JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A. Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R. Kulkarni, Ge Tan, Damir Baranasic, David J. Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier

SUPPLEMENTARY TEXT

ChIP-seq data processing

Human ChIP-seq data

Peak summit locations for human (hg38) TF ChIP-seq data were retrieved from the uniformly processed ChIP-seq datasets stored in the ReMap 2018 database at http://remap.cisreg.eu (1). Each dataset was analyzed to predict PFMs from the ChIP-seq peak summits using the following protocol:

1. Extract genomic regions of +/- 50bp around the peak summits using BedTools (2).
2. Apply the RSAT *peak-motifs* tool (3) to discover de novo motifs with the option '-disco local_words,positions'
3. For each of the discovered motifs:
    a. Compute the corresponding PWM
    b. Extract +/- 250bp around peak summits
    c. Predict TFBSs from these sequences using a 85% threshold on the relative scores of the PWM
    d. Compute a centrality p-value (following (4))
4. For each TF, select the TF binding profile from all associated ChIP-seq datasets providing the best centrality p-value.

The TF binding profiles obtained were then manually curated for inclusion into JASPAR 2018.

Plant ChIP-seq data

Raw reads from the original publications were collected (5–9). Reads were aligned using Bowtie2 (10) to the genomes of *Arabidopsis thaliana* (TAIR10) and *Zea mays* (5b.60). ChIP-seq peaks were called using MACS2 (11). Genomic regions around peak summits were analyzed the same way as for human ChIP-seq data (see above), except the option '-disco local_words,positions,oligos' was used for the RSAT *peak-motifs* tool.

TFFM computation

ChIP-seq data from ReMap 2018 (1) have been collected to construct TFFMs. JASPAR 2018 TF binding profiles were assigned to the ChIP-seq datasets wherever 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. TFFMs providing a centrality p-value < -200 were further assessed for manual curation.

Sequence logos

In this release of JASPAR, we have regenerated all sequence logos as SVG files using the R package ggseqlogo (12).

Matrix clustering

TF binding profile clusterizations were obtained using the RSAT *matrix-clustering* tool (13) 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

UCSC Genome Browser track data hubs

We generated a custom UCSC Genome Browser track data hub (14) containing genome-wide TFBS predictions from PFMs in the JASPAR CORE vertebrates collection. Specifically, for each profile, the human genome assemblies hg19 and hg38 were scanned in parallel using the TFBS Perl module (15) and FIMO (16), as distributed within the MEME suite (version 4.11.2) (17). For scanning the human genome with the BioPerl TFBS module, we converted profiles to PWMs and kept matches with a relative score ≥ 0.8. For the FIMO scan, profiles were reformatted to MEME motifs and matches with a *p*-value < 0.05 were kept. TFBS predictions that were not consistent between the two methods (TFBS Perl module and FIMO) were filtered out. The remaining TFBS predictions were converted to genome tracks and colored according to their FIMO *p*-value (scaled between 0-1000, where 0 corresponds to a *p*-value of 1 and 1000 to a *p*-value ≤ 10^{-10}) to allow for comparison of prediction confidence between different profiles. The tracks are collected as a data hub that can be visualized in the UCSC Genome Browser or downloaded for custom analysis (http://jaspar.genereg.net/genome-tracks/). Code and data used to create the UCSC tracks are available at https://github.com/wassermanlab/JASPAR-UCSC-tracks. The underlying BED files and individual matches for each TF binding profile on the human genome (hg19 and hg38 genome assemblies) are available at http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/.

References

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

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

3. 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.

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

5. Eveland,A.L., Goldshmidt,A., Pautler,M., Morohashi,K., Liseron-Monfils,C., Lewis,M.W., Kumari,S., Hiraga,S., Yang,F., Unger-Wallace,E., *et al.* (2014) Regulatory modules controlling maize inflorescence architecture. *Genome Res.*, **24**, 431–443.

6. Verkest,A., Abeel,T., Heyndrickx,K.S., Van Leene,J., Lanz,C., Van De Slijke,E., De Winne,N., Eeckhout,D., Persiau,G., Van Breusegem,F., *et al.* (2014) A generic tool for transcription factor target gene discovery in Arabidopsis cell suspension cultures based on tandem chromatin affinity purification. *Plant Physiol.*, **164**, 1122–1133.

7. Li,C., Qiao,Z., Qi,W., Wang,Q., Yuan,Y., Yang,X., Tang,Y., Mei,B., Lv,Y., Zhao,H., *et al.* (2015) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. *Plant Cell*, **27**, 532–545.

8. Cui,X., Lu,F., Qiu,Q., Zhou,B., Gu,L., Zhang,S., Kang,Y., Cui,X., Ma,X., Yao,Q., *et al.* (2016) REF6 recognizes a specific DNA sequence to demethylate H3K27me3 and regulate organ boundary formation in Arabidopsis. *Nat. Genet.*, **48**, 694–699.

9. Birkenbihl,R.P., Kracher,B. and Somssich,I.E. (2017) Induced Genome-Wide Binding of Three Arabidopsis WRKY Transcription Factors during Early MAMP-Triggered Immunity. *Plant Cell*, **29**, 20–38.

10. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

11. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

12. Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, 10.1093/bioinformatics/btx469.

13. 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.

14. 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.

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

16. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

17. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–8.