DOI 10.18699/SBB-2022-374

# CisCross web service: a gene set enrichment analysis to predict the upstream regulators for *Arabidopsis thaliana*

Lavrekha V.V.[1, 2]*, Tsukanov A.V.[1], Bogomolov A.G.[1], Omelyanchuk N.[1], Zemlyanskaya E.V.[1, 2], Levitsky V.G.[1, 2], Mironova V.V.[1, 3]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Department of Natural Sciences, Novosibirsk State University, Novosibirsk, Russia*

[3] *Faculty of Science, Radboud University Nijmegen, Nijmegen, Netherlands*

** vvl@bionet.nsc.ru*

**Key words:** DAP-seq, gene set enrichment, transcription factor binding sites, *Arabidopsis thaliana*

*Motivation and Aim*: Modern transcriptomic technology RNA-seq provides lists of differentially expressed genes (DEG), i.e. genes for which expression levels between two experimental conditions of cell types, tissues or specific treatments are statistically significant [1]. Although consequent standard analyses uncover gene networks trying to explain observed biological responses, molecular mechanisms of DEG transcription regulation are still far from being revealed. Transcription factors (TFs) are proteins that function as major players involved in this regulation. They bind genomic target sequences known as transcription factor binding sites (TFBS) to initiate gene activation or repression. Recently developed DAP-Seq (DNA Affinity Purification and sequencing) high-throughput method detects the DNA fragments from genomic DNA libraries that are bound by an *in vitro* expressed TF [2]. O'Malley and colleagues generated DAP-Seq profiles for 387 *Arabidopsis thaliana* TFs. The large collection of similarly processed in [2] DAP-Seq sets of peaks, The Plant Cistrome database, has been used for systematic identification of TF upstream regulators [3, 4]. Two web-applications for gene set enrichment analysis in Arabidopsis, TF DEACoN [4] and EAT-UpTF [5], have been developed already, which use Plant Cistrome list of TFs's targets. However, the Plant Cistrome database consists of processed by GEM tool sets of peak [2, 6]. As shown in a benchmarking study MACS2 peak calling tool may give better quality of peak sets [7, 8]. That's why we re-annotated raw DAP-Seq data [2] by the MACS2 tool so it fits better for the task of the gene set enrichment analysis. The algorithm of TF enrichment analysis, CisCross algorithm, which we used previously in [3] and new collections of TF's sets of peaks we perform on the new CisCross web service.

*Methods and Algorithms*: All 931 raw datasets [2] were processed and aligned with snakePipes (v. 2.5.6) [9]. Peak calling was done by MACS2 [8] or GEM [7]. For all samples, we used control SRR2926068 or SRR2926069 depending on the data type. We filtered out the peak sets with a FRIP value of less than 0.01. If two DAP-seq replicas were obtained for a TF, then we joined them using the IDR (irreproducible discovery rate) tool [10]. Then we select the final variant of a set of peaks and removed sets of peaks, which contained less than 200 peaks. As a result, we've got 649/650 peak sets for 426/428 TFs for CisCross MACS2/GEM collections.

*CisCross algorithm.* 5'-regulatory regions of the requested list of genes are used as a foreground and those for the rest Arabidopsis genes as a background. For each TF peak set, CisCross counts the number of genes in foreground/background which 5'-regulatory

regions possess or do not possess the TF binding peak. The gene promoter enrichment for TF binding peaks is assessed using Fisher's exact test. CisCross runs through all the peak sets in the selected DAP-Seq collection and applies the correction for multiple comparisons.

*Results*: The CisCross web service (https://plamorph.sysbio.ru/ciscross/) includes three modes: (1) CisCross-Main, (2) CisCross-Light, and (3) TF Targets. CisCross-Light gives the list of DAP-Seq binding peaks detected in the 5'-regulatory region of the requested gene. CisCross-Main implements the algorithm written above with the following options. For the background one can choose between Araport11 and TAIR10 genome versions; 500, 1000, or 1500 bp-long 5'-regulatory regions. For the DAP-Seq peak set collections: (1) Plant Cistrome [2]; (2) CisCross GEM; (3) CisCross MACS2. For the Multiple testing procedure: Benjamini–Hochberg or Bonferroni methods. TF Target mode is aimed at finding all targets of input TF among the genes submitted for input.

*Conclusion*: As CisCross collections contains the peak sets for a greater then Plant Cistrome amount of TFs, it resulted in detection of novel regulators for gene sets. Using different backgrounds and length of upstream region gives the researcher more opportunities to investigate regulation of genes by CicCross-Main. In addition, CisCross web service modes help to explore the output of the CisCross-Main in more detail.

*Acknowledgements*: The work was supported by Russian Science Foundation (21-14-00240).

*References*

1. Oshlack A., Robinson M.D., Young M.D. From RNA-seq reads to differential expression results. *Genome Biol*. 2010;11(12):220. doi: 10.1186/gb-2010-11-12-220.
2. O'Malley R.C., Huang S.S.C., Song L., Lewsey M.G., Bartlett A., Nery J.R., …, Ecker J.R. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell.* 2016;165(5):1280-1292.
3. Shi D., Jouannet V., Agustí J., Kaul V., Levitsky V., Sanchez P., ..., Greb T. Tissue-specific transcriptome profiling of the Arabidopsis inflorescence stem reveals local cellular signatures. *Plant Cell.* 2021;33(2):200-223.
4. Harkey A.F., Sims K.N., Muday G.K. A new tool for discovering transcriptional regulators of co-expressed genes predicts gene regulatory networks that mediate ethylene-controlled root development. *In silico Plants*. 2020;2(1):diaa006.
5. Shim S., Seo P.J. EAT-UpTF: Enrichment Analysis Tool for Upstream Transcription Factors of a group of plant genes. *Front. Genet*. 2020;11:1058.
6. Guo Y., Mahony S., Gifford D.K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol*. 2012;8(8):e1002638. doi: 10.1371/journal.pcbi.1002638.
7. Thomas R., Thomas S., Holloway A.K., Pollard K.S. Features that define the best ChIP-seq peak calling algorithms. *Brief Bioinform*. 2017;18(3):441-450. doi: 10.1093/bib/bbw035.
8. Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., Liu X.S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):r137. doi: 10.1186/gb-2008-9-9-r137.
9. Bhardwaj V., Heyne S., Sikora K., Rabbani L., Rauer M., Kilpert F., ..., Manke T. snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics*. 2019;35(22):4757-4759.
10. Li Q., Brown J.B., Huang H., Bickel P.J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat*. 2011;5(3):1752-1779. doi: 10.1214/11-AOAS466.