Sensitive and robust assessment of ChIP-seq read distribution using a strand-shift profile

Ryuichiro Nakato[1],* and Katsuhiko Shirahige[1]

[1]Institute of Molecular and Cellular Biosciences, The University of Tokyo, 1-1-1 Yayoi,
Bunkyo-ku, Tokyo 113-0032, JAPAN

*To whom correspondence should be addressed: rnakato@iam.u-tokyo.ac.jp

ABSTRACT

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) can detect read-enriched DNA loci for point-source (e.g., transcription factor binding) and broad-source factors (e.g., several histone modifications). Although numerous quality metrics for ChIP-seq data have been developed, the ‘peaks’ thus obtained are still difficult to assess with respect to signal-to-noise ratio (S/N) especially for broad-source factors, and peak reliability. Here we introduce SSP (strand-shift profile), a tool to assess the quality of ChIP-seq data without peak calling. SSP provides metrics to quantify the S/N for both point- and broad-source factors, and to estimate peak reliability based on the mapped-read distribution throughout a genome. We carried out an in-depth validation of our method using over 1,000 publicly available ChIP-seq datasets, along with virtual data, to demonstrate that SSP is more sensitive than existing tools for both point- and broad-source factors because of the larger dynamic range of the S/N score, and robust for various cell types and sequencing depth. We also found that SSP can identify low-quality samples that cannot be identified by quality metrics currently available. Finally, SSP provides an additional metric to avoid “hidden-duplicate reads” that cause aberrantly high S/Ns in the strand-shift profile. This metric can also contribute to estimation of peak mode (point- or broad-source) of each sample. Our

approach provides a useful way to obtain information about sample quality and traits for ChIP-seq analyses.

Introduction

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) analysis identifies DNA loci of transcriptional factors (TFs) binding (i.e., point-source) as well as broadly distributed histone modifications (i.e., broad-source) [1, 2]. In a ChIP experiment, immunoprecipitated DNA fragments are sequenced to reads, which are mapped to a reference genome, and statistically significant read enrichments (as compared with a corresponding input sample) are detected as peaks. Large consortia such as ENCODE [3], NIH ROADMAP [4] and IHEC [5] enable us to utilize thousands of ChIP-seq data for diverse cell lines and tissues. To handle such large-scale data, objective quality metrics for quantitative assessment are essential to automatically find samples which should be rejected or require a specific consideration to be included in the analysis. Numerous computational measures for ChIP-seq analysis have been developed, which include read quality, library complexity, and GC content [6, 7]. Despite great effort, however, the current approach for assessing peaks is insufficient.

To assess the success of the immunoprecipitation step, signal-to-noise ratio (S/N) is assessed, and the value should be high for ChIP samples and low for input samples. A straightforward way to evaluate the S/N is to count the number of obtained peaks and/or calculate the fraction of reads falling within peak regions (called FRiP), but these ways depend on sequencing depth and peak-calling parameters. In contrast, cross-correlation analysis [6] evaluates the S/N without the need for a peak-calling procedure. It estimates the Pearson correlation coefficient between the read densities mapped on the forward and reverse DNA strands upon shifting from one strand to the other (see Supplemental Fig. S1 for an example). Such a “strand-shift profile” typically peaks at the shift corresponding to the DNA fragment length, which increases as the S/N of the sample increases. This tendency has also been used to estimate fragment length from single-end reads. There is also a spike at the read-length shift that arises from repetitive sequences [8]. Based on this observation, cross-correlation analysis calculates two metrics, namely the normalized strand coefficient (NSC) and the relative strand correlation (RSC), which quantify the fragment length peak relative to background level and relative to the read length peak, respectively (see Results, “Method overview”, for details). These metrics have been used in the ENCODE, ROADMAP and IHEC consortia. A strand-shift profile strategy based on the Hamming distance was also proposed for rapid computation (Hansen et al. 2015). Whereas these tools are useful for point-source factors, broad-source factors (e.g., H3K9me3) often have marginal or truly low scores compared with input samples, even when the samples are of high quality [6]. Moreover, these S/N indicators do not evaluate the reliability of obtained peaks, that is, amount of false positives which are derived from read distribution bias (e.g., GC bias) [9]. Visual inspection at a limited number of sites is effective but not sufficient to explain the properties of read distribution in a whole genome. Consequently, genome-wide assessment of ChIP-seq peak quality still presents challenges that current protocols cannot circumvent.

In this work, we present a new method, SSP, which is based on a strand-shift profile using the Jaccard index to assess S/N, peak reliability and properties of read enrichment in ChIP-seq data. We evaluated the performance of SSP using an extensive dataset of ChIP-seq samples for various cell types obtained from the ENCODE, ROADMAP, and other projects, along with simulated experiments. We demonstrate that SSP provides a more sensitive S/N indicator than current methods both for point- and broad-source marks and is robust for various cell types and sequencing depth. We also found that “hidden-duplicate reads” in a sample confound the strand-shift profile because they cause unexpected enrichment, resulting in calculation of aberrantly high S/Ns. Therefore we additionally developed metrics to overcome this problem, which can also be used to estimate peak mode (point or broad source) of each sample. SSP provides a useful way to assess and obtain additional information about sample quality and traits for ChIP-seq analyses.

Results

Method overview

Fig. 1 presents an overview of SSP (see Methods for details). Using mapped reads as input, SSP generates the strand-specific vectors for forward and reverse strands (step 1). Because sequenced reads that are mapped to the same genomic position are removed as duplicate reads [6], each element of a strand-specific vector is binary, that is, either zero (unmapped) or one (mapped). This binary vector can be handled by computationally fast bit operations in C++ [10]. SSP calculates the Jaccard index between binary vectors of forward and reverse strands for each strand shift *d*, which is then normalized by total read number and chromosome length (step 2). The magnitude of the Jaccard score reflects the co-occurrence of reads mapped on the forward and reverse strands with distance *d*. Whereas the Pearson correlation and Hamming distance confer equal weight to pairs of mapped bases (1,1) and unmapped bases (0,0), the Jaccard index focuses on the mapped bases because unmapped bases can often coincide owing to the lack of sequencing depth and low-mappable regions.

A strand-shift profile is generated within –500 bp < *d* < 1 Mbp (step 3). NSC and RSC are then calculated in the same manner as a cross-correlation analysis. Whereas existing methods use ~1,000–1,500 bp as background, SSP takes the average over a range of 500 kbp to 1 Mbp because we observed that the Jaccard score still decreases up to 1 Mbp (Fig. 1, step 3). Along with NSC and RSC, SSP also calculates “background uniformity” (Bu), which evaluates the uniformity of mapped read distribution in background regions (step 4). Finally, SSP calculates a “fragment cluster score” (FCS), which estimates the cluster level of forward-reverse read pairs with each distance *d* (step 5). FCS is the maximum difference in the parameter cPNF (the cumulative proportion of neighboring fragments) at distance *d* compared at background length. The outputs of SSP are displayed in PDF format and also written to text files.

Comparison with current methods

To assess the performance of SSP for estimating S/N, we implemented three existing tools: 1) phantompeakqualtools (PPQT, https://github.com/kundajelab/phantompeakqualtools), which internally implements spp version 1.14 [11] for cross-correlation analysis and then outputs NSC and RSC; 2) Q version 1.2.0 [10], which adopts a strand-shift profile based on the Hamming distance and calculates RSC; and 3) DeepTools version 2.5.0 [12], which computes the synthetic Jensen-Shannon distance (JSD) that evaluates differences in the cumulative fraction of mapped reads between ChIP by assuming a Poisson distribution as a background model for windows of fixed length. We applied DeepTools with the “–ignoreDuplicates” option according to the instructions given in the manual. We used default parameters for each of the other tools.

Estimating fragment length

We first evaluated the performance of fragment-length estimation with SSP, PPQT and Q using 65 paired-end ChIP-seq datasets for human, mouse, chicken, and fly (Fig. 2A and Supplemental Table S1). We found that SSP could provide comparable and relatively more accurate fragment-length data than PPQT and Q for all four species investigated. PPQT and Q were nearly as accurate as SSP but could not provide a fragment-length estimate for several of the samples (e.g., samples 37 and 45). On the other hand, none of the programs could estimate an accurate fragment-length for certain samples (e.g., sample 16) for which there was no clear peak in the strand-shift profile (Fig. 2B). Because it has been reported that a high score for read-length shift can be mitigated by removing reads mapped on “blacklist regions” in the genome [8], we re-analyzed 45 human samples (no. 1–45) after removing reads mapped on blacklist regions [3] to validate the possibility that they affect the accuracy of fragment-length estimation. However, such filtering had little effect (Supplemental Fig. S2 and Supplemental Table S1). In fact, because the failure of fragment-length estimation is mainly due to a lack of enrichment at the fragment-length shift, mitigating the enrichment at read-length alone is insufficient. In this case, fragment length should be supplied by the users. In subsequent analyses, we did not remove blacklist regions because doing so could affect the RSC, and in fact detailed blacklist regions are available only for human genome build hg19.

Calculating the S/N for point and broad histone marks

Required features for good S/N metrics are the quantifiability and sensitivity of different S/Ns for both point- and broad-source factors, as well as the applicability to various cell types. To comprehensively evaluate the performance of SSP relative to other tools, we first used a compendium of 860 ChIP-seq samples of histone modifications for 127 cell types, which were obtained from the ROADMAP project [4]. These data contain information for six core histone modifications, consisting of both point-source (H3K27ac, H3K4me1, H3K4me3) and broad-source factors (H3K27me3, H3K36me3, H3K9me3) along with input samples. In the consolidated dataset, reads of each sample were truncated to 36 bp, mapped onto genome build hg19, filtered using a 36-bp mappability track, and then uniformly down-sampled to a maximum depth of 30 million reads, which is appropriate for avoiding the effect derived from different sequencing depths, parameters for mapping, and mappability.

A comparison is shown in Fig. 2C (see Supplemental Table S2 for detailed information and scores for each sample). The results revealed that SSP-NSC and JSD could achieve sufficient sensitivity both for point- and broad-source marks. The smaller difference between point- and broad-source marks for JSD compared with SSP-NSC is perhaps a consequence of score saturation, i.e., given that the maximum value of JSD is 1.0. PPQT-NSC showed little difference among three broad marks compared with input samples (~1.1 fold), indicative of insensitivity for broad marks.

As previously reported, RSCs obtained with all three tools were comparable or lower for H3K9me3 than input samples. The discrepancy between NSC and RSC is possibly because H3K9me3 is more highly enriched at the read-length shift compared with other histone modifications derived from repetitive regions, such as centromeres [13]. Because RSC amalgamates the magnitude of true peak enrichment and repeat effects, when the read-shift enrichment is high, the RSC may be small even when the S/N is sufficiently high. Furthermore, the relatively wider distribution of RSC for input samples indicates that a low S/N increases the variability of it owing to the small value of the denominator (difference between read-length value and background).

To further validate the ability of S/N indicators, we generated virtual data for histone modifications with various S/Ns by adding a fixed number of input reads to each ChIP sample in a stepwise manner. The S/N then decreased with increasing numbers of input reads. Fig. 2D shows the comparison for E072 (Brain inferior temporal lobe) and Supplemental Fig. S3 shows results for two other cell types. In most cases, the values of the indicators decreased with increasing numbers of input reads. RSC was relatively higher for H3K9me3 because, for this mark, the scores were often lower than those of the input (Fig. 2C). SSP-NSC had the superior or comparable sensitivity to changes in S/N, while PPQT-NSC lacked sensitivity for evaluating broad marks.

Evaluating the validity of the S/N for TFs for 20 cell types

The S/N estimation could be affected by multiple factors, such as sequencing depth, read length and copy number variations in cancer cell lines [14]. To validate the robustness of the

S/N indicators against these factors, we next investigated 399 ChIP-seq samples of TFs (point source) for 20 cell types obtained from the ENCODE project [15]. This dataset contains various read lengths (25, 36, and 50) and sequencing depths. Fig. 3A and Supplemental Fig. S4 depict the distribution of SSP-NSC and the other scores, respectively, for ChIP and input samples of 20 cell types (see Supplemental Table S3 for detailed information). Whereas the number of samples varied among those cell types, we found that SSP-NSC could reveal distinct differences between ChIP and input samples for all cell types. To compare the various tools in this respect, we displayed the median scores for each cell type for all indicators (Fig. 3B). For SSP-NSC and PPQT-NSC, median values for ChIP and input samples were consistently different among all cell types, indicating that a cell type–independent threshold value could be defined for these indicators. For example, SSP-NSC ≥ 3.0 may be a good candidate threshold for TF ChIP samples, whereas the averaged S/N varied among the TFs and antibodies used (Supplemental Fig. S4). Meanwhile, RSC and JSD could not sufficiently distinguish ChIP and input samples. Although ChIP samples had larger values than input samples for each cell type (Supplemental Fig. S5), the separation between the data for ChIP and input samples depended on cell type, and therefore it was difficult to determine a uniform threshold value. Consequently, SSP-NSC is a sensitive and robust estimator that can be standardized across diverse cell types.

Correlation with FRiP score

To further evaluate the performance of S/N indicators, we calculated the Spearman's correlation coefficient between the FRiP score and each S/N indicator across the ENCODE and ROADMAP datasets (Table 1). Because FRiP score depends on sequencing depth, we computed each FRiP score with and without total read normalization (see Methods for details). First, RSC yielded a low correlation, suggesting that RSC cannot be used for quantitative estimation of the S/N. In contrast, the output of each of SSP-NSC, PPQT-NSC, and JSD was highly correlated with FRiP scores. Although SSP-NSC and PPQT-NSC each correlated well with normalized FRiP scores, JSD correlated better with FRiP score without normalization, which clearly shows the dependency of JSD on sequencing depth. This conclusion is valid for the ROADMAP dataset with both point-source marks (H3K4me1, H3K4me3, H3K27ac) and broad-source marks (H3K27me3, H3K36me3, H3K9me3) (Table 1). The lesser correlation of PPQT-NSC with broad-source marks compared with point-source marks implies its lower sensitivity for broad-source marks.

To further investigate this tendency, we implemented a down-sampling analysis. We selected six samples (four ChIP and two input samples) that contained an abundant number of reads (>50 million) after removing duplicate reads. For each sample, we subsampled the reads to a fixed number (from 5 million to 50 million) and calculated the ratio of the score at each depth relative to the score for the 50 million reads (Fig. 3C and Supplemental Fig. S6A). While all indicators except for JSD did not fluctuate with sequencing depth, JSD decreased at lower sequencing depth. For input samples, each ratio fluctuated slightly (~1.1 fold) because of smaller values for the 50 million reads. The analysis of histone modification data also reached the same conclusion (Supplemental Fig. S6B). Consequently, SSP-NSC is the best predictor of S/N for both point-source and broad-source marks, independent of sequencing depth and cell types.

Background uniformity

NSC is defined as relative enrichment of the Jaccard score at each fragment-length shift compared with the background level (Fig. 1, step 3). The next question was thus “why does background level vary among samples?” By definition, the Jaccard score at background reflects the co-occurrence probability of forward and reverse reads. Ideally, the background reads should be uniformly distributed; in reality, however, the read distribution is often more congregated, or biased, owing to various potential technical or biological issues [16], resulting in a higher Jaccard score at background. Although the library complexity evaluates the percentage of duplicate reads, it does not directly reflect any potential bias in the read distribution. In fact, we observed that the background score increased ~2-fold when the mapped reads were removed in every other 10-Mbp window, whereas library complexity and NSC score remained essentially unchanged (Fig. 4A).

Based on this observation, we defined Bu, which evaluates the magnitude of the observed background score compared with the uniform distribution (see Methods). A high value of Bu indicates that the background reads are uniformly distributed even if library complexity is low. In contrast, a low Bu score indicates sparse (or biased) read distribution, which decreases the reliability of the peaks obtained.

We computed Bu scores for 860 histone modification samples from ROADMAP (Fig. 4B and Supplemental Table S2). Although most of these consolidated data had library complexity = 1.0, we noted that a small amount of data for each histone modification and input sample had a low Bu score (<0.8). A low Bu score was still observed even after filtering out samples of low sequencing depth (<20 million reads). To further investigate the various aspects of Bu, we chose 12 H3K36me3 samples as representatives, and the results are shown in Fig. 4C–E. We grouped these samples into four types: (1) low NSC and high Bu, (2) high NSC and high Bu, (3) high NSC and low Bu; this type was further classified as 3-1 (GC-rich) and 3-2 (not GC-rich). Fig. 4C illustrates the relative scores as a heatmap (see Supplemental Table S4 for details concerning scores). Fig. 4D presents data for the read distribution proximal to the housekeeping gene *IREB2* [17]. Groups 2 and 3 had high S/Ns, reflecting read enrichment at the *IREB2* locus. Samples in group 3-1, however, had an unexpectedly sparse read distribution, which is not reasonable considering that H3K36me3 is broadly distributed within genic regions. Considering the GC-rich read distribution, this read distribution may be a consequence of GC bias [18]. In contrast, group 3-2 had low Bu values without GC bias, and read distribution was reasonable compared with group 3-1. However, this group also had lower genome coverage in background region (Fig. 4E and Supplemental Fig. S7). A possible reason for this is that the DNA fragmentation of tightly packed regions, e.g., heterochromatin, did not work well, resulting in a much lower number of reads on the regions. These samples might confound the read normalization for comparative analyses that assumes comparable read depth among samples over the entire genome [19]. These results suggest that Bu is an effective criterion with which to judge whether a specific consideration is required for comparative analysis.

Interestingly, GC-biased samples (group 3-1) had a striking peak for fragment length in the strand-shift profile (Fig. 4F). This phenomenon might also facilitate the identification of read bias.

Relevance of Bu to other metrics

To ascertain whether Bu varies among other mapping statistics and cell types, we next investigated 399 ENCODE TF samples (Supplemental Table S3). We first found relatively lower Bu values for MCF-7 cells (~0.8, Supplemental Fig. S8A), possibly owing to extensive copy number variations [20]. The low-Bu samples also were more common when the S/N was extremely high (e.g., RNA pol2, Supplemental Fig. S8B). Thus, it is desirable to use a relaxed threshold value for Bu for these samples.

We next found that Bu did not correlate strongly with library complexity (Fig. 4G) or with the mapping ratio of uniquely and multiply mapped reads (Supplemental Fig. S9). This result suggests that the low values of these mapping statistics do not necessarily indicate biased read distribution. For example, sample GM12892_PAX5-C20_v041610.1 had relatively low library complexity (0.726) but a high Bu value (1.060) and no GC bias (peak = 45). The strand-shift profile of this sample clearly revealed a maximum at fragment length (Supplemental Fig. S10), indicative of sufficient quality.

On the other hand, Bu showed a moderately negative correlation with GC content (Fig. 4H), consistent with the H3K36me3 results (Fig. 4C-E), whereas several samples had a high Bu despite a highly GC-rich distribution (GC peak > 55); for instance, Rad21 sample for K562 (K562_Rad21_v041610.2) has GC peak = 56 but has an acceptable Bu (0.997). Although Rad21 binding is closely correlated with CTCF [21], CTCF sample for K562 (K562_CTCF_SC-5916_PCR1x) is not GC-rich (GC peak = 48) and had a similar Bu (0.980). In fact, this Rad21 sample had an unexpected bimodal GC distribution (Fig. 4I). Considering the remarkable peak overlap between these two samples (98.6%, Supplemental Fig. S11), the peaks of this Rad21 sample could be considered usable. This result implies that GC content alone is not always appropriate to reject a putative low-quality sample. In this respect, the Bu metric along with GC content provides a more reliable indicator of sample quality with respect to biased read distribution.

FCS can identify peak intensity and peak mode

While having verified the effectiveness of SSP-NSC for calculating the S/N, we also found that strand-shift profiles of a small number of input samples had peaks at fragment length despite having a low FRiP score (e.g., input of E024 and E058 cells, Fig. 5A). These two samples in particular had extremely high SSP-RSC (6.656 and 5.347), a phenomenon that is commonly observed in PPQT and Q (Supplemental Fig. S12). We presumed that this is due to “hidden duplicate reads”. That is, at most two reads (forward and reverse pair) that are derived from the same amplified DNA fragment can remain after PCR-bias filtering because forward and reverse strands are scanned separately for single-end reads (Fig. 5B). Such reads may often appear in low-library complexity samples and introduce a spike at the fragment length, resulting in aberrant NSC and RSC values. To examine this hypothesis, we generated strand-shift profiles for a paired-end sample in which both forward and reverse reads were mapped as ‘single-end’. As expected, the resulting profile showed a remarkable peak at the fragment length shift (Fig. 5C). While NSC increased less drastically (1.53 to 2.54), RSC increased more than three times (0.61 to 2.29). This result suggested the presence of the artifactual S/N enrichment without real peaks in a strand-shift profile, which could especially influence the calculation of RSC.

To overcome this problem, we defined FCS, which directly evaluates the cluster level of forward-reverse read pairs at each strand shift *d* (see Methods for details). The FCS value is high when read pairs with distance *d* are highly clustered as peaks (Fig. 1, step 5). Therefore, samples that contain hidden duplicate reads which are not clustered in a genome should have a low FCS score. As expected, FCS could identify read clustering in samples and was little affected by hidden duplicate reads (Fig. 5D). FCS correlated better with peak intensity (height) than did FRiP, which represents a composite of peak number and intensity (Supplemental Fig. S13).

Fig. 5E illustrates the example of five input samples from ROADMAP (see Supplemental Table S5 for details concerning scores). The E097 input sample had strong peaks and the highest FCS score among these samples (0.240). E024 and E058 (shown in Fig. 5A) had high NSC and RSC values without many peaks, resulting in a low FCS score (0.041 and 0.038, respectively). In contrast, E100 had more peaks (33,476) than E097, but the FCS score was low (0.044), indicating that the mapped reads were not highly clustered. The read distribution and relatively lower FRiP score for E100 suggested that this sample had only small peaks. Therefore, at a sufficiently high peak-calling threshold, most of the small peaks (i.e., as in E100) would be expected to disappear, in contrast to the expectation for E097. JSD was only minimally affected by hidden duplicate reads because it is not based on a strand-shift profile, while it provided E100 with the highest score, suggesting that it correlated better with peak number than did peak intensity and FRiP.

Interestingly, the FCS profile reflects the peak mode (point or broad source) for histone modifications (Fig. 5F). H3K4me3 had the highest FCS at *d* = fragment length and decreased steeply at d > 10 kbp. The broad-source marks H3K27me3, H3K36me3, and H3K9me3 each had a moderate score at fragment length, and the value was retained even at d > 10 kbp, resulting in a higher score than for H3K4me3 at d = 10 kbp. H3K27ac had a high score at fragment length and also the highest score at 10 kbp. This is not surprising because H3K27ac had high peaks for point-source marks, some of which clustered in broad genomic regions called super-enhancers [22]. This result suggested that FCS has the potential to identify peak mode without the need for peak calling.

Discussion

The quality of ChIP-seq data depend on various experimental factors such as antibody quality, crosslinking, DNA fragmentation, and PCR amplification. Although normalization using a corresponding input sample mitigates biases in a ChIP sample, input data alone cannot explain all the variability in read bias in the background [23]. It is important to assess the genome-wide properties of samples in an objective manner to validate whether each sample in the dataset requires special normalization or should be rejected for comparative ChIP-seq analysis.

In this work, we present SSP, a peak calling–free quality assessment tool for read enrichment in ChIP-seq data. We compared SSP against the existing methods PPQT, Q, and DeepTools with more than 1,000 ChIP samples in public databases and demonstrated that SSP has advantages over these methods with respect to sensitivity for both point-source and broad-source factors, correlation with normalized FRiP, and robustness for various sequencing depth and cell types. Although JSD, as utilized in DeepTools, is also sensitive and can estimate the S/N for broad marks, it has less classification power between ChIP and input samples owing to a lack of dynamic range. Moreover, because JSD depends on sequencing depth, it requires subsampling for comparison across samples, which is burdensome for a large-scale analysis.

Bu evaluates the reliability of the obtained peaks by quantifying the distribution of mapped reads in background regions. Although GC content correlates with the bias level in ChIP samples, it alone cannot be used for filtering because samples that have many GC-rich peaks (e.g., CpG islands) also have a high GC content. Bu is beneficial in this regard, especially for consolidated data, for which the mapping ratio and library complexity metrics are not generally available. While the “X-intercept” metric in DeepTools evaluates genome coverage, it also depends on sequencing depth and less robust than Bu. Finally, SSP provides FCS, which avoids the effect of hidden duplicate reads. The potential of FCS to evaluate peak mode may facilitate capturing dynamic changes of genome-wide binding patterns among samples, such as during the cell development [24].

Owing to the difficulty of assessing broad marks and peak reliability, a previous study involving large-scale sample evaluation for S/N was limited to input and negative-control samples [25]. The use of SSP enables in-depth validation using >1,000 ChIP-seq data that are publicly available, including point-source and broad-source marks, along with virtual data, and SSP provides multiple key insights for ChIP-seq analysis.

Based on our results, we recommend using NSC rather than RSC when calculating the S/N in the strand-shift profile for several reasons. First, RSC is based on the value at read length, which depends on blacklist region filtering. Second, RSC has high variance in the evaluation of low-S/N samples due to the small values at both read-length and fragment-length. Third, RSC combines the magnitude of peak enrichment and repeat effects. A strong repeat effect cancels out strong peak intensity. Finally, we observed that, compared with NSC, RSC is strongly affected by hidden duplicate reads.

One challenge that remains is to identify false-positive peaks caused by non-specific binding, such as “hyper-ChIPable regions” [26]. SSP and all existing tools cannot distinguish whether or not DNA-binding is derived from true binding, and thus a comparison with mock ChIP-seq data (e.g., IgG) is needed to avoid such false positives. Finally, the challenge remains to accurately estimate fragment length from single-end data.

Methods

Strand-shift profile using the Jaccard index

Let $\mathrm{v}_\mathrm{c}^\mathrm{fwd} = \left(x_1^{fwd} x_2^{fwd} \dots x_n^{fwd}\right)_c$and $\mathrm{v}_\mathrm{c}^\mathrm{rev} = (x_1^{rev} x_2^{rev} \dots x_n^{rev})_c$ be strand-specific binary vectors for forward and reverse strands for chromosome *c* of length *n*, respectively. $\mathrm{x}_\mathrm{k}^\mathrm{str}$ $(\mathrm{k} \in [1, \mathrm{n}], \mathrm{str} \in [\mathrm{fwd}, \mathrm{rev}])$ is the number of reads whose 5' ends map to position *k* of strand *str*, and $\mathrm{x}_\mathrm{k}^\mathrm{str} \in \{0,1\}$ after removing duplicate reads. The Jaccard index between $\mathrm{v}_\mathrm{c}^\mathrm{fwd}$ and $\mathrm{v}_\mathrm{c}^\mathrm{rev}$ at strand shift *d* is defined as follows:

$$\mathrm{J}[\mathrm{v}^\mathrm{fwd}, \mathrm{v}^\mathrm{rev}, \mathrm{d}]_\mathrm{c} = \frac{|\,\mathrm{v}_\mathrm{c}^\mathrm{fwd} \cap \mathrm{v}_\mathrm{c}^\mathrm{rev}(\mathrm{d})|}{|\,\mathrm{v}_\mathrm{c}^\mathrm{fwd} \cup \mathrm{v}_\mathrm{c}^\mathrm{rev}(\mathrm{d}))|},$$

where $\mathrm{v}_\mathrm{c}^\mathrm{rev}(\mathrm{d})$ is $(x_{d+1}^{rev} x_{d+2}^{rev} \dots x_n^{rev})_c$. Therefore, $0 \leq \mathrm{J}[\mathrm{v}^\mathrm{fwd}, \mathrm{v}^\mathrm{rev}, \mathrm{d}]_\mathrm{c} \leq 1$. This formula can be transformed as follows:

$$\mathrm{J}[\mathrm{v}^\mathrm{fwd}, \mathrm{v}^\mathrm{rev}, \mathrm{d}]_\mathrm{c} = \mathrm{N}_\mathrm{d}^\mathrm{both} / (\mathrm{N}^\mathrm{fwd} + \mathrm{N}^\mathrm{rev} - \mathrm{N}_\mathrm{d}^\mathrm{both}),$$

where $\mathrm{N}^\mathrm{str} = \sum_\mathrm{k}^\mathrm{n} \mathrm{x}_\mathrm{k}^\mathrm{strand}$, and $\mathrm{N}_\mathrm{d}^\mathrm{both} = \sum_\mathrm{k}^{\mathrm{n-d}} \mathrm{f}\left(\mathrm{x}_\mathrm{k}^\mathrm{fwd}, \mathrm{x}_\mathrm{k+d}^\mathrm{rev}\right)$, where $\mathrm{f}(\mathrm{a}, \mathrm{b}) = \begin{cases} 1\ (a = b = 1) \\ 0\ (\text{otherwise}) \end{cases}$. This score is calculated using the bitset operator in C++. The strand shift *d* ranges from –500 bp to 1,500 bp at single–base pair resolution. To standardize the value for various species having different genome lengths, this Jaccard score is then normalized per fixed number of reads ($\mathrm{N}_\mathrm{const}$, 10M default) for a fixed length of bases ($\mathrm{L}_\mathrm{const}$, 100M default):

$$\mathrm{Jnorm}[\mathrm{v}^\mathrm{fwd}, \mathrm{v}^\mathrm{rev}, \mathrm{d}]_\mathrm{c} = \mathrm{J}[\mathrm{v}^\mathrm{fwd}, \mathrm{v}^\mathrm{rev}, \mathrm{d}]_\mathrm{c} * \frac{\mathrm{N}_\mathrm{const}}{\mathrm{N}_\mathrm{c}} * \frac{\mathrm{L}_\mathrm{const}}{\mathrm{L}_\mathrm{c}},$$

where $\mathrm{N_c}$ and $\mathrm{L_c}$ are the number of mapped reads and the number of mappable positions (at which the reads starting at those positions are uniquely mapped on the genome), respectively, on chromosome *c*. We estimated $\mathrm{L_c}$ for 36-mer and 50-mer reads based on the code from Peakseq [27].

Finally, SSP assembles the Jaccard index profiles obtained from all autosomes:

$$\mathrm{Jnorm}[\mathrm{v^{fwd}}, \mathrm{v^{rev}}, \mathrm{d}]_{\mathrm{genome}} = \sum\nolimits_{\mathrm{c} \in \mathrm{C}} \frac{\mathrm{N_c}}{\mathrm{N_{genome}}} \mathrm{Jnorm}[\mathrm{v^{fwd}}, \mathrm{v^{rev}}, \mathrm{d}]_{\mathrm{c}},$$

where *C* is the set of all autosomes, and $\mathrm{N_{genome}} = \sum_{\mathrm{c} \in \mathrm{C}} \mathrm{N^c}$. SSP excludes sex chromosomes to ignore gender-specific differences. We use this $\mathrm{Jnorm}[\mathrm{v^{fwd}}, \mathrm{v^{rev}}, \mathrm{d}]_{\mathrm{genome}}$ as the Jaccard score $\mathrm{J(d)}$ for each sample in SSP. Then the fragment length $\mathrm{d_{flen}}$ can be estimated as $\mathrm{d_{flen}} = \mathrm{argmax}_{d_{readlen}*1.2<d<1500} J(d)$. To ignore a peak at the read-length shift ($\mathrm{d_{readlen}}$), SSP uses $\mathrm{d} > \mathrm{d_{readlen}} * 1.2$. Then NSC and RSC can be calculated as:

$$\mathrm{NSC} = \mathrm{J(d_{flen})}/\mathrm{J}\left(\mathrm{d_{bg}}\right) \text{ and } \mathrm{RSC} = \left(\mathrm{J(d_{flen})} - \mathrm{J}\left(\mathrm{d_{bg}}\right)\right) / \left(\mathrm{J(d_{readlen})} - \mathrm{J}\left(\mathrm{d_{bg}}\right)\right),$$

where $\mathrm{J}\left(\mathrm{d_{bg}}\right)$ is the Jaccard score for the background, which is the average from 500 kbp to 1 Mbp at steps of 5 kbp (default).

Background uniformity

Background uniformity *Bu* is defined as follows:

$$\mathrm{Bu} = \mathrm{J}\left(\mathrm{d_{bg}}\right)^{\mathrm{uniform}} / \mathrm{J}\left(\mathrm{d_{bg}}\right)^{\mathrm{observe}},$$

where $\mathrm{J}\left(\mathrm{d_{bg}}\right)^{\mathrm{uniform}}$ is the normalized Jaccard score for the background for a sample that has a completely uniform read distribution. That is, by denoting $\mathrm{E}[\mathrm{x_k} = 1]^{strand}$ as the probability of a mapped read occurring at genomic position *k*,

$$\mathrm{J}\left(\mathrm{d_{bg}}\right)^{\mathrm{uniform}} = \frac{\left(\mathrm{E}[\mathrm{x_k} = 1]^{fwd} * \mathrm{E}[\mathrm{x_k} = 1]^{rev} * \mathrm{L_c}\right)}{\left(\frac{\mathrm{N_{const}}}{2} + \frac{\mathrm{N_{const}}}{2} - \mathrm{E}[\mathrm{x_k} = 1]^{fwd} * \mathrm{E}[\mathrm{x_k} = 1]^{rev} * \mathrm{L_{const}}\right)}$$

$$= \mathrm{N_{const}}/(4\mathrm{L_{const}} - \mathrm{N_{const}}),$$

because $\mathrm{E}[\mathrm{x_k} = 1]^{fwd} = \mathrm{E}[\mathrm{x_k} = 1]^{rev} = \frac{\mathrm{N_{const}}}{2\mathrm{L_{const}}}$. Therefore, $\mathrm{J}\left(\mathrm{d_{bg}}\right)^{\mathrm{uniform}} = 1/39$ for $\mathrm{N_{const}}$ = 10M and $\mathrm{L_{const}}$ = 100M. A high *Bu* score indicates that the sample has a relatively uniform read distribution in the background region. *Bu* should range from 0 to 1, but practically, the maximum score of Bu slightly exceeds 1.0 because the estimated mappable chromosomal length $\mathrm{L_c}$ is a bit larger than the actual mappable chromosomal length.

FCS

Similar to the Jaccard score, FCS is calculated for each strand shift *d*. Let $\mathrm{rp}(d)$ represent a forward- and reverse-read pair with distance *d.* Denote all read-pair sets as $\left\{\mathrm{rp}_1(d), \mathrm{rp}_2(d), \dots, \mathrm{rp}_{\mathrm{N_d^{both}}}(d)\right\}$, which is sorted by genomic position. That is, this set consists of all $(\mathrm{x_k^{fwd}}, \mathrm{x_{k+d}^{rev}})$ pairs such that $\mathrm{f}\left(\mathrm{x_k^{fwd}}, \mathrm{x_{k+d}^{rev}}\right) = 1$.

Let $\mathrm{NF}[\mathrm{d}, s]$ represent the number of $\mathrm{rp_k}(d)$ ($\mathrm{k} \in \left[1, \mathrm{N_d^{both}} - 1\right]$) that have neighboring read pairs $\mathrm{rp_{k+1}}(d)$ within distance *s*. Then the cPNF is:

$$\mathrm{cPNF}[\mathrm{d}, \mathrm{s}] = \mathrm{NF}[\mathrm{d}, \mathrm{s}]/\mathrm{NF}[\mathrm{d}, \mathrm{L_c}]$$

$$= \mathrm{NF}[\mathrm{d}, \mathrm{s}]/\mathrm{N_d^{both}}.$$

This cPNF is calculated up to $\mathrm{s_{max}}$ (5 kbp, default). Supplemental Fig. 14A shows the typical pattern of cPNF for ChIP and input samples. If a sample has peaks, the cPNF score becomes higher at short distance *d* (Supplemental Fig. 14A left). If a sample does not have peaks, the cPNF score at short distance shows little difference from that at long distance (background) (Supplemental Fig. 14A right).

Then, we define FCS[d] as the maximum difference of cPNF[d] from background against *s*:

$$\mathrm{FCS[d]} = \arg\max_{0 \le s < s_{\max}} \left(\mathrm{cPNF[d,s]} - \mathrm{cPNF}\left[\mathrm{d_{bg}}, \mathrm{s}\right]\right).$$

Because FCS[d] depends on sequencing depth, SSP down-samples the reads to a fixed number (10M reads, default). The maximum difference strategy used here can provide more robust values for different values of parameter $\mathrm{s_{max}}$ and $\mathrm{d_{bg}}$ than a relative entropy method such as the Kullback–Leibler divergence. The resulting FCS profile (Supplemental Fig. 14B) reflects the cluster level of $\mathrm{rp}(d)$ in the sample, whereas the Jaccard score $\mathrm{J(d)}$ reflects $\mathrm{N_d^{both}}$, the number of $\mathrm{rp}(d)$. Therefore, for point-source factors, the average peak width is ~1 kbp, and $\mathrm{FCS[d](d \le 1\ kbp)}$ is high, whereas broad marks have relatively higher scores for the broad width (e.g., H3K4me3 and H3K36me3 in Fig. 5F).

Read mapping

We used bowtie 1 version 1.1.2 [28] for mapping single-end reads and extracted uniquely mapped reads. For mapping paired-end reads, we used bowtie 2 version 2.2.9 [29], which is more sensitive for longer reads than bowtie 1.

FRiP score

We used MACS2 version 2.1.1 [30] for peak calling with --nomodel option, and we also supplied --broad option for broad marks (H3K27me3, H3K36me3, H3K9me3). Because MACS2 does not have an option for read-depth normalization, we also used DROMPA3 version 3.2.6 [31] with “–n GR” option for peak calling normalized by the number of nonredundant reads. FRiP scores with and without total read normalization were calculated by peaks of MACS2 and DROMPA3, respectively. The FRiP score, peak height, library complexity for 10M reads, and GC content were also calculated with DROMPA3.

Data access

For histone modification data, we acquired the consolidated data for 117 cell types (tagAlign format, build hg19) from the ROADMAP project [32], available at http://egg2.wustl.edu/roadmap/web_portal/. For the analysis of Q and DeepTools, we converted tagAlign format to BAM format using bedtools (http://bedtools.readthedocs.io). For TF data for the 20 cell types, we acquired fastq files from the Sequence Read Archive (SRA) under accession number SRP008797, which is part of the ENCODE project [15]. Supplemental Methods describes the method for generating virtual data (Fig. 2D, Fig 3C and Fig 4A).

Software availability

SSP is open-source software that is freely available for nonprofit use. It is implemented as a C++ package with a boost library (http://www.boost.org/), and internally uses R to visualize the strand-shift profile and FCS profiles in PDF format. The user manual and examples are available at https://github.com/rnakato/SSP. The mappability tables generated for several species are also provided on the SSP website.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We are grateful to Dr. M. Suyama for valuable comments. We also thank our laboratory's members and collaborators.

FUNDING

This work was supported by a Grant-in-Aid for Scientific Research [15K18465, 17H06331 to R.N., 15H02369, 15H05970 to K.S.], The Japan Agency for Medical Research and Development, and Platform for Drug Discovery, Informatics, and Structural Life Science.

REFERENCES

1. **Park PJ: ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009, 10:669-680.**

2. **Furey TS: ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* 2012, 13:840-852.**

3. **Encode Project Consortium: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489:57-74.**

4. **Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al: Integrative analysis of 111 reference human epigenomes. *Nature* 2015, 518:317-330.**

5. **Stunnenberg HG, International Human Epigenome C, Hirst M: The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* 2016, 167:1145-1149.**

6. **Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012, 22:1813-1831.**

7. **Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J: Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* 2013, 9:e1003326.**

8. **Carroll TS, Liang Z, Salama R, Stark R, de Santiago I: Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front Genet* 2014, 5:75.**

9. **Nakato R, Shirahige K: Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform* 2016.**

10. Hansen P, Hecht J, Ibrahim DM, Krannich A, Truss M, Robinson PN: Saturation analysis of ChIP-seq data for reproducible identification of binding peaks. *Genome Res* 2015, 25:1391-1400.

11. Kharchenko PV, Tolstorukov MY, Park PJ: Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 2008, 26:1351-1359.

12. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T: deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016, 44:W160-165.

13. Guenatri M, Bailly D, Maison C, Almouzni G: Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J Cell Biol* 2004, 166:493-505.

14. Zarrei M, MacDonald JR, Merico D, Scherer SW: A copy number variation map of the human genome. *Nat Rev Genet* 2015, 16:172-183.

15. Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM: Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* 2013, 52:25-36.

16. Meyer CA, Liu XS: Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* 2014, 15:709-721.

17. Eisenberg E, Levanon EY: Human housekeeping genes, revisited. *Trends Genet* 2013, 29:569-574.

18. Benjamini Y, Speed TP: Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 2012, 40:e72.

19. Lun AT, Smyth GK: De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res* 2014, 42:e95.

20. Hampton OA, Den Hollander P, Miller CA, Delgado DA, Li J, Coarfa C, Harris RA, Richards S, Scherer SE, Muzny DM, et al: A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res* 2009, 19:167-177.

21. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al: Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* 2008, 451:796-801.

22. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA: Super-enhancers in the control of cell identity and disease. *Cell* 2013, 155:934-947.

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

24. Dahl JA, Jung I, Aanes H, Greggains GD, Manaf A, Lerdrup M, Li G, Kuan S, Li B, Lee AY, et al: Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* 2016, 537:548-552.

25. Marinov GK, Kundaje A, Park PJ, Wold BJ: Large-Scale Quality Analysis of Published ChIP-seq Data. *G3 (Bethesda)* 2014, 4:209-223.

26. Teytelman L, Thurtle DM, Rine J, van Oudenaarden A: Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci U S A* 2013, 110:18602-18607.

27. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 2009, 27:66-75.

28. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, 10:R25.

29. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012, 9:357-359.

30. **Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS).** ***Genome Biol*** **2008, 9:R137.**

31. **Nakato R, Itoh T, Shirahige K: DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data.** ***Genes Cells*** **2013, 18:589-601.**

32. **Zhou X, Li D, Zhang B, Lowdon RF, Rockweiler NB, Sears RL, Madden PA, Smirnov I, Costello JF, Wang T: Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser.** ***Nat Biotechnol*** **2015, 33:345-346.**

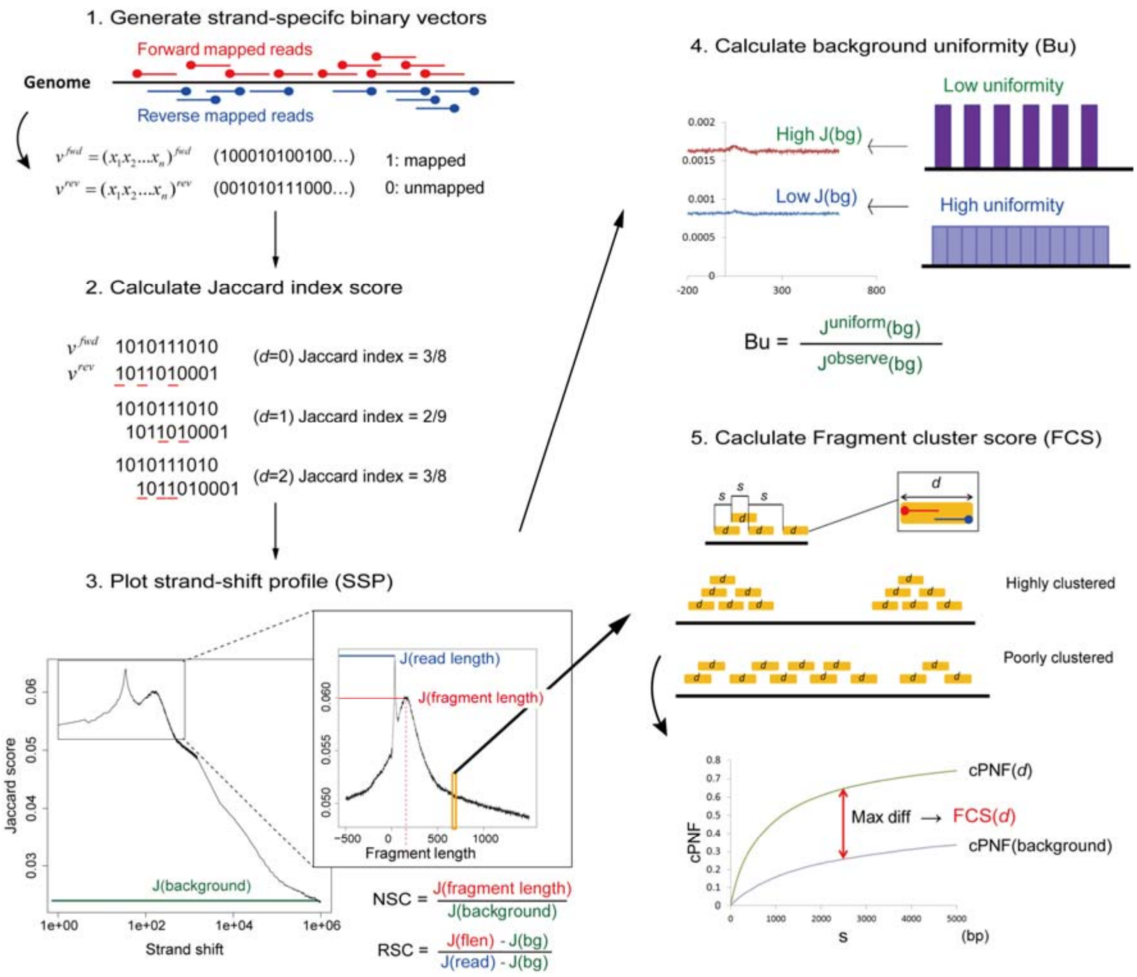


Figure 1. Workflow of SSP. Step 1: convert mapped reads to strand-specific binary vectors (*n*: chromosome length), in which ‘1’ indicates that the 5’ end of a read is mapped at the genomic position. Duplicate reads are discarded. Step 2: calculate the similarity between forward and reverse strand-specific binary vectors for each strand shift *d* based on the Jaccard index. An example calculation is shown (*n* = 10, *d* = 0, 1, 2). Step 3: plot a strand-shift profile based on the Jaccard index and calculate NSC and RSC. Fragment length is estimated as the distance *d* at which the Jaccard score is maximal except for read-length shift. Step 4: calculate background uniformity based on the background level. Step 5: calculate the fragment cluster score to evaluate the cluster level of all forward-reverse read pairs with each distance *d* (orange rectangles). These read pairs are the same as the red bars in step 2. The variable s indicates the distance to the nearest downstream read pair.

cPNF is the cumulative proportion of neighboring downstream fragments (see Methods). FCS increases as read pairs become more clustered.

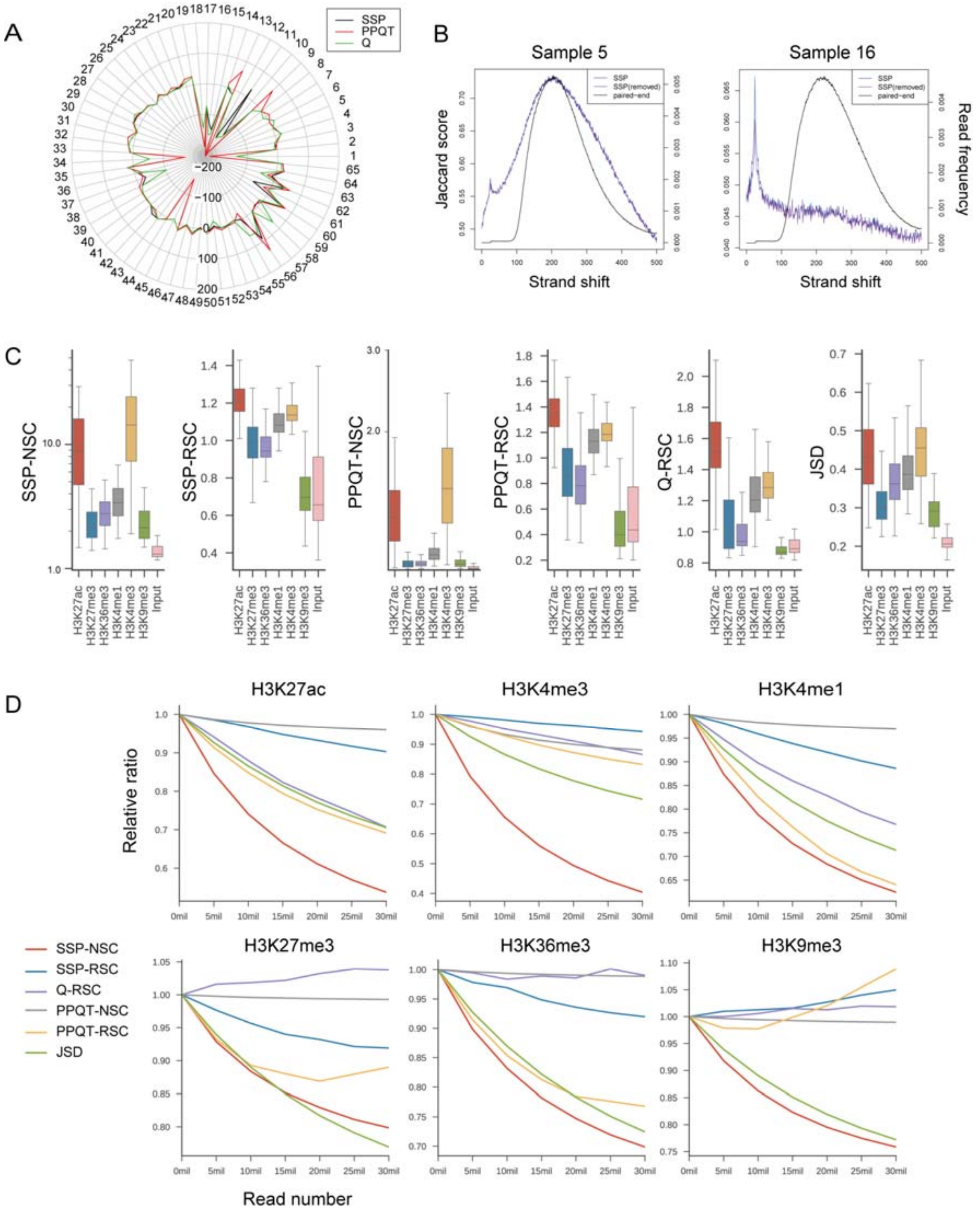
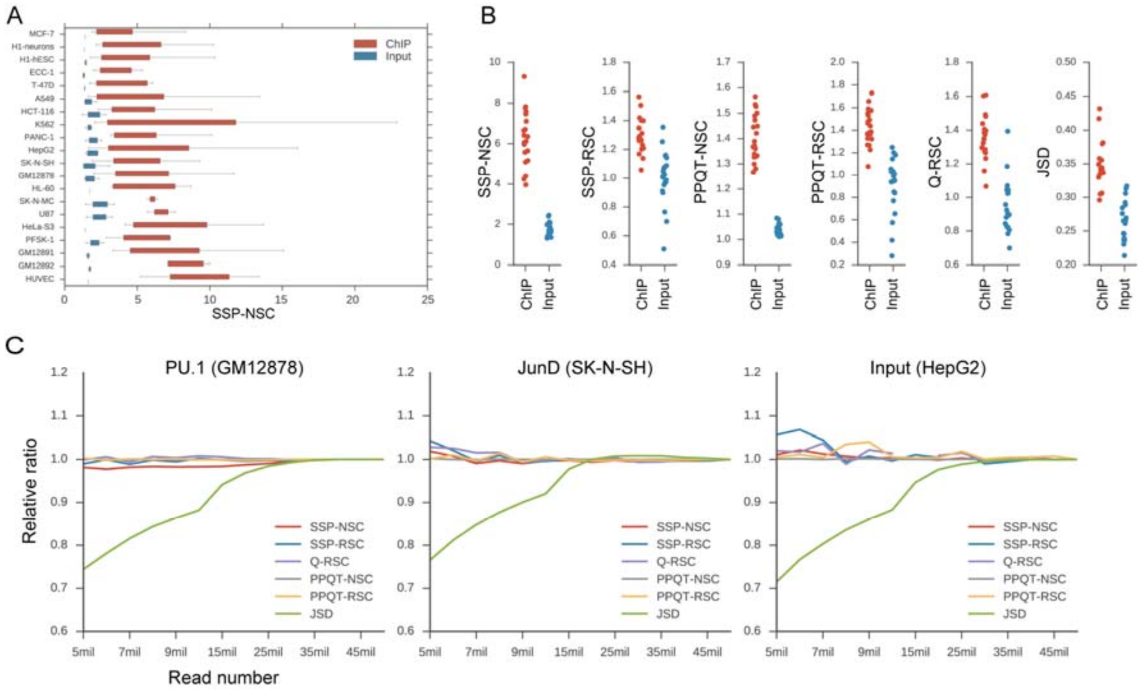


Figure 2. (A) Radar plot of the comparison between the fragment-length estimated by each tool and that from paired-end data for 65 paired-end ChIP-seq data for human (1–45, mapped to build hg38), mouse (46–54, mm10), chicken (55–61, galGal4), and fly (62–65, dm6). The y axis indicates the difference between the fragment size estimated from the

single-end (F3) reads by these tools and that derived from the paired-end reads. See Supplemental Table S1 for names and scores for each sample. (B) Examples of strand-shift profiles by SSP for original data (blue) and after removing blacklist regions (purple), with fragment length distribution inferred by paired-end data (black). Samples that have a clear peak (left) for fragment length can be estimated accurately. The estimation is less accurate if there is no peak and much repetitive enrichment in the sample (right). (C) The distribution of scores by SSP (NSC and RSC), PPQT (NSC and RSC), Q (RSC), and DeepTools (JSD). Note that the y axis is a log-scale for SSP-NSC and PPQT-NSC but a linear scale for the others. (D) Relative ratio at each different S/N (adding input reads from 5 million to 30 million) against original data. Data were obtained from cell type E072 (Brain inferior temporal lobe) of ROADMAP.



619

620 **Figure 3.** (A) SSP-NSC distribution of ChIP (red) and input (blue) for 20 cell types. (B)
621 Median values of S/N indicators for 20 cell types. (C) Relative S/Ns at each sequencing
622 depth (5–50 million) against 50 million reads. Duplicate reads were removed in advance.

623

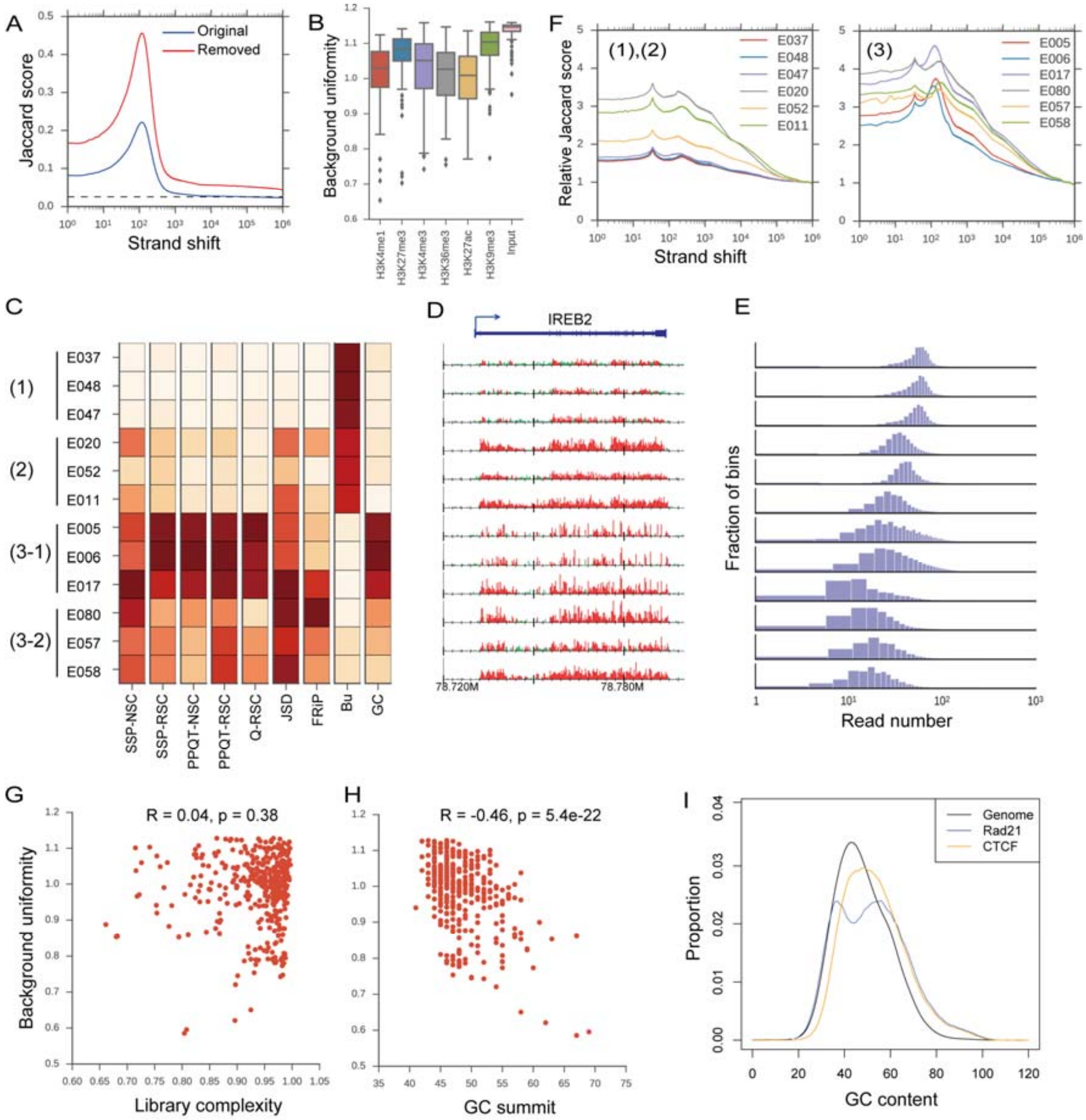


Figure 4. (A) Strand-shift profile for PU.1 data for GM12878 cells from ENCODE. Blue, original data. Red, data after removing mapped reads in every other 10-Mbp window. The horizontal dashed line indicates the expected background level. (B) Distribution of background uniformity for histone modifications. (C–F) Analysis of H3K36me3 data for 12 cell types from ROADMAP. (C) Heatmap of S/N scores alongside Bu and GC peak. Darker colors indicate higher values. See Supplemental Table S4 for the description and detailed scores of each sample. (D) The read distribution around the IREB2 locus (chromosome 15, 78.72–78.80 Mbp). Read number was normalized by the total number of nonredundant

reads. The peak regions identified by MACS2 are highlighted in red. (E) Histogram of mapped read number for each 100-kbp bin of the whole genome except chromosome Y. (F) Strand-shift profiles for the relative Jaccard score against background for groups 1 and 2 (left) and group 3 (right). (G, H) Correlation plot between Bu and library complexity (G) and GC peak (H). (I) Distribution of GC content over the entire genome (black), Rad21_v041610.2 (blue), and CTCF_SC-5916_PCR1x (yellow) for K562 cells.

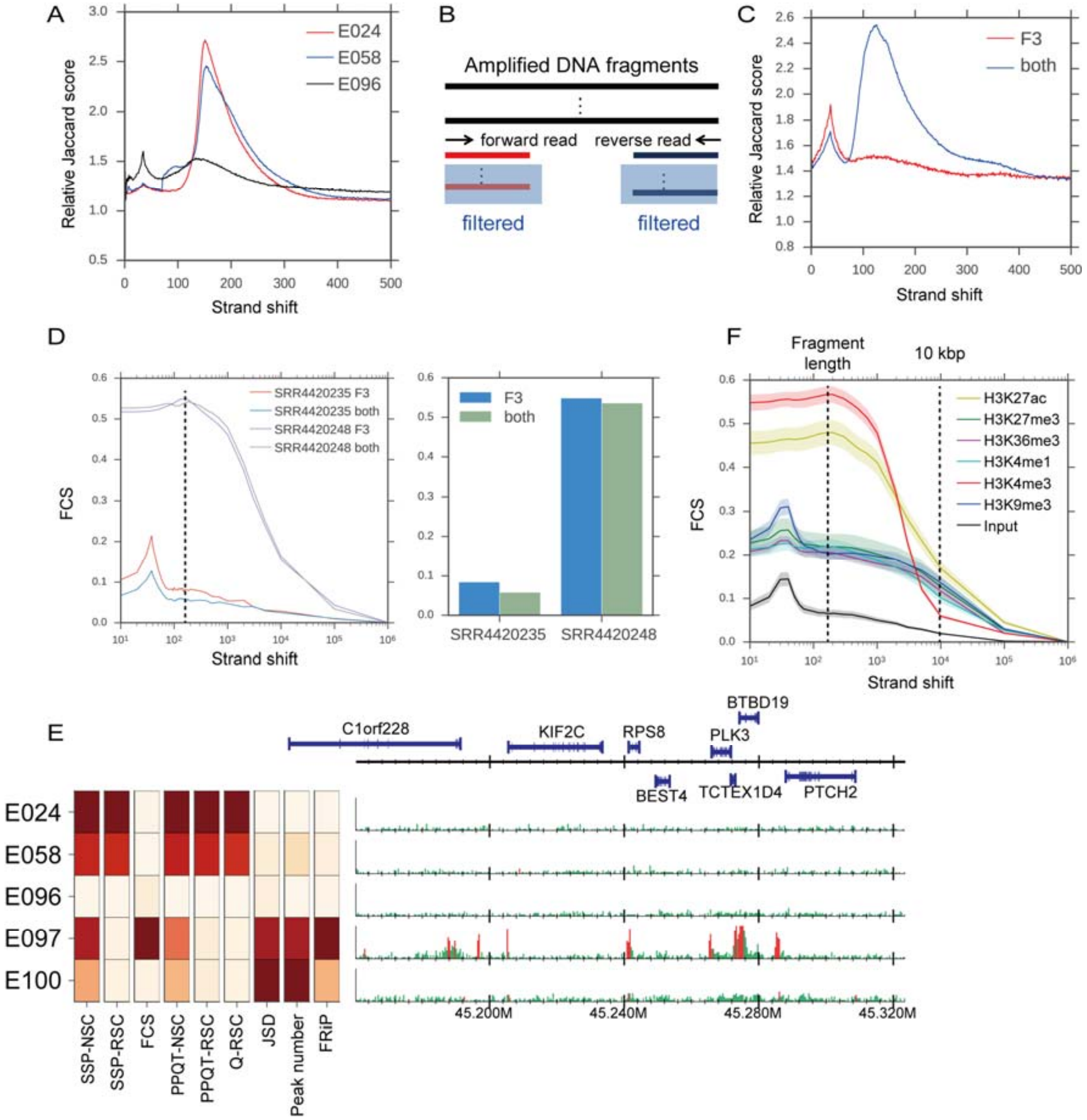


Figure 5. (A) Strand-shift profiles (0 < d < 500) for two input samples (E024, E058), which have apparent peaks at fragment length but actually have low FRiP score (< 0.01). The y-axis indicates the relative Jaccard score against the background level. A typical profile for input (E096) that has a similar FRiP score is also shown. (B) Schematic illustration of hidden duplicate reads. (C) Strand-shift profile for sample SRR4420235 (no. 25 in Fig. 2A) using F3 read only (red) and both F3 and F5 reads as single-end (blue). (D) FCS distribution for each strand shift for each of samples SRR4420235 and SRR4420248 (no. 31 in Fig. 2A) (left). The value at fragment length shown as a dashed vertical line is used as the FCS score

(right). (E) Heatmap for each S/N value and number of peaks identified by MACS2 of five input samples, and their read distribution (chromosome 1, 45.16–45.32 Mbp). The peak regions are highlighted in red. (F) FCS profile for all ROADMAP data. Lines and shaded regions indicate the mean value and 95% confidence interval, respectively, at the each strand shift (x-axis).

Table 1. Spearman's correlation between each S/N and FRiP score without read normalization (top) and with normalization (bottom). *p-value for correlation coefficient < 0.01.

	SSP-NSC	SSP-RSC	PPQT-NSC	PPQT-RSC	Q-RSC	JSD	FCS
ENCODE	*0.84 *0.90	0.10 *0.28	*0.83 *0.88	*0.15 *0.31	*0.16 *0.32	*0.93 *0.81	*0.79 *0.80
ROADMAP (point source)	*0.77 *0.94	*0.20 *0.28	*0.70 *0.89	*0.33 *0.34	*0.29 *0.30	*0.97 *0.90	*0.44 *0.73
ROADMAP (broad source)	*0.72 *0.89	*0.40 *0.31	*0.43 *0.64	*0.37 *0.32	*0.37 *0.31	*0.90 *0.79	*0.55 *0.88