Revisiting the computational analysis of DNase sequencing

Speaker: Ivan G. Costa[1,2,3,*]
Co-authors: Eduardo G. Gusmao[1,2], Manuel Allhoff[1,3], Martin Zenke[1,2]

[1] IZKF Computational Biology Research Group, RWTH Aachen University Medical School, Aachen, Germany.
[2] Department of Cell Biology, Institute of Biomedical Engineering, RWTH Aachen University Medical School, Aachen, Germany.
[3] Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Germany.
* corresponding author

Abstract

DNase-seq is a powerful technique for detection of cell-specific binding sites in a genome-wide manner. Computational footprinting methods, which search for footprint-like DNase I cleavage patterns on the DNA, allow the detection of binding sites in a base pair resolution. There is, however, a debate in the literature on the influence of experimental artifacts as DNase I cleavage bias and transcription factor residence time on computational footprint methods. We investigated these artifacts in a comprehensive panel of DNase-seq data sets employing 10 footprinting methods and 88 transcription factors. Our comparative analysis indicates the advantage of HINT, DNase2TF and PIQ in relation to other competing methods. We demonstrate that correcting the DNase-seq signal based on cleavage bias estimation significantly improves accuracy of computational footprinting. We also propose a score to detect footprints arising from transcription factors with short residence time, as footprints of such factors have low predictive performance.

Justification

DNase-seq allowed for the first time the detection of genome-wide base pair resolution transcription factor (TF) binding sites. Indeed, computational footprinting of > 50 cells was adopted in one of ENCODE's main papers for the definition of a human *cis*-regulatory lexicon (Neph, et al., *Nature*, 2012). However, recent work revisiting DNase-seq experiments and protocols (He. et al., *Nature Methods,* 2014; Sung et al., *Molecular Cell*, 2014; Yardımcı, *et al.,* Nucleic acids research, 2014) indicated that predictive power of computational footprinting varied on a TF-specific manner. This mostly stemed from DNase I bias to cleavege particular sequences and varying binding time of distinct TFs. Our paper (Gusmao et al., *Nature Methods,* 2016) revisits these and other aspects of computational footprinting problem. We show that the effect of cleavage bias is virtually eliminated by state-of-the-art methods with bias correction strategies. We propose a statistic to indicate footprint predicitons of TFs with potential short binding residence time, as their footprints have poor predictive accuracy. We believe that our paper quantified a number of disparate arguments from the literature. It also reinforces the power and limitations of computational footprinting methods in a comprehensive way. Finally, we will also include in the talk recent developments on computational footprinting analysis for ATAC-seq experiments.

Paper: Gusmao E.G., Allhoff, M., Zenke, M., Costa, I.G. (2016)**,** Analysis of computational footprinting methods for DNase sequencing experiments, **Nature Methods,** doi:10.1038/nmeth.3772, published online 22/2/2016.