Published online 10 November 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D141–D149***
https://doi.org/10.1093/nar/gkab1039

Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites

Henry E. Pratt†, Gregory R. Andrews†, Nishigandha Phalke, Jack D. Huey, Michael J. Purcaro, Arjan van der Velde, Jill E. Moore and Zhiping Weng*

Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA, USA

Received September 15, 2021; Revised October 12, 2021; Editorial Decision October 14, 2021; Accepted October 14, 2021

ABSTRACT

The human genome contains ∼2000 transcriptional regulatory proteins, including ∼1600 DNA-binding transcription factors (TFs) recognizing characteristic sequence motifs to exert regulatory effects on gene expression. The binding specificities of these factors have been profiled both *in vitro*, using techniques such as HT-SELEX, and *in vivo*, using techniques including ChIP-seq. We previously developed Factorbook, a TF-centric database of annotations, motifs, and integrative analyses based on ChIP-seq data from Phase II of the ENCODE Project. Here we present an update to Factorbook which significantly expands the breadth of cell type and TF coverage. The update includes an expanded motif catalog derived from thousands of ENCODE Phase II and III ChIP-seq experiments and HT-SELEX experiments; this motif catalog is integrated with the ENCODE registry of candidate cis-regulatory elements to annotate a comprehensive collection of genome-wide candidate TF binding sites. The database also offers novel tools for applying the motif models within machine learning frameworks and using these models for integrative analysis, including annotation of variants and disease and trait heritability. Factorbook is publicly available at www.factorbook.org; we will continue to expand the resource as ENCODE Phase IV data are released.

INTRODUCTION

The human genome includes the instructions for producing an estimated 2000 transcriptional regulatory proteins that interact with DNA in order to modulate regulatory element activity and gene expression (1). These include both sequence-specific transcription factors (TF) and co-binding factors and RNA polymerase complex members that lack sequence specificity but cooperate with sequence-specific TFs. TFs typically possess a DNA binding domain (DBD) which recognizes 6–20 base-pair (bp) long characteristic consensus binding sequences, or a *motif*, present within the TF's target regulatory elements. TFs may be grouped according to several known families of DBDs which frequently recognize similar DNA sequences. The binding specificities of these factors have been profiled both invitro, using techniques such as HT-SELEX (2), and in vivo, using techniques including ChIP-seq (3,4).

Numerous resources have been developed to catalog TF motifs. The HOCOMOCO catalog (5) indexes binding specificities for nearly 700 human TFs and >400 mouse TFs identified from ChIP-seq and HT-SELEX data, and the JASPAR catalog (6) contains more than 700 curated non-redundant binding profiles for eukaryotic TFs. The earlier UniPROBE (7) resource contains more than 700 binding profiles from *in vitro* protein binding microarray experiments, and the broader CisBP (8) incorporates data from these sources and others, including our previous release of Factorbook (9,10), to annotate both measured and inferred binding profiles for thousands of TFs across hundreds of species.

Here, we present an update to Factorbook which leverages the extensive ChIP-seq data available through Phase III of the ENCODE Project to build a comprehensive TF motif catalog for ∼750 TFs. We also provide two notable features not available in existing catalogs to our knowledge. First, we catalog motif models built using convolutional neural networks (CNNs), which are finding increasing applications in genomics including in discovering TF motifs and predicting TF binding (11–13); these will be easily integrated into future models for transfer learning. Second, we leverage the ENCODE Registry of candidate Cis-Regulatory Elements (cCREs) (14) to provide a genome-wide catalog of motif sites in regulatory sequences, with associated epigenetic and evolutionary annotations; we illustrate the usefulness of this catalog for downstream applications by using it to quantify trait heritability using partitioned LD score regression (15).

*To whom correspondence should be addressed. Tel: +1 508 856 8866; Fax: +1 508 856 0017; Email: zhiping.weng@umassmed.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

OVERVIEW

Factorbook is a transcription factor-centric database cataloging information for 694 distinct human and 62 mouse transcriptional regulatory proteins profiled in 249 and 38 human and mouse cell types. These include sequence-specific TFs, co-binding factors, and members of the RNA polymerase complex; existing knowledge about sequence specificity is provided from a recent review by Lambert *et al*. (1) and the ENCODE Portal. The primary entry point is a factor search (Figure 1A), which directs users to a detailed factor information page curated information from various external sources, including NCBI, Uniprot, HGNC and Ensembl (Figure 1B). Data tables listing the available factors and cell types are also available for browsing (Figure 1C, indicating whether a factor is sequence specific). Furthemore, we display the expression levels of each factor using ENCODE RNA-seq data in a variety of primary cells, primary tissues and cell lines (Figure 1D), and also display primary data resulting from several integrative analyses.

We additionally replaced the Wiki-based technology of the first Factorbook release (10) with a ReactJS-based frontend and GraphQL Application Programming Interface which offers improvements and novel capabilities including (i) facilitated programmatic data access, (ii) interactive visualizations and (iii) interactive analyses described in detail below, including single nucleotide polymorphism (SNP) annotation and intersection of resources with user-uploaded BED files.

A comprehensive motif catalog derived from ChIP-seq and HT-SELEX data

A cornerstone of the primary data contained within Factorbook is a comprehensive catalog of human and mouse TF recognition motifs. We expanded on existing catalogs (5,16) in two ways: first, we aimed to annotate binding sequences for as many transcription factors as possible, including TFs not profiled by other efforts to our knowledge (5,8,17,18); and second, we aimed to provide motifs in optimal formats that will integrate seamlessly into the variety of machine learning frameworks which are actively being developed to study TF binding (12,13,19) in addition to conventional downstream analysis using tools such as the MEME suite (20).

We thus designed two complementary pipelines for de novo motif identification (Methods). First, we applied our previous MEME-based pipeline (9) to the top 500 strongest ChIP-seq peaks from each TF ChIP-seq dataset produced during the first three phases of ENCODE. This pipeline identifies up to five motifs per dataset; these are subsequently filtered by quality control metrics we developed previously (9), including peak centrality and enrichment in peaks as compared with randomly-selected non-peak genomic sequences with similar GC content. Second, we applied a convolutional neural network that we recently developed, *ZMotif* (manuscript in preparation), for discovery of motifs from HT-SELEX data which can be used in downstream analysis with both other deep learning frameworks and conventional tools.

MEME identifies 6921 motifs from human and mouse ChIP-seq datasets. This includes several redundant motifs for well-profiled factors such as CTCF and REST; we therefore applied UMAP (21,22) to the motifs to map them into a reduced-dimension space (Figure 2A), revealing clusters of known and novel motifs. The number of motif clusters ranges from 100 to 300 depending on hyperparameter selection; we make several interactive UMAP plots with different hyperparameters available through Factorbook to aid users in identifying motif clusters for downstream analysis. We then applied TOMTOM (23) to compare our MEME motifs against the HOCOMOCO and JASPAR catalogs (5,16); we find that the Factorbook catalog includes nearly 100% of the motifs in these two sources, and further identifies novel motifs not present therein including candidate motifs for 101 TFs that have been previously classified as 'likely sequence-specific factors' (1) (Figure 2B). One example novel motif is that of ZNF407 (Figure 2C), which shows high evolutionary conservation (Figure 2D), prefers to reside in the center of ChIP-seq peaks (Figure 2E), and is protected from DNase I cleavage (Figure 2F). Each individual factor page contains indexed lists of all motifs identified by MEME; these and the HT-SELEX motifs can also be searched through Factorbook either by consensus sequence or by uploading motifs in MEME format to match against the catalog; visual results are provided in real time (Figure 2G).

ZMotif identifies a total of 6700 motifs from HT-SELEX two public HT-SELEX datasets (2,24); we perform similar UMAP projection and cluster annotation for these motifs (Figure 3A). SELEX motifs are also available for visualization on each TF's information page; motifs are shown for each HT-SELEX cycle along with QC statistics, including the fraction of HT-SELEX reads containing the motif and a receiver operator characteristic curve showing how well the motif distinguishes reads from control sequences (Figure 3B). Motifs from both methods are made available for download as PWMs in MEME format; deep-learned filters are further available in a Numpy format that can be directly used by models such as neural networks for transfer learning. The motif page also displays TOMTOM matches against HOCOMOCO and JASPAR for each motif.

Genome-wide instances of motifs in ChIP-seq peaks

For ChIP-seq motifs identified with MEME, we use the FIMO tool from the MEME suite (20) to scan irreproducible discovery rate (IDR) thresholded ChIP-seq peaks from human TF ChIP-seq datasets for motif instances, filtering at the standard *P*-value cutoff of 10^{-4}. We identified 110 001 176 (overlapping allowed) motif instances in total; when overlapping regions are merged, the motif sites number 6 720 871. These instances are available for download through Factorbook in BED format through the associated motif page; additionally, we have implemented a novel database-backed service allowing users to upload their own BED files for real-time intersection with motif instances, accessible through the same page.

To aid the user in evaluating motif quality, we compute three metrics for the motif instances. First, we assess the evolutionary conservation of the motif instance and

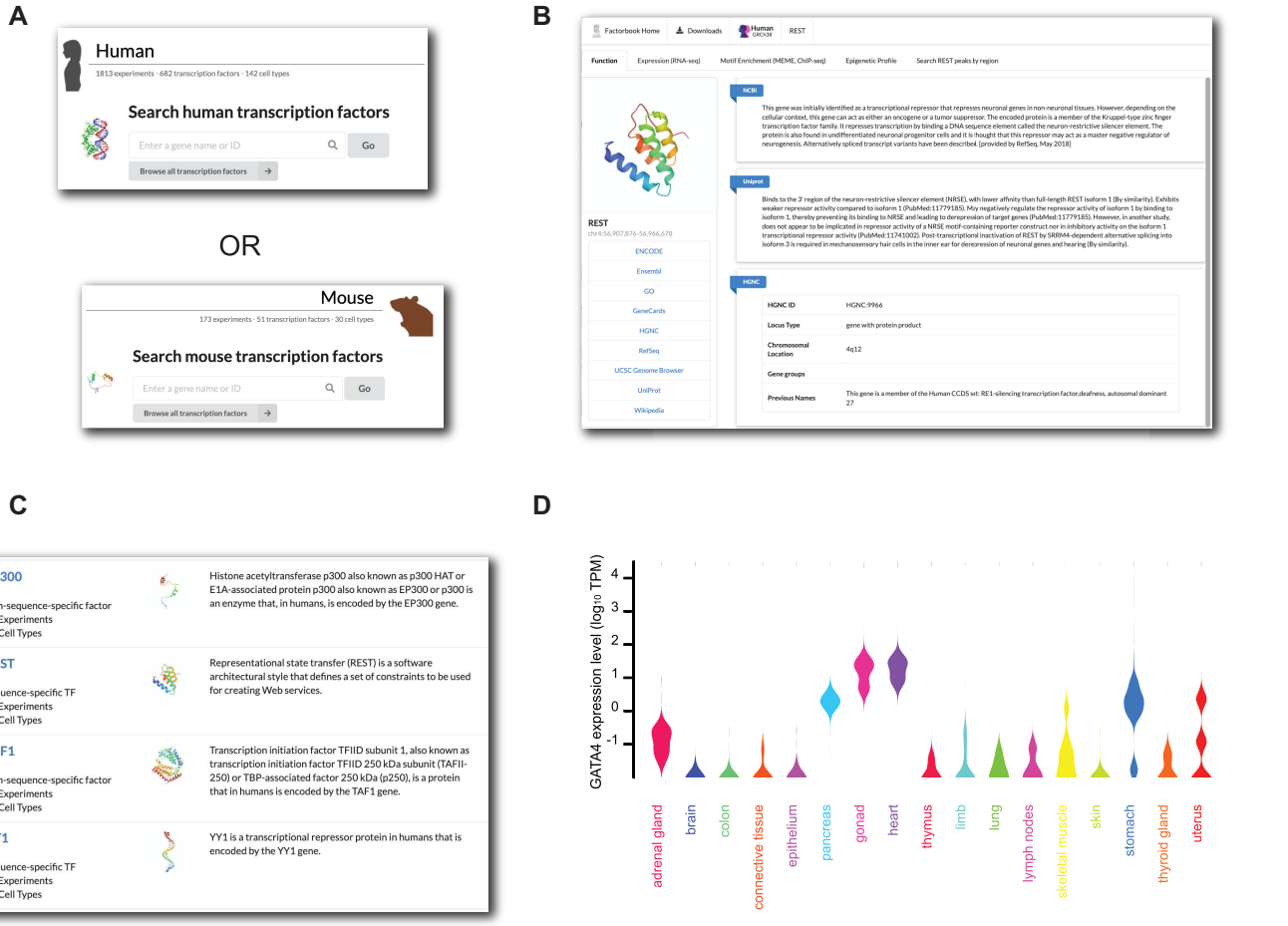

Figure 1. Overview of the main Factorbook interface. (**A**) An example of the main TF search for human and mouse. (**B**) The information page for REST, highlighting information curated from external sources. (**C**) The transcription factor table, listing all 682 human TFs with available data for browsing. (**D**) Factorbook's display of the RNA-seq expression profile for *GATA4* in human embryonic primary tissues.

surrounding peak sequence using phyloP scores across collections of vertebrates and mammals (25) (Figure 2D). Second, we compute the distance between each motif instance and the corresponding peak summit since the instances of bona fide motifs tend to be near peak summits (Figure 2E). Third, we compute the distribution of DNase-seq and ATAC-seq reads around each motif instance when matched data are available in the corresponding cell type (Figure 2F). We display histograms and aggregated signal profiles for each of these three metrics on the motif page. In general, high quality motif instances are central within peaks, are more evolutionarily conserved than the surrounding peak sequences, and are less accessible to DNase I and Tn5 than surrounding non-motif peak sequences.

Genome-wide motif instances in candidate *cis*-regulatory elements

We previously developed the ENCODE Registry of cCREs, a collection of nearly 1 million human regulatory elements. The cCREs are the subset of representative DNase hypersensitive sites (rDHSs) with high signals for H3K4me3, a promoter mark, H3K27ac, an enhancer mark, and the insulator-binding protein CTCF (14). The Registry integrates data from >1000 cell types, while the ChIP-seq data included in Factorbook derives primarily from five human cell lines (HepG2, K562, HEK293, GM12878 and MCF-7). Accordingly, motif instances in ChIP-seq peaks are most common within cCREs and rDHSs active in cell types biologically similar to these five cell lines. For example, embryonic bone marrow and liver are responsible for hematopoiesis, and nearly twice as many rDHSs active in those embryonic tissues contain peak motif instances as those active in other tissues, in line with the prevalence of ChIP-seq data in the red blood cell precursor K562 (Figure 4A, with the two tissues marked with asterisks). To capture candidate motif sites in other cell types, we applied FIMO to identify instances of all the high-quality human and mouse motifs from both our ChIP-seq MEME catalog and our HT-SELEX ZMotif catalog within cCREs and rDHSs. Given the larger scale of the rDHS set, we used a more stringent FIMO *P*-value threshold of 10^{-6} for MEME ChIP-seq motifs and 10^{-5} for ZMotif HT-SELEX motifs to reduce false positives. We also generated sets at more stringent thresholds of $P < 10^{-7}$ and $P < 10^{-8}$ for users preferring even higher confidence sets (Figure 4B).

In total, we define a catalog of 33 452 885 (overlapping allowed) candidate regulatory transcription factor motif sites

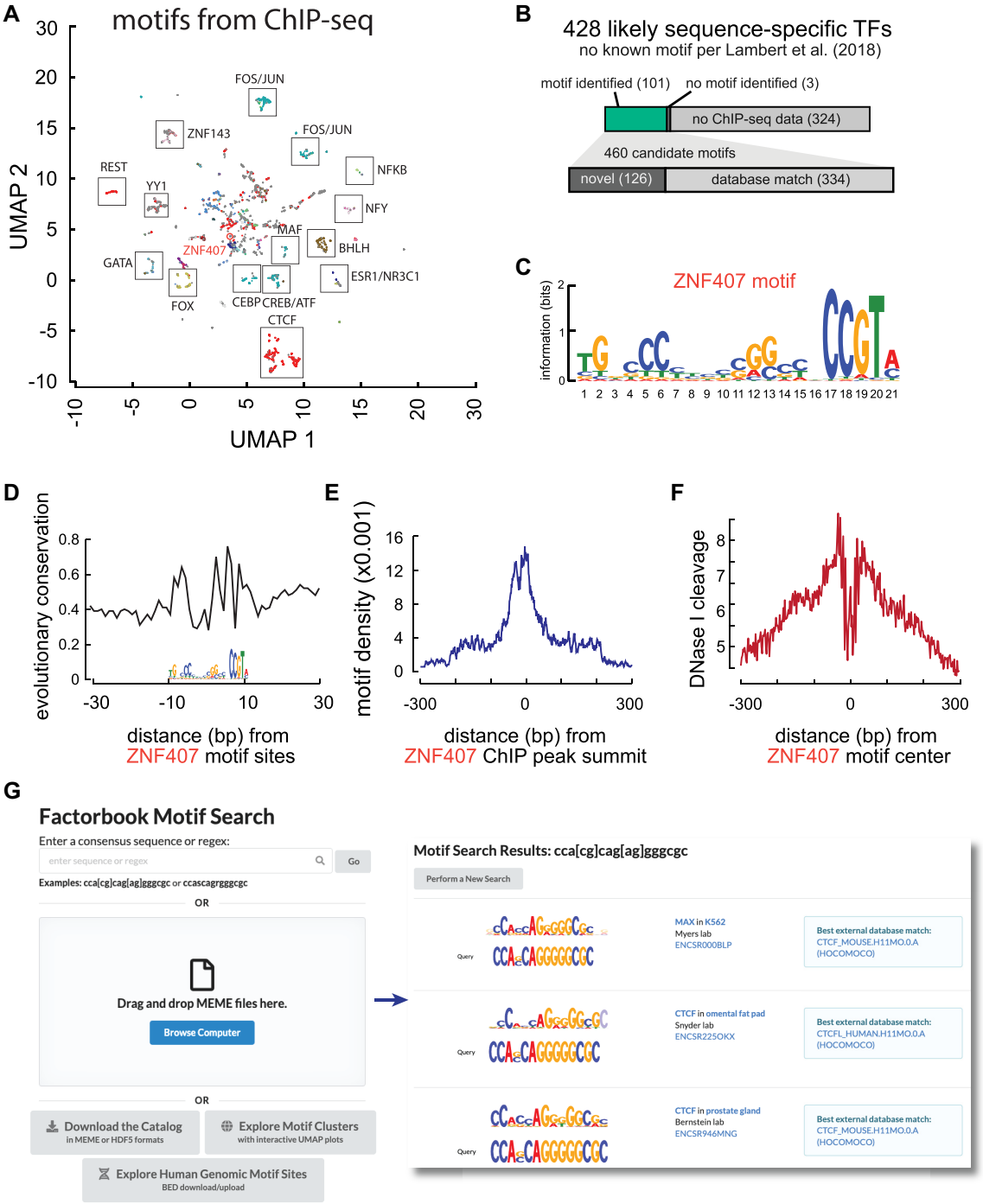

Figure 2. Overview of the Factorbook MEME ChIP-seq motif catalog. (**A**) A UMAP projection of motifs passing QC, with some DNA binding domains colored (C2H2 Zinc Finger in red, GATA in teal, nuclear receptor in dark purple) and several motif clusters annotated. (**B**) Overview of novel motifs cataloged by Factorbook. (**C**) A novel motif for ZNF407, with supporting evolutionary conservation (phyloP 100-way) (**D**), peak centrality (**E**), and DNase-seq footprint (**F**) aggregate plots. (**G**) The Factorbook motif search interface, showing matches for the consensus sequence for the CTCF motif.

within rDHSs using MEME-identified ChIP-seq motif sites at a FIMO *P*-value of $<10^{-6}$. These number 7 189 228 when overlapping motif sites are merged; of these, 4 902 200 (68.2%) are not present in the TF ChIP-seq peak motif site catalog. More than 95% of rDHSs contain at least one of these motif sites, with most having between 1 and 4; in total, the motif sites cover 30% of the base pairs in rDHS sequence (Figure 4B). The more stringent sets, numbering 2 840 049 and 1 572 634 non-overlapping motif sites, respectively, cover a smaller portion of sequence within a smaller number of rDHSs (Figure 4B). We aggregate evolutionary conservation and DNase-seq reads at each of the motif sites at the most lenient threshold of $P < 10^{-6}$; this highlights that these motifs are significantly more conserved

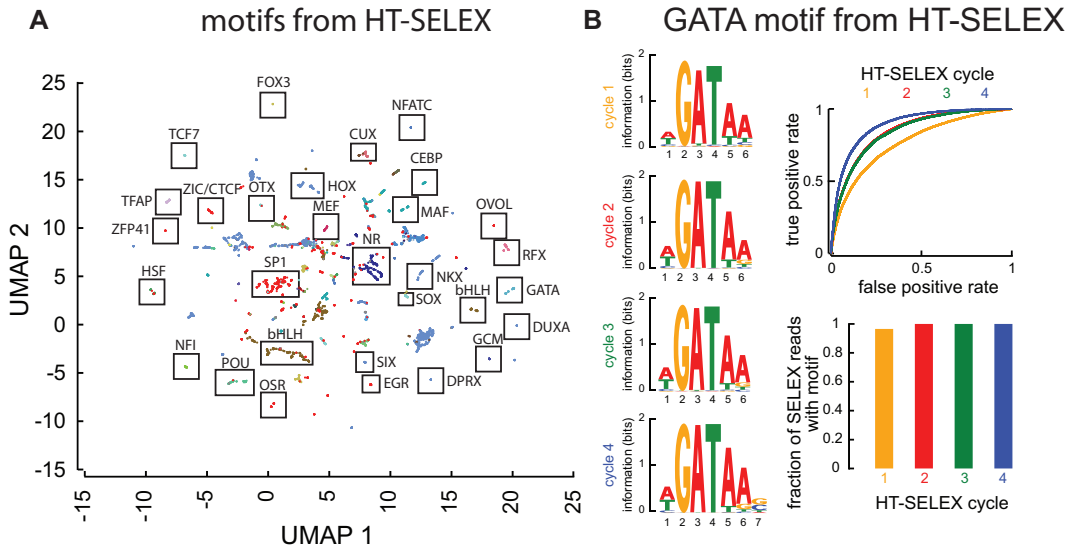

Figure 3. The Factorbook HT-SELEX motif catalog. (**A**) A UMAP projection of 6,700 HT-SELEX motifs with clusters annotated. (**B**) Example of the SELEX motif interface for GATA4, showing motifs for each of the four SELEX cycles for a GATA4 HT-SELEX experiment as well as two motif enrichment metrics, an ROC curve (top) and a readout of the fraction of reads at each cycle containing the motif (bottom).

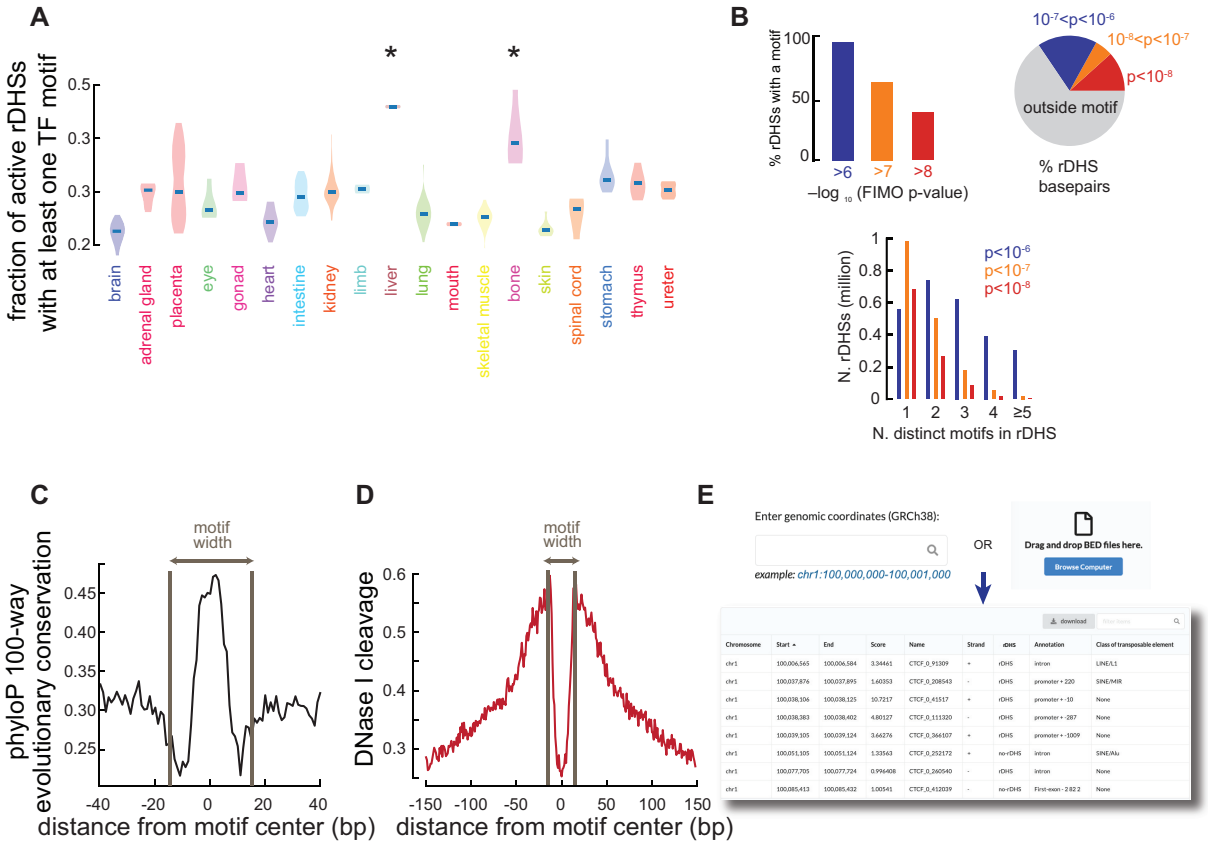

Figure 4. Overview of the Factorbook regulatory motif site catalog. (**A**) Fraction of representative DNase hypersensitive sites (rDHSs) active in a variety of embryonic cell types containing at least one motif identified in a ChIP-seq peak; cell types similar to K562 (fetal liver) and GM12878 (fetal bone) are most enriched due to the prevalence of ChIP-seq for those cell lines. (**B**) Distribution of motif sites within rDHSs and fraction of rDHS sequence covered by motifs at different thresholds. (**C**) Aggregated evolutionary conservation and (**D**) DNase I cleavage in embryonic kidney at 10,000 randomly chosen motif sites from the catalog within rDHSs active in embryonic kidney. (**E**) The motif site search interface, showing CTCF motif sites in a genomic region on chromosome 1.

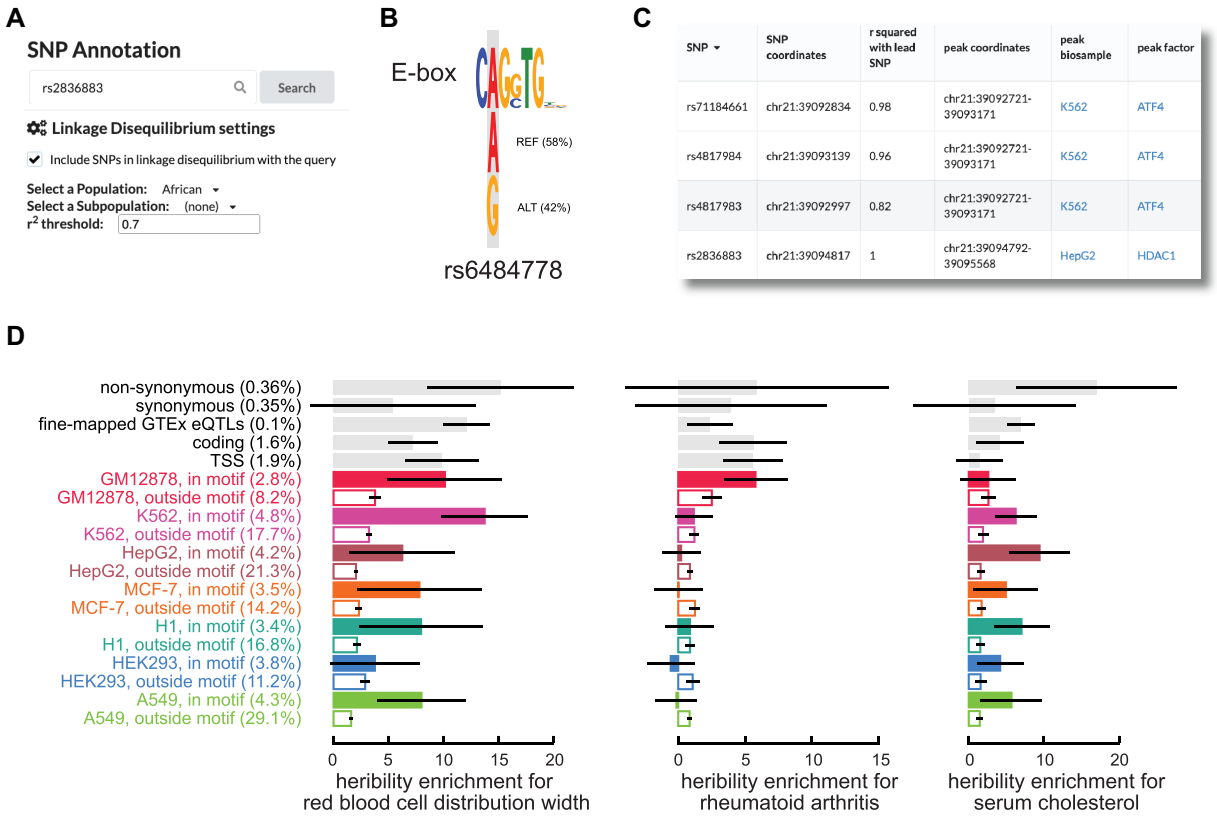

Figure 5. Annotation of variants with Factorbook. (**A**) The variant annotation interface; the user inputs a SNP rsID and can optionally select to include SNPs in LD with the searched SNP. (**B**) Example search results for a given SNP showing an impact on an E-box motif from the Factorbook rDHS motif site catalog. (**C**) The variant peak intersection view showing ChIP-seq peaks intersecting variants in LD with an example query. (**D**) Heritability enrichment for a variety of traits within motif sites identified in ChIP-seq peaks from seven distinct ENCODE cell lines, computed using partitioned LD score regression.

than surrounding rDHS sequences (Figure 4C) and are also less accessible to DNase I (Figure 4D), suggesting the protection of the associated DNA by the bound TF; these findings regarding conservation and DNase I protection hold even for motifs present in rDHSs but not ChIP-seq peaks, supporting the idea that at least a subset of these motifs are true transcription factor binding sites which would be identified if ChIP-seq were performed in the correct biological context. These metrics are available through the motif page for each TF, as are the complete sets of instances in BED format. Instances can also be searched by BED file upload (Figure 4E). We performed the same analysis for HT-SELEX, identifying 9 205 043 distinct non-overlapping motif sites within rDHSs at a FIMO *P*-value $<10^{-5}$, also accounting for ∼30% of rDHS sequence. These sites are also available for download and searching through Factorbook.

Tools for integrating motifs with GWAS results

It is hypothesized that many non-coding disease-associated variants confer risk for a given trait or disease by impacting transcription factor recognition sequences within regulatory elements. We designed an interactive platform within Factorbook to facilitate the annotation of SNPs with candidate impacts on instances of human and mouse TF motifs in our catalog. Users input a SNP's rsID and optionally select a population or subpopulation from the 1000 Genomes Project from which to include SNPs in linkage disequilibrium (LD) (Figure 5A). Factorbook intersects these SNPs with motif instances in real time and displays the results, sorted by predicted impact on the motif (Figure 5B shows an example). Simultaneously, Factorbook searches all TF peaks from ENCODE that intersect the SNPs, allowing users to determine if there is direct ChIP-seq support for any candidate TFs identified by motif analysis (Figure 5C).

Additionally, we built heritability models for partitioned LD score regression (15) from the motif sites in our catalog within ChIP-seq peaks. We provide one model which includes the complete set of motif instances as well as one which includes motif sites grouped by the cell type in which the corresponding ChIP-seq peak was identified. Heritability for traits is highly enriched within motif sequences, with enrichment generally being the strongest within motif sites identified in ChIP-seq peaks from disease-relevant cell lines. For example, heritability for red blood cell distribution width is most strongly enriched in TF motifs from K562, an erythroid cell line, heritability for rheumatoid arthritis, an autoimmune condition, is strongly enriched within TF motifs in ChIP-seq peaks from GM12878, a B-cell line; and heritability for serum cholesterol level is most strongly enriched in TF motifs from HepG2, a hepatocyte cell line (Figure 5D). These models are available for download through Factorbook for application to the summary statistics of additional genome-wide association studies (GWAS); we

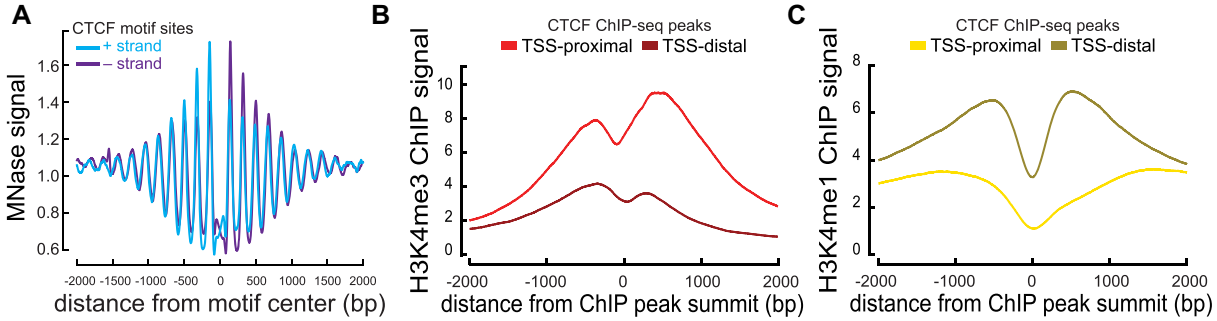

Figure 6. Epigenetic signal aggregation profiles on Factorbook. (**A**) Aggregated MNase-seq signal around CTCF motif sites, highlighting asymmetry depending on motif orientation. (**B**) Aggregated H3K4me3 ChIP-seq signal around CTCF peak summits as displayed in Factorbook. (**C**) Aggregated H3K4me1 ChIP-seq signal around CTCF peak summits as displayed in Factorbook.

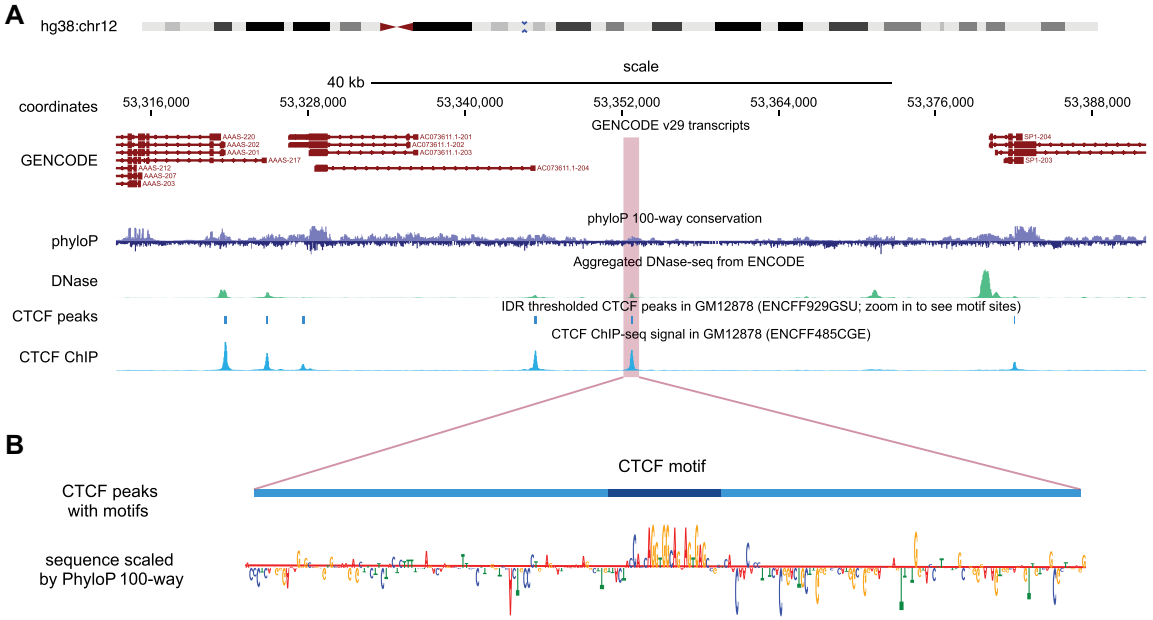

Figure 7. The Factorbook embedded genome browser view. (**A**) For a given experiment, ChIP-seq signal and IDR peaks from the ENCODE portal are displayed alongside transcripts and evolutionary conservation. (**B**) When the view is zoomed in, motif sites from the Factorbook catalog are displayed along with the underlying sequence scaled according to evolutionary conservation using a novel sequence importance track.

provide a Docker image and associated scripts for running this analysis through GitHub.

High-resolution nucleosome and epigenetic profiles around binding sites

In the previous iteration of Factorbook, we generated aggregated epigenetic signal profiles, including histone modifications and nucleosome positions from MNase-seq, around the summits of TSS-proximal and TSS-distal transcription factor ChIP-seq peaks. We find that aggregating around motif instances rather than peak summits improves the resolution and phasing of epigenetic signals; additionally, it offers a natural orientation which reveals asymmetries in the organization of features around regulatory sites which have previously been suggested to be of biological relevance (26); we highlight, for example, asymmetric positioning of nucleosomes assayed by MNase-seq around oriented CTCF motif sites (Figure 6A). Therefore, on each factor's page, Factorbook now displays aggregated signal profiles around motif instances for all cataloged motifs in addition to profiles surrounding ChIP-seq peak summits (illustrated for histone marks H3K4me3 and H3K4me1 around CTCF motif sites, Figure 6B); we separate motif sites according to TSS proximity, which highlights differences in epigenetic profiles around TSS-proximal and TSS-distal sites.

Tools for machine learning and integrative analysis

Building deep learning models which can predict regulatory readouts is a primary focus of ongoing computational efforts in regulatory genomics. Prediction targets include cross-cell type transcription factor binding (13,19) as well as epigenetic sequence profiles in a given cell type (12). Frequently, these models include one-dimensional convolutional neural network layers which learn predictive sequences including transcription factor motifs. Transfer learning, or using existing models as starting points for

new models applied to new tasks, has been proposed for sequence-based problems in biology (27,28); seeding new models with our motif features could reduce training time while improving the predictive power and human interpretability of learned models. To aid users in applying the kernels learned by our neural networks from HT-SELEX data, we provide the option to export all ZMotif-derived motifs in Numpy format. These kernels may then be loaded into Python and used to seed weights in convolutional layers in a variety of commonly-used machine learning packages including PyTorch and Tensorflow. For users interested in more conventional downstream analysis, we also offer the option to export all MEME- and ZMotif-derived motifs as PWMs in MEME format, which may then be used by a variety of downstream tools including those in the MEME suite (20).

Genomic visualization of motifs and TF binding sites

Human interaction remains essential in interpreting the biological significance of transcription factor motifs and regulatory elements. We implemented lightweight embedded genome visualizations within Factorbook which display TF peaks from ENCODE datasets alongside human and mouse motif instances from our resource. Evolutionary conservation and relevant epigenetic signal profiles from the given species are displayed alongside gene and transcript tracks (Figure 7A). Additionally, we have designed a novel sequence importance track which scales bases in the reference sequence according to a signal track of associated scores; we demonstrate the use of this track to highlight evolutionarily conserved motif instances using PhyloP as the scaling score (Figure 7B). We have engineered this track to extend easily to additional scores provided through BigWig format signal tracks. In addition, we have designed a public Factorbook trackhub for release on the UCSC Genome Browser (29).

Planned future expansion

We will further expand Factorbook to include integrative analysis of ENCODE Phase IV ChIP-seq, DNase-seq and ATAC-seq data as they are released, and will update our motif and rDHS catalog accordingly. Additionally, we will update our motif instance catalog with the release of the final version of the ENCODE Registry of cCREs at the conclusion of ENCODE Phase IV. We additionally plan to expand coverage to additional in vitro motifs derived from other assays such as protein binding microarray (PBM); for completeness, we will also offer in vitro motifs derived from existing methods, including the SELEX R package (30,31), for comparison with our neural network-derived motifs. Additionally, we will expand comparison of our motif catalog to other existing databases, including CisBP, via TOMTOM.

DATA AVAILABILITY

All data can be downloaded at https://factorbook.org/downloads.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NIH [U24HG009446]. Funding for open access charge: NIH [U24HG009446].

Conflict of interest statement. Zhiping Weng co-founded and serves as a board member and scientific advisor for Rgenta Inc.

REFERENCES

1. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
2. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
3. 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.
4. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
5. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
6. Fornes,O., Castro-Mondragon,J.A., Khan,A., van der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašić,D. *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87–D92.
7. 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.
8. 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.
9. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
10. Wang,J., Zhuang,J., Iyer,S., Lin,X.-Y., Greven,M.C., Kim,B.-H., Moore,J., Pierce,B.G., Dong,X., Virgil,D. *et al.* (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.*, **41**, D171–D176.
11. Quang,D. and Xie,X. (2019) FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, **166**, 40–47.
12. Avsec,Ž., Weilert,M., Shrikumar,A., Krueger,S., Alexandari,A., Dalal,K., Fropf,R., McAnany,C., Gagneur,J., Kundaje,A. *et al.* (2021) Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*, **53**, 354–366.
13. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
14. ENCODE Project Consortium, Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.

15. Finucane,H.K., Bulik-Sullivan,B., Gusev,A., Trynka,G., Reshef,Y., Loh,P.-R., Anttila,V., Xu,H., Zang,C., Farh,K. *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.
16. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.
17. 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–D115.
18. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
19. Chen,C., Hou,J., Shi,X., Yang,H., Birchler,J.A. and Cheng,J. (2021) DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. *BMC Bioinformatics*, **22**, 38.
20. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME suite. *Nucleic Acids Res.*, **43**, W39–W49.
21. Becht,E., McInnes,L., Healy,J., Dutertre,C.-A., Kwok,I.W.H., Ng,L.G., Ginhoux,F. and Newell,E.W. (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, **37**, 38–44.
22. McInnes,L., Healy,J. and Melville,J. (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv doi: https://arxiv.org/abs/1802.03426, 18 September 2020, preprint: not peer reviewed.
23. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
24. Yin,Y., Morgunova,E., Jolma,A., Kaasinen,E., Sahu,B., Khund-Sayeed,S., Das,P.K., Kivioja,T., Dave,K., Zhong,F. *et al.* (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, **356**, eaaj2239.
25. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
26. Kundaje,A., Kyriazopoulou-Panagiotopoulou,S., Libbrecht,M., Smith,C.L., Raha,D., Winters,E.E., Johnson,S.M., Snyder,M., Batzoglou,S. and Sidow,A. (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.*, **22**, 1735–1747.
27. Mignone,P., Pio,G., D'Elia,D. and Ceci,M. (2020) Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinformatics*, **36**, 1553–1561.
28. Heinzinger,M., Elnaggar,A., Wang,Y., Dallago,C., Nechaev,D., Matthes,F. and Rost,B. (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, **20**, 723.
29. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
30. Riley,T.R., Slattery,M., Abe,N., Rastogi,C., Liu,D., Mann,R.S. and Bussemaker,H.J. (2014) SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol. Biol.*, **1196**, 255–278.
31. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.