Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1106

HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis

Ivan V. Kulakovskiy[1,2,3,*], Ilya E. Vorontsov[2], Ivan S. Yevshin[4], Ruslan N. Sharipov[4,5,6], Alla D. Fedorova[7], Eugene I. Rumynskiy[2,8], Yulia A. Medvedeva[2,8,9], Arturo Magana-Mora[10,11], Vladimir B. Bajic[11], Dmitry A. Papatsenko[3], Fedor A. Kolpakov[5,4] and Vsevolod J. Makeev[1,2,8,*]

[1]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 119991, GSP-1, Vavilova 32, Moscow, Russia, [2]Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991, GSP-1, Gubkina 3, Moscow, Russia, [3]Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo Institute of Science and Technology, 143026 Moscow, Russia, [4]BIOSOFT.RU Ltd, 630058, Russkaya 41/1, Novosibirsk, Russia, [5]Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences, 630090, Akad. Rzhanova 6, Novosibirsk, Russia, [6]Novosibirsk State University, 630090, Pirogova 2, Novosibirsk, Russia, [7]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119234, Leninskiye Gory 1–73, Moscow, Russia, [8]Moscow Institute of Physics and Technology (State University), 141700, 9 Institutskiy per, Dolgoprudny, Russia, [9]Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, 119071, 2 Leninsky Ave. 33, Moscow, Russia, [10]National Institute of Advanced Industrial Science and Technology (AIST), Com. Bio Big-Data Open Innovation Lab. (CBBD-OIL), AIST Tokyo Waterfront Main Bldg. #323, 2-3-26 Aomi, Tokyo 135-0064, Japan and [11]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Thuwal 23955-6900, Saudi Arabia

Received September 14, 2017; Revised October 15, 2017; Editorial Decision October 22, 2017; Accepted October 31, 2017

ABSTRACT

We present a major update of the HOCOMOCO collection that consists of patterns describing DNA binding specificities for human and mouse transcription factors. In this release, we profited from a nearly doubled volume of published in vivo experiments on transcription factor (TF) binding to expand the repertoire of binding models, replace low-quality models previously based on in vitro data only and cover more than a hundred TFs with previously unknown binding specificities. This was achieved by systematic motif discovery from more than five thousand ChIP-Seq experiments uniformly processed within the BioUML framework with several ChIP-Seq peak calling tools and aggregated in the GTRD database. HOCOMOCO v11 contains binding models for 453 mouse and 680 human transcription factors and includes 1302 mononucleotide and 576 dinucleotide position weight matrices, which describe primary binding preferences of each transcription factor and reliable alternative binding specificities. An interactive interface and bulk downloads are available on the web: http://hocomoco.autosome.ru and http://www.cbrc.kaust.edu.sa/hocomoco11. In this release, we complement HOCOMOCO by MoLoTool (Motif Location Toolbox, http://molotool.autosome.ru) that applies HOCOMOCO models for visualization of binding sites in short DNA sequences.

INTRODUCTION

Models of transcription factor binding sites (TFBS) are essential tools for computational studies of transcriptional regulation from dissection of particular cis-regulatory regions and genome-wide TFBS predictions to modeling regulatory networks and functional annotation of sequence variants (1–6). Advanced TFBS models evolve rapidly (7–10), but the basic position weight matrix model remains a useful baseline for a wide range of applications (4,11–14). The wide availability of experimental data on protein-DNA interaction *in vivo* allows for systematic construction and comparison of different TFBS models.

*To whom correspondence should be addressed. Tel: +7 499 135 6000; Fax: +7 499 135 1405; Email: ivan.kulakovskiy@gmail.com
Correspondence may also be addressed to Vsevolod J. Makeev. Email: vsevolod.makeev@gmail.com

In the past few years, HOCOMOCO database of transcription factor binding models became one of the major resources for sequence analysis of transcriptional regulation in mammals. In particular, HOCOMOCO has been a useful data source in a recent DREAM-ENCODE challenge on the prediction of transcription factor binding sites (https://www.synapse.org/ENCODE) where several top-performing teams used HOCOMOCO models in their solutions.

Here we present a major update of the HOCOMOCO collection of human and mouse transcription factor binding models based on systematic motif discovery and cross-validation using more than 14 thousand ChIP-Seq data sets obtained from >5000 experiments for human and mouse transcription factors. Such large-scale analysis allowed for significant expansion and improvement of the non-redundant set of TFBS models for human and mouse transcription factors.

The diverse repertoire of experimental data sets systematically brings about alternative binding models for a particular TF. Following the original ideas used in developing of HOCOMOCO, we focus on primary binding patterns that robustly represent binding sites across multiple experiments (HOCOMOCO-11-CORE). At the same time, the alternative models are now systematically provided in the extended collection (HOCOMOCO-11-FULL), that now also contains lower-reliability binding models built from limited experimental data.

The total number of available ChIP-Seq data sets more than doubled since the previous release (HOCOMOCO v10). Each ChIP-Seq data set was processed with four different peak callers (*macs*, *gem*, *pics*, *sissrs*) (15–18), thus increasing the diversity of peak sets used for motif discovery: the total number of peak sets became more than ten fold greater compared to the previous HOCOMOCO version. Furthermore, we systematically assessed the suitability of peak callers for downstream motif discovery by comparing TFBS recognition quality of the TBFS models derived from the respective peaks. Finally, profiting from an outstandingly large number of ChIP-Seq datasets for human and mouse species we performed a cross-species validation of human/mouse binding sites models for orthologous transcription factors.

We continue to maintain the set of command-line tools to facilitate practical utilization of the HOCOMOCO models: SPRY-SARUS for motif finding (search for motif occurrences), MACRO-APE for weight matrix comparison and *P*-value estimation, and PERFECTOS-APE for annotation of regulatory variants. The accompanying tools are web-accessible at http://opera.autosome.ru including ChIP-Munk that has been used for motif discovery. Finally, in this version of HOCOMOCO, we introduce MoLoTool (Motif Location Toolbox), an interactive web-tool to identify motif occurrences in a set of sequences based on HOCOMOCO models.

MATERIALS AND METHODS

The workflow to assemble HOCOMOCO v11 consisted of the following steps: aggregation and filtering of the ChIP-Seq data sets, motif discovery, curation, and benchmarking. An overview scheme of the workflow is given in the Supplementary Figure S1; each step is described in details below.

ChIP-Seq data overview

We used GTRD (19) (http://gtrd.biouml.org, release 17 April 2017) as a source of systematically processed ChIP-Seq data. GTRD aggregated ChIP-Seq data from GEO and reprocessed it with a unified pipeline using four different peak calling tools (*macs*, *gem*, *pics*, *sissrs*) (15–18).

The peaks were called from the experimental 'data sets' (typically a pair of ChIP-Seq experiment and control samples). In some experiments, the control samples were missing, whereas replicated experiments were considered as separate data sets. Four peak sets were obtained from every data set, a peak set for each peak caller. As a starting ground, GTRD provided 3311/2623 data sets and 12612/9938 peak sets for 602/354 transcription factors for human/mouse respectively.

The numbers of peaks identified by different peak callers for the same data set were different, which is expected since GTRD used the peak callers with the default parameters and did not normalize for the number of the resulting peaks. However, for the most of the data sets the peak numbers in the peak sets were generally consistent in a sense that all four peak callers agreed in producing a relatively large or a small peak set. However, for some data sets, there were particular peak callers producing the unexpectedly large number of peaks in discordance with results of other peak callers. We considered such peak sets unreliable and decided not to use them for motif discovery and the subsequent benchmarking.

To quantitatively measure the consistency of the numbers of the peaks called by different peak callers we considered the aggregated collection of all data sets for all transcription factors. For each peak caller, an empirical distribution of the peak numbers was constructed from all the peak sets obtained by this peak caller (ignoring the peak sets containing zero peaks). This allowed us to substitute the number of peaks N in each peak set with the empirical weight of the lighter tail defined as $S = min(P\ (\geq N),\ P\ (\leq N))$, where P is the empirical probability for a peak set to contain the conditioned number of peaks. Lower values of S correspond to unlikely (extremely large or small) peak numbers. A concordant data set is expected to have similar values of S for different peak callers. Thus, from 4 peak sets of each data sets, we iteratively excluded the one with the lowest S, until the ratio between the largest and the smallest S in the data set became not >2. The entire data set was removed if only one peak set remained.

This rough filter removed nearly a third of the peak sets. Next, we removed small peak sets of less than 200 peaks since we assumed TFBS models derived from such small data sets as nonrobust. The resulting collection of 8117 / 6189 peak sets for 2885 / 2212 human/mouse ChIP-Seq data sets were used for motif discovery and benchmarking. Supplementary Figure S2 shows the number of experimental data sets and peak sets across transcription factors.

Motif discovery

The general setup of the motif discovery and analysis was inherited from the HOCOMOCO v10 pipeline (20). We utilized the top 1000 peaks from each data set: even-ranked peaks were used for motif discovery (training) and odd-ranked peaks as control data for benchmarking. To rank the peaks based on the ChIP-Seq signal strength we used the following peak caller-specific data: *macs* – 'number of tags in the peak region', *gem* – 'immunoprecipitation binding strength' (the number of immunoprecipitation reads associated with the event), *sissrs* – 'NumTags' (the number of tags supporting the strongest binding site in the reported binding region), *pics* – 'enrichment score' (normalized to the control data). Peaks of *sissrs* and *pics* were taken 'as is', 300 bp regions around peak summits were taken for *macs* and *gem* data. To perform motif discovery and construct classic and dinucleotide (21) position weight matrix models we used ChIPMunk (22) in a new 'summit' mode where the base coverage profile of a peak was approximated with a triangle with the vertex at the peak summit location and the base stretched to that of the peak. This makes the resulting motif occurrences to be correctly 'phased' against the peak summit (to reduce the chance of irrelevant patterns being identified). ChIPMunk was set to estimate the background sequence composition automatically (the nucleotide and dinucleotide frequencies for PWMs and dinucleotide PWMs respectively) and executed twice for each peak set: in the default mode and with the single-box motif shape prior (23). For all peak sets for a particular TF ChIPMunk was set to scan the model length (motif width) range from X down to 7 bp, where X was derived from HOCOMOCO v10 models by incrementing by one the maximum length of known binding models for proteins with similar DNA-binding domain architecture. To this end for each particular TF we traversed the TFClass hierarchy (24,25) from TF 'Subfamily' up to the 'Class' level stopping when binding models from the previous HOCOMOCO version were found. The maximal model length of 23 was used if none of the known models was found.

Model curation

The motif discovery results were post-processed to ensure that only reliable models entered the benchmarking. First, we used MACRO-APE to compare the newly constructed models with the previously known (HOCOMOCO v10) binding models for the same TF and for the members of the same TF subfamily and family (when such models were available). Next, the results were manually assessed to remove weak patterns (having low information content), patterns inconsistent between different peak sets, patterns belonging to different TFs (e.g. binding patterns of cofactor proteins or peculiarities of particular experimental data), and to annotate TFs for which binding models were not available or had limited reliability in HOCOMOCO v10. All TFBS models for a particular TF were classified into several categories: the primary models (similar with those previously known for the same TF or consistent within the structural family), the single-box alternative models (resembling the major box of a primary double-box model or the core pattern of the primary model if it was surrounded with long flanking regions), and the alternative models (with a clear alternative pattern reliably discovered from several peak sets).

Benchmarking models and assembling the final collection

To select the best models for each TF, we used the benchmarking workflow inherited from HOCOMOCO v10 with the following modifications: the ROC (receiver operating characteristic) curves were plotted with the false positive rate axis in the log10 scale. Thus, the AUC logROC was used instead of the AUC ROC (area under curve ROC) in a way similar to that suggested in (26) to prioritize for the ROC region with low false positive rates (see Supplementary Figure S3 for selected examples of logROC plots). This modification has the same purpose as the partial AUC used in (8).

A few additional models, e.g. ANDR (androgen receptor) binding model submitted for consideration from (27), were also considered, yet none of them succeeded in the benchmark except for the models that we specifically designed to describe binding sites subtypes, e.g. genuine binding sites of NANOG transcription factor (see Discussion).

Some TFs were notably better represented by the experimental data from either human or mouse orthologs. To benefit from all available data, we tested TF binding models from both species (human and mouse) on both human and mouse data. This allowed us to evaluate the quality of 'human' binding models taking into account mouse data and vice versa.

We used the following strategy to include cross-species AUC logROC ($lAUC$) values in the final quality estimation. For each species (human or mouse) and for each model we computed weighted AUC logROC value:

$$wlAUC_{species} = \frac{\sum_{species\ peak\ sets} lAUC(peak\ set) \cdot W(peak\ set)}{\sum_{species\ peak\ sets} W(peak\ set)}$$

where

$$W(peak\ set) = \sum_{all\ models} lAUC(peak\ set) \,/\, N_{models}$$

and

$$lAUC(peak\ set) = AUC(ROC(TPR, log_{10} FPR)).$$

FPR was estimated from the model P-values as in HOCOMOCO v10 (20).

The weight W of a peak set provides an estimate of the peak set quality aggregated from all benchmarked models.

If only the human or the mouse data were available the resulting model quality estimate was

$$Q_{human} = wlAUC_{human},\ Q_{mouse} = wlAUC_{mouse}.$$

In the case when data from both species were available, the quality estimate was corrected by including the weighted average from the data of the other species considering it as a single 'virtual' peak set, e.g. for human models:

$$Q_{human} = \frac{wlAUC_{human} \cdot N_{human\ peak\ sets} + wlAUC_{mouse}}{N_{human\ peak\ sets} + 1}.$$

This allowed us to take into account all available data for all TFs, which was especially important for the cases when

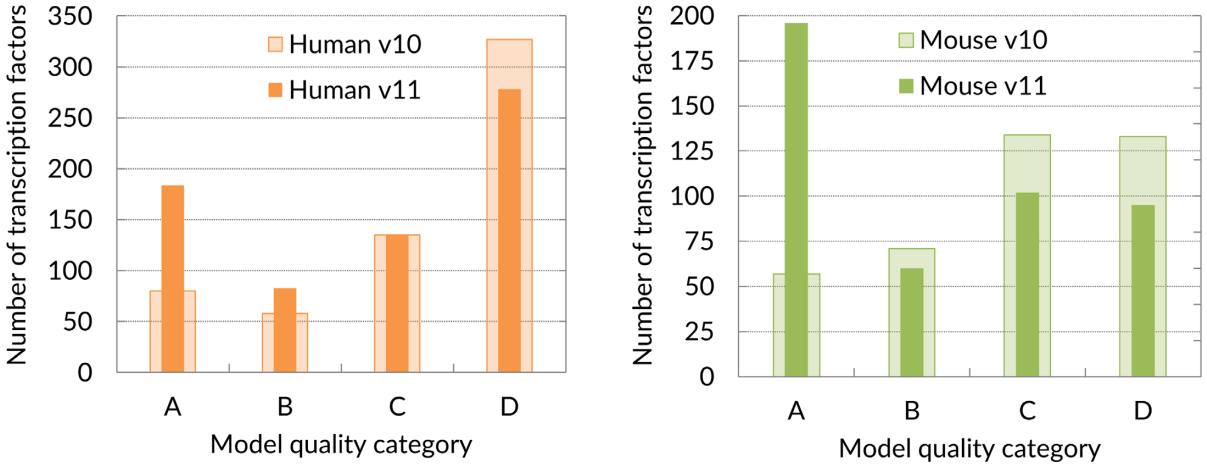

Figure 1. Number of TFs (Y-axis) with the best available model of a given quality (X-axis). The number of the most reliable A-quality models is more than doubled in HOCOMOCO v11 comparing to v10.

only a limited number of experimental data sets was available.

The resulting values of Q were used to select the best primary model for each TF and to rank alternative models.

Sometimes the model based on the mouse data was assigned to a human TFs as performing systematically better at human data in the benchmark. The same was true for some human models outperforming all mouse models at the corresponding mouse data (see Results).

In the previous releases, the HOCOMOCO models were provided with simplified A-to-D quality ratings used to distinguish models that properly characterize the TF binding specificity (A to C quality) from models of limited predictive power and those based on limited experimental evidence (D quality). We adopt the same scheme in HOCOMOCO v11: the models of A, B, and C quality are recommended to make reliable *de novo* predictions, whereas models of D quality are still useful for various exploratory purposes, e.g., family-wide TF analysis.

The final model quality assignment procedure was inspired by HOCOMOCO v10 but modified to benefit from the data of multiple peak callers (Supplementary Figure S4). For the quality assignment, as in HOCOMOCO v10, we used linear, rather than logarithmic, AUC ROC scores and AUC thresholds (minimal = 0.65, optimal = 0.8) to maintain consistency with the original HOCOMOCO v10 quality ratings.

At the benchmark stage, we performed a comprehensive cross-species evaluation of the models. For 118 out of 384 benchmarked TFs, the best TFBS model for human data was the one derived from a mouse peak set. Similarly, for 85 out of 322 TFs the best model for mouse data was a model derived from a human peak set. This means that for a particular TF differences in binding patterns between species are less obvious than between experimental conditions or cell types. This observation allowed us to make a cross-species transfer of models between human and mouse orthologous TFs.

For some TFs, the ChIP-Seq data were available only for one of the two species. In such cases, we conducted a cross-species transfer of the models if the respective TF was previously included in the HOCOMOCO v10 set for the target species. The respective quality of the new cross-species model was set one category lower (e.g., A substituted by B) to account for the uncertainty introduced by a cross-species transfer. If the resulting quality has been better or equal to that of the existing HOCOMOCO v10 model, the new model from a cross-species transfer has been adopted for v11. Otherwise, the existing v10 model has been kept. The quality values directly inherited (non-benchmarked) HOCOMOCO v10 S-models (describing secondary patterns) have been set one category lower than those of the primary models.

At this step, we re-examined D-quality models of HOCOMOCO v10 and excluded those that were not confirmed by the ChIP-Seq data such as YBX1 (28). By doing so, unreliable binding models for eight human and three mouse TFs previously included in HOCOMOCO v10 were discarded.

The HOCOMOCO v11 model identifiers have the following format: 'X.H11MO.Y.Z' where X is the UniProt ID for the TF, Y is the model rank, and Z is the model quality rating. Dinucleotide models are labeled with 'H11DI'. If the curation stage yielded several models (primary, single-box, etc.), they were ranked based on AUC logROC with the best model being ranked as 0 and 1–2 ranks assigned to alternative models.

RESULTS

As compared with the previous release (v10), in HOCOMOCO v11 we increased the number of mononucleotide models by 131/105 (for human/mouse respectively) and thus additionally covered 88/61 TFs (totaling to 771/531 models and 680/453 covered TFs). The overall performance of the models was notably improved (Figure 1), in particular, 31 human models from HT-SELEX data were replaced by the new ChIP-Seq-based models. In regards to the structural classification of transcription factors (25), the class of C2H2 zinc finger factors profited the most from the update growing from 92 to 159 binding models.

The collection of dinucleotide PWMs that covers only primary models has also been notably expanded with 227

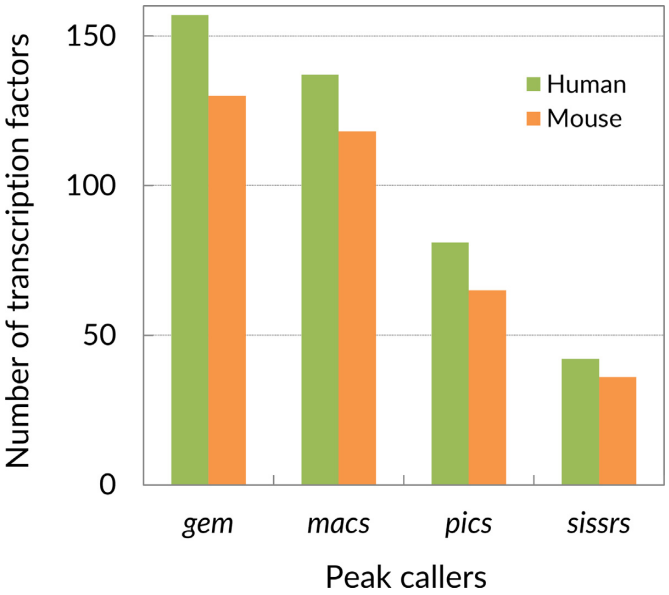

Figure 2. The number of TFs with the best-performing TFBS models (Y-axis) constructed from the peaks obtained with a particular peak caller (X-axis).

TFs (totaling to 313) for human and 211 TFs (totaling to 263 TFs) for mouse.

An interesting by-product of our analysis was the comparison of the peak callers regarding the quality of the TFBS models derived from the respective peak sets. To this end, we used the motif quality metrics already obtained at the benchmarking step and estimated the number of transcription factors for which the best available model was derived from the peak set produced by a particular peak caller. To rank the peaks produced by each peak caller we selected the measure that was the most similar to a peak height, and performed motif discovery and validation based on the highest peaks. The resulting TFBS models had notably different performance. The best peak caller was *gem*, as it was the source of the best TFBS models across the widest range of TFs. The popular *macs* appeared close second. The remaining two peak callers (*sissrs* and *pics*) performed noticeably poorer. This might be one of the reasons for the quality improvement of HOCOMOCO v11, because v10 was based solely on *sissrs*. Figure 2 shows the total number of human and mouse TFs for which the best models were derived from peaks of the particular peak caller.

DISCUSSION

For many TFs the general HOCOMOCO pipeline was successful in constructing and benchmarking the TFBS models. However, multiple questions remain open. In particular, for many TFs (including many homeodomain and putatively AT-rich binding TFs) we failed to construct reliable binding models. Some cases might be a consequence of the limited data quality, but there were dozens of TFs with multiple ChIP-Seq data sets for which our procedure failed to discover a clear pattern consistent across data sets. One of such cases is ARID3A, which is expected to bind AT-rich patterns according to previously available *in vitro* data. However, we failed to identify any common binding pattern *in vivo*, and also, no common pattern could be found in FactorBook data as well (29). Thus, a general question remains: which fraction of TFs do not exhibit clear binding patterns *in vivo* and if this is the case, is this determined by functional non-specific binding related to the architecture of the DNA-binding domains? The alternative explanation might be the prevalent indirect DNA binding or just a limited quality of the currently available antibodies.

Models of SMAD TF family make up another example. It appears that there are fuzzy but still visible patterns in aligned TFBS for these TFs. It is not clear whether these TFs require alignment-free TFBS models or if they correspond to alignable but multiple motif subtypes (Figure 3A).

Other non-trivial cases include composite elements detected as primary models. One surprising example is ANDR-FOX composite element detected as a primary model for ANDR (Figure 3B). Another non-trivial case is NANOG with the respective ChIP-Seq data stably yielding SOX2-OCT4 composite element as a primary model. Though it has been proposed that OCT4 (POU5F1), SOX2 and NANOG can potentially form a composite element (30), further analysis is needed to prove this hypothesis. The 'genuine' homeodomain-like NANOG pattern can be found in mouse ChIP-Seq data with differential motif discovery by contrasting NANOG ChIP-Seq with that of OCT4 (31). It is possible to obtain similar results using ChIPMunk by excluding the regions overlapping OCT4 and SOX2 peaks from the training NANOG sequence set (Figure 3C). The resulting 'genuine' NANOG model is in better agreement with model from SELEX experiments (32) and shows good performance reaching B/A rating on human/mouse ChIP-Seq data.

On a broader scale, one can observe, in general, a good similarity between TF binding models for TFs belonging to the same structural families. However, there are especially interesting cases such as GC-box binding Sp-family of TFs. The TFs with newly processed ChIP-Seq data include SP7, which binds a completely different pattern (Figure 3D). The model was rejected during automatic annotation but restored at the manual curation stage based on evidence from (33).

One of the distinguishing features of HOCOMOCO is the unified pipeline of motif discovery and benchmarking. For HOCOMOCO v11 we performed a careful reprocessing of the experimental data rather than the aggregation of TFBS models reported by other studies (34–36). Our approach has its disadvantages. Specifically, the compulsory curation of the discovered models is dependent on external data, e.g. UniProt (37), TF Encyclopedia (38), and the JASPAR database (39,40), which we consulted at the model curation stage.

Practical usage of HOCOMOCO models

Different practical applications of HOCOMOCO require different subsets of models. In the case of TFBS prediction for the analysis of regulatory networks, the D-quality models may notably reduce performance (41). However, other applications such as annotation of regulatory sequence variants (42,43) may benefit from additional TFs and alterna-

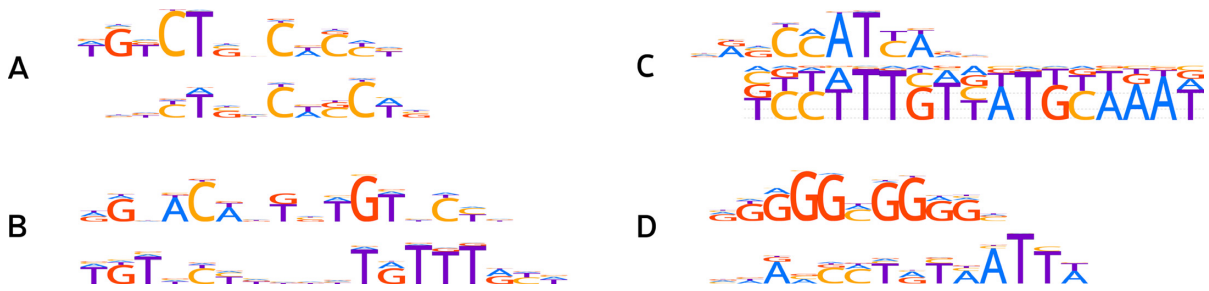

Figure 3. Examples of TFBS models included in HOCOMOCO v11: (**A**) SMAD2/3 models exhibit fuzzy patterns; (**B**) two variants of ANDR composite elements: the palindromic site and the double box ANDR-FOX; (**C**) the 'genuine' NANOG binding site along the putative OCT4-SOX2/NANOG composite element (for clarity the nucleotide pileups are shown unscaled); (**D**) SP1/2 (canonical SP-family GC-box) and SP7 binding models.

tive binding models. Thus, starting from this release we provide two collections: the CORE collection of the most precise models (ABC quality) recommended for TFBS predictions for annotation of regulatory regions, and the FULL collection that includes alternative models and models of lower quality (D), which can be used for exploratory purposes. By design, the dinucleotide collection is aimed specifically for precise TFBS prediction and includes only primary ABC quality models. To facilitate practical applications, we continue to provide *P*-value-to-score-threshold mappings for all models and the complete collection dump in various formats (including plain text, TRANSFAC and MEME).

As in HOCOMOCO v9 and v10, we continue to provide a subset of default thresholds at fixed motif *P*-values (44) with the default recommended *P*-value of 0.0005 (a single expected prediction per 1000 bp of a two-stranded random sequence). At this level HOCOMOCO v11 model reaches median sensitivity of 0.75, i.e. typically recognizes three of four binding sites in the positive control data.

When it comes to motif finding, HOCOMOCO now has several options including a command-line tool (SPRY-SARUS) and two web-interfaces: MoLoTool for visual inspection of shorter sequences, and the HOCOMOCO-in-BioUML (http://hocomoco.bioumL.org) to scan extensive genomic regions.

Marking motif occurrences with MoLoTool

Major motif analysis software (14,45) include methods for motif finding, i.e., to detect occurrences of given motifs in a user-defined set of sequences of genomic regions. Practical applications often require detailed analysis of particular DNA sequences in the vicinity of TFBS at a single-nucleotide resolution, e.g. to rationally change TFBS affinity by site-specific mutations. To facilitate such local analysis with HOCOMOCO v11 models, we designed MoLoTool (Motif Location Toolbox) which allows scanning a given set of sequences for occurrences of user-selected PWMs. The basic workflow is very simple: (a) select a desired subset of TFBS models from an interactive catalog, (b) paste or upload a set of DNA sequences and (c) submit the form to obtain not only the table view of predicted TFBS but also the colored sequence-based map of motif occurrences. The main advantage of MoLoTool is its ability to produce text-based colored markings which can be directly copied-and-pasted into any rich text editor (e.g., MS Word). The TFBS predictions depend on the motif *P*-value threshold that can be dynamically adjusted to seamlessly update the markings without reloading the web page. An example of the MoLoTool markup is shown in Supplementary Figure S5.

CONCLUSIONS

To summarize, we report a major update of the HOCOMOCO collection (v11) that contains binding models of the increased quality for 453 mouse and 680 human transcription factors and includes 1302 mononucleotide and 576 dinucleotide PWMs. The update was constructed by systematic motif discovery and benchmarking based on more than five thousand ChIP-Seq experiments processed with four ChIP-Seq peak calling tools.

We complement HOCOMOCO by MoLoTool that can perform visual mapping of HOCOMOCO binding sites in DNA sequences. We believe that the current version of HOCOMOCO can be used for diverse applications in fundamental and applied research from global TFBS prediction to the annotation of regulatory sequence variants.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Evolutionary Genomics Laboratory, Faculty of Bioengineering and Bioinformatics (Lomonosov Moscow State University) and personally Prof. A.S. Kondrashov for computational facilities. We kindly thank Nicholas Mullin for providing access to NANOG SELEX data (32). We thank Noel Alonso, Greg Wickham and Craig Kapfer from the KAUST Research Computing Core Laboratories for technical help regarding implementation of the mirror website.

FUNDING

The project was primarily supported by Russian Science Foundation [17-74-10188 to I.V.K.]; A.M.M. and V.B.B. were supported by King Abdullah University of Science and Technology (KAUST) [baseline fund BAS/1/1606-01-01 of V.B.B.]; I.E.V. was personally supported by the

Skoltech Systems Biology Fellowship. Funding for open access charge: Russian Science Foundation [17–74–10188 to I.V.K.].
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Schwartz,A.M., Demin,D.E., Vorontsov,I.E., Kasyanov,A.S., Putlyaeva,L.V., Tatosyan,K.A., Kulakovskiy,I.V. and Kuprash,D.V. (2017) Multiple single nucleotide polymorphisms in the first intron of the IL2RA gene affect transcription factor binding and enhancer activity. *Gene*, **602**, 50–56.
3. Schwartz,A.M., Putlyaeva,L.V., Covich,M., Klepikova,A.V., Akulich,K.A., Vorontsov,I.E., Korneev,K.V., Dmitriev,S.E., Polanovsky,O.L., Sidorenko,S.P. *et al.* (2016) Early B-cell factor 1 (EBF1) is critical for transcriptional control of SLAMF1 gene in human B cells. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1859**, 1259–1268.
4. Boeva,V. (2016) Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. *Front Genet.*, **7**, 24.
5. Vorontsov,I.E.I.E., Khimulya,G., Lukianova,E.N.E.N., Nikolaeva,D.D.D., Eliseeva,I.A.I.A., Kulakovskiy,I.V.I.V. and Makeev,V.J.V.J. (2016) Negative selection maintains transcription factor binding motifs in human cancer. *BMC Genomics*, **17**, 395.
6. 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.R. and Bajic,V.B. (2014) Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, **15**, 119.
7. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 375.
8. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
9. Kulakovskiy,I.V, Levitsky,V., Oshchepkov,D., Bryzgalov,L., Vorontsov,I. and Makeev,V. (2013) From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **11**, 1340004.
10. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
11. Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Lassmann,T., Itoh,M., Summers,K.M., Suzuki,H., Daub,C.O. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
12. Gursky,V.V., Kozlov,K.N., Kulakovskiy,I.V., Zubair,A., Marjoram,P., Lawrie,D.S., Nuzhdin,S.V. and Samsonova,M.G. (2017) Translating natural genetic variation to gene expression in a computational model of the Drosophila gap gene regulatory network. *PLoS One*, **12**, e0184657.
13. Balwierz,P.J., Pachkov,M., Arnold,P., Gruber,A.J., Zavolan,M. and van Nimwegen,E. (2014) ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.*, **24**, 869–884.
14. Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J.A., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C. *et al.* (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, **43**, W50–W56.
15. 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.
16. Narlikar,L. and Jothi,R. (2012) ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder. *Methods Mol. Biol.*, **802**, 305–322.
17. 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.
18. 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.
19. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2016) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
20. 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–D125.
21. Levitsky,V.G., Kulakovskiy,I.V, Ershov,N.I., Oschepkov,D.Y., Makeev,V.J., Hodgman,T.C., Merkulova,T.I., Oshchepkov,D.Y., Makeev,V.J., Hodgman,T.C. *et al.* (2014) Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genomics*, **15**, 80.
22. 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.
23. Kulakovskiy,I.V, Medvedeva,Y.A., Schaefer,U., Kasianov,A.S., Vorontsov,I.E., Bajic,V.B. and Makeev,V.J. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, **41**, D195–D202.
24. Wingender,E., Schoeps,T. and Donitz,J. (2012) TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.*, **41**, D165–D170.
25. 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.
26. Medina-Rivera,A., Abreu-Goodger,C., Thomas-Chollier,M., Salgado,H., Collado-Vides,J. and Van Helden,J. (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.*, **39**, 808–824.
27. Wilson,S., Qi,J. and Filipp,F.V (2016) Refinement of the androgen response element based on ChIP-Seq in androgen-insensitive and androgen-responsive prostate cancer cell lines. *Sci. Rep.*, **6**, 32611.
28. Dolfini,D. and Mantovani,R. (2012) YB-1 (YBX1) does not bind to Y/CCAAT boxes in vivo. *Oncogene*, **32**, 4189–4190.
29. 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.
30. Papatsenko,D., Darr,H., Kulakovskiy,I.V., Waghray,A., Makeev,V.J., MacArthur,B.D. and Lemischka,I.R. (2015) Single-cell analyses of ESCs reveal alternative pluripotent cell states and molecular mechanisms that control self-renewal. *Stem Cell Rep.*, **5**, 207–220.
31. Maaskola,J. and Rajewsky,N. (2014) Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res*, **42**, 12995–13011.
32. Gagliardi,A., Mullin,N.P., Ying Tan,Z., Colby,D., Kousa,A.I., Halbritter,F., Weiss,J.T., Felker,A., Bezstarosti,K., Favaro,R. *et al.* (2013) A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J.*, **32**, 2231–2247.
33. Hojo,H., Ohba,S., He,X., Lai,L.P. and McMahon,A.P. (2016) Sp7/Osterix is restricted to bone-forming vertebrates where it acts as a Dlx co-factor in osteoblast specification. *Dev. Cell*, **37**, 238–253.
34. Sebastian,A. and Contreras-Moreira,B. (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, **30**, 258–265.
35. Verfaillie,A., Imrichova,H., Janky,R. and Aerts,S. (2015) iRegulon and i-cisTarget: reconstructing regulatory networks using motif and track enrichment. *Curr. Protoc. Bioinformatics*, **52**, 2.16.1–2.16.39.
36. 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.
37. The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
38. Yusuf,D., Butland,S.L., Swanson,M.I., Bolotin,E., Ticoll,A., Cheung,W.A., Zhang,X.Y.C., Dickman,C.T.D., Fulton,D.L., Lim,J.S.

et al. (2012) The transcription factor encyclopedia. *Genome Biol.*, **13**, R24.
39. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
40. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
41. Schmidt,F., Gasparoni,N., Gasparoni,G., Gianmoena,K., Cadenas,C., Polansky,J.K., Ebert,P., Nordström,K., Barann,M., Sinha,A. *et al.* (2017) Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.*, **45**, 54–66.
42. Deplancke,B., Alpern,D., Gardeux,V., Adam,R.C., Yang,H., Rockowitz,S., Larsen,S.B., Nikolova,M., Oristian,D.S., Polak,L. *et al.* (2016) The genetics of transcription factor DNA binding variation. *Cell*, **166**, 538–554.
43. Afanasyeva,M.A., Putlyaeva,L.V., Demin,D.E., Kulakovskiy,I.V., Vorontsov,I.E., Fridman,M.V., Makeev,V.J., Kuprash,D.V. and Schwartz,A.M. (2017) The single nucleotide variant rs12722489 determines differential estrogen receptor binding and enhancer properties of an IL2RA intronic region. *PLoS One*, **12**, e0172681.
44. Touzet,H. and Varré,J.-S. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol. Biol.*, **2**, 15.
45. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME suite. *Nucleic Acids Res.*, **43**, W39–W49.