doi:10.1093/nar/gks1010

A co-localization model of paired ChIP-seq data using a large ENCODE data set enables comparison of multiple samples

Kazumitsu Maehara[1,2], Jun Odawara[1,3], Akihito Harada[1], Tomohiko Yoshimi[4], Koji Nagao[5], Chikashi Obuse[5], Koichi Akashi[3], Taro Tachibana[4], Toshio Sakata[2] and Yasuyuki Ohkawa[1,]*

[1]Department of Advanced Initiative Medicine, Faculty of Medicine, Kyushu University, JST-CREST, Fukuoka 812-8582, [2]Department of Human Science, Faculty of Design, Kyushu University, Fukuoka 815-8540, [3]Department of Medicine and Biosystemic Science, Faculty of Medicine, Kyushu University, Fukuoka 812-8582, [4]Department of Bioengineering, Graduate School of Engineering, Osaka City University, Osaka 558-8585 and [5]Division of Molecular Life Science, Graduate School of Life Science, Hokkaido University, Sapporo 001-0021, Japan

Received July 13, 2012; Revised September 28, 2012; Accepted October 1, 2012

ABSTRACT

Deep sequencing approaches, such as chromatin immunoprecipitation by sequencing (ChIP-seq), have been successful in detecting transcription factor-binding sites and histone modification in the whole genome. An approach for comparing two different ChIP-seq data would be beneficial for predicting unknown functions of a factor. We propose a model to represent co-localization of two different ChIP-seq data. We showed that a meaningful overlapping signal and a meaningless background signal can be separated by this model. We applied this model to compare ChIP-seq data of RNA polymerase II C-terminal domain (CTD) serine 2 phosphorylation with a large amount of peak-called data, including ChIP-seq and other deep sequencing data in the Encyclopedia of DNA Elements (ENCODE) project, and then extracted factors that were related to RNA polymerase II CTD serine 2 in HeLa cells. We further analyzed RNA polymerase II CTD serine 7 phosphorylation, of which their function is still unclear in HeLa cells. Our results were characterized by the similarity of localization for transcription factor/histone modification in the ENCODE data set, and this suggests that our model is appropriate for understanding ChIP-seq data for factors where their function is unknown.

INTRODUCTION

Chromatin immunoprecipitation (ChIP) is a quantitative measurement of protein–DNA interactions, but it is site specific. With the invention of deep sequencing technology, ChIP has extended its potential for understanding the epigenetic state in the whole genome, including histone modification, transcription factor binding and chromatin accessibility (1). The epigenome project known as Encyclopedia of DNA Elements (ENCODE) has accelerated the accumulation of ChIP by sequencing (ChIP-seq) data exponentially (2).This accumulation of ChIP-seq data has enabled the prediction of unknown protein function by comparing each ChIP-seq data. Ideally, as genome projects have been used for comparative genomics (3), these epigenomic data should be used for identifying candidate epigenomic events or identifying candidate factors for comparison.

However, comparison of different ChIP-seq data has been severely impaired by 'background' noise derived from various factor (4). This background varies in its quality and amount by experimental conditions, which is due to the specificity of antibodies or immunoprecipitation efficiency derived from fixation conditions or immunoprecipitation buffer conditions. Additionally, a deep sequencer itself also causes noise, such as bias of sequenced reads (4). Even sequenced reads that potentially map to multiple sites on the genome can also yield background (4,5). Identification of signals from a mixture of specifically immunoprecipitated signal and background noise is required.

*To whom correspondence should be addressed. Tel: +81 92 642 6216; Fax: +81 92 642 6216; Email: yohkawa@epigenetics.med.kyushu-u.ac.jp

To pick up signals from this mixture of signal and noise, various types of software for treating ChIP-seq data against control data, such as input or no antibody control, have been designed (6,7). A 'peak' is detected as a binding site of a target protein by evaluating the statistically significant accumulation of reads in this mixture. This process is called 'peak calling'. There are several types of software for call peaks, such as MACS (7) and PeakSeq (6). These peak-calling methods have been reported to detect peaks in each sample, while they also identify different qualities of peaks among various ChIP-seq data. This difference has been reported as the sensitivity of a peak caller (8).

The variety of methods for peak calling has resulted in a variety of the number of peaks as output from the same data set (4). In most software for peak calling, a parameter to set a threshold for statistical significance can be determined by users based on the experimental conditions (9,10). In the case of well-known factors, users can evaluate which is the most appropriate parameter by referencing the data obtained from ChIP-quantitative polymerase chain reaction or other experimental validations (10). However, in the case where the function or localization of a factor is unknown, it is more difficult to obtain the appropriate threshold because of a lack of reference data. In either of these cases, it is possible that the number of called peaks in a public database is overestimated or underestimated compared with the number of 'true' peaks.

The variation in peak number of ChIP-seq data affects the comparison of different ChIP-seq data. For example, to address the molecular function of a transcription factor, it has recently been reported a change in distribution, such as histone modification or chromatin accessibility, in two different ChIP/accessibility-seq data (11). To perform this type of comparison, it is critical to normalize two different called peaks from each data (12,13). However, there is no effective method to normalize two different ChIP-seq data. The ideal method to normalize two ChIP-seq data is to adjust the conditions for ChIP-seq, including antibodies, cells, controls, such as input or control antibodies, and IP protocol, and call peaks by the same peak caller with the same parameter sets. This approach is effective for comparing ChIP-seq data in-house, but it limits the data sets for comparison (in-house only). A practical approach to compare ChIP-seq data is to ignore the total number of peaks and then evaluate the change in distribution of the peaks (11). This type of qualitative evaluation could eliminate normalization of called peaks and then the change in shapes of peak distribution could be evaluated. However, if the distribution of peaks is similar, for example, a reduction in factor binding to the genome, quantitative evaluation is still required.

We report a novel method to compare different ChIP-seq data. Our method can model the relation of paired ChIP-seq data and can correct biases caused by different parameters in the protocol and software, by using a large amount of ENCODE ChIP-seq data.

MATERIALS AND METHODS

Cell culture

HeLa cells were cultured in Dulbecco's modified Eagle's medium supplemented with streptomycin (100 µg/ml; Nacalai Tesque, Kyoto, Japan) as described in (14). C2C12 cells were cultured in Dulbecco's modified Eagle's medium supplemented with 20% fetal bovine serum (Lifetechnology, CA, USA) and streptomycin (100 µg/ml; Nacalai Tesque, Kyoto, Japan), as described in (11).

Immunization of rats and production of monoclonal antibodies

Anti-rat monoclonal antibody HeLa Polymerase II C-terminal domain (CTD) serine 7 phosphorylation (S7ph) was generated against the peptide (SPTSP SYSPTSPSphYSPTSPS), as described by Sado *et al.* (15). Briefly, WKY/Izm rats (10-week-old females, Japan SLC, Shizuoka, Japan) were immunized with the peptide. After 13 days, iliac lymph nodes were isolated and the separated cells were fused to mouse myeloma Sp2/0-Ag14 cells in polyethylene glycol (PEG1500, Merck, Darmstadt, Germany) solution. At 7 days post-fusion with HAT selection, the hybridoma cells were screened by an enzyme-linked immunosorbent assay against the peptide. Positive clones that reacted to S7ph, but not HeLa Polymerase II CTD serine 5 phosphorylation (S5ph) peptide (SPTSPSYSPTSphPSYSPTSPS) and HeLa Polymerase II CTD serine 2 (S2ph) peptide (SPTSPSYSphPTSPSYSPTSPS), were established monoclonally. H3.1 monoclonal antibody (1D4F2, hybridoma supernatant, 1 µg) was used for ChIP-seq, as described in (11).

ChIP and deep sequencing

Cells were fixed by 1% formaldehyde for 5 min at room temperature. ChIP was performed as described by Odawara *et al.* (14). ChIPed DNA was sequenced by the Genome Analyzer GAIIx (Illumina K.K., CA, USA). The reads of S7ph were aligned to the human genome (hg19) and the reads of H3.1 were aligned to the mouse genome (mm9) using bowtie (version 0.12.7) software (parameter: -v 3 -m 1). ChIP-seq data for S2ph, S5ph and H3.3 in undifferentiated C2C12 cells are obtained from DNA Data Bank of Japan (DDBJ) (11,14). Peak detection and identification of binding sites of S7ph and H3.1 were obtained by MACS (version 1.4.1). The parameters for MACS were '-bw 538' for S7ph, '-bw 292' for H3.1 and others were software's default.

ENCODE data set

Called peaks of all human and mouse ENCODE data sets were obtained from the MySQL database at UCSC (mysql.cse.ucsc.edu). All tables in which the name begins with 'wgEncode' were used. Detailed information of the data was obtained from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC and http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC.

Each 'files.txt' in the directories contains detailed information of data, including cells, antibodies, type of data (e.g. ChIP-seq and FAIRE-seq), laboratory, replicates, treatment, controls (e.g. IgG, input or none), submitted date and accession number of the data.

Peak localization plot

In the peak localization plot of ENCODE data, the X-axis ranged from −0.5 to 1.5. The value of −0.5 is 5 kb downstream from the transcription start site (TSS). The region 0–1 ranges from TSS to TES. The value of 1.5 is 5 kb downstream from TES. We used 20 374 human genes as the total number of genes of which the definition was obtained from UCSC's 'knownGene' table.

Co-localization plot

A similar calculation as that for the peak localization plot was also performed for the co-localization plot. The x-axis ranged from ±5 kb from the center of the peaks in the ENCODE data set. The y-axis shows the total depth of peaks at each distance. We assumed that all called peaks had a depth of 1. Therefore, the number of stacked peaks was the total depth.

The 'fitted' line was calculated from the estimated parameters α, β and p. The estimated distribution f is then multiplied by s(sum of the depths; area) to get the right side of the fitted shape, and we flipped the curve symmetrically against $x = 0$ to draw the left side.

Computational resource

Our calculations were performed in high-performance computer at DNA Data Bank in Japan (Mishima, Japan) and at Research Institute for Information Technology, Kyushu University (Fukuoka, Japan).

The first calculation step is aggregating target peaks at reference peaks. There are 1744 human reference data in ENCODE, and we needed to calculate 1744 pairs with one target ChIP-seq data. Each process varies in the required computation time and memory resource from 11.54 to 1164.0 s and from 216 to 592 MB in case of the S2ph data. Each calculation process is independent; therefore, we could completely parallelize all processes over a high-performance computing system. The computation time is therefore ∼20 min (1164.0 s) if we can use sufficient computation nodes to run 1744 processes.

The second calculation step is the regression (curve fitting) and plotting process. It takes ∼2 h and requires 12 GB of memory resource for the S2ph data.

Software and all outputs for S2/S5/S7 phosphorylation and H3.1/H3.3

We uploaded our software and all outputs from the following: http://chromatin.med.kyushu-u.ac.jp/pol2encode/.

The outputs of human and mouse data sets contained plots and tables of 1744 (human) and 248 (mouse) factors for each analysis.

RESULTS

Co-localization model

To compare two different ChIP-seq data, first, we focused on methods for evaluating the accumulation of a factor around the TSS (11). This approach could visualize the dependency of a factor on transcriptional regulation. It is also applicable to extend this approach to compare two different ChIP-seq data. We extended their approach to estimate signal accumulation around the binding sites of various factors. In other words, we aggregated 'peaks' around 'peaks'. The 'peak' means a detected region of the occurrence of a biological event resulting from ChIP-seq data analysis software (6,7,16).

When comparing factor A with the ChIP-seq data of factor B, it is possible to calculate the accumulation of the total peaks of factor A at the locations of the peaks of factor B (Figure 1A 'distribution'). The accumulation of approaching peaks in a range of distance can be evaluated as the dependency of distance from one place to another; in other words, the probability of occurrence of some event at a certain distance ('distribution'). However, if we compare multiple ChIP-seq data (more than two), 'distribution' itself is not sufficient because it is not able to evaluate the strength of relationship. For example, there are some factors that show a similar distribution, such as H3K27ac and RNA Pol II S2ph/S5ph (17), or MyoD and myogenin in myogenic cells (18). Therefore, it is essential to estimate a reasonable measure for evaluating the relationship from the shape of the 'distribution'.

To construct a model for evaluating the relationship between two factors, we made the following assumption: if two factors have a relation, their location depends on the distance between them otherwise they are located on sites regardless of each distance. A presumable stochastic property arises from the assumption that the frequency of detection of factor A tends to be high when it is close to factor B if there is a cooperative relationship between them, and conversely, the frequency of detection of factor A is low when it is close to factor B if there is an exclusive relation (i.e. one factor is preventing binding or detecting another factor) as shown in Figure 1A 'profiling'. Otherwise, random detection of factors A around factor B results in uniform distribution. We assumed a uniform background as overlapping of invalid peaks or overlapping of unrelated peaks. This assumption has been used to evaluate co-localization of nuclear factors in the nuclei by immunocytochemistry (19).

To model such a relation between factor A and factor B, we assumed that the distribution was a mixture of a uniform distribution to represent a random approach and a geometric distribution to represent an exponential approach of two related factors (Figure 1A 'model'). The probability mass function is

$$f(x_i|\alpha, \beta, p) = \alpha p(1 - p)^{x_i} + \beta \frac{1}{N},$$

where $\alpha + \beta = 1$ to extract the mixture ratio, and $x_i = 0, 1, \ldots, N - 1$ is the distance from factor B in base pairs. f is a mixture distribution that could represent a

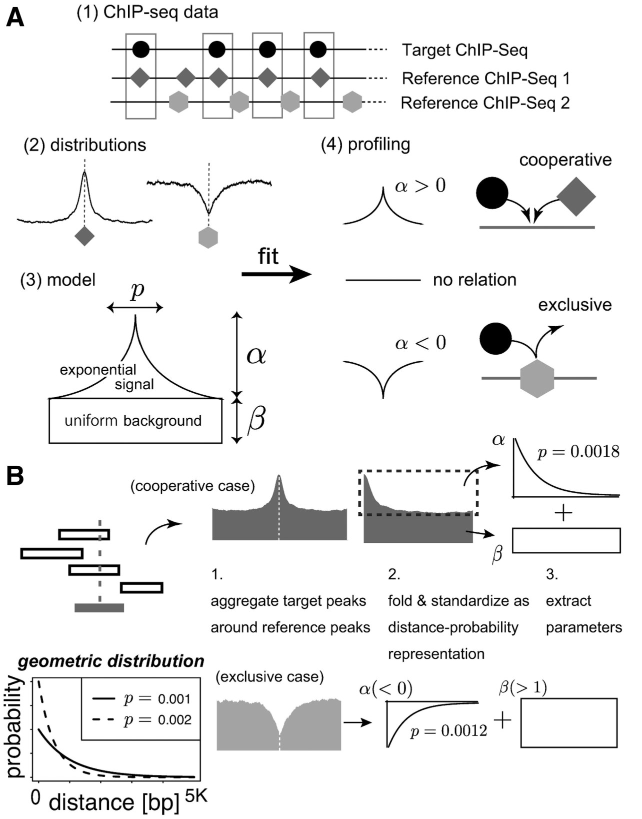

Figure 1. Schematic drawings of the co-localization model. (**A**) The concept of co-localization model (1) ChIP-seq data: called peaks of different ChIP-seq vary in their location (2) 'distributions': distribution of the target ChIP-seq peaks around the center of all peaks in other ChIP-seqs (3) model: 'distribution' is a mixture of exponential signals and uniform background. α and β represent a mixture ratio. p represents the steepness of the shape. (4) Profiling: the model represents the shape from 'cooperative' to 'exclusive'. (**B**) Workflows: the model fitting examples of 'cooperative' (upper line) and 'exclusive' (lower line) types of accumulation are illustrated. Peaks were aggregated at reference peaks and translated into distance from a reference factor (*X*-axis) and the probability of peak detection (*Y*-axis) representation. Each mixture ratio of valid signal and background was then estimated as their shape parameter in the co-localization model. Some examples of parameter p in geometric distribution are also shown in the bottom-left panel. p affects the steepness of the slope.

shape of an accumulation with parameters α, β and p. The first term $p(1-p)^{x_i}$ is a geometric distribution. A larger p results in a steeper slope, while a smaller p results in a gentler slope (Figure 1B). The second term, $1/N$, is a uniform distribution. Its height is constant over the window (from 0 to $N-1$). These two types of distribution are mixed with the ratio of $\alpha : \beta$ (Figure 1B). We used $N = 5001$, which is wide enough to reach a plateau of distribution. This range is commonly used in the field of ChIP-seq data analysis (14,20,21). Our model is characteristic in that it can represent an exclusive relation to allow for a negative α (i.e. flipped geometric distribution, as shown at the bottom line of Figure 1B) in contrast to ordinal distribution. This property comes from the presence of a uniform background.

To provide a non-negative constraint to the parameters β and p, we set $\alpha = 1 - \omega_1^2$, $\beta = \omega_1^2$ and $p = \omega_2^2$ and now $\boldsymbol{\omega} = (\omega_1, \omega_2)^{\mathrm{T}}$ is actually the estimating parameters. The sum of squared error is

$$E(\boldsymbol{\omega}) = \left\| \frac{1}{s}\mathbf{y} - \mathbf{f} \right\|^2,$$

where y_i is the total depth of the peaks at x_i and the vector of each of the total depth of the peaks at each distance is $\mathbf{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$. The formula $s = \sum_{i=1}^{N} y_i$ is the total sum of the depth and $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_N))^{\mathrm{T}}$ is the vector of fitted values (Supplementary Figure S1). We added a regularization term to obtain stability of the numerical optimization.

$$E'(\boldsymbol{\omega}) = E(\boldsymbol{\omega}) + \lambda||\boldsymbol{\omega}||^2.$$

The optimal parameters were obtained to solve:

$$\min_{\boldsymbol{\omega}} E'(\boldsymbol{\omega}),$$

with the Gauss–Newton algorithm. We used $\lambda = 0.01$, which was an arbitrarily chosen small value, as a fixed regularization parameter. It worked well for obtaining numerical stability in our case. As a result, we were able to distinguish meaningful signals and meaningless signals accumulation with this model.

We then proposed a score to evaluate the strength of the relationship between two factors. The estimated parameter α represents the percentage of related peaks of factor A in all peaks around factor B. Therefore, we defined the co-localization score as

$$\text{score} = \alpha \times \frac{\begin{pmatrix}\text{the number of peaks of factor A within} \\ N \text{ bp from factor B}\end{pmatrix}}{(\text{the total number of peaks of factor B})},$$

which can be interpreted as the 'valid' number of overlapping peaks per site (Supplementary Figure S2). The score is intuitively interpreted as follows: when it becomes a positive value, this indicates a cooperative relation, and when it becomes a negative value, this indicates an exclusive relation.

α/β can be used as the signal-to-noise ratio (S/N), but it can be a negative value. Another parameter, p, affects the steepness of the distribution's shape. Therefore, we named it the 'concentration' parameter. p is a variance parameter of geometric distribution; therefore, a larger p results in a distribution with a steeper slope, while a smaller p results in a distribution with a gentler slope. It could be biologically interpreted as the accuracy of IP techniques or two different factors having a short-range relation or a long-range relationship. We further evaluated the effect on resulting score of selecting the regularization parameter λ at two different settings and observed that the effect was small enough to ignore the differences (Supplementary Figure S3).

We designed a software program, which outputs graphs of distributions with paired factors, and a table of estimated parameters for the co-localization model.

Estimation of parameters in the model using practical data

We demonstrated how the co-localization model explains the shape of the distribution and how it was affected by relaxing the threshold of peak-calling software to evaluate our model with real data. We used ChIP-seq data of S2ph (14). We considered that S2ph is suitable to evaluate in our model, because Odawara *et al.* showed that ChIP-seq peaks of PolII-S2ph accumulate in the active transcriptional region and active transcriptional factors are well-studied targets. Currently, approximately hundreds of NGS data of different targets (i.e. histone modification, transcriptional factors and chromatin structure) are available in HeLa cells at ENCODE. These public data sets are useful to compare with Pol II-biding sites and various events in DNA. Additionally, there are data of already called peaks of data obtained in ENCODE projects (2) in UCSC. Parameters of the co-localization model were estimated from S2ph data paired with all of the data obtained in ENCODE projects. We named such extracted parameters of the model from the ENCODE data set paired with ChIP-seq data as the 'ENCODE profile'.

We selected three outputs as representatives from our software. The first output was GCN5 (Figure 2A), which had a high score relative to all scores of S2ph (score: 0.611; concentration: 0.020). The ChIP-seq data of GCN5 had 1186 peaks, which were relatively small, but 2801 peaks of S2ph overlapped with them. The fitted curve explained the exponentially decaying tail. H3K79me2 (Figure 2B) had a low score, but the peaks of S2ph were tightly concentrated around the center of H3K79me2 (score: 0.001; concentration: 0.073). Despite the shape of the background being uneven, the fitted curve only captured the shape around the center of H3K79me2. This selective capturing comes from the simplicity of our model. Figure 2C shows that S2ph accumulation was suppressed around the center of H3K36me3. Flipped distribution, which is caused by a negative α, can represent such a shape. Supplementary Table S1 shows the results of the estimated parameters for the ENCODE data set of HeLa ordered by the co-localization score. We visualized the results as points on a parameter space spanned by the co-localization score and concentration parameter (Figure 2D).

The points of each CTCF data came from three different laboratories and contained one replication and were closely located as shown in Figure 2D. It appeared that these data had a similar relation to S2ph. Some characteristic shapes of the factors, shown in Figure 2A–C, can be easily found by this plot without following each number shown in Supplementary Table S1.

Consequently, we succeeded in obtaining measurements of co-localization without a background signal and the strength of the concentration by our model. However, we could not conclude that these values are useful for determining which is the most or more related factor to S2ph in the ENCODE data set. This is because there was still the problem of scale variation, as already mentioned, in any called peaks. Therefore, the values do not directly reflect the strength of the relation.

Number of called peaks and co-localization score

To correct the problem of variation in scale and to enable the comparison of co-localization scores directly, we assessed the effect of enlargement of the total number of peaks in the S2ph data. We relaxed the threshold parameter of peak calling to obtain an enlarged number of peaks. Figure 2E shows that it was easy to change our score for the same factor by changing the threshold. This is why our scoring method depended on the number of peaks.

The parameter of 'P-value = 1e−5', which is the software's default, yielded 99 487 peaks and 'P-value = 1e−2', which is a relatively relaxed threshold, yielded 268 060 peaks. There was an ~2.7 times difference between these values, but the difference in their co-localization scores was ~1.5 times. The R^2-value was 0.9937 and there was a high linear relation. Therefore, all of the scores were amplified by a constant scaling factor caused by enlarging the total number of peaks, and it was still possible to view the relative locations of each point (Figure 2D). Furthermore, an important property of the scale difference is that two scores for one factor are only different regarding the scale caused by the different number of peaks. Similarly, scores of two different factors with different numbers of peaks are only different regarding the scale if they have equivalent relations with another factor. For example, in a situation where there are three different factors A, B and C, the relations of the pair (A,C) and the pair (B,C) are considered as equivalent if there is only a scale difference between the score of (A,C) and (B,C). Once the variance in scale in the numbers of called peaks is corrected, the true difference between them can be estimated. Comparison with two or more factors using a large ENCODE data set empirically enables correction of the scale. The situation where there is a lot of unrelated data in ENCODE with a 'target' ChIP-seq data is expected and these unrelated factors can reveal the scale difference.

Factors affected by Pol II CTD S2/S5/S7 phosphorylation changes

We discussed how to correct for the variety of scale in the previous section. We then predicted the functional relevance of unknown protein binding, by attempting to identify factors, which have changes in their 'scale-corrected' co-localization scores in response to changes in phosphorylation state, as an application for multiple sample comparison. We compared HeLa Pol II CTD S2, S5 and S7 phosphorylation (S2ph, S5ph, S7ph) against the ENCODE data set of HeLa cells.

S7ph is not well characterized compared with other types of phosphorylation, including S2ph, which is involved in active transcription, and S5ph, which is involved in active and pausing states. The distribution of S7ph in HeLa cells is still unclear. We and others previously demonstrated that S2ph preferentially recognizes active transcription states, while S5ph is localized at TSSs in HeLa cells (14,22,23). Based on this finding, we generated an antibody that specifically recognizes the S7ph state of the RNA Pol II CTD domain

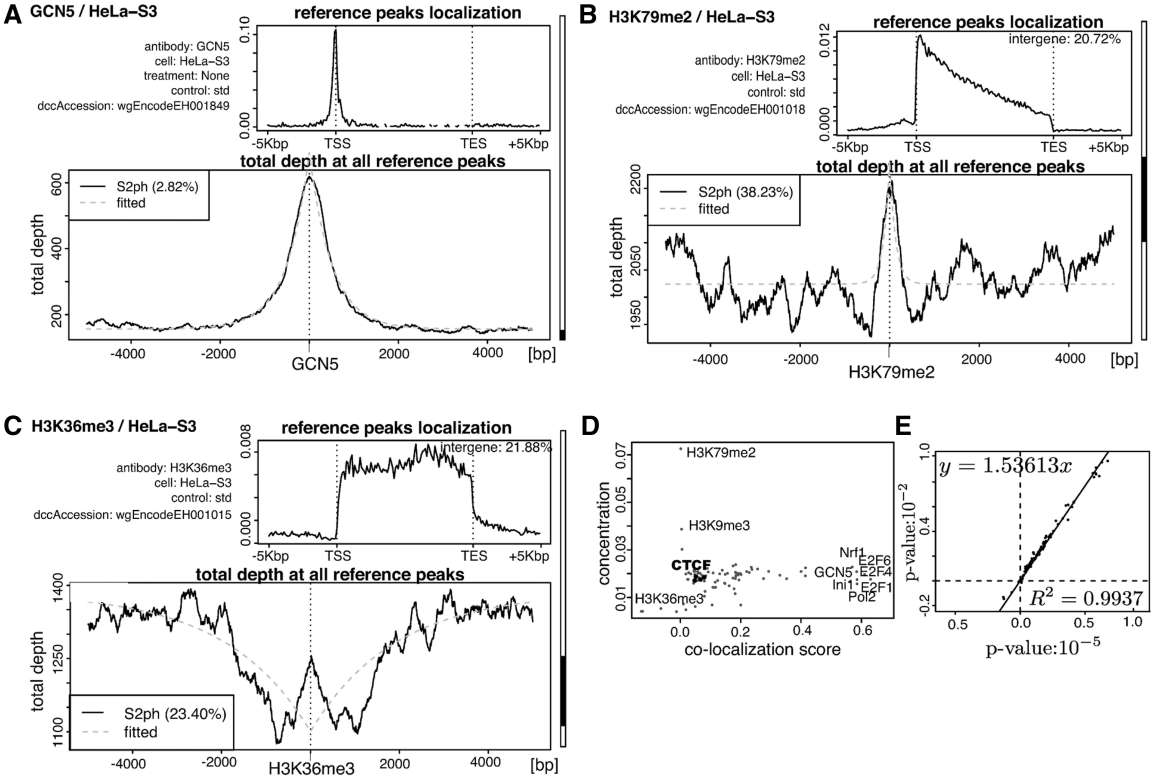

Figure 2. Evaluation of co-localization model using S2ph with ENCODE data sets. Graphs show paired analysis by our software. The used antibody/cell of ENCODE data is shown on the top of the graph (upper left). Detailed information on the ENCODE data is described under the title. Localization of the ENCODE peaks is shown. The displayed region is from 5 kb upstream to 5 kb downstream for all genes. The *y*-axis shows the total count of each peak's center divided by the number of genes for all genes in humans. Peaks that were outside of the region were noted as the percentage of the inter-gene region (upper right). The *x*-axis shows the distance from the center of the ENCODE peaks. Both sides of the positive and negative distances were plotted. The *y*-axis shows the total depth of S2ph peaks at each distance. A dotted gray line fitted to the distribution of the S2ph peaks with our model was also plotted. The shape is relatively symmetric at $X = 0$ because the orientation of transcription was ignored here, unlike localization of the ENCODE peaks plot. The percentage of S2ph peaks in the region, which is within 5 kb from the center ENCODE peaks, is noted in the left-top legend (bottom). The rightmost bar consists of black and white rectangles in each panel and is a type of Venn diagram. The vertical length from the bottom of the white rectangle to the top of the black rectangle indicates the number of ENCODE peaks. The length from the top of the white rectangle to the bottom of the black rectangle indicates the number of S2ph peaks. The length of the black rectangle indicates the number of overlaps between peaks of each set of compared data (right). (**A**) A transcriptional factor with a high score against S2ph; (**B**) a histone modification with a low score and high concentration against S2ph; (**C**) a histone modification with a negative score (exclusively related) against S2ph; (**D**) the ENCODE profile for S2ph data paired with 90 types of HeLa data sets. The points in the parameter space, spanned by the co-localization score and the concentration parameter, were plotted. The *x*-axis shows the co-localization score. The *y*-axis shows the value of concentration parameter. (**E**) A scatter plot for co-localization scores of S2ph at different threshold parameters of MACS. The *x*-axis shows the scores at *P*-value = 1*e*−2 and the *y*-axis shows the scores at *P*-value = 1*e*−5. The line is derived from linear regression analysis. The R^2 and estimated coefficient of the regression are also displayed.

(Supplementary Table S2). Our antibody for S7ph is quite specific. It does not detect phosphorylated peptides of S2ph and S5ph (Supplementary Table S2).

We obtained 99 487 peaks for S2ph, 40 355 for S5ph and 24 839 for S7ph by MACS, of which the parameters were defaults of the software, except for S7ph (*P*-value: 1e−3). We then calculated all the ENCODE profiles in HeLa on each S2/S5/S7ph. First, we created a scatter plot (Figure 3A) to determine the similarity of each ENCODE profile.

We found that the scores were all highly correlated. Although there were some 'affected' (i.e. far from the regression line) factors by changes in phosphorylation, overall, they had a similar relation to factors in the ENCODE data set. The scale difference caused by the difference between our data set in the total number of peaks was still present. However, Figure 3A shows that the scale difference was easily able to be corrected by simple linear regression. The estimated ratio of the scale of co-localization scores S2ph:S5ph:S7ph was 1.00:2.22:0.52 and we divided each score by the ratio to S2ph.

We then placed all scale-corrected scores on the surface spanned by each difference of the scores between the same factor to further focus on factors that were affected by changes in phosphorylation (Figure 3B). Each length of the points perpendicular to the regression line (Figure 3A) determines each position.

A factor 'Pol2 (phosphoS2)' was clearly located in the S2ph cooperative region (Figure 3B). The factor is suitable for validating our results, because it is expected to co-localize with S2ph. Therefore, our method successfully demonstrated the specificity of our antibody. Some of the E2 family (E2F1, E2F6 and HA-E2F1) and Ini1 were located in the S7ph exclusive region, which suggested

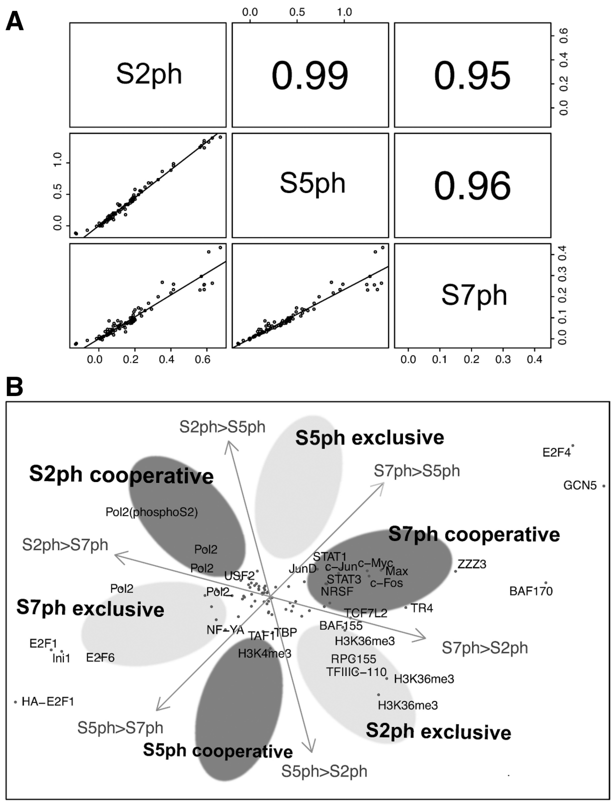

Figure 3. Factors co-localized with the phosphorylation state of Pol II CTD S2/S5/S7. (**A**) Scatter plot of co-localization scores of Pol II CTD S2/S5/S7 phosphorylation pairs. The diagonal panel shows the name of the data. The names correspond to each data pair of the off-diagonal panel. The lower triangular panels show scatter plots of co-localization scores and regression lines of each phosphorylation pair. Their *X*- and *Y*-axes are co-localization scores of the corresponding data labeled by diagonal panel. The upper triangular panel shows Pearson's correlation coefficients of each phosphorylation pair. (**B**) Three factors variation plot. This plot shows the surface spanned by the differences between three variables: three scores of one factor paired with S2ph, S5ph and S7ph. Each arrow indicates the direction of differences between scale-corrected co-localization scores. S2ph > S5ph means the direction of a difference between a co-localization factor paired with S2ph and with S5ph that is >0 (S2ph − S5ph > 0). The other axes were labeled in a similar manner. A factor located between the axes S7ph > S2ph and S7ph > S5ph is regarded as an S7ph-specific cooperative factor. The factors that were far from the center were labeled with the names of antibodies. The center was unaffected by any of the phosphorylation changes.

that their relation to S7ph was relatively exclusive compared with the other phosphorylations. E2F4, GCN5 and BAF170 were in the S7ph cooperative region. A factor 'Pol2' itself, which is 8WG16, and therefore indicates the total amount of Pol II (17), was located in the high S7ph exclusive region. This might have been caused by the low fraction of S7ph in whole Pol II, or by a property of our antibody.

To further evaluate our method, we performed histone H3.1 ChIP-seq and used the data for the analysis of the co-localization model. As a control, we also used the previously obtained histone H3.3 ChIP-seq data (11). We previously demonstrated that MyoD was preferentially associated with H3.3, but not with H3.1 in undifferentiated C2C12 cells (11). Our method revealed that the H3.3 signals co-localized with MyoD rather than H3.1 (Supplementary Figure S4), which was consistent with the previous findings.

Finally, we extracted the affected factors in the ENCODE data set by the differences of the scale-corrected co-localization scores, and we could predict factors related to S7ph.

DISCUSSION

S7ph has been shown to be localized to transcription initiation sites and promoter regions to which most transcription factors preferentially bind (24,25), which is consistent with our observation that S7ph cooperates with the transcription factors such as c-Myc, c-Fos and STAT3, or chromatin modification enzymes such as GCN5, as shown in Figure 3B. On the other hand, S7ph has been suggested to be not particularly co-localized with PolII S2ph, which was confirmed by our results. These observed co-localization patterns suggested that our method could predict not only the involvement of a specific factor but also the functions of multiple factors.

The analysis of histone H3.3 was also utilized to validate our approach. In this case, the reference data sets were obtained from mouse cells and were completely different from the data we utilized for PolII S7ph. The result confirmed the previous finding that H3.3 was preferentially co-localized with MyoD, rather than H3.1. This result suggested that our method is applicable not only for human ENCODE data sets but also for other ChIP-seq data sets.

Comparative analysis for ChIP-seq has been proposed by some research groups (12,13). There is a need to develop a more effective approach to comparing ChIP-seq data upon expansion of the public database of accumulation of ChIP-seq data. Many studies have developed methods or algorithms for peak calling. However, few reported studies have focused on the comparative approach itself for multiple ChIP-seq data sets. Taslim *et al.* (13) proposed bias-free scoring (binding quality) with a non-linear normalization. Their approach is similar to peak calling, as shown in MACS and PeakSeq and the scoring is based on read counts of control and of immunoprecipitated samples. On the other hand, our approach is to compare whole relationships between different samples after peak calling.

In the term of scale correction, in our model, it was sufficient to apply the correction by simple linear regression. Since to the identical transcription factors or histone modification, the co-localization scores indicated a sufficient linear relation in the various ChIP-seq data in ENCODE.

There are information-theoretic approaches to estimate the differences for each shape of distribution, such as cross-entropy or Kullback–Leibler divergence. These approaches are helpful to measure the differences between variable shapes of distribution. However, our approach focused mainly on obtaining the 'strength' of a relationship by using an accumulated number of peaks. We took into account the shapes of distribution with our model, and it was only used for estimating a mixture ratio of the background and the signal. Therefore, we applied a simple curve fitting approach (i.e. a non-linear regression), and it was sufficient to obtain the mixture ratio of distributions.

Our model assumes that there are sufficiently overlapped peaks to estimate parameters for the co-localization model. In the case where data only have limited overlapping, this may produce unreliable results. For example, when fewer peaks of factor A were almost overlapped by peaks of factor B, the resulting scores became higher. In that case, our model failed to capture the shape of distribution because of its coarse shape. To eliminate this misrepresentation of shape, we limited the comparable data set to have >1000 peaks.

We showed that our co-localization model can split a signal and background. The definition of background was an accumulation of randomly distanced peaks within 5 kb around a certain factor's binding site. We eliminated such background. Although there might be contained another aspect of 'distance'-dependent meaningful signal, this case is out of the scope of our model. The distance in base pairs that we used is not the only factor that determines physical distance. For example, the higher order structure of chromosomes would be one of the factors (26). However, the physical distance can be considered by comparing it with physical distance-dependent genome-wide data, such as 4C-seq data (27).

The purpose of our study was to evaluate global relations between paired data. We used our model to empirically estimate S/N after mixing all peaks, but not for filtering each peak. Our model is applicable for filtering each peak by giving each probability of co-localization for each peak after estimating the co-localization model parameters.

In Figure 2C, around the center of H3K36me3, there is a small convex shape of accumulation. H3K36me3 has been reported to be distributed on active gene bodies (28); therefore, we expected to observe a certain correlation between S2ph and H3K36me3 (i.e. they are located on the same active gene). On the other hand, from the point view of co-localization, signal enrichment is suppressed around the center relative to enrichment at >2 kb. This could represent best fitting in the sense of mean-squared error and reflects the concave/exclusive pattern that we assumed. Thus, a possible interpretation of our result is that H3K36me3 and S2ph tend to be located in active genes, according to (28); however, their positions are somewhat distant from each other.

Some shapes of distribution did not fit our model well. In the case where the mode of distribution appeared at ~500 bp, the shape could be represented by a negative-binomial distribution, which is more general than a geometric distribution. The shape of distribution is thought to be caused by maintaining a certain distance between two factors. Another case that did not fit the model is when there was a mixture of three or more distributions, such as when there were simultaneously different types of relations between factors, which were cooperative at a certain distance and exclusive at another certain distance. Other fitting approaches, such as Bayesian formulation, could also be applicable, but balance between the model's accuracy and complexity with model selection (e.g. AIC or BIC) is required.

It could be important to compare the co-localization scores across different factors. We focused on the differences between two scores with the same factor to determine the change of the functional relationship or the function of the factor synergistically. We proposed a method to normalize the differently called peaks. The alternative approach to normalize all data in ENCODE would estimate the scores of all possible pairs in the data set and to correct all scales among the scores. This enables direct comparison across all different factors. This method has the potential to restrict the data set as required, but at least a million combinations of calculations would be required.

Although we used an HeLa data set as an example in our model, because our aim was to identify synergistic factors to regulate various types of Pol II recruitment, comparison between different cells was able to be performed by our method. Our model could be a useful application to compare the same histone modifications or transcription factors between different cells or to determine changes in co-localization of factors involved in differentiation. Since the problem of variation in scale can be resolved by exhaustive scoring with assorted data sets, including related and unrelated factors by the co-localization model, this will enhance potential interest in ChIP-seq data analysis in the ENCODE project or future epigenome projects, including International Human Epigenome Consortium.

ACCESSION NUMBERS

ChIP-seq data were deposited with accession codes DRA000219 and DRA000632 (DDBJ).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2, Supplementary Figures 1–4 and Supplementary Materials and Methods.

ACKNOWLEDGEMENTS

The authors thank Drs M. Nagasaki, N. Nariai, K. Kojima and H. Kimura for their advices and discussion and Ms Ichinose, Ms Itoh and Ms Onishi for technical supports. They thank Research Institute for Information Technology, Kyushu University (sugoka) and National Institute of Genetics (NIG) for providing the high-performance computing resources.

FUNDING

The JSPS KAKENHI [23310134, 233114718, 23651195, in part]; JST; CREST and Kyushu University Interdisciplinary Programs in Education and Projects in Research Development (P&P). Funding for open access charge: Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
2. ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046+.
3. Keeling,P.J., Burger,G., Durnford,D.G., Lang,B.F., Lee,R.W., Pearlman,R.E., Roger,A.J. and Gray,M.W. (2005) The tree of eukaryotes. *Trends Ecol. Evol.*, **20**, 670–676.
4. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.
5. Zhang,Z.D., Rozowsky,J., Snyder,M., Chang,J. and Gerstein,M. (2008) Modeling ChIP sequencing in silico with applications. *PLoS Comput. Biol.*, **4**, e1000158.
6. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
7. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
8. Malone,B.M., Tan,F., Bridges,S.M. and Peng,Z. (2011) Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. *PLoS One*, **6**, e25260+.
9. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.
10. Laajala,T.D., Raghav,S., Tuomela,S., Lahesmaa,R., Aittokallio,T. and Elo,L.L. (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, **10**, 618.
11. Harada,A., Okada,S., Konno,D., Odawara,J., Yoshimi,T., Yoshimura,S., Kumamaru,H., Saiwai,H., Tsubota,T., Kurumizaka,H. *et al.* (2012) Chd2 interacts with H3.3 to determine myogenic cell fate. *EMBO J.*, **31**, 2994–3007.
12. Mendoza-Parra,M.A., Sankar,M., Walia,M. and Gronemeyer,H. (2012) POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization. *Nucleic Acids Res.*, **40**, e30.
13. Taslim,C., Wu,J., Yan,P., Singer,G., Parvin,J., Huang,T., Lin,S. and Huang,K. (2009) Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics*, **25**, 2334–2340.
14. Odawara,J., Harada,A., Yoshimi,T., Maehara,K., Tachibana,T., Okada,S., Akashi,K. and Ohkawa,Y. (2011) The classification of mRNA expression levels by the phosphorylation state of RNAPII CTD based on a combined genome-wide approach. *BMC Genomics*, **12**, 516.
15. Sado,Y., Kagawa,M., Kishiro,Y., Sugihara,K., Naito,I., Seyer,J.M., Sugimoto,M., Oohashi,T. and Ninomiya,Y. (1995) Establishment by the rat lymph node method of epitope-defined monoclonal antibodies recognizing the six different alpha chains of human type IV collagen. *Histochem. Cell Biol.*, **104**, 267–275.
16. Micsinai,M., Parisi,F., Strino,F., Asp,P., Dynlacht,B.D. and Kluger,Y. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res.*, **40**, e70.
17. Brookes,E. and Pombo,A. (2009) Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO Rep.*, **10**, 1213–1219.
18. Cao,Y., Yao,Z., Sarkar,D., Lawrence,M., Sanchez,G.J., Parker,M.H., MacQuarrie,K.L., Davison,J., Morgan,M.T., Ruzzo,W.L. *et al.* (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev. Cell*, **18**, 662–674.
19. van Steensel,B., van Binnendijk,E.P., Hornsby,C.D., van der Voort,H.T., Krozowski,Z.S., de Kloet,E.R. and van Driel,R. (1996) Partial colocalization of glucocorticoid and mineralocorticoid receptors in discrete compartments in nuclei of rat hippocampus neurons. *J. Cell Sci.*, **109**, 787–792.
20. Jin,C., Zang,C., Wei,G., Cui,K., Peng,W., Zhao,K. and Felsenfeld,G. (2009) H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat. Genet.*, **41**, 941–945.
21. Goldberg,A.D., Banaszynski,L.A., Noh,K.-M.M., Lewis,P.W., Elsaesser,S.J., Stadler,S., Dewell,S., Law,M., Guo,X., Li,X. *et al.* (2010) Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell*, **140**, 678–691.
22. Li,J. and Gilmour,D.S. (2011) Promoter proximal pausing and the control of gene expression. *Curr. Opin. Genet. Dev.*, **21**, 231–235.
23. Marshall,N.F., Peng,J., Xie,Z. and Price,D.H. (1996) Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *J. Biol. Chem.*, **271**, 27176–27183.
24. Mayer,A., Lidschreiber,M., Siebert,M., Leike,K., Söding,J. and Cramer,P. (2010) Uniform transitions of the general RNA polymerase II transcription complex. *Nat. Struct. Mol. Biol.*, **17**, 1272–1278.
25. Boeing,S., Rigault,C., Heidemann,M., Eick,D. and Meisterernst,M. (2010) RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion. *J. Biol. Chem.*, **285**, 188–196.
26. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
27. Splinter,E., de Wit,E., van de Werken,H.J.G., Klous,P. and de Laat,W. (2012) Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods*.
28. Edmunds,J.W., Mahadevan,L.C. and Clayton,A.L. (2008) Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *EMBO J.*, **27**, 406–420.