*NAR Genomics and Bioinformatics*, 2024, **6**, lqae090
https://doi.org/10.1093/nargab/lqae090
Advance access publication date: 27 July 2024
**Standard Article**

OXFORD

# Genomic background sequences systematically outperform synthetic ones in de novo motif discovery for ChIP-seq data

**Vladimir V. Raditsa[1], Anton V. Tsukanov[1], Anton G. Bogomolov[2] and Victor G. Levitsky[1,3,*]**

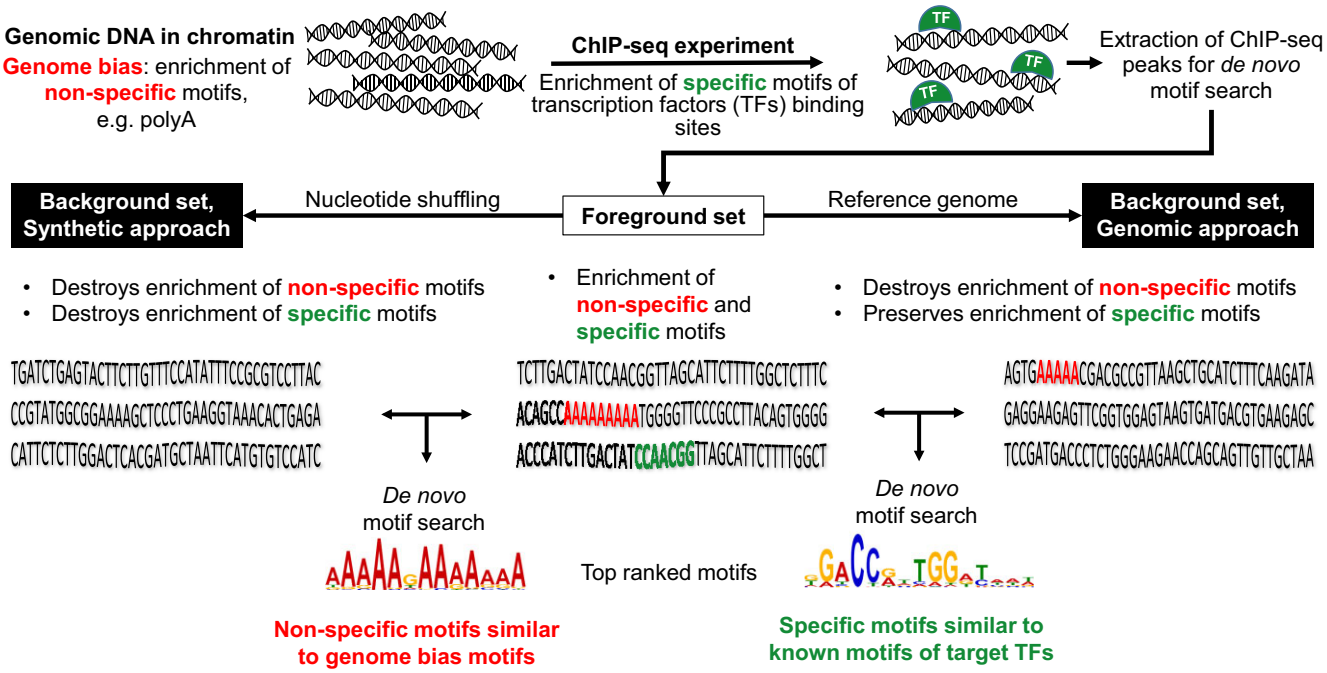
[1]Department of System Biology, Institute of Cytology and Genetics, Novosibirsk 630090, Russia
[2]Department of Cell Biology, Institute of Cytology and Genetics, Novosibirsk 630090, Russia
[3]Department of Natural Science, Novosibirsk State University, Novosibirsk 630090, Russia

[*]To whom correspondence should be addressed. Tel: +7 383 363 49 63; Fax: +7 383 333 12 78; Email: levitsky@bionet.nsc.ru
Present address: Victor G. Levitsky. Department of System Biology, Institute of Cytology and Genetics, Novosibirsk 630090, Russia.

## Abstract

Efficient *de novo* motif discovery from the results of wide-genome mapping of transcription factor binding sites (ChIP-seq) is dependent on the choice of background nucleotide sequences. The foreground sequences (ChIP-seq peaks) represent not only specific motifs of target transcription factors, but also the motifs overrepresented throughout the genome, such as simple sequence repeats. We performed a massive comparison of the 'synthetic' and 'genomic' approaches to generate background sequences for *de novo* motif discovery. The 'synthetic' approach shuffled nucleotides in peaks, while in the 'genomic' approach selected sequences from the reference genome randomly or only from gene promoters according to the fraction of A/T nucleotides in each sequence. We compiled the benchmark collections of ChIP-seq datasets for mouse, human and Arabidopsis, and performed *de novo* motif discovery. We showed that the genomic approach has both more robust detection of the known motifs of target transcription factors and more stringent exclusion of the simple sequence repeats as possible non-specific motifs. The advantage of the genomic approach over the synthetic approach was greater in plants compared to mammals. We developed the AntiNoise web service (https://denovosea.icgbio.ru/antinoise/) that implements a genomic approach to extract genomic background sequences for twelve eukaryotic genomes.

## Graphical abstract




**Received:** March 6, 2024. **Revised:** June 3, 2024. **Editorial Decision:** July 11, 2024. **Accepted:** July 19, 2024

## Introduction

Transcription factors (TFs) are proteins that control gene transcription by sequence-specific DNA binding. Chromatin immunoprecipitation (ChIP)-based high throughput technique ChIP-seq allows genome-scale mapping of TF binding sites (TFBS) (1,2). Peak calling is the primary processing of raw ChIP-seq data. It produces thousands of genomic loci or peaks enriched by protein binding (3,4). Each peak is usually at least several hundred base pairs long; it usually has a clear point of a number of maximum number of reads ('summit'), presumably correlating with positioning of motifs of target TFs (5,6). The term target TF refers to the TF studied in the ChIP-seq experiment. The most important task of secondary processing is to detect in the peaks specific motifs that are responsible for biological functions of the target TF. A motif represents the short, recurring pattern that is thought to have a biological function of sequence-specific binding sites for TF (7,6). Conventionally, a motif varies in length from 6 to 20 bp (8–10). The motif enrichment analysis, and above all, *de novo* motif discovery, is the required step to map exact positions of potential TFBS in peaks (3,11). An aim of motif discovery consists in exact definition of all parameters of a motif model.

Over the last few years, thousands of uniformly processed ChIP-seq datasets for hundreds of target TFs for major model species, including mammals, plants, insects, worms, and fungi have been collected in several focused databases (Cistrome DB, ReMap and GTRD, 12–14). Also, the massive application of *in vitro* technologies such as PBM (Protein Binding Microarray) and HT-SELEX (High-Throughput *in vitro* Selection) detected specific sequence motifs for several hundreds of TFs (15–18). As a result, the public databases JASPAR (19), Cis-BP (20) and HOCOMOCO (10) provided TF binding motifs from large-scale studies both *in vivo* and *in vitro*.

A series of studies proposed the hierarchical classification of mammalian TFs based on their DNA-binding domains (DBDs) (TFClass, 21–24). Thus, hundreds of TFs representing nine specific superclasses and several dozens of the most abundant classes were defined. Later, the TFClass framework has been applied to other eukaryotic taxa in the JASPAR database (19), and the focused study has detailed the results for plants (Plant-TFClass database (25)). This advance promotes bioinformatics analysis in plants, as no new TF superclasses have been identified at the highest level of the hierarchy in plants compared to the nine currently known superclasses in mammals (24). As previously expected (26), about half of all TF classes are common to mammals and plants.

However, enriched motifs identified by *de novo* motif search do not necessarily imply TF binding specificity. Although peaks are generally assumed to represent genomic loci with motifs of some TFs, *de novo* motif discovery tools can identify not only such motifs but also motifs enriched across the genome as a whole. These motifs are unlikely biologically relevant, and hereinafter they are referred to as non-specific. The most known examples of non-specific motifs are polyA tracts. Such motifs refer to simple sequence repeats (SSRs), short tandem repeats with a monomer length of 1–6 bp (27). All processing steps of ChIP-seq data analysis, including peak calling and *de novo* motif search require careful filtration of possible false positive non-specific motifs (28).

Long before the era of massive genome sequencing, the sizes of DNA sequence sets were very small and motif search tools took only one set of training sequences, putatively enriched with specific TFBS motifs (the generative learning principle (29)). Later, the principle of discriminant learning was proposed in much more advanced tools designed to analyze whole genomes (30–32). This principle took two sets of sequences, the first one still implied specific TFBS motifs (foreground set, peaks), while the second (background set) had to neutralize false enrichment of non-specific motifs from the foreground set. Therefore, the choice of background sequences is a key step required to estimate the significance of motifs enrichment in peaks (32,33,11). Formally, the specific and non-specific motifs are correct results of *de novo* motif discovery. Occasionally, due to missing or inadequate selection of background sequences, enrichment of non-specific motifs can compete with and even exceed enrichment of specific motifs.

So far, many studies have promoted the concept of the synthetic sequences as efficient for evaluation of the performance of motif finding in ChIP-seq data. This concept implied the generation of synthetic sequences by Markov chains of various orders (33–36), or these sequences were taken as a complete dictionary of $k$-mers, i.e. equal frequencies of nucleotides were presumed (6). Markov modeling of expected word frequencies has been very popular since the $k$-th order Markov chain captures compositional biases represented at the level of all words of lengths from 1 to $k + 1$. However, for different word lengths, spectra of word frequencies in different genomes showed behaviors, strikingly diverse from those expected in Markov modeling (37). In accordance with this, a benchmarking analysis of distinct *de novo* motif finding tools demonstrated that the synthetic approach gave too optimistic estimates of performances (32). This review suggested that background sets consisting of genomic sequences were needed to rigorously test hypotheses. This option implies randomly chosen genomic regions with basic properties of sequences (such as length, nucleotide composition, location in the genome, etc.) that match those from the foreground sequences (31,38–41).

Although several tools allowed the generation or at least application of either genomic or synthetic background sequences (38,40,42–45), none of them recommended one option as superior to the other. For example, in addition to the option of a custom user-defined set of background sequences, the popular *de novo* motif search tools, Homer (38) and STREME (43), offer the generation of genomic and synthetic background sequences, respectively. Therefore, the discrepancy in the overall results of various popular *de novo* motif search tools is related not only to the peculiarities of their algorithms, but also to different options for selecting background sequences. The application of the most reasonable approach for generating background sequences by each tool will certainly improve the quality of its results.

To date, only one study (39) has attempted to compare systematically the synthetic and genomic approaches generating background sequences for subsequent *de novo* motif discovery; a relatively small benchmark collection of 43 ChIP-seq datasets for several TFs was considered. Since last five years brought several specific databases focused on the uniform processing of ChIP-seq data (Cistrome DB, ReMap and GTRD) (12–14), a larger study comparing the results of different background sequence generation approaches is now possible. Nevertheless, no systematic large-scale studies comparing the sensitivity and specificity of applying genomic or synthetic background sequences for the *de novo* motif discovery in ChIP-seq data have been performed so far.

In the current study, we compiled the benchmark collections of ChIP-seq datasets for plants (*A. thaliana*) and mammals (*M. musculus* and *H. sapiens*). We applied these collections to compare the synthetic and genomic approaches generating the background sequences for *de novo* motif discovery. The 'synthetic' approach destroys the significant enrichment of any motifs through the shuffling of nucleotides preserving only the nucleotide composition. The 'genomic' approach selects sequences from the reference genome or certain its part randomly, using for each peak its A/T nucleotide content, thereby modeling the expected content of non-specific motifs. We aimed to clarify in massive tests which of two approaches was more sensitive in detecting the known motifs of target TFs, and simultaneously was more specific in restriction of non-specific motifs of SSRs as possible false positives.

For each ChIP-seq dataset, we generated the background sets with the genomic and synthetic approaches, and performed *de novo* motif discovery. We ranked the enriched motifs according to the significance of their enrichment. Then, we marked enriched motifs that were significantly similar to the known motifs of target TFs, as well as those for the motifs of SSRs. We concluded that the genomic approach, compared to the synthetic one, showed more reliable detection of the known motifs of target TFs and more rigorous exclusion of the SSR motifs as possible false positives. Finally, we developed the AntiNoise command-line software package and its web service. They applied the genomic approach to extract background sequences for the foreground sets from ChIP-seq data of several popular in massive analysis eukaryotic species from fungi to plants and mammals. The AntiNoise command-line software package and its web service provide the opportunity for fast extraction of the sets of background sequences that in subsequent *de novo* motif search, through careful estimation of motif enrichment, potentiates deeper insight into yet hidden mechanisms of gene transcription regulation.

## Materials and methods

### ChIP-seq data preparation

We extracted processed ChIP-seq data for *A. thaliana*, *M. musculus* and *H. sapiens* target TFs from GTRD (14). We selected ChIP-seq datasets that were preprocessed by the MACS2 peak caller (46) and had input control experiments in the primary processing pipeline. For *A. thaliana*, we extracted all available datasets, and for *M. musculus* and *H. sapiens*, we randomly selected only part of all available datasets to ensure large enough collection sizes (see Supplementary Tables S1–S3). The functionality of murine and human TFs were supported by their curated status (47). The high homology between human and murine TFs enabled this filtration (23). We used PlantRegMap (48) and TAIR (49) to validate Arabidopsis TFs. For each ChIP-seq dataset, we defined 1000 top-scoring peaks as the foreground set for subsequent analysis. Each background set consisted of 5000 sequences, i.e. five background sequences for each foreground sequence were required. Next, we applied 'genomic' and 'synthetic' approaches to generate background sets. The genomic approach chose sequences in the reference genome randomly: we allowed only the maximum deviation of 1% in the fraction of A/T nucleotides in each background sequence compared to the corresponding foreground one. The scheme in Figure 1 represents the algorithm of the genomic approach. Genomic sequence extraction requires either an unmasked or masked reference genome. The extraction of background sequences from the original unmasked version of the whole genome is designated as 'No masking'. A masked genome version can be prepared by the 'Exclusion of blacklisted regions' or 'Retention of whitelisted regions' options. The blacklisted option excludes certain genomic loci from the entire reference genome; the extraction procedure is applied to the remaining loci. The whitelisted option allows the extraction of background sequences only from certain specified regions; all other loci are excluded from the analysis. The synthetic approach performed the shuffling of nucleotides in each peak, exactly preserving its nucleotide content. Both approaches kept the lengths of sequences in the foreground and background sets unchanged.

We required for each ChIP-seq dataset that a target TF respected the known motifs (as position frequency matrices) in its class (or subfamilies only for mammalian C2H2 ZF TFs) from JASPAR (19), Hocomoco (10), or Cis-BP (20). In mammals, all TFs from the class C2H2 zinc finger factors were classified according to their subfamilies, due to the highest variability of motifs in this class (47,19). For each class or subfamily, we required the presence of at least one member motif possessing the significant enrichment ($P < 0.05$) in the foreground set compared to the corresponding background set, AME tool (50). Finally, human/murine/Arabidopsis benchmark ChIP-seq collections included 1032/706/119 datasets for 127/213/58 target TFs, respectively (see Supplementary Tables S1-S3).

### ChIP-seq data analysis pipeline

For each pair of foreground and background sets we performed *de novo* motif discovery with the traditional motif model of a position weight matrix (PWM), STREME tool, (43).We used the default parameters of the STREME tool, the motif length varied from 8 to 15 nt. For each ChIP-seq dataset, the input data of this tool included a foreground set consisting of peaks and a background set generated from peaks using a synthetic or genomic approach; the output data were a ranking of the top ten enriched motifs according to the significance of their enrichment. We estimated the sensitivity and specificity of *de novo* motif discovery for all pairs of foreground/background sets and all enriched motif as follows. To estimate the sensitivity for a particular enriched motif, we tested its significance of similarity ($P < 0.05$) to all known motifs of TFs from the same class/family/subfamily (JASPAR, Hocomoco and cis-BP) as that of the target TF (51). To estimate the specificity for the same enriched motif, we tested whether the similarity of this enriched motif to any of the SSRs motifs was significant. We proposed that the SSR motifs are false positive results of *de novo* motif search. We considered following SSRs motifs: mono-, di-, and trinucleotide repeats, i.e. two motifs of mononucleotide repeats $A_8$ and $G_8$, 10 and 32 motifs of di- and trinucleotide repeats, $(XY)_4$ and $(XYZ)_3$, respectively. We used the TomTom tool (51) to assess the significance of similarity between motifs.

We applied Fisher exact test to estimate the significance of the difference between the numbers of ChIP-seq datasets that had enriched motifs with certain ranks. First, we compared the number of datasets with enriched motifs with certain ranks between the genomic and synthetic approaches; these enriched motifs corresponded to motifs either of known TFs or SSRs (Table 1); we performed these tests separately for each collec-

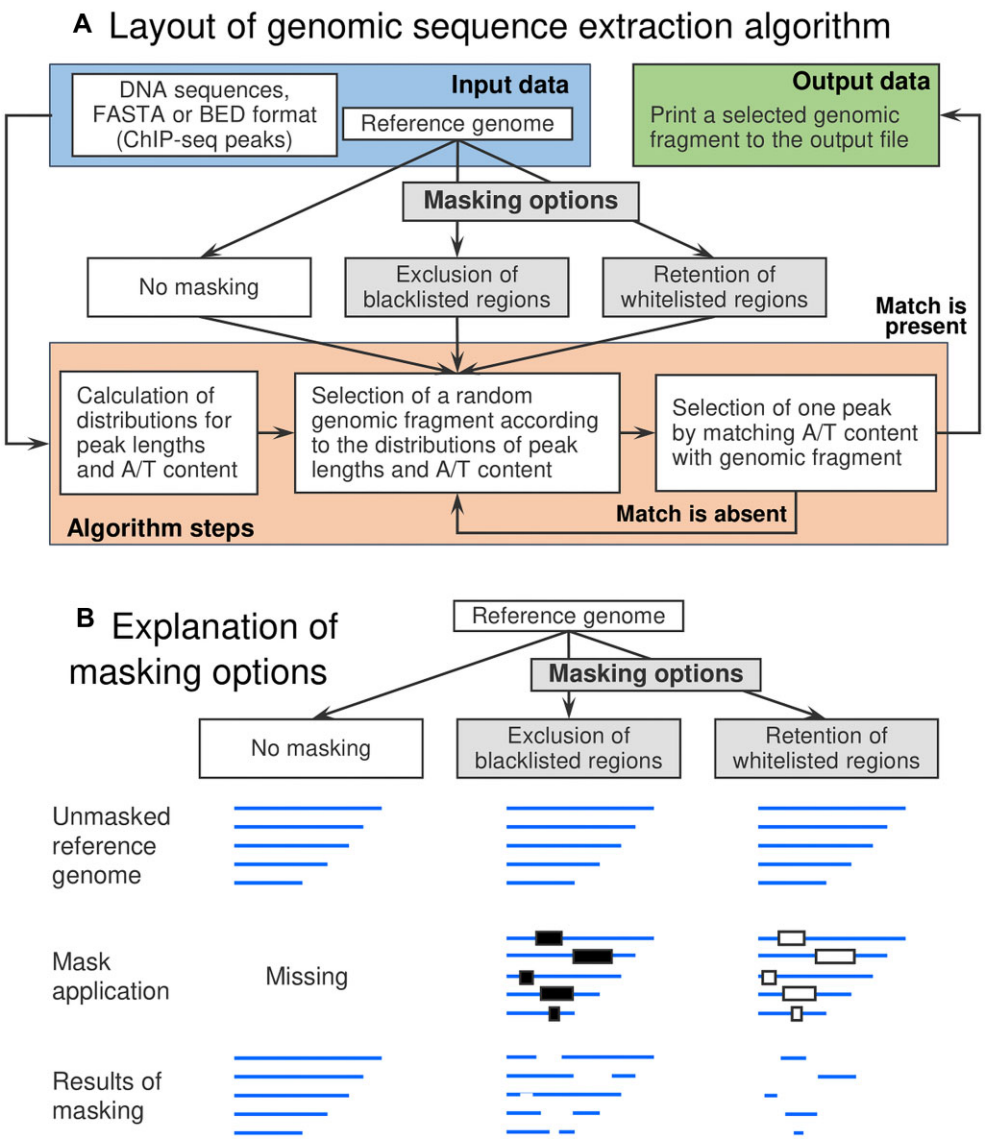



**Figure 1.** The algorithm of the genomic sequence extraction. (**A**) Layout of the algorithm. Blue/green color mark input/output data. First, input (foreground) sequences are used to compute the distributions of peak length and A/T content. Second, the reference genome is used to select proper background sequences. Hence, the output background sequences exactly match the foreground ones in length, while their A/T content allows only very small variation. (**B**) Scheme explaining two alternative masking options (grey). We either took the reference genome as is (the pathway 'No masking'), or apply 'Exclusion of blacklisted regions', removing from the reference genome particular regions, and using all remaining loci in subsequent analysis, or apply another 'Retention of whitelisted regions', preserving in the reference genome certain specific regions and removing all remaining loci.

tion. Second, we compared for each approach the ranking of enriched motifs of target TFs or SSRs motifs in *A. thaliana* and mammalian (*M. musculus*) benchmark collections (Table 2); we performed these tests separately for the genomic and synthetic approaches.

We used the hierarchical classification of TFs by their DBDs to study the relationship between the structure of DBDs of target TFs and efficacy of the genomic/synthetic background sequences application (TFClass, Plant-TFClass) (22–25). For plant TFs we also used annotations from PlantRegMap (48) and JASPAR (19). Supplementary Tables S4-S6 show for all classes of target TFs and three benchmark collections ranks of enriched motifs in the lists deduced by *de novo* motif discovery; these ranks are represented separately for the genomic and synthetic background dataset generation approaches. The third option 'promoter' means application of the whitelisted region option in the genomic approach (see Figure 1).

## Web service architecture

We proposed the web service AntiNoise promoting the genomic approach of background sequences extraction. The kernel of the web service was implemented in the C++ language. The kernel searched background sequences in the reference genome sequences (see above). The user interface (input/output data) and running of the kernel scripts were implemented in PHP language (version 7.4.3). In addition, Python code used matplotlib (52) and numpy (53) libraries to draw charts depicting distributions of the A/T content and dinucleotide frequencies.

**Table 1.** 2 × 2 contingency table 'number of ChIP-seq datasets' versus 'background set generation approach'

| | | Number of ChIP-seq datasets | |
|---|---|---|---|
| | | **Motifs have a rank or a range of ranks** | **Motifs do not have a rank or a range of ranks** |
| Background set generation approach | Genomic | $N_{G+}$ | $N_{G-}$ |
| | Synthetic | $N_{S+}$ | $N_{S-}$ |

The table counts ChIP-seq datasets with enriched motifs derived from *de novo* motif searches. These motifs have (+) or do not have (–) significant similarities to either known motifs of target TFs or SSR motifs. The table estimates the significant differences in ranking of enriched within the lists of top-ranked enriched motifs derived by *de novo* motif search for ChIP-seq datasets with application of the genomic and synthetic background sequences generation approaches. These tests were separately applied to the benchmark collections of *M. musculus* and *A. thaliana*.

**Table 2.** 2 × 2 contingency table 'number of ChIP-seq datasets' versus 'benchmark collection'

| | | Number of ChIP-seq datasets | |
|---|---|---|---|
| | | **Motifs have a rank or a range of ranks** | **Motifs do not have a rank or a range of ranks** |
| Benchmark collection | *M. musculus* | $N_{MM+}$ | $N_{MM-}$ |
| | *A. thaliana* | $N_{AT+}$ | $N_{AT-}$ |

The table counts ChIP-seq datasets with certain enriched motifs derived from *de novo* motif search. These motifs have (+) or do not have (–) significant similarities to either known motifs of target TFs or SSR motifs. Table estimates the significant differences in ranking of enriched motifs within the lists of top-ranked enriched motifs derived by *de novo* motif search for ChIP-seq datasets from the benchmark collections of *M. musculus* (MM) and *A. thaliana* (AT). These tests were separately applied for the genomic and synthetic background sequences generation approaches.

## Results

### Pipeline for ChIP-seq data analysis

We extracted ChIP-seq data for *M. musculus*, *H. sapiens* and *A. thaliana* target TFs from GTRD (14), we took each ChIP-seq dataset as the foreground set and generated for it background sets with the synthetic and genomic approaches (see Materials and methods). Figure 1 explains the layout of the algorithm applied for the genomic approach of background sequences selection. For each set of foreground sequences, we computed the enrichment of the known motifs of target TFs respective to either the genomic or synthetic set of background sequences (AME tool) (50); this left in analysis 1032/706 and 119 ChIP-seq datasets for *H. sapiens*, *M. musculus* and *A. thaliana* (see Materials and methods). Additionally, we applied the variation of the genomic approach. This option implied the preliminary masking of the whole genome (Figure 1B, retention of whitelisted regions), so that the background genomic sequences were extracted only from the promoter regions of protein coding genes, –5000/+1 relative to gene start positions. Below this option is designated as 'Promoters'. Alternatively, certain genomic loci can be excluded from analysis (Figure 1B, exclusion of blacklisted regions). The command line version of the tool enables application of arbitrary annotations in BED format as a set of whitelisted and blacklisted regions. The web version of the tool implemented the species-specific whole genome sets of promoters of protein coding genes as one of two default options (see below).

*De novo* motif discovery (STREME tool) (43) provided the top ten enriched motifs for each pair of foreground and background sets. Then, we estimated the significances of similarity between these enriched motifs and known motifs of TFs from the same class/family as a target TF. We applied at this step the JASPAR (19), *cis*-BP (20) and Hocomoco (mammals only) (10). Similarly, we estimated the significance of similarity between enriched motifs and the motifs of SSRs (see Materials and methods). Supplementary Tables S1–S3 for all datasets provide ID of ChIP-seq datasets from GTRD, target TF names, the enrichment p-value (AME), the ranks of enriched motifs, the significances of motifs similarity, and the descriptions of the respective motifs from JASPAR/*cis*-BP and Hocomoco. Finally, we compared the sensitivity and specificity between the genomic and synthetic approaches, as the ranking of the motifs respecting the known motifs of target TFs and the motifs of SSRs. We performed these comparisons for the benchmark collections for *M. musculus*, *H. sapiens* and *A. thaliana*. Supplementary Tables S1–S3 for these collections list the ranks of enriched motifs from *de novo* motif search that ensured the significant similarity to known motifs of target TFs. Supplementary Tables S4–S6 summarize the distributions of ranks for the classes of target TFs. Supplementary Tables S7–S9 compile the distributions of ranks of enriched motifs from *de novo* motif search that show the significant similarity to motifs of SSRs. All Supplementary Tables present three blocks of results for the synthetic, genomic approaches and for genomic approach restricted to gene promoters. The background sequences from whole genomes and promoters have shown very similar trend (compare columns 'Genomics' and 'Promoters' in Supplementary Tables S4–S9). Since the results for human and mouse are very similar, below in the manuscript we show the results for the murine collection due to the higher number of distinct target TFs in it compared to human collection (213 versus 127). Also, the term 'genomic' below means the genomic approach for the whole genome, the results for promoters are shown in Supplementary Tables.

### ChIP-seq data analysis pipeline

Figure 2A displays the analyses of two example ChIP-seq datasets for target TFs from *M. musculus* and *A. thaliana* (see row titles). For each dataset we applied the genomic and synthetic approaches (see column titles). Four plots show the significances of the first ten enriched motifs from the results of *de novo* motif discovery (axes Y) as a function of their A/T content (axes X). For both examples the genomic approach assigned the first ranks to the known motifs of target TFs, but the synthetic approach displayed SSR motifs at the first ranks, so that motifs of target TFs had lower enrichments and subsequent ranks. These examples suggested that the genomic approach showed both better sensitivity and specificity than the synthetic approach.

Next, we analyzed the benchmark collections of ChIP-seq data for *M. musculus* and *A. thaliana*. Figure 2B compare the distributions of ranks for *A. thaliana* and *M. musculus*. The distributions of ranks of the motifs of target TFs for these collections confirmed the conclusions derived from the analysis of two examples. The genomic approach retained motifs significantly similar to known motifs of target TFs at the first ranks in 69 and 579 datasets, correspondingly, out of total 119

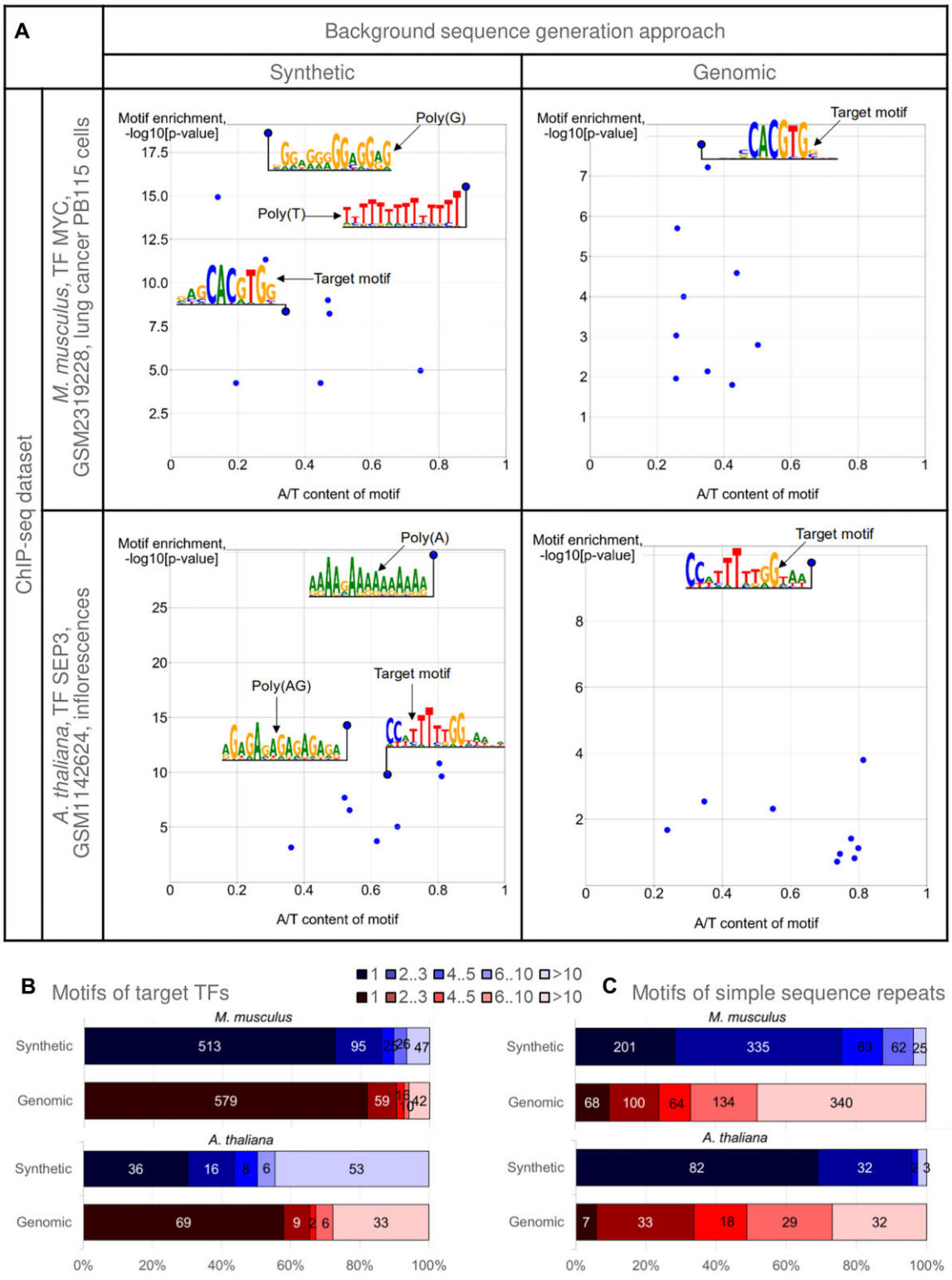



**Figure 2.** Abundances of motifs of target TFs and motifs of SSRs revealed by *de novo* motif discovery with application of the synthetic and genomic background approaches. (**A**) The ranking of motifs of target TFs and SSRs in the results of *de novo* motif discovery for two example ChIP-seq datasets. Columns show the results of *de novo* motif search for the synthetic and genomic background datasets. Rows represent the analysis of *M. musculus* and *A. thaliana* ChIP-seq datasets (TF MYC, GTRD PEAKS040291, GEO GSM2319228, lung cancer PB115 cells; TF SEP3, GTRD PEAKS042815, GEO GSM1142624, inflorescences). Axes X and Y imply A/T content of motifs and the significance of motif enrichment from the STREME tool, $-\text{Log}_{10}$[p-value]. This enrichment reflects the rank of a motif in the result of *de novo* motif search. Arrows and logos mark motifs of target TFs and motifs of SSRs. (**B**, **C**) The distributions of ranks of enriched motifs that are significantly similar to either the known motifs of target TFs (B) or SSRs (C). The distributions derived from the results of *de novo* motif discovery with application of the synthetic or genomic approaches for the benchmark collections of *M. musculus* and *A. thaliana* ChIP-seq data. Axes X mark the number of datasets possessing certain ranks of enriched motifs in the lists from *de novo* search; axes Y imply the synthetic and genomic background approaches. Red/blue colors mark the distributions computed for genomic/synthetic background sets; the darker/lighter shades of colors show the numbers of datasets possessing motifs with higher/lower ranks.

and 706 datasets. For the synthetic approach, the corresponding numbers 36 and 513 were notably smaller. Respective distributions for SSR motifs showed the opposite trend (Figure 2C). The genomic approach revealed notably fewer datasets with SSR motifs at the first ranks (7/68 for *A. thaliana*/*M. musculus*) compared to the synthetic approach (82/201).

Application of Fisher exact test (Table 1) confirmed that the fractions of motifs corresponding to target TFs and ranked first were significantly higher for the genomic approach than for the synthetic approach (for the collections of *M. musculus* and *A. thaliana* ($P < 4E\text{-}5$ and $P < 3E\text{-}5$, Figure 3A). The subsequent ranks showed the same trend, although it gradually became insignificant. The first-ranked enriched SSR motifs occurred significantly less frequently in lists derived from the genomic approach than from the synthetic approach ($P < 9E\text{-}20$ and $P < 9E\text{-}26$, Figure 3B). The subsequent ranks showed the similar and sometimes even more significant depletion.

Next, we applied Fisher exact test separately for each of the background generation approaches to compare *M. musculus* and *A. thaliana* collections (Table 2). We revealed that the fractions of datasets possessing the enriched motifs respecting target TFs are significantly higher for the *M. musculus* collection compared to those for the *A. thaliana* collection (the first rank, genomic $P < 5E\text{-}8$, synthetic $P < 3E\text{-}18$, Figure 4A). This enrichment may reflect notably higher variation of mammalian motifs compared to plant motifs (10,19,20). The comparison between the fractions of datasets containing enriched SSR motifs (Figure 4B) showed that they were significantly more abundant in *A. thaliana* than in *M. musculus*. For the highest ranks of enriched motifs (first and from first to third) this trend was significant only for the synthetic approach ($P < 1E\text{-}16$ and $P < 5E\text{-}8$, correspondingly). The genomic approach rejected the significance of this trend for the enriched SSR motifs of the first rank, and for the range of ranks from first to third (Figure 4B). Thus, for the synthetic approach, *de novo* motif search significantly higher yields top-scored enriched motifs of SSRs much more frequently in the *A. thaliana* ChIP-seq data collection than in the *M. musculus* collection.

Overall, we found that the genomic approach of background sequences generation provided both better sensitivity and specificity for benchmark ChIP-seq data collections from mammals and plants. As expected, we confirmed that the synthetic and genomic approaches guaranteed the significant enrichment of motifs for target TFs in *de novo* motif discovery results for all ChIP-seq data collections. In contrast to the genomic approach, the synthetic approach assigned the highest ranks to the enriched SSR motifs substantially more frequently in the *A. thaliana* ChIP-seq data collection compared to the *M. musculus*. Hence, *de novo* motif discovery in plant ChIP-seq data requires more careful processing taking into account possible false positives of SSR motifs enrichment.

## Genomic approach shows better sensitivity for target TFs of almost all classes

Next, we tested whether the difference between the results of applying the genomic and synthetic approaches depended on the structure of DBDs of target TFs. We used annotations of mammalian and plant target TFs from TFClass, PlantTFClass, JASPAR, *cis*-BP and Hocomoco databases, and applied their hierarchical classification, see Materials and methods, Supplementary Tables S1–S3 (21–25). Figure 5 for the most abundant superclasses and classes of target TFs of *M. musculus* and *A. thaliana* shows the distributions of the number of datasets, with enriched motifs, that are significantly similar to known motifs of target TFs with certain ranks in the lists of motifs obtained using the genomic and synthetic approaches. Supplementary Tables S4–S6 present the results of corresponding analyses for all classes. Overall, for almost all the most abundant classes of *M. musculus* and *A. thaliana* target TFs (12 of 13), the genomic approach is more sensitive than the synthetic approach. Only for the class of C2H2 zinc finger factors in *M. musculus* synthetic/genomic approaches shows the first ranks for 124/122 datasets out of 137. Since these numbers are almost equal, one exception still confirmed the general trend. Moreover, in the human collection 228/222 datasets of total 244 for the same class C2H2 zinc finger factors have the first ranks, thereby this class is not an exception to the general trend.

## AntiNoise web service and command line software package: input/output data and functionality

For a wider application of the genomic approach in further researches, we implemented it in the AntiNoise command line software package and web service. The command line software package is available at https://github.com/parthian-sterlet/antinoise. The package allows application of genomic and synthetic approaches. Additionally, the package provides perl scripts starting from the reference genome sequences in one FASTA file. There are three options for genomic sequence extraction. First, 'No masking', defines the entire reference genome as a source of background sequences. Second, 'Blacklisted region masking' means that given 'blacklisted' regions are excluded during the search. We mask these blacklisted regions, thereby they completely excluded from output data. E.g., a recent study provides examples for mouse and human (54). Third, 'Retention of whitelisted regions' implies that background sequences are restricted to only 'whitelisted' regions, and all remaining genomic loci were masked and they couldn't get into the output. We propose the promoter regions of all protein-coding genes (–5000; +100) as a default option of the whitelisted regions for all species.

The web service is available at https://denovosea.icgbio.ru/antinoise/. Figure 6A displays the main and advanced options of the web service:

- Genome release and species, among them the animals human (*H. sapiens*, hg38), mouse (*M. musculus*, mm10), rat (*R. norvegicus*, Rnor_6.0), zebrafish (*Danio rerio*, GRCz11), fly (*D. melanogaster*, dm6) and roundworm (*C. elegans*, WBcel235); the plants are arabidopsis (*A. thaliana* TAIR10), soybean (*Glycine max* v2.1), maize (*Zea mays*, B73) and liverwort (*Marchantia polymorpha*, MpTak v6.1); the fungi baker's yeast (*S. cerevisiae*, R64-1-1) and fission yeast (*S. pombe*, ASM294v2);
- Required number $R_{BF}$ of background sequences per one foreground sequence. The default value $R_{BF} = 5$ implies that that if calculations are successively completed for all $N_F$ input sequences, then totally $N_B = R_{BF} * N_F$ background sequences are found;
- Either 'no masking', or 'Retention of whitelisted regions' options are applied. The 'whitelisted regions' are the promoter regions of all protein-coding genes, (–5000; +100) relative to the 5′ ends of genes.

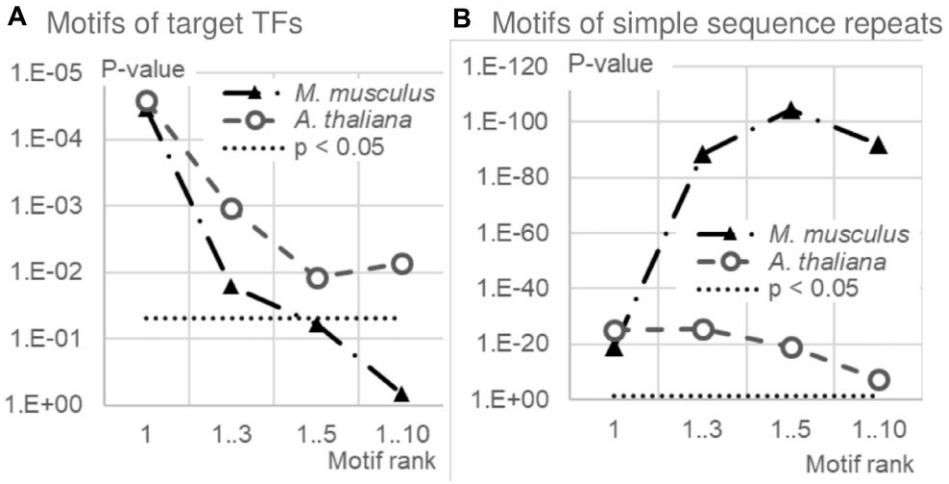



**Figure 3.** The significance of difference in ranking of the enriched motifs between the synthetic and genomic background approaches. *De novo* motif search revealed the ranks of enriched motifs. The Fisher exact test separately for the *M. musculus* / *A. thaliana* collections compared the fractions of datasets possessing enriched motifs with certain ranks. Panels (**A**, **B**) display the analysis for the enriched motifs, significantly similar to the motifs of target TFs and the motifs of SSRs. Axes X and Y show the motif rank and the significance by exact Fisher test, respectively; dotted lines mean the significance threshold ($P < 0.05$). Table 1 explains the Fisher task applied in analysis.

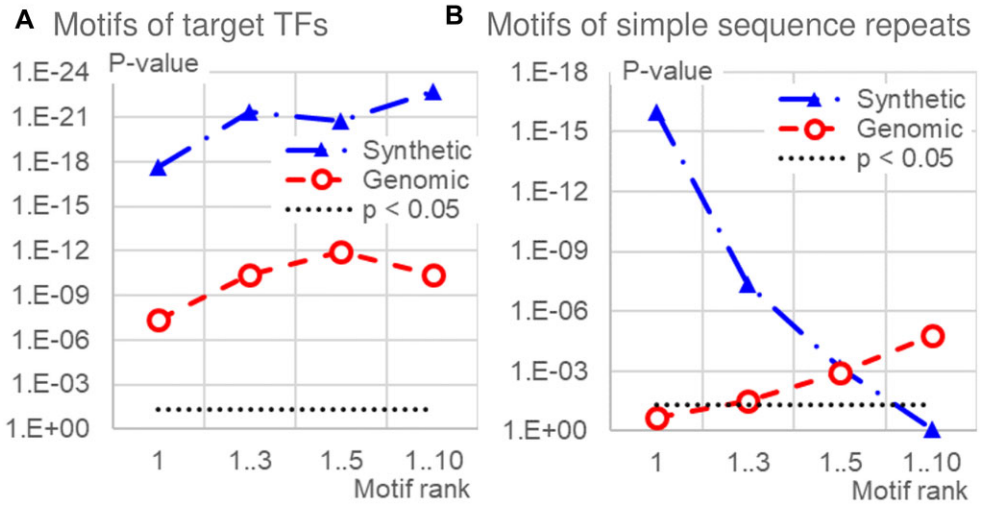



**Figure 4.** The significance of difference in ranking of the enriched motifs between the *M. musculus* and *A. thaliana* benchmark collections. *De novo* motif search revealed the ranks of enriched motifs. The Fisher exact test separately for the synthetic and genomic background approaches compared the fractions of datasets possessing enriched motifs with certain ranks. Panels (**A**) and (**B**) display the analysis for the enriched motifs, significantly similar to the motifs of target TFs and the motifs of SSRs. Axes X and Y show the motif rank and the significance by Fisher exact test, respectively; dotted lines mean the significance threshold ($P < 0.05$). Table 2 explains the Fisher task applied in analysis.

- Deviation δ of the A/T nucleotide content of each background sequence from that for the corresponding foreground sequence. The default value 0.01 allows the mismatch of one bp per a sequence length of 100 bp in a foreground sequence.
- Threshold $F_{MIN}$ for the minimum fraction of completely processed foreground sequences to stop calculations. The default value of 0.99 means that calculations stop if for 99% of all foreground sequences for each sequence at least $R_{BF}$ background sequences are found for each sequences;
- Maximal number of attempts $N_A$ to find matching background sequences in the genome. If a given number $N_A$ of last attempts to find any at least one more background sequence are unsuccessful, the algorithm terminates. The default value 50000.

Figure 6A shows the screenshot for the application page of the web service with an example. Here we analyzed the dataset of 1000 top-scored ChIP-seq peaks for *M. musculus* cortical neurons treated with 10nM purified recombinant Reelin for 1 hour, TF MEF2C, GEO GSM1629389 (55), GTRD ID PEAKS037873.

Pressing the 'SUBMIT' button starts the calculation and provides a web link to the page with its results. During the running process, the service indicates the number of found background sequences. The output data of the web service include a link to a file with the main calculation result, a set of genomic background sequences in FASTA format. These results respect the first tab 'Results/Input parameters' (Figure 6B). In addition, four charts on two tabs illustrate the validity of the background sequences search. The second tab 'Foreground set vs. Background set' shows two charts depicting

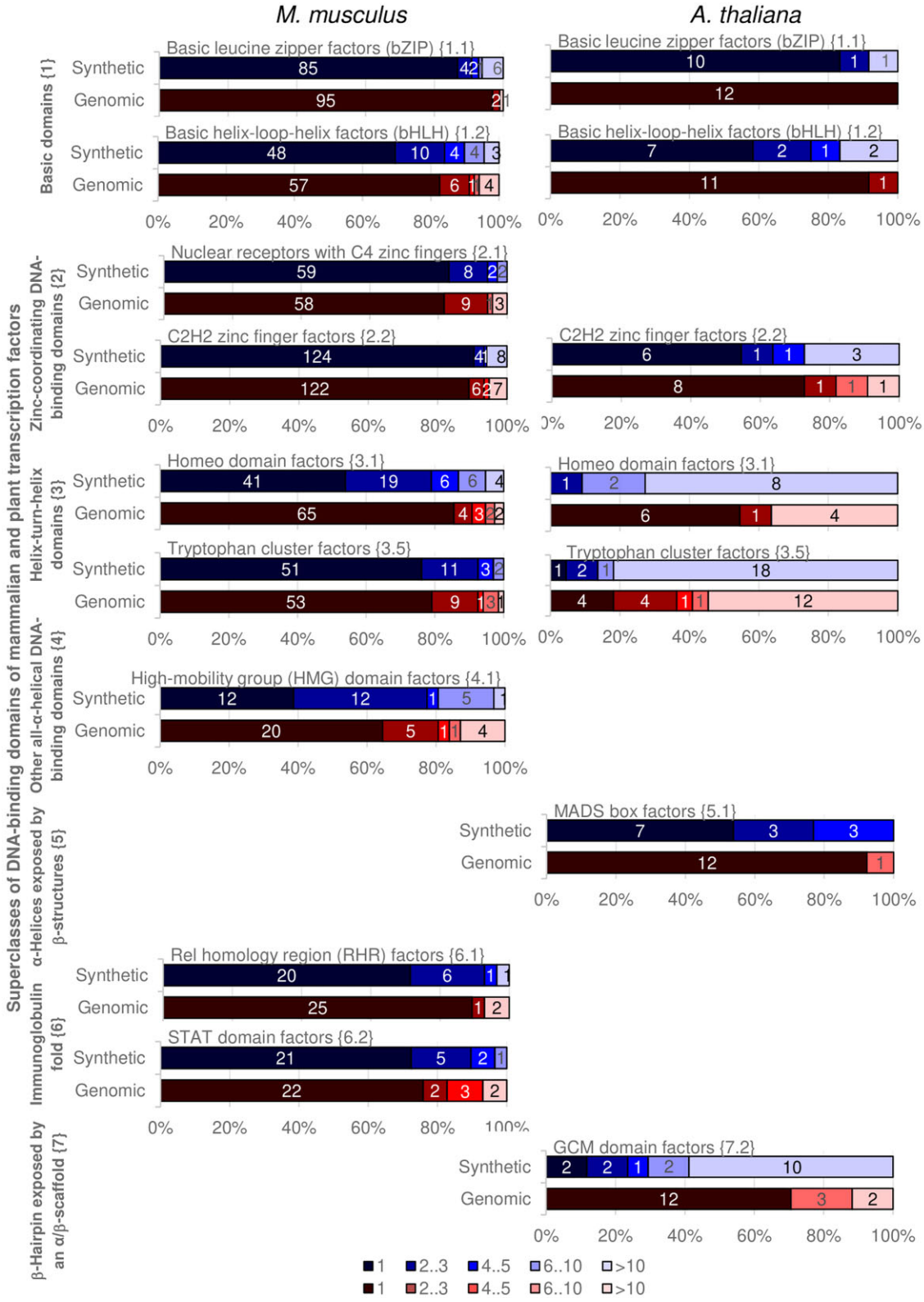



**Figure 5.** Application of the genomic and synthetic approaches for target TFs of various DBD structures. Left/right columns present analysis of ChIP-seq datasets from the *M. musculus* and *A. thaliana* collections. The hierarchical classifications of *M. musculus* and *A. thaliana* target TFs by the structure of DBDs were derived from TFclass and Plant-TFclass (22–25), see Materials and methods. Axes X mark the number of datasets possessing certain ranks of enriched motifs in the lists from *de novo* search; axes Y imply the application of the synthetic and genomic background approaches for ChIP-seq data with target TFs from various classes. The titles on the left stand for the names of nine TF superclasses from TFclass and Plant-TFclass (22–25). The ranks for TFs were assigned by the enrichment significance of motifs from *de novo* motif discovery for ChIP-seq datasets for the *M. musculus* and *A. thaliana* benchmark collections. Only the most abundant classes were represented. See Supplementary Tables S4-S6 for the respective ranking of enriched motifs for all TF classes in *H. sapiens*/*M. musculus*/*A. thaliana*.


**Figure 6.** Main screenshots of the AntiNoise web application. (**A**) Application page with an example. Here user can enter input sequences, set the parameters of a calculation task and start calculations. (**B**) The 'Results / Input parameters' output tab provides access to input (foreground) sequences and output (background) sequences, and lists major options of a calculation task. ChIP-seq dataset for *M. musculus* TF MEF2A, GEO GSM1629389, GTRD ID PEAKS037873 is used as an example.

the distributions of the A/T content and of the dinucleotide frequencies for the foreground and background sets (Figure 7A). These charts can be critical for choosing between various types of background sets. The third tab 'Selected foreground sequences versus Foreground set' demonstrate two charts listing foreground sequences that did not reach the required number of background sequences per one foreground sequence. These two charts apply the metrics of the A/T content and dinucleotide frequencies to compare the selected foreground sequence with the whole set of foreground sequences (Figure 7B). While the A/T content means the metrics applied for the genomic background sequences selection, the dinucleotide frequencies as the *k*-mers of the shortest length 2 bp show the behavior of the simplest motifs. These two charts detect foreground sequences with abnormal mono- and dinucleotide content.

The runtime for the command line version was estimated on a desktop PC. For the *A. thaliana*/*M. musculus* benchmark collections of ChIP-seq datasets (top 1000 peaks in each) the median runtime is about 3–4 minutes, and 95% of tested datasets were ready in 35 minutes/1 hour, correspondingly.

## Discussion

In the current study, we performed the massive comparison of two popular approaches to generating background sequences for subsequent *de novo* motif search in ChIP-seq data. The synthetic approach performs nucleotides shuffling that abolishes the enrichment of any motifs. This procedure radically destroys in the foreground sequences the enrichment of *k*-mers of any length. These *k*-mers represent either specific or non-specific motifs; they compete between each other at the next step of *de novo* motifs search. Applying the synthetic background approach inevitably results in a lower frequency of non-specific and specific motifs in the background sequence compared to the foreground sequence. In this case, the enrich-

**A**

Results / Input data & parameters

Foreground set vs. Background set

Selected foreground sequences vs. Foreground set

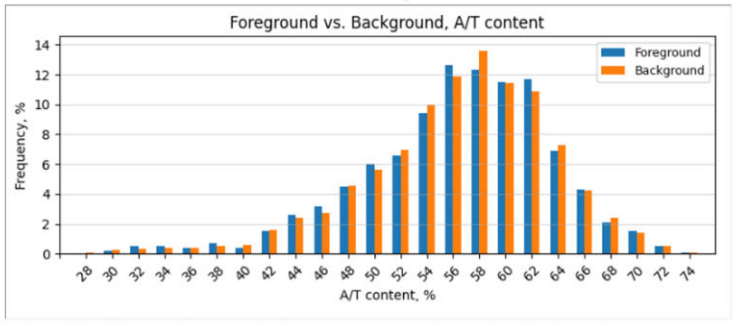



Chart 1. Distributions of the A/T nucleotide content for foreground sequences (input set) and background sequences (output set). Axis X marks the fraction of A/T nucleotides, axis Y shows the fraction of sequences (see manual)

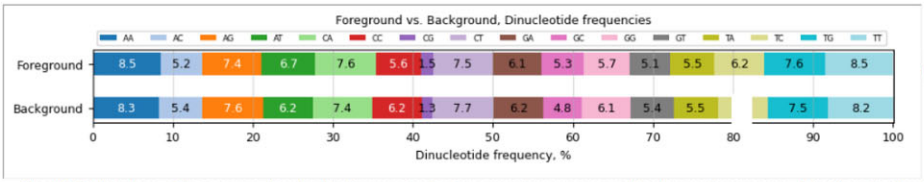



Chart 2. Distributions of the average dinucleotide frequencies for foreground sequences and background sequences. Axis X marks dinucleotide frequencies and axis Y means the sequence set (see manual))

**B**

Results / Input data & parameters

Foreground set vs. Background set

Selected foreground sequences vs. Foreground set

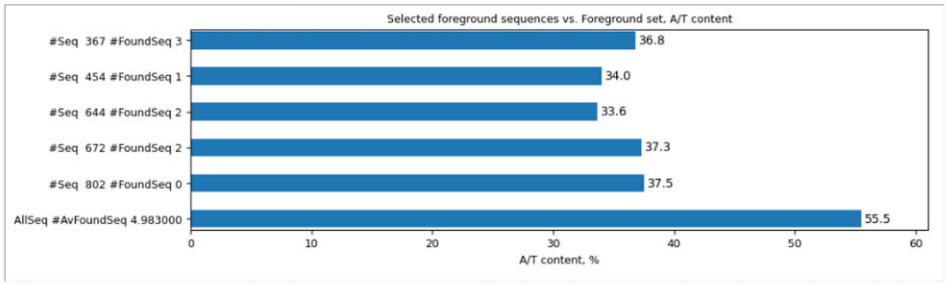



Chart 3. A/T content for selected foreground sequences and the average one for the foreground set. Axis X shows the A/T content, the Y axis lists the foreground sequences that fall short of the required number of background sequences per foreground sequence, the labels "#Seq NNN #FoundSeq K" mark for these sequences serial numbers in input file (NNN) and shows the counts of background sequences found (K). The label "AllSeq #AvFoundSeq K" implies the averaging for all foreground sequences, the value K means the average number of background sequences per one foreground sequence. (see manual)

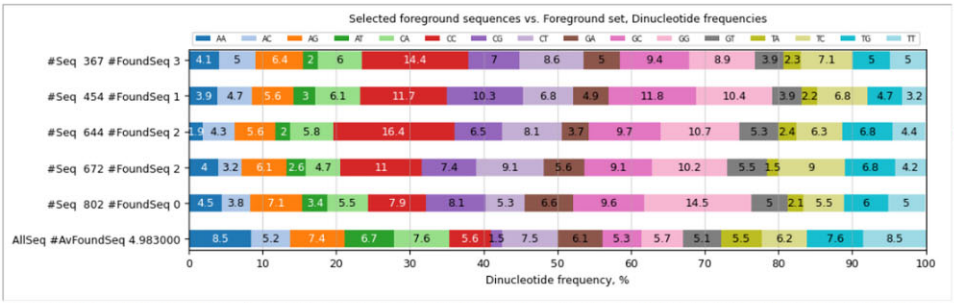



Chart 4. Dinucleotide frequencies for selected foreground sequences and the average ones for the foreground set. Axis X shows dinucleotide frequencies, the Y axis lists the foreground sequences that fall short of the required number of background sequences per foreground sequence, the labels "#Seq NNN #FoundSeq K" mark for these sequences serial numbers in input file (NNN) and shows the counts of background sequences found (K). The label "AllSeq #AvFoundSeq K" implies the averaging for all foreground sequences, the value K means the average number of background sequences per one foreground sequence. (see manual)

**Figure 7.** Additional output screenshots of the AntiNoise web application. (**A**) The 'Foreground set vs. Background set' output tab compares the input and output sequences. One chart depicts the A/T content and another shows dinucleotide frequencies. (**B**) The 'Selected background sequences vs. Foreground' output tab compares input sequences that did not reach the required threshold of found genomic background sequences and all input sequences. the Y axis lists foreground sequences that did not reach the required number of background sequences per one foreground sequence, the labels '#Seq NNN #FoundSeq K' mark for these sequences serial numbers in input file (NNN) and shows the counts of background sequences found (K). The label 'AllSeq #AvFoundSeq K' implies the averaging for all foreground sequences, the value K means the average number of background sequences per one foreground sequence. Axes X show the A/T-content (top) and dinucleotide frequencies (bottom). The ChIP-seq dataset for *M. musculus* TF MEF2A, GEO GSM1629389, GTRD ID PEAKS037873 is used as an example.

ment of non-specific motifs in the foreground set is not due to binding of specific TFs, but rather corresponds to the overall specificity of the full-length genome in terms of oligonucleotide composition. Attempting to suppress these noisy genomic bias motifs in the foreground set, one should try to preserve their content in the background set, while destroying the enrichment of context-specific motifs presumed to have TF binding functionality. Hence, the genomic approach implies the extraction of the background sequences from the reference genome sequences.

Before the beginning of the next-generation sequencing era, both TFBS and genomic mapping data were scarce, and, conventionally, motif discovery algorithms applied the synthetic sequences to model the expected frequencies of motifs (34). However, application of these algorithms for newly available large datasets derived from ChIP-seq technology (56) raised many unanswered issues, including long computation time, too redundant output data, variation in a threshold for the peak quality, filtration of genome bias motifs, etc. (6,9,28,32). In addition, the chromatin as primary source of ChIP-seq data often complicated the detection of motifs of target TFs due to their possible indirect binding through intermediate proteins, co-binding with partner TFs, and structural heterogeneity within the same TFs (57,58). Thus, even a variety of distinct algorithmic strategies gave only a limited success in higher eukaryotes (3,9,36,43,59).

Earlier, it was found that the distributions of the relative abundance of short oligonucleotides were strikingly diverse among DNA sequences from the genomes of various eukaryotic taxa, and these distributions were certainly different from those expected from Markov modeling (60). Namely, the di-/tri-nucleotide frequencies in genomic sequences differed markedly from those expected by their mono-/di-nucleotide content, etc. Authors explained these differences through distinct structural properties of short *k*-mers, in particular, the base-step stacking capacities, duplex curvature and other higher order DNA structural features of dinucleotides (61). It was concluded that the genome-wide consistency of dinucleotide relative abundance values suggested involvement of the fundamental biological processes, such as DNA replication, recombination and repair. Later, the analysis of the whole chromosomes of various species from distant eukaryotic taxa confirmed that the deviations of dinucleotide frequencies from those expected according to their nucleotide content were distinctly genome-specific (62). These studies showed that the sequence bias in whole genomes implied the species-specific pattern of enriched and depleted oligonucleotides of various lengths, including lengths typical for TFBS motifs (6–20 bp). The synthetic approach applies Markov models with various orders of a chain, which may be ranged from 0 to 5 (63,64). Therefore, even the largest order of a Markov chain can ensure the preservation of *k*-mers of short lengths, up to hexamers. However, the occurrence in the foreground sequences of *k*-mers of longer lengths, particularly those with lengths as long as the TFBS motifs (6–20 bp), can be wrongly regarded as enrichment. Hence, the synthetic approach can create an artificial enrichment of non-specific motifs in the foreground set. Thus, because the genomic approach reflects the genome-specific bias in oligonucleotide frequencies it should be superior to the synthetic approach.

The synthetic approach has been very popular in motif discovery tools to generate the background sequences (9,65,66). Some modern tools do not allow to adopt the genomic background sequences for *de novo* motif discovery (64,67–69), while others allow both synthetic or genomic background sequences (38,43,45,70). Alternative suggestion for *de novo* motif discovery compiled genome sequences flanking peaks to the background set (72,72). However, even the peak calling tool selection and its options influence on the precise positioning of peaks borders. E.g., the peak callers GEM/MACS2 provide peaks of fixed/varied lengths (71,46). Therefore, we suspect that the compilation of background sequences from areas flanking the peaks is not a completely correct methodology. Among the current sources of TFBS motifs derived from ChIP-seq data, three are the most reliable and popular, the HOCOMOCO (10), CisBP (20) and JASPAR (19) databases. Among them, only the CIS-BP was developed with the support from the various types of genomic background sequences (73).

In the current study, we took in analysis ChIP-seq data from the GTRD, since this database combined uniformly processed chromatin immunoprecipitation data with detailed annotations, such as the application of the input control experiment in the processing pipeline, descriptions of tissue/cell/treatment conditions, application of various peak caller tools, etc. We retained in the analysis only ChIP-seq datasets with enrichment of motifs of known TFs from the same hierarchical clade as the target TFs. Here, we used the AME tool; we took the clades of subfamilies for TFs from the mammalian C2H2 ZF class, and the clades of classes for the remaining TFs. Motifs of known TFs were taken from JASPAR, Cis-BP or Hocomoco. Thus, popular motif databases supported the context specificity of target TFs binding. We treated enriched SSR motifs as potential false positives. For each ChIP-seq dataset, we considered the top ten most enriched motifs from results of *de novo* motif discovery (STREME). This choice was due to the inability of ChIP-seq technology to distinguish between direct and indirect TF-DNA interactions (1,2,59), and motifs of target TFs could have lower ranks in the output lists. In confirmation of this, many massive applications of *de novo* motif discovery to ChIP-seq data showed that about a half of datasets did not reveal the known motifs for target TFs as the first-ranked enriched motifs (41,74–76). A rank of each enriched motif allowed estimating the sensitivity and specificity of each of the background generation approaches, through estimates of the significance of the similarity of this enriched motif to known motifs of target TFs or SSR motifs, respectively.

To select sequences into the background set, we applied two metrics of genomic DNA loci. The first is A/T content, the variation of which also accurately reflects the variation of G/C content, reflecting the ratio of the content of relatively strong G/C and weak A/T pairs of nucleotides with three and two hydrogen bonds in DNA. For example, it is difficult to find any G/C-reach motifs of the KLF/SP1 family of mammalian TFs in A/T-rich genomic loci. Accurate retention of the second metric, the sequence length, is required to support popular measures of accuracy for *de novo* motif search, such as the Precision-Recall curve or Area Under Curve (ENCODE-DREAM *in vivo* TF Binding Site Prediction Challenge, (77)).

Example analysis of two ChIP-seq datasets (Figure 2A) showed that the use of genomic background sequences compared to synthetic ones provided higher enrichment of known motifs of target TFs and lower enrichment of SSR motifs. We proposed that these differences were due to the complete destruction of the enrichment of any motifs by nucleotide shuffling as the generation procedure of synthetic sequences, while

the procedure of the genomic approach accounted the expected content of genome bias motifs in ChIP-seq data.

The systematic analysis of the *M. musculus*, *H. sapiens* and *A. thaliana* benchmark collections of ChIP-seq data confirmed these conclusions (Figure 2B, C, Supplementary Tables S4-S9). We came to two concordant conclusions. First, known motifs of target TFs showed significantly higher ranks in the results of *de novo* motif discovery for the genomic approach compared to those for the synthetic approach (Figure 3A). Second, SSR motifs demonstrated significantly lower ranks in the results of *de novo* motif discovery for the genomic approach compared to those for the synthetic approach (Figure 3B).

Next, we compared the results of separate applications of either genomic or synthetic approach between ChIP-seq data collections for *M. musculus* and *A. thaliana*. Both collections revealed the significant enrichment of known motifs of target TFs for the *A. thaliana* collection, although the synthetic approach showed higher significance (Figure 4A). Surprisingly, the synthetic approach provided substantially higher enrichment of SSR motifs for the *A. thaliana* collection compared to the *M. musculus* collection (motif rank 1, 1–3, Figure 4B). The genomic approach did not show the significant enrichment for the same ranks of the enriched motifs. Hence, whereas both approaches are sensitive due to enrichment of the known motifs of target TFs, the specificity appreciated through the abundances of potentially false positive motifs of SSRs is substantially worse for *A. thaliana* ChIP-seq data over those for *M. musculus*.

Finally, we considered the hierarchical classification of target TFs from ChIP-seq data of mammalian and plant TFs by their structure of DBDs (22–25,19). We showed that the higher enrichment of the known motifs of target TFs in the results from the genomic approach compared to those from the synthetic one is observed for almost all most abundant classes of murine or Arabidopsis TFs (Figure 5).

A necessary step of the processing of massive sequencing data such as ChIP-seq is the assessment of motif enrichment reflecting the binding specificity of target TFs. The choice of a particular approach has been essential for both specific tools generating background sequences (44,65), and for *de novo* motif search tools (38,67,78), and for the special databases of TFBS motifs (10,19,20). However, until now, no one has performed a massive analysis of ChIP-seq data for various eukaryotic taxa to compare two the most popular approaches - the generation of background sequences by the synthetic or genomic approaches. Since our study resulted in very strong arguments in favor of the genomic approach (Figures 2–5), we implemented it as the command line software package and the web service (Figures 6,7). They allow extract background sequences for the most popular in massive sequencing analysis eukaryotic genomes from yeasts to mammals and plants. Thus, we propose a flexible approach to robustly support the identification of specific TF targeting motifs in widely used *de novo* motif search tools from massive sequencing data.

A recent comprehensive all-against-all TF binding motif benchmarking study (79) showed that TF binding specificity correlates with the structural class of its DBD. Thus, different PWMs for TFs from the same structural classes tended to perform similarly across experiments; and the best performing PWM model often respected the same TF class. Another conclusion from the benchmarking study (79) stated that the practical set of motifs for many biological applications is much smaller than the number of motifs already contained in the most popular motifs collections. Therefore, we hope that the main conclusions of our study concerning the advantages of the genomic background sequences over synthetic sequences hold true for target TFs from a variety of eukaryotic taxa and for target TFs of any structural class.

We performed a *de novo* motif search for benchmark collections of ChIP-seq data for target TFs from mouse, human and Arabidopsis. We aimed to investigate whether background sets consisting of genomic or synthetic sequences provide both a higher prediction rate of known motifs of target TFs and a lower prediction rate of potentially false positive SSR motifs. We found that although both genomic and synthetic approaches provide pronounced enrichment of known target TF motifs, the synthetic approach compared to the genomic approach yields a very significant increase in the proportion of SSR motifs representing possible false positives. We confirmed the advantage of the genomic approach over the synthetic approach in terms of sensitivity of detection of motifs of known target TFs for all the most common classes of target TFs in mammals and plants. As for specificity, the use of the synthetic approach compared to the genomic approach resulted in higher ranks of enriched SSR motifs for plant than for mammalian ChIP-seq data. To summarize, massive analysis of mammalian and plant ChIP-seq data has shown that the genomic approach is more effective than the synthetic approach in generating background sequences for *de novo* motif discovery. Therefore, we propose to use genomic sequences extracting as a default option for generating a set of background sequences when applying *de novo* motif search to ChIP-seq data. To promote and widely apply the results of our analysis, we implemented the genomic approach of background sequences generation as the AntiNoise web service and provided its extended options in the command line version.

## Data availability

AntiNoise is implemented in C++, Python and PHP. The source code and documentation are available at https://github.com/parthian-sterlet/antinoise (permanent doi https://doi.org/10.5281/zenodo.12744549) and https://denovosea.icgbio.ru/antinoise/. Data to reproduce the results of this study are available at http://gtrd.biouml.org/#! and https://plamorph.sysbio.ru/ciscross/AboutCisCross.html.

## Supplementary data

Supplementary Data are available at NARGAB Online.

## Acknowledgements

The bioinformatics data analysis was performed in part on the equipment of the Bioinformatics Shared Access Center within the framework of State Assignment Kurchatov Genomic Center of ICG SB RAS [FWNR-2022–0020].

*Author contributions*: V.V.R.: Formal analysis, Visualization. A.V.T. Formal analysis. A.G.B.: Software, Writing—review & editing. V.G.L.: Conceptualization, Investigation, Methodology, Validation, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing.

## Funding

Russian Science Foundation project [20-14-00140].

## Conflict of interest statement

None declared.

## References


1. Nakato,R. and Shirahige,K. (2017) Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief. Bioinform.*, **18**, 279–290.
2. Lloyd,S.M. and Bao,X. (2019) Pinpointing the genomic localizations of chromatin-associated proteins: the yesterday, today, and tomorrow of ChIP-seq. *Curr. Protoc. Cell Biol.*, **84**, e89.
3. Tran,N.T.L. and Huang,C.H. (2014) A survey of motif finding web tools for detecting binding site motifs in ChIP-Seq data. *Biol. Direct*, **9**, 4.
4. Thomas,R., Thomas,S., Holloway,A.K. and Pollard,K.S. (2017) Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.*, **18**, 441–450.
5. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.
6. Kulakovskiy,I.V. and Makeev,V.J. (2013) DNA sequence motif: a jack of all trades for ChIP-Seq data. *Adv Protein Chem. Struct. Biol.*, **91**, 135–171.
7. D'Haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotechnol.*, **24**, 423–425.
8. Spitz,F. and Furlong.,E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13** , 613–626.
9. Zambelli,F., Pesole,G. and Pavesi,G. (2013) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief. Bioinform.*, **14** , 225–237.
10. Vorontsov,I.E., Eliseeva,I.A., Zinkevich,A., Nikonov,M., Abramov,S., Boytsov,A., Kamenets,V., Kasianova,A., Kolmykov,S., Yevshin,I.S., *et al.* (2024) HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.*, **52** , D154–D163.
11. Liu,B., Yang,J., Li,Y., McDermaid,A. and Ma,Q. (2018) An algorithmic perspective of *de novo* cis-regulatory motif finding based on ChIP-seq data. *Brief. Bioinform*, **19**, 1069–1081.
12. Taing,L., Dandawate,A., L'Yi,S., Gehlenborg,N., Brown,M. and Meyer,C.A. (2024) Cistrome Data Browser: integrated search, analysis and visualization of chromatin data. *Nucleic Acids Res.*, **52**, D61–D66.
13. Hammal,F., de Langen,P., Bergon,A., Lopez,F. and Ballester,B. (2022) ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.*, **50**, D316–D325.
14. 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.
15. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
16. Jolma,A. and Taipale,J. (2011) Methods for analysis of transcription factor DNA-binding specificity in vitro. *Subcell. Biochem.*, **52**, 155–173.
17. 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.
18. Franco-Zorrilla,J.M., López-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
19. Rauluseviciute,I., Riudavets-Puig,R., Blanc-Mathieu,R., Castro-Mondragon,J.A., Ferenc,K., Kumar,V., Lemma,R.B., Lucas,J., Chèneby,J., Baranasic,D., *et al.* (2024) JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **52**, D174–D182.
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. Wingender,E. (2013) Criteria for an updated classification of human transcription factor DNA-binding domains. *J. Bioinform. Comput. Biol.*, **11**, 1340007.
22. Wingender,E., Schoeps,T. and Dönitz,J. (2013) TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.*, **41**, D165–D170.
23. 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.
24. Wingender,E., Schoeps,T., Haubrock,M., Krull,M. and Dönitz,J. (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.*, **46**, D343–D347.
25. Blanc-Mathieu,R., Dumas,R., Turchi,L., Lucas,J. and Parcy,F. (2023) Plant-TFClass: a structural classification for plant transcription factors. *Trends Plant Sci.*, **29**, 40–51.
26. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R., *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
27. Srivastava,S., Avvaru,A.K., Sowpati,D.T. and Mishra,R.K. (2019) Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics*, **20**, 153.
28. Ross,M.G., Russ,C., Costello,M., Hollinger,A., Lennon,N.J., Hegarty,R., Nusbaum,C. and Jaffe,D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**, R51.
29. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
30. Redhead,E. and Bailey,T.L. (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinf.*, **8**, 385.
31. Keilwagen,J., Grau,J., Paponov,I.A., Posch,S., Strickert,M. and Grosse,I. (2011) De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Comput. Biol.*, **7**, e1001070.
32. Simcha,D., Price,N.D. and Geman,D. (2012) The limits of *de novo* DNA motif discovery. *PLoS One*, **7**, e47836.
33. Boeva,V. (2016) Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front. Genet.*, **7**, 24.
34. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J., *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
35. Jayaram,N., Usvyat,D. and Martin,A.C.R. (2016) Evaluating tools for transcription factor binding site prediction. *BMC Bioinf.*, **17**, 547.
36. Castellana,S., Biagini,T., Parca,L., Petrizzelli,F., Bianco,S.D., Vescovi,A.L., Carella,M. and Mazza,T. (2021) A comparative benchmark of classic DNA motif discovery tools on synthetic data. *Brief. Bioinform.*, **22**, bbab303.
37. Csurös,M., Noé,L. and Kucherov,G. (2007) Reconsidering the significance of genomic word frequencies. *Trends Genet.*, **23**, 543–546.
38. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H., Glass,C.K., *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
39. Worsley Hunt,R.W., Mathelier,A., del Peso,L. and Wasserman,W.W. (2014) Improving analysis of transcription factor binding sites within ChIP-seq data based on topological motif enrichment. *BMC Genomics*, **15**, 472.

40. Dang,L.T., Tondl,M., Chiu,M.H.H., Revote,J., Paten,B., Tano,V., Tokolyi,A., Besse,F., Quaife-Ryan,G., Cumming,H., *et al.* (2018) TrawlerWeb: an online *de novo* motif discovery tool for next-generation sequencing datasets. *BMC Genomics*, **19**, 238.
41. Tsukanov,A.V., Mironova,V.V. and Levitsky,V.G. (2022) Motif models proposing independent and interdependent impacts of nucleotides are related to high and low affinity transcription factor binding sites in Arabidopsis. *Front. Plant Sci.*, **13**, 938545.
42. Sharov,A.A. and Ko,M.S. (2009) Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Res.*, **16**, 261–273.
43. Bailey,T.L. (2021) STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**, 2834–2840.
44. Khan,A., Riudavets Puig,R., Boddie,P. and Mathelier,A. (2021) BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences. *Bioinformatics*, **37**, 1607–1609.
45. Santana-Garcia,W., Castro-Mondragon,J.A., Padilla-Gálvez,M., Nguyen,N.T.T., Elizondo-Salas,A., Ksouri,N., Gerbes,F., Thieffry,D., Vincens,P., Contreras-Moreira,B., *et al.* (2022) RSAT 2022: regulatory sequence analysis tools. *Nucleic Acids Res.*, **50**, W670–W676.
46. 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**, 137.
47. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
48. Tian,F., Yang,D.C., Meng,Y.Q., Jin,J.P. and Gao,G. (2020) PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.*, **48**, D1104–D1113.
49. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M., *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
50. McLeay,R.C. and Bailey,T.L. (2010) Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinf.*, **11**, 165.
51. Gupta,S., Stamatoyannopolous,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
52. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
53. Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., Smith,N.J., *et al.* (2020) Array programming with NumPy. *Nature*, **85**, 357–362.
54. Amemiya,H.M., Kundaje,A. and Boyle,A.P. (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep.*, **9**, 9354.
55. Telese,F., Ma,Q., Perez,P.M., Notani,D., Oh,S., Li,W., Comoletti,D., Ohgi,K.A., Taylor,H. and Rosenfeld,M.G. (2015) LRP8-Reelin-regulated neuronal enhancer signature underlying learning and memory formation. *Neuron*, **86**, 696–710.
56. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
57. Chumpitaz-Diaz,L., Samee,M.A.H. and Pollard,K.S. (2021) Systematic identification of non-canonical transcription factor motifs. *BMC Mol. Cell Biol*, **22**, 44.
58. Yu,C.P., Kuo,C.H., Nelson,C.W., Chen,C.A., Soh,Z.T., Lin,J.J., Hsiao,R.X., Chang,C.Y. and Li,W.H. (2021) Discovering unknown human and mouse transcription factor binding sites and their characteristics from ChIP-seq data. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2026754118.
59. Lai,X., Stigliani,A., Vachon,G., Carles,C., Smaczniak,C., Zubieta,C., Kaufmann,K. and Parcy,F. (2019) Building transcription factor binding site models to understand gene regulation in plants. *Mol. Plant*, **12**, 743–763.
60. Karlin,S. and Ladunga,I. (1994) Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 12832–12836.
61. Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290.
62. Gentles,A.J. and Karlin,S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res.*, **11**, 540–546.
63. Siebert,M. and Söding.,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
64. Eggeling,R., Grosse,I. and Grau,J. (2017) InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics*, **33**, 580–582.
65. Jiang,M., Anderson,J., Gillespie,J. and Mayne,M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinf.*, **9**, 192.
66. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
67. 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.
68. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, **43**, e119.
69. Caldonazzo Garbelini,J.M., Kashiwabara,A.Y. and Sanches,D.S. (2018) Sequence motif finder using memetic algorithm. *BMC Bioinf.*, **19**, 4.
70. Kiesel,A., Roth,C., Ge,W., Wess,M., Meier,M. and Söding,J. (2018) The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.*, **46**, W215–W220.
71. 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.
72. Samee,M.A.H., Bruneau,B.G. and Pollard,K.S. (2019) A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.*, **8**, 27–42.
73. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S., *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
74. Worsley Hunt,R. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.*, **15**, 412.
75. Levitsky,V., Zemlyanskaya,E., Oshchepkov,D., Podkolodnaya,O., Ignatieva,E., Grosse,I., Mironova,V. and Merkulova,T. (2019) A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res.*, **47**, e139.
76. Karimzadeh,M. and Hoffman,M.M. (2022) Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome. *Genome Biol.*, **23**, 126.
77. Keilwagen,J., Posch,S. and Grau,J. (2019) Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.*, **20**, 9.
78. Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.
79. 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.


**Received:** March 6, 2024. **Revised:** June 3, 2024. **Editorial Decision:** July 11, 2024. **Accepted:** July 19, 2024