doi 10.18699/bgrs2024-1.2-11

Prediction and annotation of alternative transcription starts in the chicken genome

Grushina V.*, Pintus S.
Sirius University of Science and Technology, Sirius, Krasnodar region, Russia
** grushina_v@bk.ru*

Key words: CAGE; promotor shift; *Gallus gallus* genome; transcription start sites

Motivation and Aim: The analysis of promoter regions of genes often reveals differences from the existing annotation, which can be caused both by the insufficient accuracy of the existing annotation and the existence of transcripts previously unknown and activated under certain non-standard conditions [1, 2]. Taking into account the data from CAGE experiments would make it possible to better identify the initial positions of such transcripts, which allows us to determine the position of the start of transcription, as well as to identify the events of the promoter shift. Such insights would provide a better understanding of the mechanisms of transcriptional regulation in the chicken genome and improve the quality of Gallus gallus genome annotation.
Currently, data from CAGE experiments are available for various chicken tissues, including muscles, kidneys, liver, brain, and others. This provides the opportunity to annotate enhancers and alternative initial transcription positions that are active in various tissues. This approach ensures the completeness of the annotation, taking into account the differences in the gene expression in different tissues.
In this study, we predicted transcription start sites (TSS) in the chicken genome and evaluated their expression based on data obtained from CAGE experiments, including those of the FANTOM5 project [3]. Our predictions were compared against the Ensembl and Refseq annotations. We found new transcription starts, including those presumably associated with the events of promoter shift.
Methods and Algorithms: The galGal6 and galGal7 genome assemblies and Ensemble and Refseq genome annotations were used as a referencefor comparing and validating TSS coordinates. STAR [4] (v.2.7.11b) was used to align the CAGE reads against the reference genome, then the CAGE tag peaks were clustered with DPI1 tool [5]. The TSS coordinates and alternative transcription sites (promoter shift events) were identified with the CageFightR package [6] (v. 1.24.0).
Results: Differences were found between the predicted transcription starts in FANTOM5 and Ensembl and Refseq genome annotations of the chicken genome version 6, the re-aligned readings of FANTOM5 to the 7th version genome and Ensembl and Refseq genome annotations of the chicken genome version 7. Thus, the data from the FANTOM5 project were annotated to the latest version of the chicken genome, galGal7, through read realignment. This allowed the validation of TSS obtained as a result of the analysis of new CAGE experiments.
We predicted and annotated transcription sites in the chicken genome based on new CAGE experiments for different tissues. The CAGE reads were aligned to the galGal7 chicken genome and, based on the alignment results, the CAGE peaks were counted and clustered and noise corrected. The obtained transcription starts were compared by

interval arithmetic methods of the obtained transcription starts with Ensembl and Refseq annotations for the galGal7 chicken genome.
We identified and annotated statistically significant alternative transcription starts in the chicken genome (putative promoter shifts) based on data from the FANTOM5 project and new experimental CAGE data.
Conclusion: In addition to the major transcription starts annotated in Refseq and Ensembl reference annotations, CAGE experiments reveal additional TSS. Our results suggest that some genes are characterized by the phenomenon of transcription start shifts that manifest under specific conditions.
Funding: The study is by the Russian Science Foundation (project 24-24-20106, https://rscf.ru/project/24-24-20106/)

References

1. Link V.M. Analysis of genetically diverse macrophages reveals local and domain-wide mechanisms that control transcription factor binding and function. *Cell*. 2018;173:1796-1809. doi 10.1016/j.cell.2018.04.018
2. Anderson W.D., Duarte F.M. et al. Defining data-driven primary transcript annotations with primaryTranscriptAnnotation in R. *Bioinformatics*. 2020;36(9):2926-2928. doi 10.1093/bioinformatics/btaa011
3. Lizio M., Abugessaisa I., Noguchi S., Kondo A., Hasegawa A. et al. Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res*. 2019;47(1):752-758. doi 10.1093/nar/gky1099
4. Dobin A., Davis A.C., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi 10.1093/bioinformatics/bts635
5. Forrest A., Kawaji H., Rehli M. et al. A promoter level mammalian expression atlas. *Nature*. 2014;507(7493):462-470. doi 10.1038/nature13182
6. Thodberg M., Thieffry A., Vitting-Seerup K., Andersson R., Sandelin A. CAGEfightR: analysis of 5′-end data using R/Bioconductor. *BMC Bioinformatics*. 2019;20:487. doi 10.1186/s12859-019-3029-5