DOI 10.18699/SBB-2022-044

Massive comparison of the 'genomic' and 'shuffled' background set generation approaches for efficiency of *de novo* motif search in *A. thaliana* ChIP-seq data

Raditsa V.V.*, Tsukanov A.V., Levitsky V.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
* *radicav06@gmail.com*

Key words: background data, transcription factor binding sites, genome sequence bias

Motivation and Aim: 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). Motif enrichment analyses in the foreground data of ChIP-seq loci depend on the choice of the approach of background data generation. The task of this analysis consists in detection of motifs specific for biological function of a ChIP-seq target TFs. Hence, it is critically important to pay attention to false positive genomic sequence biases, such as polyA tracts, which most probably do not respect to interactions of target TFs with genomic DNA [1, 2]. However, up to now a systematic massive analysis of efficiency of various approaches of the background set generation for motif enrichment analyses in ChIP-seq data has not been performed.
Methods and Algorithms: We used in analysis 111 ChIP-seq datasets from GTRD database [3]. We applied the STREME [4] tool for de novo motif search. To generate background data for this search we used two approaches, each of them presumes the same nucleotide content and lengths as those from peaks of foreground data. The first approach 'Shuffled' is the standard shuffling of nucleotides, i.e. the background dataset includes artificial sequences, which were generated according to the Markov chain of 0^{th} order. The second approach 'Genomic' compiles randomly chosen genomic sequences [5]. We considered in analysis ten top ranked enriched motifs for each ChIP-seq dataset. The motif enrichment we estimated with the Fisher's exact test comparing fractions of recognized sequences in the foreground and background datasets (the same recognition threshold for all motifs were applied, see [6], the recognition rate 2E-4 for the whole genome set of promoter respected this threshold). The TOMTOM tool [7] we used to prove the similarity between motifs. In particular (a) between results of de novo motif search and known motifs of target TFs and (b) between results of de novo motif search and the simple repeats (10 $(XY)_4$ and 32 $(XYZ)_3$ motifs). Here 10/32 imply the number of distinct possible di-/tri-nucleotides with respect to DNA strand.
Results: We compiled top ten enriched motifs from de novo motif search [4] for all ChIP-seq datasets and compared these motifs with (a) known motifs of target TFs and (b) motifs of simple repeats [7]. Analysis of the example GSM2898772 dataset for TF AGL8 (Fig. 1) shows that the background 'Genomic' has the better efficiency in the search for target motifs than the 'Shuffled' one. We may conclude that the first ranked motifs are the motif of target TF and polyA for the 'Genomic' and 'Shuffled' background datasets respectively. The distribution of ranks of ChIP-seq target motifs for all 111 ChIP-seq datasets confirms this conclusion (Fig. 2, A). Thus, among all 111 datasets in 75 cases (68 %) the target motifs have the first rank, in 97 (87 %) cases ranks are from one to ten. For the background 'Shuffled', 47 (42 %) target motifs possess the first rank and 89 (80 %) have ranks from one

to ten ranks. Figure 2, B shows respective distributions for motifs of simple repeats. Here we observe the opposite trend to that for motifs of target TFs.

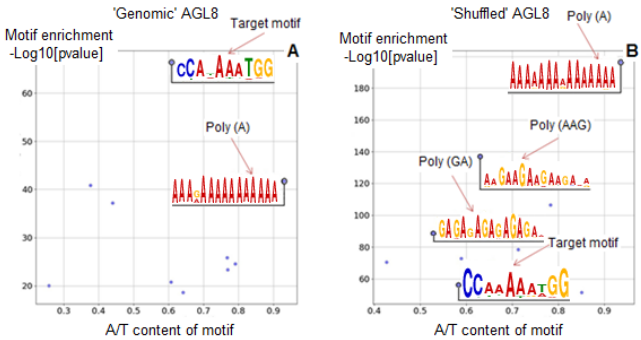

Fig. 1. The nucleotide content bias in motif representation analysis on the example ChIP-Seq dataset for TF AGL8 (GSM2898772). Axes X imply A/T content of motif. Axes Y imply the motif enrichment –Log10[pvalue]. Panels A and B shows the results of de novo motif search for the background datasets 'Genome' and 'Shuffled'

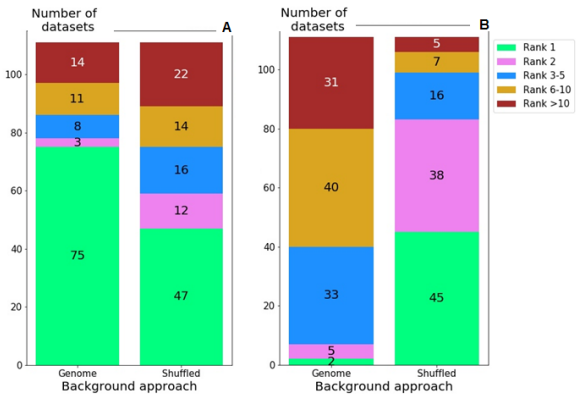

Fig. 2. The distribution of ranks of enriched motifs in the results of de novo motif search for all 111 ChIP-seq datasets. Panels A and B depict motifs of target TFs and simple repeats. Axes X shows the background dataset approaches 'Genomic' and 'Shuffled'. Axes Y marks the number of ChIP-seq datasets respecting detection of motifs of certain rank

Hence, application of the genomic background dataset avoids in the results of de novo motif search unspecific genome sequence bias motifs such as polyA. Overall, the approach for genomic background is more efficient than the approach, of nucleotide shuffling.

Acknowledgements: The study was supported by Russian Science Foundation project 21-14-00240.

References

1. Khan A. et al. BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences. *Bioinformatics*. 2021;37(11):1607-1609.
2. Worsley Hunt R. et al. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*. 2014;15(1):472.
3. Kolmykov S. et al. *Nucl Acids Res*. 2021;49(D1):D104-D111.
4. Bailey T.L. *Bioinformatics*. 2021;37:2834-2840.
5. The software package SiteGA for de novo motif search in ChIP-seq data (https://github.com/parthian-sterlet/sitega).
6. Levitsky V. et al. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res*. 2019;47:e139.
7. Gupta S. et al. Quantifying similarity between motifs. *Genom Biol.* 2007;8(2):R24.