Cremona et al. BMC Bioinformatics (2015) 16:349
DOI 10.1186/s12859-015-0787-6

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Peak shape clustering reveals biological insights

Marzia A. Cremona[1], Laura M. Sangalli[1], Simone Vantini[1], Gaetano I. Dellino[2,3], Pier Giuseppe Pelicci[2,3], Piercesare Secchi[1] and Laura Riva[4*]

Abstract

Background: ChIP-seq experiments are widely used to detect and study DNA-protein interactions, such as transcription factor binding and chromatin modifications. However, downstream analysis of ChIP-seq data is currently restricted to the evaluation of signal intensity and the detection of enriched regions (peaks) in the genome. Other features of peak shape are almost always neglected, despite the remarkable differences shown by ChIP-seq for different proteins, as well as by distinct regions in a single experiment.

Results: We hypothesize that statistically significant differences in peak shape might have a functional role and a biological meaning. Thus, we design five indices able to summarize peak shapes and we employ multivariate clustering techniques to divide peaks into groups according to both their complexity and the intensity of their coverage function. In addition, our novel analysis pipeline employs a range of statistical and bioinformatics techniques to relate the obtained peak shapes to several independent genomic datasets, including other genome-wide protein-DNA maps and gene expression experiments. To clarify the meaning of peak shape, we apply our methodology to the study of the erythroid transcription factor GATA-1 in K562 cell line and in megakaryocytes.

Conclusions: Our study demonstrates that ChIP-seq profiles include information regarding the binding of other proteins beside the one used for precipitation. In particular, peak shape provides new insights into cooperative transcriptional regulation and is correlated to gene expression.

Keywords: ChIP-seq, Transcription regulation, GATA-1, Peak shape

Background

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a widely used technique essential to study transcription factor binding and chromatin modifications. This technique has been largely used to characterize many biological processes, enabling the creation of valuable public resources of epigenomic data (i.e. ENCODE, Roadmap Epigenomics). Due to the importance of interpreting these datasets, a large number of algorithms for downstream processing of ChIP-seq experiments have been developed [1, 2]. All these methods are usually based on the evaluation of signal intensities to detect local enrichment of uniquely aligned reads on the reference genome (we refer to them as 'ChIP-seq peaks'). Peak shape shows high variability among the ChIP-seq experiments that investigate different proteins as well as among different genomic regions in a single ChIP-seq. This variability is not only related to peak intensity [3]. Indeed, the shapes of a transcription factor (TF) usually appear concentrated narrowly, while peaks that characterize histone marks can sometimes spread over a large region [4, 5].

Recently, peak shape properties different from signal intensity have been used in peak calling [6–8], peak ranking [9] and ChIP-seq differential analysis [10]. While the developed methods show that additional features of peak shape can improve peak detection, here we want to understand whether peak shape includes additional biological properties that have not been explored yet. Our hypothesis is that peak shape is influenced by the organization and interactions of the proteins bound to the DNA, hence we want to understand if the detection of differences in peak shape in a single ChIP-seq

* Correspondence: laura.riva@iit.it
[4]Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia, Milan, Italy
Full list of author information is available at the end of the article

BioMed Central

© 2015 Cremona et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

experiment can shed light on the binding of cooperative transcription factors. We are also interested in assessing whether the organization and interactions of these transcription factors is correlated to the genomic context and to gene expression. In order to address these questions, we propose an innovative analysis pipeline that distinguishes different shapes in a set of ChIP-seq peaks and relates the obtained profiles to several independent genomic datasets (other ChIP-seq experiments for different transcription factors and for histone marks, DNase-seq and RNA-Seq data). In our method, we use cluster analysis to evaluate whether the peaks of a ChIP-seq can be divided into groups, according to both the complexity and the intensity of the coverage function that defines them. To achieve this goal we select five shape indices, embedding the problem into the framework of multivariate statistical analysis. We also employ a wide range of statistical techniques to correlate the shape with a functional role. The software SIC-ChIP (Shape Index Clustering for ChIP-seq peaks), which computes the shape indices and clusters the peaks, is available online [11] as a command line R script.

To clarify the meaning of peak shape, we decide to study the erythroid transcription factor GATA-1 (GATA binding protein 1) in K562 cell line and in megakaryocytes. In this particular setting, we show that peak shapes contain information that can be used to shed light on cooperative binding and to identify up-regulated genes. Moreover, we apply the proposed methodology to a set of ChIP-seq experiments in K562 and we discover that peak shape can vary depending on the different binding proteins under investigation. Here we mainly concentrate our attention on the study of peak shape of transcription factors, but the same ideas can be generalized to other types of protein-DNA interactions.

Results and discussion

Peak shape varies among different experiments

We observe that peak shape is quite reproducible as technical and biological replicates obtained with the same library preparation protocols (see Additional file 1: Table S1 and Additional file 2) give rise to the same signal in the same genomic region (Fig. 1a). This is true even if ChIP-seq efficiencies for independent experiments can vary [12] as in Fig. 1a. In addition, if two antibodies are used to perform chromatin immunoprecipitation for the same transcription factor, a subset of peaks might have different peak shapes (Fig. 1b). The antibodies may recognize different epitopes of the same transcription factor and this fact can lead to differences in shapes. Moreover, transcription factor interactions can be cell-type specific, and we observe that ChIP-seq peaks obtained using the same antibody in different cell types can show a subset of diverse shapes (Fig. 1c). These observations suggest that the analysis of peak shape may reveal insights regarding cooperation and association of transcription factors.

It is important to point out that different library preparation protocols might affect peak shape. While the read length of a ChIP-seq experiment does not have any effect on peak shape, fragment length influences peak shape as differences in fragment lengths result in different signal resolutions (larger fragments generate smoother, less resolved and bigger peak). However, the methodology we propose is not affected by differences in library preparation and sequencing, since it considers a single ChIP-Seq at a time and clusters peaks belonging to the same experiment.

Overview of the analysis pipeline proposed

The analysis pipeline that we propose is summarized in Fig. 2. First, we perform a pre-processing step to produce

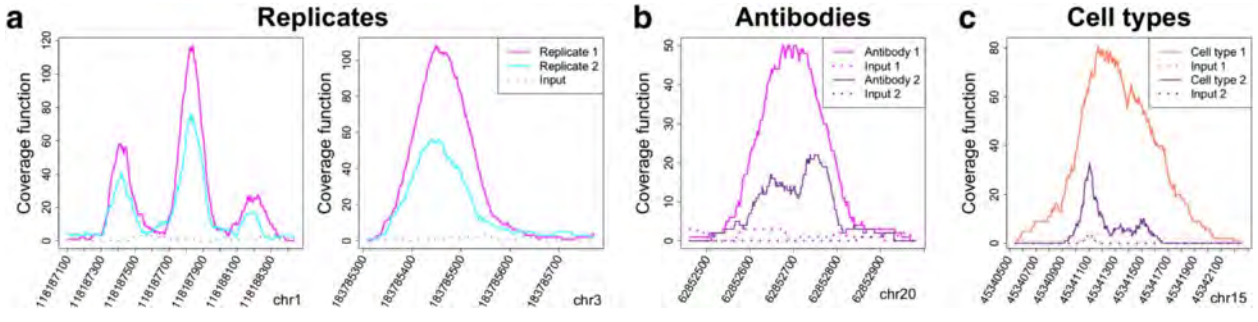

Fig. 1 Peaks shapes. **a** Random peaks in two ChIP-seq replicates - in magenta and cyan, respectively - for the transcription factor GATA-1 in K562 cells. Despite the different efficiency of the two replicates, peak shape is highly maintained. **b** Peaks in two ChIP-seq experiments in K562 cells, performed with two different antibodies for the same transcription factor GATA-1. In this situation some peaks can show different shapes. **c** Peaks in two ChIP-seq experiments for the transcription factor GATA1 in K562 and peripheral blood-derived erythroblasts - in purple and orange, respectively. Also in this case peak shape can vary. Some relevant parameters of the ChIP-Seq plotted here can be found in Additional file 1: Table S1 and in Additional file 2

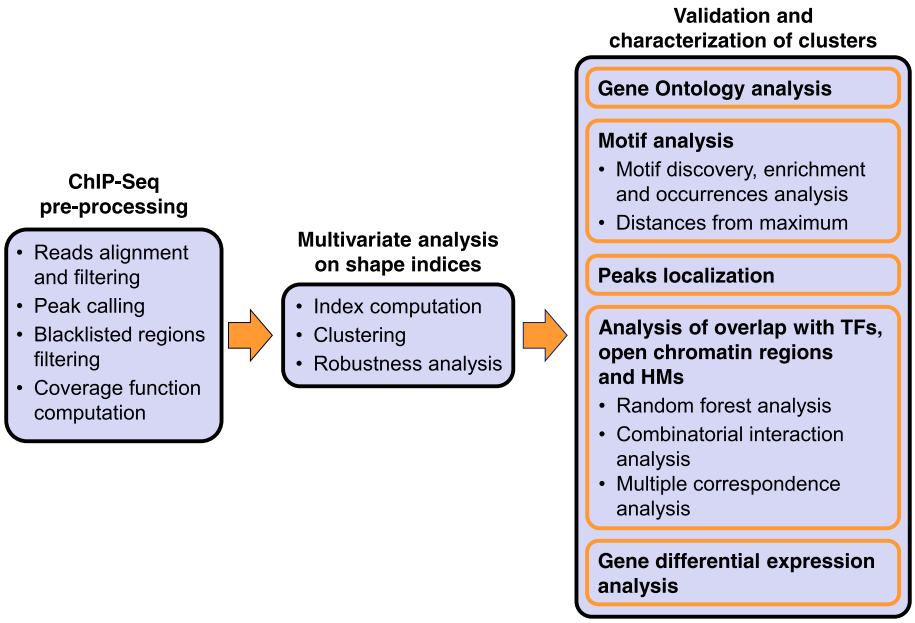

Fig. 2 A schematic overview of the analysis pipeline proposed. After a ChIP-seq pre-processing, that involves the calculation of the coverage function for each peak found by the peak caller, multivariate clustering on five indices of intensity and shape is employed to find groups of similar peaks. Afterwards, the characterization of these clusters is studied by using GO analysis and motif analysis. Moreover, clusters are related with the presence of other proteins and with gene expression

coverage function and to identify enriched peaks. In this first step, we also estimate the average size of the DNA fragments obtained during sonication. We then use this estimate to extend each tag in order to get the original fragments and to compute the coverage function, counting the number of fragments that fall over each nucleotide. The correct estimation of the fragment length is essential since, as we have previously observed, peak shape can vary based on this estimation. Next, we calculate five indices of shape: the *maximum height*, the *area*, the *full width at half maximum*, the *number of local peaks*, and the *shape index M* divided by the maximum height (Fig. 3). Afterwards, we cluster peaks in the space of these resulting shape indices. We name this central part of our method *Shape Index Clustering* [11]. Finally, the obtained clusters are validated and characterized using four steps: 1) we perform Gene Ontology analysis and motif analysis; 2) we investigate the genomic locations of the peaks; 3) we study the overlap of the peaks in each cluster with peaks of other available transcription factors and histone modifications, as well as with open chromatin regions; 4) we evaluate gene expression changes in association with the shape clustering. A detailed description of each step in the pipeline proposed is given in Methods.

GATA-1 in K562 cells

We apply the proposed analysis pipeline to ChIP-seqs for the erythroid transcription factor GATA-1 in human erythroleukemic K562 cells. The purpose of this study is to assess whether GATA-1 peak shape is associated with specific regulatory complexes and functions. GATA-1 is a transcription factor essential for erythroid and megakaryocytic development, and mutations in GATA-1 are associated with a form of leukemia found in newborns affected by Down syndrome. We select K562 cells because GATA-1 binding has been extensively characterized in this cell line [13–15] and also because K562 cell line has been widely described by several Next Generation Sequencing experiments from the Encyclopedia of DNA Elements (ENCODE) Consortium [16, 17] and from many independent investigators.

Two ChIP-seq replicates for GATA-1 in K562 human cells

The experiments under consideration consist of two ChIP-seq replicates for GATA-1 from ENCODE [16] (GEO Accession number GSM1003608, antibody used: sc-266, Santa Cruz Biotech). The signal from a normal Mouse IgG ChIP-seq (GEO Accession number GSM935631) is used as control for peak calling. Peaks are called using MACS [18]. While the number of reads after filtering is comparable and the estimated fragment length is exactly the same in the two replicates, the number of identified peaks is different: we identify 13159 peaks in Replicate 1 and 5509 peaks in Replicate 2, with 5334 overlapping peaks (Additional file 1: Table S1). Almost all the regions selected in Replicate 2 are enriched in

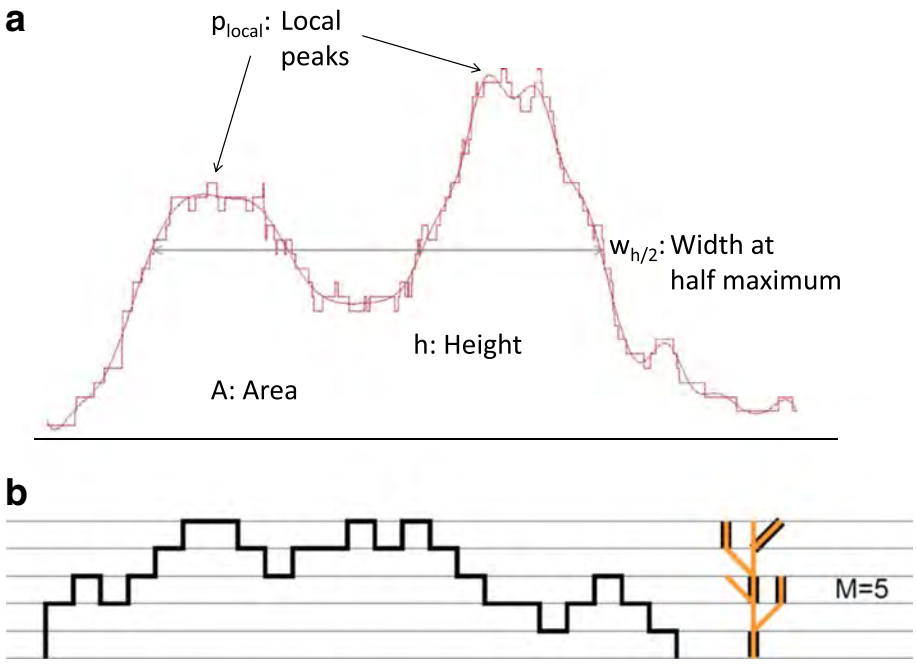

Fig. 3 Shape indices. **a** Schema of the first four shape indices. **b** A peak and its corresponding tree, constructed as suggested in [35] and in [6]; the highlighted edges represent a maximal matching for the tree, that defines the index M

Replicate 1, meaning that the second replicate is much less efficient than the first one [12]. In Fig. 1a, we show the coverage function of two random overlapping peaks - in cyan and magenta for the two replicates, respectively. Despite the different degree of efficiency of the two replicates, pairs of peaks exhibit the same shape. Notably, the whole coverage function has a similar shape in the two replicates, carrying a correlation of ~0.77 on the entire genome and a correlation of 0.95 on the common peaks (Additional file 1: Figure S1).

Clustering of shape indices leads to three clusters

We use the statistical analysis described in Methods to assess whether there are groups of peaks inside a single ChIP-seq that can be separated according to the shape, as summarized by the five selected indices. Here, we present the results obtained by running the analysis on Replicate 1. Results concerning Replicate 2 are highly similar, despite the remarkable differences of the two ChIP-seqs, and can be found in Additional file 1: Figures S4-S6. Notably, if we merge the reads of the two replicates and then we perform the analysis, we obtain highly similar results too. From the scatterplot of the shape indices (Additional file 1: Figure S2b), from principal component analysis (Additional file 1: Figure S3) and independent component analysis (Fig. 4f-g), it is clear that the five indices are not mutually independent. Actually, the two indices related to the intensity of the signal, namely the *maximum height* and the *area* are highly correlated (correlation coefficient of 0.92); the same applies to the three indices associated with peak complexity, i.e. the *full width at half maximum*, the *number of local peaks* and the *shape index M* with correlation coefficients of 0.68, 0.74 and 0.84, respectively. However, we choose not to reduce the dimensionality of the problem in order to keep all the shape variability that we can catch with the selected indices.

Running the k-mean algorithm on the standardized indices for several numbers of clusters k, we obtain the total within-clusters sum of squares plot of Additional file 1: Figure S2a: $k = 3$ seems a sensible trade-off, since the choice of a higher k is not paid off by a significant gain in the total within-clusters sum of squares. This choice leads to identify a big cluster, that comprises ~75 % of peaks (Cluster 1), and two smaller clusters including ~15 % (Cluster 2) and ~10 % (Cluster 3) of the data, respectively. According to the scatterplot of Additional file 1: Figure S2b, wherein colors indicate cluster membership, and to the boxplots of Fig. 4a-e, that display the distribution of the indices in the three clusters, Cluster 1 and Cluster 2 differ in the intensity of the peaks they contains, while Cluster 3 includes the most complex peaks. Similarly, independent component analysis (Fig. 4f-g) shows that the three clusters are well divided in the plane defined by the first two components. In particular, the component that corresponds to

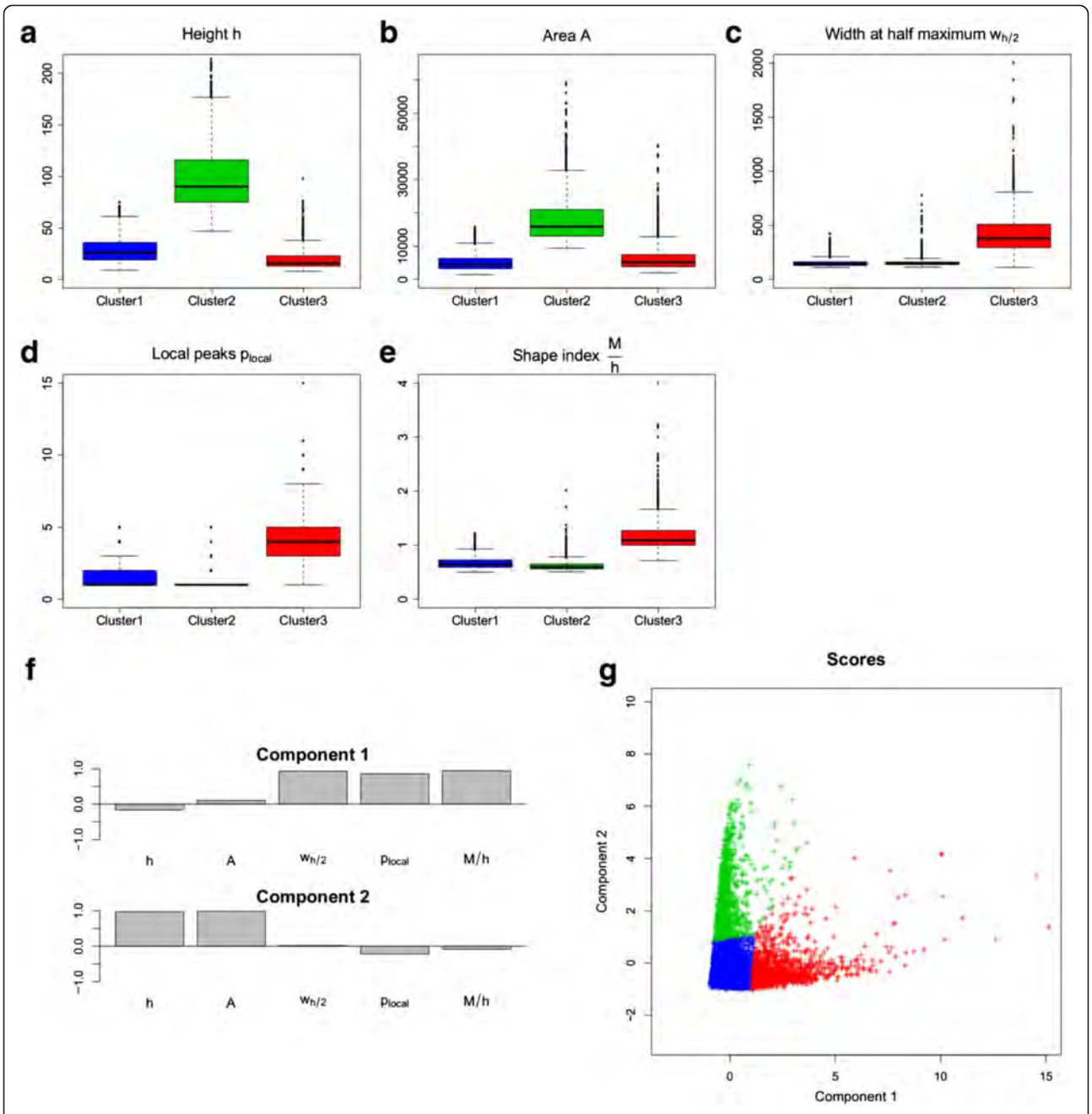

Fig. 4 Distribution of shape indices in the three clusters. Results of *k*-mean algorithm with Euclidean distance on the standardized shape indices, in Replicate 1 for K562 cells. **a**-**e** Boxplots of shape indices in the three clusters. **f** The first two components obtained with independent component analysis, in term of the initial shape indices. **g** Scatterplot of the data in the plane defined by the first two independent components, with each point representing a peak and colored according to the cluster it belongs to

peak intensity (component 2) separates Cluster 2 from the others, while the component related to the complexity of the peaks (component 1) distinguishes Cluster 3. In order to better understand the shapes selected through clustering, Fig. 5 displays the pointwise boxplots of the peak coverage function in the different clusters (top panels) and the plot of a random sample of 200 peaks (bottom panels, for visualization reasons we do not draw all the peaks simultaneously). Cluster 1 is mainly composed of unimodal and not very high peaks, while Cluster 2 comprises high, bell-shaped peaks; multimodal and wider peaks belong to Cluster 3. Interestingly, peaks of Cluster 3 are not those selected with low score (less enriched peaks) by MACS: if we further reduce the threshold in peak detection, we keep on picking a considerable subset of these peaks. Hence, the groups of peaks

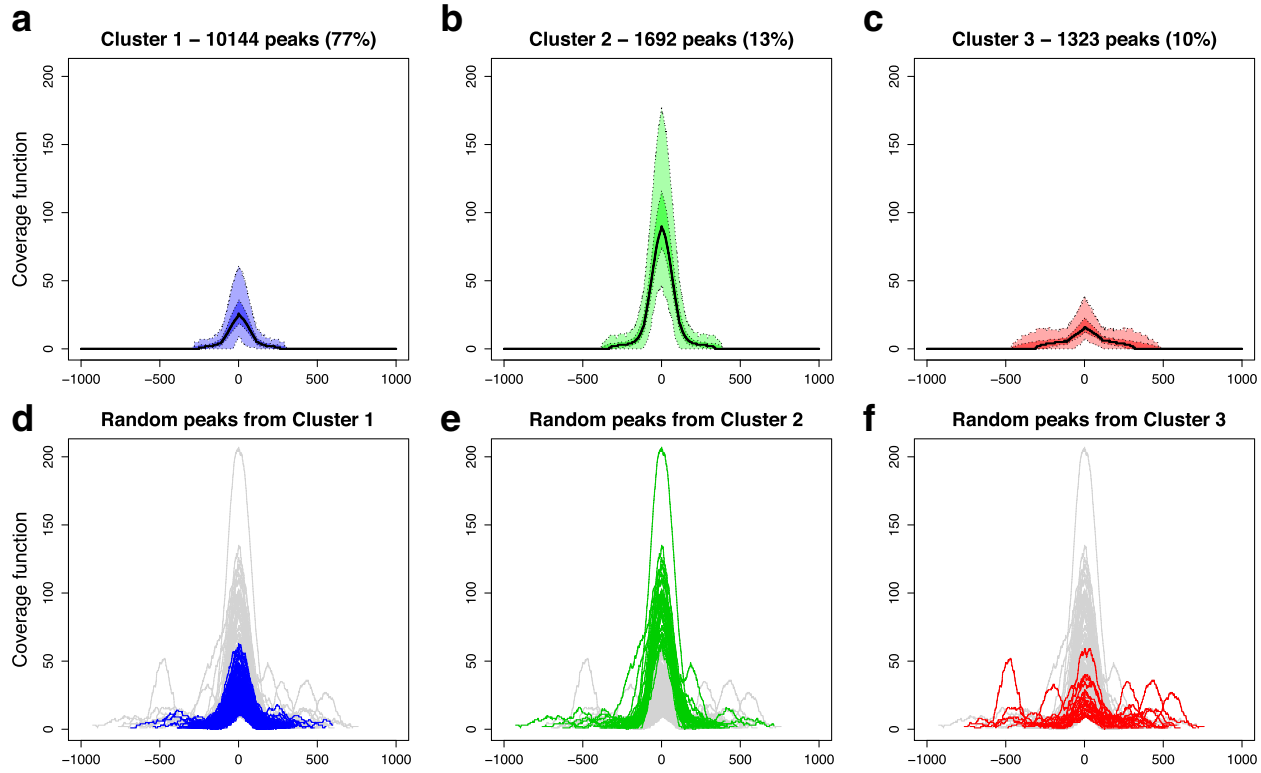

Fig. 5 The three clusters obtained on shape indices. Results of *k*-mean algorithm with Euclidean distance on the standardized shape indices, in Replicate 1 for K562 cells. **a**-**c** Pointwise boxplots of the coverage function in the three clusters. For each abscissa, *black* indicates the median value, *dark colors* highlight the central 50 % of the distribution, while *light colors* correspond to the boxplot whiskers. **d**-**f** A random sample of 200 peaks (for visualization reason not all peaks are plotted), with colors highlighting the cluster membership. In both images, peaks are registered using as landmark the location of their maximum height

obtained by our clustering of shape indices cannot be deduced by MACS output.

The comparison of the resulting classification for the two replicates, done on the common peaks, supports the robustness of the considered method. Indeed, nearly 90 % of peak pairs fall in the correspondent cluster of the two replicates. Moreover, only 18 pairs are misclassified between the two extreme shapes, namely between Cluster 2 and Cluster 3. In Additional file 1: Figure S7, we show the correspondences between cluster memberships of peaks in the two replicates. This cross-replicate robustness analysis and the relationship between the two replicates indicate that it is sufficient to consider the most efficient replicate for the evaluation and characterization of the clusters. Thus, in the following analyses, we show only results concerning Replicate 1.

Only Cluster 2 is directly associated with the typical biological processes of GATA-1

We use the genomic regions enrichment of annotations tool (GREAT) to perform Gene Ontology (GO) analysis of the three clusters. GO analysis reveals that the terms related to the typical biological processes of GATA-1 (such as erythrocyte differentiation, erythrocyte homeostasis and myeloid cell differentiation) are enriched exclusively in Cluster 2 (Additional file 1: Table S2; see Additional file 1: Table S3 for the entire list of significantly enriched GO Biological Process terms in the different groups, and Additional file 3 for the complete list of terms). Considering the complete list of terms given by GREAT, it is evident that Cluster 2 is made up of few genes and many of these genes are key hematopoietic transcription factors: GATA1, FOG, and TAL1. On the other hand, Cluster 1 also contains genes that are typically regulated by GATA-1 (i.e. GATA1, GATA-2, FOG, and RUNX1), but contains also other genes associated with secondary functions of GATA-1. Indeed we can identify many genes (e.g. BAX, BAD, CASP10, BCL10, MADD, LTA, BMF) that are related to apoptosis. Interestingly, these genes are nearly absent in Cluster 2 and present at a lower extent in Cluster 3 (e.g. BAX, CASP9, BCL2L1). GATA-1 is known to inhibit apoptosis while promoting differentiation in erythroid and megakaryocytic cells [19, 20].

Peaks of Cluster 3 contain less GATA-1 motifs

GATA-1 has been shown to recognize the consensus sequence [AT]GATAA [21, 22], so we might anticipate to find this binding motif under the vast majority of peaks, whatever cluster they belong to. This expectation is only partly fulfilled. Although GATA-1 motif is found in all cases, the significance of the enrichment is different in the three clusters (see Table 1 and Additional file 1: Table S4). Surprisingly, E-values obtained with Cluster 1 and 2 are comparable to the global one. On the contrary, GATA-1 motif is less present in the peaks of Cluster 3. Only 79 % of regions belonging to Cluster 3 contain the consensus sequence [AT]GATAA, while the motif is found in almost all peaks of Cluster 2 and in 91 % of Cluster 1 peaks (see Table 1). Furthermore, peaks in Cluster 2 tend to be associated with multiple GATA-1 motifs (see Fig. 6a). Since other members of the GATA family zinc finger proteins, namely GATA-2 and GATA-3, can bind with high affinity the same motif of GATA-1 [23], the presence of more than one motif under a peak can indicate both a multiple GATA-1 binding and the simultaneous presence of several transcription factors of the GATA family. In Fig. 6b, we show the distribution of motif distance from the peak maximum. In addition of being less associated with GATA-1 motif, Cluster 3 regions exhibit a higher distance of the found motifs from peak maxima: when the motif is present, it is usually not centered near the maximum.

Apart from GATA-1 consensus sequence, many other motifs are enriched in the three clusters (see Additional file 1: Table S4). Interestingly, these additional motifs are peculiar to the different clusters, suggesting distinctive types of gene regulation. All the three clusters are enriched for Ets motifs, including PU.1, GABPA, and FLI1 motifs, in accordance with what is shown by previous data on GATA-1 [24, 25]. Cluster 1 and 2, in contrast to Cluster 3, are enriched for TAL-1 and KLF1 motifs. Interestingly, TAL1 motif is usually enriched at GATA-1 activated genes [26]. In addition, Cluster 1 is also enriched by motifs corresponding to FOXO3, SOX7 and SRF and TEAD1, genes that are involved in regulation of apoptosis, in accordance with the functional enrichment of genes present in this cluster.

Table 1 GATA-1 motif analysis

	E-value	Peaks with motif
		
Global	3e-216 (3e-158)	90 % (11874)
Cluster 1	7e-258 (3e-159)	91 % (9187)
Cluster 2	2e-250 (1e-194)	97 % (1638)
Cluster 3	7e-21 (2e-88)	79 % (1049)

GATA-1 motif enrichment and occurrences analysis in the complete set of peaks as well as in the three clusters, in Replicate 1 for K562 cells. For the enrichment analysis, MEME-ChIP with default options is run on samples of 1323 peaks (the size of the smallest cluster) to get comparable E-values. In the global set and in Cluster 1, that are large with respect to the sample size, the median of E-values obtained by random sampling 10 times is reported. In parenthesis we indicate the E-values obtained running MEME-ChIP changing default parameters to allow the motif search in the whole peak regions (without trimming them). The number of peaks with the motif (occurrences analysis) is computed on the whole sets of peaks

Peaks of Cluster 3 frequently lie in promoter regions

Studying the association of GATA-1 peaks with known genes, we discover that about 55 % of peaks (7271 regions) are assigned to at least one gene and this percentage is similar in the three clusters (54 % in Cluster 1, 57 % in Cluster 2 and 59 % in Cluster 3). The clusters behave in the same way even considering the proportion of peaks associated to non-coding genes, as it is the same in all clusters (around 2 %). Interestingly, when we examine more deeply the genomic locations of the peaks, Cluster 3 stands out from the others because of its significantly higher association with promoters (~30 % of Cluster 3 compared to ~15 % of Cluster 1 and 2), defined as the regions within 5 kb upstream and downstream transcription start sites (Fig. 7). Testing the hypothesis that the proportion of peaks from Cluster 3 located in promoters is greater than the same proportion for peaks from Cluster 1 gives p-value = 0; we get the same p-value = 0 also considering Cluster 2, hence the association of Cluster 3 with promoter regions is statistically significant. Less and more restrictive definitions of promotor regions show the same association for Cluster 3 (Additional file 1: Figure S8).

Cluster 2 is associated with a putative protein complex

To investigate the simultaneous binding of GATA-1 with other proteins, we consider a set of 237 publicly available ChIP-seq experiments for 95 different transcription factors, as well as 38 histone modification ChIP-seqs in K562 cells (from [17], see Additional file 2 for the detailed list of used datasets). Transcription factors ChIP-seq replicates are kept separated and MACS is used to call peaks, independently in each sample, using as control the same signal used with GATA-1 ChIP-seqs. In the case of histone modifications, we use peaks called by ENCODE. In addition, we also consider DNase I hypersensitive sites in K562 cells (GEO accession number GSM816655) in order to study open chromatin regions.

We use random forest classification, as explained in Methods, to select the experiments that are more correlated with our clusterization. Specifically, seven different analyses are performed, by using as response the clusters membership and alternatively classifying: 1) all clusters; 2) one cluster versus the union of the other two; 3) two clusters one against the other. All random forest models are

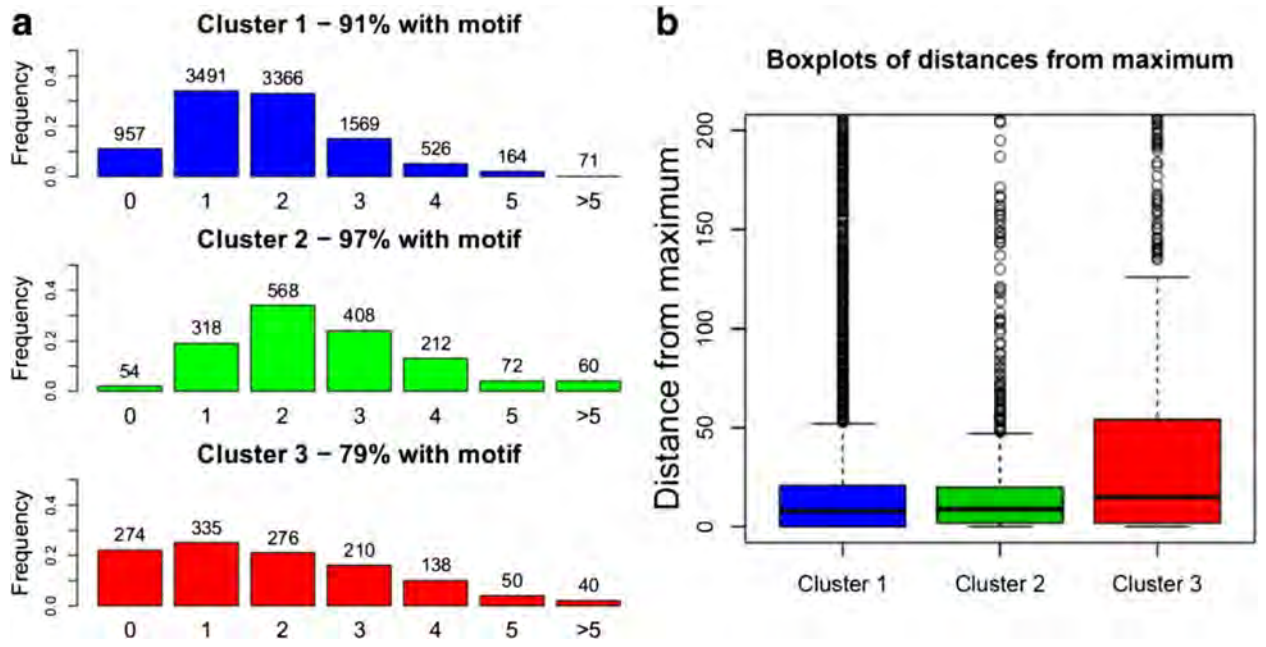

Fig. 6 Occurrences analysis for GATA-1 motif, in Replicate 1 for K562 cells. **a** Distribution of the number of occurrences of the consensus sequence [AT]GATAA under each GATA-1 peak: under many peaks of Cluster 2 the motif occurs multiple times. **b** Boxplots of motif distances from peak maxima in the three clusters; the mean distance in Cluster 3 peaks is significantly greater than the mean distances in Cluster 1 and Cluster 2 regions (test *p*-values ~0)

able to predict a considerable portion of memberships, indicating that our peak shape clustering is related to co-located proteins. Notably, the overlaps of GATA-1 peaks with open chromatin regions stand out as important in all these analyses, according to Gini index (Additional file 2). Looking more deeply to the relationship between clusters and DNase-Seq regions, we can observe that the proportion of peaks that fall in accessible regions of the genome is typical of the different clusters. Indeed, almost all Cluster 2 peaks intersect DNase I hypersensitive sites (94 %), while the portion of peaks that fall in open chromatin regions is smaller in Cluster 3 (84 %) and it is further reduced in Cluster 1 (70 %). Anyway, the three proportions are much higher with respect to the random case (see Methods), in which only the 8 % of the peaks intersect open chromatin regions, supporting the claim that none of the clusters is composed exclusively by artifacts. Interestingly, the distribution of the percentage of intersection, conditionally to the intersection being non-zero, is essentially the same in all groups, and it is highly similar to the random case. Hence, the clusters are characterized by the proportion of peaks that overlap DNase I sites, rather than by the percentages of intersection (see Additional file 1: Table S5 and Figure S9). Inspecting the rankings of transcription factors and histone modifications ChIP-seqs, according to Gini index, we are able to identify a small set of regulatory elements that emerge as influent in the seven random forest classifiers we built, with all replicates in top positions (the complete rankings are reported in the Additional file 2). Specifically, we retain for further analyses all the proteins that are, simultaneously, 1) top 15 in at least three random forests; 2) top 30 in at least five random forests; 3) top 30 with all available replicates in at least 2 random forests. The eight transcription factors selected are GATA-2, CEBPD, HMGN3, TRIM28, PML, TAL-1, ZMIZ1 and CCNT2 (see Table 2 for the full protein names). Some of these proteins are known GATA-1 interactors. In particular, GATA-2 and TAL-1 can associate with GATA-1 in complex to regulate erythroid transcription. In addition, evidences of interaction between GATA-1 and PML have recently been shown [14]. We must point out that among these important regulatory elements there are no histone modifications. Indeed, if we perform similar random forest analyses using as predictors only the 237 ChIP-seqs for transcription factors, we get nearly the same results and the same accuracy in cluster membership predictions.

The combinatorial interaction analysis on the transcription factors selected by using random forests shows that Cluster 2 is characterized by the simultaneous binding of all the eight transcription factors, in addition to GATA-1. The most significant results are obtained considering the intersections of GATA-1 peaks with at least one ChIP-seq replicate for the other regulatory elements under investigation. Nevertheless, the same conclusions

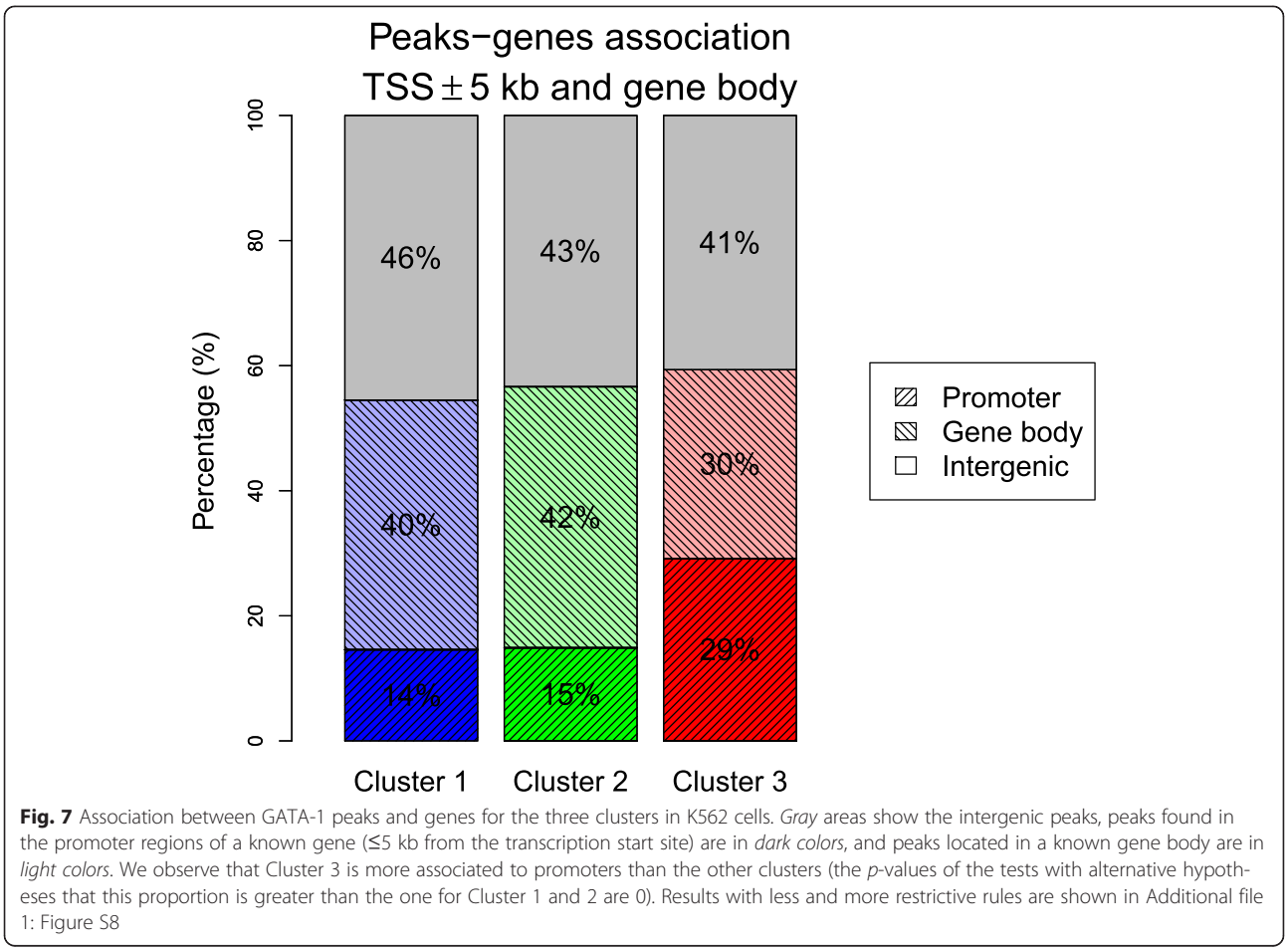

Fig. 7 Association between GATA-1 peaks and genes for the three clusters in K562 cells. *Gray* areas show the intergenic peaks, peaks found in the promoter regions of a known gene (≤5 kb from the transcription start site) are in *dark colors*, and peaks located in a known gene body are in *light colors*. We observe that Cluster 3 is more associated to promoters than the other clusters (the *p*-values of the tests with alternative hypotheses that this proportion is greater than the one for Cluster 1 and 2 are 0). Results with less and more restrictive rules are shown in Additional file 1: Figure S8

are drawn even if we require that GATA-1 peaks overlap all available ChIP-seq replicates for the other proteins. Notably, the distribution of overlaps with a combination of these eight transcription factors in Cluster 1 and Cluster 3 is very similar to the global one, in which no combination outnumbers the others.

Table 2 Transcription factors related to the clustering

GATA-2	GATA binding protein 2
CEBPD	CCAAT/Enhancer-Binding Protein Delta
HMGN3	High Mobility Group Nucleosome-binding domain-containing protein 3
TRIM28	TRIpartite Motif-containing 28
PML	ProMyelocytic Leukemia protein
TAL-1	T-cell Acute Lymphocytic Leukemia protein 1
ZMIZ1	Zinc finger MIZ domain-containing protein 1
CCNT2	Cyclin-T2

The eight transcription factors emerged as relevant for the peak shape clustering in Replicate 1, according to Gini index in all the random forest classifiers built (experiments on K562 cells). Combinatorial interaction analysis reveals that Cluster 2 is characterized by the simultaneous binding of all these eight proteins, together with GATA-1

On the contrary, ~61 % of the peaks belonging to Cluster 2 simultaneously intersect all the eight proteins considered, while only 4 regions contain GATA-1 alone (see Fig. 8 and Additional file 1: Figure S10).

The co-binding of these eight transcription factors in the genomic regions of Cluster 2 emerges also by using multiple correspondence analysis (see Methods). The advantage of this analysis is that it permits to study all the replicates simultaneously. In particular, by plotting the amount of total variation explained by an increasing number of principal coordinates (Additional file 1: Figures S11 and S12), we choose to focus on the first two components. The main effect of the first component (that explains, alone, ~40 % of the total variation in the data) is to contrast between the presence and the absence of overlaps, while the second dimension adds some variability among the different regulatory elements considered (Additional file 1: Figure S12). We observe that the various ChIP-seq replicates for the same protein are close in the space of the first two principal components, suggesting that they behave in a very similar way. Notably, Cluster 2 appears as highly different from the other two groups and the global case.