González-Vallinas et al. BMC Genomics (2015) 16:523
DOI 10.1186/s12864-015-1704-0

BMC Genomics

RESEARCH ARTICLE Open Access

A semi-supervised approach uncovers thousands of intragenic enhancers differentially activated in human cells

Juan González-Vallinas[1], Amadís Pagès[1], Babita Singh[1] and Eduardo Eyras[1,2*]

Abstract

Background: Transcriptional enhancers are generally known to regulate gene transcription from afar. Their activation involves a series of changes in chromatin marks and recruitment of protein factors. These enhancers may also occur inside genes, but how many may be active in human cells and their effects on the regulation of the host gene remains unclear.

Results: We describe a novel semi-supervised method based on the relative enrichment of chromatin signals between 2 conditions to predict active enhancers. We applied this method to the tumoral K562 and the normal GM12878 cell lines to predict enhancers that are differentially active in one cell type. These predictions show enhancer-like properties according to positional distribution, correlation with gene expression and production of enhancer RNAs. Using this model, we predict 10,365 and 9777 intragenic active enhancers in K562 and GM12878, respectively, and relate the differential activation of these enhancers to expression and splicing differences of the host genes.

Conclusions: We propose that the activation or silencing of intragenic transcriptional enhancers modulate the regulation of the host gene by means of a local change of the chromatin and the recruitment of enhancer-related factors that may interact with the RNA directly or through the interaction with RNA binding proteins. Predicted enhancers are available at http://regulatorygenomics.upf.edu/Projects/enhancers.html.

Background

Transcriptional enhancers are characterized by specific chromatin signatures, which differ depending of whether the enhancer is active or not [1–5]. Transcriptional enhancers have been generally identified by studying the genome-wide binding of the acetyl-transferase P300, a ubiquitous enhancer co-activator [1, 6, 7]. However, not all P300-bound enhancers show activity [8]. Enhancers have also been characterized by their chromatin state [1, 2, 9, 10]; and, although the mono-methylation of histone 3 lysine 4 (H3K4me1) has been identified to be an important signature for enhancers [2], this mark is not sufficient for enhancer activation [3, 11]. In fact, recent evidence shows that other marks like H3K27ac [1, 3–5] and H3K4me3 [5, 11] may be necessary for enhancer activity. Additionally, the recruitment of RNAPII and the concomitant production of enhancer-associated RNAs (eRNAs) have also been associated to active enhancers [3–5, 12, 13].

Although enhancers are typically defined to regulate gene transcription at a distance, about 50 % of potential enhancers predicted by high-throughput methods lie within protein-coding genes [2] and some overlap exons [14, 15]. Intragenic enhancers can regulate the expression of the host gene [14] or of a nearby gene [15], and have been proposed to act as alternative promoters [16]. These results raise the question of how many intragenic enhancers may be active in a cell and whether upon their activation or silencing they may affect the processing of the host gene, possibly by means of local changes of the chromatin state. In this direction, there is evidence that some enhancers upstream of a reporter gene can affect splicing *in vitro* [17], and that intragenic enhancers bound by Argonaute-1 (AGO1) protein can

* Correspondence: eduardo.eyras@upf.edu
[1] Universitat Pompeu Fabra, Dr Aiguader 88, E08003 Barcelona, Spain
[2] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010 Barcelona, Spain

affect the constitutive and alternative splicing of the host gene [18]. In this work we describe a computational method to predict active enhancers based on chromatin signals. This method, which uses the relative enrichment of chromatin signals between cell lines to the detect cell specific active enhancers, predicts thousands of intragenic active enhancers. Additionally, we find evidence that the differential activation of enhancers inside genes affect the expression and splicing of the host genes. We propose that the activation or silencing of intragenic transcriptional enhancers can modulate the regulation of the host gene through a local change of the chromatin.

Methods

Datasets

Annotated human enhancers with a mouse homologous enhancer that has been experimentally validated were downloaded from VISTA [19]. The gene set was obtained from the 7th release of GENCODE, human assembly GRCh37 (hg19). ChIP-Seq and RNA-Seq datasets were downloaded from ENCODE [20] for K562 and GM12878 cells. The datasets used were: ChIP-Seq for CTCF, EZH2, P300, RNAPII, PU.1 (SPI1), STAT1, H3K9ac, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K36me3, H3K79me2, H4K20me1 and H2A.Z; one Control ChIP-Seq experiment and one input experiment; RNA-Seq for short (<200 nt) and long (>200 nt), polyA+ and polyA-RNAs from whole cell, nucleus and cytosol; and DNaseI data for the same cell lines. All datasets were downloaded in the form of mapped reads to the reference hg19 genome in BAM format.

Relative enrichment calculation

We considered sliding windows of 1500 nt along the entire genome, as suggested by the length distribution of experimentally validated enhancers [19, 21] (Additional file 1: Figure S1), with a slide shift of 500 nt, resulting in a total of 3,086,047 overlapping windows. In order to avoid mixing enhancer signal with genic and promoter signals, we discarded windows that were closer than 500 nt to an annotated TSS. The same approach was applied to intergenic (Additional file 1: Figure S2A) and intragenic (Additional file 1: Figure S2B) regions. Although there are more intergenic windows (~$3{\cdot}10^6$ vs ~$2.2{\cdot}10^6$) in both cases the amount of windows with signal was similar (~1.5 million windows), which were then kept for further processing. The relative enrichment of chromatin signals between 2 cell lines was calculated to predict active enhancers in K562 (relative increase of activation marks in K562 with respect to GM12878) and silent enhancers in K562 (relative decrease of activation marks in GM12878 with respect to K562, i.e. active in GM12878). Full quantile normalization for counts and GC content was applied using EDASeq [22]. GC content in each region was calculated as the proportion of G + C in the 1500 nt window. After normalization, the z-score of the relative enrichment of each ChIP-Seq signals between K562 and GM12878 was calculated with Pyicoteo [23] using the *pyicoenrich* function (https://bitbucket.org/regulatorygenomicsupf/pyicoteo). A vector of z-scores per region was obtained, which we refer to as attributes, consisting of the 17 enrichment z-scores for the ChIP-Seq and Input datasets. A positive z-score for a region indicates an increased in ChIP-Seq signal in K562 relative to GM12878 in that region, whereas a negative z-score indicates a decreased signal in K562 relative to GM12878; and z-scores close to zero indicate no significant differences between the cell lines. For all datasets, except for the ChIP-Seq with non-specific antibody and for the RNA-Seq datasets, we used replicates. The relative enrichments were calculated with respect to the distribution described by the comparison between replicates. When replicates were not available, these were simulated by pooling the two conditions and dividing them using random sampling [23].

Feature selection

Feature selection was performed using Boruta [24], which finds informative features by measuring the relevance of each attribute with respect to a reference attribute, also called correlation class, and in comparison with a random model extracted from the original dataset. Boruta uses the correlation class to evaluate the other features against it using Random Forests [25]. We performed this analysis using as correlation class each of the individual marks (Additional file 1: Figure S3). In each case, the correlation was performed 10 times using normalized counts on a subset of 5000 intergenic windows, sampled randomly in each one of the 10 iterations. In order to avoid possible biases, in each analysis the correlation class was defined as the ChIP-Seq signal minus the level of Input DNA. The features used as negative controls were the ChIP-Seq sample for a non-specific antibody (Control sample) (Additional file 1: Figure S3D) and H4K20me1 (Additional file 1: Figure S3E), which has been associated to transcription repression and heterochromatin but not to enhancer activity [26, 27]. Running the selection algorithm with the H3K4me1 mark, the average Boruta score for the control increased notably, suggesting that the mark is present in many regions along the genome (Additional file 1: Figure S3C).

Window clustering

Fifteen thousand arbitrary intergenic windows of length 1500 bp were used as seed for the prediction model. Various different seed selections of the same size did not change the results significantly. These 15,000 windows were clustered using Mclust [28]. Mclust is based on finite normal mixture modeling and uses the Bayesian Information Criterion (BIC) [29] for model optimization. The BIC score plateaus at 3 clusters for most models (Additional file 1: Figure S4A). The seed windows corresponded to 552 active, 616 silent and 13,832 no-change windows. This indicates that there are mostly three main classes, two that correspond to active and silent enhancers, and a class composed of a gradient of multiple chromatin states, which show little or no relative change of chromatin activity. This is further supported by the uncertainty plot, which shows that regions classified with higher certainty are on the extreme values of the correlation (Additional file 1: Figure S4B). The final model used for clustering was the centroid type (labeled as VEV), which creates clusters with variable volume, equal shape, and variable orientation. This model was used to classify the genome-wide 1500 bp windows (Additional file 1: Figure S2) using the same clustering method Mclust to predict intergenic enhancers. Intragenic enhancers were calculated using the same seed of 15,000 intergenic windows as before. The clustering was performed in the same way as for intergenic enhancers. As controls we calculated 4 sets of randomized positions (intergenic/intragenic and active/silent putative predictions). These sets were calculated from the predicted enhancers by randomizing the positions, not closer than 500 nt to any gene, avoiding gaps, genic regions, and other random locations previously generated, and keeping the same length and the same number of regions (Additional file 1: Figure S2A). Random intragenic enhancers were generated similarly by placing the intragenic enhancers in a random location inside the same gene, avoiding regions of 1 kb around any internal TSSs and avoiding other random enhancers previously generated (Additional file 1: Figure S2B). All predicted intergenic and intragenic enhancers can be visualized in the UCSC genome browser through the link http://regulatorygenomics.upf.edu/Projects/enhancers.html.

Linking enhancers to genes

Enhancers were linked to genes by selecting the closest TSS on either direction and by using ChIA-PET data for RNAPII in K562 cells for two replicates from ENCODE [20]. An enhancer was considered connected to a gene if there were at least 3 ChIA-PET pairs connecting both the predicted enhancer and the region of 1 kb around the TSS of the gene. Random enhancers used as controls were calculated as described above. For the association of enhancers to genes, only enhancers that were between 2 and 100 kb from a TSS were considered. Genes associated to cancer were obtained from the Cancer Gene Census (http://cancer.sanger.ac.uk/cancergenome/projects/census/) [30].

Expression and splicing analysis

For every gene in GENCODE (v7) annotation [31], the most upstream TSS (TSS1) and all alternative TSSs (TSS2, TSS3, etc.) were considered. Each pair TSS1-TSS2, TSS2-TSS3, etc. was considered as an alternative transcription event. RNAPII relative enrichment levels were measured around each TSS using the same method as before. To control possible association with upstream enhancers, we discarded all alternative TSS events that had a predicted intergenic enhancer (active or silent) 100 kb upstream of the gene. We calculated the expression levels of the annotated transcript isoforms using cufflinks v2.1.1 [32] with parameters *–library-type fr-firststrand –no-effective-length-correction –min-frags-per-transfrag 5* and masking all rRNAs, tRNAs and mitochondrial sequences. The relative changes in transcript abundance were obtained using Cuffdiff with parameters *–library-type fr-firststrand –min-reps-for-js-test 1*, using the merged GTF file obtained from Cufflinks for GM12878 and K562, along with the bam files of GM12878 and K562 with replicates. This provided 3552 genes (6.68 %) with relative changes in expression between the two cell lines.

Alternative splicing events from the Gencode v7 annotation [31] were calculated using the software SUPPA (https://bitbucket.org/regulatorygenomicsupf/suppa). Only events that do not overlap any other alternative splicing event were kept, giving rise to a total of 5319 events. For exon skipping events, defined by an exon triple E1–E2–E3, the inclusion level (PSI) of the middle exon E2, was calculated as the fraction of reads that include the exon over the total number of reads that include and skip the exon:

$$PSI = \frac{n_{12} + n_{23}}{n_{12} + n_{23} + 2\mathrm{n}_{13}}$$

where n_{12}, n_{23} and n_{13} are the number of reads that span the junctions E1–E2, E2–E3 and E1–E3, respectively. PSI values were calculated using junction reads only, since enhancers can produce RNA as well, so the enhancer-related RNAs may be mistakenly included in the PSI calculation when they overlap an event. Reads at junctions were counted with sjcount [33] from the mapped RNA-Seq data, using the *-read1 1* and *-read2 0* parameters. For this analysis, RNA-Seq reads were mapped using STAR [34] with parameters *–outSJfilterOverhangMin –1 -1 -1 -1 and –sjdbScore 100* in order to use only annotated junctions. A genome index was previously generated with STAR over the Gencode.v7 annotation using the *–sjdbOverhang 75* parameter in order to adjust the splice junction database to the length of the RNA-Seq reads. Finally, only events with a total of 20 or more reads mapping at the junctions were kept. This gave a final number of 3227 and 3192 events with PSI values from the nuclear and cytosolic RNA-Seq experiments, respectively.

We defined the events to be regulated if they had |delta PSI| > 0.1 in at least one replicate comparison between cell lines. Using two pairings of the replicates, this gave rise to 339 and 293 events (148 in common) with the cytosolic samples, and 367 and 378 (210 in common) for the nuclear samples. Additionally, we defined a set of alternative events that do not change splicing by imposing |delta PSI| < 0.05 between the same replicate comparisons used before. This gave rise to 1722 and 1534 (1328 in common) for the cytosolic samples, and 1627 and 1497 (1278 in common) for the nuclear samples.

Results and discussion

Modeling and prediction of active transcriptional enhancers

We built a computational predictive model based on the relative differences in various chromatin marks between two cellular conditions. We applied this model to study the differences between the ENCODE cell lines K562, a leukemia cell line, and GM12878, a blood cell derived cell line. Using windows along the entire genome (Additional file 1: Figures S1 and S2), we considered the relative enrichment of a number of histone marks and protein factors (Methods). We then clustered windows into classes according to the chromatin features. In order to determine which features are relevant for classification, we performed a feature selection analysis in which one signal is chosen as a proxy for a classification value and is compared against the rest (Methods). We then considered two of the main epigenetic marks related to active enhancers, H3K27ac [3] and H3K4me3 [11], as proxies for enhancer activity. We found that H3K4me1 and H3K4me2, observed to be present in active and non-active enhancers [11] are strongly correlating signals (Fig. 1a and Additional file 1: Figure S3A). We also consistently found H2A.Z, which is a histone variant associated to open chromatin and H3K4 methylation [35]; and P300, which is ubiquitously present in enhancers [6].

Interestingly, when P300 or H3K4me1 were used as a correlation feature, the signals H3K27ac and H3K4me3 did not appear as the most significantly associated (Additional file 1: Figure S3B and C). Additionally, P300 seemed to associate with the largest subset of features, which is consistent with experimental evidence showing that P300 associates generally to enhancers [1, 6]. However, enhancers with H3K4me1 and/or P300 occupancy are not always active [3, 11], since H3K4me1 precedes enhancer-binding factors and P300 may be present in poised and intermediate enhancer states [36]. On the other hand, we did not find RNAPII and H3K36me3 to be strong predictors of enhancer activity (Fig. 1a and Additional file 1: Figure S3A), even though they have been previously detected on enhancers [12, 13]. Additionally, although we found a strong correlation of PU.1 (SPI1) with H3K27ac, it does not correlate with H3K4me3, hence it is likely that PU.1 associates to a subset of the putative enhancers [37]. Based on these results, we decided to keep those features that scored consistently above the technical and biological controls in the feature selection analysis using H3K27ac and H3K4me3 as correlation classes, including these two marks. That is, we used as predictors of enhancer activity the following signals: P300, H3K27ac, H3K9ac, H3k4me1/me2/me3 and H2A.Z.

Clustering the genomic windows according to the relative enrichment of the selected features (Methods) resulted in three optimal classes (Additional file 1: Figure S4). We recovered one class characterized for being enriched in H3K4me3 and H3K27ac (Fig. 1b), which we considered to be enhancers that are active in K562 cells (silent in GM12878). We recovered a second class characterized by a depletion of these same marks in K562 (Fig. 1b), which we considered to be active enhancers in GM12878 (silent in K562). Finally, the third cluster showed small or no changes in most of the signals, indicating that these regions do not have any differential activity between the two cell lines. These regions do not necessarily represent enhancers and are labeled as no-change. These three groups (active, silent, no-change) define the three predictable classes of our computational model, two of which can be identified with enhancer classes: active and silent. The genome wide classification analysis resulted in 66,079 windows predicted to be active in K562 (silent in GM12878) and 64,436 windows predicted to be active in GM12878 (silent in K562).

In-silico validation of active transcriptional enhancers

In order to evaluate the accuracy of our predictions, we first compared our predicted enhancers windows with the enhancer regions predicted in the same cell lines by ChromHMM [38]. The majority of our enhancers predicted as active in K562 or GM12878 overlap with ChromHMM windows labeled as weak or strong enhancers in the same cells (Additional file 1: Figure S5A and B). On the other hand, when we compared active windows with ChromHMM labels in the other cell line, the majority corresponds to ChromHMM silent windows (Additional file 1: Figure S5C and D), as expected. Furthermore, the overlap of our active enhancers with predicted ChromHMM enhancers increases with the posterior probability of our predictions (Additional file 1: Figure S5E and F). In contrast, when comparing the active enhancers in one cell line with the ChromHMM labels from the other cell line, we found no correlation with the posterior probability (Additional file 1: Figure S5G and H).

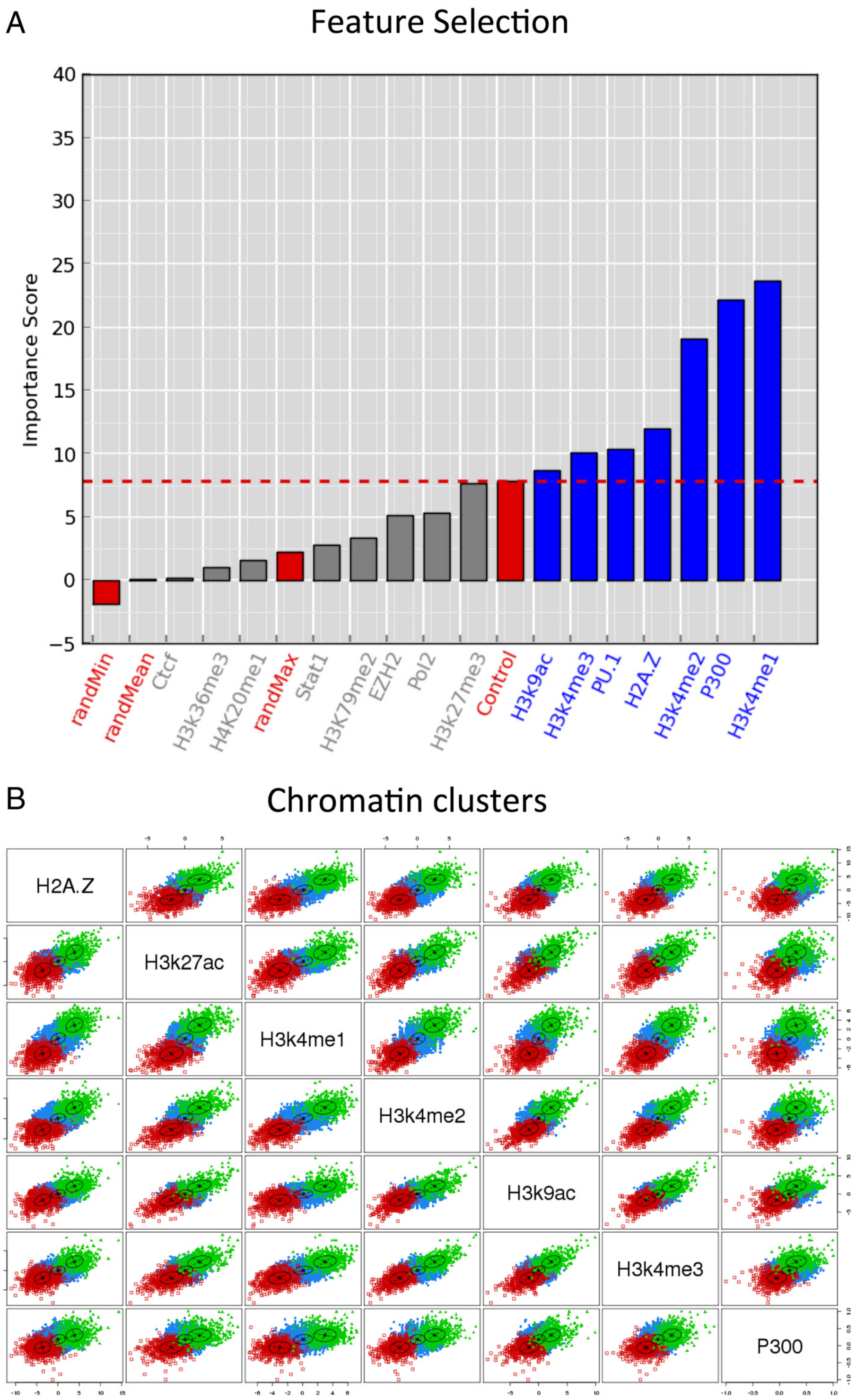

Fig. 1 A predictive model of active enhancers. **a** Feature selection using H3K27ac (minus the Input DNA) as a correlation class. The bars represent the average importance score per feature. Red labels and bars indicate the minimum (randMin), mean (randMean) and maximum (randMax) of the simulated replicates, as well as the ChIP-Seq with a non-specific antibody (Control). The red dashed line separates the relevant features (in blue) from the non-relevant features (in grey). **b** Scatter plot of the intergenic windows according to relative enrichment z-scores for every pair of selected features (x and y axis). Each dot represents a window and windows are separated according to the three classes: active enhancers (green), no-change regions (blue) silent enhancers (red). The black centroids show the centers and standard deviations of the correlations between different features

Based on these comparisons, we kept predictions with a posterior probability of > 0.95, which resulted in 36,301 active windows in K562 and 37,859 active windows in GM12878. Overlapping windows were then clustered into 16,646 active enhancers in K562 and 16,328 active enhancers in GM12878, which distribute evenly along the genome (Additional file 1: Figure S6A). These enhancers have mean length of 3053 bp and the majority of them (87.65 %) are shorter than 5 kb (Additional file 1: Figure S6B). There were also 273 (1.38 %) predictions longer than 10 kb, which may correspond to large-scale chromatin domains [39] or to clusters of enhancers [40]. We filtered out those predictions longer than 5 kb, resulting in 10,365 active enhancers and 9777 silenced enhancers, with mean lengths of 2704.6 and 2588 bp, (median lengths of 2500 and 2000 bp), respectively. These average lengths are in agreement with previous analyses of enhancers from ChIP-Seq data of histone marks and protein factors [5, 11, 15].

We next studied the association of enhancers to other signals not considered in the predictive model. PU.1 and RNAPII correlate with the predicted active enhancers, with 25.3 and 20.1 % of the active enhancers in K562 showing a significant relative enrichment (left-tailed p-value < 0.01) in PU.1 (Fig. 2a) and RNAPII (Additional file 1: Figure S7A), respectively. Similarly, we found a strong association of DNaseI to our predicted enhancers, in agreement with previous observations [2, 9, 10] and 53.6 % of the active and silent enhancers show a significant enrichment (left-tailed p-value < 0.01) (Fig. 2b). Likewise, 46.7 % of the silent enhancers in K562 (active in GM12878) show a significant depletion in DNaseI (right-tailed p-value < 0.01). In contrast, H3K27me3 shows a weak inverse correlation with enhancer activity and 6.5 % of the silent enhancers in K562 show a significant enrichment (right-tailed p-value < 0.01) of H3K27me3 (Additional file 1: Figure S7B). Although CTCF and H3K36me3 have been detected before on enhancers [12, 13], we observed a weak correlation of these signals with enhancer activity and only 7.4 and 4.6 % of active enhancers in K562 show a significant enrichment in CTCF and H3K36me3, respectively (Additional file 1: Figure S7C and D).

We additionally investigated whether enhancer-associated RNAs (eRNAs) are found in our predictions. Enhancer activity correlates with the production of polyA+ (Fig. 2c) and polyA- (Additional file 1: Figure S8A) long (>200 bp) nuclear RNAs, compared to silent enhancers. This relative enrichment is much larger than for the other RNA subclasses (Additional file 1: Figure S8B). Interestingly, there is also enrichment of cytosolic polyA+ RNAs (Additional file 1: Figure S8C), but not of cytosolic polyA-RNAs (Additional file 1: Figure S8D) or short RNAs (<200 bp) (Additional file 1: Figure S8E and F). Moreover, not all enhancers predicted as active appear to generate eRNAs: 26.4 and 32.1 % of the predicted active enhancers in K562 have a significant (left-tailed p-value < 0.01) increase of nuclear polyA+ and polyA-, respectively. In comparison, only 1.25 % of active enhancers have significant (left-tailed p-value < 0.01) increase for short nuclear RNAs. For cytosolic polyA+, 18.7 % of the predicted active enhancers in K562 have a significant (left-tailed p-value < 0.01) increase of eRNAs. In contrast, only 9.2 % of these active enhancers have a significant enrichment of short total RNAs and polyA- cytosolic RNAs, respectively.

Although enhancers can regulate genes from afar, they tend to be enriched upstream of genes (Visel et al. [6]). We therefore connected enhancers to genes by choosing for each enhancer the closest annotated transcription start site (TSS) in either direction. With this approximation, active intergenic enhancers show enrichment at distances close to TSSs compared to random regions and to silent enhancers (Fig. 2d). Using these enhancer-TSS pairs, we calculated the relative change in gene expression measured from RNA-Seq data (Methods). We observed that genes with activated enhancers at a distance between 2 and 10 kb show up-regulation, whereas genes with silenced enhancers in the same distance range show down-regulation (Fig. 2e). Moreover, this association is conserved when the distance range of the enhancers is extended to be between 10 and 100 kb from the closest gene (Additional file 1: Figure S9A). Further support for transcription activity in association to our predicted enhancers was found measuring the relative density of RNAPII around the TSS in genes close to predicted enhancers, which was found to correlate with enhancer activity (Additional file 1: Figure S9B).

We additionally searched for evidence of direct physical interactions for the enhancer-TSS pairs calculated above by using ChIA-PET data for RNAPII [41]. Although only a small fraction of active enhancers have ChIA-PET links to TSS regions (1.6 %), there is enrichment over silent enhancers and randomized regions (Additional file 1: Figure S9C), indicating that predicted active enhancers tend to have more ChIA-PET links than silent enhancers and expected by chance. Finally, we investigated whether enhancers active in K562 have any association to genes that have been involved in cancer. Using the cancer gene census [30], we found that enhancers predicted to be active in K562 are enriched for genes related to cancer, compared to random regions and to enhancers silent in K562 (active in GM12878) (Fig. 2f). Interestingly, oncogenes can be linked more frequently to active enhancers and suppressors can be linked more frequently to silent enhancers (Additional file 1: Figure S10). In summary, these analyses indicate that our predicted enhancers show properties of active enhancers. We therefore set out to predict intragenic enhancers using the same computational model.

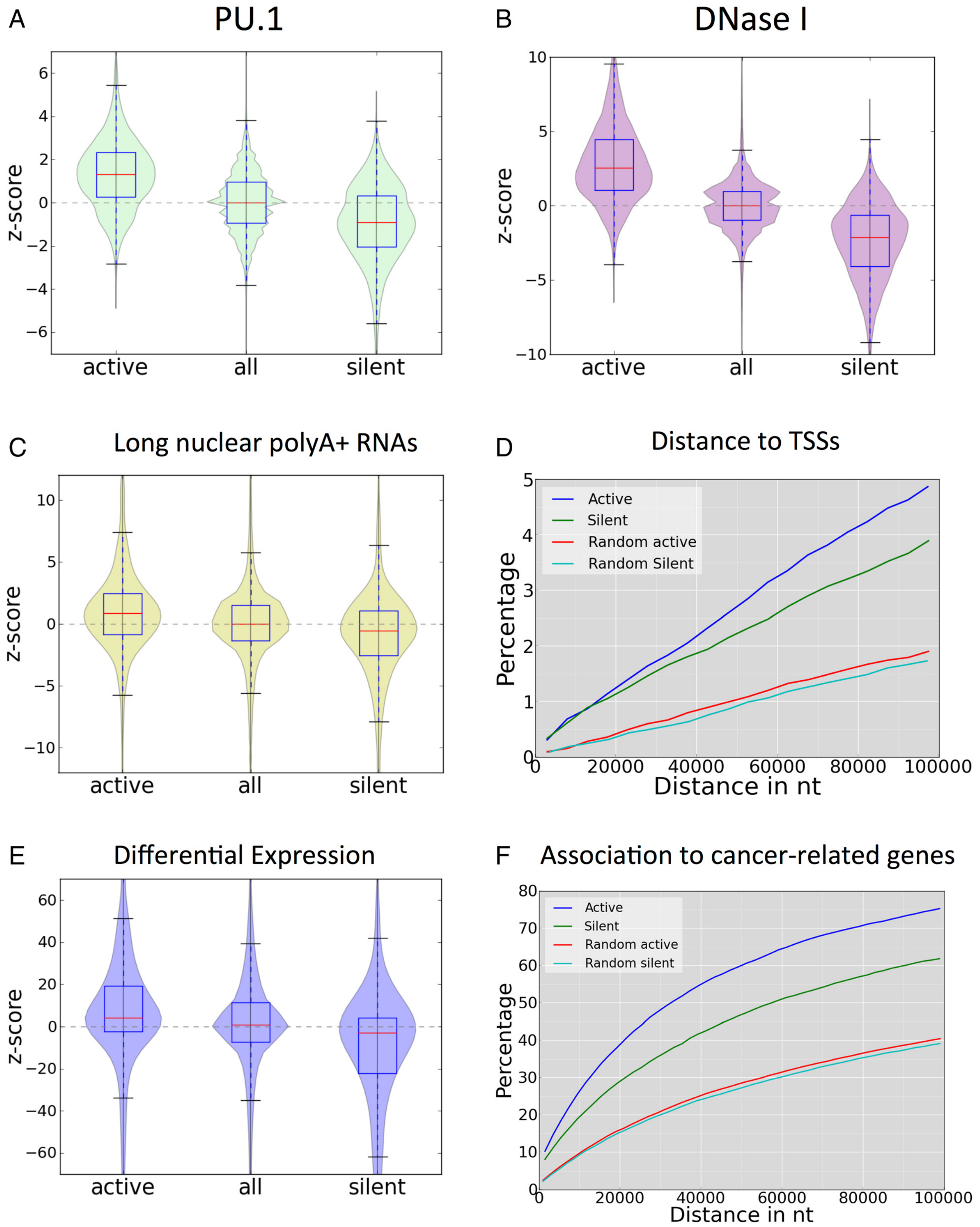

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Properties of predicted intergenic enhancers. Relative enrichment of PU.1 (**a**), DNaseI (**b**) and polyadenylated long (>200 nt) nuclear RNA (**c**) at active and silent enhancers, as well as for regions of no-change in chromatin. The violin plots describe the distributions for the z-score of the relative enrichment along the y-axis. Positive z-score values mean enrichment in K562, while negative z-scores mean enrichment in GM12878. **d** Percentage of enhancers at a given distance from the TSS, for active (blue), silent (green), as well as for the corresponding randomized sets (red and cyan) (Methods). **e** Relative expression change in genes associated to enhancers by proximity to the TSS. The violin plot describes the distributions of z-scores of the relative enrichment of RPKM values along the y-axis for genes associated to active and silent enhancers, as well as for no-change regions calculated with Pyicoteo [23]. Genes where linked to the nearest enhancers within a distance range between 2 and 10 kb. **f** Cumulative distribution of enhancer nearby genes related to cancer in terms of the distance between the TSS and the closest enhancer. The comparison is made between active and silent predicted enhancers, and the corresponding randomizations

Thousands of intragenic enhancers are differentially activated in human cells

Active enhancers regulating the expression of nearby genes have been observed in exons [14, 15] and about 50 % of enhancers predicted by high-throughput methods lie within protein-coding genes [2]. Additionally, by comparing the overlap of validated VISTA elements with the annotation in Gencode.v7 [31], we observe that there is no preference for intragenic or intergenic regions (Additional file 1: Figure S1). All these evidences indicate that intragenic enhancers represent an important regulatory component of the genome. However, it remains an open question how many intragenic enhancers may be active in a given cell. Accordingly, we decided to apply our predictive model to localize putative intragenic enhancers that are activated in K562 relative to GM12878, and vice versa.

In order to predict intragenic active enhancers, we considered 1.5 kb sliding windows inside genes, starting 500 bp downstream of the first TSS and eliminating all windows that overlap with a 1 kb region around every annotated alternative TSS (Additional file 1: Figure S2). This resulted in an initial set of 2,206,307 possible 1.5 kb windows, for which we used the same selected chromatin features as for the intergenic enhancers. Using a seed of 15,000 intergenic regions and the same clustering approach as before, we predicted 73,080 active and 92,225 silenced regions. As we did previously with intergenic enhancers, we compared our predicted intragenic predictions with ChromHMM predictions with similar results (Additional file 1: Figure S11). Accordingly, we only kept windows predicted with posterior probability > 0.95, resulting in 42,297 and 55,624 active intragenic enhancer windows in K562 and GM12878, respectively. After clustering overlapping windows, we obtained 17,791 active intragenic enhancers in K562 (relative to GM12878) and 21,108 active intragenic enhancers in GM12878 (relative to K562), falling inside a total of 5162 genes (10.11 % of all genes) and 5933 (11.61 %) genes, respectively. The mean length of these predictions is 3665 bp, with the majority (82.81 %) being shorter than 5 kb (Additional file 1: Figure S12). As before, we kept those shorter than 5 kb, resulting in 11,055 and 11,917 candidate active intragenic enhancers in K562 and GM12878, respectively.

Our predicted intragenic enhancers tend to occur in separate genes, with only 29.2 % of the genes hosting enhancers of both types. The majority of intragenic enhancers active in K562 (78.24 %) or active in GM12878 (80.61 %) fall in intronic regions, and 26.02 % in K562 (22.07 % in Gm12878) overlap at least partially with an exon. However, comparing the proportion of exonic and intronic regions covered by enhancers with the actual proportions of these regions in genes, we find no preference for exons or introns (Additional file 2). Additionally, even though we observed a preference for intragenic enhancers to fall on the first intron (Additional file 1: Figure S13), this effect can be explained by the fact that first introns are on average longer in human (Additional file 2) [42].

Intragenic enhancers affect the regulation of the host gene

As the activation or silencing of intragenic transcriptional enhancers are characterized by the differential density of chromatin marks, we hypothesized that this would lead to a modulation of the RNA processing of the host gene. To test this, we first measured whether genes hosting predicted enhancers tend to show significant differential expression between the two cell lines. Similarly as before for enhancers linked to genes, we find a correlation of the relative expression change of genes hosting active or silenced enhancers. Specifically, 23.8 % of 5162 genes with only active enhancers in K562 (34.5 % of the 5933 genes with only active enhancers in GM12878) show a significant expression up-regulation in the corresponding cell line (Methods). We then tested whether the activation or silencing of internal enhancers may produce the activation or repression of an intragenic TSS. We considered all active and silenced enhancers that fall between the most upstream TSS (TSS1) and the first internal annotated TSS (TSS2), such that the distance TSS1-TSS2 was longer than 20 kb. This resulted in a total of 870 TSS1-TSS2 pairs, from which 113 (13 %) had at least one active enhancer in K562 and 135 (15.52 %) had at least one silent enhancer in K562 (active in GM12878) located between both TSSs. When an active enhancer is present between the two alternative TSSs, we observed that generally both TSS1 and TSS2 show an increase in RNAPII density in K562 relative to GM12878 (Fig. 3a). This suggests that activation of an intragenic enhancer can affect both TSSs. Conversely, when a silent enhancer is present between both TSSs, the relative level of RNAPII tend to decrease at both TSSs relative to the other cell line (Fig. 3b). Interestingly, this effect persists for other downstream alternative TSS events (Additional file 1: Figure S14), indicating that intragenic enhancers can activate internal TSSs, but also affect transcription of the most upstream TSS to some extent. We further used ChIA-PET for

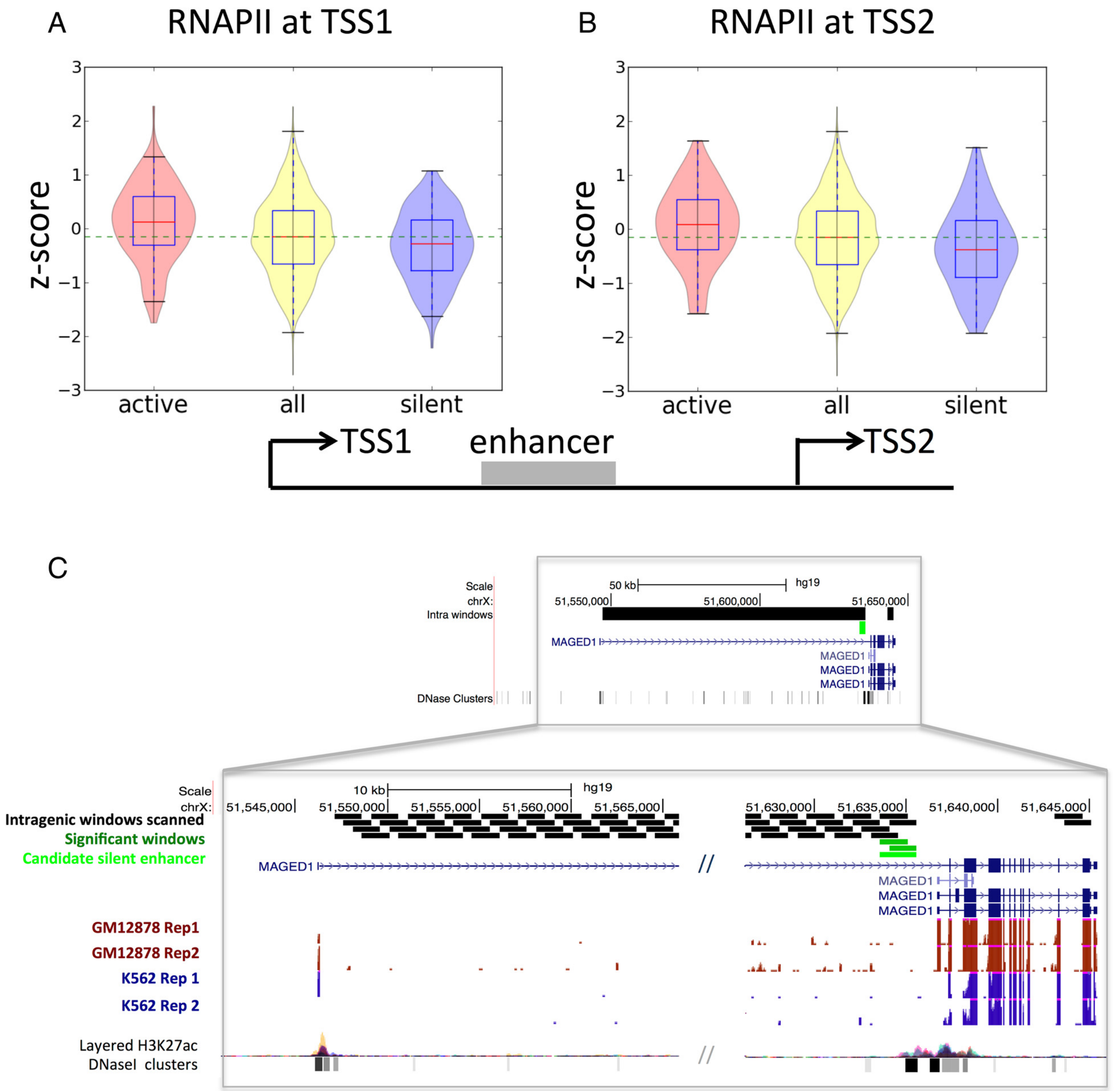

Fig. 3 Intragenic enhancers and differential transcript expression. Relative enrichment z-scores of RNAPII on the most upstream TSS (TSS1) (**a**) and the second TSS (TSS2) (**b**) when there is an active (left red violin plots) or silent (right blue violin plots) enhancer between both TSSs and it sits at a minimum distance of 1000 nt from either one. The yellow violin plots in the middle represent the z-score distribution of all TSS1-TSS2 pairs, with or without predicted enhancers and the dashed green line represents the median value of this distribution. **c** A TSS1-TSS2 event in gene MAGED1 with a silent enhancer in K562 close to TSS2 (light green box). The black boxes indicate the intragenic 1500 nt windows analyzed to search for enhancers. Dark green boxes indicate the significant windows and light green indicates the candidate silent enhancer. RNA-Seq data from K562 and GM12878 cell lines (2 replicates each) is shown below, together with the H3K27ac and DNase I hypersensitive sites tracks from ENCODE. The long intron separating the two TSSs is cut in the middle for simplicity

RNAPII in K562 to validate a possible direct interaction between our intragenic enhancers and the first TSS of each gene. Similarly as before, we observe a higher density of ChIA-PET links for active enhancers than for silent and random ones (Additional file 1: Figure S15). In this case 54.19 % of the active intragenic enhancers have ChIA-PET links, compared to 36.73 % in silent, 23.5 % in random active and 17.5 % in random silent (Additional file 1: Figure S15). This enrichment with respect to intergenic enhancers could be due to a higher density of RNAPII sites in intragenic regions. As an example of the described mechanism, we show the example of the gene MAGED1, a member of the melanoma antigen family D, which is known to have tumor-suppressor properties [43]. We predict an enhancer that is silent in K562 and active in GM12878, and is located between a distant TSS and an alternative downstream TSS (Fig. 3c). The activation of this enhancer co-occurs with the expression of the downstream first exon in GM12878 cells, whereas the silencing of the enhancer co-occurs with lack of expression of this exon in K562 cells (Fig. 3c). The RNA-Seq data suggests that the activation of this enhancer affects more strongly the usage of the TSS that is downstream.

The change in chromatin state induced by the activation or silencing of an enhancer may affect the processing of the pre-mRNA. There is evidence that localized intragenic chromatin states can produce changes in alternative splicing through various mechanisms [44–48]. Moreover, we have recently shown that active enhancers recruit Argonaute proteins to regulate the splicing of the host gene [18]. We therefore hypothesized that intragenic enhancers that are active in a cell line relative to the other one may in general be associated with relative differences in the inclusion level of nearby exons relative to the two same cell lines. To test this, we measured for all genes the variation in splicing between K562 and GM12878 using cytosolic and nuclear RNA-Seq polyA+ data from ENCODE, using only junction-reads, to avoid contributions from RNAs stemming from overlapping enhancers (Methods). We found that around 4 % of multi-exonic genes with intragenic enhancers that are active in either cell line, have a regulated alternative splicing event ($|\text{delta PSI}| > 0.1$) between both cell lines, whereas only about 1 % of all the genes without intragenic enhancers have a calculated alternative splicing event that changes between cell lines (Additional file 1: Table S2). A total of 3732 and 3908 genes with no change in expression have active or silent enhancers, respectively, and 1527 of these genes have enhancers of both types (Fig. 4a). Moreover, 325 of the genes with no change in expression have regulated events, and overlap with a total of 1046 enhancers (480 active and 566 silent) (Fig. 4a). Moreover, these genes contain 347 of the 535 (65 %) cassette events regulated between K562 and GM12878 (available as Additional file 2). Using Gorilla [49], we tested whether genes with enhancers and regulated events were enriched for any particular Gene Ontology term, and found an overrepresentation of genes encoding DNA-binding proteins implicated in gene regulation and chromatin organization (Fig. 4b) (Additional file 1: Table S3).

We next decided to evaluate whether there is any association between the presence of enhancers and regulated events in genes. To this end, we compared only genes that have one or more of the 5319 calculated alternative splicing events (Methods) and separated these genes according to whether they have one or more regulated events ($|\text{delta PSI}| > 0.1$) or not ($|\text{delta PSI}| < 0.05$) between the two cell lines. We found that in all comparisons the proportion of genes with regulated events was higher for those genes that have active enhancers (either in K562 or GM12878) (Additional file 1: Table S4), being the comparison statistically significant (Fisher $p\text{-value} < 0.05$) for both replicates for genes with active enhancers in GM12878, using nuclear RNA-Seq for the calculation of PSI values; whereas the same association for enhancers active in K562 was only significant for one of the replicate comparisons (Fisher $p\text{-value} = 0.01$) (Additional file 1: Table S4). Moreover, these associations remained significant when we considered only those genes that do not change expression between both cell lines (Additional file 1: Table S5). The regulated events in genes with active or silent intragenic enhancers present equal proportions of each pattern of PSI change, i.e. increase or decrease PSI (Additional file 1: Figure S16), which is consistent with the observed dual effect that a chromatin change can have on splicing [50]. Additionally, the direction of change of PSI does not correlate with the position, upstream or downstream, of the enhancer relative to the regulated exon (Additional file 1: Figure S17). Remarkably, the majority of the regulated events are located 5000 nt from an enhancer (Additional file 1: Figure S18). However, we did not find any significant difference with the distribution of distances of non-regulated events to nearby enhancers (Additional file 1: Figure S19).

As an example, we show the case of a regulated exon in the microtubule-actin crosslinking factor 1 gene (MACF1) (Fig. 4c). We observe a cassette exon with increased inclusion ($\text{delta PSI} = 0.72$) in K562 cells. The regulated exon is flanked by two enhancers predicted to be active in K562, one of which shows binding of PU.1 in K562, but not in GM12878 (Fig. 4c). This, together with the rest of our findings, suggests that the binding of PU.1 to a nearby enhancer, possibly in combination with other factors, could control the inclusion of this exon in MACF1. MACF1 has been implicated in the

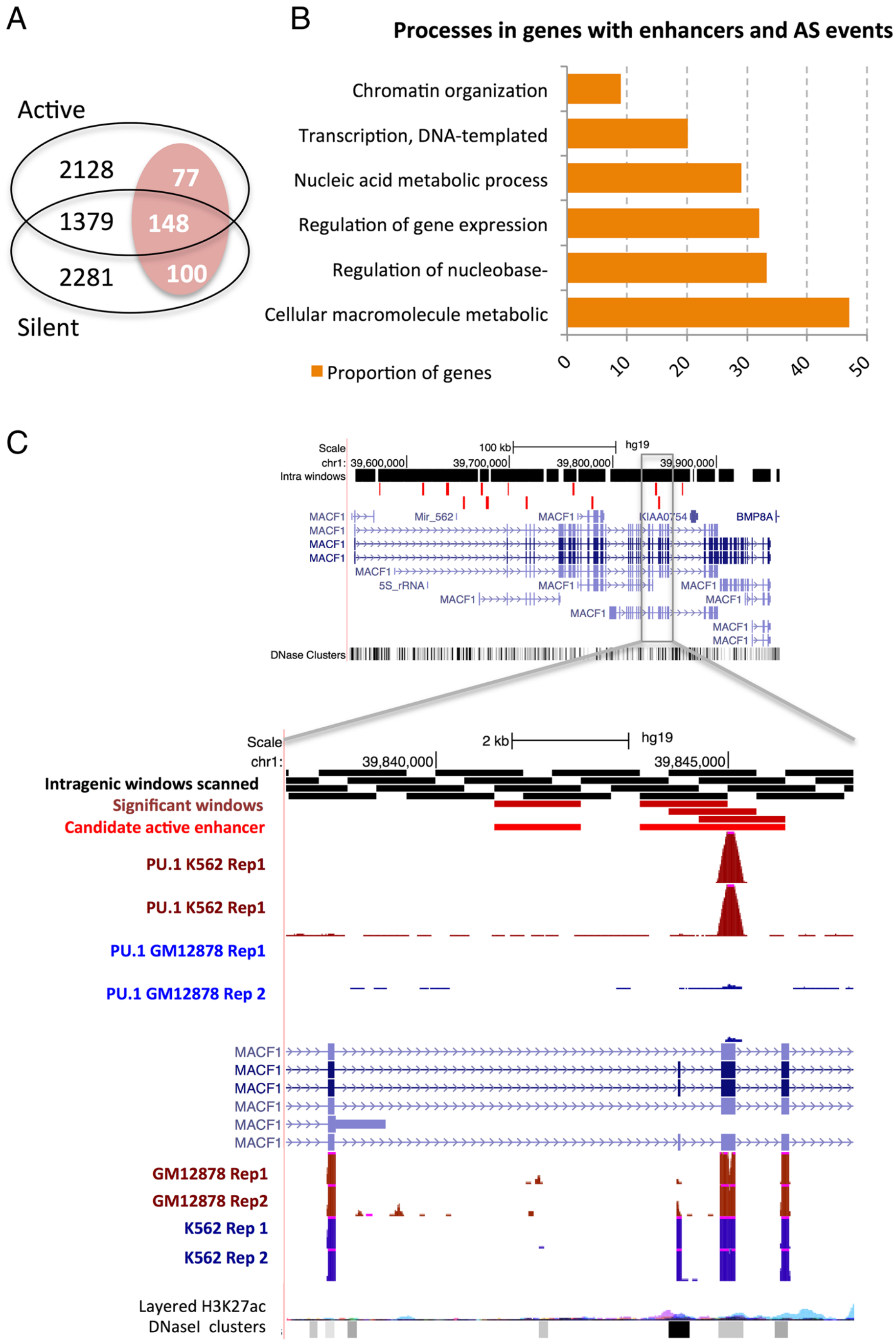

Fig. 4 Effect of intragenic enhancers on splicing. **a** From the 2205 genes found to have active enhancers, 77 of them have regulated events. From the 1527 genes with active and silent enhancers, 148 have regulated events, and from 2381 genes with silent enhancers, 100 have regulated events. **b** Enriched Gene Ontology processes in the genes with active or silent enhancers and with regulated events, compared to genes with intragenic enhancers but no regulated events (Additional file 1: Table S3). **c** Example of a regulated alternative splicing event in the gene MACF1. The cassette exon shows increased inclusion in K562 with delta-PSI = 0.72 (PSI = 0.86 in K562 and PSI = 0.14 in GM12878). Two active enhancers are predicted in K562 in the area of the regulated exon (light red boxes). Intragenic scanned windows are indicated as black boxes, and significant windows are dark red boxes. One of the predicted enhancer regions shows binding of PU.1 in K562 (2 replicates) but no biding in GM12878 (2 replicates). RNA-Seq data from K562 and GM12878 cell lines (2 replicates each) is shown below, together with the H3K27ac and DNase I hypersensitive sites tracks from ENCODE

Wnt signalling pathway [51] and the inclusion of a cassette exon in MACF1 was observed before to be associated to lung adenocarcinoma [52]. This result suggests the interesting possibility that the binding of PU.1 to an enhancer inside the MACF1 gene may affect its splicing, thereby altering Wnt signaling and contributing to the oncogenic transformation associated to PU.1 [53]. In conclusion, we have found a possible association between the activity of intragenic enhancers and the regulation of the pre-mRNA. In particular, we find evidence that the activation of intragenic enhancers, besides affecting the activity of internal TSSs, can also potentially influence the inclusion of nearby exons.

Conclusions

We have developed a novel semi-supervised method that exploits the relative enrichment and depletion of multiple signals from ChIP-Seq experiments to predict enhancers that are active in one cell line relative to another. Applying this method to ENCODE data we predicted a total of 21,420 enhancers that are active in K562 relative to GM12878 (silent in GM12878 cells) and 21,694 enhancers that are active GM12878 relative to K562 (silent in K562), including intragenic and intergenic enhancers.

The number of active enhancers is cell type specific and very much dependent on the method used to detect them [36]. Although activation of enhancers is generally associated to a number of histone modifications, only a small fraction of the many candidate enhancers previously identified using a variety of techniques may be active in a given cell. For instance, Heintzman et al. found 24,566 putative enhancers in K562 cells with approximately 20 % of them overlapping putative enhancers detected in HeLa cells [1]. In contrast, ChromHMM [38] predicts more than 60,000 non-abutting genomic regions to be strong enhancers and about three times as many for weak enhancers. There are two main reasons for the discrepancies with our predicted number of active enhancers: the resolution of the genome segmentation is very different and we only predict enhancers that are active in one condition but not in the other. That is, we do not detect enhancers active or silent in both cell types. Nonetheless, we found a good agreement between the regions we predict as active or silent enhancers and the annotations from ChromHMM for the same cell lines.

Our predicted enhancers are H3K27ac dependent and are defined almost entirely by chromatin signals. The relevant predictive features confirm that active enhancers are characterized not only by the presence of H3K4me1, but also by the presence of H3K27ac, H3K4me3 and RNAPII [4, 5, 12, 13]. We also observed that active enhancers show an enrichment of the histone variant H2A.Z, which has been identified to demarcate regulatory regions [35]. In contrast, CTCF and EZH2 and the histone marks H3K36me3 and H4K20me1 do not seem to play any prominent role in enhancer activation. H3K27me3 is the only feature that shows a pattern of depletion in active enhancers and enrichment in silent enhancers, but mainly in long enhancer-like regions (data not shown), which may be related to other regulatory mechanisms. We additionally found that predicted enhancer activity correlates strongly with production of long nuclear RNAs, rather than short ones, which can be polyA+ as well as polyA-. However, we observe that not all active enhancers produce eRNAs. Furthermore, although RNAPII and H3K36me3 have been detected on enhancers in relation to eRNA production [12, 13], we did not find them as strong predictors of enhancer activity.

When we applied the same predictive model to predict intragenic enhancers, we found a similar number of active intragenic enhancers as for intergenic ones. This result suggests that there exist in cells a considerable number of differentially activated intragenic enhancers, which may have a relevant contribution to the mechanisms of cell-specific gene regulation. Since active enhancers are characterized by a local modification of the chromatin state, we hypothesized that our predictions could be linked to relative differences between the same two cell lines in expression and splicing. We observed that intragenic transcriptional enhancers, upon activation or silencing, affect the activity of downstream alternative transcription start sites. Surprisingly, they can also affect the most upstream TSS. This generalizes previous findings indicating that intragenic enhancers can act as internal alternative promoters [16]. We also found that intragenic enhancers, upon activation or silencing, associate to the differential inclusion of nearby exons. However, a considerable proportion of splicing changes occur in genes that change expression (51.1 % for genes containing differentially included exons in K562 and 52.9 % for differentially excluded exons in GM12878). This indicates that the main effect of the activation of enhancers may be related to the activation of alternative transcription in the gene and alternative splicing may be a byproduct of that. The observed changes may be mediated by the changes in the RNAPII elongation produced due to the chromatin change. However, active intragenic enhancers show enrichment in open chromatin marks (H3K4me3, H3K27ac) that have not been associated before to changes in RNAPII elongation.

On the other hand, we found here a strong association of PU.1 (SPI1) to active enhancers in K562 cells and in particular, a significant increase in PU.1 occupancy in 26.8 % of active enhancers. PU.1 has been shown before to be an essential co-factor for enhancer activity [54] and is known to bind to H3K4me1 sites in macrophages and B cells in a cell-specific manner [55, 56]. Moreover, PU.1 has been observed to regulate alternative splicing from the promoter [57] and can interact with the RNA binding proteins FUS (TLS) and NONO (p54nrb) [58, 59]. In fact, PU.1 has been proposed to bind RNA [59] and to perform an antagonistic function to the RNA binding proteins TLS and NONO in the regulation of splicing [59, 60]. In this direction, we found enrichment of regulated events in genes with enhancers, which suggests that PU.1 could be regulating the splicing of some of these genes through its binding to intragenic enhancers, possibly interacting with RNA binding proteins [46]. In support of this model, we find that there is an enrichment of regulated events in genes with enhancers that are active or silent relative to the other cell line. We postulate that intragenic enhancers provide localized and cell-type specific mechanisms to link the chromatin state to RNA processing.

In summary, there is increasing evidence that changes in the chromatin state can affect the processing of the pre-mRNA [44–48, 61–65] and different models for this regulation have been proposed. From our analysis a picture emerges whereby localized chromatin changes inside genes can take place by means of the activation of intragenic transcriptional enhancers. We propose that the differential activation and silencing of transcriptional enhancers that fall within genes could explain the localized chromatin variation that have been observed before to affect the expression and splicing of genes, either through the modulation of RNAPII activity or through the recruitment of factors that can interfere with RNA processing, like PU.1.

Additional files

Additional file 1: Contains the supplementary figures (S1-S19) and tables (S1-S5) cited in the text.

Additional file 2: List of predicted active and silent enhancers in K562 inside genes overlapping differentially spliced events. Contains the coordinates of the enhancers and the events, as well as the inclusion values (PSI) of the events in K562 and GM12878 cells.

Abbreviations

ChIP: Chromatin Immunoprecipitation; Seq: Sequencing; CTCF: CCCTC-binding factor; RNAPII: RNA Polymerase II; H3K36me3: Histone 3 Lysine 36 tri-methylation; H3K9me2: Histone 3 Lysine 9 de-methylation; H3K27me3: Histone 3 Lysine 27 tri-methylation; H3K27Ac: Histone 3 Lysine 27 acetylation; H3K4me1/2/3: Histone 3 Lysine 4 mono/di/tri methylation; H3K79me3: Histone 3 Lysine 79 tri-methylation; H4K20me1: Histone 4 Lysine 20 mono-methylation; H3K9Ac: Histone 3 Lysine 9 acetylation; H2A.Z: H2A Histone family member Z; STAT1: Signal transducer and activator of transcription 1; polyA: Poly-adenylation; BIC: Bayesian Information Criterion; TSS: Transcription start site; ChIA-PET: Chromatin interaction analysis by paired-end tag sequencing; PSI: Percent spliced-in.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EE designed and supervised the study; JGV built the model for enhancers and carried out the benchmarking analyses; AP and BS carried out the splicing and expression analyses, respectively; EE and JGV wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank E. Furlong, Y. Barash, B. Blencowe and U. Braunschweig for useful discussions. This work was supported by grants from Plan Nacional I + D (BIO2011-23920) and Consolider (CSD2009-00080) from MINECO (Spanish Government), and by the Sandra Ibarra Foundation for Cancer (FSI 2013). JGV and BS were supported FPI grants from the MINECO (Spanish Government) BES-2009-018064 and BES-2012-052683, respectively.

Received: 3 December 2014 Accepted: 15 June 2015
Published online: 14 July 2015

References

1. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459(7243):108–12.
2. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39(3):311–8.
3. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, HannaJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. PNAS. 2010;21931–6. doi:10.1073/pnas.1016071107.
4. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann S a, Flynn R a, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011;470(7333):279–83.
5. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet. 2012;44(2):148–56.
6. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009;461(7261):199–205.
7. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama J a, Holt A, Plajzer-Frick I, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 2010;42(9):806–10.
8. Maston GA, Landt SG, Snyder M, Green MR. Characterization of enhancer function from genome-wide analyses. Annu Rev Genomics Hum Genet. 2012;13:29–57.
9. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132(2):311–22.
10. Lupien M, Eeckhoute J, Meyer C a, Wang Q, Zhang Y, Li W, Carroll JS, et al. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell. 2008;132(6):958–70.
11. Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, et al. H3K4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J. 2011;30(20):4198–210.
12. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010;8(5):e1000384.
13. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465(7295):182–7. Nature Publishing Group. doi:10.1038/nature09033.
14. Ritter DI, Dong Z, Guo S, Chuang JH. Transcriptional enhancers in protein-coding exons of vertebrate developmental genes. PLoS One. 2012;7(5), e35202.
15. Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, et al. Coding exons function as tissue-specific enhancers of nearby genes. Genome Res. 2012;1059–1068.
16. Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe J a, Sloane-Stanley J a, McGowan SJ, et al. Intragenic enhancers act as alternative promoters. Mol Cell. 2012;45(4):447–58.
17. Kadener S, Fededa JP, Rosbash M, Kornblihtt AR. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. Proc Natl Acad Sci U S A. 2002;99(12):8185–90.
18. Alló M, Agirre E, Bessonov S, Bertucci P, Gómez Acuña L, Buggiano V, Bellora N, Singh B, Petrillo E, Blaustein M, Miñana B, Dujardin G, Pozzi B,

Pelisch F, Bechara E, Agafonov DE, Srebrow A, Lührmann R, Valcárcel J, Eyras E, Kornblihtt AR. Argonaute-1 binds transcriptional enhancers and controls constitutive and alternative splicing in human cells. Proc Natl Acad Sci U S A. 2014;111(44):15622–9.

19. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35(Database issue):D88–92.
20. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
21. Pennacchio LA, Visel A. Limits of sequence and functional conservation. Nat Genet. 2010;42(7):557–8.
22. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics. 2011;12(1):480.
23. Althammer S, González-Vallinas J, Ballaré C, Beato M, Eyras E. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics. 2011;27(24):3333–40.
24. Kursa MB, Jankowski A, Rudnicki WR. Boruta–a system for feature selection. Fundamenta Informaticae. 2010;101(4):271–85.
25. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.
26. Balakrishnan L, Milavetz B. Decoding the histone H4 lysine 20 methylation mark. Crit Rev Biochem Mol Biol. 2010;45(5):440–52.
27. Beck DB, Oda H, Shen SS, Reinberg D. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. Genes Dev. 2012;26(4):325–37.
28. Fraley C, Raftery A. Mclust version 3 for R: normal mixture modeling and model-based clustering. 2007.
29. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
30. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177–83.
31. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74.
32. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5.
33. Pervouchine DD, Knowles DG, Guigó R. Intron-centric estimation of alternative splicing from RNA-seq data. Bioinformatics. 2013;29(2):273–4.
34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
35. Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nat Genet. 2009;41(8):941–5.
36. Zentner GE, Scacheri PC. The chromatin fingerprint of gene enhancer elements. J Biol Chem. 2012;287(37):30888–96.
37. Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011;12(4):283–93.
38. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–9.
39. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80.
40. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153(2):307–19.
41. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012;148(1–2):84–98.
42. Bradnam KR, Korf I. Longer first introns are a general property of eukaryotic gene structure. PLoS One. 2008;3(8), e3093.
43. Du Q, Zhang Y, Tian XX, Li Y, Fang WG. MAGE-D1 inhibits proliferation, migration and invasion of human breast cancer cells. Oncol Rep. 2009;22(3):659–65.
44. Sims 3rd RJ, Millhouse S, Chen CF, Lewis BA, Erdjument-Bromage H, Tempst P, Manley JL, Reinberg D. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell. 2007;28(4):665–76.
45. Schor IE, Rascovan N, Pelisch F, Allo M, Kornblihtt AR. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc Natl Acad Sci U S A. 2009;106:4325–30.
46. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010;327(5968):996–1000.
47. Saint-André V, Batsché E, Rachez C, Muchardt C. Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat Struct Mol Biol. 2011;18(3):337–44.
48. Zhou HL, Hinman MN, Barron VA, Geng C, Zhou G, Luo G, Siegel RE, Lou H. Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. Proc Natl Acad Sci U S A. 2011;108(36):E627–35.
49. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.
50. Dujardin G, Lafaille C, de la Mata M, Marasco LE, Muñoz MJ, Le Jossic-Corcos C, Corcos L, Kornblihtt AR. How slow RNA polymerase II elongation favors alternative exon skipping. Mol Cell. 2014;54(4):683–90.
51. Saadeddin A, Babaei-Jadidi R, Spencer-Dene B, Nateri AS. The links between transcription, beta-catenin/JNK signaling, and carcino- genesis. Mol Cancer Res. 2009;7:1189–96.
52. Misquitta-Ali CM, Cheng E, O'Hanlon D, Liu N, McGlade CJ, Tsao MS, Blencowe BJ. Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. Mol Cell Biol. 2011;31(1):138–50.
53. Moreau-Gachelin F, Tavitian A, Tambourin P. Spi-1 is a putative oncogene in virally induced murine erythroleukaemias. Nature. 1988;331:277–80.
54. Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, Gregory L, Lonie L, Chew A, Wei CL, Ragoussis J, Natoli G. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity. 2010;32(3):317–28.
55. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, and Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89.
56. Jin F, Li Y, Ren B, Natarajan R. Enhancers: multi-dimensional signal integrators. Transcription. 2011;2(5):226–30.
57. Guillouf C, Gallais I, Moreau-Gachelin F. Spi-1/PU.1 oncoprotein affects splicing decisions in a promoter binding-dependent manner. J Biol Chem. 2006;281(28):19145–55.
58. Hallier M, Lerga A, Barnache S, Tavitian A, Moreau-Gachelin F. The transcription factor Spi-1/PU. 1 interacts with the potential splicing factor TLS. J Biol Chem. 1998;273(9):4838–42.
59. Hallier M, Tavitian A, Moreau-Gachelin F. The transcription factor Spi-1/PU.1 binds RNA and interferes with the RNA-binding protein p54nrb. J Biol Chem. 1996;271(19):11177–81.
60. Delva L, Gallais I, Guillouf C, Denis N, Orvain C, Moreau-Gachelin F. Multiple functional domains of the oncoproteins Spi-1/PU.1 and TLS are involved in their opposite splicing effects in erythroleukemic cells. Oncogene. 2004;23(25):4389–99.
61. Gunderson FQ, Johnson TL. Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. PLoS Genet. 2009;5(10):e1000682.
62. Enroth S, Bornelöv S, Wadelius C, Komorowski J. Combinations of histone modifications mark exon inclusion levels. PLoS One. 2012;7(1), e29911.
63. Zhou Y, Lu Y, Tian W. Epigenetic features are significantly associated with alternative splicing. BMC Genomics. 2012;13(1):123.
64. Shindo Y, Nozaki T, Saito R, Tomita M. Computational analysis of associations between alternative splicing and histone modifications. FEBS Lett. 2013;587(5):516–21.
65. Ye Z, Chen Z, Lan X, Hara S, Sunkel B, Huang TH, Elnitski L, Wang Q, Jin VX. Computational analysis reveals a correlation of exon-skipping events with splicing, transcription, and epigenetic factors. Nucleic Acids Res. 2014;42(5):2856–69.