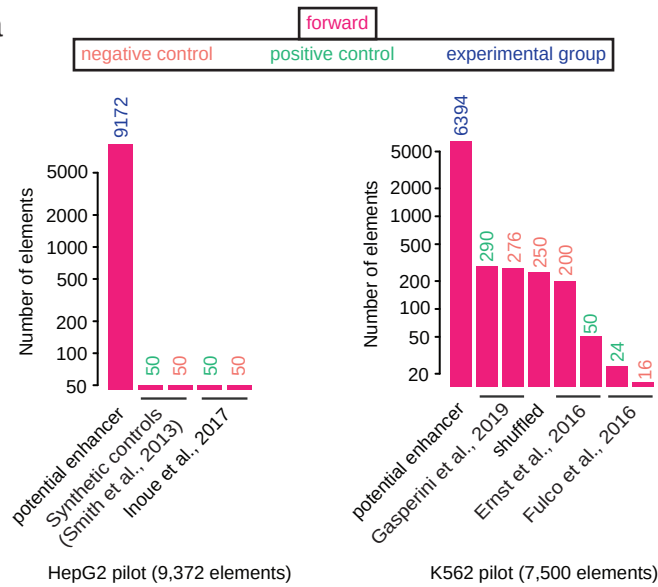
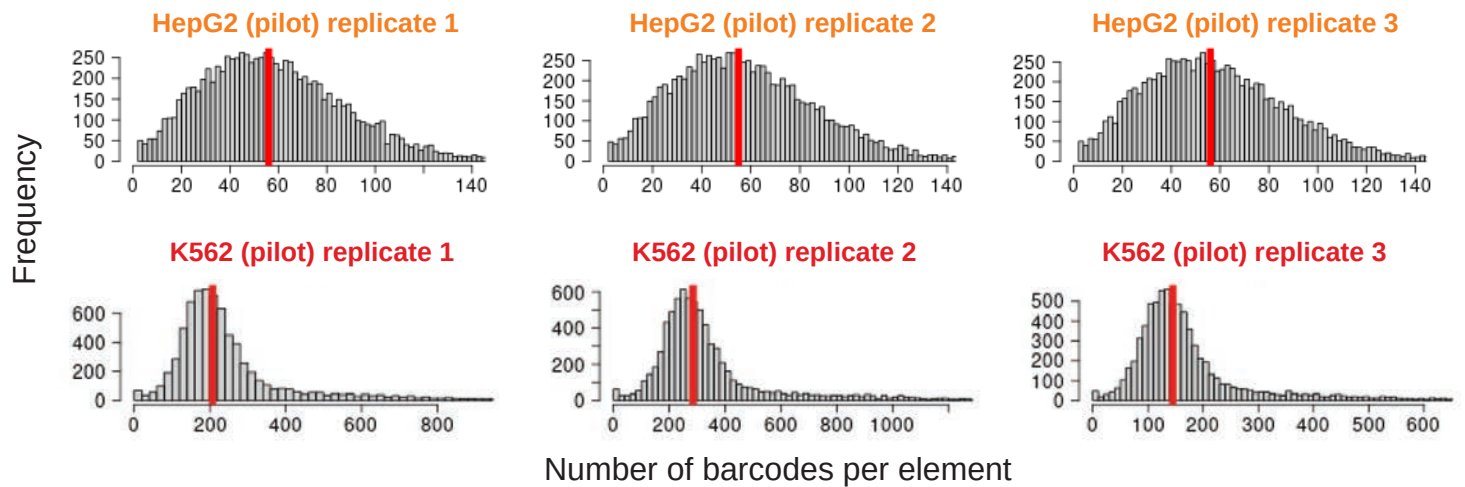


a

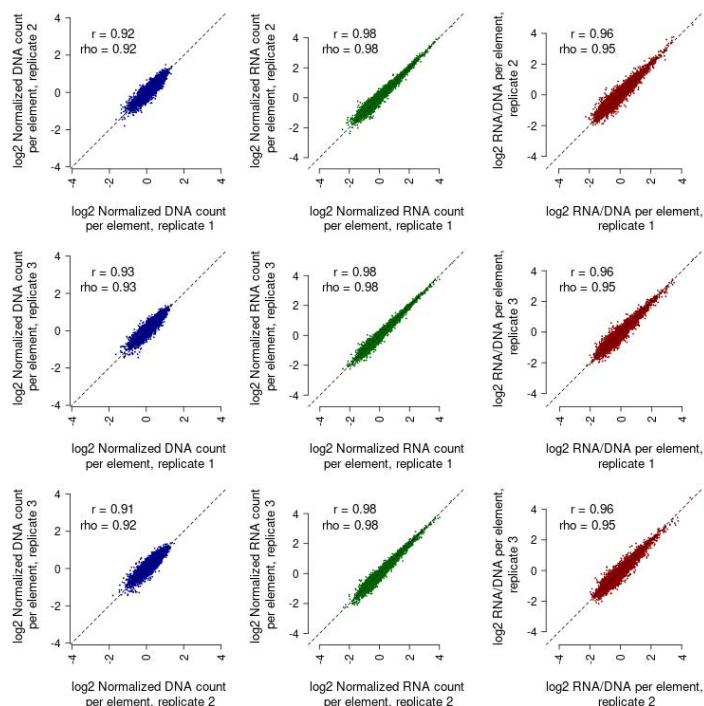


b



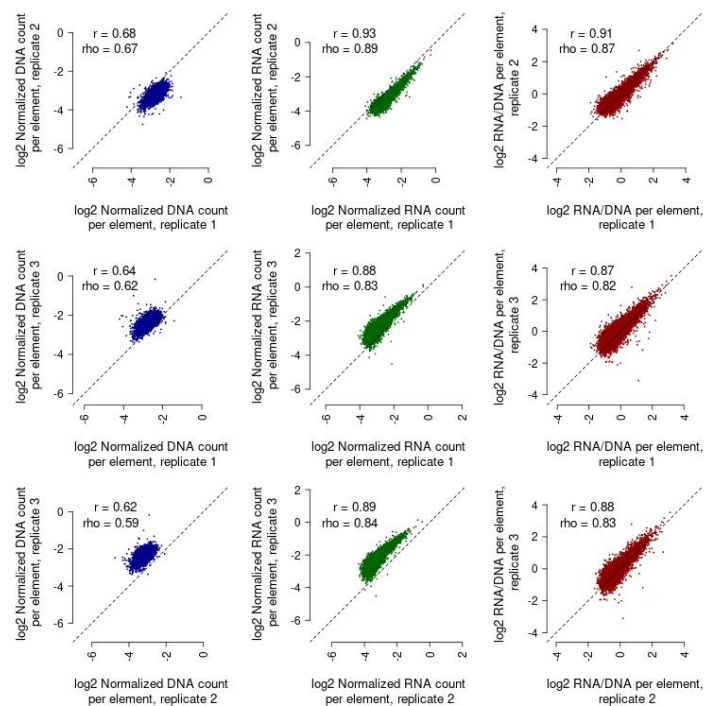
c

HepG2 (pilot)

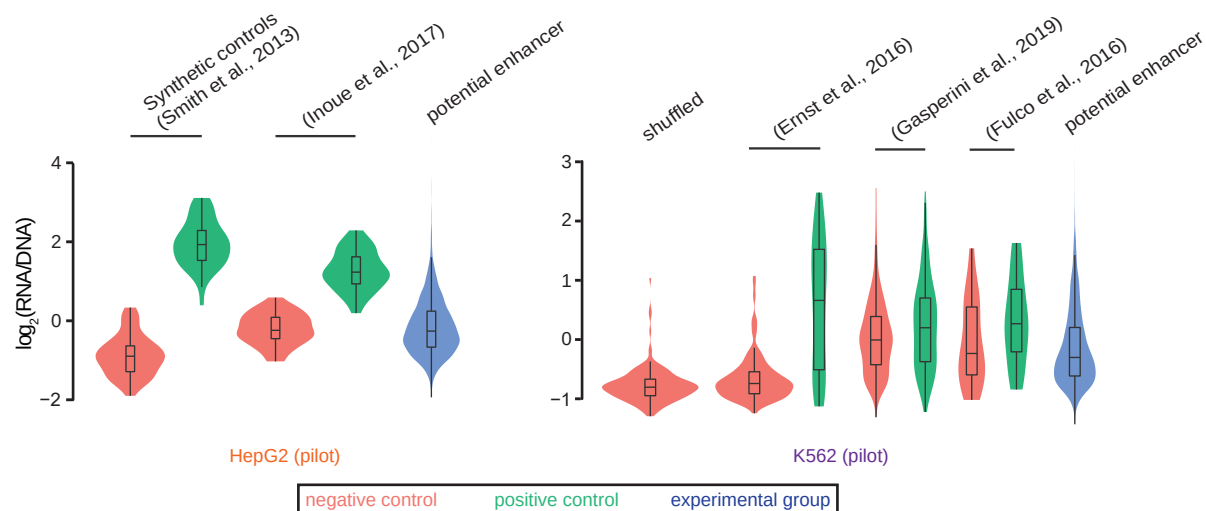


d

K562 (pilot)

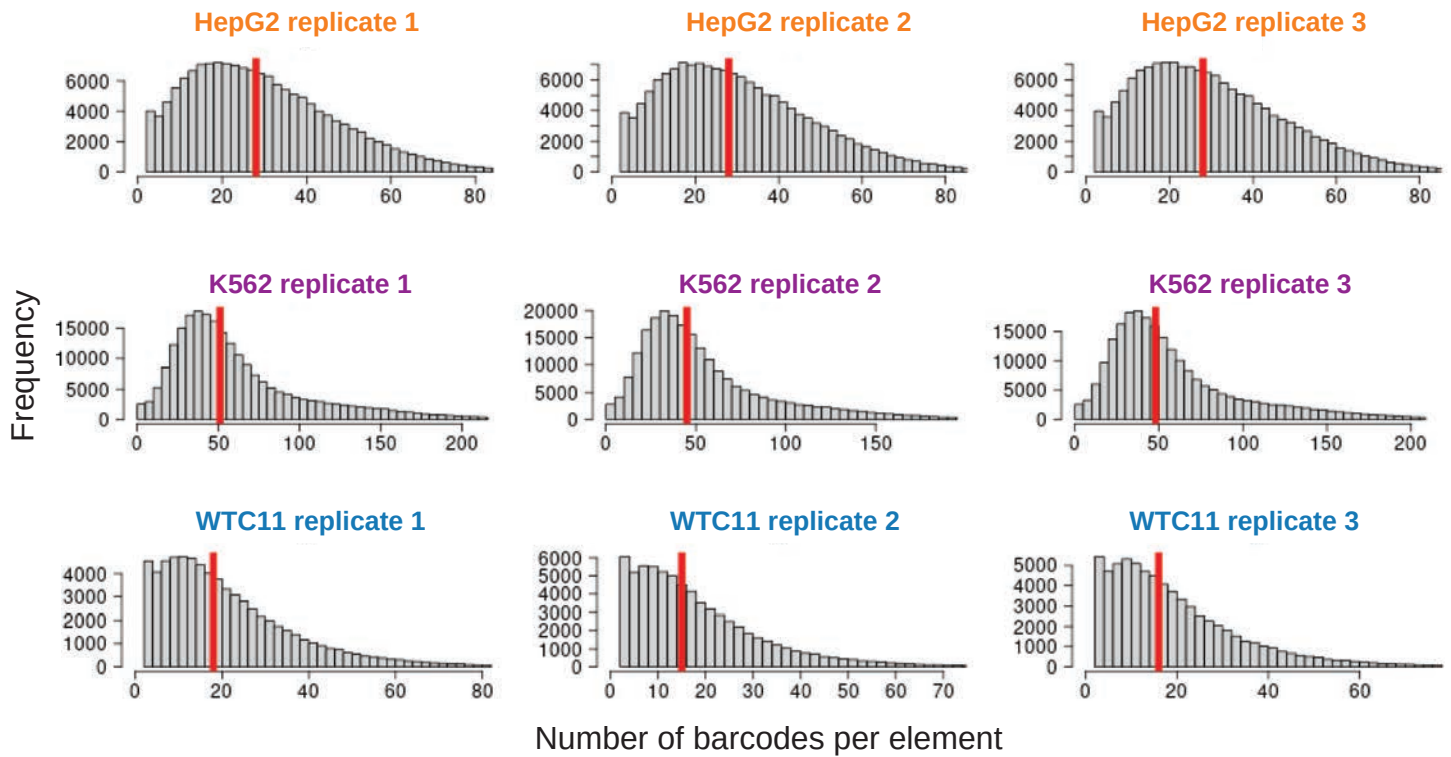


e

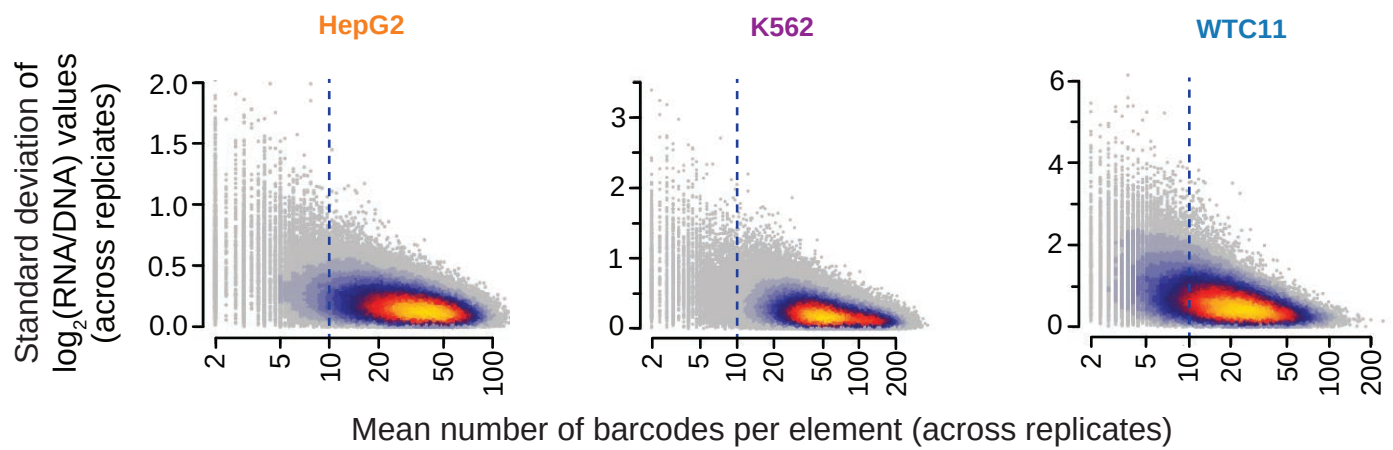


Supplementary Fig. 1: Design and quality control characteristics of two pilot MPRA libraries. **a**, Composition of the HepG2 and K562 pilot libraries. Thousands of potential enhancers, negative controls (dinucleotide shuffled sequences or elements lacking a signal from prior studies), and positive controls (elements with reported activity from prior studies) are included in each library. To maintain consistency with **Fig. 1b**, bars are colored according to orientation tested, with accompanying numbers indicating the number of elements tested in the category. Numbers are colored according to element type. **b**, Shown are histograms indicating the number of observed barcodes per element, for each of the three replicates and two pilot MPRA libraries. Shown with a vertical red line is the median number of barcodes per element. **c-d**, Shown are scatter plots displaying the relationship between observed DNA counts (blue), RNA counts (green), and RNA/DNA ratios (red) for all pairwise comparisons among replicates, for both the **(c)** HepG2 and **(d)** K562 pilot MPRA libraries. Also indicated is the Pearson (r) and Spearman (ρ) correlation values. Candidate enhancers supported by fewer than 10 barcodes were filtered out prior to this analysis to reduce the impact of technical noise. **e**, Violin plots of element activity [measured as $\log_2(\text{RNA/DNA})$] for potential enhancers, negative controls, and positive controls for each library.

a

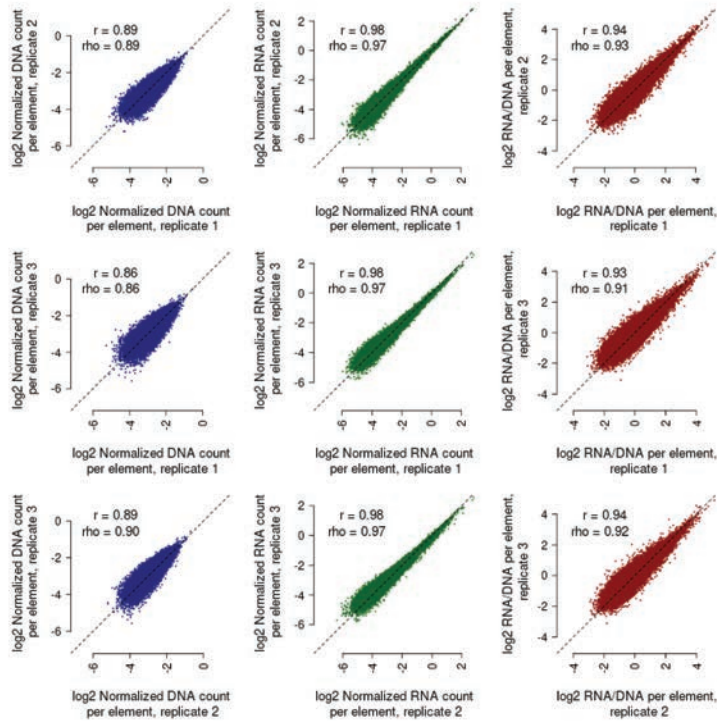


b



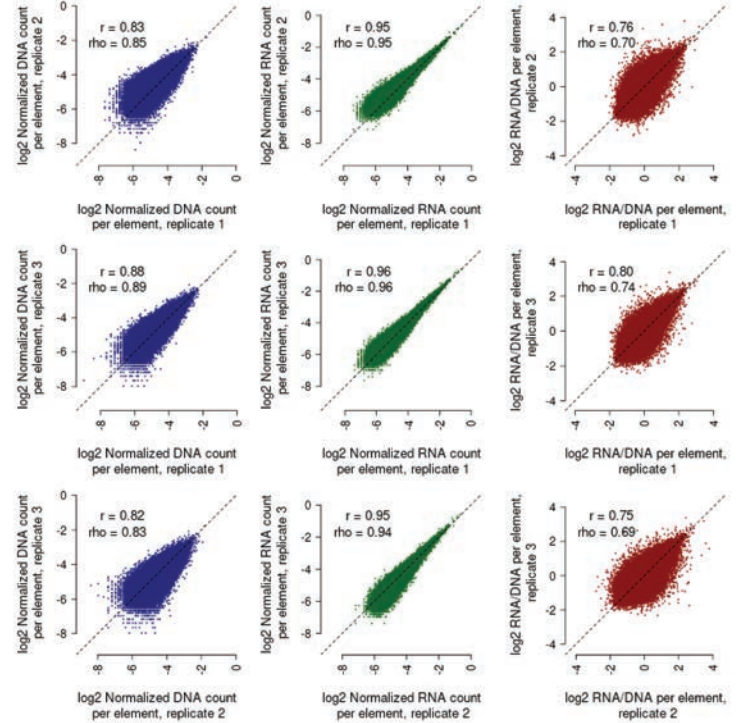
c

HepG2



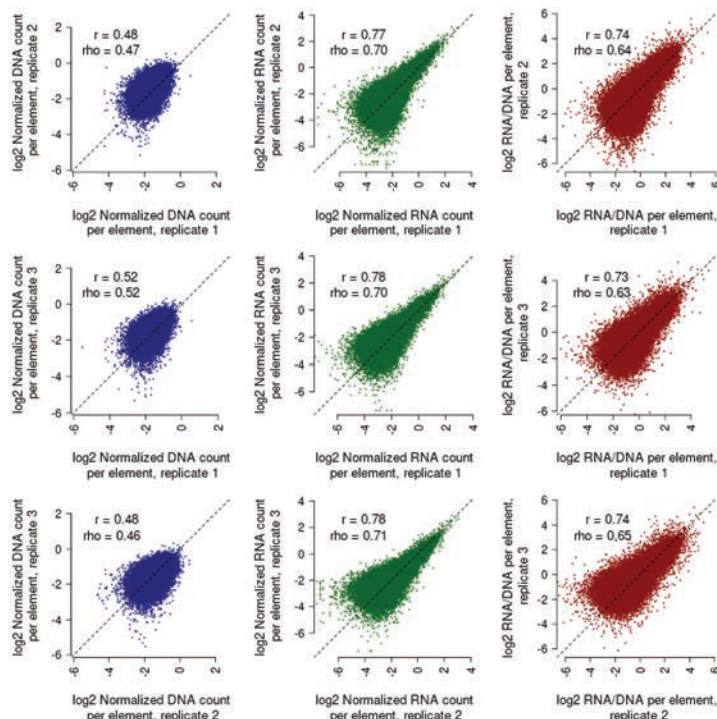
d

K562



e

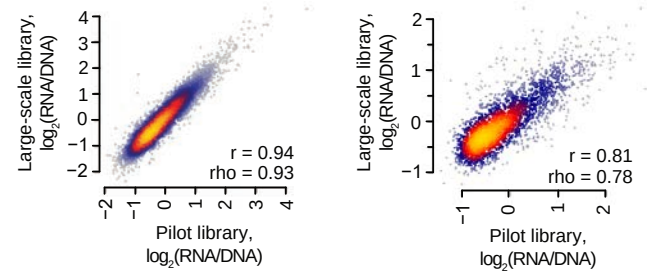
WTC11

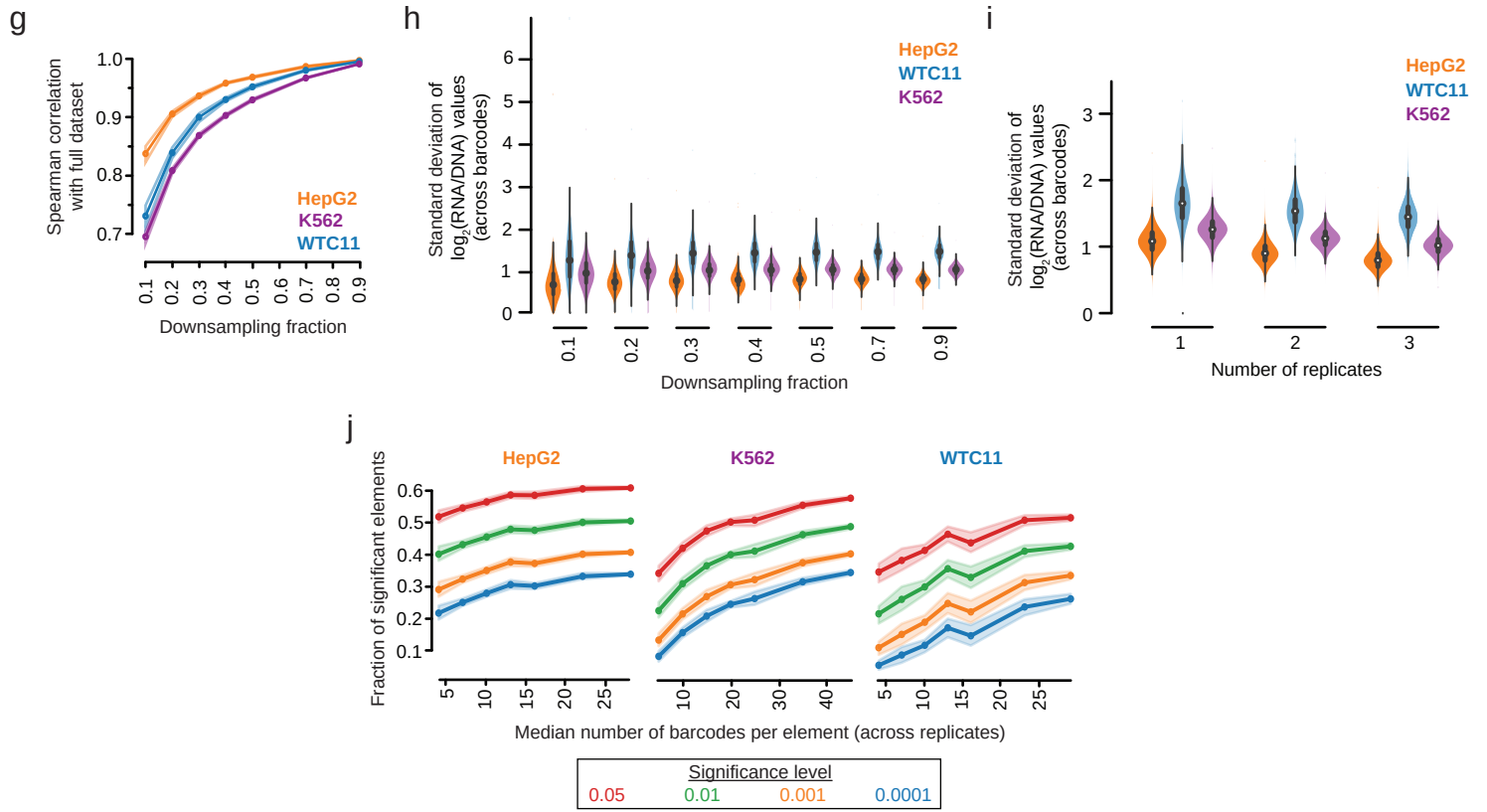


f

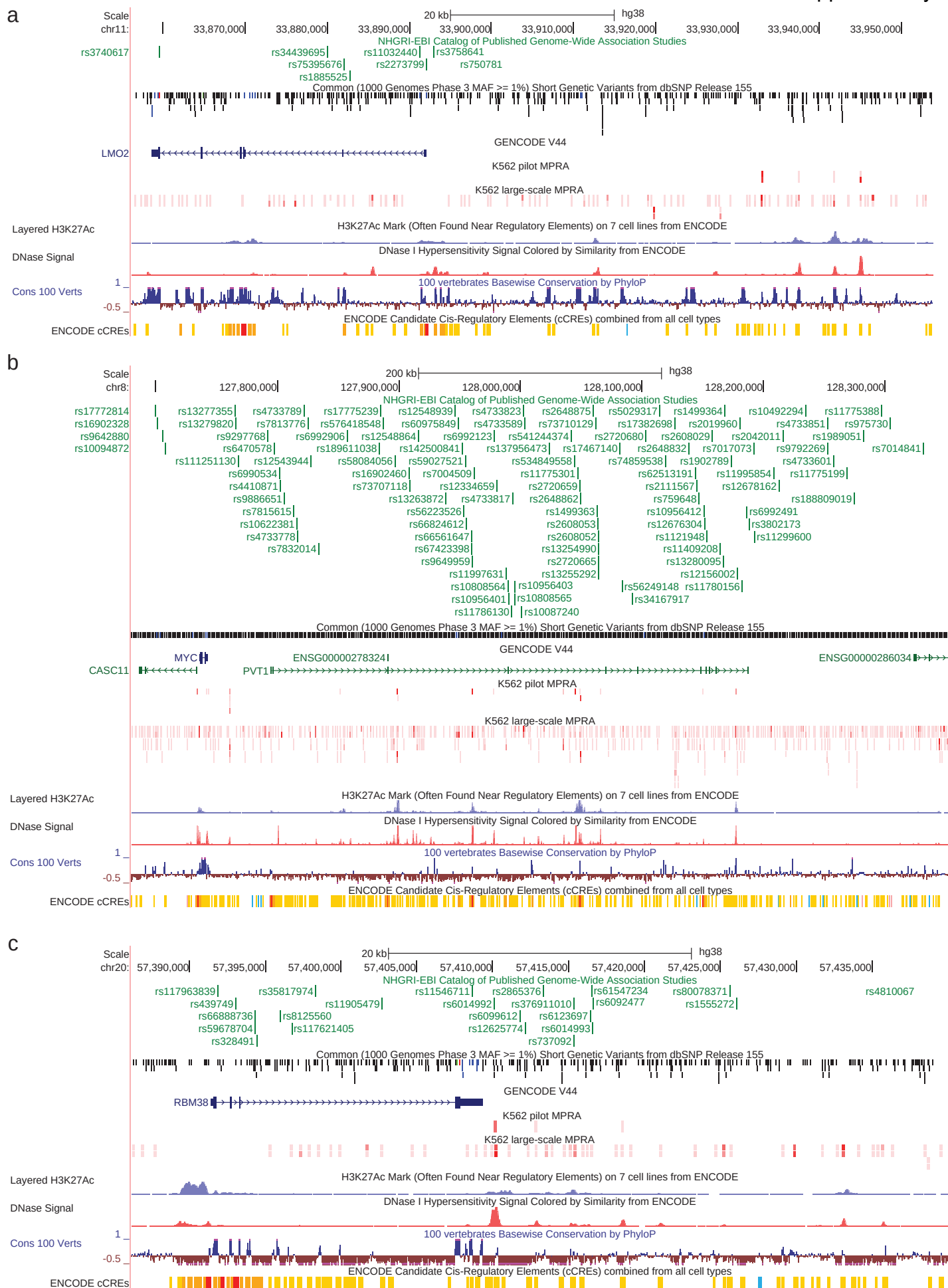
HepG2

K562

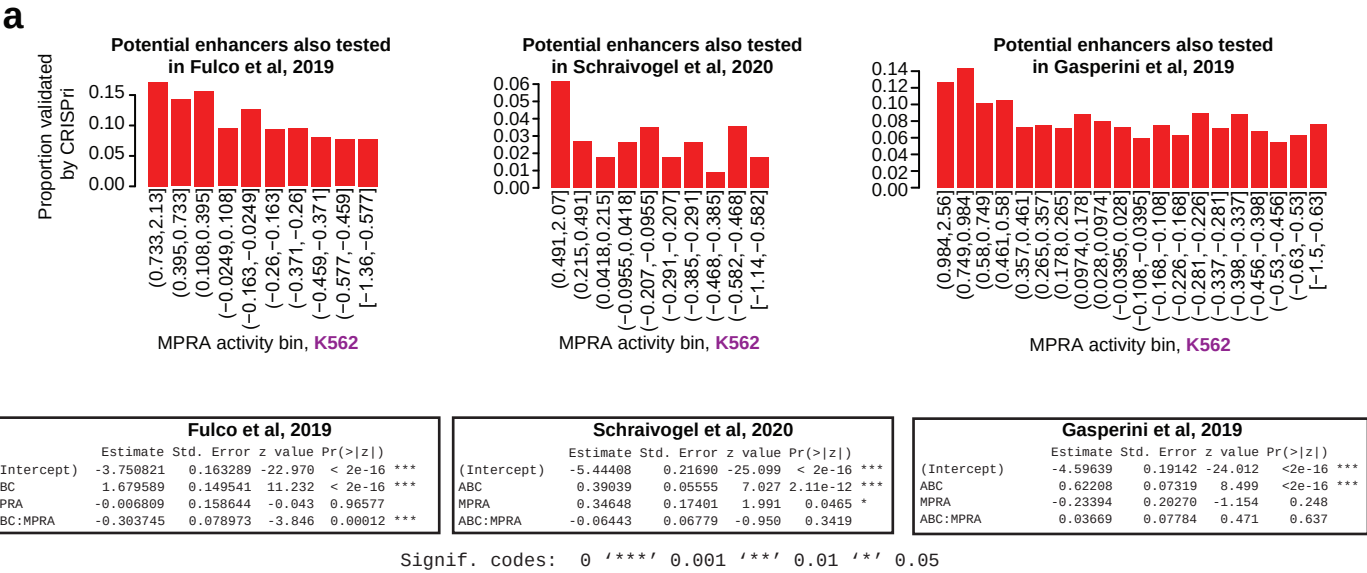




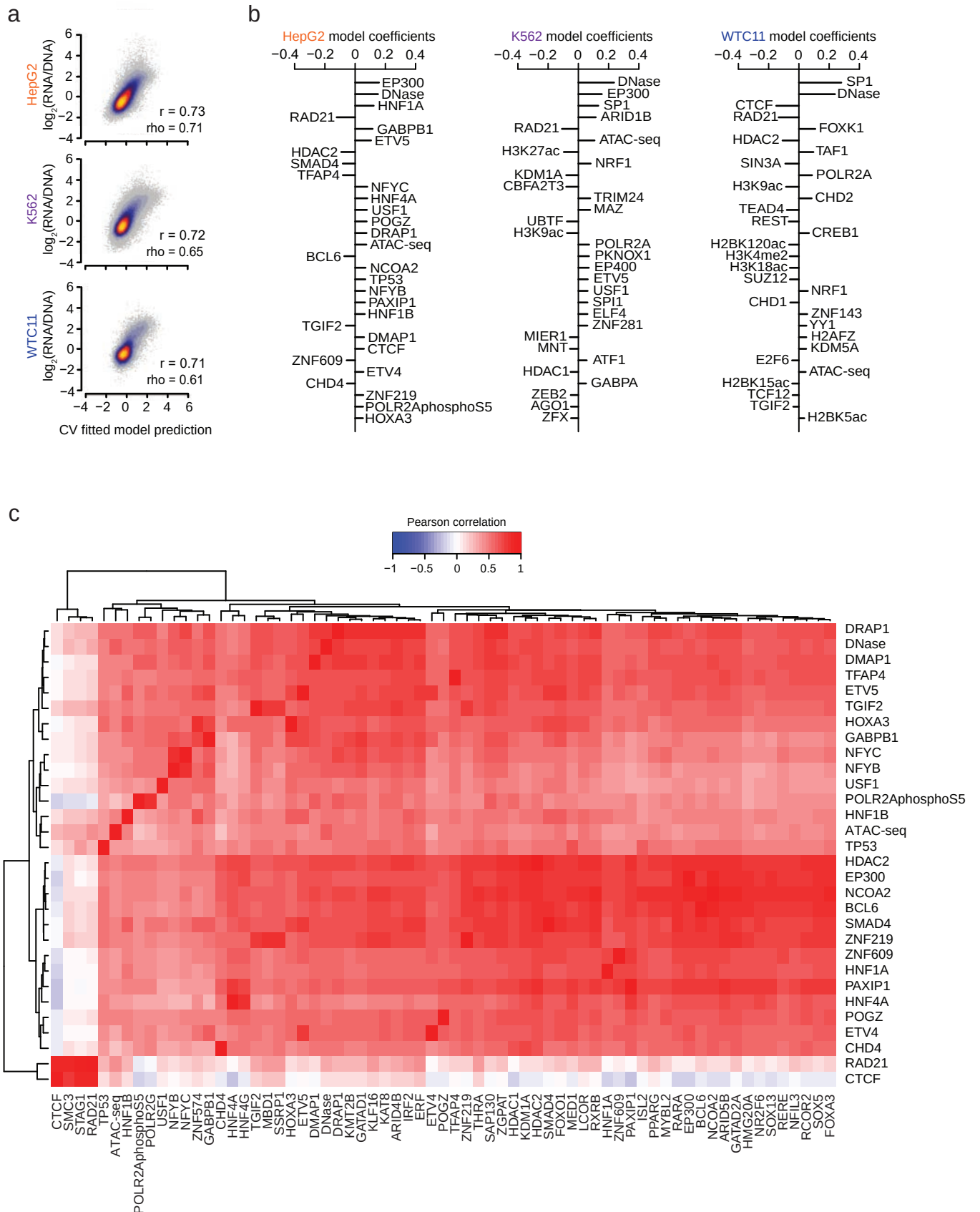
Supplementary Fig. 2: Quality control characteristics of the three large-scale MPRA libraries. **a**, Shown are histograms indicating the number of observed barcodes per element, for each of the three replicates and three large-scale MPRA libraries. Shown with a vertical red line is the median number of barcodes per element. **b**, Relationship between the mean number of barcodes in each library relative to the standard deviation of the element activity scores across replicates. The dashed line indicates our threshold of 10 minimum barcodes. **c-e**, Shown are scatter plots displaying the relationship between observed DNA counts (blue), RNA counts (green), and RNA/DNA ratios (red) for all pairwise comparisons among replicates, for the **(c)** HepG2, **(d)** K562, and **(e)** WTC11 large-scale MPRA libraries. Candidate elements supported by fewer than 10 barcodes were filtered out prior to this analysis to reduce the impact of technical noise. **f**, Scatter plots displaying the relationships between activity scores for the subset of elements common to both the pilot and large-scale libraries tested in HepG2 and K562 cells. Also indicated is the Pearson (r) and Spearman (ρ) correlation values. **g**, Correlation between the $\log_2(\text{RNA/DNA})$ ratios of the downsampled and the full data. Barcodes were downsampled at fractions of 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, and 0.9 of the total number of barcodes per sequence. The plot shows the average and the 95% confidence interval of 10 downsampling replicates for HepG2, WTC11, and K562 cells. **h**, Standard deviation between barcode $\log_2(\text{RNA/DNA})$ ratios associated to the same element at downsampling fractions of 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, and 0.9 for HepG2, WTC11, and K562. For each sequence and downsampling fraction the downsampling was performed 10 times. **i**, Standard deviation between barcode $\log_2(\text{RNA/DNA})$ ratios associated to the same element when using one, two, or three replicates for HepG2, WTC11, and K562 cells. **j**, Fraction of all sequences in HepG2, WTC11, and K562 cells that are found to be significantly active compared to the shuffled negative controls after downsampling barcodes at downsampling fractions of 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, and 0.9. On the x-axis, the median number of barcodes after downsampling is displayed. The y-axis shows the average fraction and the 95% confidence interval of 10 downsampling replicates at four different significance levels. The significance of each sequence was calculated using the empirical p-value based on the MAD-score from the distribution of the downsampled shuffled negative controls.



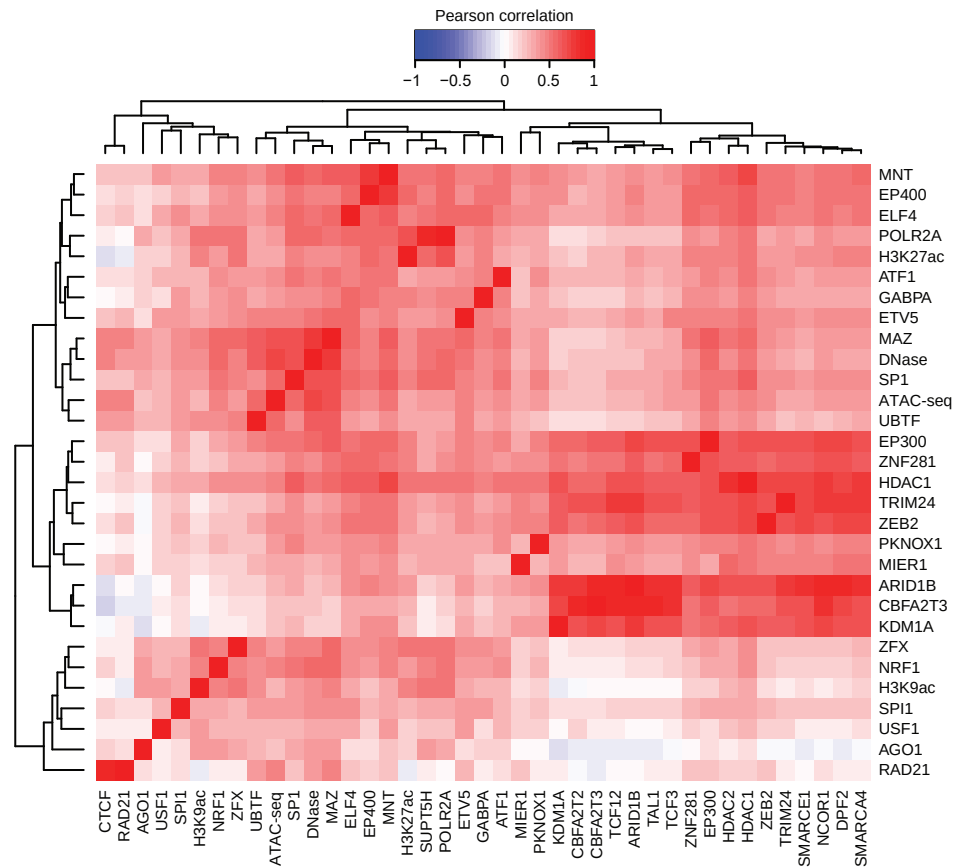
Supplementary Fig. 3: MPRA activity in selected disease loci. a-c, This figure is an extension of **Extended Data Fig. 2** but displays *LMO2* (**a**), *MYC* (**b**), and *RBM38* (**c**) as additional loci.



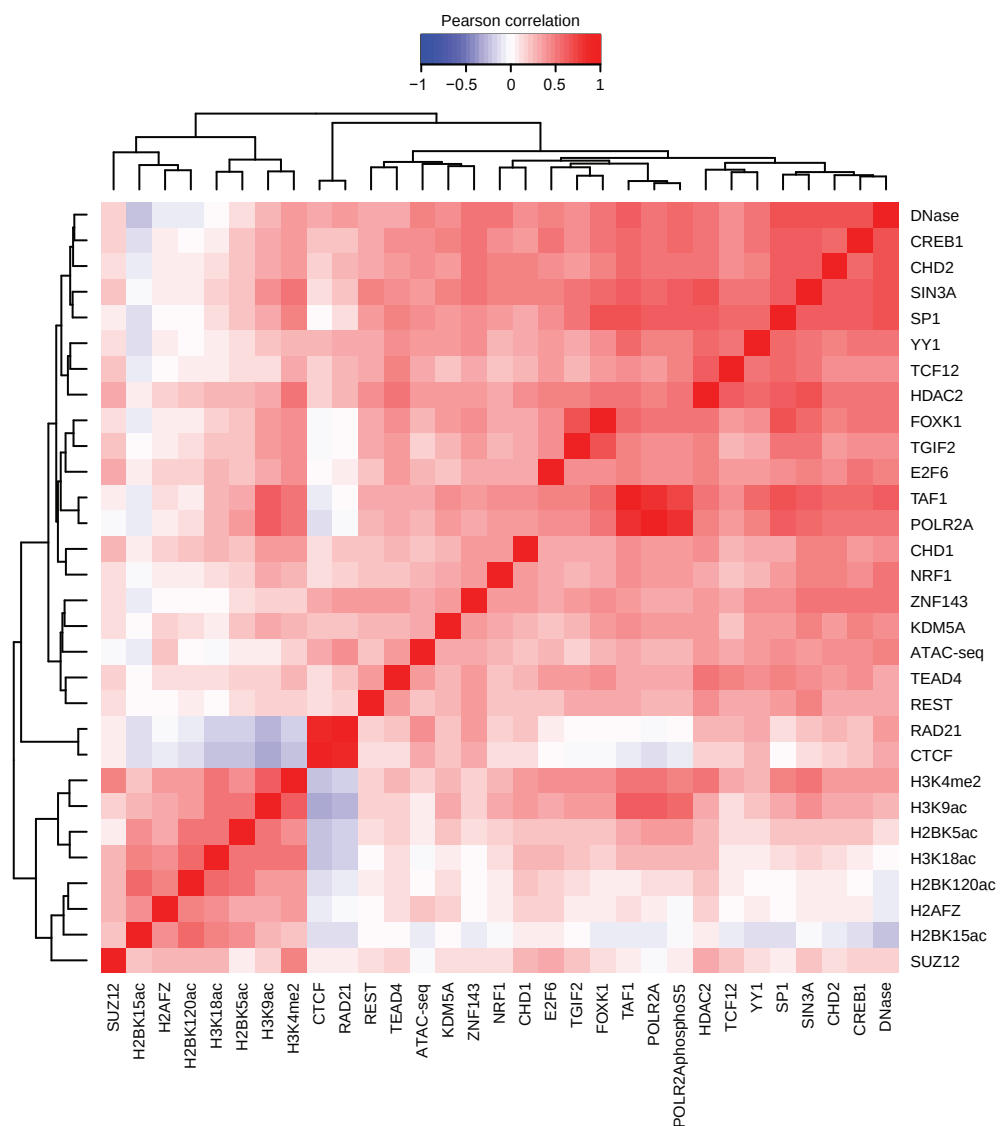
Supplementary Fig. 4: Relationship between MPRA and CRISPRi datasets. **a**, Fraction of elements validated by CRISPRi (*i.e.*, called to significantly regulate a target gene) for equally sized bins of MPRA activity. The bins with the highest activity level exhibit a greater likelihood for validation. The left and right numerical intervals of the MPRA activity bins are shown comma-separated, with open parentheses meaning the exclusion of the left interval and square brackets meaning the inclusion of the right interval. Three CRISPRi datasets were cross-referenced for this analysis (Fulco et al., 2019, Gasperini et al., 2019, Schraivogel et al., 2020). **b**, Using the datasets in part (a), we intersected our set of potential enhancers with those tested by CRISPRi. We then trained a binary classifier, using logistic regression to predict whether a specific enhancer-gene linkage was statistically significant, considering the ABC model along with an interaction term with MPRA activity. The tables shown indicate the coefficients, z-value, and statistical significance of each term. The Fulco et al., 2019 and Schraivogel et al., 2020 datasets exhibited support for at least one MPRA term while there was no evidence for the benefit of this term in the Gasperini et al., 2019 dataset.



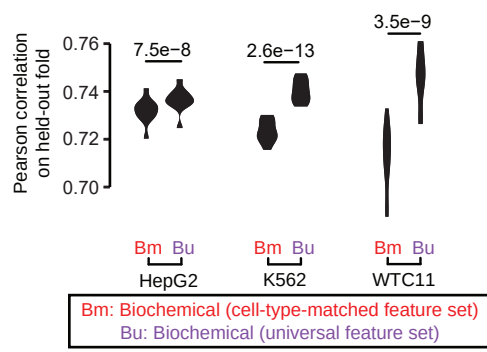
d



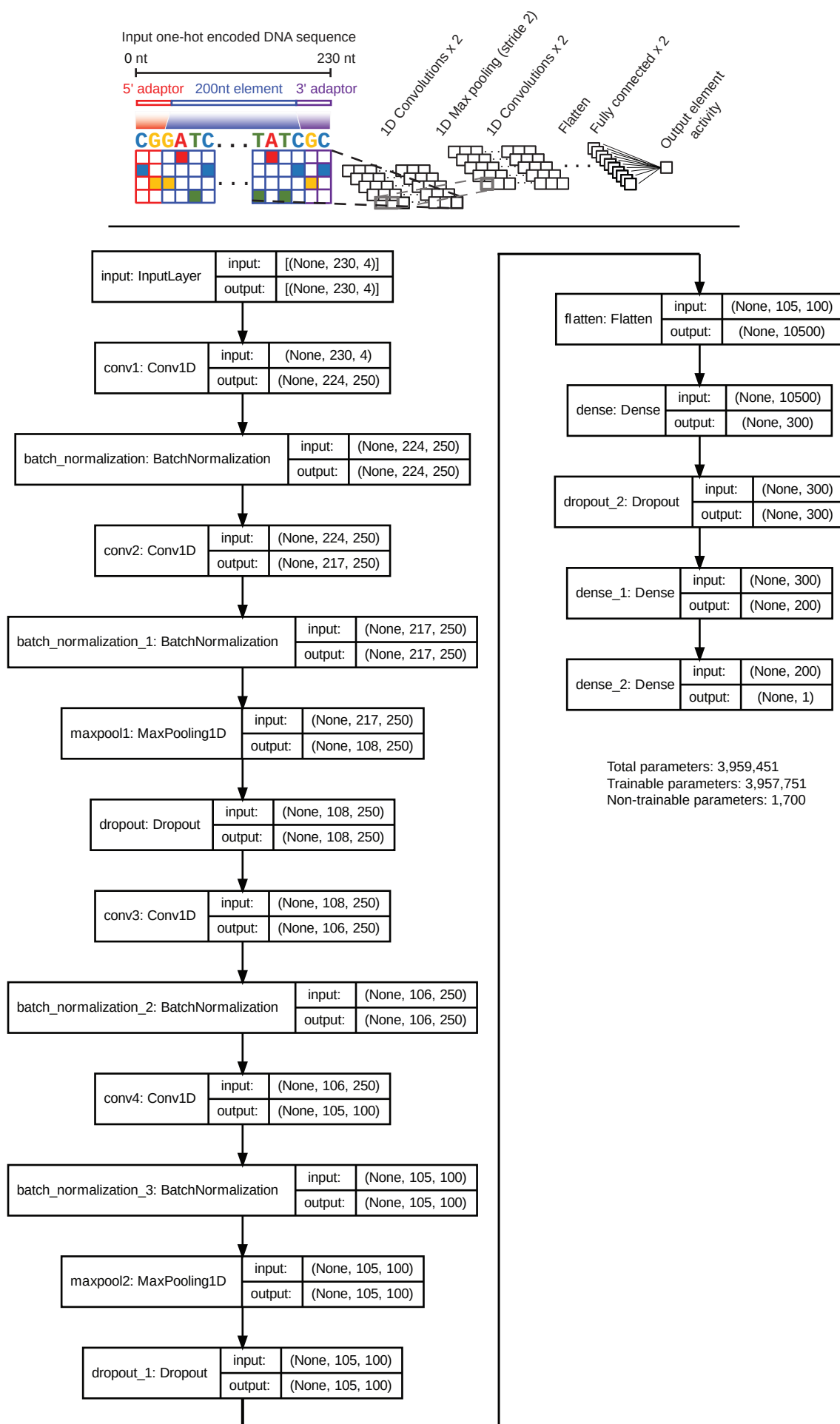
e



f

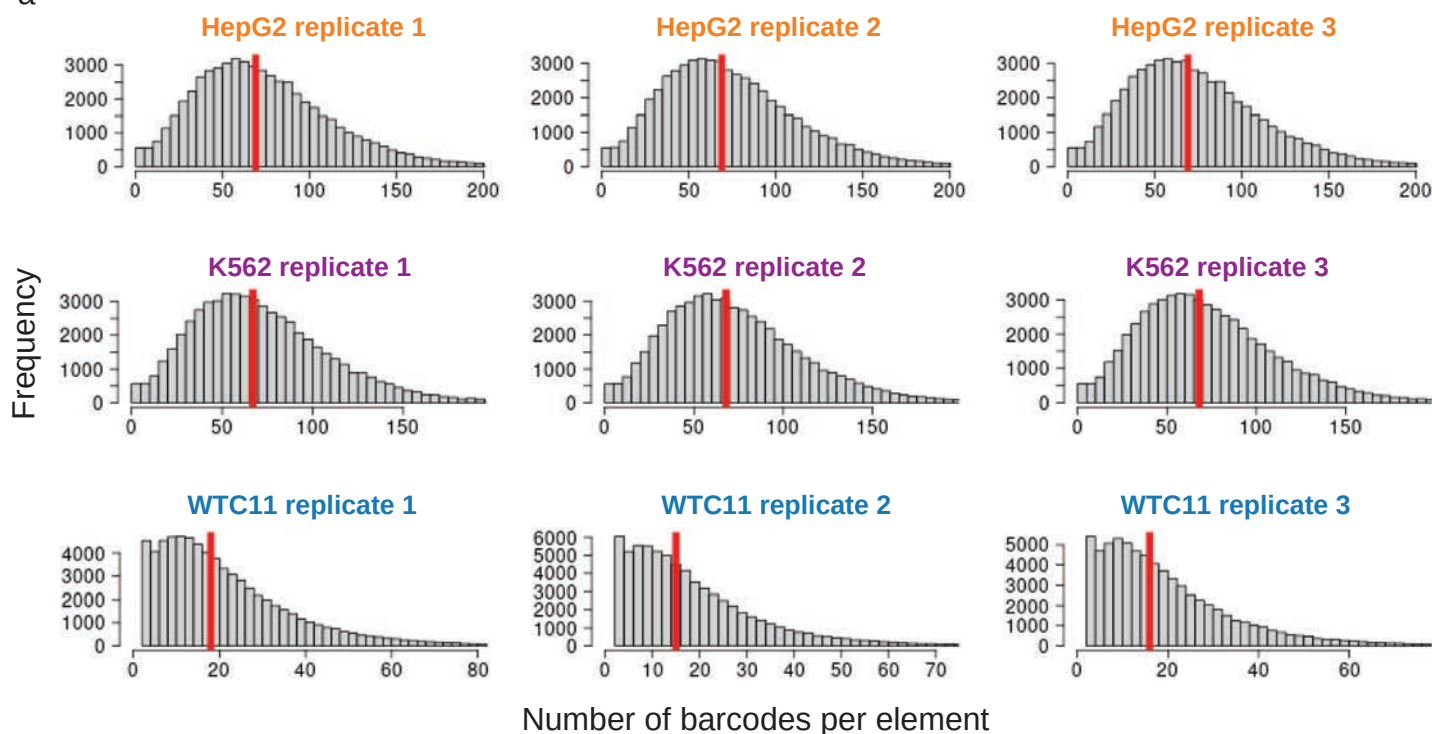


Supplementary Fig. 5: Prediction of sequence activity based on biochemical features. **a**, Scatter plot indicating relationship between model predictions and observed element activity scores for each cell type. Pearson (r) and Spearman (ρ) correlation values are shown after concatenating the observations for all 10 cross-validation folds of held-out data. **b**, The top 30 coefficients derived from lasso regression models trained independently on each cell type. **c-e**, Pearson correlation matrix between the top 30 features from **panel b**, shown as rows, and other features sharing a Pearson correlation either ≤ -0.8 or ≥ 0.8 , shown as columns, for HepG2 (**c**), K562 (**d**), and WTC11 (**e**) cells. Hierarchical clustering was used to group features exhibiting similar correlation patterns. **f**, Violin plots showing the performances of the trained biochemical models on each of the ten cross-validation folds of held-out data. The lasso regression models evaluated are those that consider cell-type matched features (Bm) for the indicated cell type, or those that consider a universal set of biochemical features across all three cell types (Bu). An improvement relative to another model was evaluated with a one-sided, paired t-test.



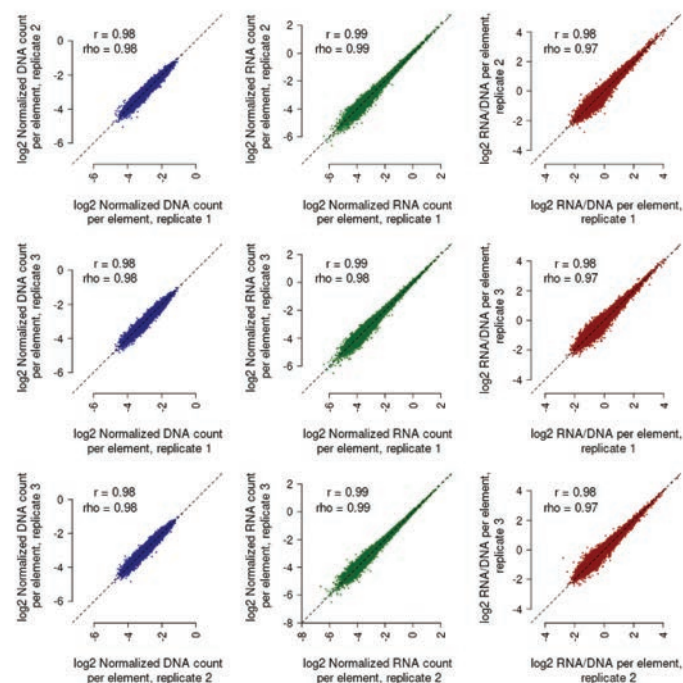
Supplementary Fig. 6: Architecture of MPRAnn. Cartoon (top) and complete architecture (bottom) of the MPRAnn model. Indicated for each layer is the layer name and dimensionality of the input and output matrices. 'None' refers to the batch size used during model training.

a



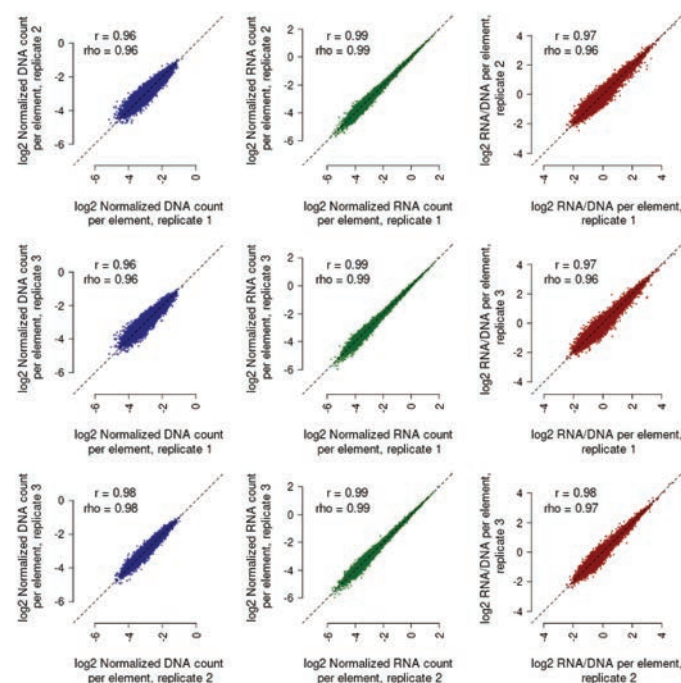
b

HepG2



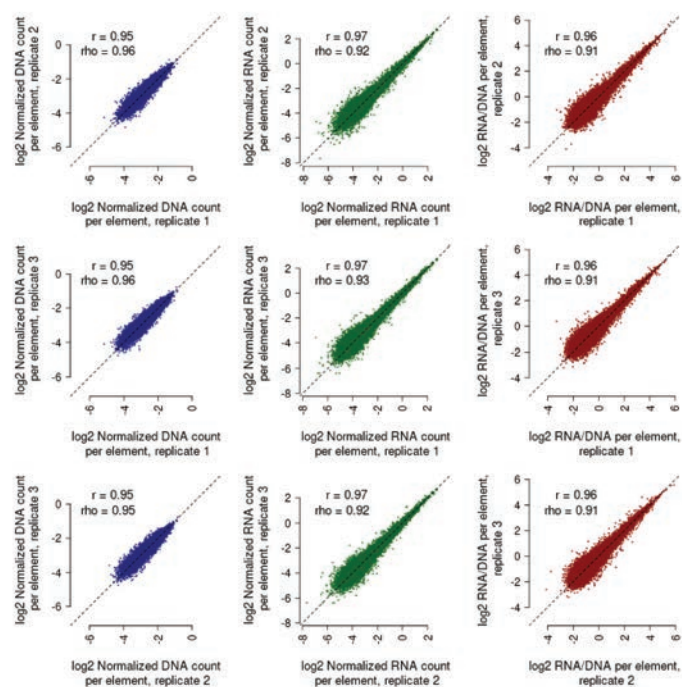
c

K562

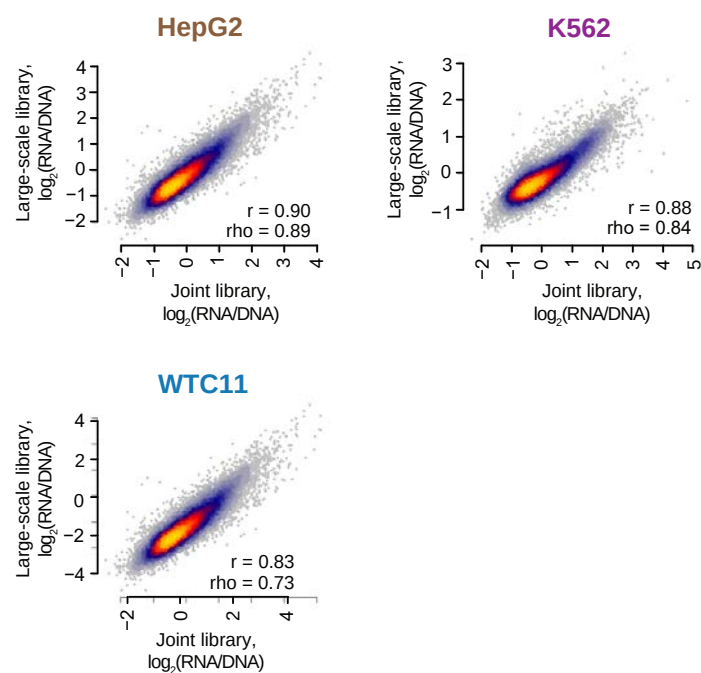


d

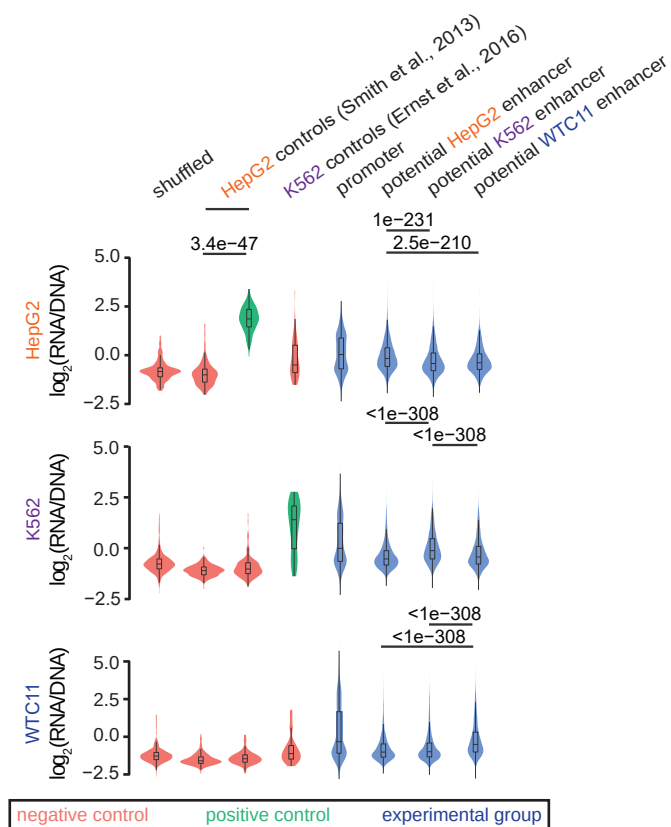
WTC11



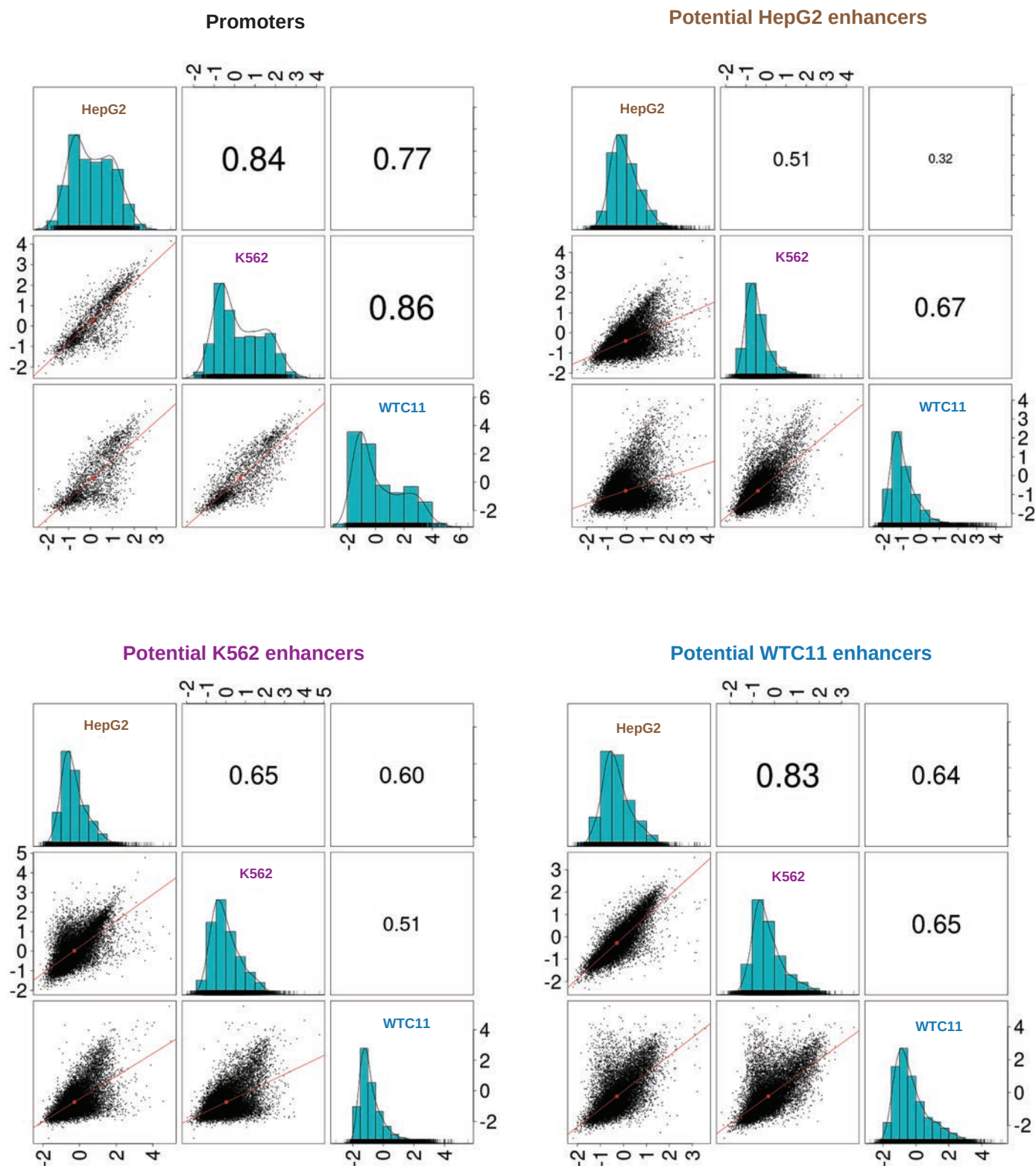
e



f

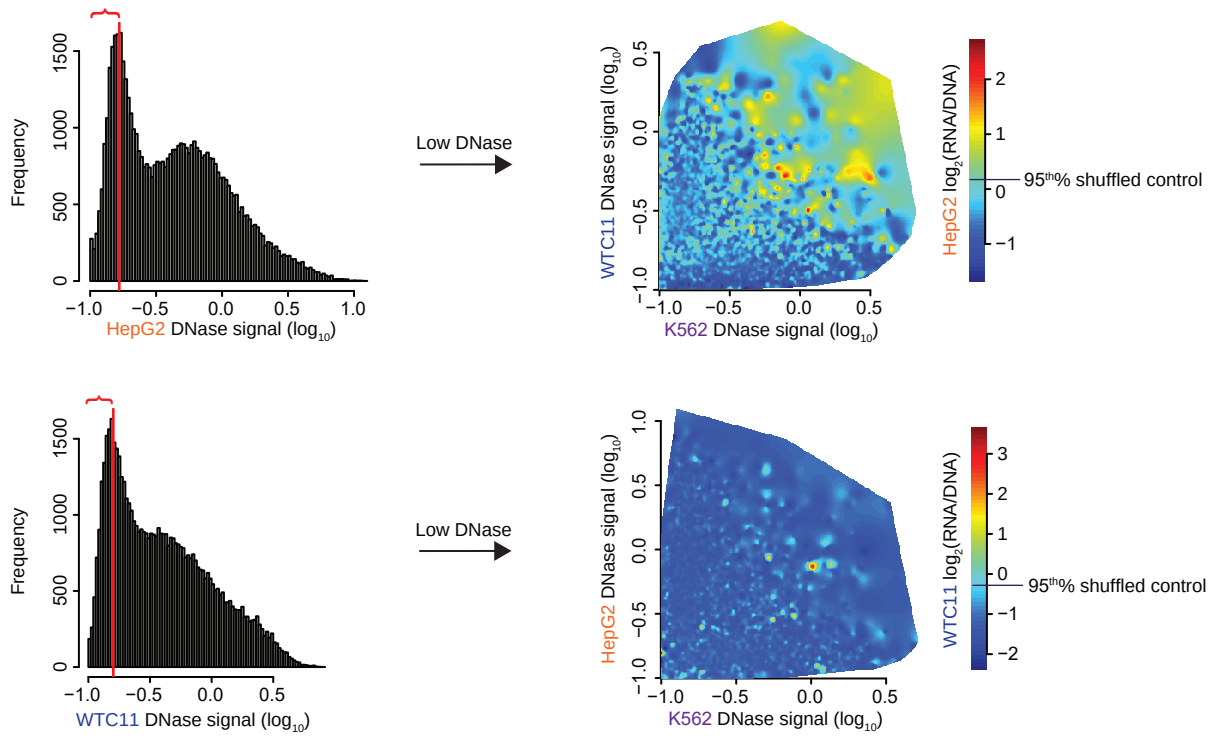


Supplementary Fig. 7: Quality control characteristics of the joint MPRA library. **a**, Histograms indicating the number of observed barcodes per element, for each of the three replicates and three cell types tested using the joint MPRA library. Shown with a vertical red line is the median number of barcodes per element. **b-d**, Scatter plots displaying the relationship between observed DNA counts (blue), RNA counts (green), and RNA/DNA ratios (red) for all pairwise comparisons among replicates, for the joint MPRA library tested in **(b)** HepG2, **(c)** K562, and **(d)** WTC11 cells. Candidate elements supported by fewer than 10 barcodes were filtered out prior to this analysis to reduce the impact of technical noise. **e**, Scatter plots displaying the relationships between activity scores for the subset of elements common to both the joint and large-scale MPRA libraries tested in HepG2, K562, and WTC11 cells. Also indicated are the Pearson (r) and Spearman (ρ) correlation values. **f**, Violin plots of element activity [measured as $\log_2(\text{RNA/DNA})$] in each of the three cell types for different element categories represented in the joint MPRA library shown in **Fig. 5a**. The difference between each pair of distributions tested was evaluated with a one-sided Wilcoxon rank-sum test, adjusted with a Bonferroni correction to account for the total number of hypothesis tests.

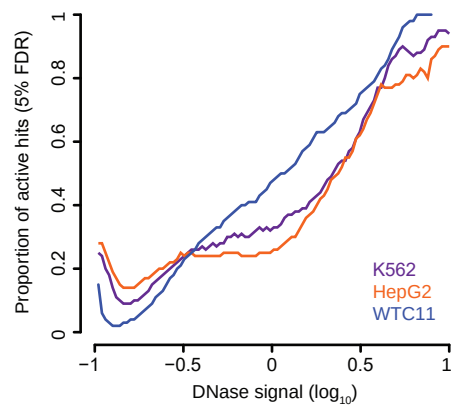


Supplementary Fig. 8: Comparison of promoter and enhancer activities from a joint MPRA library tested in three cell types. Scatter matrix displaying scatter plots corresponding to each of the three pairs of possible inter-cell-type comparisons (lower diagonal elements), for each of four element categories: i) protein-coding gene promoters, ii) potential enhancers selected from HepG2 cells, iii) potential enhancers selected from K562 cells, and iv) potential enhancers selected from WTC11 cells. Shown on the diagonal is a histogram of element activity scores [measured as $\log_2(\text{RNA/DNA})$]. Also shown are Pearson correlation values among each pair of comparisons, with the size of the text proportional to the magnitude of the correlation coefficient (upper diagonal elements).

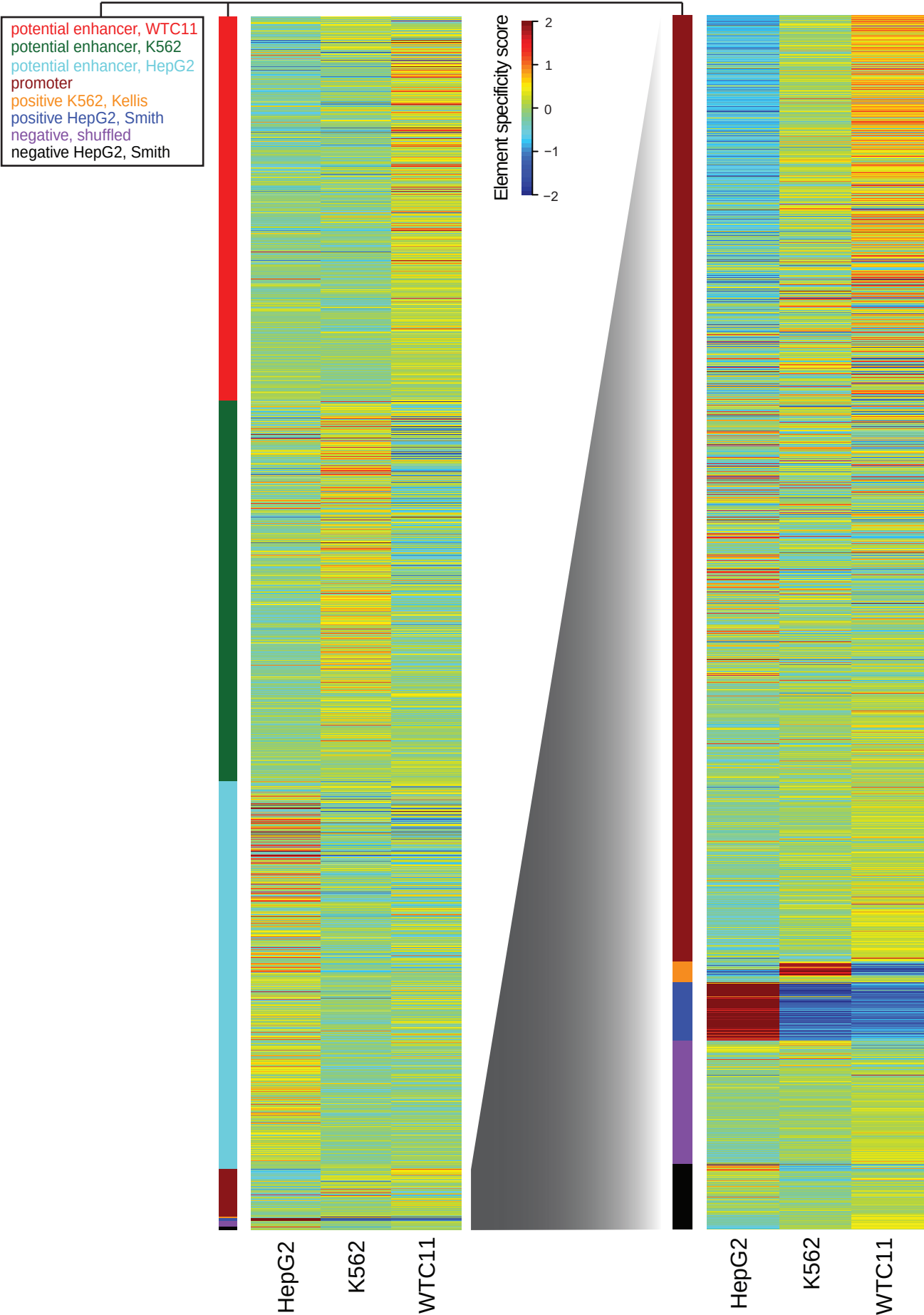
a



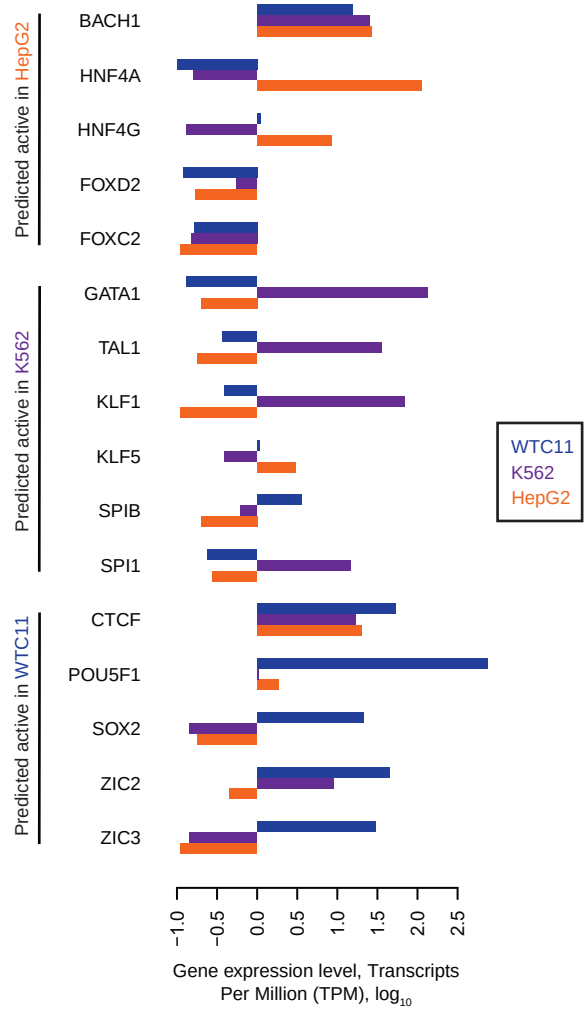
b



Supplementary Fig. 9: Activity of elements with low DNase signal. a, In each of the depicted cell types, we quantified DNase signal and selected the subset of elements with low signal (*i.e.*, having a signal below the threshold indicated by the vertical red line). For this subset of elements, we evaluated DNase signal in the other two cell types and then quantified a smoothened kernel density estimate of MPRA activity in the cell type being examined. **b,** Point estimate of the proportion of active elements (*i.e.*, exceeding the signal for shuffled negative controls with a 5% FDR) as a function of DNase signal in each of three cell types. Each point along the x-axis reflects this proportion for elements with a DNase signal in a range of $(x-0.1$ to $x)$.



Supplementary Fig. 10: Comparison of element specificity scores derived from a joint MPRA library tested in three cell types. Heatmap of element specificity scores (*i.e.*, computed as the deviation of an element's activity from its mean activity in all cell types). The heatmap shown on the right is a zoomed version of the heatmap on the left for elements other than potential enhancers. Elements are colored according to their category in the key provided.



Supplementary Fig. 11: Gene expression levels of TFs in each of the three cell types, for the members of TF families predicted to be highly active in each cell type based upon motif enrichment analysis. Shown are barplots indicating expression levels measured as Transcripts Per Million (TPM) from RNA-seq data derived from each of the three cell lines.