Integrative analysis of combinatorial chromatin interactions using high-throughput sequencing data

Yevhen Vainshtein[1,*] Karsten Rippe[2] and Vladimir B. Teif[3,*]

[1]Zentrum für Molekulare Biologie der Universität Heidelberg (ZMBH), DKFZ-ZMBH Alliance, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany;
[2]Deutsches Krebsforschungszentrum (DKFZ) & BioQuant, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany;
[3]School of Biological Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

*E-mails: y.vainshtein@zmbh.uni-heidelberg.de; vteif@essex.ac.uk

Keywords: MNase-seq, ChIP-seq, sequencing, nucleosome positioning

Abstract. Recent advancements in high-throughput sequencing experiments created an unprecedented situation when the experimental data accumulation outpaces computational analysis. While the basic analysis of each individual dataset is being conducted in the corresponding original publications, the analysis of new datasets in relation to already published data for the same cell type provides a computational problem, especially if the data are represented in significantly different formats. Here we develop a software suite called QCUMBER (Quality ClUster Maps BuildER), which allows quantitative analysis of diverse datasets based on the continuous genomic coverage concept (as opposed to peak calling algorithms which operate with discrete genomic regions). We are focusing on the interplay of nucleosome positioning, sequence-specific chromatin protein binding and DNA methylation, and describe typical workflows of data processing and integrative analysis taking into account information from diverse datasets, and discuss potential limitations and problems of such analyses.

1 Introduction

The basic bioinformatics view of chromatin is a mixture of myriads of molecules (nucleic acids, proteins, small ions, etc), sometimes characterized by 3D positions in the cell nucleus, but in most cases just the 1D genomic coordinate determining the location of a given nucleoprotein complex along the DNA. The most common nucleoprotein structure, the nucleosome, consists of eight histone proteins and 145-147 DNA base pairs (bp) wrapped around the histone octamer core. In the recent years high-throughput sequencing has become a standard way of analyzing the complexity of gene regulation in chromatin. It includes many experimental assays to map protein binding in chromatin, such as chromatin immunoprecipitation using an antibody spe-

cific for a given protein followed by sequencing (ChIP-seq) and related technologies to determine nucleosome positioning throughout the whole genome. In the latter case, antibodies are either not used, or used against core histones (e.g. H3 ChIP-seq). The most frequently used method for determining nucleosome positions is MNase-seq (chromatin digestion by micrococcal nuclease followed by sequencing). A number of complementary methods for nucleosome mapping have been proposed using MNase alone or in combination with sonication DNase (DNase-seq), transposase (ATAC-seq), CpG methyltransferase (NOME-seq), or directed chemical cleavage.

All methods listed above are based on the idea that chromatin can be cut into small fragments (with fragments being characterized by association of the DNA with a specific protein, or any protein, depending on the method) and then mapped back using the reference genome. The frequency of chromatin fragments mapped along the genome reflects the abundance of a given feature (a specific protein, or nonspecific DNA accessibility). Thus, the output of all these methods is by definition a continuous non-homogeneous distribution of protein binding along the DNA ("binding map"). Nevertheless, most existing analysis methods do not treat this as a continuous binding map, but rather as a discrete distribution of protein binding locations ("binding sites"). This is achieved with the help of peak calling methods. It is assumed that the majority of the signal is just a noise that can be disregarded, and only well-defined peaks reflect specific protein binding sites. The latter assumption is justified for mapping binding sites of transcription factors (TFs) and associating these with co-localizing proteins or genomic features such as promoters and enhancers. A number of generic computational tools has been developed to perform peak calling, including MACS/MACS2, HOMER, SICER, PeakSeq and CisGenome to name just a few. Furthermore, many specialized programs that perform peak calling to determine nucleosome positions exist, including TemplateFilter, NPC, nucleR, NOrMAL, PING/PING2, MLM, NucDe, NucleoFinder, ChIPseqR, NSeq, NucPosSimulator, NucHunter, iNPS and PuFFIN. However, in many cases the underlying biology is such that protein distribution along the DNA cannot be treated as discrete. This is the case of nonspecific protein binding, and is also applicable to the nucleosome distribution along the DNA. In this case, one operates with the continuous protein/nucleosome occupancy profile, defining regions of cell type/state specific differential occupancy (e.g. DANPOS/DANPOS2, DiNuP, NUCwave). In principle, the simplest approach to define regions of differential occupancy is to shift a window along the genome and count the number of reads at each window position. But what if we do neither peak calling, nor the differential occupancy region definition? Analyze the continuous genomic map exactly as it is produced by the sequencing is significantly more difficult, in particular, in the downstream analysis integrating this continuous binding map with discrete genomic features (promoters, enhancers, etc). Below we will consider two possible scenarios of integrative sequencing analysis: which we call "discrete" or "continuous" based on the discrete or continuous processing of chromatin binding maps. Our novel software package QCUMBER (Quality ClUster Maps BuildER) is mostly based on the continuous analysis and a mixture of these two methods. Unlike the discrete binding site analysis outlined above, the continuous

occupancy analysis does not discard the regions of low read density (which is essential in nucleosome positioning analysis). Instead, this type of analysis makes use of the large statistics and attempts to look at some averaged quantities, which characterize chromatin in different cell conditions at different genomic features.

2 Results and Discussion

The first issue in the continuous analysis is the normalization of the ChIP-seq signal. The strength of the ChIP-seq or MNase-seq signal critically depends on the quality of antibody, chromatin digestion conditions, sequencing depth and variations of the experimental protocols. Therefore, cross-platform comparison of datasets obtained in different laboratories provides a major challenge. Several solutions have been proposed in the literature, such as ChIPnorm, ChIP-Rx, NCIS, MACE and CisGenome. In QCUMBER, our normalization strategy depends on the biological situation. For example, when working with TF ChIP-seq, our normalization strategy is to do peak calling, determine common peaks, which are represented in all datasets, and normalize the datasets in such a way that the common peaks on average retain the same heights. After the normalization has been performed, one can carry out downstream analyses of the continuous occupancy profiles.

A very common type of analysis is the calculation of the coverage maps for many genomic regions aligned with respect to some common feature (e.g., the transcription start site, TSS, or the TF binding site), and then averaging them. Individual coverage maps can be combined in a heatmap, where each line represents a genomic region, and the ordering of the regions is performed according to some clustering algorithm such as GAGT or deepTools. Our software allows dissecting clusters of genomic regions which are characterized by a similar profile of ChIP-seq (MNase-seq, Ribosome-seq etc) density, then extracting the regions from these profiles and performing for them either the discrete analysis as above, or other types of continuous analysis. One example of such analysis could be to calculate differential TF binding at those genomic regions, and compare continuous profiles predicted by the theory with the experimental ChIP-seq data.

One of the tasks that are not addressed by available software packages is the integration of ChIP-seq and DNA methylation data (beyond the discrete analysis mentioned above, which simply deals with genomic coordinates of differentially methylated regions). The difficulty is that the precision of the DNA methylation data is 1 bp (as obtained e.g. with the help of bisulfite sequencing), while the precision of ChIP-seq or MNase-seq is usually much worse. QCUMBER provides a possibility to deal with all individual methylated (or unmethylated CpGs). The reverse task is also possible: one can calculate the density of DNA methylation around any genomic feature [1]. DNA methylation positions obtained from standard methylation callers such as Bismark can be converted into intermediate files with the continuous DNA methylation coverage in analogy with ChIP-seq output, thus making these datasets directly comparable.

Another integral parameter of chromatin, which changes upon cell treatment or during cell differentiation, is the Nucleosome Repeat Length (NRL). Our program allows calculating NRL based on ChIP-seq or MNase-seq data and comparing it between different cell conditions or between different types of genomic regions for the same cell type [2, 3]. These differences are usually quite small (from 1 up to 15 bp), and therefore a method needs to take into account the error of the NRL determination.

One of the novel points addressed by QCUMBER is the comparison between large datasets, sometimes obtained in different laboratories for the same cell type. For example, about 14 datasets exist where a single method, MNase-seq, was used to determine genome-wide nucleosome positioning in a single cell type, ESC, reported by about 10 different laboratories including ours [4]. Nucleosome positions derived from these datasets would overlap only partially. Thus, a discrete type of analysis of these data would mostly fail. Concerning the continuous analysis, the main question is how to deal with the fact that the occupancy profiles from these 14 datasets are very different, yet they all reflect the same underlying biology? To address this question, we have developed a window-based algorithm, which, after the normalization of individual datasets, compares the relative number of reads per regions in each dataset. As a result, genomic regions can be sorted based on the similarities or differences of their representation in ChIP/MNase-seq in different datasets. Divergently represented regions are called "fuzzy", and are further analyzed using basic discrete analysis described above, assuming that these regions undergo active chromatin redistributions (TF binding, nucleosome remodeling, etc).

The software will be made available online upon final submission of the manuscript.

3 Acknowledgements

This work was partially supported by the DKFZ grant "Developing a software suite for the analysis of epigenetic regulation from high-throughput sequencing data".

4 References

1. Teif VB, Beshnova DA, Vainshtein Y et al. Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development, Genome Research 2014;24:1285-1295.
2. Beshnova DA, Cherstvy AG, Vainshtein Y et al. Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions, PLoS Comput Biol 2014;10:e1003698.
3. Teif VB, Vainstein E, Marth K et al. Genome-wide nucleosome positioning during embryonic stem cell development, Nat Struct Mol Biol 2012;19:1185-1192.
4. Teif VB. Nucleosome positioning: resources and tools online, Brief Bioinform 2016;In press.