Bioinformatics pipeline for ChIP-seq analyses

Miklós Laczik, Jan Hendrickx, Céline Sabatel, Irina Panteleeva, Hélène Pendeville, Dominique Poncelet

Epigenetics R&D Department, Diagenode SA, Liège, Belgium

Diagenode sa, LIEGE Science Park, Rue du Bois Saint-Jean, 3 4102 Seraing Belgium

Introduction

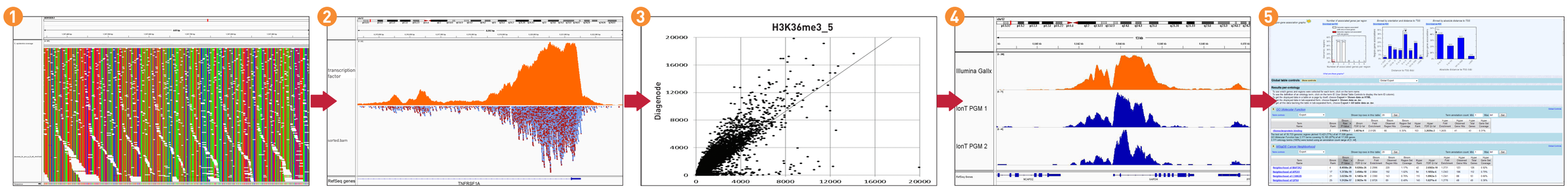
As the role of the changes in the epigenome is becoming more and more revealed there are a growing number of scientific papers reporting epigenetic/epigenomic studies. Malicious epigenomic alterations can be the reason behind widespread social problems like different types of cancers (eg. breast cancer), as well as rare diseases like ICF (Immunodeficiency, Centromeric region instability, Facial anomalies) syndrome, the like of which often not interest the medical industry therefore the diagnostic/therapeutic products are underdeveloped or non-existent at all. Such chromatin diseases (Rett syndrome, ATRX syndrome, FSHD, ICF syndrome) are the subject of the DisChrom project (FP7 project under the grant agreement PTN-GA-2009-238242) in which we participate.

Chromatin ImmunoPrecipitation (ChIP) coupled with high-throughput massively parallel sequencing as a detection method (ChIP-seq) has become one of the primary methods for epigenomic researches, namely to investigate protein-DNA interaction on a genome-wide scale. While ChIP-seq applications and the subsequent data analysis is well described for transcription factor studies, the histone modifications are far less documented: we need to set up standards and find or create the appropriate bioinformatics tools to analyse such data. It is especially important since histone modifications and the correlated chromatin structure changes could be responsible for the development of chromatin diseases like the ones mentioned above.

The principal work in our laboratory involves antibody production and quality control, and many of the candidates are also tested by ChIP-seq. We mainly focus on histone marks and besides the antibody production we regularly obtain ChIP-seq data from other projects , either from outsourced sequencing or from our own IonTorrent PGM sequencer. Thus we deal with a lot of ChIP-seq data on a regular basis, coming from various sequencing platforms and a large range of antibodies. With this heavy demand it is natural that we need an automated bioinformatics solution that is able to cope with the deluge of ChIP-seq data.

In this poster we will present a modular bioinformatics pipeline which was developed specifically for the analysis of ChIP-seq data, but can be easily modified and extended to accommodate other high-throughput sequencing and epigenetic applications, such as RNA-seq or methylation analysis and other custom data processes. We will introduce the modules and the workflow as well as the standards we set up for data quality (based on requirements of the scientific community and our own experience and research), and we will demonstrate the usage and the potential of it analysing actual ChIP-seq data from our latest experiments.

Criteria system and pipeline structure



We first check the quality of the sequencing, using the manufacturer provided software of the sequencer or a dedicated software. Then we check the peak qualities (numbers, size, significance, read fraction in peaks, genomic feature association), and compare them to reference sets and/or replicates. For overlap analysis we use the Encode guideline: we expect at least 80% of the top 40% of the peaks to have a match in the other dataset. Note that the expected value ranges and characteristics depend heavily on the histone mark (or transcription factor) in question (and of course other sample features, like the organism, cell/tissue type, treatment etc.).

Our pipeline has five core modules that cover the whole process of ChIP-seq analysis, with the option of adding further modules for custom analyses. It takes advantage of configuration files, which enable you to save and load your pipeline with your custom settings and modules.

1. Alignment module: this module starts with read files and align the reads to a reference genome (must be indexed previously). It utilizes the BWA aligner, which is a fast tool optimizable to either short or long NGS reads, which is important for us, since we use both Illumina sequencers and the IonTorrent PGM. It is also possible to skip this module if you start the pipeline with alignment files (in SAM, BAM or BED format).

Optional module: it is possible to run the command line version of FastQC automatically after the alignment to assess the read quality.

2. Peak calling module: here we utilize our own scripts and scripts from the Sicer software package, which is a genome partitioning peak caller developed specifically to detect (even long) histone marks. At this stage it is very important to use the correct settings, which are predefined for the most frequently used histone marks (but can be freely adjusted if the dataset needs it).

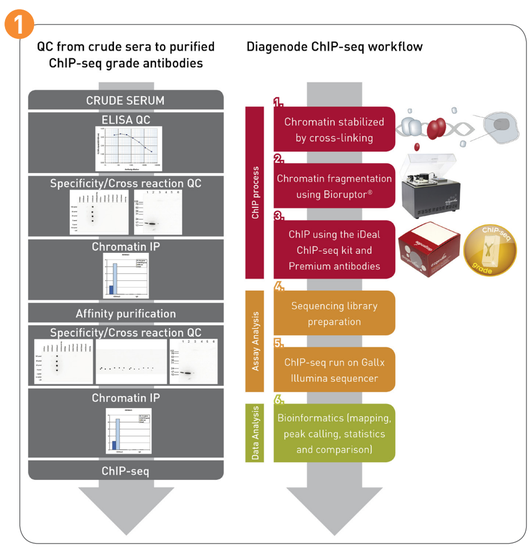
Optional module: instead of Sicer, MACS can be loaded as the second module. It is preferable for short enrichments, like transcription factors.

3. Statistics module: after the peak calling we use custom scripts and scripts from the Homer and BedTools software packages to describe the dataset statistically. We count the reads, the duplicate read ratio, the number of peaks, the reads that are in peaks and their ratio against the background (signal-to-noise ratio), we measure the average size of peaks and their average significance, and the genome coverage. Also we do a general annotation to identify whether the peaks have an affinity to a genomic feature.

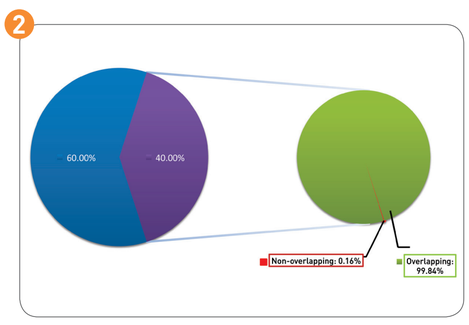
4. Comparison module: the aim of this module is to compare the analysed peak set to one or more reference datasets, utilizing Homer and BedTools, as well as custom scripts. The reference datasets could be the predefined Encode/modEncode datasets, or any other custom datasets, replicates. The overlaps are checked in terms of the total peak set as well as the top40% of the peaks.

Optional modules for 3. and 4.: because the aim of the analysis can differ from project to project, and we cannot prepare for every possible case, there are no predefined optional modules; although it is easy to add a custom script for further analyses (eg. the annotation in module 3. can be exploited to perform a pathway analysis). Note however that the results of these custom modifications won't be involved in the final report, unless you modify module 5. as well, so the best practice is to prepare your scripts to produce their own reports too.

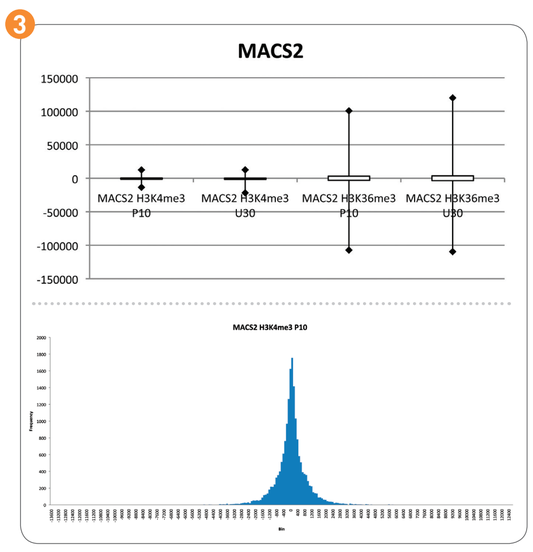
5. Reporting module: this is the final module, designed to retrieve all the necessary information from the outputs of the previous modules and create a report file in XLS spreadsheet format, which is user friendly and easy to review and understand. Naturally you can open it in eg. Excel of Microsoft Office or Calc of OpenOffice to produce further charts, tables and analyses.



1. We successfully use the pipeline on a weekly basis in the assessment and comparison of the qualities of antibody candidates. This is crucial to produce high quality antibodies that are proven to be effective for ChIP-seq. The picture shows the antibody validating process.



2. We also use the pipeline in various research and development projects. With this pipeline we managed to prove that our proprietary ChIP-seq kit and protocol for extremely low amount of cells is able to produce high quality results from as low starting amount as only a few picograms.



3. The flexibility of the pipeline enables us to use it for a wide range of analyses. The histogram and box-and-whiskers plot are excerpts from the report file of a study where we examined the ChIP-seq results generated with a novel shearing method, to see if we could manage to produce even fragment and read distribution.

Future directions

The pipeline is still in an alpha phase, it needs testing, bugfixing, refactoring and some minor developments. Later we plan to make it publicly available either as a Galaxy pipeline or as an R package, to make it accessible and user friendly. Also we are always trying the improve the pipeline and we are actively looking for further quality control methods to implement,such as the IDR software tool which estimates the consistency of peaks between replicates (and pseudo-replicates).

For more information, please contact dominique.poncelet@diagenode.com • For Europe, contact Orders@diagenode.com • For the US and Canada, contact Orders.NA@diagenode.com