Nucleic Acids Research, 2023, 1–10
https://doi.org/10.1093/nar/gkad1077
Database Issue

HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors

Ilya E. Vorontsov[1,†], Irina A. Eliseeva[2,†], Arsenii Zinkevich[1,3,†], Mikhail Nikonov[3], Sergey Abramov[1,4], Alexandr Boytsov[1,4], Vasily Kamenets[1,5,6], Alexandra Kasianova[7,8], Semyon Kolmykov[9], Ivan S. Yevshin[10], Alexander Favorov[1,11], Yulia A. Medvedeva[12], Arttu Jolma[13], Fedor Kolpakov[9,14], Vsevolod J. Makeev[1,5,6,*] and Ivan V. Kulakovskiy[1,2,15,*]

[1]Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991 Moscow, Russia
[2]Institute of Protein Research, Russian Academy of Sciences, 142290 Pushchino, Russia
[3]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991 Moscow, Russia
[4]Altius Institute for Biomedical Sciences, 98121 Seattle, WA, USA
[5]Moscow Institute of Physics and Technology, 141700 Dolgoprudny, Russia
[6]FSBIS Institute of Biochemistry and Genetics of the Russian Academy of Sciences, 450054 Ufa, Russia
[7]Skolkovo Institute of Science and Technology, 121205 Moscow, Russia
[8]Institute for Information Transmission Problems of the Russian Academy of Sciences. 127051 Moscow, Russia
[9]Department of Computational Biology, Sirius University of Science and Technology, 354340 Sirius, Krasnodar region, Russia
[10]Biosoft.Ru LLC, 630090 Novosibirsk, Russia
[11]Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[12]Research Center of Biotechnology RAS, Russian Academy of Sciences, 119071 Moscow, Russia
[13]Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada
[14]Bioinformatics Laboratory, Federal Research Center for Information and Computational Technologies, 630090 Novosibirsk, Russia
[15]Laboratory of Regulatory Genomics, Institute of Fundamental Medicine and Biology, Kazan Federal University, 420008 Kazan, Russia

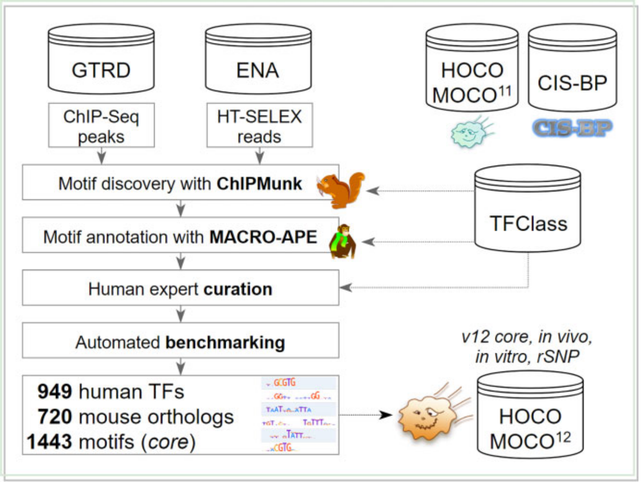
[*]To whom correspondence should be addressed. Tel: +7 495 514 02 18; Email: ivan.kulakovskiy@gmail.com
Correspondence may also be addressed to Vsevolod J. Makeev. Email: vsevolod.makeev@vigg.ru
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Abstract

We present a major update of the HOCOMOCO collection that provides DNA binding specificity patterns of 949 human transcription factors and 720 mouse orthologs. To make this release, we performed motif discovery in peak sets that originated from 14 183 ChIP-Seq experiments and reads from 2554 HT-SELEX experiments yielding more than 400 thousand candidate motifs. The candidate motifs were annotated according to their similarity to known motifs and the hierarchy of DNA-binding domains of the respective transcription factors. Next, the motifs underwent human expert curation to stratify distinct motif subtypes and remove non-informative patterns and common artifacts. Finally, the curated subset of 100 thousand motifs was supplied to the automated benchmarking to select the best-performing motifs for each transcription factor. The resulting HOCOMOCO v12 core collection contains 1443 verified position weight matrices, including distinct subtypes of DNA binding motifs for particular transcription factors. In addition to the core collection, HOCOMOCO v12 provides motif sets optimized for the recognition of binding sites *in vivo* and *in vitro*, and for annotation of regulatory sequence variants. HOCOMOCO is available at https://hocomoco12.autosome.org and https://hocomoco.autosome.org.

Received: September 22, 2023. **Revised:** October 17, 2023. **Editorial Decision:** October 25, 2023. **Accepted:** October 26, 2023

Graphical abstract



Introduction

Computational annotation of transcription factor binding sites (TFBS) remains an essential pillar supporting the dome of gene regulation studies. The most common context is the recognition of individual TFBS in genome regulatory regions (1), e.g. as supportive evidence for transcription factor (TF) target genes (2). Predicted TFBS can reveal the genomic locations and the structure of regulatory regions and thus provide information on the composition of transcriptional regulatory complexes (3,4). In synthetic biology, the information on possible locations of TFBS is needed to create biologically neutral spacers in designed CRISPR-Cas guide RNAs for controlled transcription modulation of target genes (5). Finally, TF binding motifs are widely used to interpret regulatory sequence variants within TFBS (6), including non-coding single-nucleotide polymorphisms associated with predisposition to hereditary syndromes (7–10) and somatic mutations occurring in stem cells and cancer (11,12).

Lots of advanced approaches to model and predict TFBS using high-throughput data were presented in the past decade (13–18) yet the classic position weight matrices (PWMs) (19), also called the position-specific scoring matrices, remain the off-the-shelf solution that is widely applied in practice. More than ten years ago we introduced the HOCOMOCO collection of transcription factor binding models. Since then, HOCOMOCO became one of the key resources in the field along with CIS-BP (20) and JASPAR (21), and powered multiple studies in human and mouse gene regulation and epigenetics (22–25). However, the last release of HOCOMOCO was dated 2018, and a recent accumulation of high-throughput data and improvements in transcription factor annotation demanded substantial upgrading of the collection. It became both necessary and possible to cover more transcription factors and alternative binding motif subtypes. The wealth of data also made it possible to improve the motif quality through human expert curation based upon the significantly expanded compendium of studied TFs and the volume of available experimental information for each of the TFs. Here we present HOCOMOCO v12, which was rebuilt from scratch utilizing *in vivo* and *in vitro* high-throughput data on TF binding obtained with ChIP-Seq and HT-SELEX, respectively. By systematic reanalysis and careful curation, we assembled an updated catalog of binding motifs for 949 human TFs and 720 mouse orthologs. The motifs were computationally benchmarked in different scenarios, including recognition of sites bound *in vivo*, *in vitro*, and detection of altered TF binding at regulatory single-nucleotide variants and polymorphisms (rSNPs).

Materials and methods

HOCOMOCO transcription factors master list

The first step of the HOCOMOCO reassembly was populating the catalog of human and mouse transcription factors. We have parsed, merged, and verified the gene-ID mapping of the existing TFClass classification (26,27) available online (https://tfclass.bioinf.med.uni-goettingen.de/, https://genexplain.com/tfclass/huTF_classification_Classes.html) for orthologous human and mouse TFs. The classification was extended by adding ten methyl-CpG-binding domain proteins and supplemented with external protein IDs. The resulting HOCOMOCO master list (see Supplementary Table ST1) contains 2681 entries describing 1104 human + mouse orthologous pairs and 473 human-only TFs. Additionally, we annotated the master list with the information from (28) highlighting 1378 (human) and 921 (mouse) proteins with strong documented evidence of being the genuine DNA-binding transcription factors.

Overview of the HOCOMOCO pipeline

An overview of the pipeline used for constructing HOCOMOCO is shown in Figure 1. To construct this release, we utilized two major sources of high-throughput data on DNA-protein recognition *in vitro* and *in vivo*: peak calls from chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) and high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX). ChIP-Seq peaks were extracted from the GTRD database (29). HT-SELEX reads obtained in (30–32) were downloaded from the European Nucleotide Archive (ERP001824, ERP001826, PRJEB14744, PRJEB9797). After preprocessing (see the details below), the resulting sequences were supplied to ChIPMunk motif discovery software (33,34), followed by (I) automated

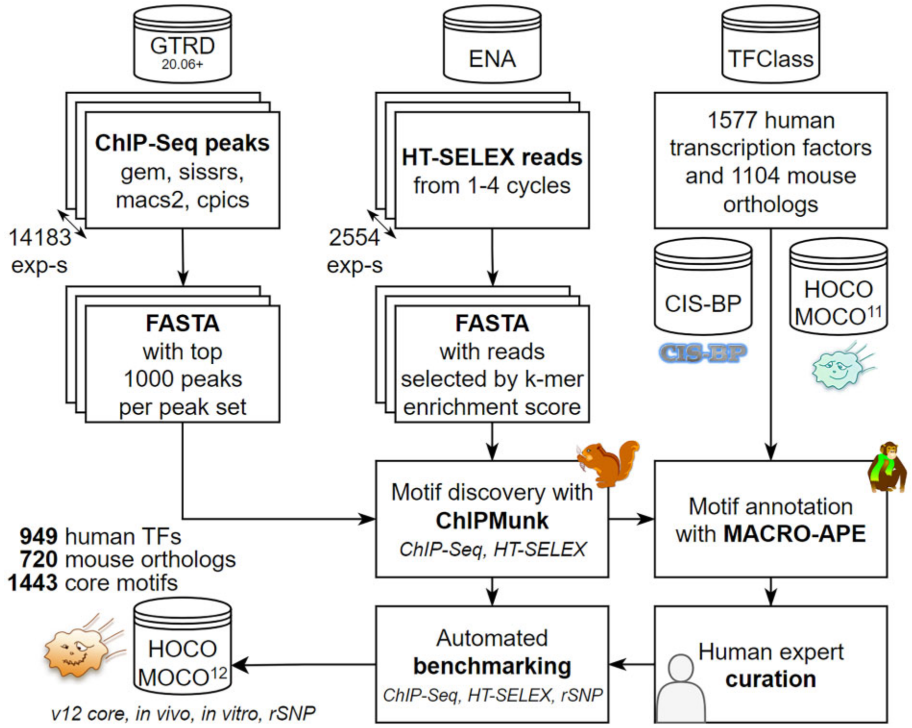


Figure 1. An overview of the HOCOMOCO v12 pipeline. ChIP-Seq peaks from GTRD and HT-SELEX reads from ENA are preprocessed and supplied to motif discovery with ChIPMunk, followed by motif annotation with MACRO-APE using known CIS-BP and HOCOMOCO v11 motifs as reference. The final collection is assembled by human expert curation and automated benchmarking.

annotation of resulting motifs by similarity with known motifs present in CIS-BP (20) and HOCOMOCO v11 (35) using MACRO-APE (36) within the TFClass family and subfamily, (II) human expert curation, and (III) automated benchmarking. The benchmarking results were used to re-visit and improve curation of particular motif subtypes, and then assemble the final motif collection.

Experimental data overview

ChIP-Seq data. ChIP-Seq and ChIP-exo peak sets were extracted from GTRD (mainly from ver. 20.06 and partly from ver. 21.12), see Supplementary Table ST2. GTRD provides peak calls from four different peak calling tools: macs2, gem, pics, sissrs (37–40). In total, the source data from the 14 183 experiments covered 1022 human and 468 mouse TFs with more than 50 thousand peak sets, more than doubling the size of the ChIP-Seq data volume used for assembling HOCOMOCO v11.

HT-SELEX data. We used the data from 1810 traditional HT-SELEX experiments: 546 experiments from Jolma *et al.*, 520 from Yang *et al.* and 744 from Yin *et al.* (30–32). In addition, we considered the results of 744 methyl-CpG HT-SELEX experiments of Yin *et al*. The comprehensive list of the datasets is provided in the Supplementary Table ST3.

Motif discovery

Motif discovery from ChIP-seq peaks

Data preparation. For each peak set, we prepared subsets of the top 1000 peaks ranked (a) by the peak height and (b) if provided by a particular peak calling tool, peak statistical significance. Peak regions were used ‘as is’ except gem, for which 301 bp regions around peak summits were taken for analysis.

Motif discovery. For each peak subset, we generated fasta files using `bedtools getfasta` and ran ChIPMunk (33) twice searching for short single-box motifs and longer motifs of arbitrary structure. From the ChIPMunk output, we excluded motifs from the consequent analysis if constructed from less than 50 words or covering <25% of the peak set. In total, >300 thousand ChIP-Seq motifs entered the curation stage.

Motif discovery from HT-SELEX reads

Data preparation. Reads from later cycles of HT-SELEX are likely to resemble the consensus better. However, in some experiments, the later cycles are over-enriched with identical consensus reads, while extra information might be available in the earlier cycles. Thus, two read sets were defined for each HT-SELEX experiment: full (all experiment cycles combined) and late cycles only (starting from the third cycle). In each case, the reads from the selected cycles were pooled and the sequences were ranked by 5-mer enrichment against the dinucleotide-shuffled background. The ‘singleton’ reads that were found only once in a pooled dataset were removed. To allow the motif occurrences to partially overhang the constant HT-SELEX adapters, each sequence was extended by 5′-NNX$_1$- and -X$_2$NN’-3′, where X$_{1,2}$ are the constant nucleotides of experiment-specific adapters flanking the HT-SELEX random inserts. For each read set, we generated up to four sequence sets using top 1000, 2500, 5000 and 10 000 reads, or all available unique reads if there were not enough available. In total, up to eight sequence sets were produced from a single HT-SELEX experiment.

Motif discovery. As HT-SELEX sequences are 10–100 times shorter than those from ChIP-Seq, it was computationally feasible to use the dinucleotide version of ChIPMunk (34) to better account for the background composition, although in the end, we have produced mononucleotide PWMs from the resulting sequence alignments. In total, nearly 20 thousand motifs from HT-SELEX entered the curation stage.

Please refer to the Supplementary Table ST4 for exact ChIPMunk command-line parameters.

Motif annotation and human expert curation

To facilitate human curation, the motifs were annotated with known motifs of the same TF, TFClass subfamily, and family, using the MACRO-APE similarity comparison tool (36) and HOCOMOCO v11 (35) along with CIS-BP (20) as the reference motif collections. CIS-BP served as a source of diverse putative motifs when checking the motif similarity at the TF family level. In turn, HOCOMOCO v11 was used as a direct primer for curation to pre-sort and group the candidate motifs based on similarity to previously annotated subtypes, if any. As in a previous large-scale benchmarking study (41) the motif performance was not affected by the motif information content, the latter was not used as a curation criterion.

Overview of the curation workflow

The motif logos of the annotated motifs of each TF were independently inspected by two junior curators who decided upon grouping motifs into known subtypes, introducing new subtypes, and discarding non-relevant patterns or common artifacts, such as low-complexity poly-A patterns or ETS motifs commonly found in ChIP-Seq peaks of other TF families. The curation decision was guided by motif annotation and TF information from HOCOMOCO v11, CIS-BP and JASPAR, taking into account known motifs of the TFClass structural family and subfamily. Particularly, a motif subtype was introduced when similar alternative motifs were discovered recurrently in multiple ChIP-Seq peak sets, or in both ChIP-Seq and HT-SELEX, or for related TFs of the same family. The curation results from junior curators were assessed and further refined by two senior curators, and the disagreements were resolved on a case-by-case basis. As a general rule, vague DNA patterns (e.g. unstructured low-complexity poly-T tracts) found in a single experiment with missing external confirmation (considering CIS-BP, JASPAR, or publications which yielded the source data) were discarded, while well-defined motifs (e.g. having high information content core regions with less conserved flanking regions) and motifs resembling the known patterns of related TFs were kept. In this setting, Zinc-finger TFs proved to be the most difficult case due to their affinity to genomic repeats (42) making it non-trivial to distinguish genuine binding sites from common artifactual signals such as various parts of ALU repeats. In the end, many of Zinc-finger TFs' motifs received the C quality due to being found only in a single experiment.

Curating distinct motif subtypes

HOCOMOCO v11 already contained distinct motif subtypes for many TFs. In this release, we took a step further and systematically considered motif subtypes merging ChIP-Seq- and HT-SELEX-derived subtypes, when possible. For many TFs, the motif subtypes were observed only with ChIP-Seq, e.g. for different TF heterodimers binding composite elements, or only with HT-SELEX, e.g. for various combinations of homodimers binding palindromic or tandem repeat sequences. Those cases were kept separately by construction. Of note, we did not try to decouple individual binding sites within composite elements found in ChIP-Seq, and the respective motifs were linked with the particular TFs for which the experimental data were produced.

The CpG methylation-related motif subtypes represented a particularly difficult curation case. On the one hand, in the genomic ChIP-Seq, such motifs usually carry a mix of CpG/TpG possibly arising from several mingled sources: TF-specific binding preference for TGs, CGs, or methylated CGs; varying methylation status of the bound regions; and CpG mutation hotspots. On the other hand, the preference for CpG over TpG and vice versa can be interpreted in HT-SELEX if results from both traditional HT-SELEX and methyl-HT-SELEX (with methylated oligonucleotides) are available. Thus, to keep the HOCOMOCO subtypes comprehensive and consistent, the curators introduced separate CpG and/or TpG motif subtypes when a preference for one or both variants was detected in HT-SELEX/methyl-HT-SELEX experiments. The motifs from HT-SELEX and ChIP-Seq were grouped into a single subtype, when possible, but the information regarding the type of the HT-SELEX experiments contributing to the subtype was kept explicitly.

Motif quality ratings and subcollections

A popular HOCOMOCO feature is the subjective motif quality or reliability on the A-B-C-D scale where A denotes the most reliable motifs. In the HOCOMOCO v12 core collection, the motif quality ratings were assigned as follows: A for motifs and subtypes found both *in vitro* and *in vivo*, B for those reproducible between individual experiments but *in vitro* or *in vivo* only, and C for all other remaining cases that passed the curation stage. Adapting to the growth of the data and HOCOMOCO usage scenarios, we have excluded from this release suspicious motifs that could have a D quality rating in HOCOMOCO v11. In the v12 core collection, D quality (unclear reproducibility) was assigned to a few motifs belonging to the motif subtypes missed by the v12 pipeline and directly inherited from v11.

In addition to the core collection, we have introduced three subcollections specialized for TFBS prediction *in vivo* ('v12-invivo'), *in vitro* ('v12-invitro'), and for rSNP annotation ('v12-rsnp'), which were based on the benchmarking results from individual data types (see below). In these subcollections, the D rating was assigned to the 'untested' motifs of TFs lacking the respective experimental data to perform the specific benchmark.

Motif benchmarking

To assess the reliability of motif models and select the best-performing model for each motif subtype, we used three sets of computational benchmarks assessing the models' performance in three scenarios: recognizing (I) TFBS bound *in vivo* using ChIP-Seq data, (II) TFBS bound *in vitro* using HT-SELEX data; (III) regulatory SNPs altering TF binding *in vivo* (allele-specific binding in ChIP-Seq (8)) and *in vitro* (SNP-SELEX (43)). Several performance metrics were computed in each category followed by log-rank-sum aggregation across obtained motif ranks for each performance metric, each benchmark, and each test dataset to obtain the final rank of each of the tested motifs for each TF in each of the three benchmarking categories. The top-ranking motifs for each TF formed the three specialized motif collections: HOCOMOCO-invivo (benchmarked on ChIP-Seq), HOCOMOCO-invitro (benchmarked on HT-SELEX), and HOCOMOCO-rSNP. On top of that, another round of log-rank-sum aggregation was performed to obtain the overall best motifs forming the unified HOCOMOCO v12 core collection.

ChIP-seq benchmarking

In HOCOMOCO v11 we used independent peak subsets for motif discovery and benchmarking. However, for transcription factors with smaller peak sets this approach reduced the sequence space for motif discovery. In this release, we did not explicitly separate the training data used for motif discovery from the test data for a particular experiment, which introduced some information leakage between model training (motif discovery) and testing (benchmarking). However, (I) for the most studied TFs there were multiple independent experiments to rely on cross-validation across experiments, (II) for many TFs there were both ChIP-Seq and HT-SELEX data available for testing, (III) for a single experiment we used multiple dependent but non-identical peak sets from alternative peak callers, (IV) the dataset of origin was rarely yielding the top-ranked motif in (41) and (V) for poorly studied TFs with a single peak set the performance ratings have limited value in any case.

To reduce the risk of introducing a technical bias in the final collection related to a particular motif performance benchmarking protocol, for HOCOMOCO v12 we used 3 different ChIP-Seq benchmarks.

1. The area under the receiver operating characteristic (auROC) using the Ambrosini *et al.* (41) implementation of the Orenstein-Shamir protocol (44) with the following modifications: up to top 1000 peaks were used as ‘positives’, both downstream and upstream regions were included in the negative set. In addition to auROC, we computed the area under the precision-recall curve (auPRC). For each peak set the benchmark was run twice: the top peaks were selected either by signal value (e.g. peak height) or by statistical significance (e.g. P-value), as reported by the peak calling tools.
2. Asymptotic pseudo-au-logROC as in HOCOMOCO v11 (35) and pseudo-auROC as in HOCOMOCO v10 (45). The benchmark identifies the best PWM hits in 301 bp genomic windows (‘positives’) centered at the peak summits and in random sequences of the same lengths (‘negatives’) following the dinucleotide composition of the positive sequence set. Up to 1000 top peaks from the test data were used, and each benchmark was run twice using the top peaks yielded by score- or significance-based sorting, as above.
3. CentriMo -log-*E*-value of motif centrality (46) measuring how motif hits are located relative to the peak summits. Up to 1000 top peaks from test data were used, and each benchmark was run twice using top peaks yielded by score- or significance-based sorting, as above.

Only peak sets yielding at least one motif that was approved and assigned to one of the motif subtypes during the curation stage were used in benchmarking. In the final motif rankings, for each motif subtype, we used only peak sets (I) comprising ≥100 peaks and (II) for which at least one of the tested motifs reached auROC ≥ 0.6. Subtypes for which no motifs reached auROC ≥ 0.6 were considered not applicable to the ChIP-Seq data and were excluded from the ‘v12-invivo’ subcollection.

Globally, ChIP-Seq benchmark had increased complexity relative to the number of datasets and motifs, thus the most studied TFs with the largest number of ChIP-Seq experiments were bottlenecking the computations. To reduce the computational cost, for the top five of such TFs (CTCF, ESR1, AR, SPI1, FOXA1) we ran the pseudo-auROC benchmark on a randomly selected subset of around 500 datasets. The results were used to pre-rank the motifs and select those ranking from 1 to 250, which were then used in the full-scale benchmarking.

HT-SELEX benchmarking

For HT-SELEX, we used the strategy described in (41). The reads from different cycles (for HT-SELEX) were pooled and a maximum of 500 000 randomly sampled unique reads per dataset were used for benchmarking: 10%, 25% or 50% of top-scoring reads were designated as ‘positives’ for each tested PWM. In addition to auROC we also computed auPRC. In the final motif rankings, for each motif subtype, we used only benchmarks where at least one of the tested motifs reached auROC ≥0.6. The subtypes for which no motifs reached auROC ≥0.6 were considered not applicable to the HT-SELEX data and were excluded from the ‘v12-invitro’ subcollection.

Benchmarking with regulatory SNPs

With rSNP benchmarking we aimed to identify the motifs suitable for assessing transcription factor binding altered by regulatory single-nucleotide variants. To this end, we employed two data sets: (I) artificial rSNP-carrying oligonucleotides assessed with SNP-SELEX for differential TF binding (43) and (II) sites of allele-specific TF binding *in vivo* from ADASTRA (8). The overview of the benchmarking data is available in Supplementary Table ST5.

Assessing rSNP prediction with SNP-SELEX data. We followed the benchmarking protocol described in (47) using the SNP-SELEX data obtained in two batches for 270 and 487 TFs, respectively. Briefly, as in (43), for each TF we distributed TF-bound SNPs between the ‘positive’ set of rSNPs affecting TF binding and ‘negative’ variants. Next, we computed auROC and auPRC of binary classification using the absolute log-ratio of PWM hit *P*-values for the reference and alternative alleles as the predicted classification score. Additionally, for the positives, we used the *P*-value log-ratio to compute Kendal τ_b and Pearson ρ against the log-*P*-value reported by SNP-SELEX that reflects the experimentally determined variant-dependent differential binding. In the final evaluation, we included only the test sets with at least ten positive class labels and only the TFs and batch combinations where at least one model reached auROC ≥ 0.6 with both $\tau_b > 0$ and $\rho > 0$. The subtypes for which no motifs fulfilled those criteria were considered not applicable to the SNP-SELEX data.

Assessing rSNP prediction with ADASTRA allele-specific binding

Compared to SNP-SELEX, the data on allele-specific binding *in vivo* lacks explicit true negatives and does not necessarily reflect direct TF binding, so the true positives, in fact, are mixed with neutral variants. Thus, for each TF, we used HOCOMOCO v11 motif (this annotation was already present in ADASTRA) as a starting filter for selecting candidate SNPs directly overlapping motif occurrences (PWM

P-value < 0.0005). Of those, we selected the ASB 'positive' set (minimal FDR < 0.05 for ChIP-Seq allelic bias towards reference or alternative allele) and the 'neutral' set (maximal FDR > 0.5). For the positive set, we computed Kendal τ_b and Pearson ρ comparing the log-ratio of PWM *P*-values against the ASB -log-FDR. Additionally, as in ADASTRA (8), we assessed the concordance of the allelic preferences between the predicted (PWM *P*-value log-ratio, log-P_{Alt} versus log-P_{Ref}) and observed (ASB $\mathrm{FDR}_{\mathrm{Alt}}$ versus ASB $\mathrm{FDR}_{\mathrm{Ref}}$) difference, and applied the Fisher's exact test to the 2 × 2 contingency table built by counting separately concordant and discordant SNPs in the positive and the neutral ASB subsets. Only TFs with at least 10 'positive' rSNPs and for which at least one model reached Fisher's *P*-value < 0.05 with both $\tau_b > 0$ and $\rho > 0$ were considered in the final evaluation. The subtypes for which no motifs fulfilled those criteria were considered not applicable to the ASB *in vivo* data.

In the end, the subtypes inapplicable to both SNP-SELEX and ASB were discarded from the 'v12-rsnp' subcollections, and the subtypes inapplicable to any of the available data type were excluded from the core collection.

Results and discussion

Here we present HOCOMOCO v12, the major update of the curated database of human and mouse transcription factor binding models. HOCOMOCO was constructed by motif discovery in peak sets from 14 183 ChIP-Seq experiments and sequenced reads from 2554 HT-SELEX experiments yielding >400 thousand candidate motifs in total. Of those, around 100 thousand motifs of 949 human TFs were approved, assigned to particular motif subtypes, and supplied to the automated benchmarking pipeline. In result, the HOCOMOCO v12 core collection contains 1443 motifs and covers 949 human TFs, 720 of which have mouse orthologs and 229 are human-exclusive. This is 40% and 60% more TFs than in v11 for the human and mouse, respectively. 241 motifs of 1443 were fully concordant between ChIP-Seq and HT-SELEX for the respective TFs and received the A quality rating, while 819 motifs were concordant between at least two experiments of the same type and received the B quality. Overall, the database update significantly improves the coverage of the motif dictionary across transcription factor families, particularly, with a major increase in binding motifs for TFs with zinc-finger DNA-binding domains (Figure 2A, B; see also the interactive version of the Figure 2B on the HOCOMOCO website). In addition to the core collection, HOCOMOCO v12 provides three extra subcollections fine-tuned for genomic TFBS prediction (v12-invivo), locating sites preferably bound *in vitro* (v12-invitro), and for identifying TFs with altered binding at rSNPs (v12-rsnp).

Due to the increased data volume and improved curation pipeline, in HOCOMOCO v12 we have eradicated most of the unreliable models of D quality which constituted more than a third of the HOCOMOCO v11 full collection. In the core collection, there are now only three D quality motifs inherited from v11 which were approved by curators but not reproduced directly at the motif discovery step of v12. Also, the D quality motifs remain in dedicated subcollections aimed at particular applications of motif libraries (v12-invivo, v12-invitro, v12-rSNP), where motifs untested at the respective target data (HT-SELEX, ChIP-Seq, or rSNPs) are included but explicitly marked by D-s. The v12 core collection includes 380 models of C quality with support from only a single experiment, through expert curation we ensured that these models are consistent either with the DNA binding patterns of the respective TF structural family or known patterns present in other motif collections. Quantitatively, motifs produced by the updated pipeline are better scoring across a diverse set of new benchmarks by design. In the v12 core collection only 11 TFs have motifs inherited from v11, which is probably related to the much larger volume of new data and more rigorous benchmarking setup of the current update. Given the possibly confusing diversity of quality ratings, subtypes, and subcollections, we have designed a flowchart to guide the user in selecting a proper motif collection and a suitable subset of motifs (Supplementary Figure SF1).

With multiple other motif collections covering human and mouse TFs, we consider HOCOMOCO as having its distinctive value for several reasons. First, all motifs were derived by the same motif discovery tool in the same pipeline, making them uniform and comparable across the board. Second, all motifs and multiple motif subtypes were manually curated to reduce redundancy and discard ambiguous patterns and shared artifactual signals. Third, in this version, we provide not only the complete core collection but specifically optimized subsets suitable for different practical needs, from genomic TFBS prediction to designing high-affinity artificial oligonucleotides and regulatory SNP annotation.

Compared to v11, in this release, we do not provide dinucleotide models. Producing and testing these models demanded extra computational expenses for motif discovery due to dramatically increased data volume, and required major changes in the benchmarking protocols, which are not directly applicable to models other than PWMs. Thus, at the current stage, dinucleotide models will remain available in HOCOMOCO v11, while HOCOMOCO v12 can provide a better baseline to derive and test not only dinucleotide PWMs but also more complex models.

The starting ChIP-Seq and HT-SELEX data covered 1022 and 609 human TFs, respectively, but in the end, not all of them are listed in the v12 collection. Unfortunately, our setup does not allow us to state whether these failures have arisen due to the TF itself having limited DNA-binding specificity or technical issues with the original data, data preprocessing, or motif discovery. Data from other experimental technologies may fill this gap in the future.

In this release, we are providing a joint motif collection covering human TFs and their mouse orthologs with a shared set of motifs, motivated by the fact that the TF binding specificities are conserved between human and mouse (30) and for some TFs extend largely even to the level of the fruit fly (48). In HOCOMOCO v11 we cross-validated the motifs between human and mouse and realized that high-performing motifs are performing well across datasets, regardless of the species on which the motif discovery was conducted. Further, in the SNP-SELEX benchmark (47), we observed many cases, where a motif obtained for an ortholog TF of other species outperformed the motif obtained directly from human data. Regarding particular subtypes, there were examples of human- and mouse-exclusive motif subtypes spotted at the curation stage. Yet, in general, the respective TFs were profiled in different cell types and with different antibodies, making it hard to attribute the differences specifically to species and not to technical features of experiments. Thus, in this release, we kept all alternative motifs as species-agnostic motif subtypes, but

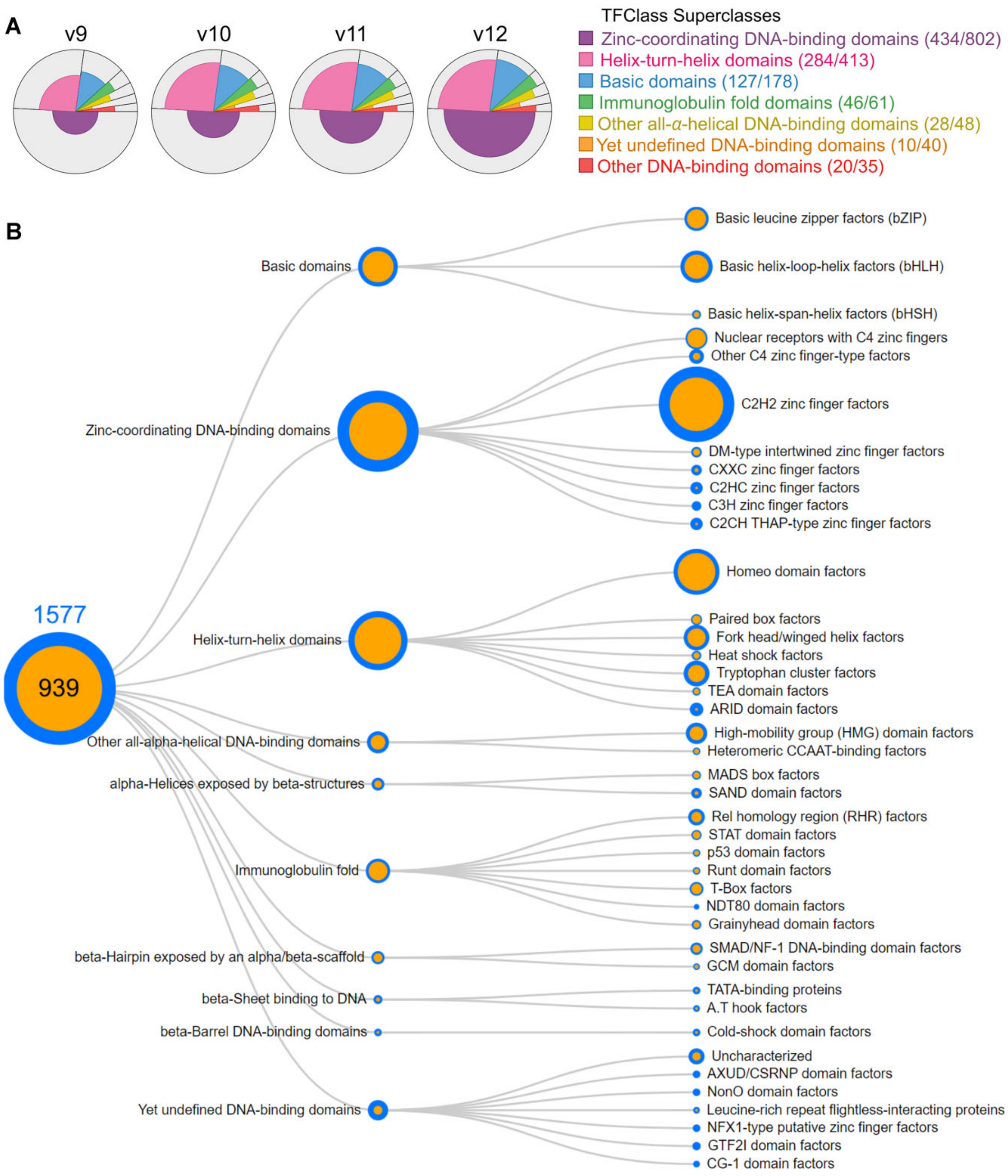


Figure 2. The coverage of different TF classes by HOCOMOCO v12 motifs. **(A)** Improvements in the coverage of transcription factors by HOCOMOCO motifs across the largest DNA-binding domain superclasses, HOCOMOCO from v9 to v12. The pie chart slices denote the contribution of superclasses to the total set of TFs, and the colored parts of the slices denote the fraction of TFs with motifs. The total number of TFs and the number of TFs with HOCOMOCO v12 motifs in each superclass are given in the legend in brackets. **(B)** Relative coverage of different TF classes by HOCOMOCO v12 motifs. Blue: total number of TFs in a superclass or a class, orange: TFs with motifs.

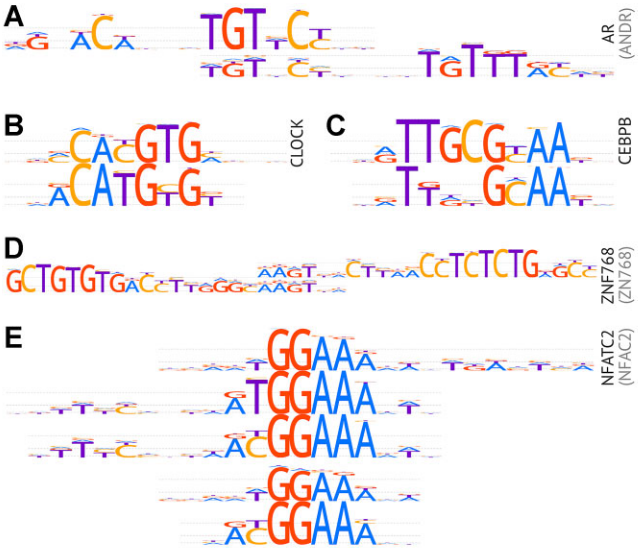


Figure 3. Illustrative examples of motif subtypes included in HOCOMOCO v12. The plots show motif logos, the TF gene symbols are labeled on the right with UniProt entry name prefixes in brackets, if different from the gene symbol. **(A)** The AR palindromic motif (top), and its shorter version making a composite element with FOXA1 (bottom). **(B)** CLOCK motifs derived from ChIP-Seq (top) and methyl-HT-SELEX (bottom). The methyl-HT-SELEX motif prefers TG instead of CG in the central position, which is followed by [C/T]G instead of pure TG in ChIP-Seq. **(C)** CEBPB motifs derived from HT-SELEX (top) and ChIP-Seq (bottom). **(D)** ZNF768 motif subtypes representing overlapping regions of the same repetitive element. **(E)** NFATC2 multiple motif subtypes.

explicitly annotated subtypes coming exclusively from human or mouse data.

On redundancy of motif subtypes

The selection of top-performing motifs from multiple alternatives has been guided by formal criteria of benchmarking measures, which is a major feature of the HOCOMOCO collection. However, the presence of distinct motif subtypes is relying solely on human expert curation. The original idea of including subtypes in HOCOMOCO was to attribute individual motifs to different modes of binding, e.g. distinguish AR palindromic binding sites from AR-FOXA1 (Figure 3A) composite elements. We did not introduce motif subtypes automatically via TF-level motif clustering (49) as many intermediate motifs are blurring the identity of clusters and thus complicating the selection of universal similarity thresholds across the whole range of TFs. Further, this release was designed to be as inclusive as possible and we explicitly included subtypes even with minor motif alterations, if biologically substantiated, e.g. if they arise from different experimental layouts (methyl- vs normal HT-SELEX), different modes of binding (e.g. tandem binding sites in HT-SELEX), or protein-protein complexes (composite elements in ChIP-Seq). Notably, as expected, ChIP-Seq-derived subtypes describing composite elements perform strongly in ChIP-Seq-based benchmarks. Yet, their relative ranks in the core collection vary across TFs, in particular, depending on the composition of the benchmarking data. For example, OCT4-SOX2 composite element is the major first-ranked motif subtype for OCT4 but a secondary motif subtype for SOX2.

A typical example of a relevant but small difference between the subtypes is the CG/TG substitution mostly related to the methylated CpGs. The changeable position arises from two linked but distinct scenarios: (I) TF preferring or avoiding mCpG within the binding sites revealed by methyl-HT-SELEX, and (II) [C > T]G mutation hotspots, leading to depletion of CG pairs in genomic sites and thus affecting ChIP-Seq-derived motifs. The first case is illustrated e.g. by the CLOCK motif subtypes found in HT-SELEX data (Figure 3B). The second case is clearly exhibited for C/EBP motifs (Figure 3C), as C/EBP TFs bind the frequent T-G mismatches at CpGs within its binding sites with increased affinity and by doing so impair base excision repair (11).

A complicated case comes with the repeat-binding zinc-finger proteins, for which ChIP-Seq motif discovery often fails to capture the exact location of the binding site within a longer repetitive element, thus the alternative subtypes do not explain any different modes of binding but rather belong to different parts of the longer consensus, see ZNF768 motifs in Figure 3D.

The diversity of the annotated subtypes in some cases might be excessive, e.g. there are 5 motif subtypes for NFATC2 (Figure 3E), including methylation-specific subtypes and a composite element found in ChIP-Seq. Yet, we nonetheless included all recurrently found motif variants as they could have different practical applications. For example, longer motifs from ChIP-Seq likely capture extended sequence context or cofactor binding patterns and thus are specifically useful for genomic TFBS recognition. Shorter single box motifs might be inefficient in predicting the complete binding sites but useful to decouple TFBS composite elements. Finally, tandem or palindromic motifs from HT-SELEX often represent optimally spaced binding sites with high-affinity and might be suitable to optimize oligonucleotides for binding affinity *in vitro*.

Of note, the motif subtypes in HOCOMOCO v12 are ordered (0, 1, …) according to their performance in benchmarking, thus in the case a user requires a single motif per TF, the first subtype (0) motifs can be safely selected as TF representatives.

Concluding remarks

Summing up, this release brings HOCOMOCO close to a thousand TFs with reliably described binding specificities. In v12 we have eradicated the legacy non-benchmarked models and models built from low-throughput data, significantly expanded the benchmarking setup, and performed rigorous annotation of motif subtypes. As in the previous release, HOCOMOCO update is accompanied by renewed MoLoTool (motif location toolbox), an interactive JavaScript web application for visualizing motif hits in user-supplied sequences. The online version of our tool for rSNP analysis, PERFECTOS-APE, was also updated to use v12-rsnp collection by default.

We believe HOCOMOCO v12 will serve as a solid knowledge base empowering molecular biology and genetics of gene regulation and also establish the ground for reaching a complete and reliable collection of human and mouse TF motifs in the future.

Data availability

The HOCOMOCO v12 database is freely available at https://hocomoco12.autosome.org and https://hocomoco.autosome.org.

The HOCOMOCO v12 motif sets and benchmarking results are available at ZENODO [doi:10.5281/zenodo.10012937].

The implementation of ChIP-Seq and HT-SELEX benchmarking protocols is available at GitHub: https://github.com/autosome-ru/motif_benchmarks.

The implementation of the HT-SELEX k-mer enrichment estimation is available at GitHub: https://github.com/autosome-ru/kmer-motif-enrichment.

The implementation of rSNP-based benchmarking protocols is available at GitHub: https://github.com/autosome-ru/hocomoco_rsnp_benchmarks.

The online-only supplementary data are available at the NAR website.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We personally thank Philipp Bucher for suggesting the strategy for filtering HT-SELEX reads using k-mer enrichment and Oriol Fornes for his technical assistance with setting up the CentriMo benchmark. We thank Maria Korlyakowa, Renata Betenkova, Daria Nikolaeva, and Arina Kukhani for their assistance at the initial motif curation stage. We thank Niagara Computers LLC and Valery Egorshev personally for valuable technical support in setting up computational resources.

Funding

Russian Science Foundation [20-74-10075 to I.V.K.]; A.Z. (HT-SELEX cross-validation) was supported by a personal fellowship from the Non-commercial Foundation for Support of Science and Education 'INTELLECT'; Y.M. (mCpG motif annotation) was supported by the Ministry of Science and Higher Education of the Russian Federation; V.K. and V.J.M. (motif ranking engine) were supported by the Megagrant from the Government of the Russian Federation [075-15-2021-595]; Analysis of selected TF families was supported by the Ministry of Science and Higher Education of the Russian Federation [075-152-021-601]. Funding for open access charge: Personal funds.

Conflict of interest statement

None declared.

References

1. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
2. Garcia-Alonso,L., Holland,C.H., Ibrahim,M.M., Turei,D. and Saez-Rodriguez,J. (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, **29**, 1363–1375.
3. Georgakopoulos-Soares,I., Deng,C., Agarwal,V., Chan,C.S.Y., Zhao,J., Inoue,F. and Ahituv,N. (2023) Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nat. Commun.*, **14**, 2333.
4. Yi,X., Zheng,Z., Xu,H., Zhou,Y., Huang,D., Wang,J., Feng,X., Zhao,K., Fan,X., Zhang,S., *et al.* (2021) Interrogating cell type-specific cooperation of transcriptional regulators in 3D chromatin. *Iscience*, **24**, 103468.
5. Crone,M.A., MacDonald,J.T., Freemont,P.S. and Siciliano,V. (2022) gDesigner: computational design of synthetic gRNAs for Cas12a-based transcriptional repression in mammalian cells. *NPJ Syst Biol Appl*, **8**, 34.
6. Vorontsov,I.E., Kulakovskiy,I.V., Khimulya,G., Nikolaeva,D.D. and Makeev,V.J. (2015) PERFECTOS-APE - predicting regulatory functional effect of SNPs by approximate P-value estimation. In: *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*. SCITEPRESS - Science and and Technology Publications, pp. 102–108.
7. Vuckovic,D., Bao,E.L., Akbari,P., Lareau,C.A., Mousas,A., Jiang,T., Chen,M.-H., Raffield,L.M., Tardaguila,M., Huffman,J.E., *et al.* (2020) The polygenic and monogenic basis of blood traits and diseases. *Cell*, **182**, 1214–1231.
8. Abramov,S., Boytsov,A., Bykova,D., Penzar,D.D., Yevshin,I., Kolmykov,S.K., Fridman,M.V., Favorov,A.V., Vorontsov,I.E., Baulin,E., *et al.* (2021) Landscape of allele-specific transcription factor binding in the human genome. *Nat. Commun.*, **12**, 2751.
9. Boytsov,A., Abramov,S., Aiusheeva,A.Z., Kasianova,A.M., Baulin,E., Kuznetsov,I.A., Aulchenko,Y.S., Kolmykov,S., Yevshin,I., Kolpakov,F., *et al.* (2022) ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs. *Nucleic Acids Res.*, **50**, W51–W56.
10. Uvarova,A.N., Stasevich,E.M., Ustiugova,A.S., Mitkin,N.A., Zheremyan,E.A., Sheetikov,S.A., Zornikova,K.V., Bogolyubova,A.V., Rubtsov,M.A., Kulakovskiy,I.V., *et al.* (2023) rs71327024 Associated with COVID-19 hospitalization reduces CXCR6 promoter activity in Human CD4+ T cells via disruption of c-myb binding. *Int. J. Mol. Sci.*, **24**, 13790.
11. Ershova,A.S., Eliseeva,I.A., Nikonov,O.S., Fedorova,A.D., Vorontsov,I.E., Papatsenko,D. and Kulakovskiy,I.V. (2021) Enhanced C/EBP binding to G·T mismatches facilitates fixation of CpG mutations in cancer and adult stem cells. *Cell Rep.*, **35**, 109221.
12. Vorontsov,I.E., Khimulya,G., Lukianova,E.N., Nikolaeva,D.D., Eliseeva,I.A., Kulakovskiy,I.V. and Makeev,V.J. (2016) Negative selection maintains transcription factor binding motifs in human cancer. *Bmc Genomics [Electronic Resource]*, **17**, 395.
13. Tognon,M., Giugno,R. and Pinello,L. (2023) A survey on algorithms to characterize transcription factor binding sites. *Brief. Bioinform*, **24**, :bbad156.
14. Isakova,A., Groux,R., Imbeault,M., Rainer,P., Alpern,D., Dainese,R., Ambrosini,G., Trono,D., Bucher,P. and Deplancke,B. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods*, **14**, 316–322.
15. Grau,J., Posch,S., Grosse,I. and Keilwagen,J. (2013) A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Res.*, **41**, e197.
16. Novakovsky,G., Fornes,O., Saraswat,M., Mostafavi,S. and Wasserman,W.W. (2023) ExplaiNN: interpretable and transparent neural networks for genomics. *Genome Biol.*, **24**, 154.
17. Khamis,A.M., Motwalli,O., Oliva,R., Jankovic,B.R., Medvedeva,Y.A., Ashoor,H., Essack,M., Gao,X. and Bajic,V.B. (2018) A novel method for improved accuracy of transcription factor binding site prediction. *Nucleic Acids Res.*, **46**, e72.
18. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
19. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
20. 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.
21. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A., Manosalva Pérez,N., *et al.* (2022) JASPAR

2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **50**, D165–D173.

22. Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., De Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M., Itoh,M., *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
23. Medvedeva,Y.A., Khamis,A.M., Kulakovskiy,I.V., Ba-Alawi,W., Bhuyan,M.S.I., Kawaji,H., Lassmann,T., Harbers,M., Forrest,A.R. and Bajic,V.B. (2014) Effects of cytosine methylation on transcription factor binding sites. *Bmc Genomics [Electronic Resource]*, **15**, 119.
24. Alam,T., Medvedeva,Y.A., Jia,H., Brown,J.B., Lipovich,L. and Bajic,V.B. (2014) Promoter analysis reveals globally differential regulation of Human long non-coding RNA and protein-coding genes. *PLoS One*, **9**, e109443.
25. Lioznova,A.V., Khamis,A.M., Artemov,A.V., Besedina,E., Ramensky,V., Bajic,V.B., Kulakovskiy,I.V. and Medvedeva,Y.A. (2019) CpG traffic lights are markers of regulatory regions in human genome. *Bmc Genomics [Electronic Resource]*, **20**, 102.
26. Wingender,E., Schoeps,T. and Dönitz,J. (2013) TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.*, **41**, D165–D170.
27. Wingender,E., Schoeps,T., Haubrock,M. and Dönitz,J. (2015) TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.*, **43**, D97–D102.
28. Lovering,R.C., Gaudet,P., Acencio,M.L., Ignatchenko,A., Jolma,A., Fornes,O., Kuiper,M., Kulakovskiy,I.V., Lægreid,A., Martin,M.J., *et al.* (2021) A GO catalogue of human DNA-binding transcription factors. *Biochim. Biophys. Acta (BBA) - Gene Regul. Mech.*, **1864**, 194765.
29. Kolmykov,S., Yevshin,I., Kulyashov,M., Sharipov,R., Kondrakhin,Y., Makeev,V.J., Kulakovskiy,I.V., Kel,A. and Kolpakov,F. (2021) GTRD: an integrated view of transcription regulation. *Nucleic Acids Res.*, **49**, D104–D111.
30. 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.
31. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.
32. 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.
33. Kulakovskiy,I.V., Boeva,V.A., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
34. Kulakovskiy,I.V., Levitsky,V.G., Oschepkov,D.G., Vorontsov,I.E. and Makeev,V.J. (2013) Learning advanced TFBS models from Chip-seq data - diChIPMunk: effective construction of dinucleotide positional weight matrices. In: *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*. SciTePress - Science and and Technology Publications, pp. 146–150.
35. 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.
36. Vorontsov,I.E., Kulakovskiy,I.V. and Makeev,V.J. (2013) Jaccard index based similarity measure to compare transcription factor binding site models. *Algorith. Mol. Biol.*, **8**, 23.
37. 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.
38. Narlikar,L. and Jothi,R. (2012) ChIP-seq data analysis: identification of protein–DNA binding sites with SISSRs peak-finder. Wang,J., Tan,A. and Tian,T. (eds). *Next Generation Microarray Bioinformatics. Methods in Molecular Biology*. Vol. **802**, Humana Press, pp. 305–322.
39. Guo,Y., Mahony,S. and Gifford,D.K. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.
40. Zhang,X., Robertson,G., Krzywinski,M., Ning,K., Droit,A., Jones,S. and Gottardo,R. (2011) PICS: probabilistic inference for ChIP-seq. *Biometrics*, **67**, 151–163.
41. Ambrosini,G., Vorontsov,I., Penzar,D., Groux,R., Fornes,O., Nikolaeva,D.D., Ballester,B., Grau,J., Grosse,I., Makeev,V., *et al.* (2020) Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol.*, **21**, 114.
42. Schmitges,F.W., Radovani,E., Najafabadi,H.S., Barazandeh,M., Campitelli,L.F., Yin,Y., Jolma,A., Zhong,G., Guo,H., Kanagalingam,T., *et al.* (2016) Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.*, **26**, 1742–1752.
43. Yan,J., Qiu,Y., Ribeiro dos Santos,A.M., Yin,Y., Li,Y.E., Vinckier,N., Nariai,N., Benaglio,P., Raman,A., Li,X., *et al.* (2021) Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, **591**, 147–151.
44. Orenstein,Y. and Shamir,R. (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.*, **42**, e63.
45. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Soboleva,A.V., Kasianov,A.S., Ashoor,H., Ba-Alawi,W., Bajic,V.B., Medvedeva,Y.A., Kolpakov,F.A., *et al.* (2016) HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, **44**, D116–D25.
46. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
47. Boytsov,A., Abramov,S., Makeev,V.J. and Kulakovskiy,I.V. (2022) Positional weight matrices have sufficient prediction power for analysis of noncoding variants. *F1000Res*, **11**, 33.
48. Nitta,K.R., Jolma,A., Yin,Y., Morgunova,E., Kivioja,T., Akhtar,J., Hens,K., Toivonen,J., Deplancke,B., Furlong,E.E.M., *et al.* (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife*, **4**, e04837.
49. 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.

Received: September 22, 2023. **Revised:** October 17, 2023. **Editorial Decision:** October 25, 2023. **Accepted:** October 26, 2023