BIOENGINEERED
2017, VOL. 8, NO. 3, 248–255
https://doi.org/10.1080/21655979.2016.1226714

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS

Check for updates

Genomic feature extraction and comparison based on global alignment of ChIP-sequencing data

Binhua Tang[a,b]

[a]Epigenetics & Function Group, College of the Internet of Things, Hohai University, Jiangsu, China; [b]School of Public Health, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Enhanced accuracy and high-throughput capability in capturing genetic activities lead ChIP-sequencing technology to be applied prevalently in diverse study for tackling DNA-protein interaction problems. Till now, such questions as deciding suitable ChIP-seq arguments and comparing sample quality still haunt biologists. We propose the methods for answering such questions as deciding optimal argument pairs in global alignment of ChIP sequencing data; then we employ a modern signal processing approach to extract inherent genomic features from the global alignments of transcriptional binding activities; together with pairwise comparison from intra- and inter-sample perspectives; thus we can further determine alignment quality and decide the optimal candidate for multi-source heterogeneous high-throughput sequences. The work provides a practical approach to quantitatively compare the alignment quality for heterogeneous sequencing data, especially in determining the efficiency of transcriptional binding from replicate samples, thus it helps to exploit the potentiality of ChIP-seq for deep comprehension of inherent biological meanings from the high-throughput genomic sequences.

ARTICLE HISTORY
Received 5 August 2016
Revised 12 August 2016
Accepted 16 August 2016

KEYWORDS
ChIP-seq; comprehensive analysis; genomic feature; optimal argument

Introduction

Next generation sequencing (NGS) combined with ChIP technology provides a genome-wide study perspective for current biomedical research and clinical diagnosis applications.[1-3]

Data quality and inherent characteristics of those NGS sequencing profiles are directly related to the reliability and authenticity of the analysis results. For example, ChIP-seq data characterize alteration evidence for transcription factor (TF) binding activities in response to chemical or environmental stimuli, but if the ChIP-seq data quality is below normal standard, any follow-up analysis may lead to inaccurate TF binding results, for example, inevitable loss of biological meaningful sites.[4,5]

And secondly, mostly investigated items in ChIP-seq peak-calling procedures are peak number, False Discovery Rate (FDR) and corresponding bin size selected in each analysis. Without exception such arguments form impenetrable barriers for biologists and bioinformaticians to select suitable pair conditions for analyzing experimental results. And to our knowledge, few literatures focus on such topics. Thus herein we propose a flexible data feature detection algorithm for solving such an argument-optimization problem in peak-calling.

In breast cancer cell content, a specific estrogen receptor α (ERα) is recognized as mediating genetic regulation through diverse nuclear signaling pathways and protein kinase cascades. Within those biological pathways, ER binds to corresponding estrogen response elements (ERE) at target genes' regulatory areas, and combines other components to control downstream transcriptional processes.

Thus, elucidating innate regulatory mechanisms of those specific ERs facilitates comprehensive understanding of ER-specific regulation in most breast cancer pathways and networks under investigation.

CONTACT Binhua Tang bh.tang@outlook.com 200 N Jinling Rd, Hohai University, Jiangsu 213022, China.

Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/kbie.

Supplemental data for this article can be accessed on the publisher's website.

Within analysis of the ERα ChIP-seq data, we seek to detect underlying genetic transcription factors from the corresponding genome sequence, and also employ a frequency-based signal processing method for extracting inherent features from the ChIP-sequencing data sources.

Results

Biological background and data source description

For our experiments of ChIP-seq data generation, the tissue samples were collected from patients diagnosed with breast cancer. Then the corresponding ChIP-seq data were generated with Illumina NGS platform. We collected MCF-7 ChIP-seq data with and without estrogen (E2) treatment with replicate measurements, respectively.

The NGS sequencing platform provides short read length sequences of ~36 base pairs capable of capturing ChIP-derived fragments. Then sequences are mapped back to a corresponding reference genome, where those frequently sequenced fragments will form peaks at specific regions.

Through global computational analysis of those identified peaks, it facilitates our further understanding of underlying genetic regulatory activities.

Analysis methods and measures for ChIP-seq data

Normally for examining data quality, one needs to analyze peak numbers under specific argument constraints. And we attempt to acquire optimal peak numbers by constraining specific arguments, which might be formalized as a class of optimal track analysis, illustrated as follows,

$$\begin{aligned} &\arg\max_{i} \mathrm{P}_i \circ i \in N \\ &s.t.: f_i \le \chi,\ b_i = \beta,\ p_i \le \delta \end{aligned} \tag{1}$$

where P_i denotes a set of optimal peak numbers under corresponding argument constraints, f_i stands for argument FDR, b_i for bin size and p_i for p-value threshold, χ, β and δ represent the presupposed argument values, respectively.

Herein we define a track rate function (TR) to characterize underlying data features from diverse argument pair sets (peak number and FDR), depicted as follows,

$$TR_i = \frac{ATS_i}{STS_i} = \frac{\sum_{j=1}^{M} S(j)}{\sum_{k=1}^{N} S(k)},\ i \in N \tag{2}$$

where ATS represents actual track scoring function, STS represents the shortest track scoring function, and $S(\cdot)$ denotes corresponding score value for each track step, respectively.

Through the above optimal track estimation, we extract underlying genomic features in those ChIP-seq data, as shown in the below section. Moreover, we detect transcription factors binding sites; then through a frequency power spectrum method we attempt to acquire related genomic characteristics.

For a finite random variable sequence, its power spectrum is normally estimated from its autocorrelation sequence by use of discrete-time Fourier transform (DTFT), denoted as,[6-8]

$$P(\omega) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} C_{xx}(n) e^{-jn\omega} \tag{3}$$

where C_{xx} denotes autocorrelation sequence of a discrete signal x_n, defined as,

$$C_{xx}(i,j) = \frac{E[(X_i - \mu_i)(X_j - \mu_j)]}{\sigma_i \sigma_j} \tag{4}$$

where μ and σ stand for mean and variance, respectively. In our study, for consideration of the investigated data characteristics, we use 128 sampling points to calculate discrete Fourier transform, with the related sampling frequency 1 KHz.

For two ChIP-seq samples, C_1 and C_2, to compare their alignment feature or data quality, supposing each sample has L homogeneous subfeatures quantified as CF_{1i} and CF_{2j} ($i, j \le L$), respectively; we define a new pairwise sample ratio (SR) function as below,

$$SR(C_1, C_2) = \sum_{i,j=1}^{L} \frac{CF_{1i}}{CF_{2j}} \Big/ L\ (i,j,L \in N) \tag{5}$$

where the higher the SR rate is, the more peaks ChIP-seq sample C_1 has, and vice versa; the ideal condition is that both C_1 and C_2 contains equal peak count with respect to each bin size, then the SR rate equals 1. Thus we can quantitatively compare the underlying genomic subfeatures using multi-scale SR function.

Each subfeature *CF* represents the local genomic property, and the *SR* function can characterize the global feature distribution with respect to diverse bin size or other argument of interest.

Analysis and results

Optimal track analysis with arguments' constraints

For the ChIP-seq dataset, we detect several optimal argument pairs for peak number and FDR with the corresponding argument constraints of bin size and p-value threshold.[9-13] Normally, we seek to acquire the highest peak numbers under specific argument constraints following an optimal track set. The algorithmic pseudo-code for detecting an optimal track set is illustrated in Table 1.

With the algorithmic analysis on ERα ChIP-seq data at the time point 4 hours, thus we get the optimal tracking result in Fig. 1, where black solid dot denotes a starting track point, arrows for intermediate points, and black solid square for the track end.

Based on the totally 153 pairs, there are 7 intermediate points, one start and one end, respectively, with incremental step 0.0044. From Fig. 1, the maximum peak number is 15,597 with its FDR at 21.498%; and contrarily, the minimum peak number is 508 with its corresponding FDR to 0%.

Figure 1 also indicates that under condition of current arguments, the most suitable peak number exists when bin size is selected as 100, no matter which FDR constraint is chosen in the follow-up peak-calling procedures.

According to the predefined track rate in Eq. 2, MCF-7 ChIP-seq data's track rate values are 0.6117 for peak number and 0.39 for FDR, with its interval number $N = (\pi_x - \pi_n)/\delta = 50$. Figure 2 illustrates the track rate distribution with respect to an interval number. As depicted, when interval number exceeds 40, both track rates will eventually stabilize to equilibrium, respectively. The equilibrium denotes the estimated optimal status for the peak number and FDR.

Furthermore, for further analyzing NGS sequences, we seek to detect underlying sequence-based transcription factors, which actually facilitate understanding diverse biological regulatory mechanisms.

Table 1. Algorithmic pseudo-code for detecting an optimal track set from ChIP-seq data.

Input:
π_x: maximum FDR value;
π_n: minimum FDR value;
δ: incremental step;
Output:
optimal track set: *P*.
Begin:
index $\leftarrow \pi_n$;
while index $\leq \pi_x$ do
1. search a maximum peak number candidate,
s.t. index and other arguments;
2. index $\leftarrow$ index $+ \delta$;
3. return index's information to *P*.
End

Frequency-based genomic feature extraction from ChIP-seq

The basic idea for identifying transcription factor binding sites from genome-wide NGS sequences by use of the position weight matrix concept has been presented in the references.[14,15] Using the position weight matrix, we have identified 487 transcription factor binding sites from MCF-7 ChIP-seq data at the time point 4 hours. Although all of those candidates need further experimental validation, from the computational perspective we can detect their corresponding frequency spectrum of peak number and identify the inherent genomic features.

Here we use the sample rate of 1,000 Hz, chop the signal into overlapping slice at every 5 ms, window each slice with the size of 40 ms, then apply a Fourier Transform to determine the frequency components at each slice. For standardized comparison, we shrink the frequency within the magnitude (0, 500] Hz, and normalize the spectrum within $[-40, -3]$ dB for all the genomic data spectrums.

Figure 3 illustrates the corresponding spectrum distributions for the identified peak number candidates for the original and randomized cases, respectively. We can easily find that in the original power spectrum distribution there exists regular frequency spectrum regions at 100, 200, 300 and 400 Hz (left panel), i.e. the spectrum has periodic signs of wave-like fluctuations; while for the randomized spectrum on the right panel, we cannot find such regular frequency area. These features are also indicative of the corresponding occurrence of transcription factors candidates in binding activities.

Together we analyze the relationship between those TF candidates' calculated average scores and corresponding frequency of their occurrence, as shown in Fig. 4. We find that the calculated average scores for those TF candidates mostly remain around 0.965. And

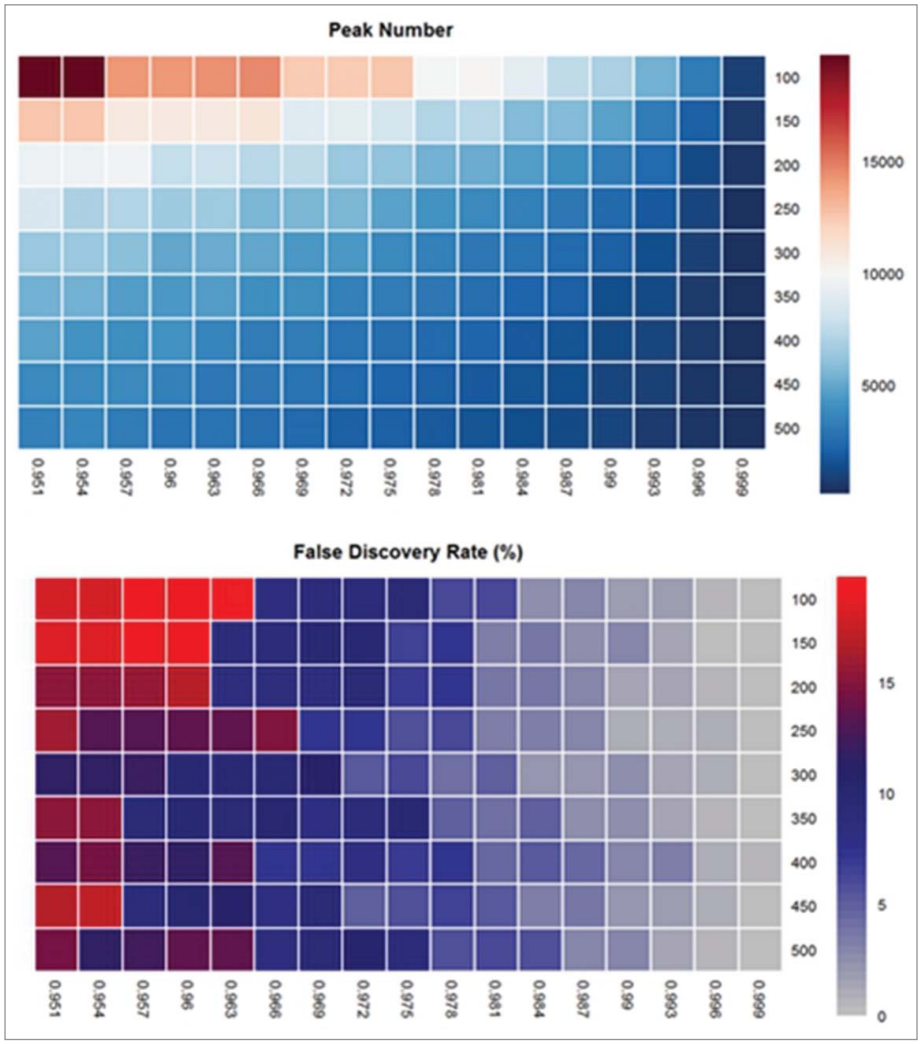

Figure 1. (Case: GSE35109): the distributions of peak number (upper panel) and corresponding False Discovery Rate (FDR, in percentage, lower panel) subject to 2 argument constraints, namely bin size (vertical axis) and p-value threshold (horizontal axis) for both panels.

the most part of TF's frequency of occurrence is lower than 0.01. Thus it provides clues for further study and validation for those candidates by use of other statistical and computational methods, for example, integrative analysis with additional profiling, PCR information, histone modification, and other epigenetic level information.

Validation of the proposed method with public ChIP-seq data

To validate the proposed method with the other public data, in the following we refer to the other public ChIP-seq data sets by Welboren et al.[16] and Hurtado et al.,[17,18] respectively.

Welboren et al. used non-sequential ChIP-seq data to map ERα-binding sites and profiled changes in

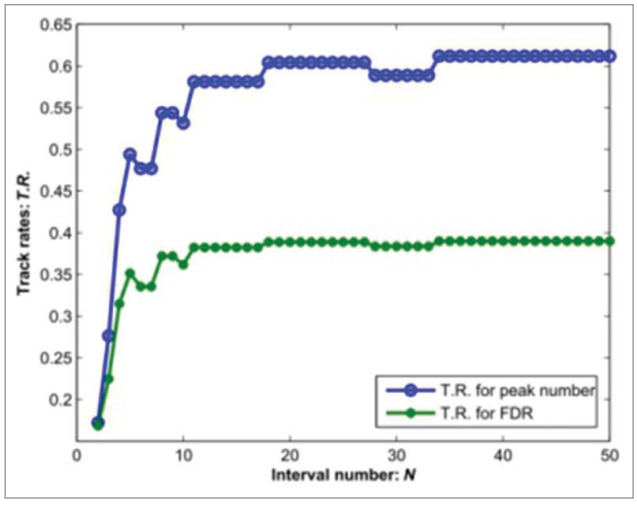

Figure 2. The track rate (T.R.) distribution plot for peak number (dark blue) and FDR (green) with respect to interval number *N*.

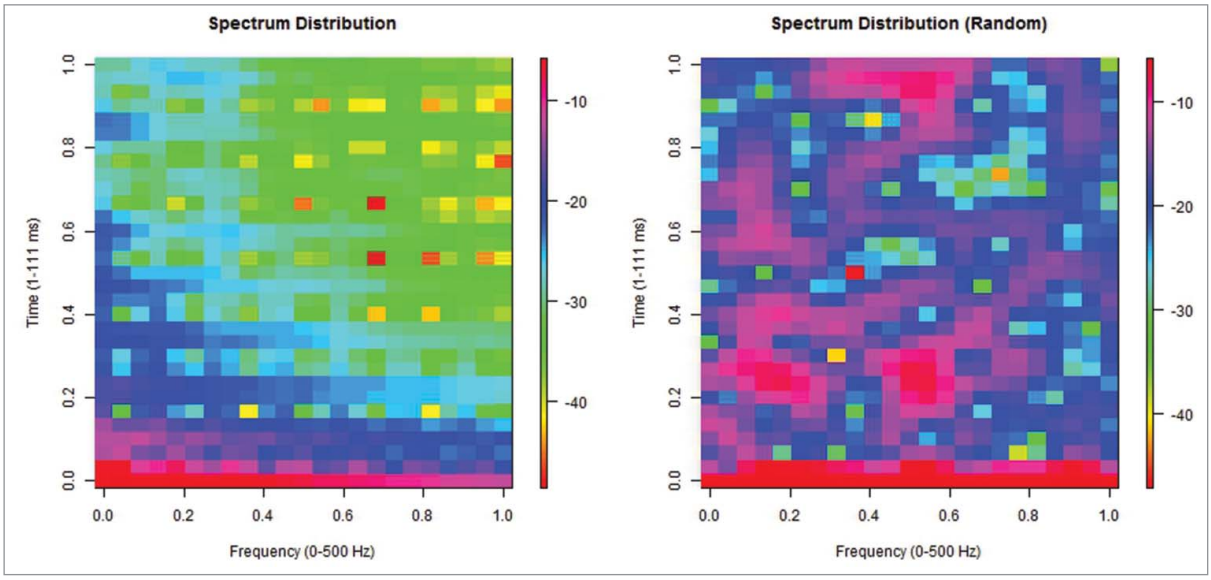

Figure 3. (Case: GSE35109): The peak number spectrum distribution based on their statistical occurrences. The sampling frequency is 1 KHz in Fourier analysis for both cases. Within the original spectrum distribution (left panel) there exists regular frequency spectrum regions along the 500 Hz range; while for the randomized case, there does not exist such regular sign.

RNA polymerase II (RNAPII) occupancy in MCF-7 cells in response to estradiol (E2), tamoxifen or fulvestrant.

We interrogate the ChIP-seq data by E2 treatment, and the supplemental Fig. S1.1 (A) illustrates the statistical distribution of peak number and corresponding FDR from totally 153 pairs. Through the global comparison, we can easily determine the optimal peak number and statistically meaningful FDR. And interestingly, we find there also exists the regular frequency spectrum pattern after normalization; but there is no regular sign for the randomized case on the right panel (supplemental Fig. S1.1 (B)).

Similarly, after a systematic investigation on the recent work by Hurtado, et al.,[17,18] we identify 153 pairs of peak number and corresponding FDR subject to diverse argument constrains (bin size and p-value threshold). The supplemental Fig. S1.2 (A) depicts the global statistical distribution of the detected 153 pairs. Furthermore, similar as the above 2 cases, we also find the regular spectrum pattern along the frequency axis of 500 Hz range on the left subplot, but not in the randomized case on the right panel (supplemental Fig. S1.2 (B)).

Feature extraction within single and multiple ChIP-seq samples

We firstly interrogate the feature from single ChIP-seq sample. Generally we perform the association analysis between signal reads number and peak number with respect to bin size, together with that between FDR and peak number.

Figure 5 illustrates the association analysis results. The left panel shows that peak number is approximately proportional to the square of signal reads number, especially for the case of steep slope with bin size = 100. The result shows that peak number amplification ratio

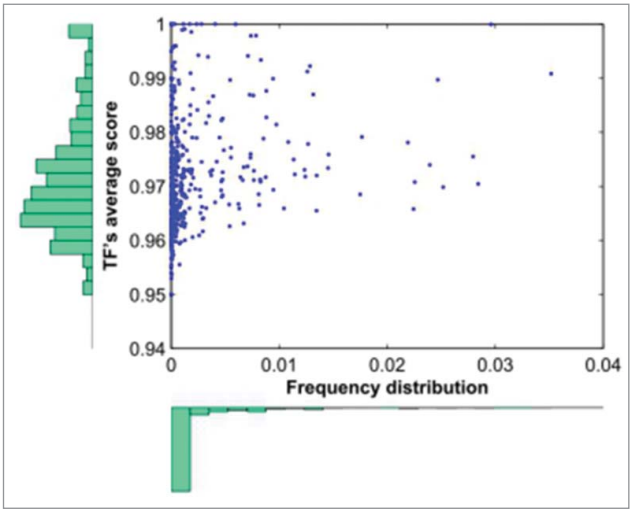

Figure 4. The statistical distribution for the identified transcription factor candidates' average score information. The horizontal axis denotes frequency range distribution of their occurrence, with maximum 0.04; and the vertical axis illustrates the TF candidates' average score distribution, with range between 0 and 1.

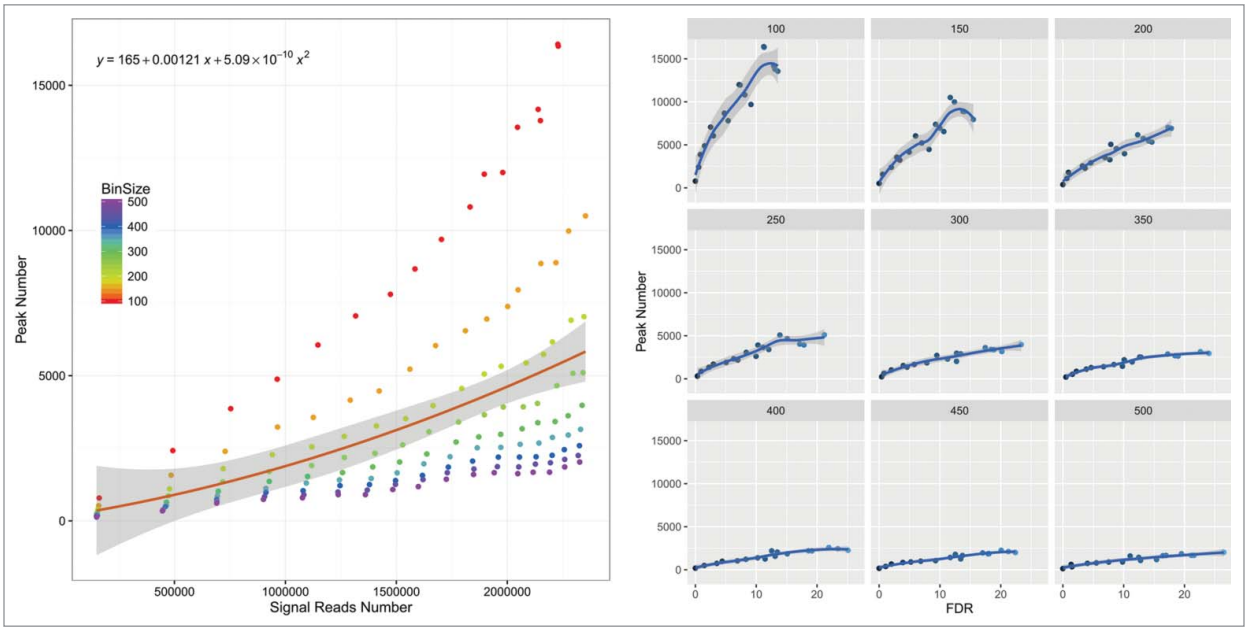

Figure 5. Association analysis on aligned peak, signal reads and FDR within single ChIP-seq sample (Case: GSE35109). Left panel illustrates the peak number is approximately proportional to the square of signal reads number, with the fitted quadratic equation on the top left; right panel indicates the association between FDR and peak number with respect to bin size.

decreases when the bin size increases. The fitted quadratic equation is given on the top left.

We find the same regular ratio decreases with the increase of bin size for the case of FDR and peak number, see the right panel of Fig. 5. Peak number changes dramatically with respect to FDR when its bin size is set at 100; on the contrary is the case with its bin size at 500. Other analyses between noise level and thresholds are given in the supplemental file.

We further perform the analyses on multiple ChIP-seq datasets in order to compare the sample differences and investigate the quality issues among replicates. We investigate the 3 public ChIP-seq data ERR022052, SRR015350 and SRR399019, and compare the peak number grouped within each bin size.

Based on Eq. 5, we can further interrogate the underlying association properties between those ChIP-seq samples. Figure 6 illustrates the analysis results between the samples, ERR022052 and SRR015350, with respect to the 9 conditions of bin size. We can find that basically peak number in ERR022052 is comparatively less than that in SRR015350, with the *SR* value ranging from 0.8310 (bin size, 500) to 0.8512 (bin size, 400); meanwhile their correlation coefficient, *CC*, is comparatively high, ranging from 0.9766 (bin size, 150) to 0.9986 (bin size, 450 and 500). Such results indicate that the peak aligned for sample ERR022052 is comparatively higher than that in sample SRR015350 at each bin size; and the aligned peak differences for the 2 samples are from 14.88% to 16.9%, with averaged 15.78%.

Other pairwise analyses between ERR022052, SRR015350 and SRR399019 are given in the supplemental Section 3.

Discussion

Within the work, we have analyzed transcription factor binding and its relevant regulatory characteristics through extracting underlying genomic data features. We utilized high-throughput ChIP-seq profiling technology to generate the NGS sequences from MCF-7 cell line, and also validated the prosed method with public ChIP-seq data.

With globally determining an optimal peak number set under relative argument constraints, we discovered the inherent genomic features, thus we can quantitatively compare and determine their qualities from diverse data sources, which guarantees the statistical confidence level of further analysis for genomic sequences, especially in the downstream interrogation of biologically meaningful results.

Through exhaustion searching for all possible peak number candidates subject to argument pairs, our method proposes an applicable approach for facilitating global investigation of underlying biological

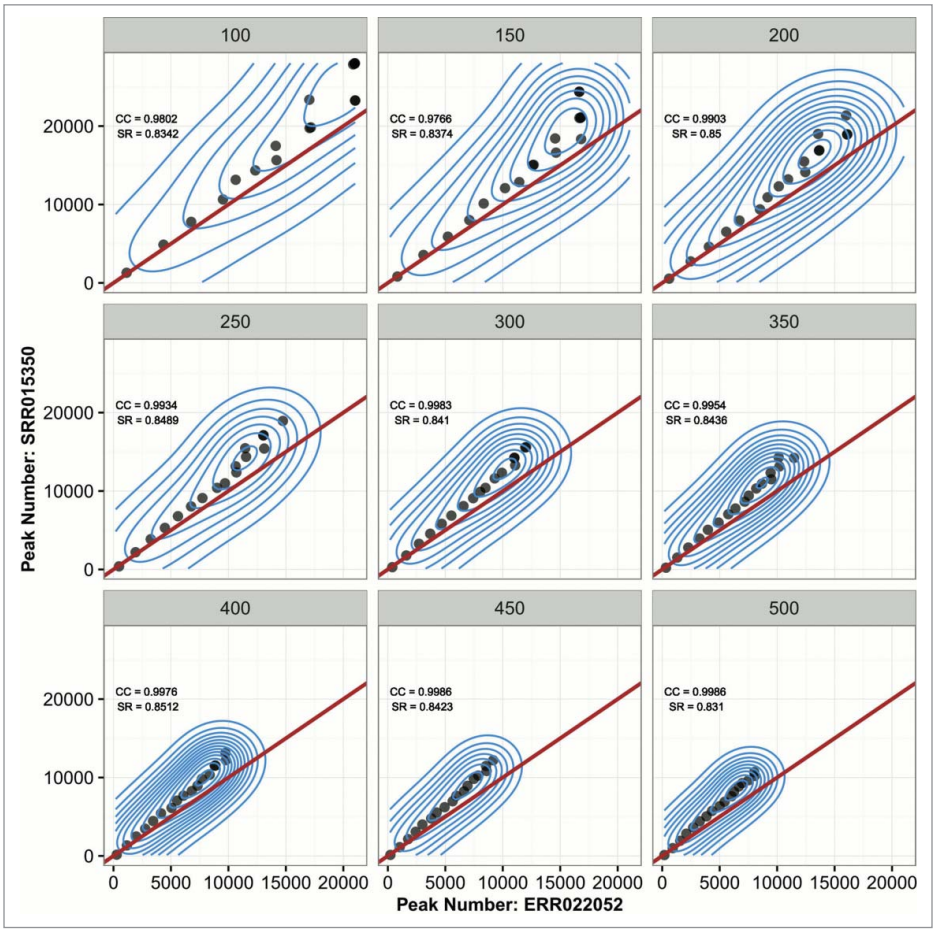

Figure 6. The association analysis between the ChIP-seq samples, ERR022052 (E-MTAB-223) and SRR015350 (GSE14664), with respect to bin size, with the corresponding *Pearson* correlation coefficient (*CC*) and comparative sample rate (*SR*) are given on the top left.

mechanisms and providing clues for experimental validation.

Furthermore, we extracted genomic signal features by use of a spectrum-based method. We identified statistical characteristics from aligned peak number candidates by use of the frequency distribution of genome-wide occurrence, and we found that there existed regular regions of high frequency spectrum, which is indicative of the existence of high occurrence of specific transcription factors candidates, but not in the randomized study cases.

We further proposed a statistical measure in quantitative comparison of peak-related differences within a single sample or among multiple samples, e.g. aligned peak count, signal reads and FDR, etc., thus it provides a practical approach for biologists in their NGS experiment and analysis.

Lastly we validated such conclusions with 2 recent public ChIP-seq data sets, and the analysis results are basically in accordance with the results from our own data.

Besides flexibility and enhanced accuracy in high-throughput genomic sequencing, NGS technology demands comprehensive analysis of genomic features from the generated sequences to exploit its potentialities. Thus, as to future directions, we will further combine those findings with other information through integrative analysis of the underlying transcription regulatory characteristics, thus it will put forward biological meaningful and clinically applicable conclusions.

Availability

The ChIP-seq data interrogated in this work are deposited at NCBI GEO with the accession: GSE35109 and GSE14664; and EMBL-EBI with accession: E-MTAB-223.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This work was supported by the Natural Science Foundation of Jiangsu, China (BE2016655 and BK20161196), the Fundamental Research Funds for China Central Universities (2016B08914) and Changzhou Science & Technology Program (CE20155050). This work made use of the resources supported by the NSFC-Guangdong Mutual Funds for Super Computing Program (2nd Phase), and the Open Cloud Consortium sponsored project resource, which supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation (USA) and major contributions from OCC members.

References

[1] Mardis ER. ChIP-seq: welcome to the new frontier. Nat Meth 2007; 4:613-4; https://doi.org/10.1038/nmeth0807-613

[2] Martinez GJ, Rao A. Cooperative transcription factor complexes in control. Science 2012; 338:891-2; PMID: 23161983; https://doi.org/10.1126/science.1231310

[3] Kilpinen H, Barrett JC. How next-generation sequencing is transforming complex disease genetics. Trends Genetics 2013; 29:23-30; PMID:23103023; https://doi.org/10.1016/j.tig.2012.10.001

[4] Chikina MD, Troyanskaya OG. An effective statistical evaluation of ChIPseq dataset similarity. Bioinformatics 2012; 28:607-13; PMID:22262674; https://doi.org/10.1093/bioinformatics/bts009

[5] Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 2012; 13(12):840-52; PMID:23090257

[6] Oppenheim AV, Schafer RW. Discrete-time signal processing, 3rd edition. 2010. Upper Saddle River, NJ: Prentice Hall.

[7] Tang B, Hsu HK, Hsu PY, Bonneville R, Chen SS, Huang TH, Jin VX. Hierarchical modularity in ERα transcriptional network is associated with distinct functions and implicates clinical outcomes. Scientific Reports 2012; 2:875.

[8] Wang SL, Zhu YH, Jia W, Huang DS. robust classification method of tumor subtype by using correlation filters. IEEE/ACM Transactions Computational Biol Bioinformatics 2012; 9:580-91; https://doi.org/10.1109/TCBB.2011.135

[9] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol 2008; 9:R137; PMID:18798982; https://doi.org/10.1186/gb-2008-9-9-r137

[10] Lan X, Bonneville R, Apostolos J, Wu W, Jin VX. W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Bioinformatics 2011; 27:428-30; PMID:21138948; https://doi.org/10.1093/bioinformatics/btq669

[11] Spyrou C, Stark R, Lynch A, Tavare S. BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics 2009; 10:299; PMID:19772557; https://doi.org/10.1186/1471-2105-10-299

[12] Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 2008; 24:1729-30; PMID:18599518; https://doi.org/10.1093/bioinformatics/btn305

[13] Zhu L, Guo WL, Deng SP, Huang DS. ChIP-PIT: Enhancing the analysis of ChIP-seq data using convex-relaxed pair-wise interaction tensor decomposition. IEEE/ACM Transactions Computational Biol Bioinformatics 2016; 13:55-63; https://doi.org/10.1109/TCBB.2015.2465893

[14] Cheng AS, Jin VX, Fan M, Smith LT, Liyanarachchi S, Yan PS, Leu YW, Chan MW, Plass C, Nephew KP, et al. Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-α responsive promoters. Mol Cell 2006; 21:393-404; PMID:16455494; https://doi.org/10.1016/j.molcel.2005.12.016

[15] Ou YY, Chen SA, Gromiha MM. Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. Proteins: Structure Function Bioinformatics 2010; 78:1789-97

[16] Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG. ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. EMBO J 2009; 28:1418-28; PMID:19339991; https://doi.org/10.1038/emboj.2009.88

[17] Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet 2011; 43:27-33; PMID:21151129; https://doi.org/10.1038/ng.730

[18] Cheung E, Ruan Y. Determination of transcription factor binding. Nat Genet 2011; 43:11-12; PMID:21217639; https://doi.org/10.1038/ng0111-11