Bioinformatics Advance Access published May 28, 2015

Bioinformatics, 2015, 1–9
doi: 10.1093/bioinformatics/btv293
Advance Access Publication Date: 7 May 2015
Original Paper

OXFORD

Genome analysis

Using combined evidence from replicates to evaluate ChIP-seq peaks

Vahid Jalili[1], Matteo Matteucci[1], Marco Masseroli[1] and Marco J. Morelli[2,*]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133, Milan, Italy and [2]Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), 20139 Milan, Italy

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 6, 2014; revised on April 24, 2015; accepted on May 4, 2015

Abstract

Motivation: Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) detects genome-wide DNA–protein interactions and chromatin modifications, returning enriched regions (ERs), usually associated with a significance score. Moderately significant interactions can correspond to true, weak interactions, or to false positives; replicates of a ChIP-seq experiment can provide co-localised evidence to decide between the two cases. We designed a general methodological framework to rigorously combine the evidence of ERs in ChIP-seq replicates, with the option to set a significance threshold on the repeated evidence and a minimum number of samples bearing this evidence.

Results: We applied our method to Myc transcription factor ChIP-seq datasets in K562 cells available in the ENCODE project. Using replicates, we could extend up to 3 times the ER number with respect to single-sample analysis with equivalent significance threshold. We validated the 'rescued' ERs by checking for the overlap with open chromatin regions and for the enrichment of the motif that Myc binds with strongest affinity; we compared our results with alternative methods (IDR and jMOSAiCS), obtaining more validated peaks than the former and less peaks than latter, but with a better validation.

Availability and implementation: An implementation of the proposed method and its source code under GPLv3 license are freely available at http://www.bioinformatics.deib.polimi.it/MSPC/ and http://mspc.codeplex.com/, respectively.

Contact: marco.morelli@iit.it

Supplementary information: Supplementary Material are available at *Bioinformatics* online.

1 Introduction

Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) is the most commonly used method to study genome-wide chromatin modifications or protein–DNA interactions. Computational tools like MACS (Zhang *et al.*, 2008) or ZINBA (Rashid *et al.*, 2011) are applied on aligned ChIP-seq reads to detect enriched regions (ERs) over the genome (often called 'peaks'), where the local accumulation of sequencing fragments exceeds that of a background distribution, typically estimated from randomly fragmented chromatin or by performing the ChIP-seq protocol with a control antibody (Bailey *et al.*, 2013). As the protocol is subjected to multiple sources of noise (Chen *et al.*, 2012), some low-intensity accumulation of reads is possible even in absence of a true interaction with the target. These low-intensity peaks, which are usually present in large amounts, contain therefore a mixture of false positives and true, although weak, interactions; they are typically discarded when stringent thresholds on the peak call are used. This approach leads to the discovery of the strongest interactions only, and might distort the genome-wide picture of the genomic locations of the transcription factor binding sites or histone modifications of interest.

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Given the intrinsic noise of the ChIP-seq protocol, it is good practice to repeat every experiment at least twice, as the guidelines of the ENCODE project indicate (Landt *et al.*, 2012). The information contained in replicates can then be used to assess the validity of the peaks obtained from a single sample, especially of those with low-intensity.

In this paper, we propose a novel method to rigorously combine the results of peak calls in ChIP-seq replicates and to obtain new, sample-specific, peak lists taking into account their combined evidence. Our method takes as input, for each replicate, a list of enriched genomic regions and a measure of their individual significance in terms of a *P*-value. Starting from a permissive call, we divide the initial ERs in 'stringent' (highly significant) and 'weak' (moderately significant), and we assess the presence of overlapping enriched regions across multiple replicates. Non-overlapping regions can be penalised or discarded according to specific needs. The significance of the overlapping regions is rigorously combined with the Fisher's method to obtain a global score. Finally, this score is assessed against an adjustable threshold on the combined evidence, and peaks in each replicate are either confirmed or discarded (a schematic view of our method is given in Fig. 1 and a visualisation of the results of the method on data from the ENCODE project is shown in Fig. 2). In other words, we are able to 'rescue' weak peaks, which would probably be discarded in a single-sample analysis, when their combined evidence across multiple samples is sufficiently strong.

We applied our method to ENCODE datasets from ChIP-seq experiments of the Myc transcription factor in K562 human cells, for which multiple samples with replicates are available. As Myc preferentially binds to a well-defined motif, the choice of this TF allowed us to validate our results through motif analysis and DNase-seq data; finally, we compared our findings with other state-of-art methods. The strong aspects of our method, besides the validity and relevance of the results that it provides, are its simplicity and flexibility, together with its efficiency (few minutes for 2–3 replicates with a few tens of thousands of peaks each).

2 Methods

Here, a brief description of the method and datasets used is given. For more details, see the Extended Methods section in the Supplementary Material.

2.1 Data collection and peak calling

ChIP-seq enriched regions are read from data files in standard Browser Extensible Data (BED) format; besides standard ER format specifications (columns 'chromosome', 'start', 'end', 'ID'), we require a *P*-value quantifying the significance of each ER, which is usually computed by the peak caller used to identify the ER.

Binary Alignment/Map (BAM) files for the transcription factor Myc in human K562 cells (myelogenous leukaemia) were taken from the ENCODE project repository, for a total of 15 samples obtained in 7 different experiments as summarised in Table 1. Each experiment contained 2 or 3 biological replicates of the same ChIP-seq. Technical replicates were artificially created to test our method, as they were not directly available in the ENCODE repository. Technical replicates were obtained by merging the ENCODE alignment files relative to biological replicates for each of the seven conditions considered above, and then by randomly splitting their reads in two new alignment files. Details about technical replicates, and the process used to generate them, are collected in the Supplementary Table S1.

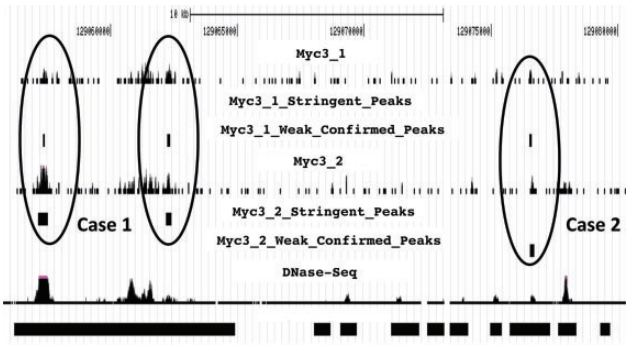


Fig. 2. Genome browser view of a result of the proposed method. Tracks for two ChIP-seq replicates are shown along with the position of the stringent peaks, the weak peaks confirmed by our method and the open chromatin regions (measured as DNase-seq enriched regions). Case 1 refers to a weak peak rescued by a stringent peak, while Case 2 refers to two weak peaks validating each other

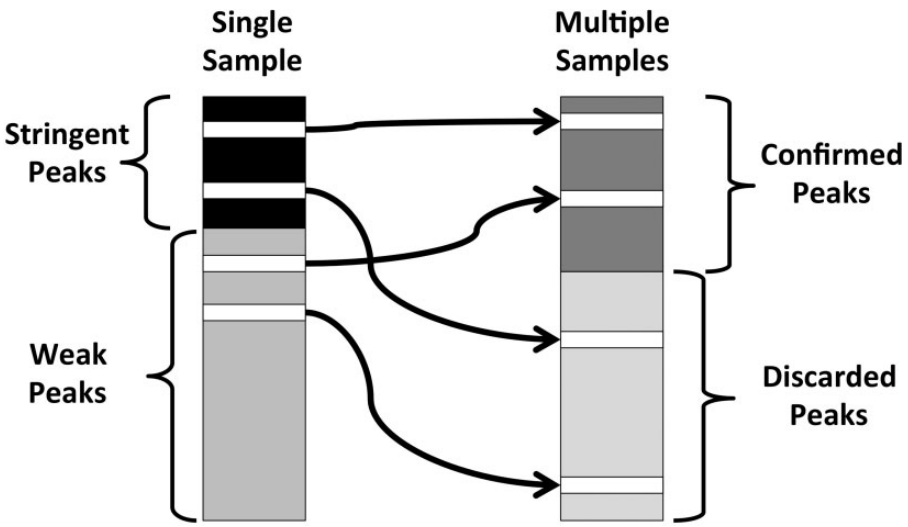


Fig. 1. Pictorial schematic view of the proposed method. First, with a permissive call, we divide peaks of a single individual sample in stringent and weak. Then, combining the evidence of multiple replicates, the peaks in each replicate are confirmed or discarded

Table 1. ENCODE alignment files used and their quantitative features

Sample name	Short name	Aligned reads	R^s	R^w
wgEncodeOpenChromChip K562CmycAlnRep1	Myc1_1	10 719 209	19 171	287 651
wgEncodeOpenChromChip K562CmycAlnRep2	Myc1_2	8 763 362	32 850	311 409
wgEncodeOpenChromChip K562CmycAlnRep3	Myc1_3	9 649 688	13 623	104 911
wgEncodeSydhTfbs K562CmycIggrabAlnRep1	Myc2_1	17 507 194	42 456	64 016
wgEncodeSydhTfbs K562CmycIggrabAlnRep2	Myc2_2	22 256 240	33 015	54 773
wgEncodeSydhTfbs K562CmycStdAlnRep1	Myc3_1	6 077 198	5473	22 965
wgEncodeSydhTfbs K562CmycStdAlnRep2	Myc3_2	5 897 211	12 832	18 753
wgEncodeSydhTfbs K562CmycIfna30StdAlnRep1	Ifna30_1	10 115 596	1901	13 654
wgEncodeSydhTfbs K562CmycIfna30StdAlnRep2	Ifna30_2	18 600 414	2527	97 620
wgEncodeSydhTfbs K562CmycIfna6hStdAlnRep1	Ifna6h_1	9 377 798	5852	12 087
wgEncodeSydhTfbs K562CmycIfna6hStdAlnRep2	Ifna6h_2	19 334 518	4547	102 168
wgEncodeSydhTfbs K562CmycIfng30StdAlnRep1	Ifng30_1	11 602 299	8227	13 190
wgEncodeSydhTfbs K562CmycIfng30StdAlnRep2	Ifng30_2	16 666 560	30 524	25 484
wgEncodeSydhTfbs K562CmycIfng6hStdAlnRep1	Ifng6h_1	14 019 564	2485	13 376
wgEncodeSydhTfbs K562CmycIfng6hStdAlnRep2	Ifng6h_2	19 666 823	27 728	25 118

Peaks were called with the software package MACS2. R^s: stringent ER set (ERs with P-value $< T^s$). R^w: weak ER set (ERs with $T^s \leq P\text{-value} < T^w$). $T^s = 10^{-8}$, $T^w = 10^{-4}$.

Peak calling was performed with the software package MACS2 (Zhang *et al.*, 2008) with the following parameters: '–auto-bimodal -p 0.0001 -g hs' (thus setting a p-value threshold of 10^{-4}), using alignment files available in the ENCODE repository, together with the corresponding background (standard input for all samples except for Myc2, for which the input signal from rabbit IgG ChIP-seq was used). In each sample-input pair, the total number of reads was made equal by randomly down-sampling the largest alignment file. The performed call determined between ∼15 k and ∼345 k peaks across the different samples (see Table 1).

2.2 Definitions

Given a set of J replicates, each sample j is associated with a set R_j of I enriched regions r_{ji}: $R_j = \{r_{j1}, r_{j2}, \ldots, r_{ji}, \ldots, r_{jI}\}$. Each region r_{ji} is defined by $(chromosome_{ji}, start_{ji}, end_{ji}, ID_{ji}, p_{ji})$, where p_{ji} denotes a measure of the significance of r_{ji} (i.e. its P-value). T^s is a stringent threshold on P-values, defining a set R_j^s of stringent (highly enriched) ERs; R_j^s: $r_{ji} \in R_j^s$ *iff* $p_{ji} < T^s$. Similarly, we define a set R_j^w of weak (moderately enriched) ERs, containing all regions whose P-value is between T^s and a weak threshold T^w, with $T^w > T^s$, i.e. R_j^w: $r_{ji} \in R_j^w$ *iff* $T^s \leq p_{ji} < T^w$. Clearly, $R_j^w \cap R_j^s = \emptyset$ and, if T^w is the maximum P-value allowed for an ER to be associated with sample j, $R_j^w \cup R_j^s = R_j$.

For each region i of each sample j, let $r_{ji,k}$ denote the region of sample k overlapping with r_{ji}, if any. If sample k has multiple regions overlapping with r_{ji}, we choose the most significant one, i.e. the one with the lowest p-value. Let R_{ji} be the collection of $r_{ji,k}$ for $k \in \{1, \ldots, J\}$, including r_{ji} itself. Let $K = |R_{ji,*}|$ be the cardinality of $R_{ji,*}$, the set of the ERs intersecting with r_{ji}, with $1 \leq K \leq J$ by definition.

We distinguish between technical and biological replicates of an experiment. Technical replicates aim at controlling the variability of the experimental procedure used to obtain the data and should yield exactly the same results in absence of experimental noise. In a ChIP-seq experiment, this corresponds to performing multiple times the same ChIP protocol on the same biological sample, followed by independent sequencing on the same platform; we expect to observe a significant overlap between ER lists in these samples. Conversely, biological replicates are obtained by applying the same protocol on biologically equivalent samples, what could give rise to different binding profiles of a transcription factor, as in the case of tumor samples; here, the variability in the data can also stem from the 'true' biological variation of the phenomenon of interest. Consequently, the lack of overlap between ERs in biological replicates does not necessarily correspond to a false positive result, as it could reflect a true biological interaction occurring only in some samples. With our method, the user is able to control for the required level of overlap and combined significance, according to the specificities of the dataset.

2.3 Algorithm: overall procedure

The main idea behind our method is that repeated evidence across replicates can compensate for a lower significance in a single sample, which is implemented through the Fisher's method. The Fisher's method combines the P-values of each test in a global test statistics that follows a chi-squared distribution with 2 k degrees of freedom (where k is the number of tests combined); therefore, it can be used to falsify the statement 'all null hypotheses are true', i.e. 'all overlapping ERs are due to background noise'. Comparing intersecting ERs from a set of J replicates is equivalent to test the same genomic region in independent experiments against the same null hypothesis H_0, i.e. 'the number of reads in the region under study is sampled from the background distribution', and obtaining independent probabilities of rejecting H_0 (i.e. independent P-values). Here, we briefly outline the structure and motivation of our algorithm, following the flowchart given in Figure 3, while we discuss its details in the Extended Methods (data structures, search algorithms and combining test statistics sections).

We assign every ER r_{ji} in a given sample j to either R_j^s or R_j^w according to its significance. For a given ER, we then determine $R_{ji,*}$ as the set of ERs in the replicates that overlap with r_{ji}, including r_{ji} itself (see Definitions subsection). The cardinality K of $R_{ji,*}$ represents a measure of the reproducibility of the signal in the region spanned by r_{ji}, while the significance of $r_{ji,k} \in R_{ji,*}$ is a measure of the intensity of the signal in a specific replicate k, given the background. We rigorously combine the significance of the overlapping ERs in $R_{ji,*}$ with the Fisher's method, as described in the Extended Methods, and define a new score for their combined evidence p_{ji}^{comb}. Then, we compare this new score with an adjustable threshold γ: if the desired stringency is obtained, we assign r_{ji} to the set R_j^c of confirmed peaks for sample j; if the condition is not met, i.e. the combined evidence is not strong enough, we assign r_{ji} to the set R_j^d of discarded peaks for sample j. All the ERs in $R_{ji,*}$ are assigned to the corresponding confirmed R_k^c or discarded R_k^d set, respectively.

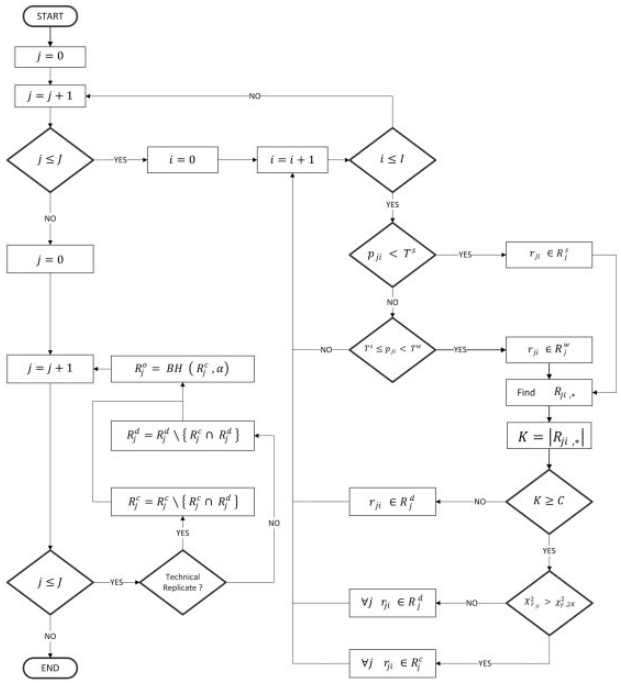


Fig. 3. Flowchart of the proposed method. For the definition of the symbols, see the Definitions subsection of the Methods

We leave the possibility to distrust a region r_{ji}, regardless of its significance, when it is not backed up by the presence of overlapping ERs in a minimum number of samples C. C is an adjustable parameter ranging between 1 and J, with different default values for biological and technical replicates. In summary, for a given sample j, $R_j^c = \{r_{ji} \mid p_{ji}^{comb} \leq \gamma \bigwedge (K \geq C)\}$ and $R_j^d = \{r_{ji} \mid (p_{ji}^{comb} > \gamma) \bigvee (K < C)\}$. We repeat this procedure for each sample.

We note that an ER can participate in different sets of overlapping regions, as we discuss in detail in the Extended Methods. As a consequence, it is possible that an ER is assigned to both the confirmed and discarded sets as a result of different tests. These peaks are assigned to the confirmed set if replicates are biological and to the discarded set if replicates are technical. In other words, as technical replicates are supposed to be very similar, for an ER it is enough to fail the test once to be removed from the confirmed set, while for biological replicates this condition is relaxed and it is enough to pass the test at least once for an ER to be confirmed.

After applying the proposed method, each peak has two properties: its initial significance, which can be either *stringent* (s) or *weak* (w), and the result of the multiple replicate comparison, which can be either *confirmed* (c) or *discarded* (d). Then, for each sample we define four mutually exclusive sets on the basis of these property values: $R^{s,c}$, $R^{w,c}$, $R^{s,d}$, $R^{w,d}$, with $R^c = R^{s,c} \cup R^{w,c}$ and $R^d = R^{s,d} \cup R^{w,d}$. The final output set R_j^o of each sample j is obtained by applying the Benjamini–Hochberg correction procedure to R_j^c, independently from the choice of the other parameters, in order to account for multiple testing (Benjamini and Hochberg, 1995), and keeping only the ERs with false discovery rate smaller than an adjustable threshold α.

2.4 Validation

In order to validate the peaks obtained after combining replicates, we first checked whether the peaks we rescued fell within open chromatin by intersecting their genomic coordinates with enriched regions in DNase-seq data obtained from ENCODE (see Extended Methods). Then, we looked for enriched motifs in the nucleotide sequences spanned by the sets of ERs that we obtained. To perform motif analysis, we used the software package DREME (Bailey, 2011) with parameters '-e 0.00001 -m 10'. Results were scored against the JASPAR database (Mathelier *et al.*, 2014) using the software package TOMTOM (Gupta *et al.*, 2007). Myc strongly binds to one specific motif, called the canonical Enhancer-box or E-box, corresponding to the Position Weight Matrices (PWMs) MA0058 (MAX), MA0059 (MYC::MAX), MA0093 (USF1), MA0104 (Mycn) and MA0147 (Myc) in the Jaspar Core Vertebrata database (the corresponding sequence logos are shown in the Supplementary Fig. S1); thus, an ER set is validated when at least one of these PWMs is found significantly enriched in the ER set. We note that Myc has a weaker affinity for other versions of the E-box, but we chose to exclude these other motifs from the validation to achieve maximum stringency.

2.5 Comparison with other methods

Irreproducibility Discovery Rate (IDR) (Li *et al.*, 2011). It is a metric quantifying the reproducibility of a peak across two ChIP-seq replicates by comparing the two lists of ERs, ranked according to their significance. In essence, after calling the peaks, the IDR pipeline uses a bivariate rank distribution to separate the signal (reproducible peaks) from noise (irreproducible peaks) in an experiment (or pairwise comparison). Each peak is associated with an IDR value, which quantifies the probability that the peak belongs to the irreproducible set. IDR was computed for our validation with the scripts provided by Anshul Kundaje at the URL https://sites.google.com/site/anshulkundaje/projects/idr

jMOSAiCS (Zeng *et al.*, 2013). It is a generic tool for joint analysis of multiple ChIP-seq samples, which can be also used to find common patterns of enrichment between ChIP-seq replicates. First, the MOSAiCS peak caller pre-processes replicates and corresponding control samples by binning the mapped read counts on the genome (default width of 200 bp), and applies the MOSAiCS model fit to each replicate-control pair individually. Afterwards, the jMOSAiCS model is applied to the data fitted with MOSAiCS: region-specific enrichment patterns are determined by posterior probabilities assigned to the internal variables, and a binary variable denotes the potential enrichment of a region based on dependencies among samples. jMOSAiCS was executed with default parameter values as described at http://www.bioconductor.org/packages/release/bioc/vignettes/jmosaics/inst/doc/jmosaics.R. The complete script is included in the Extended Methods.

3 Results

In this section, we report the results obtained by applying our method on either technical or biological ChIP-seq replicates. The method takes as input the genomic coordinates and a measure of significance (i.e. the P-value) of each of the ERs for each replicate considered. For each input sample j, it outputs the lists of confirmed (R_j^o) and discarded (R_j^d) ERs, as well as the lists of stringent confirmed ($R^{s,c}$), weak confirmed ($R^{w,c}$), stringent discarded ($R^{s,d}$) and weak discarded ($R^{w,d}$) ERs.

Adjustable parameters of the method are: T^s (maximum P-value to consider a peak as 'stringent'), T^w (maximum P-value to consider a peak as 'weak'), C (minimum number of samples with intersecting peaks needed to apply the combined evidence evaluation), γ (maximum combined significance to confirm a peak), α (maximum false

discovery rate after the Benjamini-Hochberg correction), together with the choice of 'technical replicate' versus 'biological replicate' mode. For our evaluations we used: $T^s = 10^{-8}$, $T^w = 10^{-4}$, $\gamma = 10^{-8}$, $\alpha = 0.05$ for all comparisons, $C = 1$ for biological replicates and $C = J$ for technical replicates. Required time was a few minutes for 2 samples with 100 000 peaks each on a standard desktop computer.

3.1 Technical replicates

Technical replicates are used to evaluate and remove the noise introduced in the experimental procedure. In the case of ChIP-seq experiments, they are usually generated by performing the same ChIP protocol on the same biological sample, and then performing the sequencing independently. As the ENCODE datasets include only biological replicates, we tested our method on artificial technical replicates, simulated as described in the Methods section. Details about these samples can be found in the Supplementary Table S1. An alternative to our strategy to generate technical replicates would be to randomly split the reads in each original alignment file in replicates rather than merging biological replicates in ENCODE first. However, this procedure gives rise to a much poorer signal, preventing the identification of most ERs. The statistics of these alternative technical replicates are described in the Supplementary Table S2.

Results for $T^s = 10^{-8}$, $T^w = 10^{-4}$, $\gamma = T^s$, $\alpha = 0.05$ and $C = 2$ are shown in Figure 4. For each replicate sample (panels A–G), we show two bars: the left bar (SS) represents the peaks called in a single-sample analysis (R^s in light gray and R^w in dark grey), while the right bar (MS) classifies the same peaks, according to the output of our algorithm, in the four sets described above: $R^{s,c}$ (light gray), $R^{w,c}$ (medium-light grey), $R^{s,d}$ (medium-dark grey) and $R^{w,d}$ (dark grey).

As expected, the number of R^s stringent and R^w weak peaks called in the same technical replicates is always very similar, even if the absolute numbers differ significantly across the different conditions considered. Each output set has a consistent fraction of $R^{w,c}$ weak confirmed ERs, which ranges from 20% to 98% (mean 46%, standard deviation 30%) of the starting number of stringent peaks ($R^{w,c}$ / R^s, panel H); thus by combining evidence in replicates, our method 'rescues' (i.e. confirms) a large amount of weak co-localised peaks that would otherwise be discarded through a usual single sample evaluation. The percentage of stringent discarded peaks ($R^{s,d}$ / R^s, panel H) is very low and varies from 0% in Myc2 to 12% in Myc3 (mean 5.6%, standard deviation 3.8%). The output set R^o corresponds to the set of confirmed peaks $R^c = R^{s,c} \cup R^{w,c}$, where the significance of each peak has been adjusted for multiple testing; combining the evidence present in replicates increases the number of obtained peaks up to almost the double of what obtained with a single sample at the same stringency (R^o/ R^s, panel H).

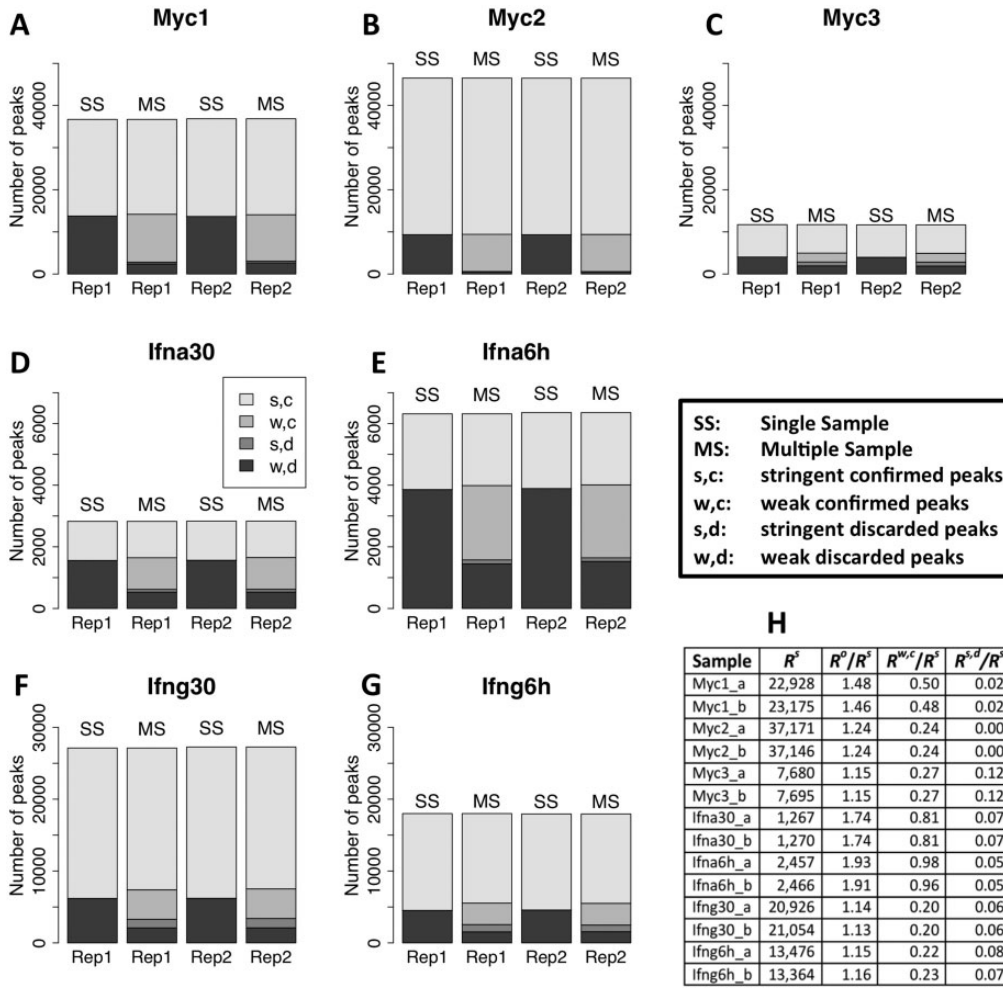
For technical replicates, we expect the output of each replicate to be similar, and therefore the parameter C was set to $C = J = 2$ for all technical replicate comparisons. Setting $C = 1$ would be instead equivalent to trust even isolated peaks, which are not present in the other replicate. With the latter choice, and $\gamma = T^s$, no stringent peaks would be discarded.

As a preliminary evaluation of the results obtained, we considered the overlap of the peaks with the enriched regions in DNase-seq data. On average, 95.4% of the peaks in R^o, 95.4% of peaks in R^s and 94.6% of peaks in $R^{w,c}$ were in open chromatin regions, while this fraction was only 89.4% for $R^{s,d}$ and 93.0% for $R^{w,d}$ (Supplementary Table S3). The overlap with open chromatin, however, is not yet a validation of a specific binding event. We performed then motif analysis on the nucleotide sequences corresponding to the ERs in the four sets: R^s, R^o, $R^{w,c}$ and $R^{s,d}$. Myc is known to bind a large number of sites on the DNA, particularly with high affinity to those with the 6-nucleotide motif called *Enhancer-box* or *E-box*. This protein-binding region has the generic consensus nucleotide sequence *CANNTG* (with N representing any nucleotide; Murre *et al.*, 1989). In particular, Myc binds with maximum strength to the *CACGTG* motif (also called the 'canonical' Myc E-box [Walhout *et al.*, 1997]). Therefore, we consider the enrichment of the E-box in a set of peaks a sufficient condition to consider the set as containing 'true' binding sites. Panel I in Figure 4 shows that the E-box is always enriched in the R^s stringent and R^o output sets, as well as in the $R^{w,c}$ weak confirmed set. This result confirms that in the large majority of cases the weak peaks overlapping in replicates identify real binding sites, which are missed by a stringent single-sample call. In 4 out of 14 cases, the $R^{s,d}$ stringent discarded set is enriched for the E-box, although at much lower significance (Supplementary Table S5), while in the remaining cases the number of peaks in the $R^{s,d}$ set is low. This analysis suggests that the default value $C = J$ used for our artificial technical replicates may be too conservative and still discards a small fraction of real binding sites.

3.2 Biological replicates

The ENCODE data repository always includes one or more biological replicates for each ChIP-seq experiment. For the transcription factor Myc, multiple data sources are available, either obtained in independent experiments, or scored against different backgrounds (in Myc2, the input was derived from immuno-precipitating normal rabbit IgG, while in all the other samples the standard input for the K562 cell line was used). We applied our method to biological replicates obtained from each of the ENCODE experiments considered, and we also combined replicates from 2 experiments (Myc2 and Myc3). Parameters for the method were the same as for the technical replicate evaluation reported in the previous section (i.e. $T^s = 10^{-8}$, $T^w = 10^{-4}$, $\gamma = T^s$, $\alpha = 0.05$); for the additional parameter C, in the case of biological replicates we adopted the permissive choice of $C = 1$ (default for the analysis of biological replicates). With these values (i.e. $\gamma = T^s$ and $C = 1$), our method never discards a stringent peak (we consider that a single strong evidence is enough for biological replicate evaluation). Results are shown in Figure 5.

The number of peaks in biological replicates of the same experiment can be very different (panels A-H), reflecting the different efficiency of the ChIP-seq protocol, and the number of weak peaks (R^w) is usually much larger than the number of stringent peaks (R^s). In the considered cases, the number of confirmed weak peaks ($R^{w,c}$) is often much bigger (up to ~4 times) than the number of stringent peaks (R^s) (column $R^{w,c}/R^s$ in panel I), confirming that the evidence in a 'good' replicate allows the rescue of many peaks in a 'bad' replicate. We observe a similar situation when we combine samples obtained with different inputs. For example, by combining together the four replicates of the Myc2 and Myc3 cases (panel D), we increase massively the number of peaks in the output set for the samples with lower peak counts (Myc3) by confirming a number of their weak peaks much larger than in the evaluation performed without Myc2. Therefore, the presence of high-quality replicates can be of great help in improving the call on many low-quality replicates. The average overlap with open chromatin regions of the $R^{w,c}$ weak confirmed sets is 91.0% (compared with the 51.6% for the weak discarded peaks), and motif analysis confirms that in all the samples the ERs contain the canonical Myc binding site (panel J and Supplementary Tables S6 and S8).



H

Sample	R^s	R^o/R^s	$R^{w,c}/R^s$	$R^{s,d}/R^s$
Myc1_a	22,928	1.48	0.50	0.02
Myc1_b	23,175	1.46	0.48	0.02
Myc2_a	37,171	1.24	0.24	0.00
Myc2_b	37,146	1.24	0.24	0.00
Myc3_a	7,680	1.15	0.27	0.12
Myc3_b	7,695	1.15	0.27	0.12
Ifna30_a	1,267	1.74	0.81	0.07
Ifna30_b	1,270	1.74	0.81	0.07
Ifna6h_a	2,457	1.93	0.98	0.05
Ifna6h_b	2,466	1.91	0.96	0.05
Ifng30_a	20,926	1.14	0.20	0.06
Ifng30_b	21,054	1.13	0.20	0.06
Ifng6h_a	13,476	1.15	0.22	0.08
Ifng6h_b	13,364	1.16	0.23	0.07

I

Sample	R^s	R^o	$R^{w,c}$	$R^{s,d}$
Myc1	Y	Y	Y	-
Myc1	Y	Y	Y	-
Myc2	Y	Y	Y	-
Myc2	Y	Y	Y	-
Myc3	Y	Y	Y	-
Myc3	Y	Y	Y	-
Ifna30	Y	Y	Y	-
Ifna30	Y	Y	Y	-
Ifna6h	Y	Y	Y	-
Ifna6h	Y	Y	Y	-
Ifng30	Y	Y	Y	Y
Ifng30	Y	Y	Y	Y
Ifng6h	Y	Y	Y	Y
Ifng6h	Y	Y	Y	Y

Fig. 4. Technical replicates. For each of the 7 experiments considered, two technical replicates were obtained by pooling reads from the biological replicates of the conditions and then randomly splitting the resulting alignment files in two equal parts. A-G: ER sets for the technical replicates considered. SS, single sample analysis; MS, multiple sample analysis. In each panel, the SS stacked bars represent R^s (light gray) and R^w (dark grey) in the two replicates, while the MS bars show the same peaks, confirmed or discarded according to the output of our method: $R^{s,c}$ (light gray), $R^{w,c}$ (medium-light gray), $R^{s,d}$ (medium-dark gray) and $R^{w,d}$ (dark grey). H, general statistics on the cardinality of the ER sets; I, validation of the sets with the Myc binding motif (Myc canonical E-box); 'Y', presence of the E-box; '-', set too small to find any enriched motif. See Supplementary Table S5 for E-box enrichment *P*-values

We applied our method also using $C=2$ on all biological replicates considered: most of the times that the $R^{s,d}$ stringent discarded peak set had a substantial size, the E-box motif was present, although the average overlap with open chromatin was only 75.9% (Supplementary Fig. S2 and Tables S9 and S11). This further confirms that, in biological replicates, a lack of overlapping with peaks in other replicates does not necessarily correspond to an artifactual interaction.

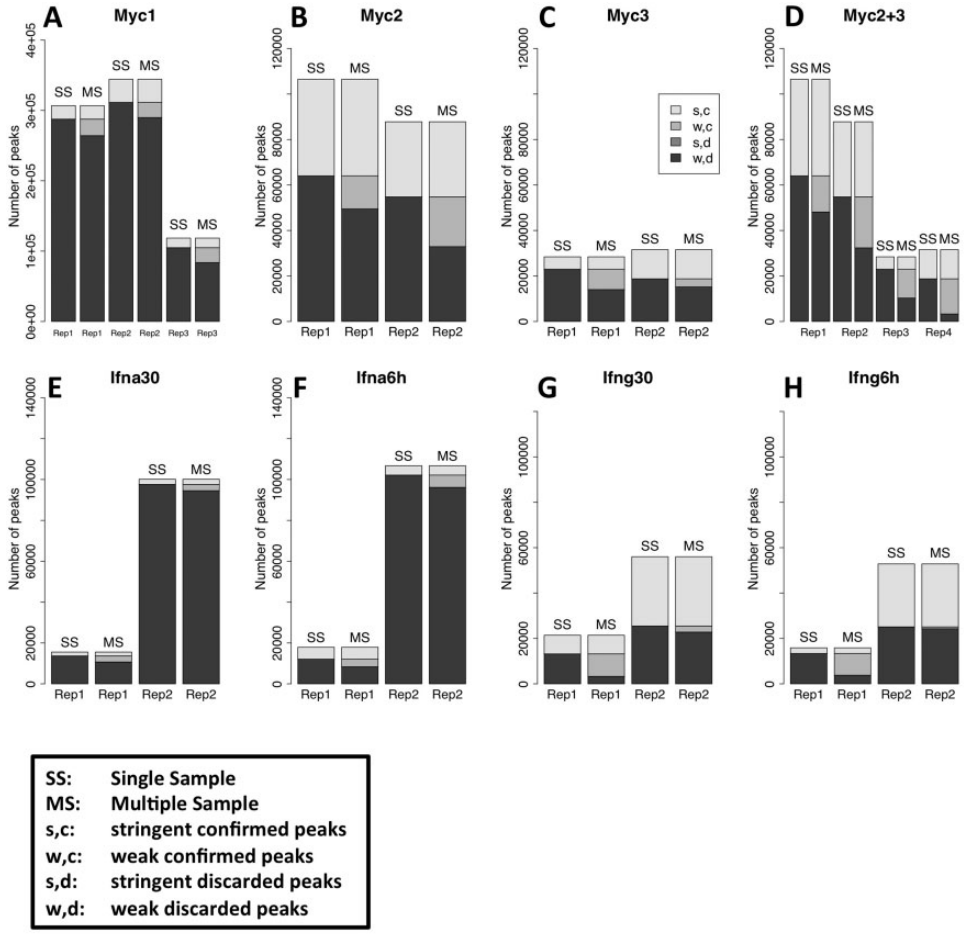
3.3 Comparison with alternative strategies

3.3.1 Alignment read merging

An intuitive way to combine evidence in replicates is to merge the alignment reads, and use a peak caller on the combined data. As the combined dataset corresponds to the sum of the two signals, weak, co-occurring peaks should increase their significance. We have merged alignment files from replicates for each considered experiment, using merged backgrounds when available (Myc3 only), and considered only peaks with *P*-value smaller than $T^s=10^{-8}$. In almost all the cases, the number of peaks called from the merged replicates was substantially lower than the number of peaks obtained by our algorithm (Table 2, third and fourth columns). Moreover, a large fraction of the peaks detected in the merged samples overlapped with at least one of the peaks obtained by our method (Table 2, fifth column). The only cases when this fraction was below 70% were those where the replicates exhibited a very large difference in the number of called peaks (Myc3, Ifng30, Ifng6h). In these cases, the output set of the replicate with the higher number of peaks always showed a very high overlap with the merged sample peaks. Merging the alignment files therefore 'averages' replicates with different sets of peaks, whereas our method 'rescues' a sample with few ERs with the help of a sample with many ERs. Besides, the merging strategy has no user-defined parameter to tune the results, whereas our method provides a rigorous way to weight co-occurrence and significance of ERs.

3.3.2 Irreproducibility discovery rate

The IDR (Li *et al.*, 2011) is a measure of the consistency of ERs identified in replicates, which has been systematically assessed in the ENCODE project. We computed the IDR for the ERs in our samples and used an IDR threshold of 0.05. The results are shown in the



I

Sample	R^s	R^o/R^s	$R^{w,c}/R^s$
Myc1	19171	2.23	1.23
Myc1	32850	1.66	0.66
Myc1	13623	2.56	1.56
Myc2	42456	1.34	0.34
Myc2	33015	1.66	0.66
Myc3	5473	2.63	1.63
Myc3	12832	1.28	0.28
Myc2+3	42456	1.37	0.37
Myc2+3	33015	1.68	0.68
Myc2+3	5473	3.31	2.31
Myc2+3	12832	2.2	1.2
Ifna30	1901	2.55	1.55
Ifna30	2527	2.22	1.22
Ifna6h	5852	1.64	0.64
Ifna6h	4547	2.32	1.32
Ifng30	8227	2.2	1.2
Ifng30	30524	1.09	0.09
Ifng6h	2485	4.85	3.85
Ifng6h	27728	1.03	0.03

J

Sample	R^s	R^o	$R^{w,c}$
Myc1_1	Y	Y	Y
Myc1_2	Y	Y	Y
Myc1_3	Y	Y	Y
Myc2_1	Y	Y	Y
Myc2_2	Y	Y	Y
Myc3_1	Y	Y	Y
Myc3_2	Y	Y	Y
Myc2+3_1	Y	Y	Y
Myc2+3_2	Y	Y	Y
Myc2+3_3	Y	Y	Y
Myc2+3_4	Y	Y	Y
Ifna30_1	Y	Y	Y
Ifna30_2	Y	Y	Y
Ifna6h_1	Y	Y	Y
Ifna6h_2	Y	Y	Y
Ifng30_1	Y	Y	Y
Ifng30_2	Y	Y	Y
Ifng6h_1	Y	Y	Y
Ifng6h_2	Y	Y	Y

Fig. 5. Biological replicates. A–H: ER sets in the biological replicates considered. SS, single sample analysis; MS, multiple sample analysis. In each panel, the SS stacked bars represent R^s (light gray) and R^w (dark grey) in the replicates, while the MS bars show the same peaks, confirmed or discarded according to the output of our method: $R^{s,c}$ (light gray), $R^{w,c}$ (medium-light gray), $R^{s,d}$ (medium-dark gray) and $R^{w,d}$ (dark grey). I, general statistics on the cardinality of the ER sets; J, validation of the sets with the Myc binding motif (Myc canonical E-box); 'Y', presence of the E-box. See Supplementary Table S8 for E-box enrichment *P*-values

Table 2. Comparison with merged alignment files and IDR

Sample	R^s	R^o	Merged	$R^o \cap$ Merged	IDR < 0.05	$R^o \cap$ IDR < 0.05
Myc1_1	19 171	42 663	27 958	20 359 (73%)	3097	3097 (100%)
Myc1_2	32 850	54 420	27 958	22 265 (80%)	4618	4618 (100%)
Myc1_3	13 623	34 858	27 958	20 112 (72%)	4966	4964 (99%)
Myc2_1	42 456	56 989	39 805	38 958 (98%)	24 066	23 767 (99%)
Myc2_2	33 015	54 889	39 805	37 765 (95%)	24 066	23 767 (99%)
Myc3_1	5473	14 411	15 404	9 252 (60%)	2356	2237 (95%)
Myc3_2	12 832	16 401	15 404	12 483 (81%)	2356	2237 (95%)
Ifna30_1	1901	4839	3650	2918 (80%)	1171	811 (69%)
Ifna30_2	2527	5605	3650	2914 (80%)	1171	823 (70%)
Ifna6h_1	5852	9570	6633	5913 (89%)	2274	2078 (91%)
Ifna6h_2	4547	10 547	6633	5671 (85%)	2274	2078 (91%)
Ifng30_1	8227	18 112	32 363	15 828 (49%)	5586	5586 (100%)
Ifng30_2	30 524	33 203	32 363	30 307 (94%)	5586	5586 (100%)
Ifng6h_1	2485	12 052	21 145	8506 (40%)	5181	4352 (84%)
Ifng6h_2	27 728	28 564	21 145	20 187 (96%)	5181	5079 (98%)

Comparison of the output set R^o (3 rd column) with ERs obtained by merging replicates (4th column) and with ERs having Irreproducibility Discovery Rate (IDR) < 0.05 (6th column). The number of overlapping peaks between R^o and the two other methods (5th and 7th columns, respectively) is shown, together with the fraction of the other method peaks overlapping with peaks in R^o. In general, R^o is larger and comprises the majority of the peaks obtained by the other methods.

Table 3. Comparison with jMOSAiCS

Sample	R^s	R^o	jMOSAiCS	R^o\jMOSAiCS	E-box	jMOS AiCS\R^o	E-box
Myc1_1	19 171	42 663	50 539	7649	Y	31 750	N
Myc1_3	13 623	34 858	33 867	4156	Y	17 249	N
Myc2_1	42 456	56 989	91 252	1346	Y	46 979	N
Myc2_2	33 015	54 889	92 423	799	Y	50 428	Y
Myc3_1	5473	14 411	18 244	2801	Y	9061	N
Myc3_2	12 832	16 401	27 116	1390	Y	13 181	Y
Ifna30_1	1901	4839	32 711	111	-	28 752	Y
Ifna30_2	2527	5605	30 695	527	-	26 616	Y
Ifna6h_1	5852	9570	38 517	114	Y	29 703	N
Ifna6h_2	4547	10 547	36 258	408	Y	28 028	Y
Ifng30_1	8227	18 112	49 843	257	Y	33 763	N
Ifng30_2	30 524	33 203	69 128	150	-	39 996	N
Ifng6h_1	2485	12 052	31 160	483	Y	22 678	Y
Ifng6h_2	27 728	28 564	70 888	98	-	45 601	Y

Comparison of the output set R^o (3rd column) with ERs obtained by jMOSAiCS (4th column). The 5th and 6th columns show the number of peaks that are present in R^o but not in the jMOSAiCS output, and the enrichment of the Myc canonical E-box in this last set, respectively. Vice versa, the 7th and 8th columns show the number of peaks present in the jMOSAiCS output but not in R^o, and the corresponding enrichment of the E-box, respectively. For Myc1, the comparison was done only with replicates 1 and 3. jMOSAiCS outputs a large number of ERs, including most of the peaks identified by our method, but only a fraction of the jMOSAiCS-specific peaks contains the Myc binding motif. Each E-box column refers to the set described in the previous column and is marked as follows: 'Y': presence of the E-box; 'N': absence of the E-box; '-': set too small to find any enriched motif.

sixth and seventh columns of Table 2. First of all, the IDR-validated peaks are few and often entirely contained in our output sets. This highlights that the IDR method is rather stringent and generates only a small set of validated (reproducible) peaks. We repeated our analysis for the Myc2 sample for a lower T^w threshold and found similar results (see Supplementary Material). Our method does not score the reproducibility of a peak, but it rather combines evidence in replicates, and has the option to accept very stringent peaks even if they are not found in other replicates. We conclude that our method confirms weak peaks that are considered 'irreproducible' by IDR (with a 0.05 threshold), and are validated by motif analysis (Figs 4 and 5, Supplementary Tables S3–S11). Finally, differently from our method, which accepts any number of replicates, the IDR can be directly computed only for pairs of replicates.

3.3.3 jMOSAiCS

We compared our results also with those obtained with jMOSAiCS (Zeng *et al.*, 2013), a tool designed to detect combinatorial patterns of enrichment in multiple ChIP-seq samples. Even if jMOSAiCS is conceived to integrate different ChIP-seq datasets that profile distinct features on the same biological sample, it can also be applied to replicates of the same ChIP-seq. Applying jMOSAiCS to our biological replicates (Table 3) resulted in a very large amount of ERs in each experiment, which were on average around 5 times larger than our peaks (ER average size data are not shown). These sets of peaks contained by far the majority of the ERs identified by our method. We checked for the enrichment of the Myc E-box in the peaks identified by our method, but not by jMOSAiCS and vice versa (peaks identified by jMOSAiCS, but not by our method). While in the former case the Myc binding motif was enriched in all the sets with an enough number of peaks to detect any enriched motif, the latter case showed the presence of the Myc canonical E-box only in half of the samples. Moreover, the running times of jMOSAiCS for two replicates were in the order of 3 hours, with about 40 GB of memory consumption, on a server with two Intel Xeon E5-2650 processors and 64 GB of RAM, as this tool starts from alignment files and finds ERs independently. On the same platform, our method ran in a few minutes; the preliminary peak calling step needed to obtain the sets of enriched regions required about 40 min for each of the replicates with MACS2 (Zhang *et al.*, 2008). We conclude that jMOSAiCS, when applied to replicates of the same ChIP-seq experiment, has the tendency of introducing a large amount of extra peaks, which are not always validated by motif analysis, and it requires significant computational resources. On the other hand our method is much faster, as it allows to start from pre-determined ER lists, and at least equally specific.

4 Discussion

We introduced a novel and rigorous method to combine evidence in ChIP-seq replicates and we applied it to several ENCODE datasets for the transcription factor Myc in the K562 cell line. Our results confirmed that a considerable number of ERs, which display a weak significance in single-sample analysis, can be 'rescued' by the help of co-localised evidence in multiple replicates. We proved that the nucleotide sequences spanned by these weak peaks are almost always found in open chromatin regions and enriched for the Myc canonical E-box motif, for which the Myc protein has the highest affinity. Surprisingly, even in the case of technical replicates, where reproducibility should be high, we found that discarding ERs only on the base of the lack of overlap often results in the dismissal of true binding sites. This can be due to the fact that our technical replicates were simulated using a computational procedure starting from the biological replicates available; nonetheless, we recommend to be careful in setting the overlapping parameter to high stringency (i.e. $C = J$).

We stress that our method works as a post-processing of a permissive peak call and it does not question the reliability of the output of the peak caller (any peak caller providing a P-value score can be used; we recommend using the same peak caller with the same parameters on all the replicates). The method has three main strengths: (i) rigour: single-sample evidence from each replicate is combined through the Fisher's method; (ii) versatility: with the choice of a few parameters, it can be decided to weight co-localisation (C) and combined significance (γ) differently; (iii) efficiency: typically, the

required time is in the order of few minutes on a standard desktop computer and does not require special hardware.

Comparing our method with two other common approaches (replicate merging and IDR) confirmed the stable identification of a core of stringent, reproducible peaks. Besides this, our tests demonstrated that less stringent evidence consistently present across replicates can be combined, leading to the 'rescue' of sets of ERs corresponding to real binding sites, e.g. of the transcription factor (TF) Myc. In particular, our results are compatible with those found with IDR, a method widely used in ENCODE to assess the consistency of each detected peak: IDR works by comparing peak rankings and inferring the proportion of reproducible and irreproducible signal in the replicates, while our algorithm provides complementary information by computing the combined significance of a number of overlapping peaks. Despite comparable running times, we differ from IDR as we do not automatically discard non-overlapping peaks and we can directly apply our method to more than two replicates without relying on multiple pairwise comparisons. We stress that our method should not be considered an alternative to IDR, but rather complementary to it.

A further comparison with a tool designed for more complex analysis (identifying combinatorial patterns of enrichment across different ChIP-seq experiments performed over the same biological sample), jMOSAiCS, revealed that, in the specific task of comparing replicates of ChIP-seq experiments performed against the same target, this last tool confirms more peaks, which however do not always enrich for the E-box.

Recently, JAMM (Ibrahim *et al.*, 2015), a tool based on local multivariate Gaussian mixture models for directly finding ERs on ChIP-seq replicates, has been introduced. JAMM confirms that pooling replicates can blur the specific spatial resolution of single-sample peaks and lead to less accurate calls in terms of peak width and intensity.

In summary, our strategy represents a promising trade-off between stringent techniques (IDR) and permissive techniques (jMOSAiCS).

Acknowledgements

We thank Mattia Pelizzola and Stefano de Pretis for useful discussions.

Funding

This work was supported by the Fondazione Istituto Italiano di Tecnologia and by AIRC [grant no. IG_13182], and it is part of and supported by the 'Data-Driven Genomic Computing (GenData 2020)' PRIN project (2013–2015), funded by the Italian Ministry of the University and Research (MIUR).

Conflict of Interest: none declared.

References

Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.

Bailey,T. *et al.* (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003326.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B*, 289–300.

Chen,Y. *et al.* (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods*, **9**, 609–614.

Gupta,S. *et al.* (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

Ibrahim,M.M. *et al.* (2015) JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, **31**, 48–55.

Landt,S.G. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

Li,Q. *et al.* (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

Mathelier,A. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.

Murre,C. *et al.* (1989) A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. *Cell*, **56**, 777–783.

Rashid,N.U. *et al.* (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.*, **12**, R67.

Walhout,A.J. *et al.* (1997) c-Myc/Max heterodimers bind cooperatively to the E-box sequences located in the first intron of the rat ornithine decarboxylase (ODC) gene. *Nucleic Acids Res.*, **25**, 1493–1501.

Zeng,X. *et al.* (2013) jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol.*, **14**, R38.

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