JASPAR 2020: update of the open-access database of transcription factor binding profiles

Oriol Fornes[1,†], Jaime A. Castro-Mondragon[2,†], Aziz Khan[2,†], Robin van der Lee[1], Xi Zhang[1], Phillip A. Richmond[1], Bhavi P. Modi[1], Solenne Correard[1], Marius Gheorghe[2], Damir Baranašić[3,4], Walter Santana-Garcia[5], Ge Tan[6], Jeanne Chèneby[7], Benoit Ballester[7], François Parcy[8], Albin Sandelin[9,*], Boris Lehnard[3,4,*], Wyeth W. Wasserman[1,*], and Anthony Mathelier[2,10*]

[1]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 W 28th Ave, Vancouver, BC V5Z 4H4, Canada.
[2]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway.
[3]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK.
[4]Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W120NN, UK.
[5]Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.
[6]Functional Genomics Centre Zurich, ETH Zurich, Zurich, Switzerland.
[7]Aix Marseille Univ, INSERM, TAGC, Marseille, France.
[8]CNRS, Univ. Grenoble Alpes, CEA, INRA, IRIG-LPCV, 38000 Grenoble, France.
[9]The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, DK2200 Copenhagen N, Denmark.
[10]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway.

†These authors contributed equally to this work as first authors

SUPPLEMENTARY TEXT

ChIP-seq data processing

We downloaded and processed ChIP-seq peak datasets from the following resources:

- ReMap (1): human (hg38) and *A. thaliana* (TAIR10)
- ChIP-atlas (2): *C. elegans* (ce10), *D. melanogaster* (dm3), *S. cerevisiae* (sacCer3)
- GTRD (3): *A. thaliana* (TAIR10), *C. elegans* (WBcel235), *D. melanogaster* (dm6), *S. cerevisiae* (R64), *S. pombe* (ASM294v2)
- CistromeDB (4): human (hg38) and mouse (mm10)
- ModERN (5): *D. melanogaster* (dm6), *C. elegans* (PRJNA13758.WS245)
- DAP-seq (Zea mays, genome version AGPv3.31) datasets from Galli *et al (6)*.

Each dataset was used to predict de novo enriched motifs around the ChIP-seq peak summits using the following protocol:

1. Extract genomic regions of +/- 50bp around the peak summits using BedTools (7).
2. Apply the RSAT peak-motifs tools (8) to discover overrepresented motifs with the following options: `*-markov auto -disco oligos,dyads,positions,local-words -nmotifs 2 -minol 6 -maxol 7 -no_merge_lengths -ci 25 -noov -2str -origin center*`
3. For each of the discovered motifs:
    a. Compute the corresponding PWM using the TFBS perl module (9)
    b. Extract genomic regions of +/- 250bp around peak summits
    c. Predict TFBSs within this sequences using an 85% threshold on the relative scores of the PWMs (9)
    d. Compute the TFBS centrality enrichment p-value (following (10))
4. For each TF, select the PWM associated with the lowest centrality p-value, which satisfies $\log_{10}$(p-value) <= -200.

The TF-binding profiles obtained were then manually curated for inclusion into JASPAR 2020.

This pipeline was developed using Snakemake (11) and the code is available at bitbucket.org/CBGR/jaspar_2020/.

Unvalidated collection

A TF-binding profile was introduced into the unvalidated collection of JASPAR if no further supporting evidence was found in the literature and the PFM satisfied the following criteria:

1. the centrality p-value of predicted TFBSs (see above) was such that $\log_{10}$(p-value) <= -200, if the profile was derived from ChIP-seq;
2. through visual inspection, the profile was found with enough information content and did not correspond to a repetitive motif.

TF binding profiles clustering

TF binding profiles clusterization were obtained using the RSAT matrix-clustering tool (12) with the following parameters: `*-hclust_method average -calc sum -metric_build_tree Ncor -lth w 5 -lth cor 0.6 -lth Ncor 0.4 -label_in_tree name -return json -radial_tree_only*`

The clusters were generated for each taxon. For aesthetics, the radial tree radium and font size were manually adapted according to the number of profiles on each taxon. Code available at bitbucket.org/CBGR/jaspar_2020/.

Sequence logos

All the PFM weblogos were generated as SVG files using the R package ggseqlogo (13)**.**

Transcription Factor Flexible Models (TFFMs)

TF binding profiles in JASPAR 2020 were associated with ChIP-seq data sourced from ReMap (1), ChIP-atlas (2), GTRD (3), and CistromeDB (4) whenever possible. These profiles were used to initialize TFFMs that were trained on genomic regions of -/+ 50bp around the corresponding peak summits. Centrality enrichment p-values were computed using genomic regions +/- 250bp on each side of the peak summits, as for PFMs by considering the best hit per sequence. TFFMs providing a $\log_{10}$(p-value) <= -200 were further introduced into JASPAR.

Profile inference tool

We updated the profile inference tool with the recently described similarity regression approach (14). We followed the methods as indicated, with a few exceptions:

- For the dependent variable (*i.e.* Y), we used Tomtom (version 5.0.5) (15) *e*-values instead of E-score overlaps as values;
- Due to our choice for Y, we did not train logistic regression models; and
- Instead of regressing on individual DBD classes (*e.g.* C2H2 zinc fingers), we regressed on individual DBD compositions (*e.g.* 3x C2H2 zinc fingers such Zif268, 11x zinc fingers such as CTCF, etc.).

For a given TF sequence (*i.e.* query), the updated profile inference tool:

1. Identifies the query's Pfam (16) DBD(s) using hmmscan (version 3.2.1) (17) with the "--domtblout" option and E-value thresholds for models and domains of 10^{-5} and 10^{-2}, respectively (thresholds informed by (18) and (19), respectively);
2. Searches for JASPAR TFs homologous to the query using BLAST+ (version 2.9.0) (20);
3. And selects homologs:
    a. with the same DBD composition than the query; and
    b. whose BLAST+ alignment with the query is above the Rost's sequence identity curve (21).
4. Then, for each selected homolog, evaluates the amino acid sequence similarity of the query and homolog DBDs on the best regression model for that DBD composition; and
5. If the query and homolog DNA-binding specificities are predicted as "highly similar" (*i.e.* Tomtom *e*-value ≤ 10^{-8}), returns the profiles associated with the homolog.

Steps 2 and 3b are performed as previously described in (22). Moreover, in step 4, DBDs are aligned using hmmalign (version 3.2.1) (17). The updated inference tool, regression models and training data are available at https://github.com/wassermanlab/JASPAR-profile-inference.

Genomic tracks

We extended our custom UCSC Genome Browser track data hub (23) beyond human (hg19 and hg38 assemblies) to provide coverage for 6 additional organisms: mouse (mm10), zebrafish (danRer11), *Drosophila melanogaster* (dm6), *Caenorhabditis elegans* (ce10), *Arabidopsis thaliana* (araTha1), and baker's yeast (sacCer3).

DNA sequences were scanned with JASPAR CORE TF-binding profiles for each taxa independently using PWMScan (24). We selected TFBS predictions with a PWM relative score ≥ 0.8 and a *p*-value < 0.05. *P*-values were scaled between 0 (corresponding to a *p*-value of 1) and 1000 (*p*-value ≤ 10^{-10}) for colouring of the genome tracks and to allow for comparison of prediction confidence between different profiles.

Instructions on how to use the genomic tracks on the UCSC Genome Browser are provided on the JASPAR website (http://jaspar.genereg.net/genome-tracks/). Code and data used to create the genomic tracks are available at https://github.com/wassermanlab/JASPAR-UCSC-tracks and http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2020/, respectively.

References

1. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2017) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.

2. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.*, **19**.

3. Yevshin,I., Sharipov,R., Kolmykov,S., Kondrakhin,Y. and Kolpakov,F. (2019) GTRD: a database on gene transcription regulation—2019 update. *Nucleic Acids Res.*, **47**, D100–D105.

4. Zheng,R., Wan,C., Mei,S., Qin,Q., Wu,Q., Sun,H., Chen,C.-H., Brown,M., Zhang,X., Meyer,C.A., *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.*, **47**, D729–D735.

5. Kudron,M.M., Victorsen,A., Gevirtzman,L., Hillier,L.W., Fisher,W.W., Vafeados,D., Kirkey,M., Hammonds,A.S., Gersch,J., Ammouri,H., *et al.* (2018) The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. *Genetics*, **208**, 937–949.

6. Galli,M., Khakhar,A., Lu,Z., Chen,Z., Sen,S., Joshi,T., Nemhauser,J.L., Schmitz,R.J. and Gallavotti,A. (2018) The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat. Commun.*, **9**, 4526.

7. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

8. Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.

9. Lenhard,B. and Wasserman,W.W. (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

10. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.

11. Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

12. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.

13. Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645–3647.

14. Lambert,S.A., Yang,A.W.H., Sasse,A., Cowley,G., Albu,M., Caddick,M.X., Morris,Q.D., Weirauch,M.T. and Hughes,T.R. (2019) Similarity regression predicts evolution of transcription factor sequence specificity. *Nat. Genet.*, **51**, 981–989.

15. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

16. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A., *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.

17. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

18. Xu,Q. and Dunbrack,R.L.,Jr (2012) Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics*, **28**, 2763–2772.

19. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

20. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

21. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, **12**, 85–94.

22. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R., *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–5.

23. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D., *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.

24. Ambrosini,G., Groux,R. and Bucher,P. (2018) PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, **34**, 2483–2484.