DOI 10.18699/BGRS/SB-2020-065

# Peak caller comparison through quality control of ChIP-Seq datasets

Ruslan N. Sharipov
BIOSOFT.RU, LLC;
Novosibirsk State University
Novosibirsk, Russia
shrus79@biosoft.ru

Yury V. Kondrakhin
Institute of Computational
Technologies SB RAS;
BIOSOFT.RU, LLC
Novosibirsk, Russia
yvkondrat@mail.ru

Semyon K. Kolmykov
FRC Institute of Cytology and Genetics
SB RAS;
Institute of Computational
Technologies SB RAS
Novosibirsk, Russia
kolmykovsk@gmail.com

Ivan S. Yevshin
Institute of Computational
Technologies SB RAS;
BIOSOFT.RU, LLC
Novosibirsk, Russia
ivan@biosoft.ru

Anna S. Ryabova
Institute of Computational
Technologies SB RAS;
BIOSOFT.RU, LLC
Novosibirsk, Russia
anna@biosoft.ru

Fedor A. Kolpakov
Institute of Computational
Technologies SB RAS;
BIOSOFT.RU, LLC
Novosibirsk, Russia
fedor@biosoft.ru

*Abstract* — **Chromatin immunoprecipitation followed by high throughput sequencing, i.e. ChIP-Seq, is a widely used experimental technology for the identification of functional protein-DNA interactions. Nowadays, such databases as GTRD, ChIP-Atlas and ReMap systematically collect and annotate a large number of ChIP-Seq datasets generated by distinct peak callers, including MACS2. The quality control of such datasets is currently indispensable, since the peak callers may produce different results for the same ChIP-seq experiment. We have performed a comparative analysis of intensively used peak callers with the help of two metrics that control false positive/negative rates. We have found that MACS2 outperformed its competitors.**

***Keywords — quality control, ChIP-Seq datasets, peak caller, false positives, false negatives, GTRD database***

## Introduction

Understanding the basic mechanisms of transcription regulation is a big problem in modern biology. Regulation of transcription is a complex process in which transcription factors play a key role. Nowadays, ChIP-Seq experiments are widely used to detect protein-DNA binding in whole genomes. To date, several databases, such as GTRD [1], ChIP-Atlas and ReMap, accumulate ChIP-Seq datasets obtained by applying various peak callers to the primary ChIP-Seq data. To control the quality of accumulated datasets distinct control metrics are used. For example, such well-known metrics as Non-Redundant Fraction (NRF), PCR Bottlenecking Coefficient 1 and 2 (PBC1 and PBC2), Normalized Strand Cross-correlation coefficient (NSC), and Relative Strand Cross-correlation coefficient (RSC) evaluate the quality of read alignments for individual genomes. These metrics were developed as part of the ENCODE project [2]. However, these metrics do not control false positive and false negative rates. Recently, two quality control metrics, namely, the False Positive Control Metric (FPCM) and the False Negative Control Metric (FNCM), were developed on the base of the population size estimation approach [3].

Several tens peak callers have been developed to generate transcription factor binding regions (TFBRs) from aligned ChIP-Seq data. [4] However, a comparative analyses [5, 6] of peak callers did not reveal so far the best among them. We performed comparative analysis of the most popular peak callers – GEM [7], MACS, MACS2 [8], SISSRs [9] and PICS [10]. For this purpose, we used the FPCM and FNCM metrics, as well as 8982 TFBR datasets from the GTRD database. The conducted comparative analysis showed that MACS2 outperformed its competitors.

In this study we applied some rank aggregation (RA) methods for a meta-analysis of transcription factor binding sites (TFBSs) obtained from ChIP-seq data publically available in the GTRD database. Additionally, we are introducing a new RA method based on the Borda method utilizing values of FPCM and FNCM quality metrics.

## Materials And Methods

To generate TFBR datasets, we used two distinct scenarios. According to scenario 1, see Fig. 1(A), the four peak callers - GEM, MACS, PICS, and SISSRs – were applied independently to the same ChIP-Seq set of reads aligned to the reference genome. Then the obtained four sets of peaks were merged into a final dataset. According to scenario 2, MACS was replaced by MACS2, and the final TFBR dataset was obtained by overlapping the peaks instead of merging them. The processes of merging and overlapping peaks are demonstrated in Fig. 1.

To compare peak callers, we used FPCM and FNCM metrics [3]. FNCM for each peak caller was defined as the ratio of the observed number of its peaks to the estimated number of genuine peaks. FNCM varies in the range [0.0; 1.0]. The closer the FNCM value to 1.0, the lower the false-negative rate, and the values closer to 0.0 indicate that a large number of genuine peaks have been missed. FPCM was defined as the ratio of the observed number of orphans in the TFBR dataset to the estimated number of true orphans. In turn, orphans were defined as separate peaks that did not overlap with other initial peaks. If the difference between the observed and estimated number of orphans is insignificant, then the FPCM should be close to 1. Such FPCM values indicate that erroneously formed peaks are practically absent. However, if the FPCM considerably exceeds 1 (e.g., FPCM > 2.0 or FPCM > 3.0), then at least half or more orphans are classified as false positives.

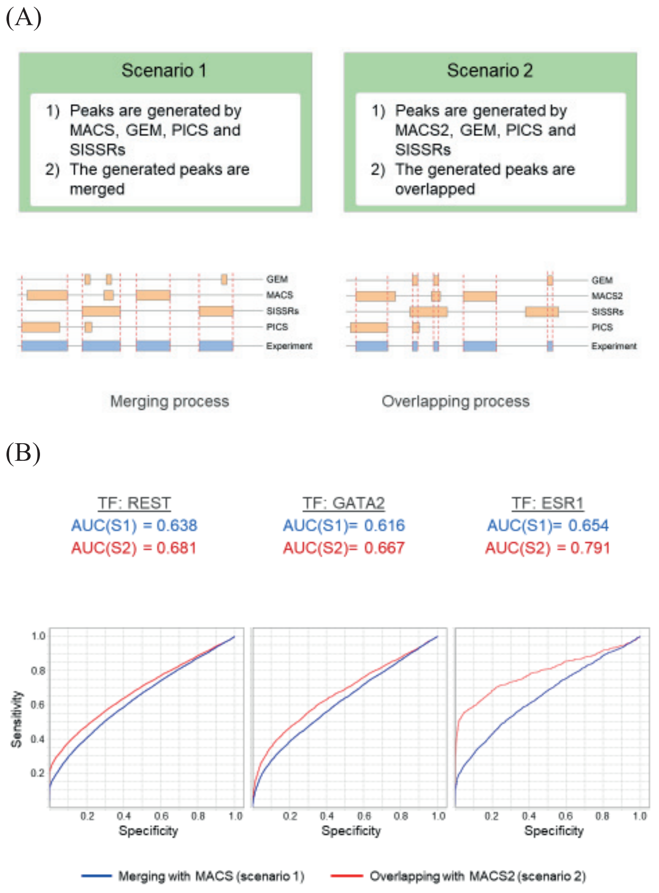

Fig. 1. Two scenarios for peak caller comparison. (A) scenario definition; (B) ROC curves and AUC values in two scenarios

## Results And Discussion

To obtain reliable conclusions, we performed a comparative analysis of 8982 human ChIP-Seq experiments stored in GTRD. For each experiment, we calculated two FPCM values, say FPCM1 and FPCM2, which corresponded to scenario 1 and scenario 2, respectively. In general, in 70.9% of experiments, replacing MACS with MACS2 resulted in improved quality by lowering FPCM. In particular, for experiments PEAKS034562 (REST), PEAKS033231 (GATA2) and PEAKS034509 (ESR1), the pair of values (FPCM1, FPCM2) is equal to (14.431, 1.167), (10.778, 1.132) and (11.514, 1.204).

We performed a direct comparison of peak callers by comparing their FNCM. Thus, for experiment PEAKS034562 in scenario 1 the following FNCM values were achieved: MACS – 0.967, SISSRs – 0.909, GEM – 0.807 and PICS – 0.04. Hence, MACS has outperformed its competitors because its FNCM is maximal. Overall, in 56.1% of the experiments MACS showed better results, while GEM, SISSRs, and PICS outperformed in 18.5%, 16.2% and 9.2% of experiments, respectively. In scenario 2, MACS2 showed better results in 69.5%, while GEM, SISSRs and PICS outperformed in 8.4%, 13.9% and 8.2% of experiments, respectively.

Finally, the usefulness of transition from scenario 1 to scenario 2 can be confirmed by increasing the accuracy of identifying site motifs in the three ChIP-Seq datasets mentioned above. To identify the motif, we used the following models of position weight matrix from the HOCOMOCO database [11]:

REST_HUMAN.H11MO.0.A,
GATA2_HUMAN.H11MO.0.A,
ESR1_HUMAN.H11MO.0.A.

Fig. 1(B) demonstrates that the transition from scenario 1 to scenario 2 increased the accuracy of site identification. This increase in accuracy is in good agreement with the decrease in FPCM values.

## Conclusions

Comparative analysis of GEM, MACS, MACS2, SISSRs and PICS revealed that MACS2 outperformed its competitors in terms of the FPCM and FNCM metrics.

## Acknowledgment

This work was supported by the Russian Science Foundation, grant agreement 19-14-00295.

## References

[1] I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov F, “GTRD: a database on gene transcription regulation-2019 update”. Nucleic Acids Res., vol. 47(D1), pp. D100–D105, January 2019.

[2] ENCODE Project Consortium, “An integrated encyclopedia of DNA elements in the human genome”, Nature, vol. 489, pp. 57–74, September 2012.

[3] S. K. Kolmykov, Y. V. Kondrakhin, I. S. Yevshin, R. N. Sharipov, A. S. Ryabova, and F. A. Kolpakov, “Population size estimation for quality control of ChIP-Seq datasets”, PLoS One, vol. 14(8), e0221760, August 2019.

[4] R., Thomas, S. Thomas, A. K. Holloway, and K. S. Pollard, “Features that define the best ChIP-Seq peak calling algorithms”, Brief Bioinform, vol. 18(3), pp. 441–450, May 2017.

[5] T. D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. L. Elo, “A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments”, BMC Genomics, vol. 10(1), 618, December 2009.

[6] H., Koohy, T. A. Down, M. Spivakov, and T. Hubbard, “A comparison of peak callers used for DNase-Seq data”, PLoS ONE, vol. 9(5), e96303, May 2014.

[7] Y. Guo, S. Mahony, and D.K. Gifford, “High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints”, PLoS Comput. Biol., vol. 8(8), e1002638, August 2012.

[8] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, and et al., “Model-based analysis of ChIP-Seq (MACS)”, Genome Biol., vol. 9(9):R137, September 2008.

[9] L. Narlikar and R. Jothi, “ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder”, Methods Mol. Biol., vol. 802, pp. 305–322, November 2011.

[10] X. Zhang, G. Robertson, M. Krzywinski, K. Ning, A. Droit, S. Jones, and et al., “PICS: probabilistic inference for ChIP-seq”, Biometrics, vol. 67(1), pp.151–163, March 2011.

[11] I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, A. V. Soboleva, A. S. Kasianov, H. Ashoor, and et al., “HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models”, Nucleic Acids Res., vol. 44(D1), pp. D116–D125, January 2016.