Wavelet-based processing of ChIP-Seq signals

M.P. Orihuela Martínez-Costa[1], M. Braga-Monteiro[2], A. Muñoz-Barrutia[3], V. Segura[2]

[1] Departamento de Ingeniería Biomédica, Universidad de Navarra (TECNUN), San Sebastián, Spain
a108372@alumni.tecnun.es

[2] Unidad de Proteómica, Genómica y Bioinformática and [3] Laboratorio de Imagen del Cáncer, Centro de Investigación Médica Aplicada, Universidad de Navarra, Pamplona, Spain

{mbmonteiro, arrmunoz, vsegura,mbmonteiro}@unav.es

Abstract

Chromatin Immunoprecipitation followed by sequencing (ChIP-Seq) is a powerful technology that enables genome-wide detection of epigenetic phenomena (i.e., histone modifications). The analysis of these experiments involves the use of complex computational methods ranging from the data acquisition to the functional analysis. The challenge for proper peak detection is to distinguish enriched regions from noise. In this work, we have implemented three established peak calling methods and proposed a new one based on wavelets. The quantitative evaluation of the methods performance against an expert annotation shows an improved specificity and sensitivity of the proposed method.

1. Introduction

Sequencing technology has progressed far beyond the analysis of DNA sequences. In particular, this technology is routinely used today to analyze other biological components such as RNA and protein sequences, as well as their interaction in complex networks. Furthermore, Next Generation Sequencing (NGS) technologies can quantify chromatin features, locate DNA modifications and identify a number of steps in the cascade of information that goes from transcription to translation. Moreover, these technologies are key tools to perform health-related discoveries such as the regulatory mechanisms and expression profiles distinguishing non-tumoral from malignant cells [1].

The process of sequencing DNA starts with the fragmentation of the DNA sample into a library of small segments. The identified strings of bases, called reads, are then reassembled using a known reference genome as a scaffold. The raw data from high-throughput sequencing experiments are images obtained as the output from the next generation sequencing platform. A base caller converts the image data to sequence tags, which are then aligned to the genome.

Genome-wide mapping of protein-DNA interactions and epigenetic marks is essential for a full understanding of transcriptional regulation. The main tool for investigating these mechanisms is chromatin immunoprecipitation (ChIP). In ChIP, antibodies are used to enrich DNA fragments bound to the recognized proteins. In a ChIP-Seq assay two signals are obtained: the enriched immunoprecipitated signal and the non-enriched input signal or control.

The analysis of ChIP-Seq experiments involves the use of complex computational methods. The process goes all the way from data acquisition to the functional interpretation (Figure1). After mapping the reads to the reference genome, a peak detection algorithm is needed to identify enriched regions [2]. The peak detection algorithm either ranks the detected regions by their absolute signal or by the statistical significance of the enrichment in order to detect significant peaks.

In this manuscript, we focus in peak calling on histone modification experiments. The characteristics of histone modification profiles are diverse ranging from sharp well-defined peaks surrounding the genome transcription start sites to broad diffuse marks on large genomic regions. This inherent variability difficulties the differentiation of true enriched regions from background noise. To address this issue, a group of 'spectral' methods capture the shape of the histone modification and identify the potential enriched regions in the frequency domain, while others rely in 'matched filtering' to identify the peaks on the genomic domain. In this manuscript, we present a novel wavelet-based method belonging to the second class and demonstrate its suitability for the analysis of this type of signals.

2. Materials and Methods

2.1. Datasets

The algorithms implemented have been applied to a ChIP-Seq experiment for histone modification analysis (H3K4me3) of K562 cell line from the ENCODE project. The dataset was downloaded from Gene Expression Omnibus (GEO) database (accession number GSM733708) [3].

2.2. Analysis of ChIP-Seq data

As shown in Figure 1, the ChIP-Seq experiments can be broadly divided in biological sample preparation and analysis. In this section, we focus in the description of the steps involve in the analysis, in particular, normalization, peak detection and annotation. Functional analysis is a critical step but no part of the current discussion.

2.2.1. Normalización

Before the peak detection, both ChIP and input signals are normalized [4] to make them comparable. Then, the ChIP

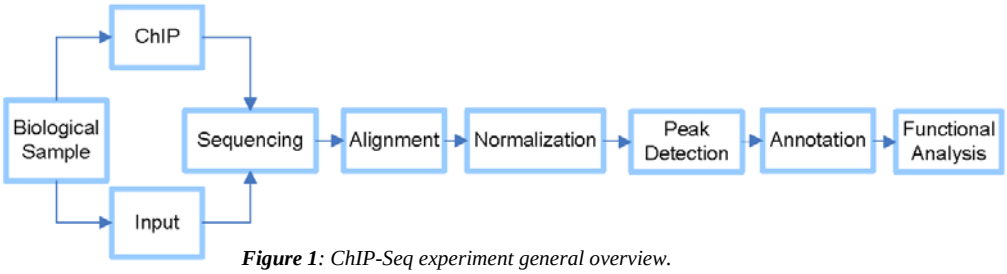

***Figure 1**: ChIP-Seq experiment general overview.*

and input signals are divided into non-overlapping windows of a fixed width and read counts are calculated for each window.

2.2.2. Peak detection

In this manuscript, we have implemented four different methods devoted to solve the ChIP-Seq peak calling challenge. Two of them belong to (1) the 'spectral' group: the Model-based Analysis of ChIP-Seq data (MACS) [5] and the modified WaveSeq [6]; and the other two (2) to the 'matched filtering' group: PeakDetection [7] and the proposed Zero Crossing Lines (ZCL) detection which was inspired on an algorithm for the study of transcriptional activity on tiling microarray data [8].

As discussed above, ChIP-Seq data is composed by a mixture of peaks with diverse characteristics (pike-like vs. wide and smooth embedded in a noisy background), making this type of data well suited for multi-resolution analysis [6]. For this reason, three of the implemented methods (i.e., WaveSeq, PeakDetection and ZCL) rely in the computation of the continuous wavelet transform (CWT).

The CWT is defined as the convolution of a continuous signal s(x) with a translated and scaled mother wavelet ψ(x) as given by [6]:

$$\mathrm{CWT}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} s(x) \psi^* \left(\frac{b-x}{a} \right) dx,$$

where a is the scale parameter, b is the translation parameter, ψ(x) is the mother wavelet, ψ(x)*((b-x)/a) is the complex conjugated, scaled and translated wavelet and CWT is the 2D matrix of wavelet coefficients. The wavelet decomposition produces a series of real coefficients that measure the correlation between the mother wavelet and the signal at a given position and for a given scale.

Model-based Analysis of ChIP-seq data

In MACS, peaks are marked as candidate when two conditions are satisfied: the number of reads in the ChIP signal is greater than a user-defined threshold and simultaneously, greater than twice the normalized read counts in the input signal for that given location.

The shape parameter (i.e., lambda) of a Poisson distribution indicates the average number of events in a given time interval. In MACS, instead of using a unique value estimated from the whole genome, a shape parameter is defined for each candidate peak. The Poisson test is then applied and only peaks with a statistically significant p-value ($p<0.01$) are considered true peaks. In order to control the number of false positive peaks, we perform a False Discovery Rate (FDR) analysis following the guidelines given in [9].

Finally, a clustering algorithm is applied to the remaining candidate peaks. All peaks separated a distance below 1000 base pairs are aggregated together defining the final enriched regions. As ChIP-seq experiments often present broad peaks, the algorithm typically clusters several regions into a single peak.

Modified WaveSeq algorithm

The WaveSeq algorithm we have implemented, is a modification of the one presented in [6] eliminating the use of Montecarlo sampling for the selection of candidate peaks. The implemented algorithm is briefly described next. The CWT is computed using the Morlet wavelet as mother wavelet. Next, the spectrum is estimated from the computed wavelet coefficients as the square of their absolute value divided by the variance of the signal. The regions with the highest spectrum values (belonging to the fifth percentile) are identified as candidate peaks. The Binomial distribution is used to statistically include them as potentially enriched regions. A p-value smaller than 0.01 is considered statistically significant. A FDR is then applied and finally, peaks separated by a user-defined number of windows with no significant wavelet power are clustered together.

PeakDetection algorithm

In the PeakDetection algorithm, the CWT of both ChIP and input data sets is computed using the second derivative of a Gaussian as mother wavelet. The Signal to Noise Ratio is estimated as the ratio between the square of the ChIP signal wavelet coefficients and those of the input signal [7]. If this value is greater than an empirically defined threshold the region is considered as a candidate peak. Finally, the regions considered as enriched are the output of the candidate peak clustering step defined as for the WaveSeq algorithm.

Zero-Crossing Lines Detection

As the modified WaveSeq and PeakDetection, the detection of the enriched regions in the ZCL is based on the computation of the CWT. In this case, the first derivative of the Gaussian function is used as mother wavelet. Then, the peaks locations are identified as the zero-crossing lines of the CWT decomposition. Those are obtained by connecting the points across which the Gaussian derivative changes sign. In particular, in this case, it corresponds to the peak position.

After zero-crossing lines calculation only those lines with a length greater than a pre-defined threshold are

considered as putative peaks. Next, a signal filtering is performed to eliminate low abundance peak candidates. Then, a peak expansion algorithm is applied to improve the definition of the peaks (i.e., locate the start and end sites). Namely, if the candidate peak is surrounded by regions with a number of reads over a user defined threshold the peak expands its width. Next, the statistical analysis using the Poisson distribution ($p<0.01$) and FDR correction is performed. Finally, the enriched regions are identified as those resulting from clustering together those candidate regions separated by one window.

As a distinguish feature, this algorithm can detect enriched region for experiments with or without an input sample. For experiments with control data, input read counts are used to estimate the parameters of the distribution. For experiments without it, an estimation of the ChIP background is calculated as the mean intensity of a sample region where no candidate peaks are found.

2.2.3. Annotation

The output from the peak calling algorithms is a list of peaks that are significantly enriched. The annotation of these peaks to the genes of the reference genome has been carried out using the Bioconductor package ChIPpeakAnno. Only genes with peaks located in their promoter region are considered for further analysis. The output of the analysis pipeline is a file in Browser Extensible Data (BED) format, which can be easily visualized using a genomic browser (for example. The Integrated Genomic Viewer, IGV [10])

3. Results and Discussion

In order to compare the peaks and annotated genes obtained using the algorithms previously described, we have analyzed the chromosome 13 of the ChIP-Seq experiment for the study of H3K4me3 histone modification in K562 cell line. We consider it to be a good representative as its length and the number of coding genes present is on the average range of the human chromosomes.

Evaluation of results

The algorithms were implemented in R and executed using default parameters defined for each method. The annotation of peaks was performed using the human gene code version 19 as reference.

A robust comparison of the methods is hard to accomplish due to the lack of a validated dataset that can be considered as the ground-truth. To face this issue, an expert has performed manual peak detection. The expert carried out two annotations: non-restricted annotation, which includes all the observed peaks, and restricted annotation which eliminates all the uncertain peaks.

The visualization of the results in IGV reveals important aspects of each method's performance (see Figure 2). MACS detects with high precision narrow but well-defined peaks. Nevertheless, the method misses some true peaks reducing its sensitivity. WaveSeq identifies more true peaks than MACS and interestingly, their width is accurately estimated. However, the number of false positive peaks is higher than for MACS. PeakDetection achieves a fairly good performance for the peaks location, but overestimates their width. ZCL shows robust peak location detection with an accurate width estimation of both broad and narrow peaks.

Figure 3 shows the intersections between the gene lists obtained by the annotation of the peaks detected using all the computational methods and the two expert annotations. The maximum overlap with the expert's gene annotation is achieved by the ZCL algorithm (238 annotated genes of which 218 are in the expert list). Interestingly, although the number of genes detected using MACS is lower (110 genes) almost all of them are included in the list provided by the expert (94 genes).

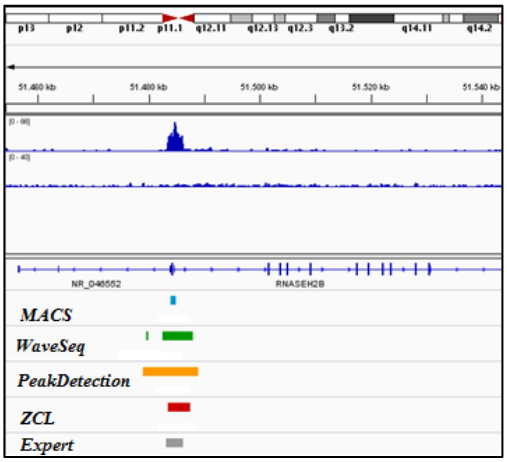

Figure 2. *Example of peak detection performed by each method for chromosome 13. All methods detect the peak, but there significant difference in the width estimation. MACS detects a very narrow peak; WaveSeq and PeakDetection, a much broad peak while ZCL adjust better to the real peak's width. The latter gives the most similar outcome to the expert's annotation.*

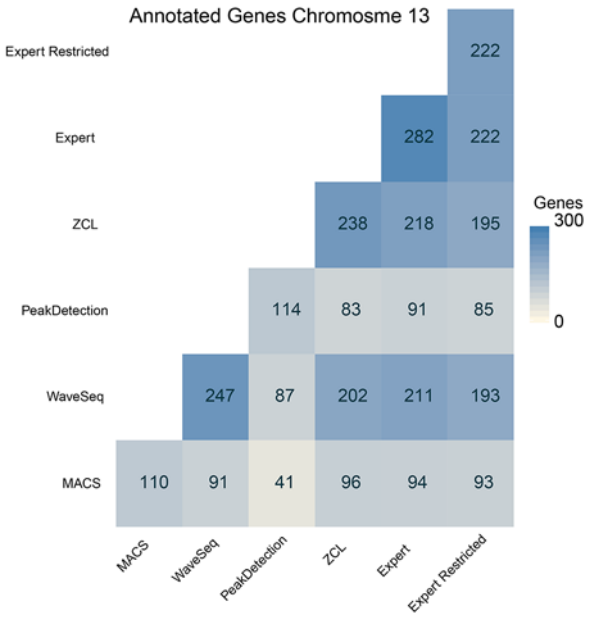

Figure 3. *Coincidence of genes between methods for chromosome 13. The method that shows more overlap with the expert's annotation is ZCL. PeakDetection performs a very poor annotation with not so many coincidences with the expert's annotation.*

We have calculated the sensitivity and the specificity of the algorithms using these results. The best performance metrics are obtained by MACS and ZCL (see Table 1). The proposed method is the most sensitive considering both expert annotations. MACS has also a very high specificity, but its sensitivity is very poor.

	Method	Sensitivity	Specificity
Expert	MACS	0.33	**0.98**
	WaveSeq	0.74	0.96
	PeakDetection	0.32	0.97
	ZCL	**0.77**	**0.98**
Expert Restricted	MACS	0.4	**0.98**
	WaveSeq	0.86	0.94
	PeakDetection	0.62	0.97
	ZCL	**0.88**	**0.96**

Table 1. *Sensitivity and specificity of the annotated genes for the four implemented methods with respect to both expert annotations.*

Both MACS and ZCL are well suited for the functional analysis of annotated genes (i.e., transcription factor analysis). However, other applications (i.e., comparison of ChIP signal intensity among different experimental conditions, sequence motif detection) require accurate peak area quantification and size determination. For those applications, ZCL would be the method of choice.

Computational performance

ChIP-Seq experiments were executed in an Intel Xeon® processor server (64 bits, 4 cores, 2 GHz) with 32 Gb installed memory running Red Hat Enterprise Linux AS release 4 and R 2.15.0. The mean time to execute the peak detection algorithms varies between 5 and 15 min.

4. Conclusions and further work

In conclusion, all the implemented methods perform fair peak detection. In despite of that fact, in our hands, MACS and ZCL present the highest specificity while ZCL achieves the highest sensitivity. MACS gives a precise gene annotation which is very useful for functional analysis. ZCL excels also on this but on addition, performs an accurate peak width estimation which is well-suited for other applications such as comparison of experiments, or motif detection.

All the algorithms have been implemented to be executed in a High Performance Computing (HPC) platform. In particular, the implementation of the CWT is very efficient as multi-scale wavelet-based calculations are especially convenient for parallel computing.

As further work, the functional analysis of the annotated genes could be carried out. Such analysis will provide key information about how a given histone modification affects the gene expression. The systems biology applications that such a study would entail are very numerous. We could imagine to construct an overview of all histone modification profiles existent in the ChIP-Seq data of the ENCODE project.

Acknowledgments

We would like to thank the Proteomics, Genomics and Bioinformatics Unit of CIMA, especially to Elizabeth Guruceaga for technical support.

References

[1] Soon, W. W., Hariharan, M. and Snyder, M. P. High-throughput sequencing for biology and medicine. Mol Syst Biol, Department of Genetics, Stanford University School of Medicine, Alway Building, 300 Pasteur Drive, Stanford, CA 94305, USA., 2013, Vol. 9, pp. 640

[2] Schones, D. E. and Zhao, K. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet, Laboratory of Molecular Immunology, The National Heart, Lung and Blood Institute, National Institutes of Health,, Maryland 20892, USA. schonesde@nhlbi.nih.gov, 2008, Vol. 9(3), pp. 179-191

[3] Consortium, E. P. and others The ENCODE (ENCyclopedia of DNA elements) project Science, American Association for the Advancement of Science, 2004, Vol. 306(5696), pp. 636-640

[4] Diaz, A., Park, K., Lim, D. A. and Song, J. S. Normalization, bias correction, and peak calling for ChIP-seq. Stat Appl Genet Mol Biol, University of California, San Francisco, USA., 2012, Vol. 11(3), pp. Article 9

[5] Zhang, Y., Liu, T., Meyer et al. Model-based analysis of ChIP-Seq. Genome Biol, Department of Biostatistics and Computational Biology,USA., 2008, Vol. 9(9), pp. R137

[6] Mitra, A. and Song, J. WaveSeq: a novel data-driven method of detecting histone modification enrichments using wavelets. PLoS One. Vol. 7(9), pp. e45486

[7] Wu, H.-Y., Zhang, J. and Huang, K. Peak detection on ChIP-Seq data using wavelet transformation. IEEE International Conference on 2010, pp. 555-560

[8] Segura, V. Toledo-Arana, A., Uzqueda, M. and Muñoz-Barrutia, A. Wavelet-based detection of transcriptional activity on a novel Staphylococcus aureus tiling microarray. BMC bioinformatics 2012, Vol. 13, pp. 222

[9] Storey, J. D. and Tibshirani, R. Statistical significance for genomewide studies Proceedings of the National Academy of Sciences, National Acad Sciences, 2003, Vol. 100(16), pp. 9440-9445

[10] Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics, Oxford Univ Press, 2012, pp. bbs017