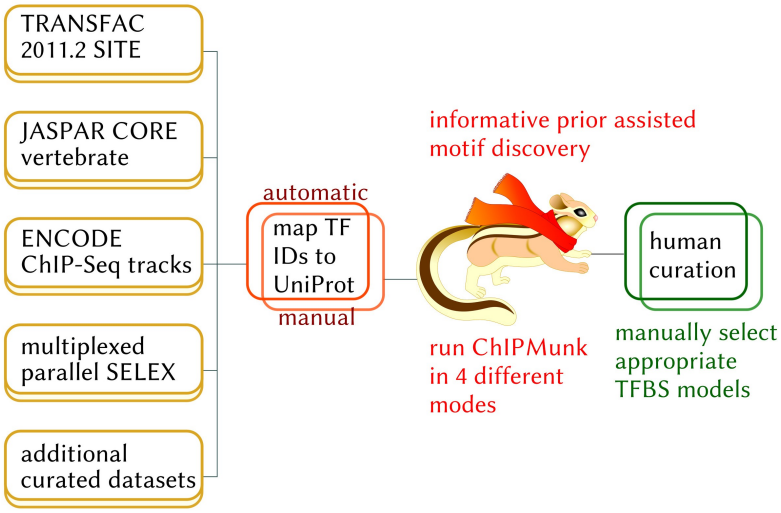
Supplementary section 1. Data preprocessing strategies

Supplementary section 1 figure 1. Simplified representation of the HOCOMOCO workflow



TRANSFAC and JASPAR

Both TRANSFAC and JASPAR contain segments of DNA binding regions with uppercase-marked binding sites. We use this information to provide ChIPMunk with the sequence positional prior. A binding region obtains a positional profile value equal to 1.0 for the uppercase positions and 0.5 for the remaining positions. Additionally, for TRANSFAC data, we downscale the profile by multiplying it by a coefficient varying from 0.5 to 1.0 depending on the TRANSFAC quality ratings as annotated in the TRANSFAC SITE table, a value that varies from 1 to 6 respectively (from the worst to the best site quality).

For all data sources except TRANSFAC, we made TF ID mapping to UniProt IDs by hand. The TRANSFAC collection was automatically processed in the following way. We collected all synonyms for each TRANSFAC and for UniProt protein records. The TF ID from TRANSFAC was considered as mapped when this TF had the highest number of common synonyms with a particular UniProt entry. In a number of cases this type of mapping was ambiguous due to different inconsistencies in both databases, so we tried to select the most appropriate UniProt ID by an additional deeper inspection of synonym lists.

For some TFs like GABP1/2 or ATF2/4 it was impossible to distinguish UniProt entries based on synonym lists because those TFs were subunits of a protein complex or isoforms of the same protein; hence, they had identical synonymous lists. Such IDs in the collection are listed as the GABP1+GABP2 or ATF2+ATF4. For the cases when TRANSFAC binding site entries were linked with several TFs, we tried to link this binding site to all corresponding UniProt IDs.

Text-mining and manually curated data sets

For all TFs having no more than 8 binding sites in all the sources mentioned above, we

have added binding sites found in different PubMed articles, but not presented in TRANSFAC or JASPAR.

For TFs mentioned in the TcoF database (Schaefer et al., 2011) but not having binding sites in other processed sources an association network was created based on a co-occurrence of a word-combination “binding site” and any of the synonyms for particular TF available in UniProt within one sentence in the PubMed abstract. The papers thus obtained were manually curated and available binding sites were added to the collection. For binding sites obtained by text-mining and manual selection, we provide PubMed IDs/paper references directly in filenames or fasta-file headers.

We also specifically looked for several specific TFs and their binding sites. One such TF group contains methyl-binding proteins and TFs whose binding sites are known to be affected by DNA methylation, particularly KAISO, MBD2, MECP2, UHRF1, and ZBTB4. In the case when experimental conditions allowed us to conclude if binding sites were methylated or not, the methylated and unmethylated sets of binding sites were processed separately resulting in two models (e.g., ZBTB4 and ZBTB4!METH). We also have added manually curated data for HIF1A and P53.

If any information about TF affinity to some site was available (e.g., for EMSA), we used it to assign weight for particular binding sequences.

Removing redundant flanking sequences

TRANSFAC entries for some TFs contained redundant identical flanking sequences possibly related to the experimental design and not removed when the sequences were included in the database. These identical flanks aligned better than the remaining sequences, thus reducing the effectiveness of motif discovery. We iteratively removed the identical flanking sequences of length from 6 to 8 bp from both sequence ends if they were presented in no less than 4 sequences in the set.

Multiplexed parallel SELEX

We used only the data from the last SELEX cycle. We used only the sequences found more than once. We took no more than 2000 sequences with top copy numbers; when some sequences had the same copy number as those in the top 2000 we included all of them. The sequence weights were assigned as the log(reads count).

All sequences coming from data sources except ChIP-Seq were additionally extended by "N" letters at both ends (to avoid possible truncation of binding sites found at the end of short DNA fragments).

Yale ChIP-Seq (Release3, ENCODE data)

We have extracted the base coverage data to provide ChIPMunk with the sequence positional prior. Data for all available experiments/cell lines was merged. We have created the set of sequences by taking no less than 2000 of the top peaks (in cases where there were several more peaks having the same height as those already included, we also included all such peaks); for sets with less than 2000 peaks all sequences were taken. To

speed-up the motif discovery, the peaks were truncated by discarding flanks with a signal lower than 10% of peak height. The peaks shorter than 25bps were discarded.

HudsonAlpha ChIP-Seq (Release2, ENCODE data)

The base coverage data was unavailable for HudsonAlpha ChIP-Seq, so we used the signalValue as the sequence weight. ChIP-Seq peak subset for each TF was created using the following multi-step approach:

(1) We selected all peak segments no longer than 1000bp having a signalValue higher than 2.

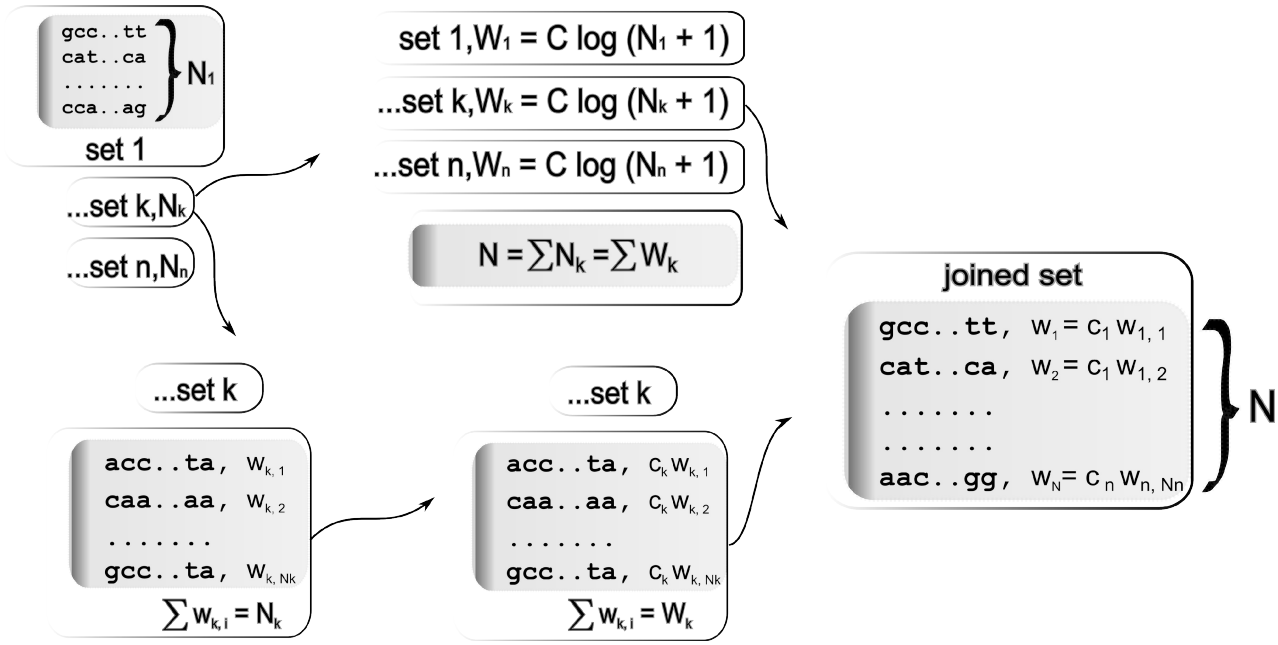
(2) If the number of sequences was less than 200, we added sequences no longer than 2000bp having a signalValue higher than 1.

(3) If the number of sequences was still less than 200, we selected the whole set of peaks.

(4) Finally the signalValue threshold was selected corresponding to no less than 2000 peaks or the whole set if it contained less than 2000 sequences.

Supplementary section 2. Sequence weighting strategy, ChIPMunk workflow and parameter settings

Supplementary section 2a. Sequence weighting strategy



Here $N_1 .. N_k .. N_n$ are the sizes of n given sequence sets; $W_1 .. W_k .. W_n$ are the corresponding weights of these sequence sets; $w_{1,1} .. w_{k,i} .. w_{n,Nn}$ are the initial weights of sequences. The total weight of the joined set is always equal to its overall size N. The final normalized weights, w_i will be used in the motif discovery procedure.

Supplementary section 2b. ChIPMunk PCM to PWM transformation

Position count matrices are converted to position weight matrices according to the strategy used in [Lifanov et al., 2003]:

$$S_{\alpha,j}=\log\left(\frac{x_{\alpha,j}+a\,q_\alpha}{(N+a)\,q_\alpha}\right),$$

where N is the total number of sequences, $S_{\alpha,j}$ is PWM score for nucleotide α in position j, a is the pseudocount value which was taken equal to $\sqrt{N}$, and $x_{\alpha,j}$ is the position count matrix (PCM) element.

Selecting optimal alignment

ChIPMunk searches (see the next section) for the gapless multiple local alignment with the maximum KDIC value.

KDIC refers to the Kullback Discrete Information Content:

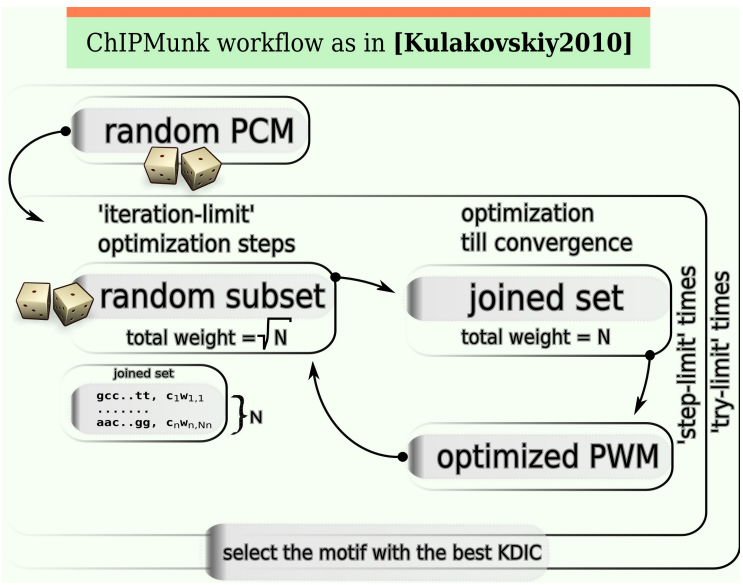
$$KDIC=DIC-\sum_{j=1}^{l}\sum_{\alpha\in\{A,C,G,T\}}x_{\alpha,j}\log q_\alpha,$$

where l it the PWM width (the motif length), q_α is the background probability for nucleotide α, and

$$DIC=\sum_{j=1}^{l}\sum_{\alpha\in\{A,C,G,T\}}\left(\log(x_{\alpha,j}!)-\log(N!)\right).$$

Actually, to compute factorials we used the Stieltjes approximation [Abramowitz1972], which also holds for the real x. The details are given in [Kulakovskiy2010 Supplementary] section 1.

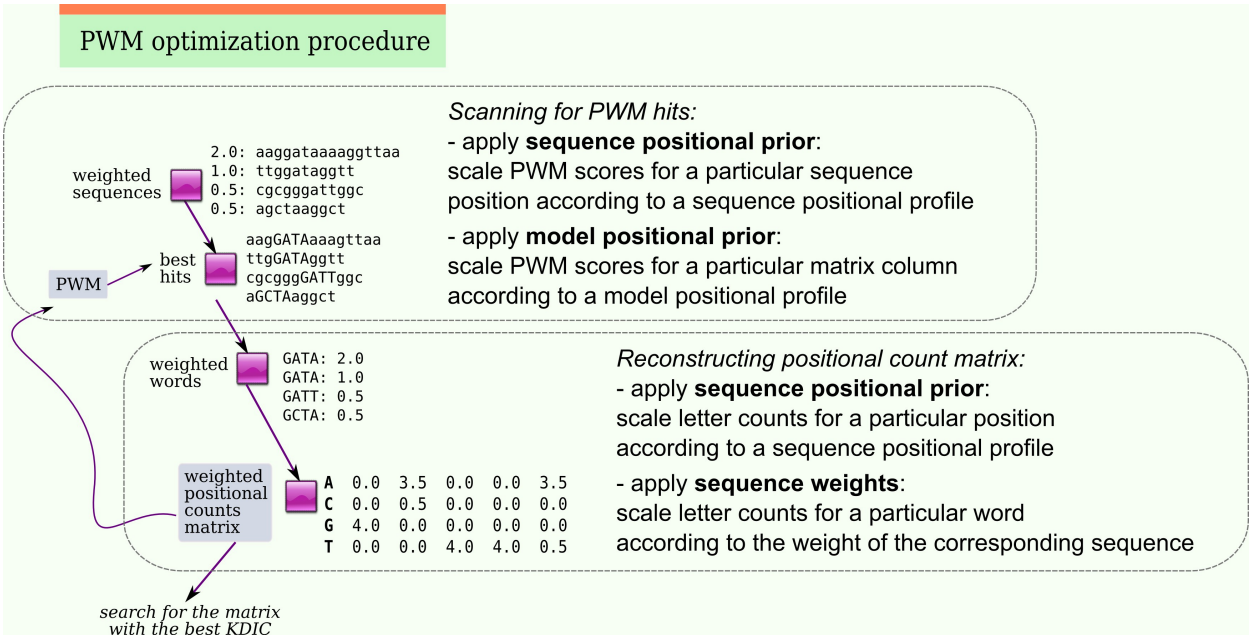
Supplementary section 2c. Overall ChIPMunk workflow



ChIPMunk is a subsampling-based greedy motif discovery algorithm based on simple PWM optimization procedure. ChIPMunk iteratively optimizes a starting random matrix and a corresponding alignment using a data subset and then optimizes the PWM and alignment using the whole data set. Different random matrices are used as seeds at the start and the best optimal model is finally selected according to the highest KDIC value.

Supplementary section 2d. PWM optimization accounting for positional priors

PWM optimization iteratively selects the best PWM hit from each sequence and then reconstructs the PWM. Details on applying positional priors are given below.



Sequence positional prior (or a sequence positional profile) defines a real positive value for each particular nucleotide position. For example such a profile can be derived from the ChIP-Seq peak shape ("base coverage") data. An initial weight for each sequence is the corresponding peak height (i.e. the maximum value of the profile). Then the profile is normalized to fit in [0,1] interval. The sequence then receives a normalized weight according to the weighting scheme described in the previous section.

The sequence without positional profile can be imagined with the flat positional profile filled with 1.0 values.

Model positional prior is the array of the same length as PWM having values from 0.0 to 1.0 where 1.0 corresponds to unchanged scoring in position and 0.0 would correspond to position that should have no impact on scoring (i.e. gap position). While arbitrary model priors are not supported by ChIPMunk at this moment there are two predefined model priors for single- and double-box motifs with the box size adapted for DNA helix pitch as it is described in the corresponding section in the manuscript.

Iterative PWM optimization includes two alternating steps: (i) the alignment is rebuilt from words with maximal PWM scores in each sequence and (ii) the PWM is rebuilt from the new alignment from the previous step.

At the first step PWM scores for putative hits are weighted based on sequence positional profiles (*spp*) and model positional prior (*mpp*):

$$score(\mathrm{PWM}, \mathrm{word}) = \sum_{j=1}^{l} S_{\mathrm{word}[j],j} \cdot \mathrm{spp}[j] \cdot \mathrm{mpp}[j]$$

where l it the PWM width (the motif length) and, $S_{\mathrm{word}[j],j}$ is the PWM element for the *j*-th letter in the word. Thus the positions with larger profile values (either *mpp* or *spp*) contribute more (either positively or negatively) into the PWM score.

Let us have the *i*-th sequence from the total joined set having the final weight of w_i. When the best PWM hits (best words) are selected from each sequence one can estimate the positional weighted count $x_{\alpha,j}$ (which are in this case real) for each letter α for each motif position j:

$$\alpha \in \{A, C, G, T\}: \; x_{\alpha,j} = \sum_i w_i \cdot \mathrm{profile}_i[j] \cdot \delta(\alpha, \mathrm{word}_i[j])$$

$$\delta(a,b): \begin{cases} a=b: \delta(a,b)=1 \\ a \neq b: \delta(a,b)=0 \end{cases}$$

$$x_{\mathrm{N},j} = \sum_i w_i \cdot (1 - \mathrm{profile}_i[j])$$

$$\forall j: \sum_{\alpha \in \{A,C,G,T,\mathrm{N}\}} x_{\alpha,j} = \sum_i w_i$$

Note, that here $\mathrm{profile}_i[j]$ is the sequence positional profile of *i*-th word at *j*-th position and **N** denotes an 'unknown nucleotide'. After collecting the data from all words, $x_{\mathrm{N},j}$ values are uniformly split between $x_{\alpha,j}$ so the total sum is equal to the number of words.

Supplementary section 2e. ChIPMunk parameters and length ranges

The default alignment length range was from 7 to 25bp (ChIPMunk “motif length” parameters). The maximum length of 25bp was reduced in the following cases.

(a) For sets of 8 or less sequences, the maximum alignment length was taken as the length of the shortest available binding sequence.

(b) For sets of 16 or less sequences, the maximum alignment length was taken as the length of the second shortest available binding sequence.

(c) For all other sets, the maximum alignment length was selected so that no less than 70% of available sequences were longer than the maximum length.

The number of starting seeds (the “try_limit” ChIPMunk parameter) was increased for smaller sets (where the signal is not so stable and therefore is more difficult to detect correctly): 2000 seeds for the sets containing from 5 to 200 sequences, 200 seeds for the sets containing from 200 to 1000 and 100 seeds (default) for the extra-large sets of more than 1000 sequences. The default number of bootstraps (the ChIPMunk “step_limit” parameter, 10 by default) was increased up to 20 for smaller sequence sets (having 200 and less sequences).

The background was assumed to be uniform for all the cases with the exception for ChIP-Seq data, where ChIPMunk was allowed to automatically determine the background nucleotide distribution.

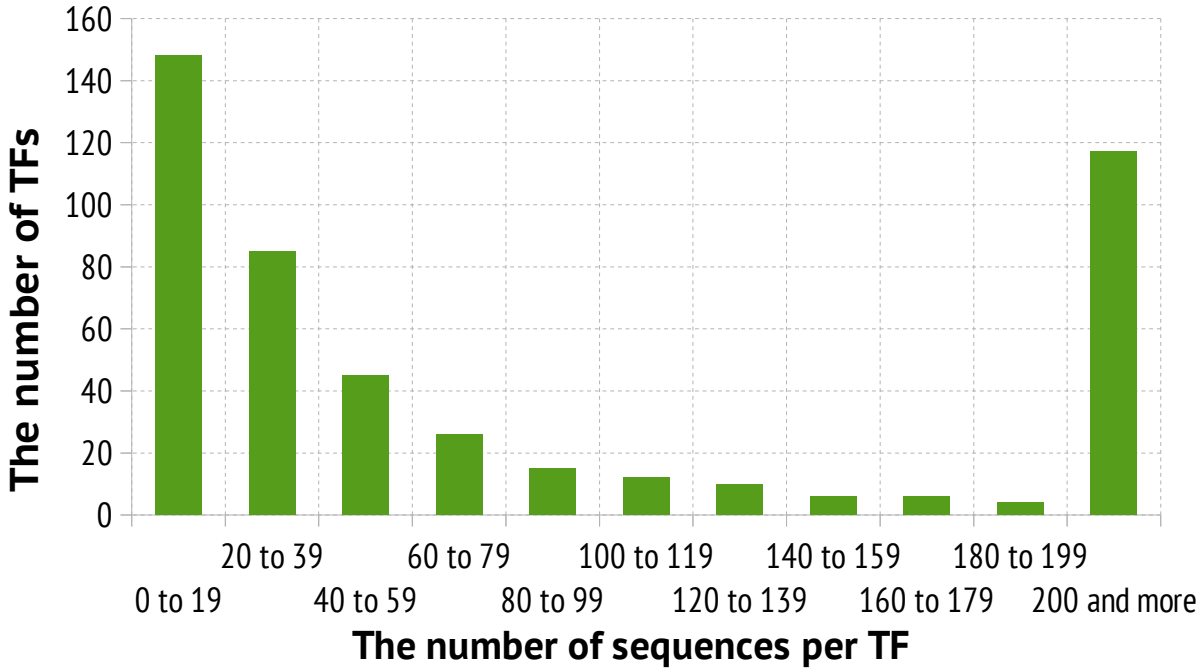
The alignment mode (one-occurrence-per-sequence, no noise filtering) was used for small sequence sets containing 16 and less sequences. The zero-or-one-occurrence-per-

sequence (ZOOPS) mode was used with the zoops_factor=1.0 (ChIPMunk "zoops-factor" parameter, 1.0 recommended default) for all cases except for the parallel SELEX data, where it was set to a more strict level of 0.0 to filter out noisy reads. This parameter is used to scale pseudocount values during a PWM self-consistency test. More details on ZOOPS mode are given in the corresponding section in the Supplementary text of the ChIPMunk paper [Kulakovskiy2010].

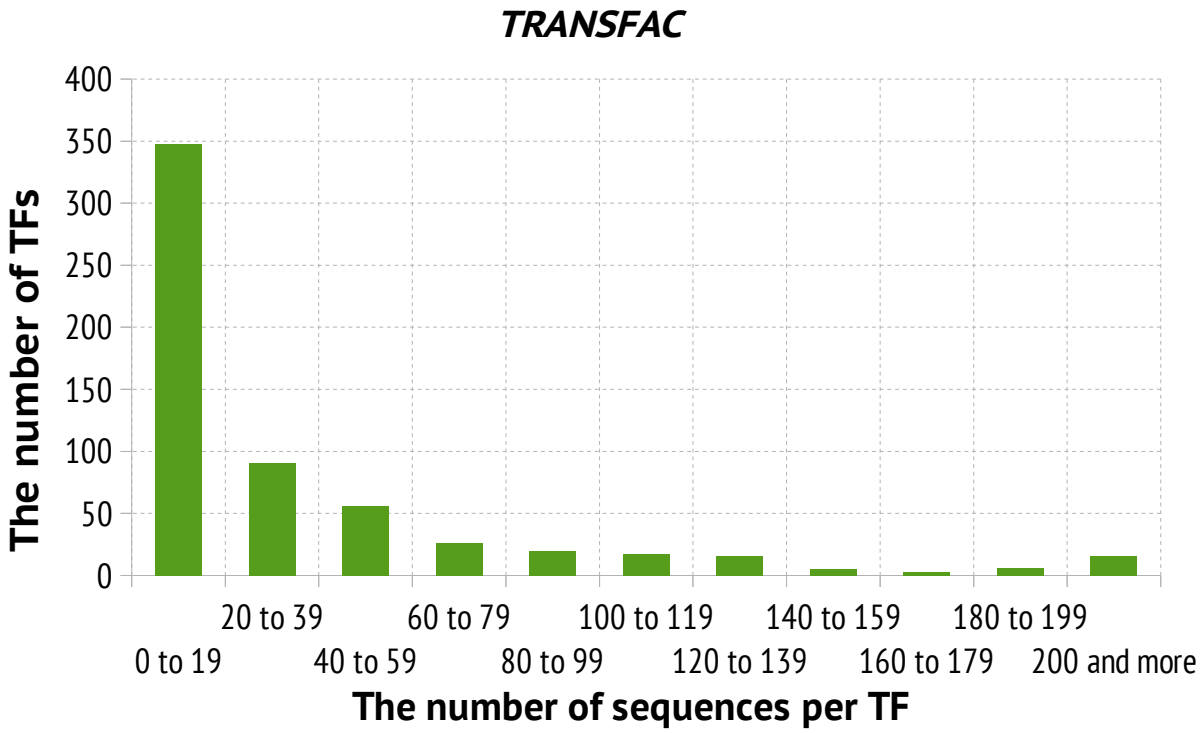
See ChIPMunk supplementary material [http://autosome.ru/ChIPMunk/supplement.html] and manual [http://autosome.ru/smbsm/librettos/libretto_chipmunk/chipmunk_v33_manual.txt] for more details on parameter settings.

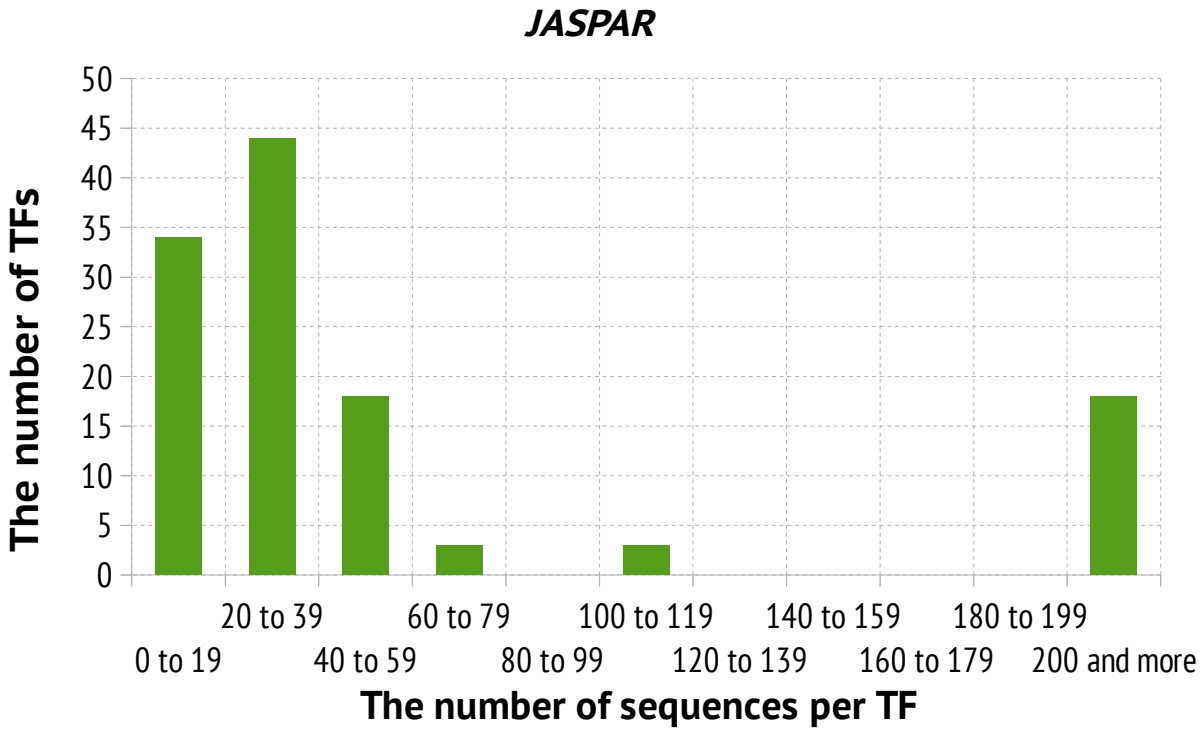
Supplementary section 3a. Overview of the initial data

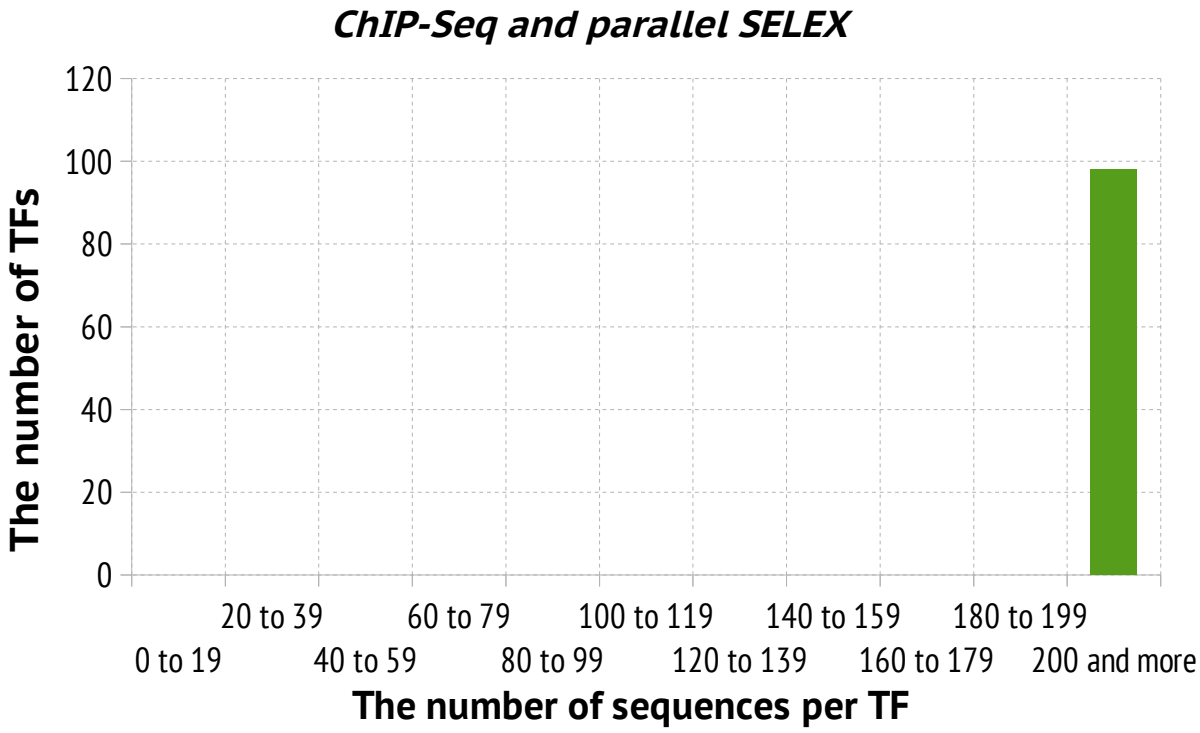
The number of TFs with a given number of known binding sequences (the total dataset)

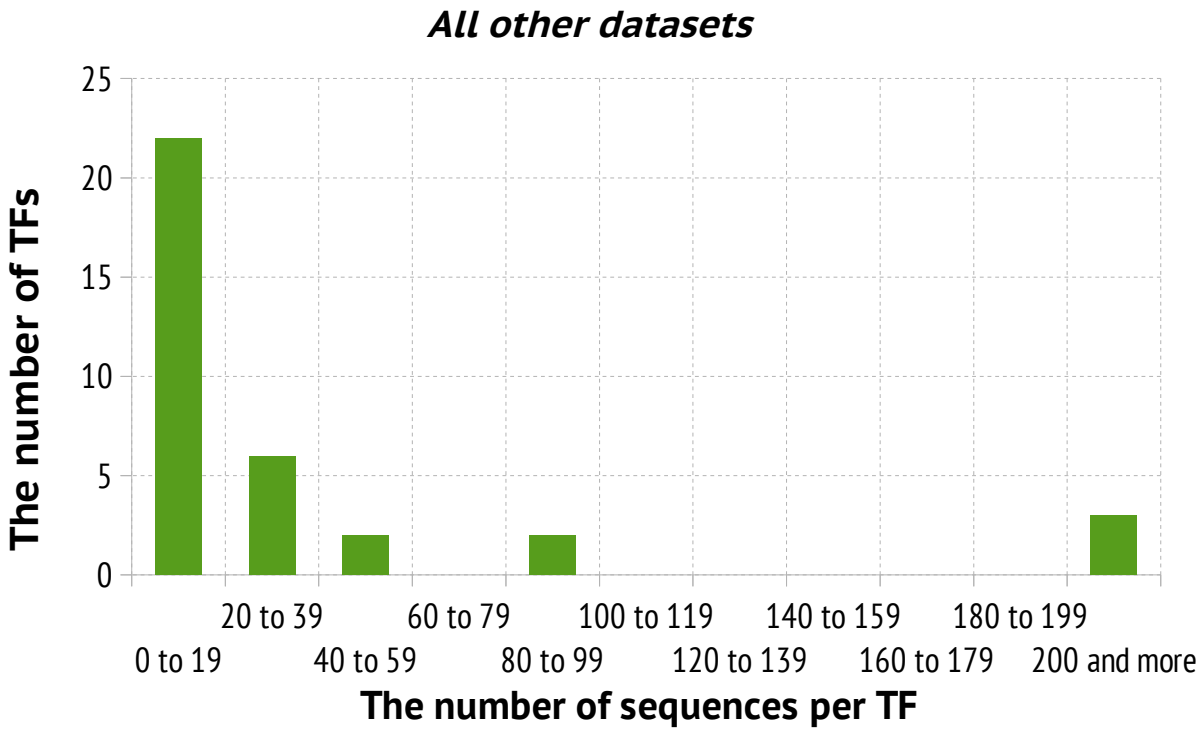


Separate graphs for each data source

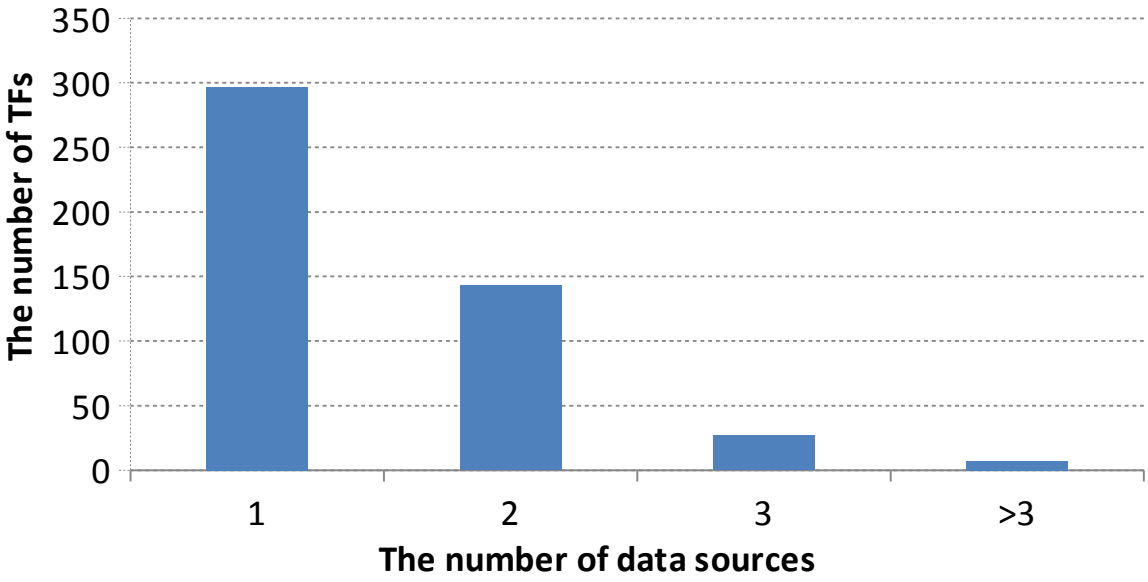





ChIP-Seq and parallel SELEX
The number of TFs
120
100
80
60
40
20
0
0 to 19
20 to 39
40 to 59
60 to 79
80 to 99
100 to 119
120 to 139
140 to 159
160 to 179
180 to 199
200 and more
The number of sequences per TF


All other datasets
The number of TFs
25
20
15
10
5
0
0 to 19
20 to 39
40 to 59
60 to 79
80 to 99
100 to 119
120 to 139
140 to 159
160 to 179
180 to 199
200 and more
The number of sequences per TF

The number of TFs with a given number of independent data sources



The number of TFs per data source

TRANSFAC	442
JASPAR	108
Yale ChIP-Seq	50
HudsonAlpha ChIP-Seq	28
Parallel SELEX	19
All other data sets	29

Supplementary section 3b. Venn diagram for data sources

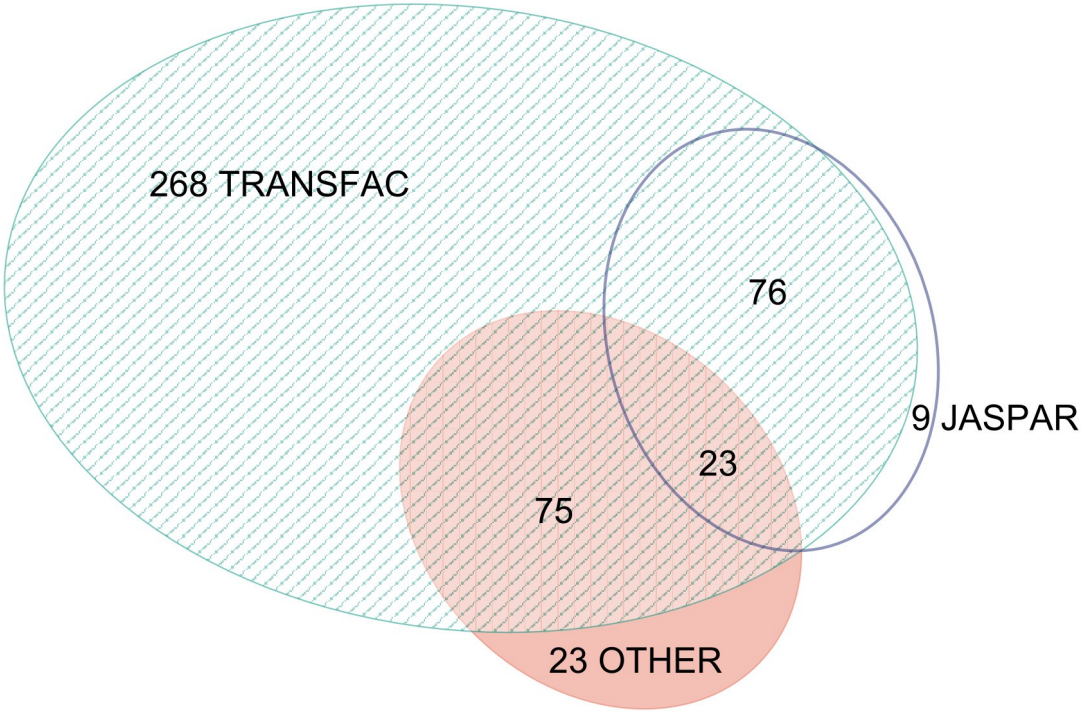


Diagram shows the number of TFs with binding sequences available in data sources used in construction of HOCOMOCO collection.

For TRANSFAC and JASPAR here we show only the data that we were able to map to UniProt IDs. "OTHER" set corresponds to all other data sources (ChIP-Seq, textmining, etc).

The diagram was produced with the help of eulerAPE [http://www.eulerdiagrams.org/eulerAPE/].

Supplementary section 4. Comparison of HOCOMOCO-AD models to the whole TRANSFAC PWM collection

Here we show that our TFBS models are significantly different from those in TRANSFAC even when based on TRANSFAC-only binding sequences. To this end we selected the most similar (i.e. the closest) model for each TRANSFAC PWM from HOCOMOCO-AD collection, and vice versa. The distance between PWMs was measured as the Jaccard distance between sets of words recognized by two PWMs at given threshold. To compute Jaccard distance we used MACRO-APE software [http://autosome.ru/macroape/] which constructs the optimal mutual alignment of two PWMs and computes the Jaccard similarity as a ratio of the

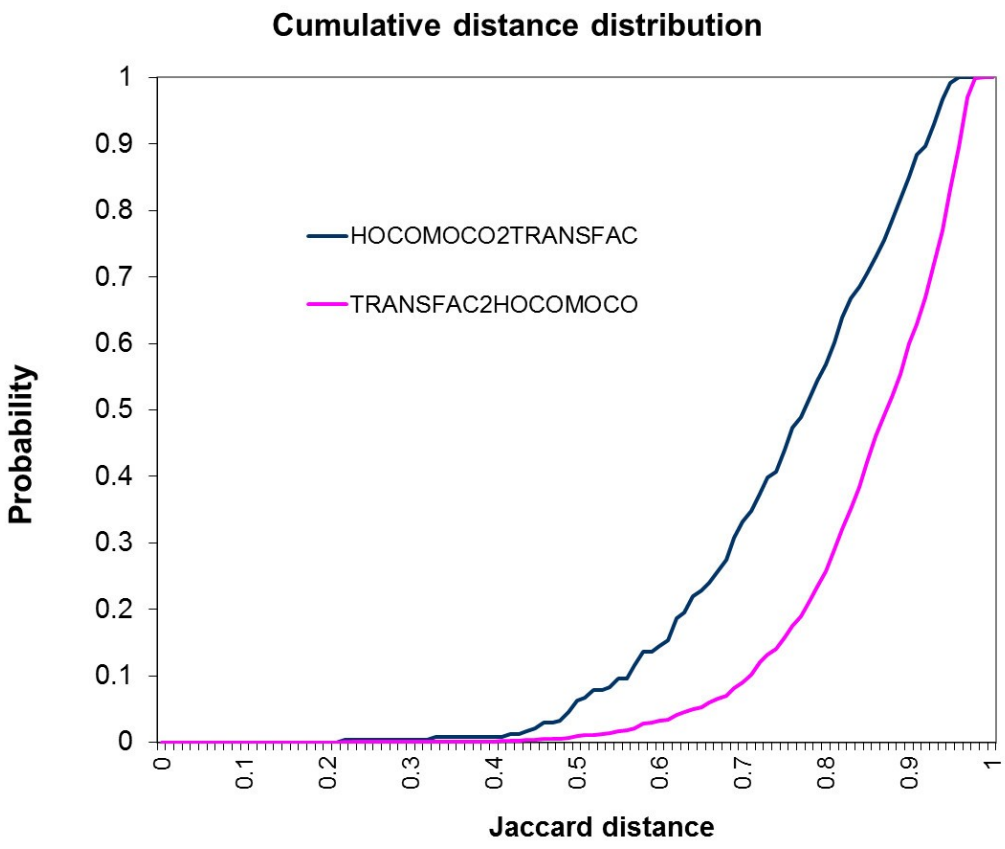
number of words recognized by both PWMs to the number of words recognized by at least one of PWMs. The Jaccard distance is defined as 1 – Jaccard similarity.

Then we computed a distribution of distances from HOCOMOCO-AD PWMs to their best matches in TRANSFAC (HOCOMOCO2TRANSFAC) and vice versa (TRANSFAC2HOCOMOCO). The figure below displays the fraction of HOCOMOCO PWMs having at least one TRANSFAC PWM at distances not greater than that given at X axis (the blue curve) and the fraction of TRANSFAC PWMs having at least one HOCOMOCO AD PWM at distances no greater than that given at X axis (the magenta curve). TRANSFAC contains much more PWMs than HOCOMOCO-AD, including those for invertebrates. Basically, for vast majority of HOCOMOCO-AD models there is at least one TRANSFAC PWM at a rather close distance, most likely representing the same TF. Conversely, TRANSFAC contains many PWMs, which are dissimilar to any HOCOMOCO-AD PWM. Hence, the magenta curve is located to the right from the blue curve.

HOCOMOCO-AD has many PWMs constructed from the TRANSFAC TFBS as a single source. However both curves clearly show that only few PWMs from HOCOMOCO-AD have a TRANSFAC PWM at a distance closer than 0.5. The distance of 0.5 means that only half of words recognized by any of PWMs is recognized by both PWMs.

Thus, there are **no** identical or highly similar models (zero and close to zero distances) when comparing HOCOMOCO-AD to TRANSFAC.

The P-value of 0.0005 was used to set the PWM thresholds.



References

Abramowitz,M. and Stegun, I.A. (1972) Handbook of Mathematical Functions. Dover Publications, New York

Kulakovskiy,I.V. et al. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, 26, 2622-3.

Lifanov,A.P. et al. (2003) Homotypic regulatory clusters in Drosophila. *Genome Res.*, 13, 579-88.

Schaefer,U. et al. (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res.*, 39, D106-10.