A comprehensive framework for ChIP-seq data quality assessment

Livi CM(1)[†] , Pal K(1), Sebestyén E(1) and Ferrari F(1,2)

(1) IFOM the FIRC Institute of Molecular Oncology, Milan, 20139, Italy
(2) Institute of Molecular Genetics, National Research Council, 27100, Pavia, Italy

[†] Email: carmen.livi@ifom.eu

Motivation
Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is a widely-used technique to detect genome-wide binding sites of transcription factors (TFs) and distribution of chromatin marks [1]. Because of its widespread adoption, an increasing number of datasets is nowadays available in public repositories [2,3]. The datasets are highly variable in data quality despite criteria for quality control (QC) have been previously proposed [4-6]. There is still a lack of an consensus on objective assessment parameters due to several reasons. First, the proposed QC measures are not always taking into account the fundamental differences in the sharp vs broad enrichment profiles of ChIP-seq for different chromatin marks. As a result, the proposed QC metrics are often biased by the different enrichment profiles. Second, the data curator expertise and the visual inspection of the enrichment profiles are still relevant elements in the final decision on whether the experiment was successful, thus hampering objectivity and reproducibility of quality control. The lack of a consensus on reliable QC procedures for ChIP-seq data has a negative impact in the epigenomics field. In this work we implemented a computational framework for the comprehensive and automated quality assessment of ChIP-seq data.
Methods
We collected a comprehensive set of QC-metrics including previously proposed measures based on ENCODE guidelines (EM) (4), metrics based on the global enrichment profile (GM) [5], as well as novel metrics based on local features of the enrichment profile shape (LM). The latter are based on metagene profiles that capture and quantify in detail the characteristics of different sharp or broad binding profiles. Using this comprehensive set of metrics we analysed more than 3700 ChIP-seq and more than 500 input samples from the ENCODE project [2] and Roadmap Epigenomics Consortium (Roadmap) [3] to obtain a compendium with reference values. To train a random forest classifier that can be applied to score sample quality, we filtered highly correlating metrics via hierarchical clustering. The most representative elements per cluster were finally used to train the classifier on putative good quality versus problematic/failed ChIP-seq samples. The artificial set of problematic ChIP-seq data was created by simulating a lower signal to noise ratio.
Results
Our results show that previously proposed metrics vary between narrow and broad peak profiles, thus being biased by the underlying type of binding profile. In our

work we provide a comprehensive set of QC-metrics and a compendium with reference values for a broad range of ChIP-seq profiles, including TFs and multiple types of chromatin marks. Our resulting compendium that covers a large number of enrichment profile subtypes, can be used to compare QC-metrics and provides an easy solution for quality assessment. In alternative our random forest model can be applied to screen the sample quality by integrating all the QC-metrics into one single score. Finally, we incorporated all of these tools into a user-friendly R package, available on Bioconductor (named ChIC) that can be useful to analyse novel and pre-existing ChIP-seq samples. Our package is a powerful and flexible resource useful both for less experienced users and experts, to easily perform a comprehensive assessment of a ChIP-seq experiment or to quickly screen the quality of many samples.

References

1. Park,P.J. (2009) ChIP-seq: Advantages and challenges of a maturing technology. Nat. Rev. Genet., 10, 669–680.

2. Sloan,C.A., Chan,E.T., Davidson,J.M., Malladi,V.S., Strattan,J.S., Hitz,B.C., Gabdank,I., Narayanan,A.K., Ho,M., Lee,B.T., et al. (2016) ENCODE data at the ENCODE portal. Nucleic Acids Res., 44, D726–D732.

3. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330.

4. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P., et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res., 10.1101/gr.136184.111.

5. Diaz,A., Nellore,A. and Song,J.S. (2012) CHANCE: comprehensive software for quality control and validation of ChIP-seq data. Genome Biol., 13, R98.

6. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-Scale Quality Analysis of Published ChIP-seq Data. G3 Genes|Genomes|Genetics, 4, 209–223.