checked the available literature for references that support their function as a true sequence-specific DNA-binding RNA polymerase II TF (DbTF). The results are available in the TFcheckpoint database, an exhaustive collection of TFs annotated according to experimental and other evidence on their function as true DbTFs. TFcheckpoint.org provides a high-quality and comprehensive knowledge source for genome-scale regulatory network studies.

Availability: The TFcheckpoint database is freely available at www.tfcheckpoint.org

Contact: martin.kuiper@ntnu.no

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 8, 2013; revised on July 8, 2013; accepted on July 23, 2013

1 INTRODUCTION

Transcription factors (TFs) lie at the basis of gene-expression diversity in different cell types and conditions. TFs constitute key gene regulatory components that usually participate in large multiprotein-DNA complexes, where they guide RNA polymerase (i.e. RNAP I, II and III) activity and regulate the onset and rate of RNA synthesis. These protein complexes may include general transcription factors that bind to core-promoter DNA; general cofactors that bind to general transcription factors to form a pre-initiation transcription complex; and specific DNA-binding transcription factors and factors that lack DNA-binding domains but exert their regulatory roles through interaction with other proteins in the transcription complex. This last class of protein-interacting transcription regulators includes coactivators, corepressors, histone modifiers and chromatin remodeling proteins (Lee and Young, 2000).

The DNA-binding transcription factors (DbTFs) play a central role in specifying which genes are transcribed, as they guide the transcription machinery to distinct target genes by binding to specific gene regulatory elements located in proximal promoters as well as in distal enhancer regions Kadonaga (2004). The DbTF proteins that regulate RNAP II enjoy a special focus in gene regulatory network building because of their strong ability to explain the protein-coding gene-expression landscape of biological responses. Access to accurate and genome-scale knowledge concerning these DbTFs, therefore, is of key importance. Multiple resources with knowledge about mammalian TF exist (Fulton *et al.*, 2009; Harris *et al.*, 2004; Kummerfeld and Teichmann, 2006; Messina *et al.*, 2004; Ravasi *et al.*, 2010; Sandelin *et al.*, 2004; Schaefer *et al.*, 2011; Vaquerizas *et al.*, 2009; Zhang *et al.*, 2012). However, we observed that (i) most of them do not distinguish well between true DbTFs, protein-interacting TFs and general TFs and (ii) only in a minority of cases do they provide standardized evidence for the functional role of the TFs. Because of this, users of these resources will only have an obscured view at the domain of DbTFs. Here we present TFcheckpoint (www.tfcheckpoint.org), a comprehensive repository of human, mouse and rat TF candidates. All entries have been manually checked for literature information pertaining to their potential biological function as DbTFs. The database serves as a checkpoint for TF information, is freely available and supports ID or name searching, browsing and bulk download.

2 RESULTS

2.1 Database content

TFcheckpoint contains the cumulative inventory of nine major TF information sources (Fig. 1 and Supplementary Material), and we manually checked each of these entries for literature describing evidence for RNAP II-regulating DNA-binding transcription factors, for human, mouse or rat. The evidence that we selected should at least support Gene Ontology (GO) term 'Sequence-specific DNA-binding RNA polymerase II transcription factor activity (GO:0000981)', taking this as the minimum defining term for a true RNAPII-regulating DbTF. In general,

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

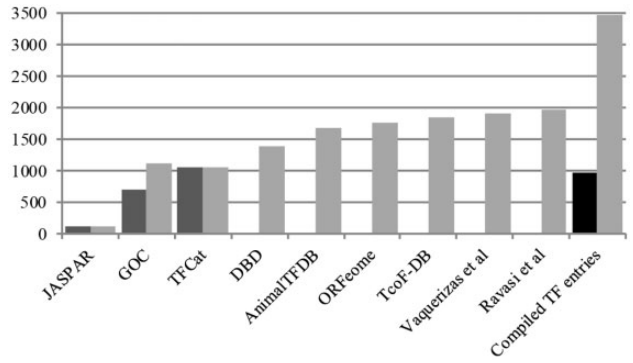

Fig. 1. TF candidate and associated literature references. For each resource, the total numbers of TF entries (light gray) and TF entries with literature references (dark gray) are given. For GOC data, all unique proteins annotated to 'Sequence-specific DNA-binding transcription factor activity (GO:0003700)' or any of its children, are listed. The bar to the far right indicates 3462 unique TF entries in TFcheckpoint; 984 of these (adjacent black bar) were deemed to be true DbTFs.

we selected the first PubMed article(s) that showed satisfactory evidence for a specific TF (for details see Supplementary Material).

We assembled a list of 3462 putative TFs from the aforementioned resources (Fig. 1). We have used orthology mappings from UniProt to identify corresponding gene Entrez IDs from human, rat and mouse. For 984 proteins, we could identify one or more relevant articles with the evidence for a DbTF, yielding 1073 unique PubMed references. Eight hundred twenty-four DbTFs are supported by literature references with experimental evidence. Just to be as comprehensive as possible in our coverage of current knowledge, we included a further 155 DbTFs that are supported by author statements and a final 5 that are supported by sequence-based analysis. The full list and the literature reference results are available from the TFcheckpoint database.

The availability of high-quality and exhaustive information at one central place facilitates the access by the global scientific community. We are currently working together with the Gene Ontology Consortium (GOC) to develop and apply general standards for TF annotation and merge our findings with the GO database (Harris *et al.*, 2004). Our efforts should help the GOC to solve the current backlog in DbTF curation.

2.2 Database user interface

TFcheckpoint is powered by MySQL and accessible through a web interface created with Joomla (http://www.joomla.org), implementing HTML and PHP scripts. The database is hosted on an apache server at the Norwegian data infrastructure Norstore (http://www.norstore.no) and available at www.tfcheckpoint.org. The database can be used for simple browsing of all 3462 candidate TFs as well as the subset of DbTFs with literature evidence. For each DbTF, the literature reference(s) and information about the original TF candidate resource that we obtained it from are provided. All TF entries are linked to Entrez and UniProt IDs. The NCBI official gene symbol is used as a primary key, but the data can also be searched for any of the NCBI provided synonyms, as well as Entrez and UniProt identifiers. All data are also downloadable as a tab-delimited text file.

3 CONCLUSION

A literature-based exhaustively curated list of transcription factors is an invaluable resource for researchers working on gene regulatory mechanisms. The ENCODE project for instance is targeting the generation of evidence for some 1900 putative DbTFs (ENCODE Project Consortium *et al.*, 2012), as estimated by one of our sources (Vaquerizas *et al.*, 2009). TFcheckpoint provides a reference for both small-scale experiments and genome-scale studies. Researchers may verify predicted lists of TFs even before characterizing the role of these regulatory proteins (Choi *et al.*, 2006; Gray *et al.*, 2004) or they can use the background knowledge of TFs to infer gene regulatory networks (Ye *et al.*, 2009). By ensuring that our annotations become part of the GO database, this knowledge will become available to all analysis approaches based on GO knowledge.

Funding: This work was supported by The Norwegian Cancer Society, The Liaison Committee between the Central Norway Regional Health Authority (RHA), the Norwegian University of Science and Technology (NTNU) and Sør-Trøndelag University College (HisT).

Conflicts of Interest: none declared.

REFERENCES

Choi,M.Y. *et al.* (2006) A dynamic expression survey identifies transcription factors relevant in mouse digestive tract development. *Development*, **133**, 4119–4129.

ENCODE Project Consortium *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Fulton,D.L. *et al.* (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.

Gray,P.A. *et al.* (2004) Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science*, **306**, 2255–2257.

Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Kadonaga,J.T. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, **116**, 247–257.

Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res.*, **34**, D74–D81.

Lee,T.I. and Young,R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, **34**, 77–137.

Messina,D.N. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.

Ravasi,T. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Schaefer,U. *et al.* (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res.*, **39**, D106–D110.

Vaquerizas,J.M. *et al.* (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

Ye,C. *et al.* (2009) Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast. *PLoS Comput. Biol.*, **5**, e1000311.

Zhang,H.M. *et al.* (2012) AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.*, **40**, D144–D149.