DOI 10.18699/SBB-2018-46

New method for estimation of number of transcription factor binding sites using results of processing of ChIP-seq data by different peak callers

S. Kolmykov[1, 2], Yu. Kondrakhin[1, 3], F. Kolpakov[1, 3]*

[1] *Biosoft.Ru Ltd., Novosibirsk, Russia*

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

[3] *Institute of Computational Technologies SB RAS, Novosibirsk, Russia*

* *e-mail: fkolpakov@gmail.com*

Key words: ChIP-seq, quality control, peak caller, transcription factor binding sites

Motivation and Aim: Identification of transcription factor binding sites in genomes has been one of the most important tasks of modern biology. The accumulation of a large number of ChIP-seq data sets worldwide has led to the establishment of dedicated databases, in particular, ENCODE, GTRD [1] and ReMap [2]. Obviously, it is necessary to perform quality analysis of collected data sets because the quality of ChIP-seq experiments can vary significantly. The common practice for assessment of ChIP-seq data sets quality is to use well known quality criteria developed within ENCODE project. For example, NRF, PBC1, PBC2 is extensively used to control the quality of alignments while IDR and FRiP are exploited for assessment of peak callers outputs quality [3]. In addition to these quality metrics we have proposed two novel quality metrics FNCM (False Negative Control Metric) and FPCM (False Positive Control Metric) to control false negatives and false positives rates, respectively.

Methods and Algorithms: Both developed metrics are based on assessment of transcription factor binding sites number with the help of population size estimation approach. In particular, for creation of these metrics we used Chao's method, Lanumteang-Bohling method, Zelterman's method and maximum likelihood method.

Results: We applied proposed metrics to obtain refined data sets of transcription factor binding sites and for peak callers comparison of ChIP-seq data sets from GTRD database. In particular, the refined data sets allow to perform more reliable comparative analysis of position weight matrix methods for binding sites prediction. In the case of peak caller comparison, we revealed the following ranking of peak callers, when ChIP-Seq input controls are available: MACS > GEM > SISSRS > PICS. On the other hand, the altered ranking MACS > SISSRS > PICS > GEM was obtained when ChIP-Seq input controls were absent.

Conclusion: The proposed metrics are appeared to be fruitful for obtaining refined data sets of transcription factor binding sites and for peak callers comparison.

Acknowledgements: Supported by the RFBR (No. 17-00-00296 KOMFI).

References

1. Yevshin I. et al. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Research. 45(Database issue):D61-D67.
2. Chèneby J. et al. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Research. 46(Database issue):D267-D275.
3. Stephen G.L. et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research. 22(9):1813-1831.