

RESEARCH ARTICLE

Open Access

Peak shape clustering reveals biological insights



Marzia A. Cremona¹, Laura M. Sangalli¹, Simone Vantini¹, Gaetano I. Dellino^{2,3}, Pier Giuseppe Pelicci^{2,3}, Piercesare Secchi¹ 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

⁴Center 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

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

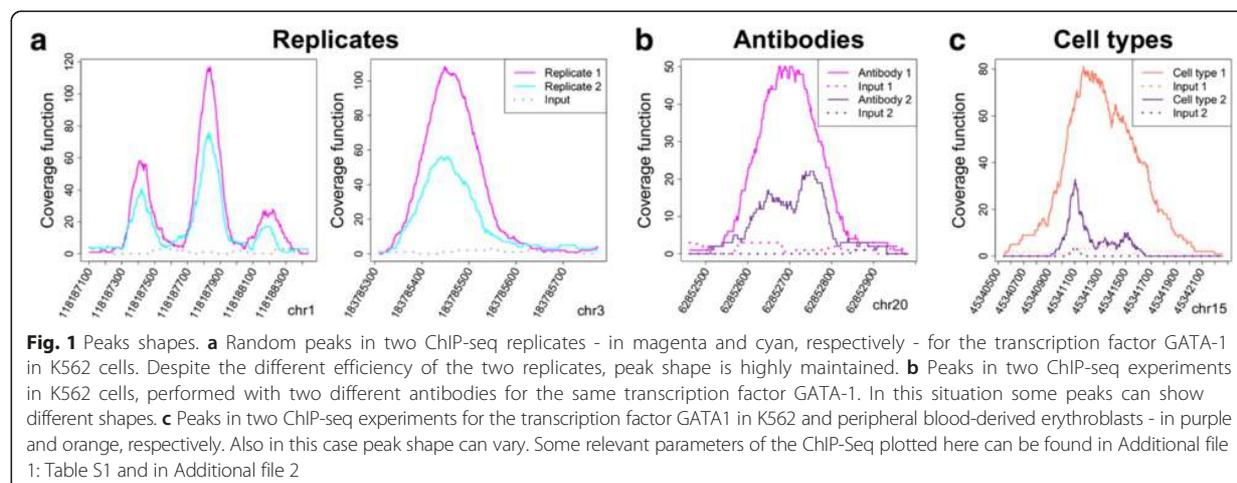
Peak shape varies among different experiments

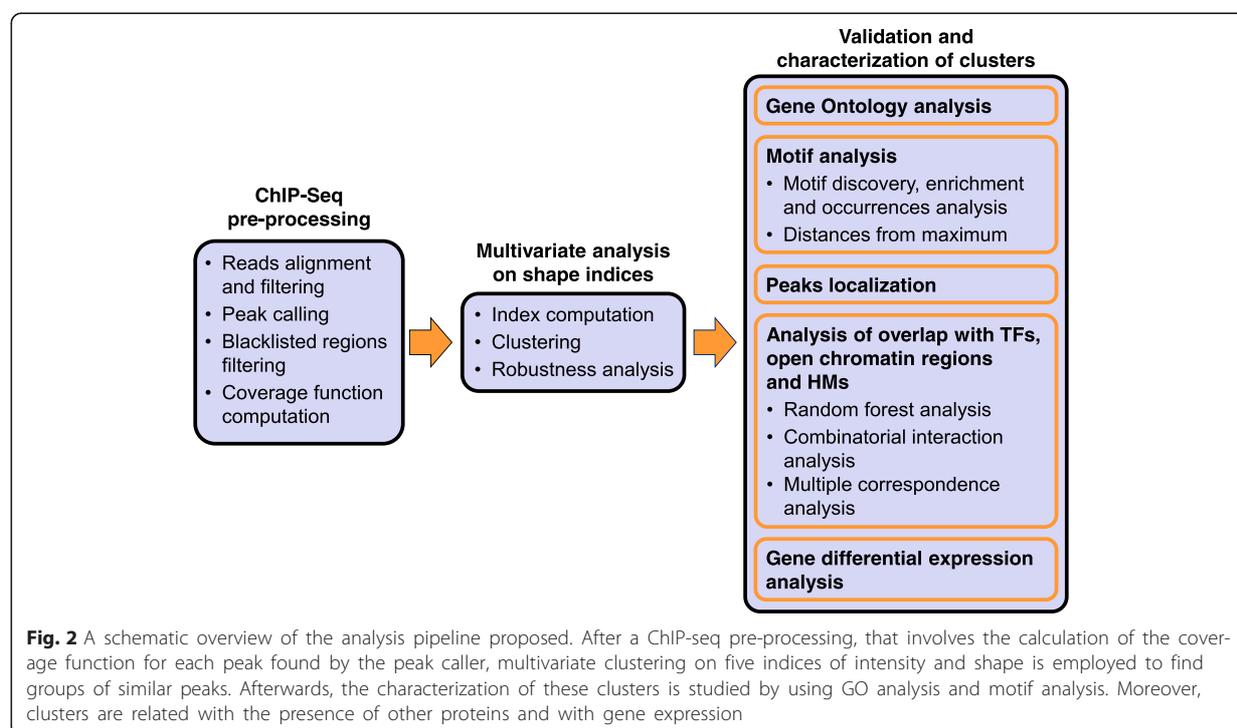
We observe that peak shape is quite reproducible as technical and biological replicates obtained with the same library preparation protocols (see Additional file 1: Table S1 and Additional file 2) give rise to the same signal in the same genomic region (Fig. 1a). This is true even if ChIP-seq efficiencies 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





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.

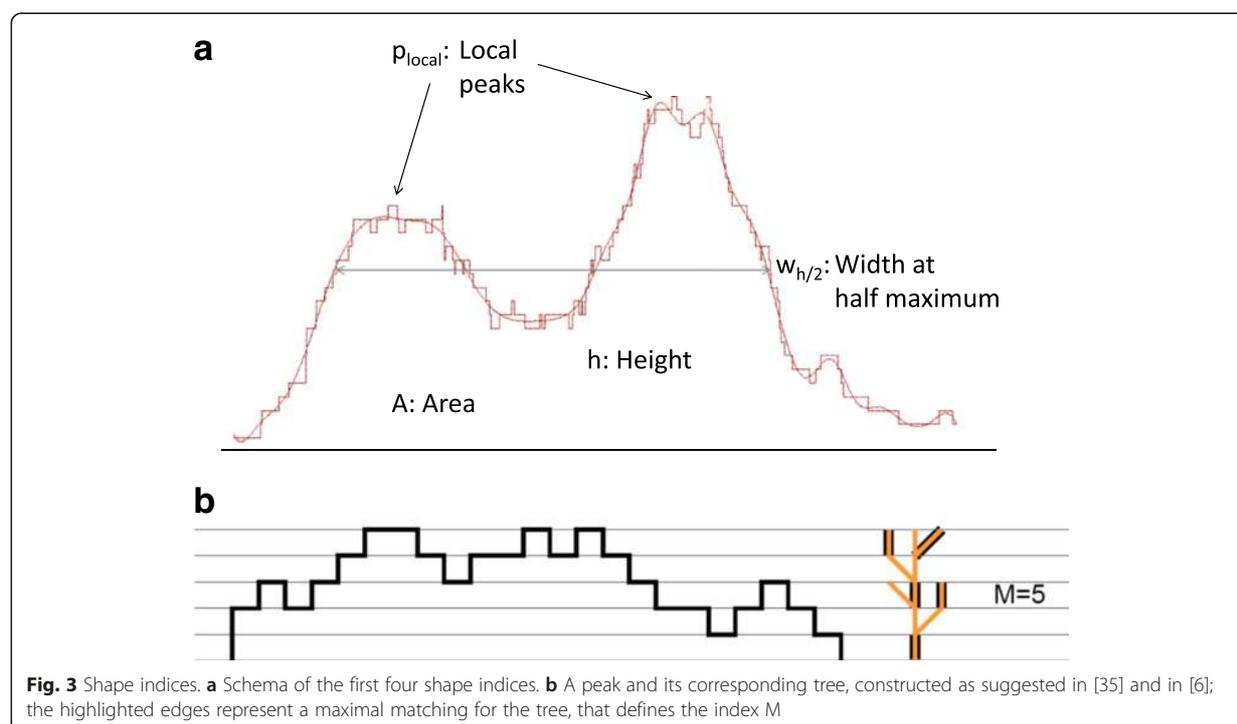
GATA-1 in K562 cells

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



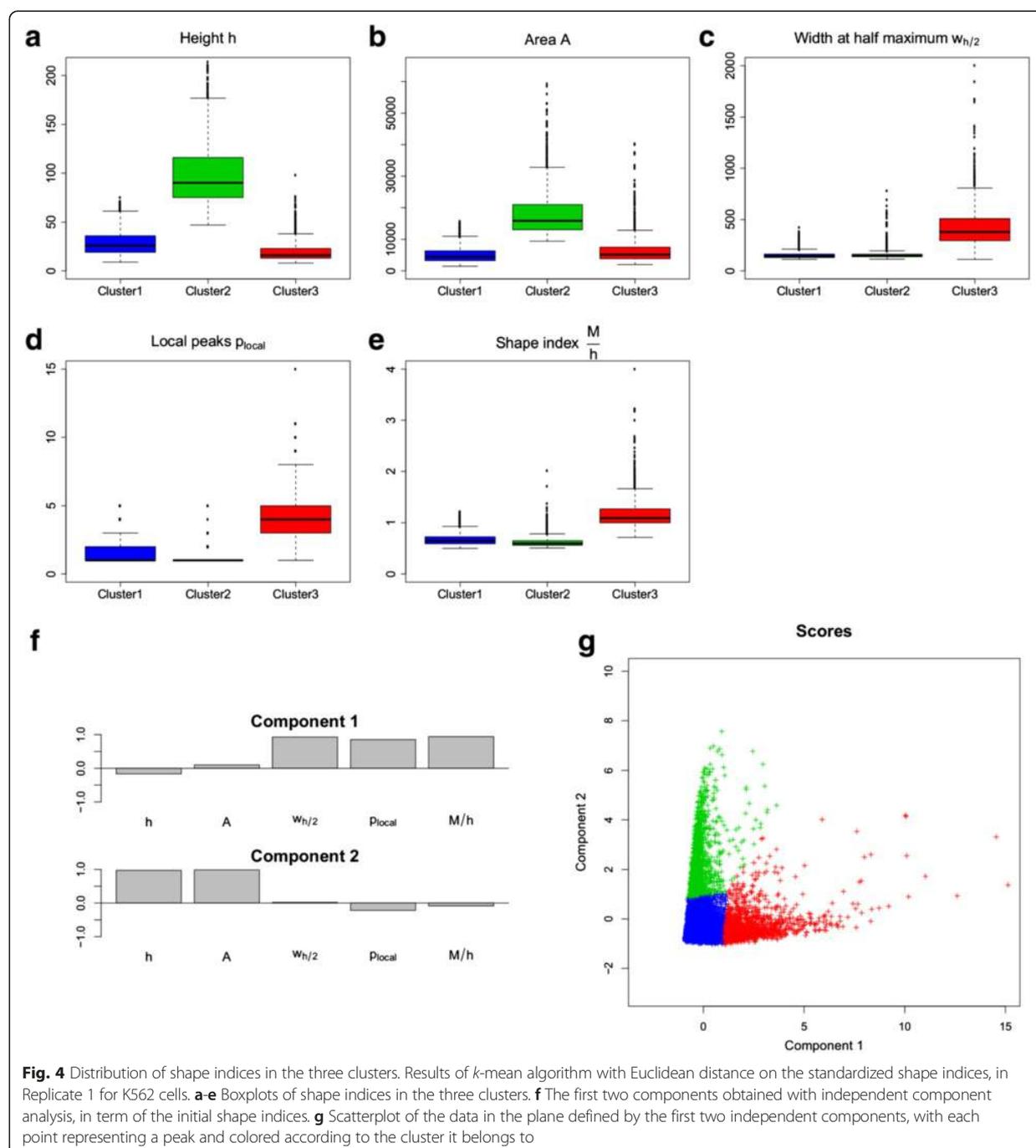
Replicate 1, meaning that the second replicate is much less efficient than the first one [12]. In Fig. 1a, we show the coverage function of two random overlapping peaks - in cyan and magenta for the two replicates, respectively. Despite the different degree of efficiency of the two replicates, pairs 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,

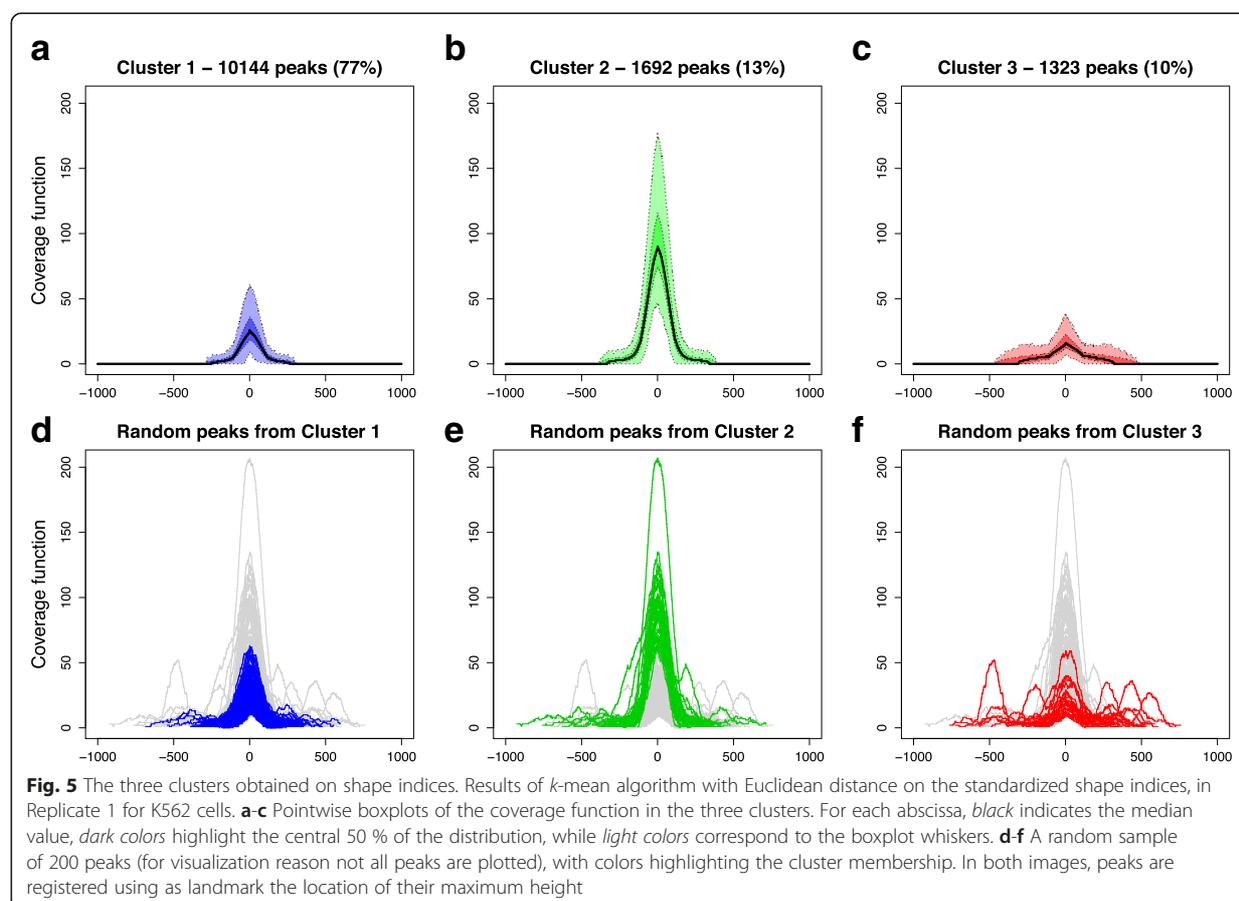
namely the *maximum height* and the *area* are highly correlated (correlation coefficient of 0.92); the same applies to the three indices associated with peak complexity, i.e. the *full width at half maximum*, the *number of local peaks* and the *shape index M* with correlation coefficients of 0.68, 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 $\sim 75\%$ of peaks (Cluster 1), and two smaller clusters including $\sim 15\%$ (Cluster 2) and $\sim 10\%$ (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



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



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].