Performing chromatin immunoprecipitation with minimal hands-on time using the ChIPettor, a novel semi-automated system, validated by ChIP-seq and rigorous bioinformatics analysis

Introduction

ChIP-seq is an excellent method for studying protein-DNA interactions, providing a genome-wide profile versus the limited insights from ChIP-PCR; it is also useful as it addresses both important biological or technical questions such as how chromatin structure changes due to differential histone modifications, or, for example, it can help validate the specificity and sensitivity of an antibody or monitor the bias of an NGS library preparation kit. Routine ChIP-seq experiments and rigorous bioinformatics analysis are an integral part of Diagenode's product development and diligent product quality control. In this application note we present our comprehensive analysis pipeline for analyses of ChIP-seq data. We then show a case study where we apply the introduced stringent analysis methods for validating a novel semi-automated ChIP method, providing an elegant and viable alternative to more costly automation systems.

Methods

In general every NGS data analysis workflow can be split into three major parts: primary, secondary and tertiary analysis.

The primary analysis involves the processing and QC of the raw read signals (e.g. base calling, filtering ambiguous signals, etc.), which are either a light signal or electric signal for most types of sequencing machines. The primary analysis is usually done automatically by the built-in software of the sequencer, although some algorithms have been developed for some parts of the workflow like improving the base calling over the standard base caller provided by Illumina. As we have found the differences negligible, we use the default base caller outputs. At the end of the primary analysis the read files are generated.

The secondary analysis consists mostly of alignment, where reads are mapped to a reference genome. In addition for ChIP-seq, peak calling is also performed subsequently. Thus secondary analysis yields alignment files and the detected peaks: sites where enrichments occurred.

The tertiary analysis is sometimes referred as "making sense" as it is usually during this step where researchers try to answer their specific scientific questions. Thus, by nature tertiary analysis methods are diverse and specific, can be different for each study. The aim of this study was to validate Diagenode ChIP tools and methods by analyzing ChIP-seq data with rigorous standards and by using comparisons, annotations, statistical tests and other stringent bioinformatics methods.

Each phase of the analysis workflow contains various steps for data processing and QC. The following section provides a step-by-step guide, starting with the secondary analysis, from processing the read files.

1. Alignment

For mapping the reads we usually use BWA, although we tried out other tools as well (e.g. ELAND, TMAP, Bowtie). Usually we have an abundance of reads, so the goal is not to align as many reads as possible, it is more important to avoid introducing bias by misalignments. Default settings serve this purpose well, they are stringent enough to allow only 1 or 2 mismatches for most aligners.

2. Controlling the read quality

We use FastQC for a reliable QC for reads: it is very informative as it reports the general base quality, read length distribu-

tions, GC content, adapter contamination and duplicate read level among others. It can be used either with reads or alignments; we use it with the alignment file, which gives additional information.

3. Cross correlation analysis

We implemented the method described in the ENCODE guidelines. To summarize briefly, this method slides the positive and negative strand reads along the genome. Theoretically the sliding reads should meet at the size of the fragment length (i.e. an accumulation should occur in a distance roughly equal to the average length of fragments). Using this accumulation point the general enrichment level can be characterized. This method usually works well with short and uniform peaks (e.g. transcription factor ChIP-seq data), but produces highly questionable results for long and diverse enrichments, like histone marks. However, we found that with some shorter histone marks like H3K4me3, the method also usually gives reliable results.

4. Peak calling

The peak calling and its settings are crucial to obtain correct results. Inappropriate settings can easily lead to false conclusions. Therefore we pay extreme attention for using and adjusting peak callers. Every dataset can be different, so finding the optimal settings can be an arduous trial and error procedure. As starting points we have pre-defined settings for each histone mark/transcription factor we use, but in many cases the special characteristics of the dataset require unique settings. Usually we use MACS2 for short peaks (like most transcription factors), Sicer for long peaks (like H3K36me3) and an in-house developed method for ubiquitous peaks (like H3K9me3).

The subsequent methods are all part of the tertiary analysis.

5. Characterization of datasets and annotation

The tertiary analysis involves custom procedures that we use to control the quality of ChIP-seq datasets and compare them to each other. Below we describe the most critical parameters. These figures should always be interpreted in context, not individually. For example, only comparing the read numbers cannot inform which sample is better, despite the conception that more reads should yield more and better peaks. However if the sample with a higher read number has a lower peak number with less enriched peaks (shown by the lower scores) and the reads-in-peaks ratio is also lower, then certainly that sample is the one of lower quality.

- Read number, unique read number, duplicate read ratio: important information about library size and complexity
- Ratio of positive and negative strand reads: ideally the same number of reads comes from both strands
- Reads in peaks, genome coverage ratio: give a general idea about the enrichment profile
- Peak number, average peak width, average peak probability score: characterize the peaks, their abundance and their average dimensions, and the signal-to-noise ratio
- Average peak profile: shows the average read counts across the peak length (given in percentage); used to monitor and compare the general enrichment level and peak shape (the latter can be characteristic of certain transcription factors / histone marks)
- Annotation: certain histone marks (or transcription factors) have an affinity to certain genomic features, e.g. H3K4me3 peaks are dominantly associated with promoters; with annotation we can monitor these associations
- Visual inspection: the peak / alignment files are loaded into a viewer and the quality of the enrichments are checked visually at selected control regions, in addition to the above statistics

6. Overlap analysis

When we assess the quality of a ChIP-seq dataset, the most important step is perhaps the comparison to a reference dataset, and not just by checking the above mentioned parameters, but also by determining whether the peak positions match in the two datasets. This truly indicates success in reproducing the expected peaks. Choosing the right reference dataset is the most crucial part of this analysis – we take extreme care to avoid unreliable datasets and subsequent false conclusions. Below we describe the comparison analyses:

- Overlap matrix: for every peak of each sample we show how many of the peaks overlap with other peak(s) in the other sample(s), i.e. how many peaks have a match
- Top40% overlap analysis: similar to the previous matrix, but inspecting how many of the best 40% of the peaks have a match; this metrics was also adapted from the ENCODE guidelines, which require at least an 80% match
- Correlation analysis: to control the consistency of the matching datasets we also calculate the coefficient of correlation between overlapping peaks, to see if the matching peaks are also similar in size; in this way we can control random matches with false peaks
- Average overlap profile: a graph is created showing the average shift between overlapping peaks, with this we can control if the peaks match centrally (or e.g. only their ends overlap)
- Bias control: similar to the above analysis, we control the distribution of peaks if they cover the 5' or 3' end of the reference peaks more; ideally there should be no bias, peaks should overlap centrally each other

Results

The ChIPettor system

Diagenode has developed a novel system for chromatin immunoprecipitation based on a programmable multichannel pipette. Automation or semi-automation of ChIP allows for high reproducibility by minimizing human error with minimal hands-on time and convenient processing of large numbers of samples. This unique ChIPettor System contains reagents for start-to-finish histone or transcription factor ChIP including controls, purification reagents, and a semi-automated pipettor with special resin pipette tips. The ChIPettor is designed to stand independently on a 96-deepwell plate while it automatically dispenses, pipettes, and mixes reagents using a pre-programmed ChIP protocol.

Validating the ChIPettor

In order to control the quality of the ChIPettor system we performed three ChIP-seq experiments on HeLa cells: two with antibodies against H3K36me3 and one with antibody against the H3K4me3 histone mark (plus an input control was also sequenced). Then we performed the analyses described above and compared our samples to the respective datasets of the ENCODE project, produced by the Broad Institute. The reads were mapped to the hg19 genome version with BWA, and Sicer was used for peak calling with the appropriate settings for the different histone marks. For the visual inspection we used the IGV software from the Broad Institute.

The results showed high quality and excellent consistency. The ChIP-seq results proved to be reproducible, and in some parameters outperformed the datasets of the Broad Institute. Below we describe the highlights.

1. Sequencing and mapping quality

The FastQC reports show excellent base quality (Figure 1) and read quality. The duplicate levels are higher than for the ENCODE datasets but still acceptable.

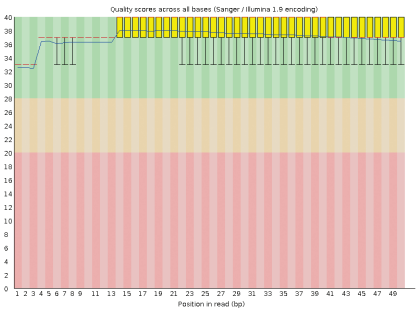

Figure 1

2. Descriptive statistics

As Table 1 shows we have higher duplicate ratios, therefore we have similar (H3K4me3) or less (H3K36me3) unique reads than the ENCODE datasets.

Table 1	H3K4me3	H3K36me3_1	H3K36me3_2	Broad H3K4me3	Broad H3K36me3
Total reads	50985396	46445119	60862159	35897578	60030600
Positive reads	25492691	23221540	30440149	17943221	30012039
Negative reads	25492705	23223579	30422010	17954357	30018561
Unique reads	31392576	29358920	38475441	32853368	57334020
Duplicates	19592820	17086199	22386718	3044210	2696580
Duplicate %	38,43%	36,79%	36,78%	8,48%	4,49%
Reads in peaks	13436448	16918757	23542681	12934460	22523693
RIP %	42,80%	57,63%	61,19%	39,37%	39,29%
Total peaks	20050	10695	10316	28615	10800
Average score	42,78	36,13	36,64	42,38	33,40
Average peak width	3422,16	41849,77	45206,25	3599,58	40670,02

Despite this we achieved high quality enrichments: for the H3K4me3 we detected somewhat less peaks than ENCODE, but the average score, size and reads-in-peaks ratio is similar, while for H3K36me3 we clearly outperform the ENCODE results, the better enrichments and signal-to-noise ratios are evident if you compare the similar peak numbers and sizes and the higher probability scores and reads-in-peaks ratios that we have.

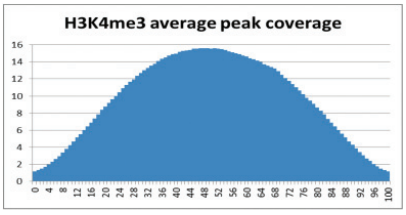

Figure 2A

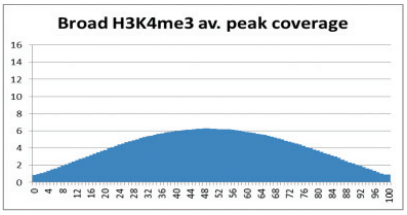

Figure 2B

3. Average enrichment profile and read distribution

The average peak profile graphs (Figure 2A: Diagenode data, 2B: Broad Institute data) also show a generally better enrichment for Diagenode samples. The ratios of positive/negative strand reads are equal as expected.

4. Overlap matrices

In the matrices you can observe the outstanding reproducibility and consistency between the datasets for both histone marks (Table 2A: H3K4me3 data, 2B: H3K36me3 data). The required 80% for the top 40% of peaks is surpassed by a great extent, for our H3K36me3 datasets we achieved 100% overlap.

K4	Overlap w/ H3K4me3	%	Overlap w/ Broad H3K4me3	%
Peaks of H3K4me3	20050	100,00%	17031	84,94%
Top40 of H3K4me3	8020	100,00%	7929	98,87%
Peaks of Broad H3K4me3	16855	58,90%	28615	100,00%
Top40 of Broad H3K4me3	11113	97,09%	11446	100,00%

Table 2A

K36	overlap w/ H3K36me3_1	%	overlap w/ H3K36me3_2	%	overlap w/ Broad H3K36me3	%
Peaks of H3K36me3_1	10695	100,00%	10385	97,10%	8954	83,72%
Top40 of H3K36me3_1	4278	100,00%	4278	100,00%	4175	97,59%
Peaks of H3K36me3_2	9730	94,32%	10316	100,00%	8599	83,36%
Top40 of H3K36me3_2	4126	100,00%	4126	100,00%	4029	97,65%
Peaks of Broad H3K36me3	9424	87,26%	9535	88,29%	10800	100,00%
Top40 of Broad H3K36me3	4274	98,94%	4278	99,03%	4320	100,00%

Table 2B

5. Correlation analysis

The scatterplots (Figure 3: Correlation of Diagenode peaks (y axis) and Broad Institute peaks (x axis) for H3K4me3) and the Spearman's correlation coefficients close to 1 proves that consistency is also achieved in regards to the size of overlapping peaks.

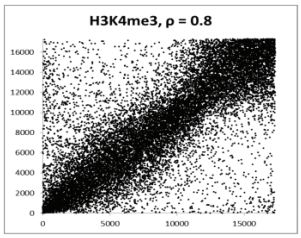

Figure 3

7. Bias control

The graphic display (Figure 4A: Average overlap profile for one of the H3K36me3 samples) and the box-and-whiskers plots (Figure 4B: Distribution of position shifts of overlapping peaks of Diagenode and Broad Institute datasets) show no bias, and the matching peaks overlap each other centrally.

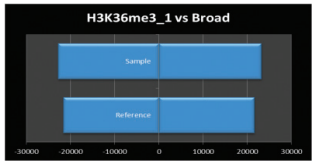

Figure 4A

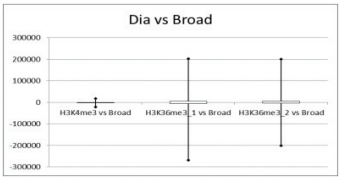

Figure 4B

8. Visual inspection

The genome viewer tool visualizes the excellent overlaps and high quality enrichments. Figure 5A: A close view of the gene GAPDH in the H3k4me3 datasets. Figure 5B: A more distant view to compare the H3K36me3 datasets. The blue graph is the Diagenode data, the red one is the Broad Institute data.

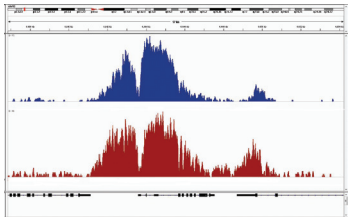

Figure 5A

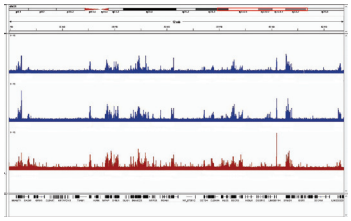

Figure 5B

Discussion

We have described the ChIPettor System, a semi-automated chromatin immunoprecipitation solution, and the rigorous bioinformatics and stringent ChIP-seq QC criteria applied to optimize this system. Our validation shows that the ChIPettor System is indeed capable of producing high quality ChIP-seq results with excellent signal-to-noise ratios, consistency and reproducibility. In many instances, the generated data outperformed the ENCODE datasets. The ChIPettor also requires only minimal preparation and hands-on time. Ultimately it is a true alternative to more costly robotic automation systems.