Published online 11 November 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D105–D110***
doi:10.1093/nar/gkp950

JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles

Elodie Portales-Casamar[1], Supat Thongjuea[2], Andrew T. Kwon[1], David Arenillas[1], Xiaobei Zhao[3], Eivind Valen[3], Dimas Yusuf[1], Boris Lenhard[2,*], Wyeth W. Wasserman[1,*] and Albin Sandelin[3,*]

[1]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, 950 West 28th Avenue, Vancouver BC, V5Z 4H4, Canada, [2]Computational Biology Unit – Bergen Center for Computational Science, and Sars Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway and [3]The Bioinformatics Centre, Department of Biology and Biomedical Research and Innovation Centre, Copenhagen University, Ole Maaloes Vej 5, DK-2200, Denmark

Received September 15, 2009; Revised October 11, 2009; Accepted October 12, 2009

Downloaded from http://nar.oxfordjournals.org/ by guest on October 7, 2013

ABSTRACT

JASPAR (http://jaspar.genereg.net) is the leading open-access database of matrix profiles describing the DNA-binding patterns of transcription factors (TFs) and other proteins interacting with DNA in a sequence-specific manner. Its fourth major release is the largest expansion of the core database to date: the database now holds 457 non-redundant, curated profiles. The new entries include the first batch of profiles derived from ChIP-seq and ChIP-chip whole-genome binding experiments, and 177 yeast TF binding profiles. The introduction of a yeast division brings the convenience of JASPAR to an active research community. As binding models are refined by newer data, the JASPAR database now uses versioning of matrices: in this release, 12% of the older models were updated to improved versions. Classification of TF families has been improved by adopting a new DNA-binding domain nomenclature. A curated catalog of mammalian TFs is provided, extending the use of the JASPAR profiles to additional TFs belonging to the same structural family. The changes in the database set the system ready for more rapid acquisition of new high-throughput data sources. Additionally, three new special collections provide matrix profile data produced by recent alternative high-throughput approaches.

INTRODUCTION

The wide availability of TF affinity data is becoming essential for an increasing number of research efforts to understand gene regulation in the post-genomic era. The increasing amount of assembled genome sequences, transcriptome data (1), as well as high-throughput studies revealing genome-wide locations of core promoters (2) and enhancer elements (3,4) have resulted in the greatest demand for TF binding site content analyses.

TF binding affinities are typically modeled as position frequency matrices (PFMs, also known as raw count matrices or simply binding profiles), summarizing nucleotide counts in an alignment of active binding sites These can be used to scan genomes for new binding sites (5). Since the first official release of JASPAR in 2004 (6), the research community has embraced it as the leading open-access database of such matrix profiles for TF binding sites. From the beginning, the aim of its core collection has been to provide a non-redundant set of curated, high-quality matrix profiles derived from experimental binding data in the form of position frequency matrices (7); in other words, the goal is to present the best currently available DNA binding model for a given TF, decided by expert curators.

The availability of potentially useful matrices derived by other means (e.g. using a number of genome-wide computational approaches) as well as non-TF binding profiles, prompted the addition of separate JASPAR Collections in the second release (8): the intention was to provide those matrix profiles in the same format and hence usable with the same tools as the core JASPAR

*To whom correspondence should be addressed. Tel: +47 555 84362; Fax: +47 555 84295; Email: boris.lenhard@bccs.uib.no
Correspondence may also be addressed to Wyeth W. Wasserman. Tel: +1 604 875 3812; Fax: +1 604 875 3819; Email: wyeth@cmmt.ubc.ca
Correspondence may also be addressed to Albin Sandelin. Tel: +45 353 21 285; Fax: +45 353 21281; Email: albin@binf.ku.dk

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

© The Author(s) 2009. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

database, while keeping the latter reserved for profiles representing experimentally derived data.

While the community has valued the open-access policy and non-redundant nature of JASPAR, a common complaint was that the size of the core collection was small compared to the commercial TransFac database (9), currently the only comprehensive alternative to JASPAR. In this update, our goal was to make this gap smaller by performing a major expansion of the core database, while maintaining the popular non-redundant, curated quality. As a result, this fourth major release introduces a wealth of new and improved matrix profiles and represents the largest expansion of the core database since its inception, with new data coming either from high-throughput methods like Chip-seq, or assembled from TF binding site databases particularly PAZAR (10) described below.

NEW AND IMPROVED MATRIX PROFILES IN JASPAR CORE DATABASE

Profiles from ChIP-seq

Several recent genome-wide studies have revealed thousands of TF binding sites for individual TFs. Compared to the original matrices, the larger number of representative target sequences provides potentially more accurate profiles and brings the added benefit that (unlike in DNA SELEX), all the binding sites come from the actual genome sequence to which the TFs in question are bound *in vivo*.

To make the derivation of matrices uniform, we extracted the original sets of bound regions from published experiments (11–19). We retrieved 200 bp sequences centered on each peak and performed *de novo* motif discovery on them using parallelized MEME (20) on a Cray XT4 supercomputing platform, which can handle inputs of many thousands of sequences in manageable time. In most cases, the resulting matrices closely resemble those reported in the original publications, produced using various motif discovery tools. The single exception was the Zfx profile, where our profile obtained with MEME from sites reported in (13) differed reproducibly from the profile reported therein. In this case, we chose to include the newly derived matrix.

In most cases, the ChIP-seq data resulted in improved matrices with higher information content than the original ones derived from either compiled single promoter assays or from DNA SELEX (Figure 1). This contradicts the widely held view that SELEX is prone to producing over-specified models since many selection rounds are commonly used. Also, somewhat surprisingly, the resulting matrices did not differ much as thresholds were varied for the inclusion of ChIP identified regions (e.g. top 100 highest confidence bound regions versus top 1000).

Profiles from ChIP-chip experiments

The ChIP-chip derived TF binding sites, while not providing the resolution of the ChIP-seq data, are a rich source of binding data. Even though they are currently being superseded by ChIP-seq (21), the published sets contain a number of high-quality binding data currently unavailable in the ChIP-seq version. As with ChIP-seq, we use the enriched regions reported by the authors of the study in question, and then apply MEME to find the pattern.

Yeast profiles in core collection

Previous versions of JASPAR did not include any matrix profiles for yeast TFs. Responding to community requests, we have compiled results from several large-scale binding profile projects to produce a non-redundant set of matrix profiles for TFs from *Saccharomyces cerevisiae*. The sources used, in order of preference, were a recent *in vitro* binding screen (22), a protein-binding microarray (PBM) experiment (23), the compiled SCPD binding profile database (24), the SwissRegulon computational re-analysis of multiple data collections (25) and a motif discovery-based collection from a widely used ChIP-chip data collection (26). The prioritization of the contributions, as well as the indicated deviations, reflect the curators' personal perspective. The preferred set, from Badis *et al*. (22), appeared to offer matrices of consistently high-quality, likely reflecting the curated nature of the effort (new experimental data were compared against existing data for consistency). All matrices were manually curated to remove redundancies and converted to count matrices. In curating the collection, the curators identified a few instances in which profiles were preferred in contradiction with the source priority: GAL4 (SwissRegulon), GCR1 (SwissRegulon), MATALPHA2 (SCPD), PHO4 (UniProbe with the six leftmost and rightmost nucleotides trimmed) and ROX1 (SCPD). The resulting non-redundant set represents a comprehensive open-access compilation of yeast binding profiles, facilitating genome-wide computational studies of yeast regulatory inputs. We are grateful to the commitment of all of the data providers to open information, without which the compilation would have been impossible.

New literature-based profiles from PAZAR

Recently, annotations of hundreds of experimentally validated TF binding sites from published studies have accumulated in the PAZAR database (27), allowing us to produce additional matrices similar in nature to the original JASPAR release (DNA SELEX or compiled from multiple studies on individual binding sites). The PAZAR database was mined to identify TFs with more than 15 annotated binding sites. The resulting data was manually curated, selecting only the results from the most high-quality data collections (i.e. collections manually annotated from the literature by specialists) and discarding any redundant sequences to build the profiles. The resulting set of compiled binding sites for each TF was used as input to the MEME software to obtain a profile. If non-informative positions were obtained on the edges of the matrices, the profiles were trimmed accordingly.

Additional model organism core profiles

For this new release, two major sources of *Drosophila melanogaster* matrix profiles have been used: DNaseI

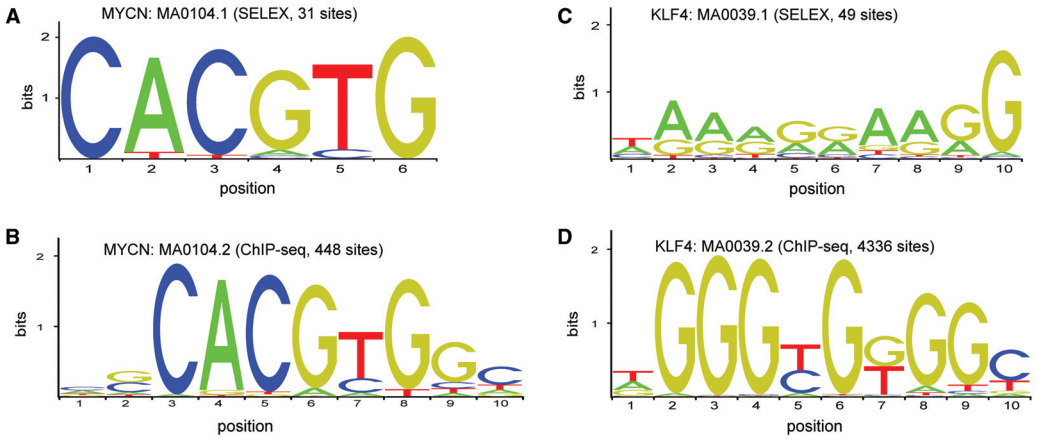

Figure 1. Examples of SELEX-derived matrix profiles replaced by ChIP-seq-derived profiles. (**A**) The previous MYCN matrix profile (MA0104.1) derived by DNA SELEX. (**B**) The new MYCN profile (MA0104.2) derived from ChIP-seq binding shows general agreement with the SELEX profile, with additional information derived from hundreds of sites at flanking positions. (**C**) The previous KLF4 profile (MA0039.1) is an example of SELEX-derived profile that did not correspond well to the handful of individually characterized KLF4 sites. (**D**) The new KLF profile derived from ChIP-seq data (13) shows a dramatic increase in information content and a good agreement with individually characterized binding sites.

footprinting data by Bergman *et al.* (28) and bacterial one-hybrid data by Wolfe and colleagues (29–31). The profiles from these data sets have been curated by the authors to remove redundancies among the results and with the existing profiles in the previous version of JASPAR database. In addition, any profile based on less than 10 sequences has been discarded. This new insect sub-section of JASPAR core includes 123 curated profiles; however, these are heavily dominated by the homeodomain profiles (29). For *Caenorhabditis elegans*, no large sources of data are currently available. Through literature searches, we identified only five profiles suitable for inclusion in the core database (32–36).

In summary, the JASPAR core database now numbers 457 non-redundant matrix profiles (Table 1). New core profiles are summarized in Supplementary Table S1.

NEW COLLECTIONS

In addition to the expansion of the core database, we remain committed to providing other collections of matrix profiles within JASPAR.

Recently, the PBM technology has emerged as a new *in vitro* method for the characterization of TF binding affinities (37). The UniPROBE database hosts the PBM datasets and makes the derived matrix profiles available to the community (38). We have selected three of these new datasets as new collections in JASPAR:

- PBM, the set derived by (39) from binding preferences of 104 mouse TFs. For each TF, both the primary and secondary motifs identified in the study were incorporated.
- PBM_HOMEO, the set derived by (40) includes 176 profiles from mouse homeodomains. From the original 168 TFs analyzed, two were discarded because they could not be identified (Dobox4 and Dobox5) and ten have two alternative profiles.
- PBM_HLH, the set derived from binding preferences of dimers of *C. elegans* bHLH TFs, including nine homodimers and ten heterodimers (41).

With these additions, JASPAR now holds 840 profiles within collections outside of the core database.

GENERAL ORGANIZATIONAL CHANGES

Version control and taxonomic catagories

In line with our goal of presenting the best currently available binding model for any TF, we updated some previous JASPAR entries motivated by new available data. Seventeen entries of the previous release were updated. The replacement of existing matrices with the new ones led us to the introduction of version numbers in matrix IDs, in a manner equivalent to the management of sequence versions in GenBank. For example, the old GATA1 profile MA0035 is replaced with a new one, and the full identifier of the new matrix is MA0035.2, while the old one becomes MA0035.1. By default, the latest version of non-redundant database includes the latest version of each profile. A search for 'MA0035' also retrieves the newest version, with an option to view older versions. Older versions can also be downloaded from the JASPAR web site.

The addition of 177 yeast matrices to the core collection means that the JASPAR matrices now span the entire eukaryote crown group. Even before that, a typical user scenario included the selection of only a subset of matrices derived from a particular taxonomic category

Table 1. Summary of the content and growth of the JASPAR database

JASPAR	Brief description	Subset	Number of profiles in JASPAR 3.0	New profiles in JASPAR 4.0	Updated profiles	Removed profiles	Total profiles (including all versions)	Total profiles (non-redundant)
Core								
	Non-redundant, literature-derived, curated models	Vertebrates	101	29	16	1	145	130
		Plants	21		–	–	21	21
		Insects	14	109	1	–	124	123
		Nematoda	–	5	–	–	5	5
		Fungi	–	177	–	–	177	177
		Urochordata	1	–	–	–	1	1
Total core			137	321	17	1	474	457
Collections								
POLII	Core promoter element profiles	–	13	–		–	13	13
FAM	Familial 'consensus' profiles for major structural families of transcription factors	–	11	–	–	–	11	11
CNE	Profiles overrepresented in vertebrate highly conserved non-coding elements	–	233	–	–	–	233	233
PHYLOFACTS	Evolutionary conserved profiles in 5′ promoter regions	–	174	–		–	174	174
SPLICE	Splice sites	–	6	–	–	–	6	6
PBM	Protein binding microarray profiles	–	–	208	–	–	208	208
PBM_HOMEO	Protein binding microarray profiles focused on homeodomain TFs	–	–	176	–	–	176	176
PBM_BHLH	Protein binding microarray profiles focused on bHLH domain TFs	–	–	19	–	–	19	19
Total collections			437	403	–	–	840	840

of organisms, across which the TFs are strictly orthologous and their binding activities largely unchanged (e.g. vertebrates). For that reason, both the JASPAR web interface and the download section now present the database content split into major taxonomic categories—vertebrates, insects, nematodes, (higher) plants and fungi—within which most of the binding sites are transferable across species. The option to search with and download the entire core collection is still available and behaves as before.

A standardized TF classification

Up to now, JASPAR used an *ad hoc* structural class annotation for the TFs associated with each matrix profile. In this release, we have updated the structural class annotation using our recently published catalog for mouse and human TFs (42) in which DNA binding proteins are associated with a structural classification system. We adopted the two-level classification described by Luscombe *et al.* (43) and extended it to accommodate additional binding domain structures. For the TFs from other species, we extrapolated the structural class and family based on the PFAM annotation of the DNA-binding domains. This addition to JASPAR provides a standardized system for the classification of TFs and allows a better grouping into families (or sub-families) with potentially similar binding preferences. A curated list of putative mouse/human DNA-binding proteins is provided at the JASPAR web site. It is also possible to browse the catalog by structure, to see what profiles that are available within the web interface.

Changes in the underlying database structure and interface

The underlying database schema was updated to accommodate matrix versions and to allow multiple species and TF accession numbers, as well to allow the storage of multiple collections in the same sql database. A Perl API (JASPAR5) for the new schema is available as part of the open-source TFBS Perl framework (44).

FUTURE DEVELOPMENTS

In the forthcoming months and years, a large amount of whole-genome binding data from ChIP-seq and related techniques will become available. We have created the first steps towards a standardized way of including this new data into JASPAR, which is expected to expand significantly with the concomitant increase in the quality of matrix data. At the same time, JASPAR collections outside the core will continue to include interesting matrix sets derived by other means.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Frank Grosveld and Eric Soler for permission to include into JASPAR of ChIP-seq derived profiles prior to publication. We thank the laboratories of Yair Benita, Martha Bulyk, Richard Gronostajski, Steven Jones, and Zhiping Weng for suggestions and/or contributions of

data reviewed for inclusion in the new release. We are grateful to Debra Fulton for her efforts to make the TFCat catalog available to the community.

FUNDING

EU Framework Programme 6 integrated project EuTRACC (to S.T.); YFF grant 180435 from the Norwegian Research Council (NRF), and by Bergen Research Foundation (BFS) (to B.L.). Novo Nordisk Foundation to the Bioinformatics Centre (to X.Z., E.V. and A.S.); The European Research Council under the EU 7th Framework Programme (FP7/2007-2013)/ERC grant agreement 204135 (to A.S.); Scholar of the Michael Smith Foundation for Health Research (to W.W.); Canadian Institutes for Health Research, GenomeCanada (via the Pleiades Promoter Project), GenomeBritishColumbia and the Canada Foundation for Innovation (to W.W. research laboratory). Funding for open access charge: Norwegian Research Council (NFR) (project no. 180435).

Conflict of interest statement. None declared.

REFERENCES

1. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
2. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A.M., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al*. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet*., **38**, 626–635.
3. Heintzman,N.D., Heintzman,N.D., Hon,G.C., Hawkins,R.D., Hawkins,R.D., Kheradpour,P., Kheradpour,P., Stark,A., Stark,A. *et al*. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108.
4. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al*. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
5. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*., **5**, 276–287.
6. Sandelin,A., Alkema,W., Engström,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.
7. Sandelin,A. and Wasserman,W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol*., **338**, 207–215.
8. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*., **34**, D95–D97.
9. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*., **34**, D108–D110.
10. Portales-Casamar,E., Arenillas,D., Lim,J., Swanson,M.I., Jiang,S., Mccallum,A., Kirov,S. and Wasserman,W.W. (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res*., **37**, D54–D60.
11. Tuteja,G., White,P., Schug,J. and Kaestner,K. (2009) Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res*., **37**, e113.
12. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
13. 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.
14. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*., **9**, R137.
15. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
16. 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.
17. Guillon,N., Tirode,F., Boeva,V., Zynovyev,A., Barillot,E. and Delattre,O. (2009) The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *PLoS ONE*, **4**, e4932.
18. Nielsen,R., Pedersen,T.A., Hagenbeek,D., Moulos,P., Siersbaek,R., Megens,E., Denissov,S., Børgesen,M., Francoijs,K.-J., Mandrup,S. *et al*. (2008) Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev*., **22**, 2953–2967.
19. Welboren,W.-J., van Driel,M.A., Janssen-Megens,E.M., van Heeringen,S.J., Sweep,F.C., Span,P.N. and Stunnenberg,H.G. (2009) ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO J*., **28**, 1418–1428.
20. Bailey,T., Boden,M., Buske,F., Frith,M., Grant,C., Clementi,L., Ren,J., Li,W. and Noble,W. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*., **37**, W202–W208.
21. Shendure,J. (2008) The beginning of the end for microarrays? *Nat. Methods*, **5**, 585–587.
22. Badis,G., Chan,E.T., van Bakel,H., Pena-Castillo,L., Tillo,D., Tsui,K., Carlson,C.D., Gossett,A.J., Hasinoff,M.J., Warren,C.L. *et al*. (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell*, **32**, 878–887.
23. Zhu,C., Byers,K., McCord,R., Shi,Z., Berger,M., Newburger,D., Saulrieta,K., Smith,Z., Shah,M., Radhakrishnan,M. *et al*. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*., **19**, 556–566.
24. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, **15**, 607–611.
25. Pachkov,M., Erb,I., Molina,N. and van Nimwegen,E. (2007) SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*., **35**, D127–D131.
26. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
27. Portales-Casamar,E., Kirov,S., Lim,J., Lithwick,S., Swanson,M.I., Ticoll,A., Snoddy,J. and Wasserman,W.W. (2007) PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol*., **8**, R207.
28. Bergman,C.M., Carlson,J.W. and Celniker,S.E. (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, **21**, 1747–1749.
29. Noyes,M.B., Christensen,R.G., Wakabayashi,A., Stormo,G.D., Brodsky,M.H. and Wolfe,S.A. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–1289.
30. Noyes,M.B., Meng,X., Wakabayashi,A., Sinha,S., Brodsky,M.H. and Wolfe,S.A. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res*., **36**, 2547–2560.

31. Meng,X., Brodsky,M.H. and Wolfe,S.A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.*, **23**, 988–994.
32. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
33. Yi,W. and Zarkower,D. (1999) Similarity of DNA binding and transcriptional regulation by Caenorhabditis elegans MAB-3 and Drosophila melanogaster DSX suggests conservation of sex determining mechanisms. *Development*, **126**, 873–881.
34. Hristova,M., Birse,D., Hong,Y. and Ambros,V. (2005) The Caenorhabditis elegans heterochronic regulator LIN-14 is a novel transcription factor that controls the developmental timing of transcription from the insulin/insulin-like growth factor gene ins-33 by direct DNA binding. *Mol. Cell Biol.*, **25**, 11059–11072.
35. 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.
36. Etchberger,J.F., Lorch,A., Sleumer,M.C., Zapf,R., Jones,S.J., Marra,M.A., Holt,R.A., Moerman,D.G. and Hobert,O. (2007) The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. *Genes Dev.*, **21**, 1653–1674.
37. Berger,M.F. and Bulyk,M.L. (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol. Biol.*, **338**, 245–260.
38. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
39. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
40. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Peña-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
41. Grove,C.A., De Masi,F., Barrasa,M.I., Newburger,D.E., Alkema,M.J., Bulyk,M.L. and Walhout,A.J.M. (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell*, **138**, 314–327.
42. Fulton,D.L., Sundararajan,S., Badis,G., Hughes,T.R., Wasserman,W.W., Roach,J.C. and Sladek,R. (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.
43. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, REVIEWS001.
44. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.