Chapter 18

Analyzing ChIP-seq Data: Preprocessing, Normalization, Differential Identification, and Binding Pattern Characterization

Cenny Taslim, Kun Huang, Tim Huang, and Shili Lin

Abstract

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a high-throughput antibody-based method to study genome-wide protein–DNA binding interactions. ChIP-seq technology allows scientist to obtain more accurate data providing genome-wide coverage with less starting material and in shorter time compared to older ChIP-chip experiments. Herein we describe a step-by-step guideline in analyzing ChIP-seq data including data preprocessing, nonlinear normalization to enable comparison between different samples and experiments, statistical-based method to identify differential binding sites using mixture modeling and local false discovery rates (fdrs), and binding pattern characterization. In addition, we provide a sample analysis of ChIP-seq data using the steps provided in the guideline.

Key words: ChIP-seq, Finite mixture model, Model-based classification, Nonlinear normalization, Differential analysis

1. Introduction

How proteins interact with DNA, the genomic locations where they bind to DNA, and their influence on the genes regulation have remained the topic of interests in the scientific community. By studying protein–DNA interactions, scientists are hopeful that they will be able to understand the mechanism of how certain genes can be activated while the others are repressed or remain inactive. The consequence of activation/repression/inactive will in turn affect the production of specific proteins. Since proteins play important roles for various cell functions, understanding protein–DNA relations is essential in helping scientists elucidate complex biological systems and discover treatment for many diseases.

Junbai Wang et al. (eds.), *Next Generation Microarray Bioinformatics: Methods and Protocols*,
Methods in Molecular Biology, vol. 802, DOI 10.1007/978-1-61779-400-1_18, © Springer Science+Business Media, LLC 2012

There are several methods commonly used for analyzing specific protein–DNA interactions. One of the newer methods is ChIP-seq, an antibody-based chromatin immunoprecipitation followed by massively parallel DNA sequencing technology (also known as next-generation sequencing technology or NGS). ChIP-seq is quickly replacing ChIP-chip as the preferred approach for generating high-throughput accurate global binding map for any protein of interest. Both ChIP-seq and ChIP-chip goes through the same ChIP steps where cells are treated with formaldehyde to cross-link the protein–DNA complexes. The DNA is then sheared by a process called sonication into short sequences about ~500–1000 base-pair (bp). Next, an antibody is added to pull down regions that interact with the specific protein that one wants to study. This step filters out DNA fragments that are not bound to the protein of interest. The next step is where it differs between ChIP-chip and ChIP-seq experiments. In ChIP-chip, the fragments are PCR-amplified to obtain adequate amount of DNA and applied to a microarray (chip) spotted with sequence probes that cover the genomic regions of interest. Fragments that find their complementary sequence probes on the array will be hybridized. Thus, in ChIP-chip experiment, one needs to predetermine their regions of interest and "place" them onto the array. On the other hand, in ChIP-seq experiment, the entire DNA fragments are processed and their sequences are read. These sequences are then mapped to a reference genome to determine their location. Figure 1 shows a simplified workflow of ChIP-seq and ChIP-chip and the different final steps. Both ChIP-seq and ChIP-chip experiments require image analysis steps either to determine their probe binding intensities (DNA fragment abundance) or to read out their sequences (base calling). Some of the advantages of using ChIP-seq versus ChIP-chip include: higher quality data with lower background noise which is partly due to the need of cross hybridization for ChIP-chip, higher specificity (ChIP-chip array is restricted to a fixed number of probes), and lower cost (ChIP-seq experiments require less starting material to cover the same genomic region). Interested readers can find more information regarding ChIP-seq in refs. 1 and 2.

In a single run, ChIP-seq experiment can produce tens of millions of short DNA fragments that range in size between 500 and 1,000 bp long. Each fragment is then sequenced by reading a short sequence on each end (usually 35 bp or longer, newer illumina genome analyzer can sequence up to 100–150+ bp) leading to millions of short reads (referred to as tags). Sequencing can be done as single-end or paired-end reads. In single-end reads, each strand is read from one end only (the direction depends on whether it is a reverse or forward strand) while in paired-end each strand is read from both ends in opposite directions. Because of the way the sequences are read, some literatures either extend the reads

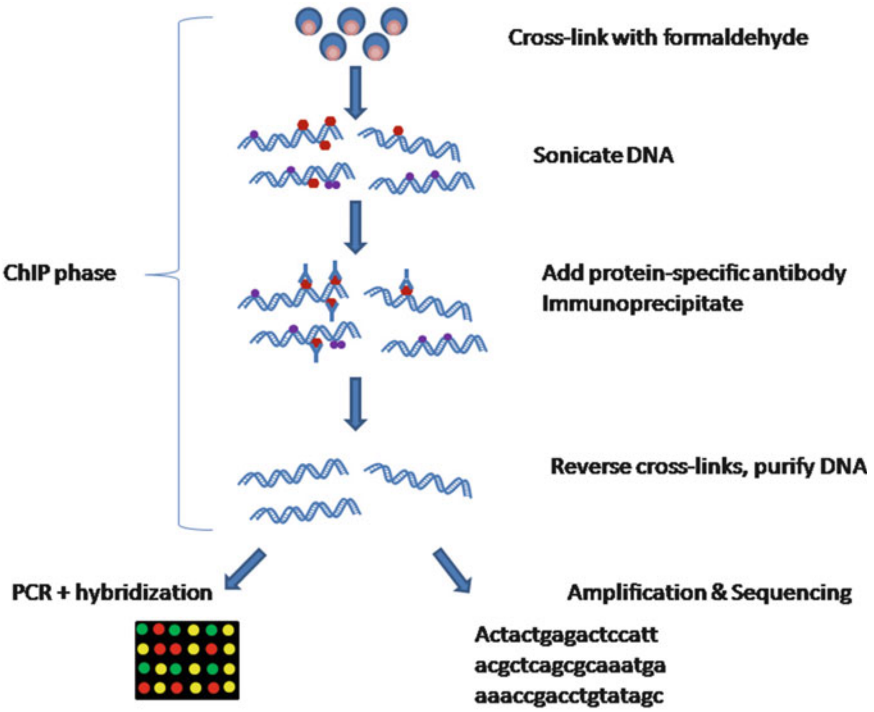
Cross-link with formaldehyde
Sonicate DNA
ChIP phase
Add protein-specific antibody
Immunoprecipitate
Reverse cross-links, purify DNA
PCR + hybridization
Amplification & Sequencing
Actactgagactccatt
acgctcagcgcaaatga
aaaccgacctgtatagc

Fig. 1. Schematic of the ChIP-seq and ChIP-chip workflow. First the cells are treated with formaldehyde to cross-link the protein of interest to the DNA it binds to in vivo. Then the DNA is sheared by sonication and the protein–DNA complex is selected using antibody and by immunoprecipitation. Reverse cross-links is done to remove the protein and DNA is purified. For ChIP-chip, the fragments continue on to be cross hybridized. In ChIP-seq, they go through the sequencing process.

or shift the reads to cover the actual binding sites (see Note 1). In the sample analysis provided in this chapter, since the RNA polymerase II (Pol II) tends to bind throughout the promoter and along the body of the activated genes, it is unnecessary to shift or extend the fragments to cover the actual binding sites. Once all the tags are sequenced, they are aligned back to a reference genome to determine their genomic location. To prevent bias in the repeated genomic regions, usually only tags that are mapped to unique locations are retained. Preprocessing of ChIP-seq usually includes dividing the entire genome into *w*-bp regions and counting the number of short sequence tags that intersect with the binned region. The peaks of the binned regions signify the putative protein binding sites (where the protein of interest binds to the DNA). Figure 2 shows an example visualization of binned Pol II ChIP-seq data in MCF7, a breast cancer cell line.

Even though ChIP-seq data has been shown to have less error compared to ChIP-chip, they are still prone to biases due to variable quality of antibodies, nonspecific protein binding, material differences, and errors associated with procedures such as DNA library preparation, tags amplification, base calling, image processing,

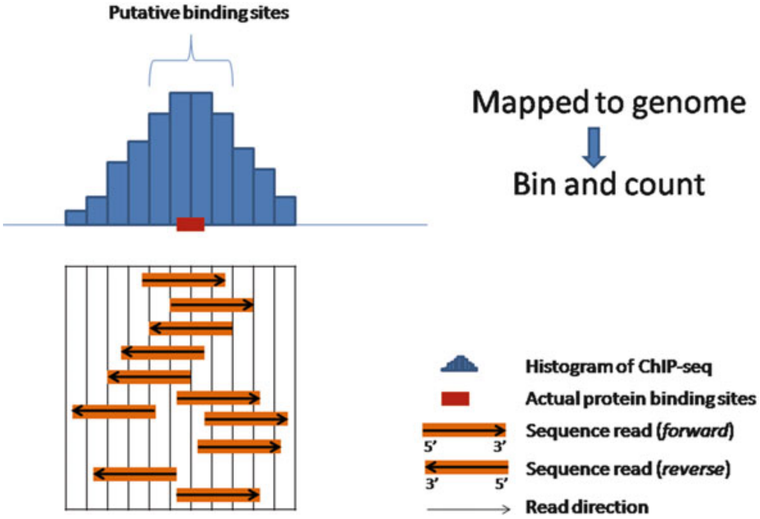

Fig. 2. An example visualization of the binned data with respect to the actual Pol II binding sites from ChIP-seq data. The single-end sequences are read from 5′ end or 3′ end depending on the direction of the strand. Note that since Pol II tends to bind throughout larger region, the peak is unimodal. For other protein, the histogram may be bimodal and hence some shifting or extension of the sequence read may be needed to identify the actual binding sites.

and sequence alignment. Thus, innovative computational and statistical approaches are still required to separate biological signal from noise. One of the challenges is data normalization which is critical when comparing results across multiple samples. Normalization is certainly needed to adjust for any systematic bias that is not associated with any biological conditions. Under ideal, error-free environment where every signal is instigated by its underlying biological systems, even a difference of one tag in a certain region can be attributed to a change in the conditions of the samples. However, various source of variability that is out of the experimenter's control can lead to differences that are not associated with any biological signal. Hence, normalization is critical to eliminate such biases and enable fair comparison among different experiments.

Our goal is to provide a general guideline to analyze ChIP-seq data including preprocessing, nonlinear data normalization, model-based differential analysis, and cluster analysis to characterize binding patterns. Figure 3 shows the flow chart of the analysis methods.

2. Methods

Given a library of short sequence reads from ChIP-seq experiment, the following steps are performed to analyze the data. We illustrate the process using the data generated from the Illumina Genome Analyzer platform, it nevertheless is applicable to data generated

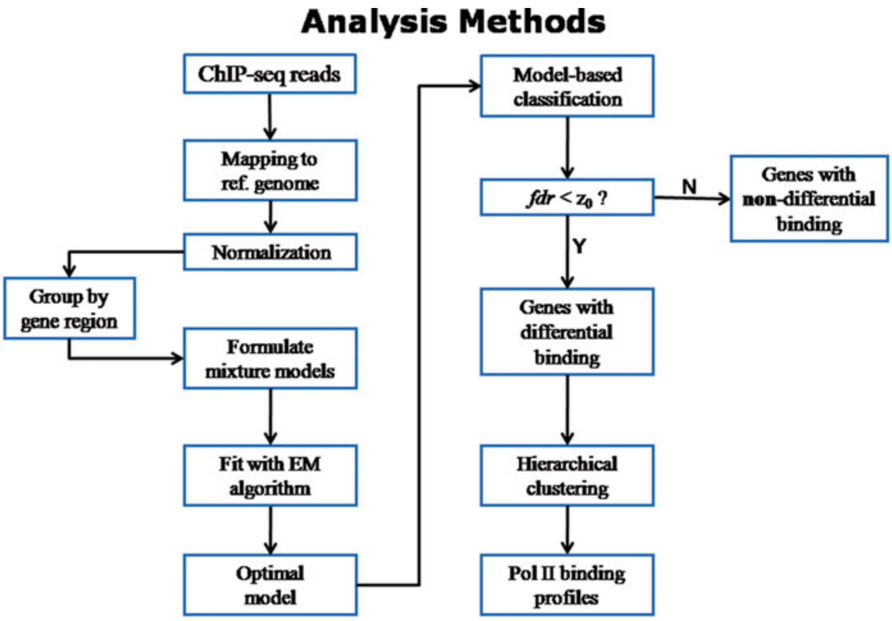

Fig. 3. Flow chart of the ChIP-seq analysis. The main steps of the methods to analyze ChIP-seq data including preprocessing are summarized in this figure.

from other sequencing platforms such as the Life Technology SOLiD sequencer.

2.1. Data Preprocessing

1. Determining genomic location of tags:
   (a) ELAND module within the Illumina Genome Analyzer Pipeline Software (Illumina, Inc., San Diego, CA) is used to align these tags back to a reference genome, allowing for a few mismatches.
   (b) After mapping, each tag will have its residing chromosome, starting and ending location. Depending on the software used, there may be a quality score associated with each base calling.
2. Filtering and quality control:
   (a) Filter out tags that are mapped to multiple locations.
   (b) Tags with low-quality score is filtered out internally in the Illumina pipeline.
   (c) Additional filtration maybe done as well. See Note 2.
3. Dividing genome into bins:
   (a) To reduce data complexity, the genome is divided into nonoverlapping w-bp regions (commonly called bins). The number of tags that overlap with each bin is then counted. We define x_{ij} as the sum of counts of tags that intersect with bin i in sample j.
   (b) Alternatively, one can use overlapping windows; see Note 3.

2.2. Normalization

1. When comparing multiple samples/experiments, normalization is critical. Normalization is needed so that the enrichment is not biased toward a sample/region because of systematic errors.

2. Sequencing depth normalization.

   Sequencing depth is a method used for normalization in SAGE (serial analysis of gene expression) and has been adapted for the analysis of NGS data by some authors; *See*, for example, ref. 10. The purpose of this normalization is to ensure the number of tags in each bin is not biased because the total number of tags in one sample (x_1) is much higher than in the other sample (x_2). Without lose of generality, let $x_1 > x_2$ and define $s = x_1/x_2$. Then, each bin in the other sample is multiplied by the scale factor s, that is $x_{i2}^{'} = s \times x_{i2}$. This is a (scaling) linear normalization, where x_{i2} is the tag count in bin i.

3. Nonlinear normalization.

   When comparing samples with stages of disease progression or samples before and after a treatment in which it is expected that many genes will not be affected, nonlinear normalization may be used. The nonlinear normalization is done in two stages. In the first stage, the data is normalized with respect to the mean. In the second stage, the data is normalized with respect to the variance.

   (a) Mean-normalization:

$$\hat{y}_i = \text{loess}\left((x_{i2} - x_{i1}) \sim \left(\frac{x_{i2} + x_{i1}}{2}\right)\right),$$

$$D_{i.\text{mean}}(x_{i2} - x_{i1})\hat{y}_i, \qquad (1)$$

   where $\hat{y}_i$ is the fitted value from regressing the difference on the mean counts using loess (locally weighted regression) proposed by Cleveland (3), and x_{i2} and x_{i1} are tag counts (may be after sequencing depth normalization) in bin i for control and treatment libraries, respectively. In this analysis, we assume no replicates are available. See Note 4 if replicates are available. This normalization step will find nonlinear systematic error and adjust them so the mean difference of unaffected genes becomes zero. $D_{i.\text{mean}}$ is the mean-normalized difference between reference and treatment libraries in bin i.

   (b) We choose to use the binding quantity for each sample directly (i.e., difference counts) rather than transforming it and using log-ratios for several reasons. First, it enables us to distinguish sites which have the same log-ratios but with vastly different magnitude. Furthermore, in ChIP-seq experiment, zeros indicate our protein of interest does not bind to the specific region. If we take log-ratios, these zero counts will be filtered out. In addition to those reasons, using difference counts will also help minimize problem with unbounded variance when fitting a mixture model; see Note 5.

(c) Wean-variance normalization:

$$\hat{z}_i = \text{loess}\left(|D_{i.\text{mean}}| \sim \left(\frac{x_{i2} + x_{i1}}{2}\right)\right),$$

$$D_{i.\text{var}} = \frac{D_{i.\text{mean}}}{\hat{z}_i}, \tag{2}$$

where $\hat{z}_i$ is the fitted value from regressing the absolute of mean-normalized difference on the mean counts. This step will find nonlinear and nonconstant variability in each region and adjust them so the spread is more constant throughout the genome. $D_{i.\text{var}}$ is the mean and variance normalized difference counts in bin i.

(d) For more detailed information including the motivation on nonlinear normalization for ChIP-seq analysis, the reader may refer to ref. 4.

4. Grouping tags into meaningful regions.

(a) To study how the changes in the binding sites affect specific region of interest, we can sum the tags into grouped regions as follows:

$$R_g = \sum_{i \in I_g} D_{i.\text{var}}, \tag{3}$$

where $D_{i.\text{var}}$ is the normalized difference in bin i as defined above. I_g is the index set specifying the bins belonging to group g. Thus, R_g is the sum of normalized tag-counts difference in region g for a total of G groups.

5. Although we did not scale our data based on the length of the groups, it may be a good idea to do further scaling normalization. See Note 6.

2.3. Differential Analysis: Modeling

1. With the normalized difference of grouped region (R_g) as input, we are now ready to perform statistical analysis. To determine whether there is a significant change in the tag counts of region g, we fit a mixture of exponential-normal component on R_g and apply a model-based classification. Assume that the data come from three groups, i.e., positive differential (genes that show increased bindings after treatment), negative differential (genes that have lower counts after treatment), and non-differential (those that do not change).

2. These three groups are assumed to follow certain distributions:

(a) Positive differential: an exponential distribution.

(b) Negative differential: the mirror image of exponential.

(c) Nondifferential: a combination of one or more normal distribution.

(d) See Note 7 for special cases.

3. The choice of these distributions is based on observation that the characteristics of these distribution match well with the biological data (5).

4. The modeling are done by fitting a mixture of exponential (a special case of gamma) and normal components. This model is called GNG (Gamma-Normal^k-Gamma) which is described in ref. 5 and used in the analysis of ChIP-seq (4). The superscript k indicates the number of normal component in the mixture which will be estimated. Model fitted by GNG is as follows:

$$f(R_g;\psi)=\sum_{k=1}^{K}\left(\gamma_k\phi\{R_g,\mu_k,\sigma_k^2\}\right)+\pi_1E_1\left(-R_g\times I\{R_g<-\xi_1\},\beta_1\right)+\pi_2E_2\left(R_g\times I\{R_g>\xi_2\},\beta_2\right), \tag{4}$$

where ψ is a vector of unknown parameters of the mixture distribution. The first component $\sum_{k=1}^{K}\left(\gamma_k\varphi\{R_g,\mu_k,\sigma_k^2\}\right)$ is a mixture of k normal component, where $\varphi\{.\}$ denotes the normal density function with mean μ_k and variance σ_k^2. Parameters γ_k indicate the proportion of each of the k normal components.

5. E_2 and E_1 each refers to an exponential component with π_2 and π_1 denoting their proportions and beta parameters β_2 and β_1, respectively. $I\{.\}$ is an indicator function that equals to 1 when the condition is satisfied and 0 otherwise; $\xi_2,\xi_1>0$ are the location parameters that are assumed to be known. In practice, we can set $\xi_1=\left|\max(R_g<0)\right|$ and $\xi_2=\left|\min(R_g>0)\right|$.

6. EM algorithm is used to find the optimal parameters by calculating the conditional expectation and then maximizing the likelihood function. See Note 5.

7. Akaike information criteria (AIC) (6), a commonly used method for model selection, is used to select k, the order of the mixture component that best represents the data.

2.4. Differential Analysis: Model-Based Classification

1. The best model selected by EM algorithm provides a model-based classification approach. Using this model, we can classify regions as differential and nondifferential binding sites.

2. Local false discovery rate (fdr) proposed by Efron (7) will be used to classify each binding sites based on the GNG model.

$$\mathrm{fdr}(R_g) = \frac{f(R_g;\psi_0)}{f(R_g;\psi_0) + f(R_g;\psi_1)}, \tag{5}$$

where $f(R_g;\psi_0)$ is the function of the k normal components and $f(R_g;\psi_1)$ is the function of the exponential components.

3. Ultimately, one can adjust the number of significantly different sites by setting the fdr value that they are comfortable with.

2.5. Binding Pattern Characterization

1. To further investigate the importance of protein binding profiles, one can perform clustering on the genes binding patterns which show significant changes.
2. Genes' lengths are standardized to enable genome-wide profiling.
3. The binding profiles for each gene are interpolated with optimum interpolator designed using direct form II transposed filter (8). As a result of this interpolation, all genes have the same length artificially.
4. Hierarchical clustering is then performed to group genes based on their binding profiles.

2.6. A Sample Analysis

In this section, we show a sample ChIP-seq analysis applying the above methodologies. Details on where to download the sample data and the software are provided in Subheading 2.7. The protein that we are interested in is RNA polymerase II (Pol II) and we are comparing MCF7, a normal breast cancer cell line before and after 17 β-estradiol (E2) treatments. We define MCF7 as the control sample and MCF7 + E2 as the treatment sample. The first part of the analysis is to discover genes that are associated with significant Pol II binding changes after E2 treatment. Because it is expected that the E2 treatment on cancer cell does not affect a large proportion of human genome, the above nonlinear normalization can be applied. See Note 8. Finally, significant genes are clustered to characterize their binding profiles.

2.6.1. Data Preprocessing

1. Sequence reads are generated by Illumina Genome Analyzer II. Reads are mapped to reference genome using ELAND provided by Illumina, allowing for up to two mismatches per tag.
2. Only reads that map to one unique location are used in the analysis. The total number of uniquely mapped reads (also known as sequence depth) for MCF7 sample is 6,439,640 and for MCF7 + E2 is 6,784,134. Table 1 shows details of the mapping result.
3. Nonoverlapping bins of size 1 kbp are used to divide the genome. 1 kbp is chosen to balance between data dimension and resolution. Thus, we set window size, $w = 1{,}000$ (bp).

Table 1
Reads of Pol II ChIP-seq data

Samples	Number of reads	Unique map	Multiple location	No match
MCF7	8,192,512	6,439,640 (79%)	1,092,519 (13%)	660,353 (8%)
MCF7 + E2	8,899,564	6,784,134 (76%)	1,233,574 (14%)	881,415 (10%)

The number of reads gives the raw counts from Solexa Genome Analyzer. Those under unique map are the reads that are used in our analysis. Those that are not uniquely map are either mapped to multiple loci or there is no match in the genome even allowing for two bases mismatches

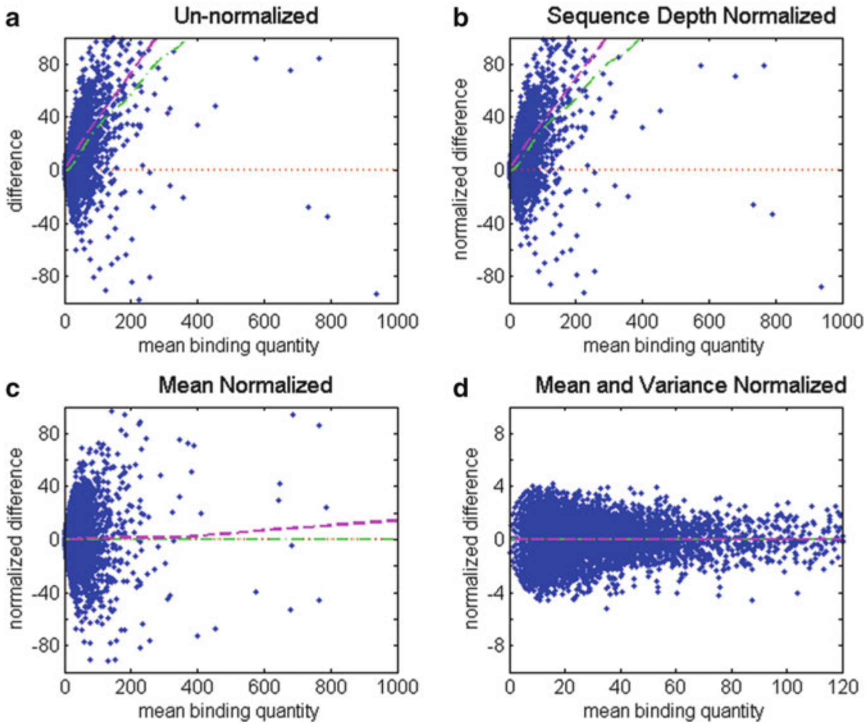

Fig. 4. Normalization process. The effects of the different normalization for chromosome 1 in MCF7 sample are shown. (**a**) The unnormalized data shows biases toward positive difference and nonconstant variance. (**b**) Data normalized using sequencing depth. (**c**) Data after normalization with respect to mean. (**d**) Data after two-stage normalization (with respect to mean and variance). *Dot-dashed* (*green*) *line* is the average of the difference counts estimated using loess regression. *Dashed* (*magenta*) *line* is the average absolute variance estimated using loess. *Dot* (*red*) *line* indicates the zero difference.

2.6.2. Sequencing Depth and Nonlinear Normalization Detailed in Subheading 2 Is Applied

1. We define MCF7 sample as the reference ($j = 1$) and MCF7 + E2 data as the treatment ($j = 2$). Figure 4a (raw data) shows that a large proportion of regions in treatment sample have Pol II binding that are higher than the control sample (indicated by the green dot-dashed line, estimated mean difference $D_{i.\mathrm{mean}}$ in (1), which is always above zero). Sequencing depth normalization is commonly used for normalizing ChIP-seq

data. This normalization method scales the data to make the total sequence reads the same for both samples. As shown in Fig. 4b, since the total number of reads in control versus treatment sample is about the same, normalization based on sequencing depth has little effect. Figure 4b depicts the data after applying sequence depth (linear) normalization which still show biases toward positive difference and unequal variance, hence it is not sufficient as a normalization method. Figure 4c, d show the effect of the nonlinear normalization. In our application, we use a span of 60% and 0.1 to calculate loess estimate of the mean and variance, respectively. Since E2 treatment should only affect a small proportion of binding sites, i.e., most regions should have zero difference, normalization with respect to mean is applied to correct for this bias. Figure 4c shows the data after the mean adjustment. In addition, since the spread of the region increases with the mean as shown in Fig. 4c (indicated by the magenta dashed line, $D_{i.\text{var}}$ in (2)), we apply normalization with respect to variance. Figure 4d shows that the data after mean and variance normalization is spread more evenly around zero (difference) which indicate the systematic error caused by unequal variance and bias toward positive difference has been corrected.

2. Grouping. In our application, we are interested in the Pol II binding quantities changes in the gene regions. Thus, after normalization, we summed tags count differences that fall into gene region based on RefSeq database (9). Hence, in Equation 3 above, I_g is the index of bins that overlap with gene region g and R_g is the sum of normalized tag-counts difference in gene region g for all 18,364 genes. The number of genes is small enough for a whole genome analysis.

2.6.3. Differential Analysis: Modeling

1. We fit GNG on the normalized difference R_g for all $g = 1,\ldots,$ G genes (genome-wide). In Fig. 5, the fit of the best model superimposed on the histogram is plotted in panel a, which shows the model fits the data quite well. The individual component of the best GNG model with two normal components is shown in Fig. 5b. The QQ plot of the normalized data versus the GNG mixture in Fig. 5c, where most of the points scatter tightly around the straight line, further substantiates that the model provides a good fit for the data. The EM algorithm was re-initialized with 1,125 random starting points to prevent it from getting stuck in the local optimum. The EM algorithm is set to stop when the maximum iteration exceeding 2,000 or when the improvement on the likelihood functions is less than 10^{-16}.

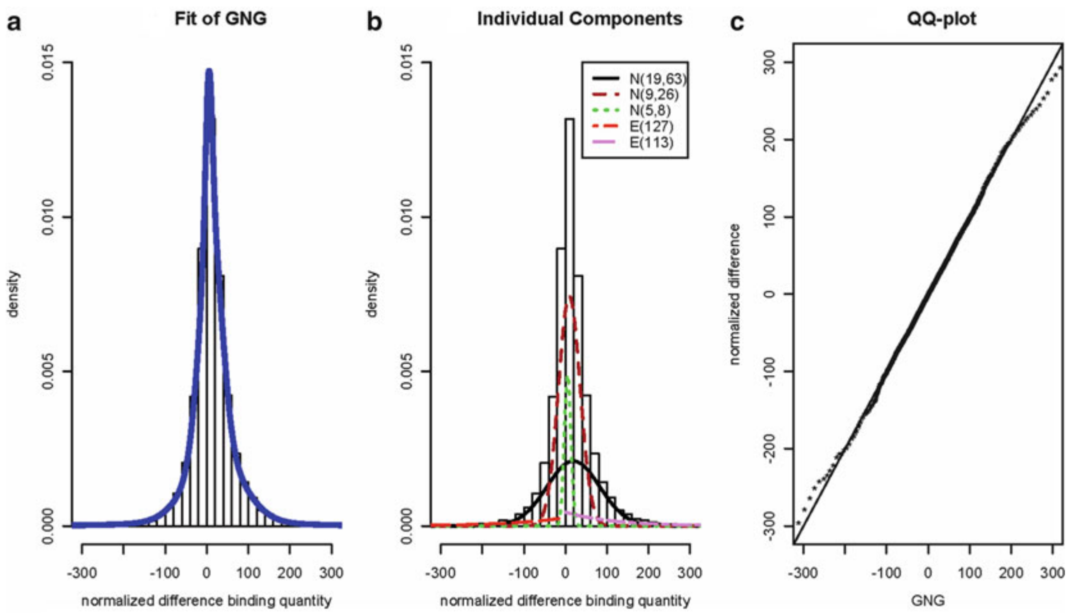

Fig. 5. The goodness of fit of the optimal GNG mixture to ChIP-seq data. (**a**) The fit of the best model imposed on the histogram of the normalized data (**b**) Plot of the individual components of the best GNG model. The best mixture has three normal components with parameters: $(\mu_1 = 5, \sigma_1 = 8)$, $(\mu_2 = 9, \sigma_2 = 26)$, and $(\mu_3 = 19, \sigma_3 = 63)$ represented by *dot* (*green*), *brown* (*dashed*), and *solid* (*black*) *lines*, respectively. The parameters for each of the exponential components are $\beta_1 = 127$ and $\beta_1 = 113$ represented by (*dot-dashed*) *red* and (*long-dash*) *magenta lines*, respectively. (**c**) QQ plot of the data versus the GNG model. All together these plots show that the optimal GNG model estimated by EM algorithm provide a good fit to the data.

2.6.4. Differential Analysis: Classification

1. Genes which have local fdr less than 0.1 are called to be significant. Using this setting, we find 448 genes to be associated with differential Pol II binding quantities in MCF7 versus MCF7 + E2 where around 60% of them are associated with increased bindings.

2. This finding is consistent with previous breast cancer study where the treatment of E2 appears to make more genes to be upregulated. Furthermore, we find PGR and GREB1 to be associated with significant increase of Pol II bindings (after E2 treatment) which are also found to be ER target genes that are upregulated in refs. 10 and 11.

3. A functional analysis on the genes associated with increased Pol II bindings is done using Ingenuity Pathway Analysis (17) (see Note 9) and shown in Fig. 6. The top network functions associated with these genes are cancer, cellular growth and proliferation, and hematological disease. Our finding thus suggests a regulation of nervous system development, cellular growth and proliferation, and cellular development in E2-induced breast cancer cells.

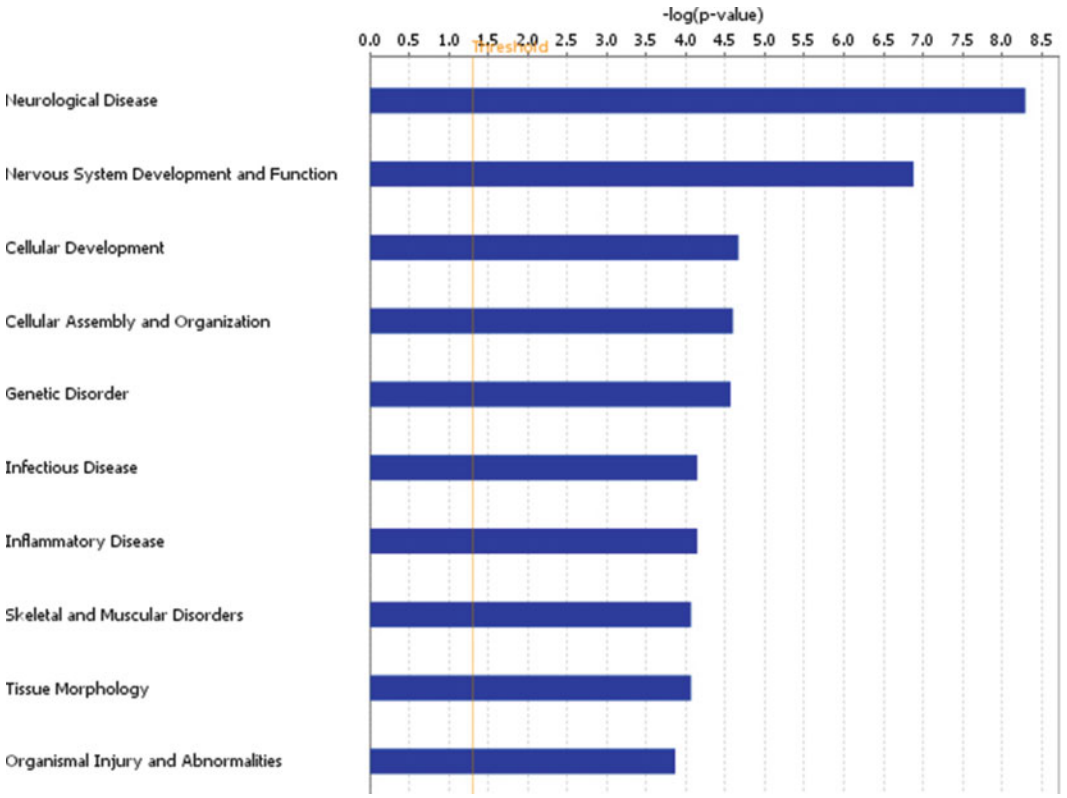

Fig. 6. The top ten functional groups identified by IPA. Analysis is done on the 264 genes which are found to show significant increase of Pol II binding in E2-induced MCF7. The *bar* indicates the minus log10 of the *p*-values calculated using Fisher's exact test. The *threshold line* indicates $p = 0.05$.

2.6.5. In Order to Characterize Pol II Binding Profiles of the Significant Genes Found in Previous Step, We Perform Hierarchical Clustering on These Regions

1. First, we filter out all the tags associated with introns retaining only those falling into exons regions. We did this filtration because the protein we are studying mainly acts on the exons regions.
2. Pearson correlation is used as the similarity distance in the hierarchical clustering procedure.
3. Binding profiles for each of the genes is interpolated to artificially make all genes length to be the same.
4. We find distinct clusters of genes with high Pol II binding sites at 5′ end (yellow, cluster 1) and genes with high Pol II binding quantity at 3′ end (blue, cluster 2), see Fig. 7.
5. Interestingly, there are more genes associated with high Pol II binding sites at 5′-end in MCF7 after E2 treatment.
6. This seems to indicate that different biological conditions (specifically treatment of E2) not only lead to changes in the Pol II binding quantity but it can also induce modification in the Pol II dynamics and patterns.

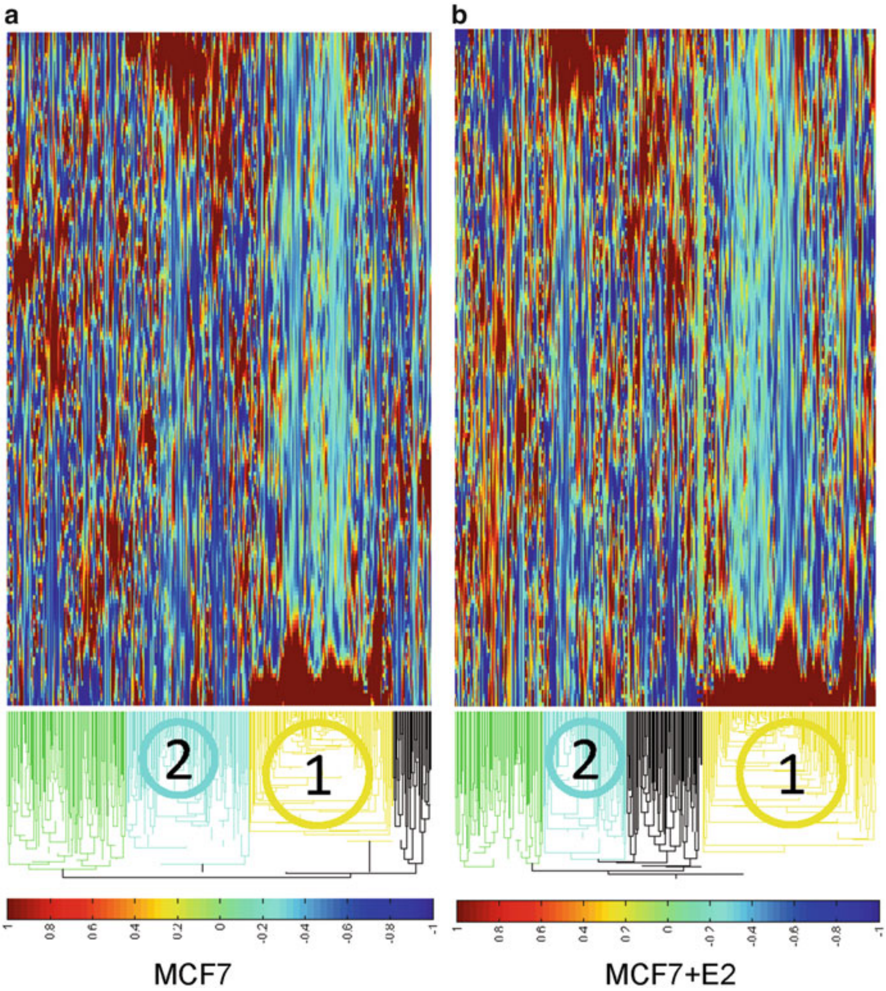

Fig. 7. Clustering of Pol II binding profiles in genes with significant changes in MCF7 after being treated with E2. Each column represent the Pol II binding profiles in each gene. Cluster 1 shows genes that are associated with high Pol II binding at the 5′ end and cluster 2 shows genes that are associated with high Pol II binding quantity at the 3′ end. (**a**) Binding profiles in MCF7; (**b**) binding profiles in MCF7 after E2 treatment. This indicate that E2 stimulation on MCF7 cell line not only change the Pol II binding quantity but it also modify its binding dynamics.

2.7. Software

The model fitting (GNG) is implemented as an R-package and is publicly available (21). The data used in the sample analysis is also downloadable from the same Web site.

3. Notes

1. Because the sequencing process cannot read the sequence of the entire tag length, some literature extends the sequenced tags to *x*-bp length and others shift each tag *d*-bp along the direction it was read in an attempt to cover the actual protein binding sites. For example, Rozowsky et al. (12) extend each

mapped tag in the 3′ direction to the average length of DNA fragments (~200 bp) and Kharchenko (13) shift the tags relative to each other. In our sample analysis, since Pol II tends to bind throughout the promoter and the body regions of a regulated gene, it is unnecessary to do shifting or extension. Readers should consider doing extension or shifting for any other protein binding analysis.

2. By combining number of mismatches with QC values of each base, one may be able to filter out low-quality/high mismatch reads from the analysis. On the other hand, one can also include more sequence reads with reasonable number of mismatches that are associated with high-quality score.
3. Instead of a fixed bin, some literature, for example, Jothi et al. (14) use a sliding window of size w where each consecutive window overlapped by $w/2$.
4. The methodology outlined here focus on analyzing ChIP-seq data without any replicates. When replicates are available, the same methodology can be applied by treating each replicates as individual independent samples or by taking the average of the replicates.
5. By allowing more than one normal components and not restricting them to have constant variances, the EM algorithm can have spurious solution where the variance becomes closer to zero and the model achieve artificially higher likelihood. We advise readers to use difference counts which would have a larger range than log-ratios in the modeling to minimize this problem. Re-initializing EM with multiple starting points will also help minimize this problem and prevent it from being trapped in a local optimum. For more information regarding the unboundedness problem of the likelihood function, see ref. 15.
6. A scaling normalization method known as RPKM (reads per kilobase per million mapped), proposed in ref. 16, is commonly used for ChIP-seq because of its simplicity. The main goal of this normalization is to scale all counts based on the length of the region and the total number of sequence reads. Although we did not apply this in our sample analysis, it may be a good idea to further scale our normalized data to minimize bias due to genes length and sequence depth. In this case, we can apply it on our normalized data as follows

$$y_g = \frac{R_g}{L_g \times \mathrm{SD}} \times 10^3 \times 10^6, \quad g = 1, \ldots, G;$$

where R_g is the number of loess-normalized tags in region g of a set of G regions, L_g is the gene length (in bp) of region g, and SD is the loess-normalized sequence depth (the total number of tags after loess normalization).

7. In the special situation where a normal component have either a large variance (say > 2IQR) or a large mean (say > 1.5 IQR), then such normal components should also be classified as differential components.
8. The nonlinear normalization described above is applicable when comparing samples in which the majority of genes do not show significant changes in treatment versus control samples. This assumption is satisfied for application in which the difference between the samples (i.e., effects of a drug treatment) is not expected to influence a large proportion of binding sites.
9. IPA is proprietary. There are free programs that provide similar information such as KEGG (18), GO (19), WebGestalt (20).

Acknowledgments

This work was partially supported by the National Science Foundation grant DMS-1042946, the NCI ICBP grant U54CA113001, the PhRMA Foundation Research Starter Grant in Informatics and the Ohio State University Comprehensive Cancer Center.

References

1. Johnson DS, Mortazavi A, Myers R et al (2007) Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316: 1441–1442
2. Liu E, Pott S, Huss M (2010) Q&A: ChIP-seq technologies and the study of gene regulation. BMC Biology 8: 56
3. Cleveland WS (1988) Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. J. Am. Stat. Assoc. 85: 596–610
4. Taslim C, Wu J, Yan P et al (2009) Comparative study on ChIP-seq data: normalization and binding pattern characterization. Bioinformatics 25: 2334–2340
5. Khalili A, Huang T, Lin S (2009) A robust unified approach to analyzing methylation and gene expression data. Computational Statistics and Data Analysis 53: 1701–1710
6. Akaike H (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR*: 267–281.
7. Efron B (2004) Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. Journal of the American Statistical Association 99: 96–104
8. Oetken G, Parks T, Schussler H (1975) New results in the design of digital interpolators. IEEE Transactions on Acoustics, Speech and Signal Processing [see also IEEE Transactions on Signal Processing] 23: 301–309
9. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Research 35: D61–65
10. Lin CY, Strom A, Vega V et al (2004) Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. Genome Biology 5, R66
11. Feng W, Liu Y, Wu J et al (2008) A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology. BMC Genomics 9: S23

12. Rozowsky J, Euskirchen G, Auerbach RK et al (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotech 27: 66–75
13. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology 26: 1351–1359
14. Jothi R, Cuddapah S, Barski A et al (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucl. Acids Res. 36: 5221–5231
15. McLachlan G, Peel D (2000) Finite Mixture Models. Wiley-Interscience, New York
16. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 5:621–628
17. The networks and functional analyses were generated through the use of Ingenuity Pathways Analysis (Ingenuity® Systems), see http://www.ingenuity.com
18. KEGG pathway analysis, see http://www.genome.jp/kegg/
19. Gene Ontology website, see http://www.geneontology.org/
20. WEB-based GEne SeT AnaLysis Toolkit, see http://bioinfo.vanderbilt.edu/webgestalt/
21. Software and datasets used can be downloaded, see http://www.stat.osu.edu/~statgen/SOFTWARE/GNG/