5 Epigenetic regulation of RNA processing

Many novel and previously known non-coding RNA species are characterized in ENCODE

Many of the ENCODE assays directly or indirectly provide information about the action of promoters. Focusing on the TSSs of protein-coding transcripts, we investigated the relationships among different ENCODE assays, in particular testing the hypothesis that RNA expression ("output") can be effectively predicted from patterns of chromatin modifications or TF binding ("input"). Consistent with previous reports[45], we observe two relatively distinct types of promoters: (1) broad, mainly C+G rich, TATA-less promoters; and (2) narrow, TATA-box-containing promoters. These promoters have distinct patterns of histone modifications, and TF-binding sites are selectively enriched in each class (Supplementary Figure Z1).

We developed predictive models to explore the interaction between histone modifications and measures of transcription at promoters, distinguishing between modifications known to be added as a consequence of transcription (such as H3K36me3 and H3K79me2) and other categories of histone marks[59]. In our analyses, the best models had two components: an initial classification component (on/off) and a second quantitative model component. Our models showed activating acetylation marks (H3K27ac and H3K9ac) are roughly as informative as activating methylation marks (H3K4me3 and H3K4me2) (Figure 2A). Although repressive marks, such as H3K27me3 or H3K9me3, show negative correlation both individually and in the model, removing these marks produces only a small reduction in model performance. However, for a subset of promoters in each cell line repressive histone marks (H3K27me3 or H3K9me3) must be used to accurately predict their expression. We also examined the interplay between the H3K79me2 and H3K36me3 marks, both of which mark gene bodies, likely reflecting recruitment of modification enzymes by polymerase isoforms. As described previously, H3K79me2 occurs preferentially at the 5' ends of gene bodies and H3K36me3 occurs more 3', and our analyses support the previous model in which the H3K79me2 to H3K36me3 transition occurs at the first 3' splice site[60].

We determined the relative importance of each feature for predicting expression datasets (see Materials and methods). We observed that histone modifications like H3K9ac and H3K4me3 are more important in identifying genes that are 'on' or 'off,' while histone modifications like H3K79me2, H3K36me3 are more important for regression of expressed genes (Figure 2B). DNase I hypersensitivity is the third important for both classification and regression. We also observed that the normalized CpG score is more important for gene 'on' or 'off' status classification than for regression of the expression levels of 'on' genes.

We summarized the correlation coefficients between predicted and measured expressions for all 78 RNA expression experiments from the seven cell lines in our analysis (Figure 2C). It shows that most experiments show a strong correlation (median $r = 0.83$) between predicted and measured expression levels by both TSS-based CAGE and RNA-PET and transcript-based RNA-Seq techniques.

To further explore whether the models are generalizable across different cell lines, we applied the model trained in one cell line to other cell lines, using the values of chromatin features in those cell lines as inputs to the models to determine if the prediction accuracy dramatically changed. Figure 4B shows an example of this cross-cell line prediction, wherein we learned a prediction model from CAGE-measured PolyA+ cytosolic RNA from K562 cells and applied it to CAGE-measured PolyA+ cytosolic RNA from four other cell lines. The prediction accuracy remains high, with $r = 0.82$, 0.86, 0.87, and 0.84 in GM12878, H1-hESC, HeLa-S3, and NHEK cell lines respectively. These results indicate that our models accurately captured the relationships among the various chromatin features and are broadly applicable to predicting expression in all cell lines.

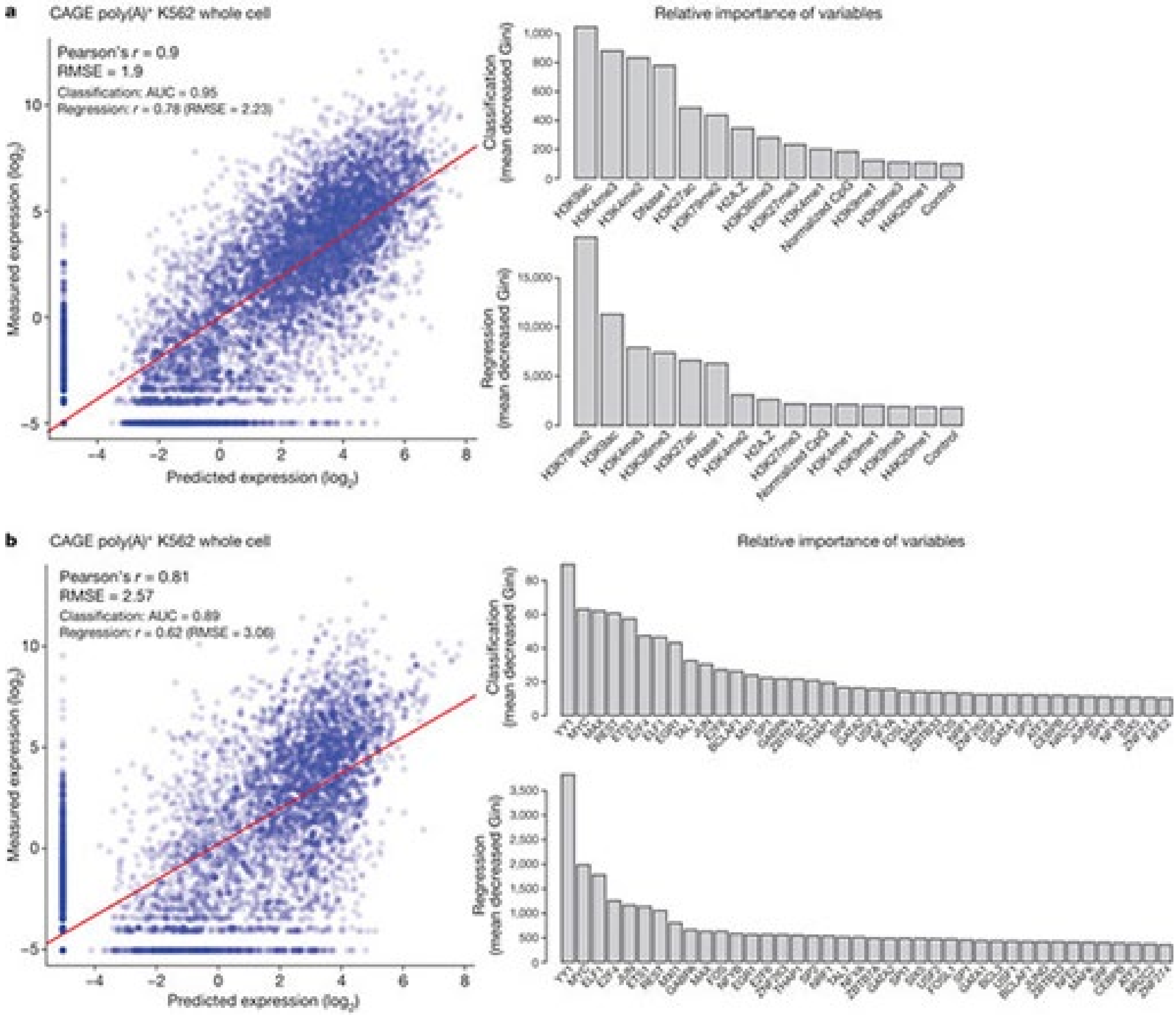

Figure 2 | Modelling transcription levels from histone modification and transcription-factor-binding patterns. (**a**,**b**) Correlative models between either histone modifications or transcription factors, respectively, and RNA production as measured by CAGE tag density at TSSs in K562 cells. In each case the scatter plot shows the output of the correlation models (*x* axis) compared to observed values (*y* axis). The bar graphs show the most important histone modifications (**a**) or transcription factors (**b**) in both the initial classification phase (top bar graph) or the quantitative regression phase (bottom bar graph), with larger values indicating increasing importance of the variable in the model. Further analysis of other cell lines and RNA measurement types is reported elsewhere[59,79]. AUC, area under curve; Gini, Gini coefficient; RMSE, root mean square error.

By comparing the prediction accuracy using marks from each category or a combination of two categories (Figure 5), we show that for CAGE TSS-based gene expression, promoter marks are the most predictive, while for RNA-Seq Tx-based expression data, structural marks are better predictors. For CAGE-measured PolyA+ cytosolic RNA, promoter marks as a group have high correlation coefficients (median r = 0.86). Promoter marks combined with another category of chromatin features give equally high prediction accuracy. However, non-promoter mark categories have lower prediction accuracy (e.g. median r = 0.84 for structural marks only; median r = 0.35 for repressive marks only). On the other hand, structural marks like H3K79me2 and H3K36me3 are more predictive for RNA-Seq expression data.

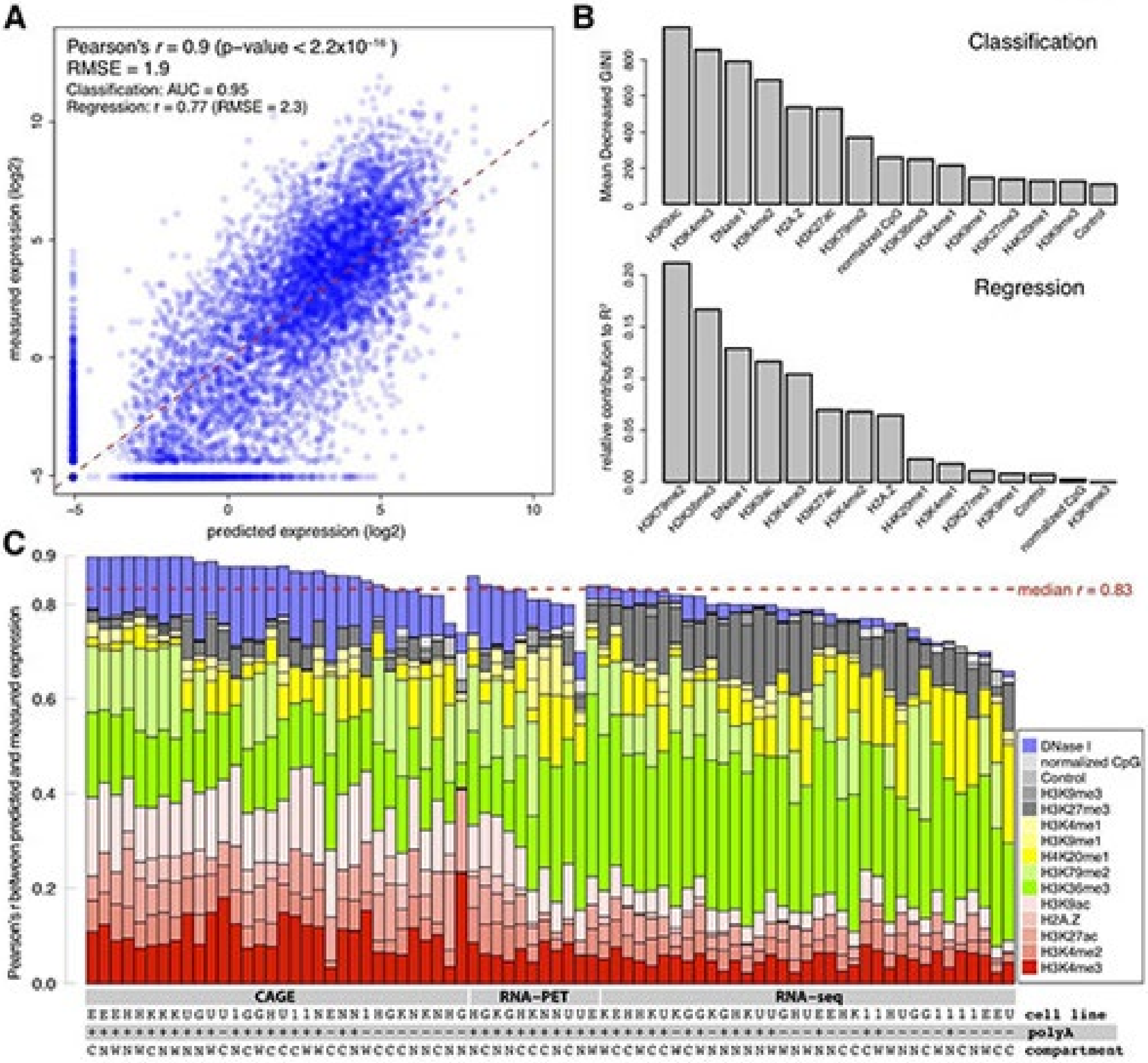

Figure 2 | Quantitative relationship between chromatin feature and expression. (a) Scatter plot of predicted expression values using the two-step prediction model (Random Forests classification model and linear regression model) vs. the measured PolyA+ cytosolic RNA from K562 cells measured by CAGE. Each blue dot represents one gene. The red dashed line indicates the linear fit between measured and predicted expression values, which are highly correlated (Pearson's correlation coefficient r = 0.9, p-value < 2.2×10-16), indicating a quantitative relationship between chromatin features and expression levels. The accuracy for the overall model is indicated by RMSE (root-mean-square error), which is 1.9. Accuracy for the classification model is indicated by AUC (area under the ROC curve), which is 0.95. The accuracy for regression model is r =0.77 (RMSE = 2.3). (b) The relative importance of chromatin features in the two-step model. The most important features for the classifier (upper panel) include H3K9ac, H3K4me3, and DNase I hypersensitivity, while the most important features for the regressor (bottom panel) include H3K79me2, H3K36me3, and DNase I hypersensitivity. (c) Summary of overall prediction accuracy on 78 expression experiments on whole cell, cytosolic or nuclear RNA from seven cell lines. The bars are sorted by correlation coefficient in decreasing order for each high throughput technique (CAGE, RNA-PET and RNA-Seq). Each bar is composed of several colors, corresponding to the relative contribution of each feature in the regression model. The red dashed line represents median Pearson's r = 0.83. Code for cell lines: K (K562), G (GM12878), 1 (H1-hESC), H (HepG2), E (HeLa-S3), N (NHEK) and U (HUVEC); Code for RNA extraction: + (PolyA+) and - (PolyA-); Code for cell compartment: W (whole cell), C (cytosol) and N (nucleus).

We then assess how well histone modifications can predict PolyA+ and PolyA- RNA levels. PolyA+ RNA is significantly better predicted than PolyA- RNA, regardless of the technique with which RNA levels are measured and the location from which the RNA molecules are extracted (Figure 7A and 7B), indicating that the PolyA- fraction might be

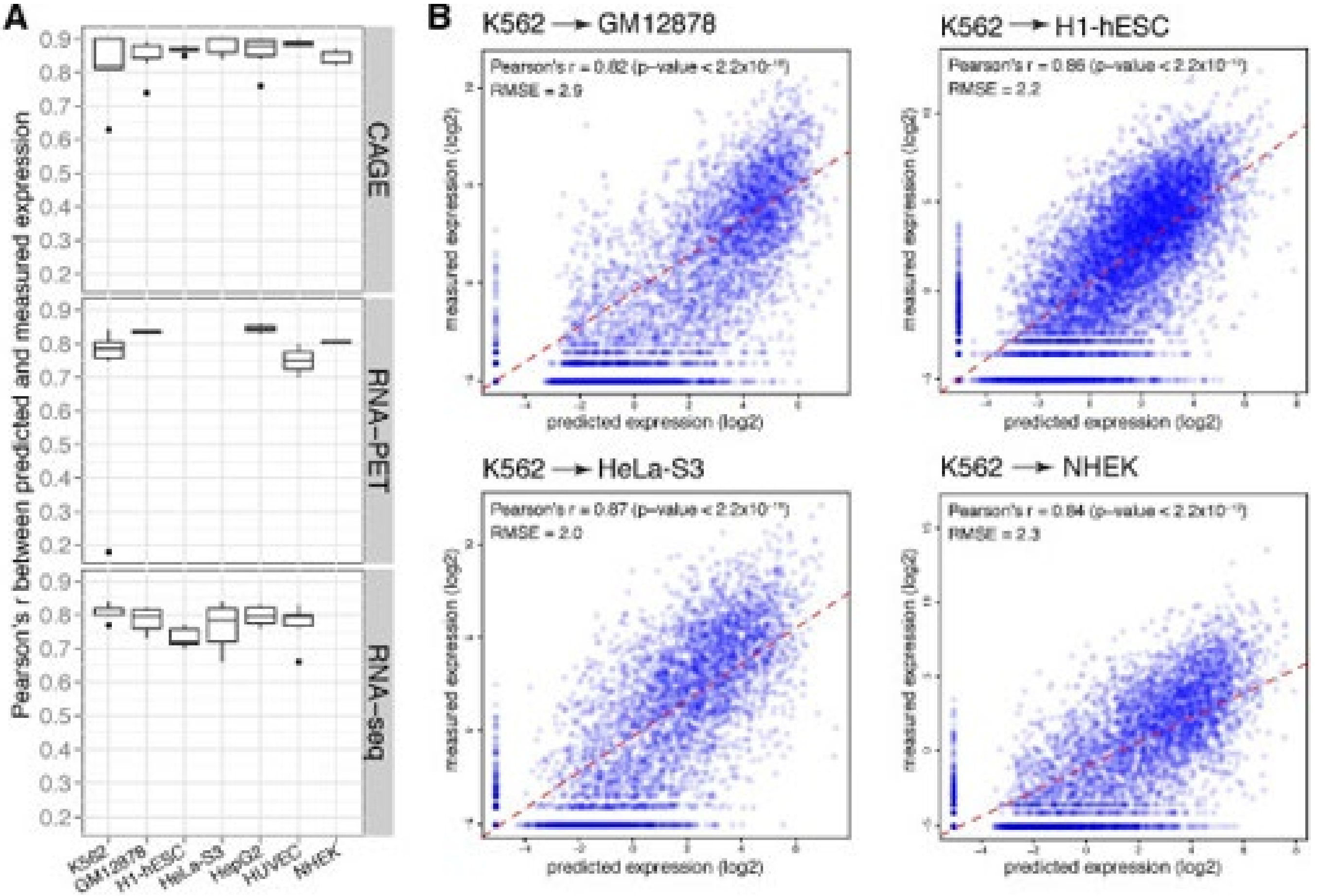

Figure 4 | Comparison of prediction accuracy across different cell lines. (a) Boxplot of correlation coefficients for seven cell lines (K562, GM12878, H1-hESC, HeLa-S3, HepG2, HUVEC and NHEK) in different types of expression quantification (CAGE, RNA-PET, and RNA-Seq). It shows that the high quantitative relationship between chromatin features and expression exist in various cell lines and different expression quantification methods. Paired wilcox tests between H1-hESC and other cell lines show that H1-hESC has significantly lower prediction accuracy (p-value = 0.02, 0.02, 0.07, 0.02, and 0.05 for K562, GM12878, HeLa-S3, HepG2 and HUVEC, respectively). (b) Application of the model learned from K562 to other cell lines (GM12878, H1-hESC, HeLa-S3 and NHEK) indicates that the model performs well across cell lines (r = 0.82, 0.86, 0.87 and 0.84, respectively). This indicates that the quantitative relationship between chromatin features and gene expression is not cell line-specific, but rather a general feature.

regulated by different mechanisms from the PolyA+ fraction. We also compared the performance for RNAs extracted from different compartments. The analysis based on RNA-Seq datasets showed that for polyadenylated RNAs (left panel of Figure 7B), cytosolic RNA is significantly better predicted than nuclear RNA (Paired Wilcox test p-value = 0.01) and the reverse is true for non-polyadenylated RNA (p-value = 0.03).

The interrogation of different subcellular RNA fractions provides snapshots of the status of the RNA population along the RNA processing pathway. Thus, by analysing short and long RNAs in the different subcellular compartments, we confirm that splicing predominantly occurs during transcription. By using RNA-seq to measure the degree of completion of splicing (Fig. 2a), we observed that around most exons, introns are already being spliced in chromatin-associated RNA-the fraction that includes RNAs in the process of being transcribed (Fig. 2b). Concomitantly, we found strong enrichment specifically of spliceosomal small nuclear RNAs (snRNAs) in this RNA fraction (see 'Short RNA expression landscape' below). Co-transcriptional splicing provides an explanation for the increasing evidence connecting chromatin structure to splicing regulation, and we have observed that exons in the process of being spliced are enriched in a number of chromatin marks[17,18].

We introduce a measure, based on the RNASeq reads mapping to the exon junctions, to assess the degree of completion of splicing around internal exons. We simply count the number of reads mapping across the exon

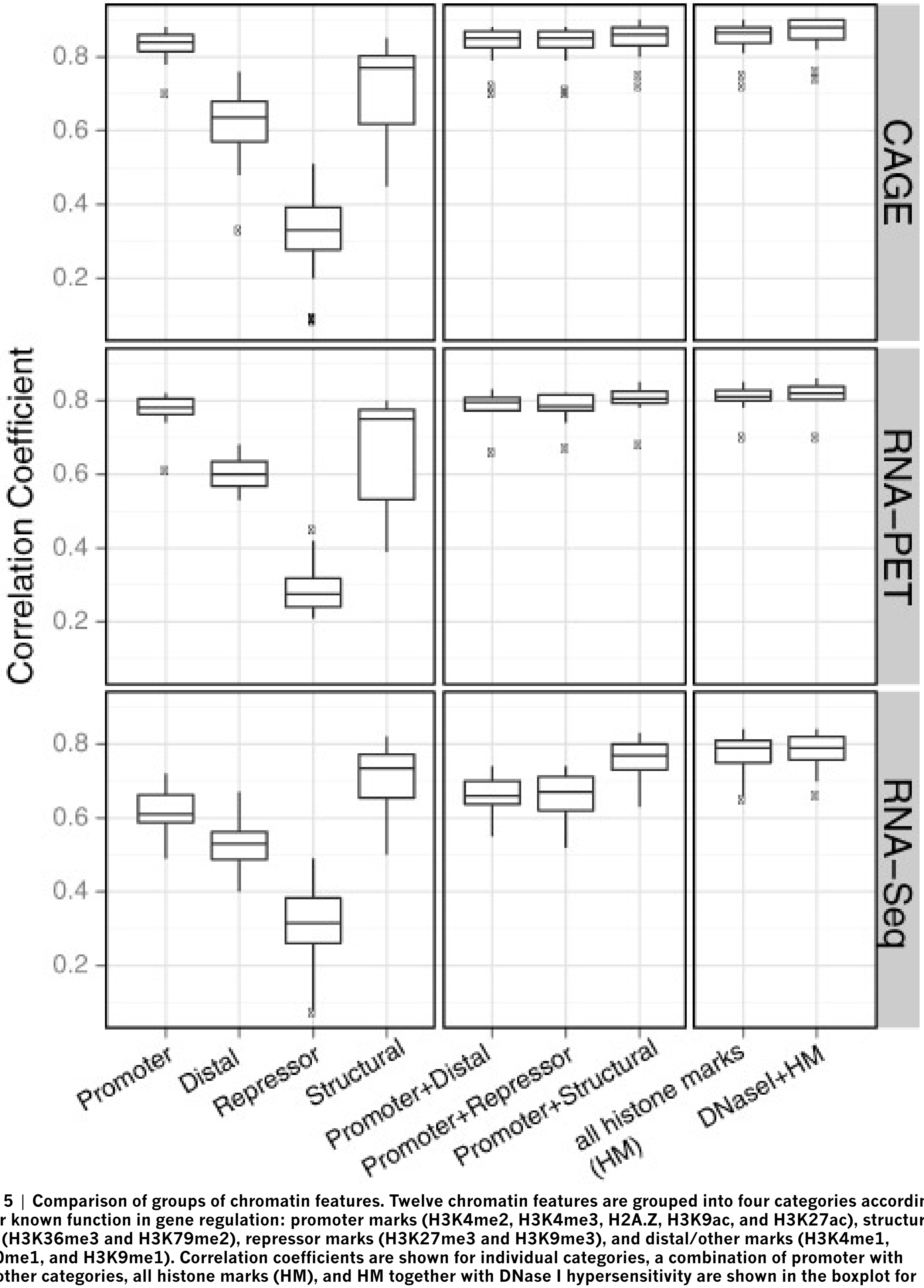

Figure 5 | Comparison of groups of chromatin features. Twelve chromatin features are grouped into four categories according to their known function in gene regulation: promoter marks (H3K4me2, H3K4me3, H2A.Z, H3K9ac, and H3K27ac), structural marks (H3K36me3 and H3K79me2), repressor marks (H3K27me3 and H3K9me3), and distal/other marks (H3K4me1, H4K20me1, and H3K9me1). Correlation coefficients are shown for individual categories, a combination of promoter with three other categories, all histone marks (HM), and HM together with DNase I hypersensitivity are shown in the boxplot for CAGE (TSS-based), RNA-PET (TSS-based), and RNA-Seq (Tx-based) expression data. It indicates that for TSS-based data, promoter marks are the most predictive among the four categories, while for Tx-based expression, structural marks are the most predictive.

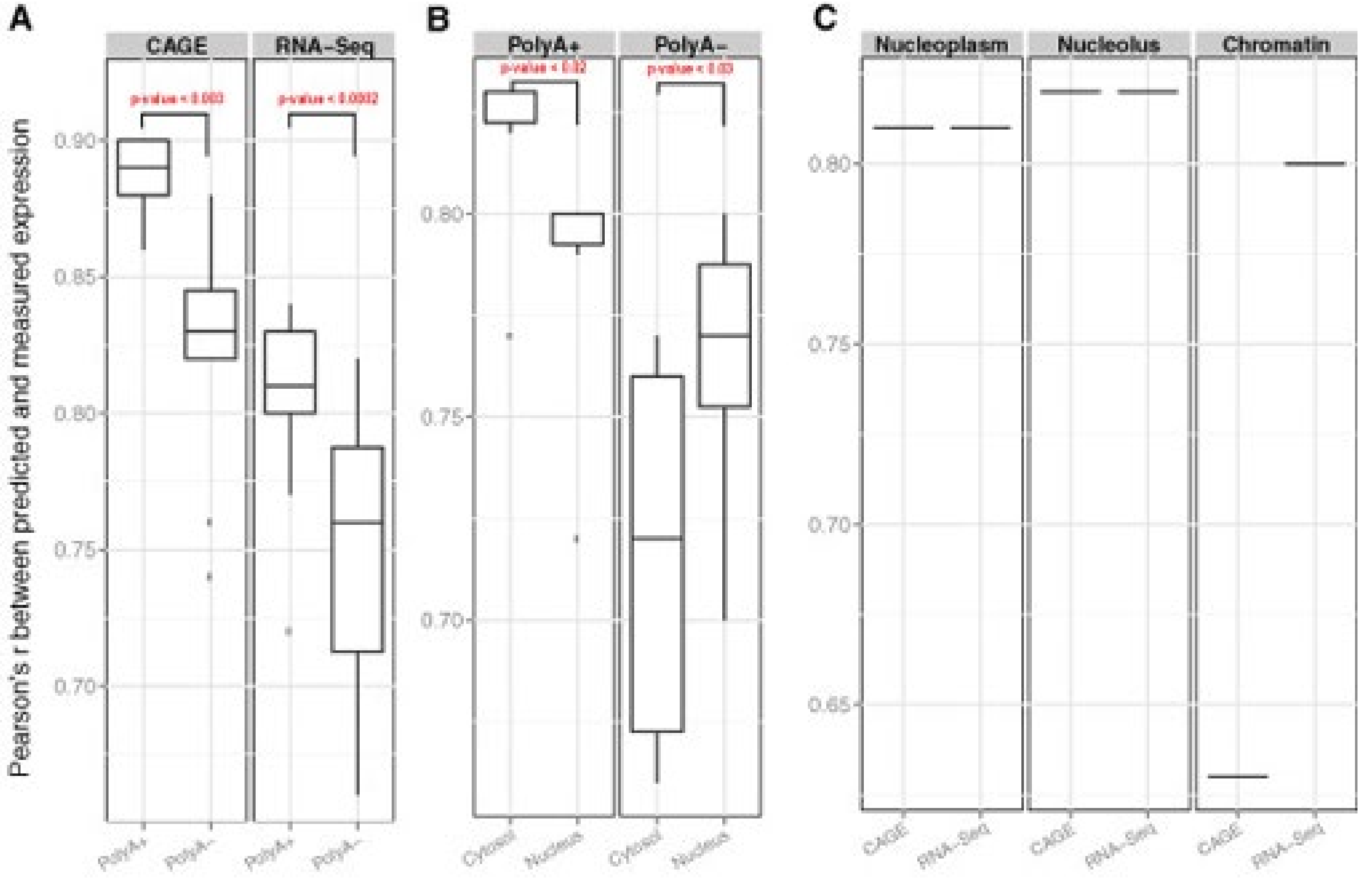

Figure 7 | Comparison of prediction accuracy among different RNA extractions and different cell compartments. (**a**) Prediction accuracy of PolyA+ and PolyA- RNA for allgenes measured with CAGE and RNA-Seq technique. This shows that PolyA+ RNA are better predicted than PolyA- RNA (p-value of paired Wilcox test between PolyA+ and PolyA-). (**b**) Prediction accuracy of PolyA+ and PolyA- RNA from different cell compartments for all genes measured with RNA-Seq technique (p-value of paired Wilcox test between cytosol and nuclues). (**c**) Prediction accuracy of total RNA indifferent nuclear sub-compartments, measured by CAGE or RNA-Seq.

boundaries into the adjacent intron sequence (which originate from primary, unspliced mRNA molecules), as well as the number of reads split-mapping across exon-exon junctions, either from the exon to another exon of the same gene, or between an upstream and a downstream exon (both types of read originating from a successfully completed splicing event, Figure 1b). Based on these numbers, we compute the completed Splicing Index (coSI) of a given exon, corresponding thereby to the weighted percentage of reads supporting splicing completion around the exon. The coSI value can be broadly assumed to correspond to the fraction of exon-containing RNA molecules in which splicing in the region around the exon has already been carried out. A coSI value of 1 means entirely completed splicing, while coSI=0 indicates that the exon is still completely included in the sequence of the primary transcript.

We have computed coSI scores for human internal exons (see Supplemental Methods) in all analyzed K562 RNA fractions (Supplemental Table S1).We have observed a higher correlation of coSI values ($R = 0.82$) between two replicates of the chromatin fraction than between the chromatin fraction and other fractions (R of between 0.35 and 0.71) (Supplemental Fig. S1A-G), confirming that overall coSI values are reproducible within a given experiment. Figure 2 shows the distribution of coSI scores in the different RNA fractions that we have interrogated. As expected, for most exons, splicing of the corresponding introns is fully completed in the cytosolic polyA+ fraction (92%of the exons have a $coSI > 0.95$), as well as the cytosolic polyA- fraction (data not shown). Of even more interest with respect to splicing is the polyA- nuclear fraction, in which the median coSI is 0.84. For 16% of the exons, their surrounding introns are completely spliced in this fraction, and only for a vanishing fraction ($<0.2\%$ with coSI < 0.05) do the corresponding introns remain completely unspliced. The polyA- nuclear fraction contains RNA molecules of three types: first, RNAs that are still being transcribed and for which transcription has not yet reached the polyA-site; second, RNAs that have been released from the transcribing Pol II, before it could reach the polyA-site; and, third, products of aborted transcription. The high degree of splicing completion in this fraction therefore suggests that splicing is mostly initiated before completion of transcription. Even more enriched for RNAs in the act of being transcribed is the chromatin-associated fraction. With a median coSI of 0.75 in this fraction, around most exons we see large amounts of completed splicing. For 5.6% of the exons, we see absolutely completely spliced introns (coSI > 0.95); however, as in the polyA- nuclear fraction, only a tiny fraction of exons ($<0.3\%$, coSI < 0.05) are surrounded by completely unspliced introns, further suggesting that splicing is intimately coupled and occurs almost simultaneously with transcription.

If splicing occurs mostly co-transcriptionally and therefore in proximity to the chromatin template, one would expect that RNAs of the splicing machinery would also reside in proximity to chromatin. We have therefore investigated the sub-cellular location of U1-U6 and U6atac (UxRNAs) based on RNAseq of small RNAs performed in five different sub-cellular locations (Nucleus, Cytosol, Nucleoplasm, Nucleoli and Chromatin, ENCODE RNA consortium, Djebali *et al.* 2012, Supplemental Methods). As predicted, all spliceosomal UxRNAs, that is U1, U2, U4, U5, U6 and U6atac, but not U3, are clearly enriched in the chromatin-associated fraction compared to the other fractions (Figure 4a,b,d-g). In contrast, U3 and snoRNAs (excluding U RNAs), both of which are thought not to be involved in splicing, were highly enriched in the nucleoli fraction (Figure 4c, h), as expected from their known functions.

A number of sequence features characterizing the exons and their surrounding regions seem to weakly correlate with exon coSI values in chromatin (Figure S10). The most notable correlation is with distance to the PolyA-site (See also Figure 2) and, albeit somehow weaker with the distance to the TSS. In addition, exon coSI values correlate positively with the strength of the acceptor sites, GC content, and anti-correlate with the length of the downstream intron-this is supposedly because reads spanning the exon-intron border can be observed once the donor is transcribed, while splicing can only be carried out once the entire downstream intron is transcribed. It also appears that the presence of binding sites for some splicing factors weakly correlate with coSI scores (data not shown). We have further investigated the exonic behavior of a number of chromatin modifications (Ernst *et al.* 2011) (monitored through ChIP-seq in K562, see Supplementary Information and Methods) depending on the exon coSI value in chromatin associated RNA. All chromatin marks monitored, as well as nucleosome (Kundaje *et al.* 2012) and Pol II occupancy negatively correlate with chromatin coSI values (Figure S10). That is, there is a general enrichment of chromatin marks in exons with low coSI values, consistent with the DNA in these exons being still in chromatin status prior to or during transcription

Published online 1 January 2019.