doi 10.18699/bgrs2024-1.2-23

Identification of potential transcription factors from whole transcriptome data by taking into account enrichment and classification of motifs of known transcription factors, and their expression levels

Tsukanov A.V.*, Levitsky V.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
** tsukanov@bionet.nsc.ru*

Key words: transcription factors; motif enrichment; RNA-seq; transcription; regulation of transcription

Motivation and Aim: To date, RNA-seq remains the most popular technology for the whole transcriptome analysis identifying differentially expressed genes (DEGs). An important challenge in the analysis of RNA-seq data is the search for transcription factors (TFs) associated with changes in gene expression levels in response to external stimuli. A straightforward approach to solve this issue is to define the enriched motifs of known TFs in promoters of DEGs. We previously proposed an ESDEG tool (https://github.com/ubercomrade/esdeg) [1] that implemented this approach, but its efficiency may be further improved. While the number of known motifs is approached the number of known TFs (above 1400 and 1600 for human [2, 3]), the number of structurally different motifs is substantially smaller [4], and structurally-related TFs often have very similar motifs. Hence, a number of distinct enriched motifs can reach several hundreds, and each of them often respects at least several TFs. Thus, now we are still too far from a reliable definition of key TFs associated with the change of the whole transcriptome. Here we apply information on the classification of TFs [5] and their expression levels to refine the lists of reliable TFs from RNA-seq data.
Methods and Algorithms: As input data, we use a counting matrix in which genes are rows, the test and control samples are columns, and cells are the numbers of reads mapped to genes. Next, the *pyDEseq2* package [6] determines DEGs (we apply the criteria $|log_2(Fold\ Change)| > 1$ and $padj < 0.05$), the *rnanorm* tool (https://github.com/genialis/RNAnorm) calculates *FPKM* values for genes (normalized read counts based on gene length and the total number of mapped reads). Further, the ESDEG tool is applied. It takes an annotation of DEGs from *pyDEseq2* and the list of all motifs as positional frequencies matrices from the HOCOMOCO database [2]. The ESDEG identifies the list of enriched motifs in promoters of DEGs using a Monte Carlo approach. For each motif the enrichment (odds) ratio ($log_2(OR)$) and significance (*em_padj*) are defined as described earlier [1]. Here we introduced a few changes to ESDEG as follows. For each motif, class and family information for TFs [2, 5] is provided. Then, the expression level values (*FPKM*) as well as expression level changes ($log_2(Fold\ Change)$) and theirs significances (*padj*) are given for TFs of all motifs. As a result, a common table combines expression levels and motif enrichment information for each TF. The final list of TFs is determined as follows. First, we restrict the most significant enriched motifs, $em\_padj < 0.05$ and $log_2(OR) > 1.0$. Second, we test whether the most reliable TF genes are found among DEGs. Third, only those TFs with *FPKM*

values above a specified threshold are left. The threshold was set as the third quartile (Q_3) of the *FPKM* distribution over all TFs.

Results: The count matrix for the RNA-seq data of the E-MTAB-6598 experiment (human K562 cell line, interferon alpha treatment) we obtained from the Expression Atlas database (https://www.ebi.ac.uk/gxa/home). We used 5'-regions (–400; +100) relative gene transcription start sites in analysis. The *pyDEseq2* determined 159 DEGs, and ESDEG found totally 226 significantly enriched motifs for 173 TFs among 1443 motifs for 949 TFs. They were classified into 42 families [2, 5] (Fig. 1).

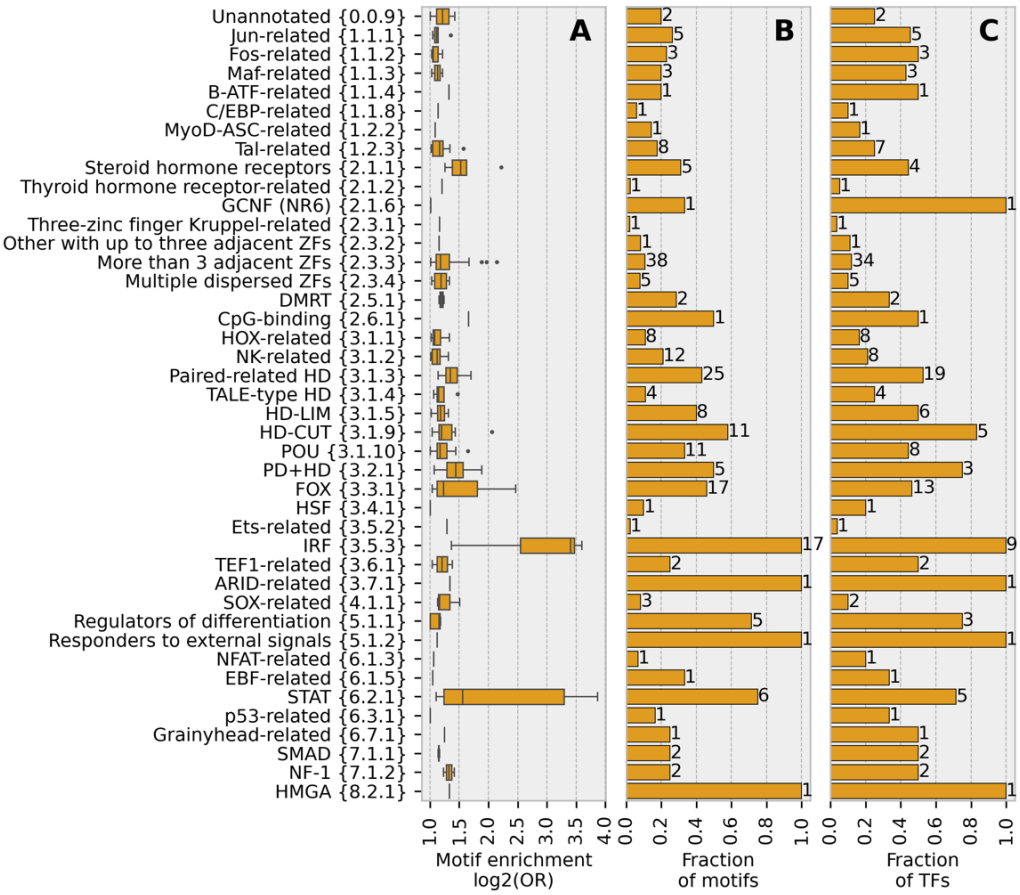

Fig. 1. Results of ESDEG tool application for the RNA-seq data of the E-MTAB-6598 experiment. **A**, Distribution of the motif enrichment for TF families. **B** and **C**, Fractions of enriched motifs and their TFs per TF family, for each family the numbers of enriched motifs and TFs are marked

The highest enrichment we found for the motifs of TFs from the IRF {3.5.3} and STAT {6.2.1} families (Fig. 1A). However, the fractions of enriched motifs notably varied between different families (Fig. 1B). Motifs of all nine TFs from the IRF family are enriched, five out of total seven TFs from the STAT family have enriched motifs (71 %, Fig. 1C). TFs from the IRF and STAT families are involved in signaling pathways related to the regulation of type I interferon [7]. Unfortunately, the results of motif enrichment analysis are insufficient to identify specific TFs exactly. The TF-DEG matching criterion left only four most reliable TFs from three families (Table 1, bold).

The criterion on the gene expression level gave a few families and TFs with enriched motifs. Here we kept only TFs with FPKM > 4.3 (the value of Q_3). This allowed us to predict 31 additional TFs from 20 TF families. Table 1 shows TFs from 12 families with the highest enrichment of motifs. Totally, among 173 TFs predicted by ESDEG tool, the TF-DEG matching and gene expression level criteria mark 35 TFs as reliable.

Table 1. Application of the gene expression level and TF-DEG matching criteria to the results of ESDEG analysis. Bold font marks DEGs representing TFs

TF family	Number of TFs before / after filtration		List of TFs
Maf-related {1.1.3}	3	1	MAFG
Tal-related {1.2.3}	7	1	TAL1
Steroid hormone receptors {2.1.1}	4	1	NR3C1
Thyroid hormone receptor-related {2.1.2}	1	1	RARA
Three-zinc finger Kruppel-related {2.3.1}	1	1	KLF3
Paired-related HD {3.1.3}	18	1	**HESX1**
FOX {3.3.1}	13	1	FOXM1
IRF {3.5.3}	9	3	**IRF9**, **IRF7**, IRF3
TEF1-related {3.6.1}	2	2	TEAD3, TEAD1
ARID-related {3.7.1}	1	1	ARID1A
STAT {6.2.1}	5	4	**STAT1**, STAT6, STAT5B, STAT2
HMGA {8.2.1}	1	1	HMGA1

Conclusion: We have developed an approach that applied additional information on TF classification and their expression levels to deduce TF regulators from RNA-seq data. Our approach allowed us to identify the most reliable TFs involved in DEG regulation among a large number of TFs whose motifs can be enriched in DEG promoters.

Funding: The study is supported by Russian Science Foundation project No. 20-14-00140.

References

1. Oshchepkov D. et al. Transcription factors as important regulators of changes in behavior through domestication of gray rats: quantitative data from RNA sequencing. *Int J Mol Sci*. 2022;23(20):12269. doi 10.3390/ijms232012269
2. Vorontsov I.E. et al. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* 2024;52(D1):D154-D163. doi 10.1093/nar/gkad1077
3. Lambert S.A. et al. The human transcription factors. *Cell*. 2018;172(4):650-665. doi 10.1016/j.cell.2018.01.029
4. Ambrosini G. et al. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol.* 2020;21(1):114. doi 10.1186/s13059-020-01996-3
5. Wingender E. et al. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* 2018;46(D1):D343-D347. doi 10.1093/nar/gkx987
6. Muzellec B. et al. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics.* 2023;39(9):btad547. doi 10.1093/bioinformatics/btad547
7. Mogensen T.H. IRF and STAT transcription factors – from basic biology to roles in infection, protective immunity, and primary immunodeficiencies. *Front Immunol.* 2019;9:3047. doi 10.3389/fimmu.2018.03047