A Brief Survey of ChIP-seq Protocols and Peak Detection Methods

ALDEN HUANG, University of California, Los Angeles

General Terms: ChIP-seq

Additional Key Words and Phrases: ChIP-seq, MACS, ZINBA, STATS

ACM Reference Format:

Alden Huang. 2014. A Brief Survey of ChIP-seq Peak Detection Methods. *STATS 254* 2014, Spring, Article 1 (June 2014), 7 pages.

DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Mapping the genome-wide landscape of epigenetic modifications and protein-DNA interactions has become a burgeoning field of bioinformatics research essential to our understanding of the regulation of transcription within the cell. It is becoming increasingly apparent that in order to fully appreciate the dynamic mechanisms that mediate basic biological processes like differentiation, development, as well as disease states, we will ultimately require a systematic and comprehensive profile of epigenomes in multiple cell types and stages [Berstein et al, 2007].

Chromatin immunoprecipation followed by next-generation short-read sequencing (ChIP-seq) has become the de facto standard method to understanding the way that genetic information is encoded within chromatin structure. The recent technological advances in next-generation sequencing (particularly in terms of cost) have made this methodology much more accessible to investigators wanting to examine protein-DNA interactions directly.

1.1 ChIP-seq standards established by the ENCODE Consortium

A key driver to the rapid advance of the methodology has been the work done by the Encyclopedia of DNA Elements (ENCODE) Consortium. In a monumental effort to discover all functional elements of the genome, they have conducted hundreds of different ChIP-seq experiments. They have published a set of guidelines for conducting such experiments, and this has been invaluable in establishing a set working standards [Landt et al, 2012]. Because ChIP-seq, like many applications in bioinformatics, is a very data-driven process, it is of value to review these working guidelines to give a firm understanding of the nature of the analytical challenge at hand.

Author's address: A. Huang, 1524 Gonda Building, 695 Charles E. Young Drive, Los Angeles, C.A. 90095; email: alden.huang@gmail.com.

DOI: http://dx.doi.org/10.1145/0000000.0000000

Perhaps the single most important factor of a successful ChIP-seq experiment is quality of the antibody used for the chromatin immunoprecipation. The model organism ENCODE (mod-ENCODE) projects have done perhaps the most comprehensive study of antibody efficacy in the context of ChIP-seq studies to date. Collectively, they characterized over 200 antibodies from different species and found that there was a huge variation in efficiency, even among the same antibody taken from different lots. A quarter of the antibodies that they tested failed specificity requirements, which is remarkably high. Therefore, the ENCODE Consortium guidelines recommends that a prior assessment of particular antibody quality be done before using it to undertake a ChIP-seq experiment. The methods for doing so are largely outside the scope of this review, but include basic immunoblotting and chemiluminescence assays in both a sample of interest and a negative control, where the antigen of interest has been knocked-down by molecular methods, when such a method is applicable.

The ability to detect enrichment peaks requires sequencing at different depths depending on the type of peaks that are expected. The ENCODE Consortium guidelines distinguish between punctate sources (e.g. transcription factor binding), broad sources (e.g. HK36me3 assays), and mixed-type of data sources, and provides recommended guidelines for the sequencing depth that is necessary to accurately detect these types, which also varies with organism type (since genome complexity naturally varies between species). Specifically for human data, recommendations vary between 20 and 40 million reads, depending on whether punctate peaks or broad peaks, respectively, are expected. In addition, they recommend that a single biological replicate be carried out. From personal experience, it has become obvious that technical replicates for most every NGS experiment is essentially useless, as this has always shown to be highly reproducible for any given sample. It is of interest here to note that, unlike other NGS applications like whole-transcriptome sequencing (RNA-seq), performing additional biological replicates beyond two has actually been shown to be of little value in ChIP-seq. The ENCODE Consortium has set a baseline for success of between biological replicates: either they must have 75% of the identified targets in common, or 80% of the top 40% of targets must be the same.

The rest of the ENCODE Consortium guidelines are with regard to quality assessment and data reporting. In short, following these guidelines serve to unify the way ChIP-seq is performed across different laboratories and with various experimental setups. From a biological perspective, this type of experimental standardization and open data reporting has been invaluable to both developing the methodology as well as deciphering meaningful biological insight from it. It is the opinion of the author that perhaps one unexpected and striking aspects of our biological understanding of chromatin modifications that has been gleaned from the recent advances of the ChIP-seq methodology is just how cell and tissue-specific the various chromatin marks are. Although it is outside the scope of our discussion here, a growing and highly successful field of intense study is devoted to actually predicting the presence of certain chromatin marks in the absence of experimental data, based on a previously established standard set of data like that produced by the ENCODE Consortium project. The ability to do this in successful manner also underscores the biological relevance of the ChIP-seq assay in general.

2. CHARACTERISTICS OF CHIP-SEQ DATA

2.1 Types of experimental assays

If there is one point I wish to illustrate with the discussion of working guidelines for ChIP-seq experiments, it is essentially that ChIP-seq analysis boils down to detecting enriched signal, and this is challenging because there are many elements of experimental noise inherent to the data. The aim, ultimately, of any ChIP-seq experiment is to discover enrichment of sequencing signal over background The guidelines set forth by the ENCODE Consortium were established in hopes of reducing sources of inter-experimental variability that can arise when performing any of the widely available experimental

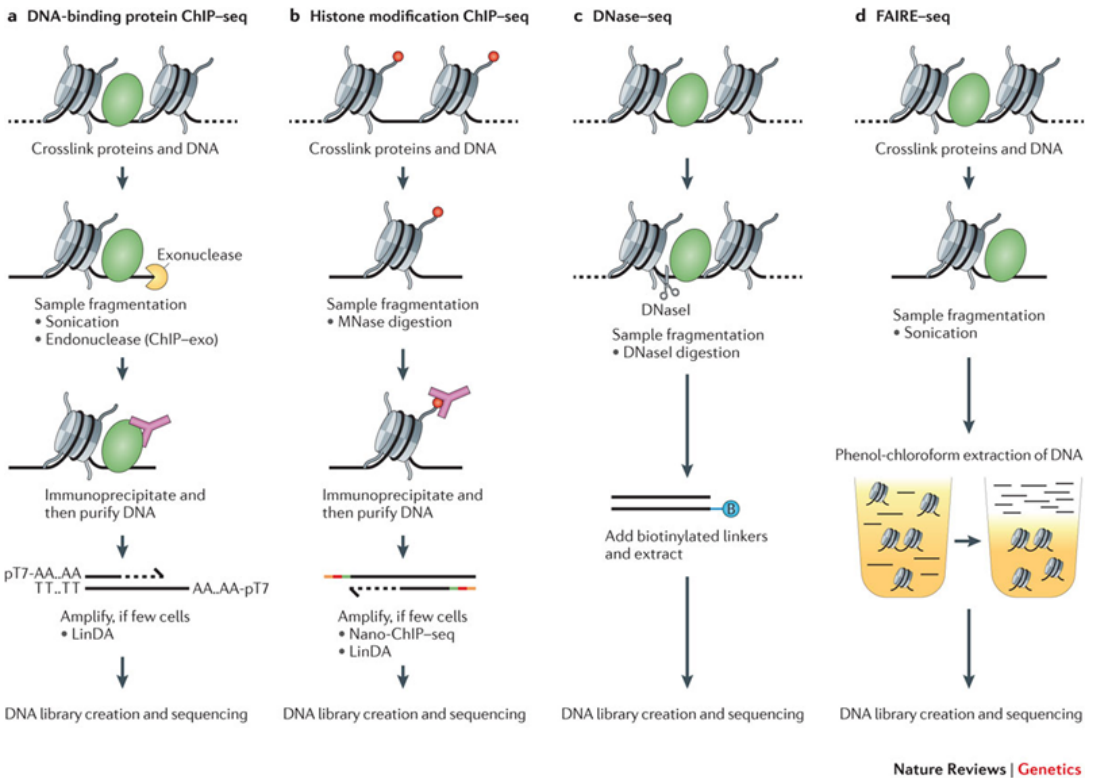

Fig. 1. Experimental overview of different types current of ChIP-seq assays. From: Furey, TS 2012

protocols currently in common use to detect an ever-growing catalog of biologically relevant chromatin modifications present in the cell. Figure 1 highlights some of the different experimental assays used to assay some of the most common chromatin modifications. Each of the types of experimental protocols produces a slightly different expected peak profile. Since typical DNA-binding proteins that directly interact target segments of DNA based on short, specific sequences (a DNA-binding motif), this type of experimental assay typically produces punctate peaks. Histone modifications shown in (b) typically produce a variety of peaks depending on the type of modification being assayed, but most are fairly broad in nature, due to the wider prevalence of histone modifications. FAIRE-seq, the most recently developed experimental protocol depicted in the figure, also produces very wide-spread broad peaks, as it basically assays all of the open chromatin. The resultant tag alignments as shown in the Interactive Genomics Viewer [Robinson, et al 2011] for several of these types of assays from public data sources can be seen in Figure 2.

2.2 Control input data

The use of a control input is an issue of particular importance that warrants further discussion in regards to ChIP-seq. There are many potential sources of artifacts that complicate peak detection and make effective peak-calling methodology not only an analytical task but also an experimental one [Park, PJ 2009]. It has been observed early on that even the most advanced shearing methods (the current state-of the art standard method is acoustic shearing using a Covaris instrument) do not fragment even nascent DNA in a completely uniform manner. This non-uniform fragmentation is often

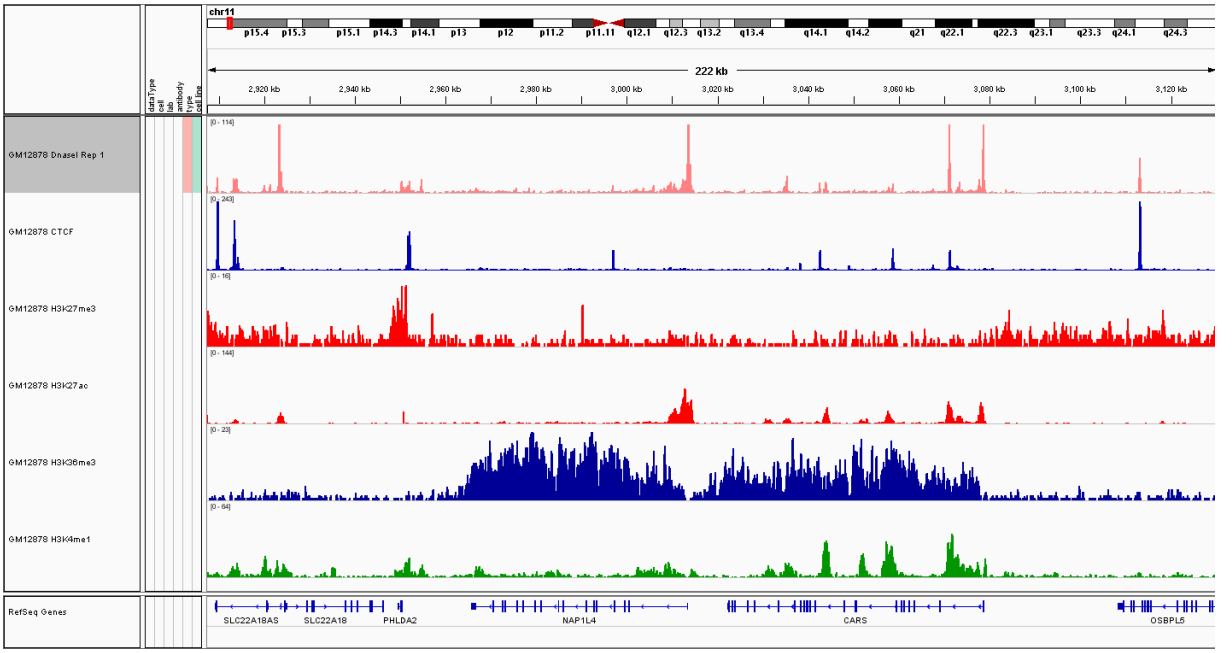

Fig. 2. Examples of aligned and processes ChIP-seq data taken from the same cell type generated by the ENCODE project (depicted in the Interactive Genome Viewer).

exacerbated in the case of ChIP-seq data, because in the typical case, the DNA chromatin structure is somewhat preserved, e.g. by cross-linking to the proteins bound to it. In summary, DNA in open chromatin tends to fragment easier than those in compacted regions, and this can lead to bias when the fragment length is assumed to be uniform genome wide. Moreover, the sequence of the particular genome of interest is itself inherently biased. Highly repetitive stretches of genomic sequence, for example, will inherently map with less precision than unique regions, presenting another potential source of bias when tags of a certain sequence can map to multiple genomic locations. In order to account for these systematic and structured biases, ChIP-seq experiments almost always include a background control signal. All experimental parameters are the same, except the antibody step is omitted and the sample is sequenced without enrichment.

3. METHODS FOR DETERMINATION OF REGIONS OF ENRICHMENT

The peak profiles garnered from ChIP-seq protocols vary widely by the type of experimental protocol used. Rather than present a wide survey comparing the individual disparities between the myriad of available peak-calling methods, I choose to focus the discussion on two different popular peak-callers, each of which is tailored towards a different experimental paradigm.

3.1 Model-based analysis of ChIP-seq data

Model-based Analysis of ChIP-seq data (MACS) is a very popular peak-finding program developed specifically for the classical ChIP-seq experiment whereby a specific transcription factor of interest is pulled down by targeted immunoprecipitation using a particular antibody, and the resultant peaks are punctate and well defined [Zhang, et al 2008]. Its main contributions are two-fold. It was one of the first methods to explicitly model fragment size to give an accurate delineation of the precise binding pattern

of a particular DNA-binding protein of interest. Second, it recognized that the nature of the noise or background signal for a particular ChIP-seq experiment is highly context dependent, and makes an effort to explicitly determine true peaks within its genomic context by measuring enrichment relative to a local, rather than global, background signal.

Central to the MACS peak finding method is the realization that ChIP-seq tags represent only the 5 ends of the actual DNA fragment that is pulled down. Therefore, to precisely locate the actual region of a DNA-protein interaction, estimation fragment length is crucial. However, this is generally an unknown parameter, and moreover, ChIP-seq data (at least in its earlier manifestations) is typically single-ended. MACS takes a user-inputted fragmentation size (the average size at which the DNA from the sample is fragmented by sonication) and this value is used to scan the genome with a sliding window twice the fragmentation size only to check initially for regions of putative enrichment.

Instead of using a user-inputted fragment size explicitly to help precise the accurate boundaries of DNA regions bound to protein, MACS exploits a particular consequence of the ChIP-seq protocol. Because there is an equal likelihood that a particular fragment will be sequenced from either end, the pattern of ChIP-seq tags exhibit a bimodal enrichment pattern surrounding the DNA region bound to the protein. It quantifies the distance between each bimodal pair of peaks and calculates the distance between them d, then shifts the tags by a distance of d/2 and then extends them by a length d from the center. Note that d is effectively an estimate of fragment size empirically derived from the spacing of bimodal peak pairs surrounding a putative DNA-binding motif.

With these shifts and manipulations of the tag sizes MACS effectively centers the in-silico the sequencing tags at the middle of a putative DNA-binding motif. It then quantifies the relative enrichment of these tags centered at their peaks relative to a control signal. For the count-based nature of ChIP-seq data, the Poisson distribution is a very natural choice, and many earlier peak-callers utilize it. However, MACS was one of the first methods in wide use to recognize that the background levels in ChIP-seq are non-uniform and well-structured. The second major contribution of MACS to the then current calling algorithms was to employ a local λ in modeling the background signal from control. Rather than sampling from the entire control signal, MACS determines a local λ dynamically:

$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k}], \lambda_{5k}, \lambda_{10k})$$

For each putative peak, it takes the local λ_1 where λ_{1k}, λ_{5k} and λ_{10k} are estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample, as well as the overall background signal, and it chooses the maximum λ upon which to base its test for significance. In the absence of a control sample, MACS measures enrichment based on a background distribution that is centered on the putative peak window instead. Similar to other methods of peak detection, MACS assesses significance using a false discovery rate based on a sample swap. The empirical FDR is calculated as the (no. of peaks detected in control)/(no. of peaks detected in the sample). This simple modified assessment of background signal is in practice quite effective. Using a local λ instead of a global one reduces the FDR rate especially in the case where no control input is available. In their evaluations using real data on Forkhead box protein A1 (FOXA1), the authors of MACS note using a local λ is very effective in reducing the FDR rate. It is only 0.4% when control data is available and 3.8% when it is not. This is in stark contrast to when global λ_{BG} is used, where the FDR increases to over 40%. This demonstrates that the background in ChIP-seq data is highly structured, and this structure is accurately modeled in a context-specific manner.

3.2 Shortcomings of the MACS Model

MACS is a particularly effective method for detecting enriched peaks in the classical ChIP-seq context, when an antibody targeting a specific DNA-bound protein (e.g. a transcription factor) is used and well-

defined, punctate peaks are expected. However, as is clear by the way in which it explicitly models peaks by inspection of the bimodal distributions of tags present surrounding well-defined motifs of DNA-protein interactions, the MACS model is clearly not applicable for newer enrichment protocols like Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)-seq [Song et al, 2011], or for certain histone modifications, both of which enrich for much larger regions of the genome compared to those typically seen for specific transcription factors (5 and 10% respectively, versus 1%) [Park, PJ 2009]. Moreover, the non-specific peaks produced by these less punctate enrichment methods tend to be much broader and variable in their size range, often spanning entire genes. The need for a method that can accurately model the wide variability in enrichment patterns generated from the ever-increasing variety of experimental paradigms is something that has been recognized early in the development of modern ChIP-seq methods.

Additionally, the statistical methods used in MACS are, by current measure, somewhat elementary. As in the case of other next-generation sequencing methods, the Poisson and binomial distributions were obvious first choices for distributions with which to model count data. However, as in the case of RNA-seq, the simplified use of Poisson distribution, in which both the mean and variance are captured in a single parameter, λ, is inaccurate. It has long been known in the case of RNA-seq that the negative binomial distribution, which incorporates the variance as a separate parameter, allowing for it to account for over-dispersion, is a much more appropriate in the case of NGS [Anders S Huber W 2010]. Moreover, a natural extension of this in the case of ChIP-seq data, where the vast majority of the genomic space is indeed background, is the zero-inflated form of the negative binomial, as zero-count data is considered separately.

3.3 The Zero-Inflated Negative Binomial Algorithm

The Zero-Inflated Negative Binomial Algorithm (ZINBA) is a recently developed methodology aimed at addressing many of the aforementioned shortcomings [Rashid et al 2011]. It is a much more sophisticated statistical model and its particular utility is that it is widely applicable to the range of possible peaks types that may be present in any of the variety of experimental methods used to enrich protein-bound DNA signals. The method consists essentially of three separate steps: data preprocessing, determination of enrichment, and peak refinement; we will discuss the first two here.

The first step tallies the reads across the genome in non-overlapping contiguous regions and is used to score covariate information. Any number of covariates is allowed to co-vary with the signal, and includes things like local G/C composition, read-counts, and mappable-bases, to name a few. Similar to MACS, these covariates are determined in a local manner. Since the background signal in ChIP-seq data has been shown to be highly structured, accurately accounting for this structure using a mixed linear regression model results in a much more sophisticated ability to discern regions of enrichment above noise that are biologically relevant.

The second step makes use of a novel mixture regression model in order to classify each region into one of three categories defining the type of signal present in a ChIP-seq sample: background signal, enrichment signal, or zero-inflated regions. The background is essentially experimental noise. The enrichment signal is specifically regions that are enriched by the experimental protocol, basically the regions we are biologically interested in characterizing. The zero-inflated region consists of regions that are, by various caveats of NGS, unrepresented by sequencing tags in the data. A common cause of these zero-inflated regions is, for example, a lack of sufficient sequencing depth.

Statistically, ZINBA uses a three-part mixture of distributions to more accurately model count data typically encountered in ChIP-seq data:

$$p(Y_i = y_i|\mu_i, \theta, \pi_i) = \begin{cases} \pi_{i0} + (1-\pi_{i0})\pi_1(\frac{\theta_1}{\mu_{i1}+\theta_1})^{\theta_1} + (1-\pi_{i0})\pi_2(\frac{\theta_1}{\mu_{i2}+\theta_2}) & y_i = 0 \\ (1-\pi_{i0})\pi_1 \frac{\Gamma(y_i+\theta_1)}{y_i!\Gamma(\theta_1)}(\frac{\theta_1}{\mu_{i1}+\theta_1})(\frac{\mu_{i1}}{\mu_{i1}+\theta_1})^{y_i} & \\ +(1-\pi_{i0})\pi_2 \frac{\Gamma(y_i+\theta_2)}{y_i!\Gamma(\theta_1)}(\frac{\theta_2}{\mu_{i2}+\theta_2})(\frac{\mu_{i2}}{\mu_{i2}+\theta_2})^{y_i} & y_i > 0 \end{cases}$$

Here, $\mu_i - (\mu_{i1}, \mu_{i2})$ are the means of the negative binomial distributions for the background and enrichment portions of the data, respectively, for window i, and $\theta - (\theta_1, \theta_2)$ are the dispersion parameters. π_{i0} is the prior probability that a window i belongs to either the zero-inflated, background, or enriched component of the data. Essentially, $\pi_0 = (\pi_{10}, ..., \pi_{i0}, ..., \pi_{n0})$ is an $n \times 1$ vector indicating zero-inflated prior probabilities, and π_1 and π_2 are set as scalars such that $\pi_1 + \pi_2 = 1$. The ZINBA algorithm then utilizes an Expectation Maximization algorithm to estimate model parameters and posterior probabilities. These are used to partition windows into the three separate component memberships. Readers interested in the mathematical details of the EM algorithm are invited to refer to the supplementary methods provided with the ZINBA paper.

By using mixture of distributions to model ChIP-seq tag data, and the ability to include covariates in the model, ZINBA is able to overcome the rather specific nature of MACS and other previous models and accurately models the wide variety of data that results from disparate enrichment protocols that collectively comprise modern ChIP-seq. It has been shown to be particularly useful in calling broad-ranged peaks like those derived from histone modifications, and also with considerable success with newer experimental methods like FAIRE-seq.

4. CONCLUSION AND FUTURE DIRECTION

ChIP-seq has rapidly supplanted previous methods of assessing DNA-protein interactions and has become the de facto standard for studies of chromatin structure. The growing recognition of the chromatin structure on nearly every aspect of cellular biology, spurned by advances in NGS, has motivated a rapid development of novel experimental protocols, which in turn has driven the need for new analytic techniques.

Although the results from initial experiments were plagued by inconsistent methodologies, projects like the ENCODE Consortium have been instrumental in establishing universal guidelines to help mitigate the experimental and methodological inconsistencies between studies. Even still, inherent experimental noise, while structured, presents many challenges (or opportunities) which requires the development refined models to further elucidate novel biological insights from the intricate repertoire that regulates chromatin structure, and ultimately, gene expression.

5. WORKS CITED

Anders S Huber W (2010) Differential expression analysis for sequence count data. Genome biology 11(10):R106.

Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128(4):669-681.

Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature reviews. Genetics 13(12):840-852.

Landt SG, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consor-

tia. Genome research 22(9):1813-1831.

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nature reviews. Genetics 10(10):669-680.

Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nature methods 6(11 Suppl):S22-32.

Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome biology 12(7):R67.

Robinson JT, et al. (2011) Integrative genomics viewer. Nature biotechnology 29(1):24-26.

Song L, et al. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome research 21(10):1757-1767.

Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome biology 9(9):R137.