Methods. Briefly, BWA is used to map high quality reads to the human reference genome hg19. Afterward, MACS is employed to call peaks, obtaining 3399 GATA-1 peaks (after blacklisted regions filtering). The estimated fragments length of 57 nucleotides is then used to compute the coverage function for each peak.

Similarly to what we obtained with GATA-1 in K562, also in megakaryocytes the multivariate analysis on shape indices leads to three clusters (see the total within-clusters sum of squares plot in Additional file 1: Figure 16a). Notably, examining the nature of the resulting clusters, we conclude that the selected shapes have the same characteristics of the ones selected within K562 peaks, with the only difference being the proportions of peaks in the different clusters. Clustering results are shown in Additional file 1: Figures S16b-S17 and Fig. 10 (that report the scatterplot of indices, their distributions and the coverage function in the different groups), where for the sake of clarity the clusters are named and colored as their matching pair in K562. In summary, Cluster 1 includes the majority of the peaks (~66 %), that are bell-shaped and quite low. Cluster 2 is the smallest group (~13 % of the peaks) and is composed by high, bell-shaped peaks. Finally, Cluster 3 comprises the most complex and wide peaks (~21 % of GATA-1 regions).

### Key hematopoietic transcription factors simultaneously bind regions of Cluster 2

To study the genome-wide binding of multiple transcription factors and identify regulatory complexes, Tijssen et al. generated ChIP-seqs for the hematopoietic transcription factors GATA-1, FLI1, GATA-2, RUNX1 and TAL-1. By analyzing all the regions bound by at least one of these proteins, the authors showed that the five transcription factors in megakaryocytes bind the DNA together more frequently than expected at random [24]. To examine the potential relationship between the three peak shape clusters and the binding of GATA-1 with different sets of proteins, we pre-process the four ChIP-seq experiments available (GEO accession number GSE24674) and we use MACS to identify the binding sites of each transcription factor (see Methods). Afterwards, we perform combinatorial interaction analysis on the whole set of GATA-1 peaks, as well as on the three clusters separately. We observe (see Fig. 11) that only Cluster 2 is

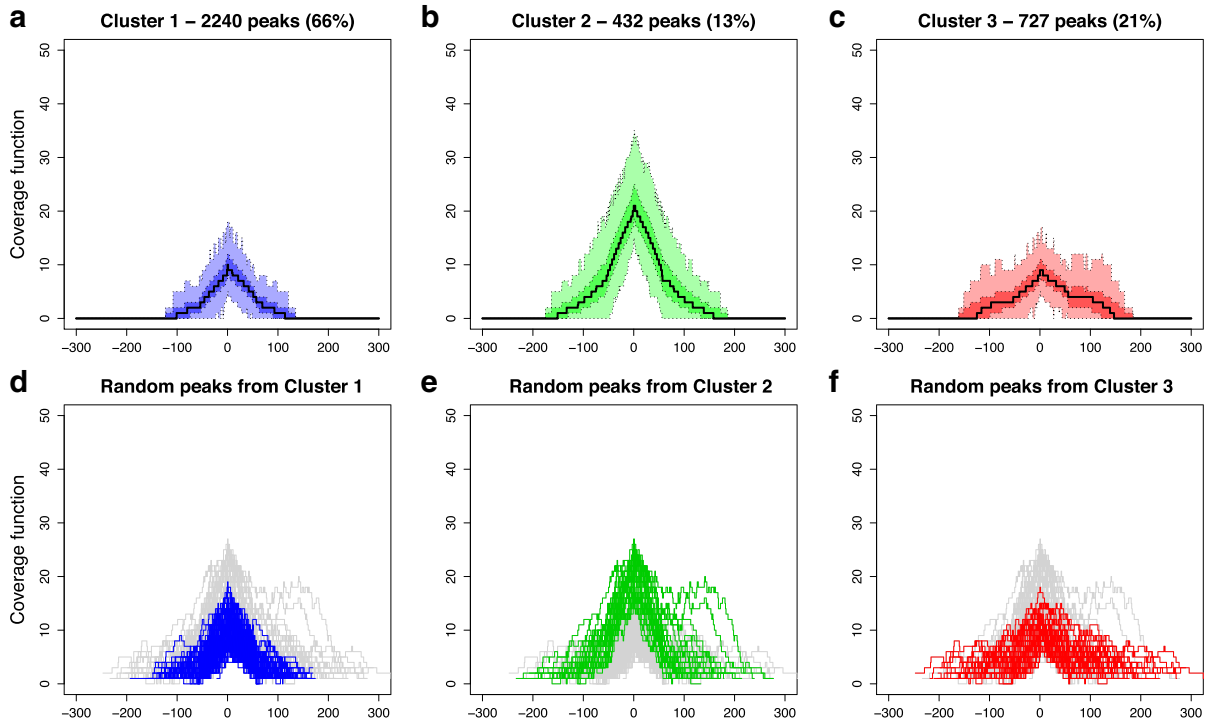


**Fig. 10** The three clusters obtained on shape indices in megakaryocytes. Results of *k*-mean algorithm with Euclidean distance on the standardized shape indices, in megakaryocytes. **a**-**c** Pointwise boxplots of the coverage function in the three clusters. For each abscissa, *black* indicates the 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), with colors highlighting the cluster membership. In both images, peaks are registered using as landmark the location of their maximum height

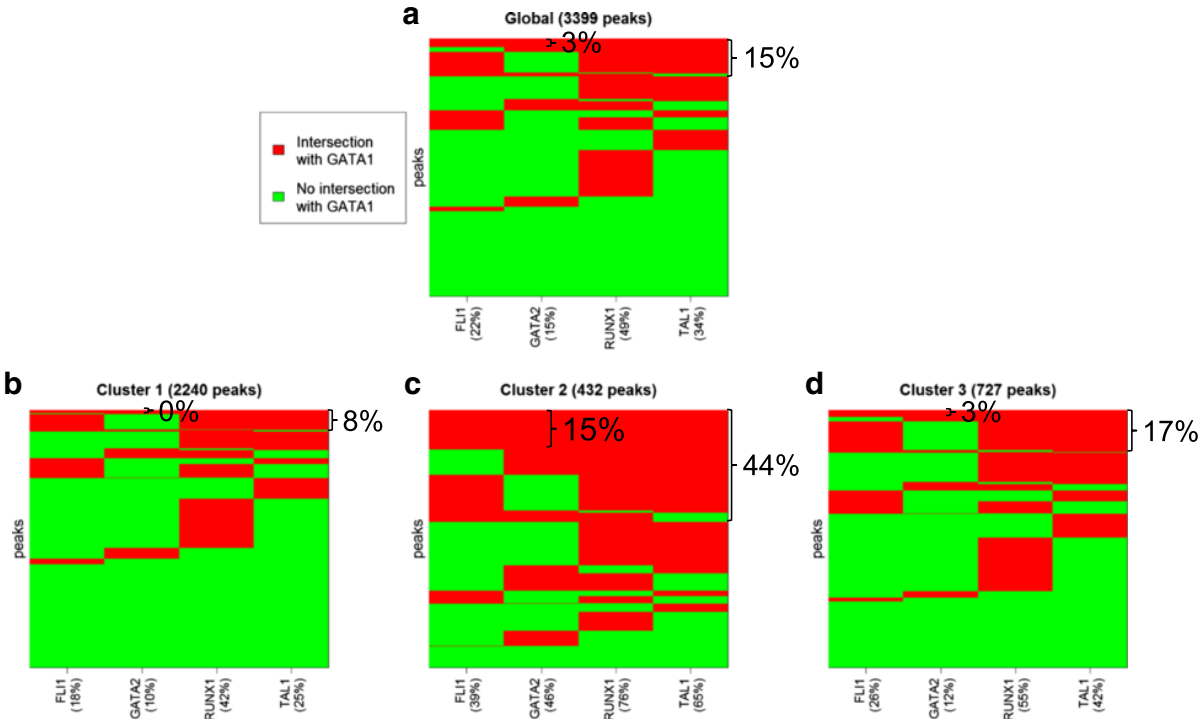


**Fig. 11** Combinatorial interaction analysis on the five hematopoietic transcription factors studied in human megakaryocytes. **a** Results considering all the peaks simultaneously. **b**-**d** Results concerning the peaks that belong to three clusters separately. In all the plots, each row represents the overlap of a GATA-1 peak with the eight protein considered. The *red* color means that there is an intersection, while *green* stands for the absence of the protein in the correspondent GATA-1 region

characterized by the simultaneous presence of all the five proteins. Notably, about 15 % of the peaks belonging to Cluster 2 overlap all five transcription factors together, while this co-binding is present in only ~1 % and ~3 % of peaks from Cluster 1 and Cluster 3, respectively. Moreover, at least three proteins among FLI1, GATA-2, RUNX1 and TAL-1 simultaneously bind only ~15 % of GATA-1 sites, but this combination is present in ~44 % of peaks in Cluster 2. These findings suggest that we can actually use GATA-1 peak shape to highlight the co-localization preference of the five hematopoietic transcription factors in megakaryocytes. Furthermore, 5 of the 6 genes whose depletion resulted in a severe hematological phenotype in zebrafish (i.e. MARCH2, MAX SMOX, EMILIN1 and SUFU), that were identified by the co-localization of the five TFs in [24], are part of Cluster 2. Thus, peak shape conveys biologically relevant findings.

### GATA-1 peaks in mouse and human megakaryocytes have similar shapes

We apply our analysis pipeline to an ENCODE ChIP-seq for GATA-1 in mouse megakaryocytes (GEO Accession number GSM923586 Replicate 1, antibody used: sc-265, Santa Cruz Biotech) [16]. We start our preprocessing on the reads already aligned to mm9 reference genome, then we call peaks using MACS as detailed in Methods. After filtering our blacklisted regions we obtain 2586 peaks and an estimated fragments length of 52 nucleotides, comparable with the length obtained in human megakaryocytes. The clustering of shape indices produces four clusters in this case. However, one cluster is made of only 13 peaks, and 10 out of 13 lie in repetitive regions (they have more than 80 % overlap with RepeatMasker track), thus they are false positive peaks. Importantly, these 13 peaks show extreme shapes (they are quite high, wide, multimodal and really complex peaks) and our clustering methodology is able to recognize and group them together. After excluding this artifact cluster from further analyses, GATA-1 in mouse and human megakaryocytes have the same number of clusters, with very similar shapes and indices distributions. The sole difference we can spot is that mouse Cluster 2 peaks are a bit more complex than human Cluster 2 peaks. This result suggests that peak shape might be the highly similar in the same cell types of different organisms.

### Clustering of shape indices for other ChIP-seq experiments

We examine a set of ChIP-seqs for nine different proteins in K562 cells (see Additional file 4 for details), to understand whether the clustering of shape indices always give rise to the same number of clusters and to the same shapes that are found in GATA-1. All the analyses are performed using the same methodology described in Methods. Interestingly, we observe that the number of clusters varies depending on the protein under consideration. In particular, we obtain four clusters for TRIM28, CCNT2, Z-MIZ1 and PML, and three clusters for GATA-2, TAL-1, c-Fos, SP1 and NF-YA. Differences in the number of clusters depend on the complexity of the data under analysis. Indeed, in all the experiments that give rise to four clusters, one of them is quite small (less than 6 % of peaks belong to it) and contains broad and complex peaks; we exclude that these peaks are alignment artifacts (as we observe for GATA-1 in mouse megakaryocytes) because they are not located in repetitive regions and they show a central stronger signal. In these cases the peak shapes appear to be generally more complex than the shapes we obtained in GATA-1 (Additional file 1: Figures S18-S20 show the representative example of CCNT2 peak shape clustering). Moreover, different TFs have various rates of peaks in the different clusters. Among the TFs with three clusters, we observe that some proteins (such as GATA-2, c-Fos and SP1) produce almost the same clustering of GATA-1, while other TFs (TAL-1 and NF-YA) have slightly more complex shapes, with characteristic rates (see Additional file 4).

## Conclusions

We have developed a novel analysis method that studies ChIP-seq enriched regions focusing both on the complexity and the signal intensity of peaks. Shape Index Clustering for ChIP-seq peaks has the ability to identify different groups of peaks in a single ChIP-seq, based on differences in peak shape. These differences could not be identified using either tag counts or peak enrichment alone. In addition, the proposed pipeline involves several downstream analyses able to investigate possible relationships between the peak shape clusters identified by SIC-ChIP and biological properties.

By applying the proposed analysis pipeline to ChIP-seq experiments for the transcription factor GATA-1 in K562 cells and in primary human megakaryocytes, we have demonstrated that statistically significant different peak shapes are correlated with several cooperative transcriptional regulators. We have shown that GATA-1 peak shape is associated with characteristic regulatory complexes and changes in gene expression profiles. Moreover, peak shape can shed light on previously described GATA-1 occupancy profiles. Specifically, considering GATA-1 ChIP-seqs in K562 cells, peaks belonging to Cluster 2 emerge as part of a putative protein complex that comprises well known GATA-1 interactors such as GATA-2 and TAL-1. The target genes of these peaks appear to be mostly down-regulated after knockdown of GATA-1, suggesting that GATA-1 in Cluster 2 behaves primarily as a transcriptional activator. This result is in agreement with previous studies that reported the ability of GATA-1 to act both as an activator and as a repressor, and that also highlighted a positive correlation between activated GATA-1 target genes and binding of TAL-1 [28, 29]. In conclusion, our study demonstrates that ChIP-seq shapes include information regarding the binding of other proteins beside the one used for precipitation and it is correlated with gene expression. Moreover, studying other ChIP-seq experiments with the same methodology, we showed that peak shape clustering depends on the protein under investigation. Thus, ChIP-seq profiles carry much more information than previously suspected.

Although we presented our methodology applied mainly to ChIP-seq experiments for transcription factors, the same principles can be applied to the investigation of other ChIP-seq data, e.g. histone modification peaks. Furthermore, a generalization of these methods may be employed to analyze ChIP-exo data [30]. We expect that applying the "peak shape concept" to ChIP-exo peaks can lead to even more interesting and clear correlations between shape and biological properties, thanks to the high resolution reached by this technique.

## Methods

### ChIP-seq pre-processing

#### *Reads alignment and filtering*

Mapping reads back on a reference genome is the first pre-processing step that must be done when analyzing ChIP-seq data. We perform it by using Burrows-Wheeler Aligner (BWA) [31], unless we are dealing with ENCODE ChIP-seqs. In this case, ENCODE mapping is taken as the starting point of the analysis pipeline. In all cases the experiments are made up of single-end reads and only high quality tags that maps uniquely to the genome are retained for further analysis. Moreover, only autosomes and X chromosome are considered. Reads duplicates are discarded too.

#### *Peak calling*

The peak caller MACS [18], which is one of the best ChIP-seq callers [32] and optimally estimates the spatial resolutions of binding events [2], is run with the aim of detecting significantly enriched regions in the genome (namely the peaks) with respect to a control signal. Peaks represent the areas where the protein of interest interacts with the DNA. We use MACS default options except for the p-value cutoff for peak detection, that we

set to 1e-8 (more stringent than the default 1e-5). In this step the average length of the initial fragments from which tags are sequenced is estimated from reads positions in the two DNA strands.

#### *Filtering of blacklisted regions*

The ENCODE Data Analysis Consortium Blacklisted regions [16] (a set of artifact regions in the genome), is used to filter the resultant peaks, in order to obtain a purified collection of peaks.

#### *Coverage function computation*

Short reads are extended in 3' direction to the average fragment length estimated before. The coverage function, defined as the base by base count of the elongated reads, is then computed.

### Multivariate analysis on shape indices

#### *Index computation*

Consider the genomic region $R = \{x_1,\ldots,x_L\}$ of $L$ contiguous nucleotide positions found by the peak caller. We define the corresponding peak as the function $f$ that associates to each nucleotide $x_i$ the coverage function, i.e. the count of elongated tags, calculated at that position $f(x_i)$. We summarize the shape of each peak with five indices: the first two are related to the intensity of the signal, while the others are connected with the complexity of the peak (Fig. 3). In particular, for each peak we calculate:

1. The *maximum height* of the peak, i.e. $h = \max_{xi \in R} f(x_i)$;
2. The *area* subtended by the function, i.e. $A = \sum_{xi \in R} f(x_i)$;
3. The *full width at half maximum*, that is the width of the peak (the projection on genome positions) at half of its maximum height, i.e. $w_{h/2} = \max G - \min G$, where $G = \{x_i \in R : f(x_i) \geq h/2\}$;
4. The *number of local peaks* $p_{local}$ of the smoothed function, as detailed below;
5. The *shape index M* (computed as explained below), divided by the maximum height of the peak.

To calculate the number of local peaks $p_{local}$ we need, first of all, to smooth the function $f$ in order to filter out noise (Fig. 3a). A cubic B-splines basis with knots every 20 nucleotides is fitted by using ordinary least squares (see, e.g., [33]). The index is the number of local maxima of the resulting smoothed function, provided that they are at least 50 nucleotides apart and their difference in height from the two nearest local minima is more than the 20 % of the maximum height of the peak.

The shape index $M$ is a measure of the complexity of the peak that is robust to noise, computed as suggested in [6, 34]. Each peak is associated with a rooted tree, built by following the profile of its function $f$. In particular, we start with constructing the root of the tree. Then we look at the value of $f$ at the first nucleotide $x_1$ and we create a new node of depth $f(x_1)$. At this point, for $i \in 2,\ldots,L$, the nucleotide $x_i$ is considered. A new node is created in correspondence to an increase of the function (when $f(x_i) > f(x_{i-1})$). When the function decreases (that is $f(x_i) < f(x_{i-1})$), we move toward the root to the parent of the current node. Finally, if the function keeps constant (we have $f(x_i) = f(x_{i-1})$), nothing is done. An example of a peak and the corresponding tree resulting from this procedure is shown in Fig. 3b. The index $M$ is the number of edges in a maximal matching for the constructed tree, that is the highest number of edges of the tree without common nodes. It is clearly extremely dependent on the height of the tree, that turns out to be the maximum height of the peak. Consequently, we consider the index $M$ divided by the maximum height $h$. The resulting index is related to the complexity of the peak meaning that, height being equal, it is bigger when the peak is multimodal so that the tree has multiple branches. In addition, note that noise in the peak converts to high degree nodes, hence it does not affect the maximal matching for the tree.

#### *Clustering*

We use the $k$-mean algorithm with Euclidean distance on the five standardized indices to cluster the peaks (see, e.g., [35, 36] for details on $k$-means as well as other multivariate clustering techniques). In this step, each replicate is considered separately, with the aim of finding statistically significant differences in peak shape inside a single ChIP-seq. The "correct" number of clusters $k$ is estimated through the analysis of total within-clusters sum of squares plot. Each resulting cluster is characterized by a specific distribution of the five indices, representing its typical shape. This characterization is illustrated by the scatterplot of the indices and by the scatterplots on the first components obtained with PCA and ICA, all colored according to the clustering obtained with $k$-means. Moreover, the different peak shapes in the resulting clusters are shown by boxplots of the shape indices and pointwise boxplots of the coverage function in the different clusters, besides the plots of a random sample of peaks.

#### *Robustness analysis*

We take advantage of the multiple replicates, when they are available, to evaluate the robustness of the proposed technique. First, only peaks that are present in all replicates are selected. The overlapping regions are defined as the contigs of the peaks in all replicates, that is the union of the genomic areas corresponding to peaks with non-empty intersection in different replicates. Each of

these regions $R = \{x_1,\ldots,x_L\}$ is associated with its corresponding peaks, i.e. the coverage functions $f_j$ for each replicate $j = 1,\ldots,J$. Intensity and shape indices are computed on all peaks as explained above. Subsequently, the peaks are clustered independently in each replicate, as illustrated in the last paragraph. Finally, the robustness of the method is evaluated by means of the correspondences between cluster memberships of a peak in the different replicates.

## Validation and characterization of clusters

### *Gene ontology analysis*

Gene Ontology and other annotation ontologies enrichment analysis is done by using GREAT version 2.0.2 [37] with default association rule, for the whole set of ChIP-seq peaks found, as well as for the peaks in each cluster. This study permits to correlate clusters, and consequently peak shape, with an inferred biological meaning.

### *Motif analysis*

Next step consists in different types of motif analysis. De novo motif discovery and motif enrichment analysis are performed with MEME-ChIP version 4.9.1 [38] (with default parameters), to provide a comprehensive view of the sequence motifs under the peaks in each cluster. Furthermore, we also run this tool after sampling the same number of peaks from each cluster, in order to compare the significance of the motifs found in the different groups. Afterwards, we analyze motif occurrences in order to understand how many times the same motif is present under each peak. We also compute, in each cluster, the distribution of motif distance from peak maximum (when more than one motif is found under a single peak, we consider the lowest distance of the motifs from the maximum). This points out whether the motif is central in the peak or it is a side motif.

### *Peaks localization*

To investigate the genomic locations of the different clusters, we annotate each ChIP-seq peak as lying in a promoter, in a gene body or in a intergenic region. We define the promoter region as the area ≤ 2.5 kb, ≤ 5 kb or ≤ 10 kb from a transcription start site, using RefGene annotation database and considering both coding and non-coding genes. Plots and hypothesis tests are used to evaluate whether any difference in the proportion of peaks associated to genes in the different clusters exists, either in promoters or in gene bodies. Moreover, the potential association of the clusters with a specific gene type (coding or non-coding) is inspected too.

### *Analysis of overlap with transcription factors, open chromatin regions and histone modifications*

An important characterization of a group of ChIP-seq peaks is given by their intersection with other transcription factors, indicating the co-occurrence of different protein bindings in the same site and suggesting the existence of protein complexes or co-regulatory activities. Relationships with chromatin accessibility and histone modifications are also significant. In particular, we observe that transcription factors are more likely to bind DNA in open chromatin regions. Therefore, DNase I hypersensitive sites are considered and used to compute the proportion of peaks in each cluster that overlap open chromatin regions. The distribution of the percentage of intersection, conditionally to be non-zero, is evaluated too. Both are compared to the random case, obtained by randomly shuffling the peaks among the chromosome where they lie, after excluding repetitive elements given by the RepeatMasker track of UCSC [39]. This allows us to characterize the different clusters, excluding at the same time that the procedure described in the previous subsection selects some clusters of noisy regions with artifact signals.

Moreover, we use random forest analysis (see [40] for details about this method) to assess which co-occurrences are more correlated with the regions of interest and to select the most important regulatory elements in relation with the clusters identified as explained before. Specifically, for each single peak of the starting ChIP-seq, we compute the percentage of intersection with the set of regions selected by a ChIP-seq for a different transcription factor or histone modification. Doing this calculation for all the available proteins, we obtain a matrix $P$ whose rows represent the different peaks, while the columns correspond to the other considered experiments. We add also a column with the percentage of intersection with DNase I hypersensitive sites. The various ChIP-seq (and DNase-Seq) overlaps are the predictors of a random forest classification with response the categorical variable of clusters membership. Priors for the classes are used to handle imbalanced classification problems. Once the random forest classifier is built, variable importance is estimated through the mean decrease in Gini index [40, 41]. Briefly, for each node within a tree of the random forest classifier, the Gini index is computed as $1\text{-}p^2(c_1)\text{-}\ldots\text{-}p^2(c_k)$ where $p^2(c_i)$ is the proportion of the samples assigned to the node belonging to category $c_i$. Every time a node is splitted using a certain variable, there is a decrease in Gini index. The variable total decrease is given by the sum of these Gini index reductions over all nodes of a tree in which the variable is used to split. Finally, the mean decrease in Gini index is obtained averaging over all the trees. By ranking regulatory elements according to their importance in classifying all the different clusters, or one cluster against the others,

and by looking at concordance between replicates, we can select a small number $N$ of proteins on which to implement what we term *combinatorial interaction analysis*.

The combinatorial interaction analysis consists of evaluating the number (and the percentage) of peaks of the starting ChIP-seq that overlap each of the $2^N$ possible combinations of the transcription factors and histone modifications selected by using random forests. The same count is done for the different clusters obtained in the previous steps. Here we can say that a peak of the starting ChIP-seq intersects a regulatory element when it overlaps at least one ChIP-seq replicate for that protein (less stringent request) or when it overlaps all replicates (more stringent rule). In both cases we require a minimum overlap of 1 bp. Here we are interested in differences between clusters regarding specific co-occurrences, that can give a functional explanation to the statistically significant different shapes found.

Lastly, after converting the matrix $P$ of the percentage of intersections to a boolean matrix that indicates overlap versus non-overlap, multiple correspondence analysis (see, for example, [42] for a detailed explanation of this technique) is used to study the relationships between overlaps with different proteins, by evaluating all the replicates simultaneously. The map showing the different levels of each variable in principal coordinates allows us to assess the similarity between ChIP-seq replicates for the same protein and to detect associations in overlapping elements in the different clusters obtained by using the methodology we proposed before.

### *Gene differential expression analysis*

To assess whether or not peak shape is related to the regulation of gene expression, we combine the results of previous clustering and RNA-seq experiments. We assign each ChIP-seq peak to a known gene (using RefGene database) if the peak is located less than 5 kb from the transcription start site of the gene or if it falls within the gene body. We inspect less and more restrictive association rules too, considering peaks located less than 10 kb or 2.5 kb from the transcription start site, besides gene body. Once the association is established, we inspect the possible correlations between clusters and gene expressions by plotting in logarithmic scale the reads per kilobase per million (RPMK), computed normalizing HTSeq count results on a RNA-Seq without treatment. Moreover, we are interested in recognizing peak shapes involved in regulating some genes targeted by the protein of interest. This goal is achieved by analyzing changes of gene expression in a RNA-Seq after knockdown of the protein under investigation. After running DESeq with default options to select significantly differential expressed genes (the ones with false discovery rate $< 0.05$), we analyze the percentages of up and down regulated genes in the silenced cells, for each cluster. This permits to identify shapes that are more related to the protein acting as repressor or activator, respectively. To assess the significance of the findings, Fisher's exact test is performed to test the null hypothesis of independence between this percentages and the association of the genes with a particular cluster, versus the alternative hypothesis that the odd ratio is less than one. Resulting p-values are then corrected for multiple testing by using Bonferroni method. To avoid mixing of different clusters, in these analyses we contemplate only the genes that are univocally associated to a single cluster.

## Additional files

**Additional file 1: Supplementary Figures and Tables.** (PDF 2223 kb)
**Additional file 2: Random forest analysis.** (XLSX 74 kb)
**Additional file 3: GREAT GO Biological Process.** (XLSX 364 kb)
**Additional file 4: ChIP-seq data.** (XLSX 14 kb)

#### Abbreviations

ChIP-exo: Chromatin ImmunoPrecipitation exonuclease; ChIP-seq: Chromatin ImmunoPrecipitation sequencing; DNase-seq: DNase I hypersensitive sites sequencing; GO: Gene Ontology; RNA-seq: RNA sequencing; RPKM: Reads Per Kilobase per Million; TF: Transcription Factor.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

MAC designed the study, performed the statistical analysis, carried out the calculations, wrote the code, interpreted the results and wrote the manuscript. LR designed and supervised the study, interpreted the results and wrote the manuscript. LMS, SV and PS designed and supervised the statistical analysis, interpreted the results and participated in manuscript writing. PGP and GID participated in the design of the study, in the interpretation of results and in manuscript writing. All authors read and approved the final manuscript.

#### Acknowledgements

This work was carried out within the "Genomic Computing Project", a research program involving the European Institute of Oncology, the Istituto Italiano di Tecnologia and the Politecnico di Milano. We wish to thank Lucilla Luzi and Alice Parodi for their collaboration to the project. We are grateful to Giancarlo Ferrari for his help in implementing the computation of the shape index *M*, and to Luciano Giacò for providing us his assistance in analyzing the RNA-Seqs. The work was in part supported from the "Data-Driven Genomic Computing (GenData 2020)" PRIN project (2013–2015).

#### Author details

[1]MOX - Dipartimento di Matematica, Politecnico di Milano, Milan, Italy. [2]Department of Experimental Oncology, European Institute of Oncology, Milan, Italy. [3]Dipartimento di Scienze della salute, Università degli Studi di Milano, Milan, Italy. [4]Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia, Milan, Italy.

Received: 22 May 2015 Accepted: 20 October 2015
Published online: 28 October 2015

#### References

1. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nat Methods. 2009;6:S22–32.
2. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One. 2010;5, e11471.

3. Kulakovskiy IV, Makeev VJ. Motif discovery and motif finding in ChIP-Seq data. In: Poptsova, editor. Genome analysis: current procedures and applications. Norfolk: Caister Academic Press; 2014. p. 83–100.
4. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10:669–80.
5. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009;25:1952–8.
6. Hower V, Evans SN, Pachter L. Shape-based peak identification for ChIP-seq. BMC Bioinform. 2011;12:15.
7. Mendoza-Parra MA, Nowicka M, Van Gool W, Gronemeyer H. Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. BMC Genomics. 2013;14:834.
8. Mahony S, Edwards MD, Mazzoni EO, Sherwood RI, Kakumanu A, Morrison CA, et al. An integrated model of multiple condition ChIP-seq data reveals predeterminants of Cdx2 binding. PLoS Comput Biol. 2014;10, e1003501.
9. Hu H, Ji H. PolyaPeak, detecting transcription factor binding sites from ChIP-seq using peak shape information. PLoS One. 2014;9, e89694.
10. Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G. MMDiff: quantitative testing for shape changes in ChIP-seq data sets. BMC Genomics. 2013;14:826.
11. SIC-ChIP software. http://cgsb.genomics.iit.it/wiki/projects/SIC-ChIP.
12. Bao Y, Vinciotti V, Wit E, AC't Hown P. Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. BMC Bioinform. 2013;14:169.
13. Papadopoulos GL, Karkoulia E, Tsamardinos I, Porcher C, Ragoussis J, Bungert J, et al. GATA-1 genome-wide occupancy associates with distinct epigenetic profiles in mouse fetal liver erythropoiesis. Nucleic Acids Res. 2013;41:4938–48.
14. Wu J, Zhou LQ, Yu W, Zhao ZG, Xie XM, Wang WT, et al. PML4 facilitates erythroid differentiation by enhancing the transcriptional activity of GATA-1. Blood. 2014;123:261–70.
15. Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol Cell. 2009;36:667–81.
16. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
17. Encyclopedia of DNA Elements (ENCODE) Project. https://www.encodeproject.org.
18. Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, et al. Model-based analysis of ChIP-seq (MACS). Genome Biol. 2008;9:R137.
19. Shimizu R, Engel JD, Yamamoto M. GATA1-related leukaemias. Nat Rev Cancer. 2008;8:279–87.
20. Pan X, Ohneda O, Ohneda K, Lindeboom F, Iwata F, Shimizu R, et al. Graded levels of GATA-1 expression modulate survival, proliferation, and differentiation of erythroid progenitors. J Biol Chem. 2005;280:22385–94.
21. Evans T, Reitman M, Felsenfeld G. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. Proc Natl Acad Sci U S A. 1988;85:5976–80.
22. Martin DI, Orkin SH. Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. Genes Dev. 1990;4:1886–98.
23. Ko LJ, Engel JD. DNA-binding specificities of the GATA transcription factor family. Mol Cell Biol. 1993;13:4011–22.
24. Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, et al. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev Cell. 2011;20:597–609.
25. Chlon TM, Doré LC, Crispino JD. Cofactor-mediated restriction of GATA-1 chromatin occupancy coordinates lineage-specific gene expression. Mol Cell. 2012;47:608–21.
26. Yu M, Riva L, Xie H, Schindler Y, Moran TB, Cheng Y, et al. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. Mol Cell. 2009;36:682–95.
27. Lan X, Witt H, Katsumura K, Ye Z, Wang Q, Bresnick EH, et al. Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic Acids Res. 2012;40:7690–704.
28. Boes M, Dake BL, Bar RS. Interactions of cultured endothelial cells with TGF-beta, bFGF, PDGF and IGF-I. Life Sci. 1991;48:811–21.
29. Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, et al. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res. 2010;20:1064–83.
30. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011;147:1408–19.
31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754–60.
32. Rye MB, Sætrom P, Drabløs F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. Nucleic Acids Res. 2011;39, e25.
33. Ramsay JO, Silverman BW. Functional data analysis. 2nd ed. New York: Springer; 2005.
34. Evans SN, Hower V, Pachter L. Coverage statistics for sequence census methods. BMC Bioinform. 2010;11:430.
35. Johnson RA, Wichern DW. Applied multivariate statistical analysis. 6th ed. Upper Saddle River: Pearson; 2007.
36. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009.
37. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501.
38. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011;27:1696–7.
39. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010. http://www.repeatmasker.org. Accessed 25 Feb 2015.
40. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
41. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinform. 2009;10:213.
42. Greenacre M. Correspondence analysis in practice. 2nd ed. London: Chapman & Hall/CRC; 2007.