nature methods

ANALYSIS

https://doi.org/10.1038/s41592-020-0965-y

Check for updates

A systematic evaluation of the design and context dependencies of massively parallel reporter assays

Jason C. Klein[1,9,11], Vikram Agarwal[1,2,11], Fumitaka Inoue[3,4,10,11], Aidan Keith[1,11], Beth Martin[1], Martin Kircher[1,5,6], Nadav Ahituv[3,4] ✉ and Jay Shendure[1,7,8] ✉

Massively parallel reporter assays (MPRAs) functionally screen thousands of sequences for regulatory activity in parallel. To date, there are limited studies that systematically compare differences in MPRA design. Here, we screen a library of 2,440 candidate liver enhancers and controls for regulatory activity in HepG2 cells using nine different MPRA designs. We identify subtle but significant differences that correlate with epigenetic and sequence-level features, as well as differences in dynamic range and reproducibility. We also validate that enhancer activity is largely independent of orientation, at least for our library and designs. Finally, we assemble and test the same enhancers as 192-mers, 354-mers and 678-mers and observe sizable differences. This work provides a framework for the experimental design of high-throughput reporter assays, suggesting that the extended sequence context of tested elements and to a lesser degree the precise assay, influence MPRA results.

Spatiotemporal control of gene expression is orchestrated in part by distally located DNA sequences known as enhancers. The first enhancers were identified by cloning fragments of DNA into a plasmid with a reporter gene and promoter[1–4]. Transcriptional enhancement in such reporter assays continues to be widely used for evaluating whether a putative regulatory element is a bona fide enhancer. However, conventional, one-at-a-time reporter assays are insufficiently scalable to test the >1 million putative enhancers in the human genome[5–8].

MPRAs modify in vitro reporter assays to facilitate simultaneous testing of thousands of putative regulatory elements[9–11] per experiment. MPRAs characterize each element through sequencing-based quantification of transcribed, element-linked barcodes[9–15]. MPRAs have facilitated the scalable study of putative regulatory elements for goals, including functional annotation[16–18], variant effect prediction[10–15,19] and evolutionary reconstruction[20,21].

Over the past decade, diverse designs for enhancer-focused MPRAs have emerged. Major differences include whether the enhancer is upstream[10,11] versus within the 3′ untranslated region (UTR) of the reporter[16] and whether the construct remains episomal versus integrated[18]. Additionally, most MPRAs test sequences in only one orientation, effectively assuming enhancer activity is independent of orientation. Finally, while sheared genomic DNA[16,22], PCR amplicons[12] or hybrid captured sequences[23,24] have been used in MPRAs, most studies synthesize libraries of candidate enhancers on microarrays, generally limiting them to <200 base pair (bp).

Unfortunately, we have, as a field to date, largely failed to systematically evaluate how these design choices impact or bias the results of MPRAs; previous work in this vein is briefly discussed in Supplementary Note 1. Particularly as efforts to validate a vast number of putative enhancers[5–8] take shape, a clear-eyed understanding of the biases and tradeoffs introduced by MPRA experimental design choices is needed. We performed a systematic comparison by testing the same 2,440 sequences for regulatory activity using nine MPRA strategies, including conventional episomal, self-transcribing active regulatory region sequencing (STARR-seq) and lentiviral designs. We further tested the same sequences in both orientations. Finally, we improved multiplex pairwise assembly[25] and applied it to test differently sized versions of the same enhancers. Our results quantify the impact of MPRA experimental design choices and provide further insight into the nature of enhancers.

Results

Implementation and testing of nine MPRA strategies. We sought to systematically compare nine MPRA strategies (Fig. 1). A first strategy is related to the 'classic' MPRA, using the pGL4.23c vector, wherein the enhancer library resides upstream of a minimal promoter and the associated barcodes reside in the 3′ UTR of the reporter gene (pGL4)[10,26]. A second pair of strategies is related to STARR-seq, wherein the enhancer library resides in the 3′ UTR of the reporter gene, either as originally described (human STARR-seq; HSS)[16] or using the bacterial origin of replication for transcriptional initiation (ORI)[22]. In both cases, we introduce barcodes immediately adjacent to the enhancers in the 3′ UTR to facilitate consistent procedures with other assays. A third set of strategies is related to LentiMPRA, wherein lentiviral integration is used to mitigate concerns about potential differences in chromatin between episomes versus chromosomes, either with the enhancer library upstream of

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Calico Life Sciences LLC, South San Francisco, CA, USA. [3]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA. [4]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA. [5]Berlin Institute of Health (BIH), Berlin, Germany. [6]Charité - Universitätsmedizin Berlin, Berlin, Germany. [7]Howard Hughes Medical Institute, Seattle, WA, USA. [8]Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA, USA. [9]Present address: Memorial Sloan Kettering Cancer Center, New York, NY, USA. [10]Present address: Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan. [11]These authors contributed equally: Jason C. Klein, Vikram Agarwal, Fumitaka Inoue, Aidan Keith. ✉e-mail: nadav.ahituv@ucsf.edu; shendure@uw.edu

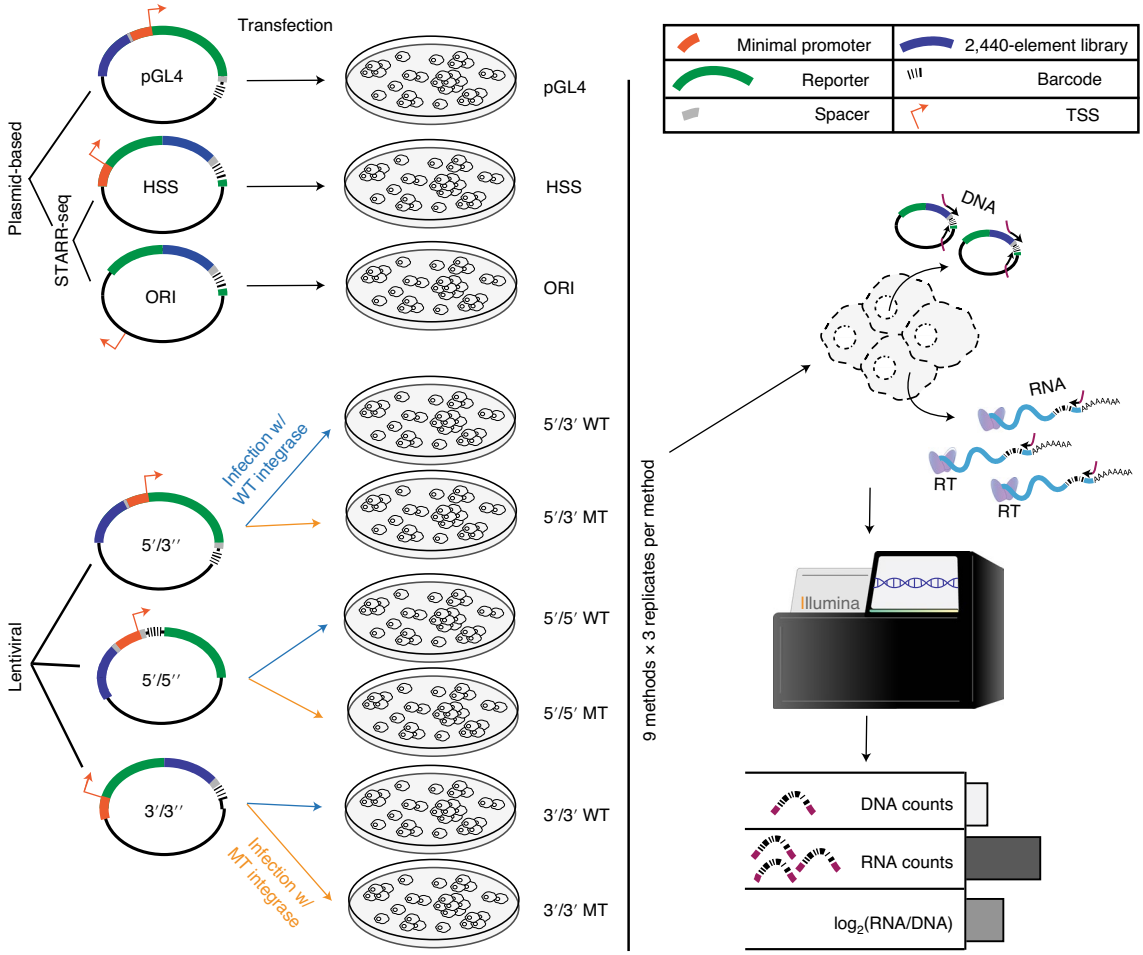

Fig. 1 | Nine MPRA strategies and experimental workflow. Nine different MPRA designs are schematically represented on the left and, from top to bottom, include: pGL4.23c (pGL4); the original STARR-seq vector (HSS); STARR-seq with no minimal promoter (ORI); and lentiMPRAs with the enhancer library upstream of the minimal promoter and the associated barcodes in the 3′ UTR of the reporter gene (5′/3′), the enhancer library upstream of the minimal promoter and barcodes in the 5′ UTR of the reporter (5′/5′), or both the enhancer library and the barcodes in the 3′ UTR of the reporter (3′/3′). The episomal designs (pGL4, HSS and ORI) were transfected into HepG2 cells, while 5′/5′, 5′/3′ and 3′/3′ were packaged with either WT or MT integrase and infected into HepG2 cells. DNA and RNA were extracted from the cells, the enhancer-associated barcodes were amplified and sequenced, and a normalized activity score for each element was computed on the basis of the counts.

the minimal promoter and the associated barcodes in the 3′ UTR of the reporter (5′/3′ wild-type (WT))[18], the enhancer library upstream of the minimal promoter and the barcodes in the 5′ UTR of the reporter (5′/5′ WT), or both the enhancer library and the barcodes in the 3′ UTR of the reporter (3′/3′ WT). The 5′/5′ WT design was developed to address distance-dependent template switching before lentiviral integration[27,28], as it reduces the distance between the enhancer and barcode from 801 to 102 bp. The 3′/3′ WT design is analogous to STARR-seq, but is integrated into the genome and also addresses template switching by positioning the enhancer and barcode immediately adjacent to one another. A fourth set of designs are identical to the three lentiMPRA designs, except that the vector harbors a mutant (MT) integrase, such that the constructs remain episomal (5′/3′ MT, 5′/5′ MT, 3′/3′ MT)[18].

For a common set of sequences to test, we turned to a previously developed library[18] consisting of 2,236 candidate enhancer sequences based on HepG2 chromatin immunoprecipitation sequencing (ChIP-seq) peaks, along with 204 controls (Supplementary Table 1). Of these, 281 overlapped promoters (±1 kb of the transcription start sites (TSSs) of protein-coding genes). The controls consist of synthetically designed sequences that previously demonstrated enhancer MPRA activity (100 positives) or lack thereof (100 negatives)[29] in HepG2 cells, along with 2 positive and 2 negative controls derived from endogenous sequences that were previously validated with luciferase assays[18]. All sequences were 171 bp and synthesized on a microarray together with common flanking sequences. A 15-bp degenerate barcode was appended during PCR amplification and amplicons were cloned to the HSS vector. The enhancer/barcode region of the HSS library was amplified and used for two purposes; first, it was sequenced to link barcodes to enhancers; second, the amplicons were cloned at high complexity into other vectors to create libraries for the remaining eight MPRA designs (Supplementary Fig. 1). As such, the relative abundances of enhancers and barcodes, as well as the enhancer–barcode associations, were consistent across all MPRA libraries. Cloning details and references for each of the nine assay designs are provided in the Methods.

Plasmid libraries were transfected into HepG2 cells in triplicate (three different days). LentiMPRA libraries were packaged with either WT or MT integrase lentivirus and infected into HepG2 in triplicate (three different days). We extracted DNA and RNA, amplified barcodes via PCR and PCR with reverse transcription (RT), respectively, and sequenced amplicons to generate barcode

counts (Fig. 1). An activity score for each element was calculated as the $\log_2$ of the normalized count of RNA molecules from all barcodes corresponding to the element, divided by the normalized number of DNA molecules from all barcodes corresponding to the element (Supplementary Table 2). For each of the 27 experiments (nine assays with three replicates), only barcodes observed in both RNA and DNA were considered. For 26 of 27 experiments (all but 3′/3′ MT replicate 1), the median number of barcode counts per element was >100 (Supplementary Fig. 2).

Comparing results from different MPRA designs. We first sought to evaluate the technical reproducibility of each assay. Most assays were highly correlated between the three replicates. Specifically, intra-assay Pearson correlations for pairwise comparisons of activity scores of replicates exceeded 0.90 for all assays except for 5′/3′ MT (mean $r=0.87$) and 3′/3′ MT (mean $r=0.54$) assays (Fig. 2a and Supplementary Fig. 3a). We also confirmed correlations for 5′/3′ WT and 5′/3′ MT between this and our previous study[18] ($r=0.92$ for 5′/3′ WT and $r=0.81$ for 5′/3′ MT; Supplementary Fig. 3b).

We next sought to compare the results of the various assay designs to one another. We calculated the average activity score for each element across all technical replicates of a given assay (Supplementary Table 2) and then compared the assays to one another. Six of the nine assays demonstrated inter-assay Pearson and Spearman correlations of >0.7 with all other members of this group (Fig. 2b and Supplementary Figs. 4 and 5). These were the ORI and pGL4, together with both WT and MT versions of the 5′/5′ and 5′/3′ assays. The remaining three assays (3′/3′ MT, 3′/3′ WT and HSS) did not show good agreement with the other six assays, nor with one another.

As a different approach to compare assays, we subjected activity scores from all 27 experiments (nine assays with three replicates) to principal component analysis (Fig. 2c). The aforementioned six assays with inter-assay correlations of >0.7 clustered closely to one another. Notably, principal component 1 (PC1) tended to separate the assays wherein the enhancer resides upstream of the minimal promoter (5′/5′, 5′/3′, and pGL4) from those wherein it resides 3′ of the reporter gene (3′/3′, HSS and ORI). In contrast, principal component 2 (PC2) tended to separate lentiviral designs (5′/5′, 5′/3′ and 3′/3′) from plasmid-based designs (pGL4, HSS and ORI). This suggests systematic differences in the enhancer activity measurements that relate to aspects of MPRA design. It also highlights that the location of the candidate enhancer on the plasmid backbone plays a larger role in differential activity than does the episomal versus integrated aspect of the assay.

Next, we examined the dynamic range of activity scores (Fig. 2d). Of note, 3′/3′ MT was removed from further analyses due to comparatively poor technical reproducibility (mean $r=0.54$). The classic enhancer reporter vector (pGL4) and the promoterless STARR-seq assay (ORI) exhibited the greatest dynamic range, with pGL4 showing the largest separation between positive and negative controls (two-sided t-statistic = 37.46). Among the lentiviral assays, the 5′/5′ WT design exhibited the greatest dynamic range and separation of controls (two-sided t-statistic = 30.92).

We generated lasso regression models based on 915 biochemical, evolutionary, and sequence-derived features (Supplementary Tables 3 and 4) using tenfold cross-validation. We were able to predict enhancer activities for the six aforementioned assays (Pearson r ranging from 0.59 for 5′/3′ WT to 0.71 for pGL4) (Supplementary Fig. 6a,b). In general, strong enhancers tended to be underpredicted by the model, whereas weak enhancers tended to be overpredicted.

Many of the top coefficients fitted by these models correspond to ChIP-seq signal or sequence-based binding site predictions for transcriptional activators, coactivators and repressors (Supplementary Fig. 6c,d and Supplementary Table 5). We caution that the interpretation of feature selection and coefficient-based ranking is inherently limited by substantial multicollinearity among features (Supplementary Table 4), which in turn limits the determination of which features are mechanistically or causally involved. Potential reasons for inter-feature correlations are summarized in Supplementary Note 2.

We next sought to ask whether we could predict differences in enhancer activity between the assays on the basis of the same 915 features. For models predicting pairwise differences between the results of the pGL4, 5′/5′ WT, 3′/3′ WT and ORI assays, we were able to achieve correlations of 0.4–0.5 (Fig. 3a and Supplementary Fig. 7a). We were particularly interested in whether features corresponding to RNA-binding proteins and splicing factors would be especially predictive of promoterless STARR-seq (ORI) or 3′/3′ WT results, as in these assays the enhancer itself is included in the 3′ UTR. Indeed, SRSF1/2, BRUNOL4, PTBP1, PPRC1, KHDRBS2, SYNCRIP and MBNL1, which are known to modulate mRNA stability and splicing, predict differences in measured activity in ORI or 3′/3′ WT versus 5′/5′ WT or pGL4 (Fig. 3b,c, Supplementary Fig. 7b,c and Supplementary Table 5). Of note, SRSF1/2, PTBP1, PRPC1, SYNCRIP and MBNL1 are all expressed in liver[30] and could therefore influence MPRA results in HepG2. Additionally, several promoter-binding proteins (TEAD1, TEAD3, NRSF1, JUN and YY1), all expressed in the liver, favor pGL4 and 5′/5′ WT, whereas CCAAT-enhancer-binding proteins favor HSS and ORI. This may correspond to a tradeoff wherein conventional MPRAs are biased toward testing for promoter-like activity, whereas STARR-seq MPRAs are biased by mRNA stability and splicing factors.

Next, we examined differences between episomal versus integrated assays. We note that FOXP1 is more predictive of integrated activity, while ETS-variant transcription factors are more predictive of episomal activity, suggesting that these or correlated factors play a differential role in episomal versus integrated contexts (Fig. 3b,c and Supplementary Fig. 7b,c).

Notably, general transcriptional activity, as measured by cap analysis gene expression (CAGE)[31], was among the most predictive features of the 3′/3′ WT assay (Supplementary Fig. 6c). As this is the only assay where the tested elements are both genomically integrated and distally located from the promoter, this observation suggests that CAGE-based transcriptional activity may be a good predictor of distal enhancer activity[32,33].

Enhancer activity is largely, but not completely, independent of sequence orientation. We next set out to test a key aspect of the canonical definition of enhancers, that they function independently of their orientation with respect to the promoter. We directionally cloned 2,336 sequences (2,236 candidates described above extended out to a 192-bp genomic reference sequence, along with 50 positive and 50 negative controls from Vockley et al.[12]), in both orientations into the pGL4 vector, pooled these libraries, and transfected HepG2 cells in quadruplicate (Fig. 4a). The median number of barcode counts per element was >100 (Supplementary Fig. 8) and the measured activities were reproducible (Pearson $r>0.98$; Fig. 4b, Supplementary Fig. 9 and Supplementary Table 6). Notably, enhancer activities for the same elements cloned in forward versus reverse orientation to the pGL4 vector were also highly correlated (mean $r=0.88$) but less so than same-orientation comparisons ($r>0.98$; Fig. 4b). This suggests that enhancer activity in reporters is largely, but not completely, independent of orientation.

In contrast with enhancers, promoters are established to be directional[34,35]. Overall, 266 of 281 promoter-overlapping elements were successfully measured in both orientations. We tested whether these behaved differently than 1,953 more distally located elements. Indeed, the promoter-overlapping sequences exhibited greater differences in activity between the two orientations than distal elements, supporting the conclusion that they contain signals to promote transcription in an asymmetric fashion (Fig. 4c,d).

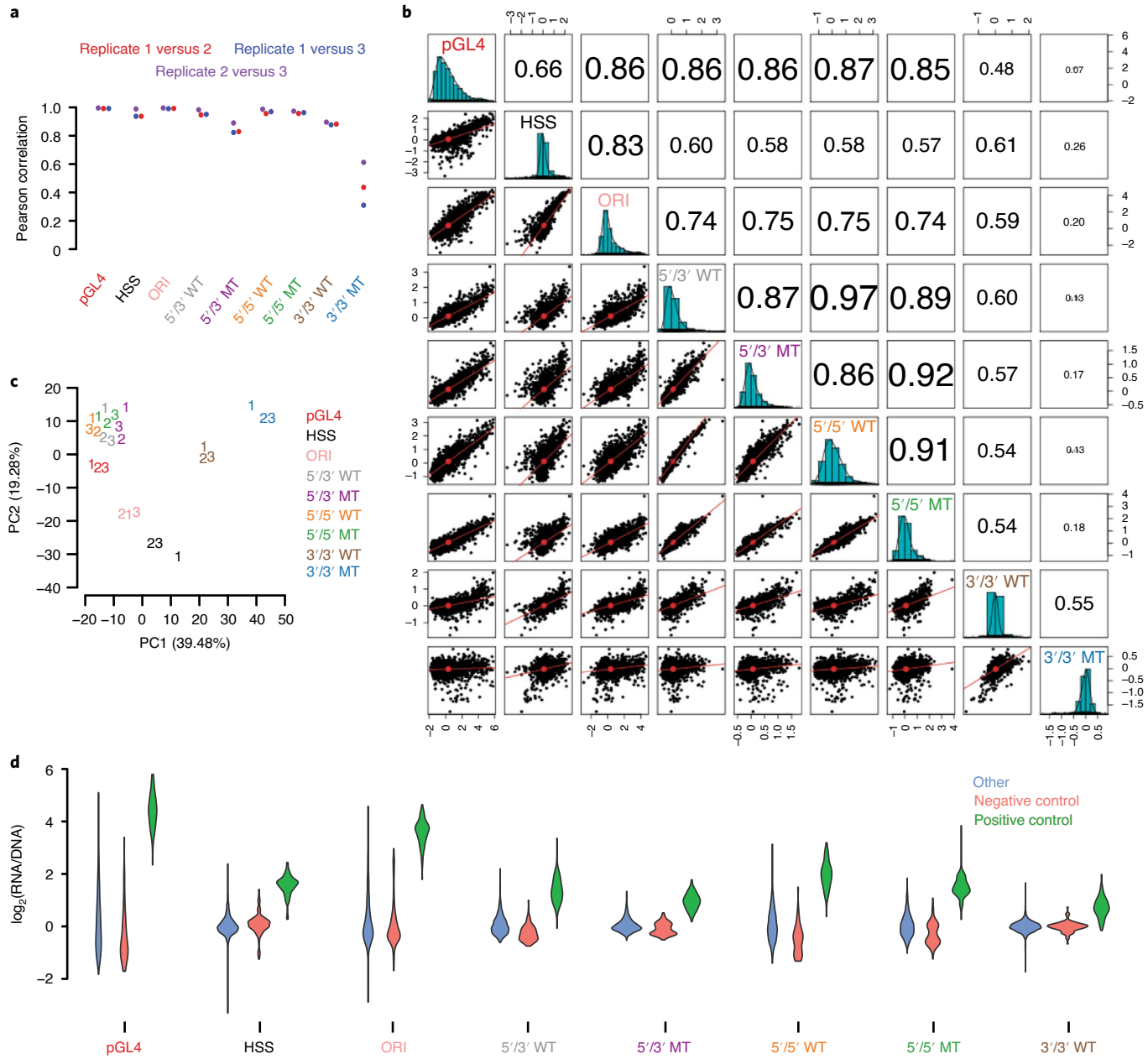

Fig. 2 | Quantitative comparison of different MPRA strategies. a, Beeswarm plot of the Pearson correlation values for each of the three possible pairwise comparisons among the replicates of each MPRA technique. **b**, Scatter matrix displaying scatter plots corresponding to each of the 36 pairs of possible inter-assay comparisons (lower diagonal elements). Shown on the diagonal is a histogram of the $\log_2$(RNA/DNA) ratios, averaged among replicate samples. 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). See Supplementary Fig. 5 for equivalent but with Spearman correlations. **c**, Principal component analysis of 27 experiments (three replicates of nine different MPRA designs). Shown are the first two PCs that together explain over half of the variation. Slight jitter was added to each data point to enhance readability. **d**, Violin plots displaying the distribution of average $\log_2$(RNA/DNA) ratios across independent transfections for positive controls, negative controls and putative enhancer sequences tested, for each of the nine assays.

Appending sequence context leads to differences in the results of MPRAs. Most MPRAs use array-synthesized libraries that are, for technical reasons, limited in length, typically to fewer than 200 bp. To evaluate the impact of this length restriction, we designed 192-bp ('short'), 354-bp ('medium') and 678-bp ('long') versions of our candidate enhancer library, centered at the same genomic position and corresponding to the equivalent 2,236 candidate enhancers tested above (including more flanking sequence from reference genome; Supplementary Table 1). We also included 50 high- and low-scoring putative elements from Vockley et al.[12] in the short and medium libraries (excluded from long libraries because they were all shorter than 678 bp).

The 192-bp versions of these candidate enhancers were synthesized directly on a microarray; sequencing showed a 100% yield (2,336 of 2,336) and a 3.8-fold interquartile range (IQR) for relative abundance (Supplementary Fig. 10a). To generate 354-bp versions, we performed our previously published multiplex pairwise assembly[25] on overlapping pairs of array-synthesized 192-bp fragments (95% yield (2,241 of 2,336); 4.9-fold IQR; Supplementary Fig. 10a). Finally, to generate the 678-bp versions, we developed

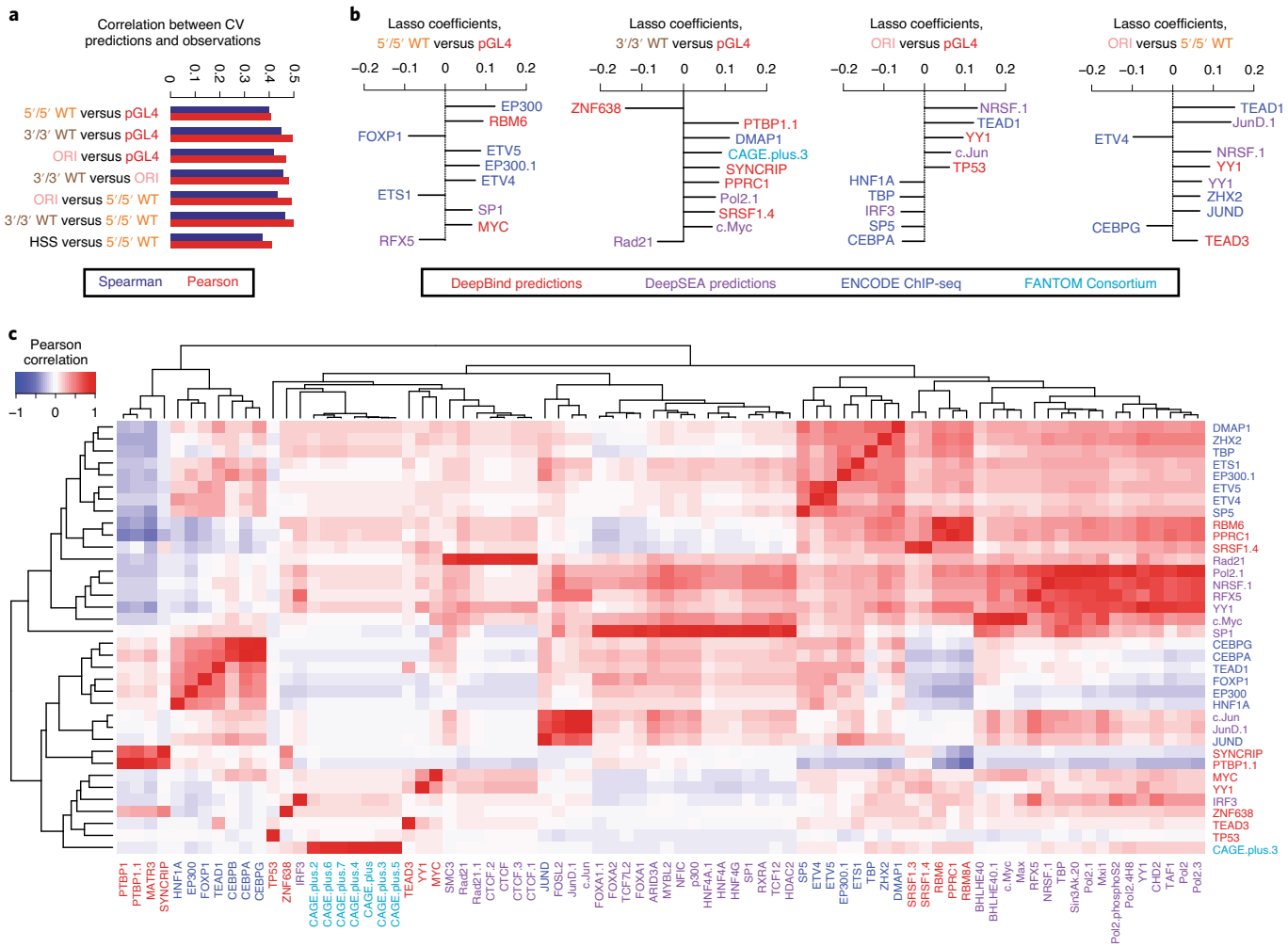

Fig. 3 | Predictive modeling of the ratios and differences between MPRA methods. a, Pearson and Spearman correlation coefficients for tenfold cross-validated (CV) predictions derived from lasso regression models and the observed RNA/DNA ratios, for each of the seven indicated differential comparisons tested. Also indicated are the Pearson (r) and Spearman (ρ) correlation values. **b**, The top ten coefficients derived from lasso regression models trained on the full dataset to predict observed differences in the indicated pairs of MPRA methods. Features with the extensions .1, .2, etc. allude to redundant features or replicate samples. **c**, Pearson correlation matrix between the union of all top ten features from **b**, shown as rows and other features sharing a Pearson correlation either ≤ -0.8 or ≥ 0.8, shown as columns. Feature names are colored according to the origin of the feature as shown in the boxed key. Hierarchical clustering was used to group features exhibiting similar correlation patterns.

a 'two-round' version of MPA that we call hierarchical multiplex pairwise assembly (HMPA) (Supplementary Figs. 10b and 11). HMPA of overlapping pairs of array-synthesized 192-bp fragments yielded overlapping pairs of 354-bp fragments, which were further assembled to generate 678-bp fragments (84% yield (1,887 of 2,236); 27.9-fold IQR; Supplementary Fig. 11a). We verified a subset of our long enhancers with PacBio sequencing (Supplementary Fig. 10c,d; chimera rate of 16.5%).

We cloned all three libraries into the pGL4 vector, then pooled and transfected them in quadruplicate to HepG2 cells (Fig. 5a and Supplementary Table 7). Requiring each element to be detected with at least ten unique barcodes, there were 651 candidate enhancers tested at all three lengths. Technical replicates within any given length class were highly reproducible, albeit modestly less so for long elements (mean Pearson $r = 0.94$; Fig. 5b and Supplementary Figs. 12 and 13). However, there was less agreement for the same candidate enhancers tested at different lengths (short versus medium, mean $r = 0.78$; medium versus long, mean $r = 0.67$; short versus long, mean $r = 0.53$; Fig. 5b,c). Finally, we observed that the positive control sequences were significantly more active than the negative controls when tested as either 192-bp or 354-bp fragments ($P < 0.01$, Wilcoxon signed-rank test; Fig. 5d).

We chose ten MPRA-active candidate enhancers to test in individual luciferase assays: five that showed differential activity between their long and medium forms (cyan; Supplementary Fig. 14a) and five that did not (green; Supplementary Fig. 14a). Of the five that showed differential activity, three were active in the luciferase assay (2–4), all concordant with MPRA results (Supplementary Fig. 14b–d). Of the five that did not show differential activity in the MPRA, all were active in the luciferase assay in at least one form and four had differential activity, possibly due to greater sensitivity of the luciferase assay or subtle differences between the constructs (Supplementary Fig. 14b–d). We also tested versions of all ten of these MPRA-active candidates in their long form but with the middle 354 bp deleted; all of these showed insignificant ($n = 8$) or reduced ($n = 2$) activity in the luciferase assay (Supplementary Fig. 14b). Overall, these results highlight the relevance of the lengths and boundaries of elements tested in MPRAs in influencing measured activity.

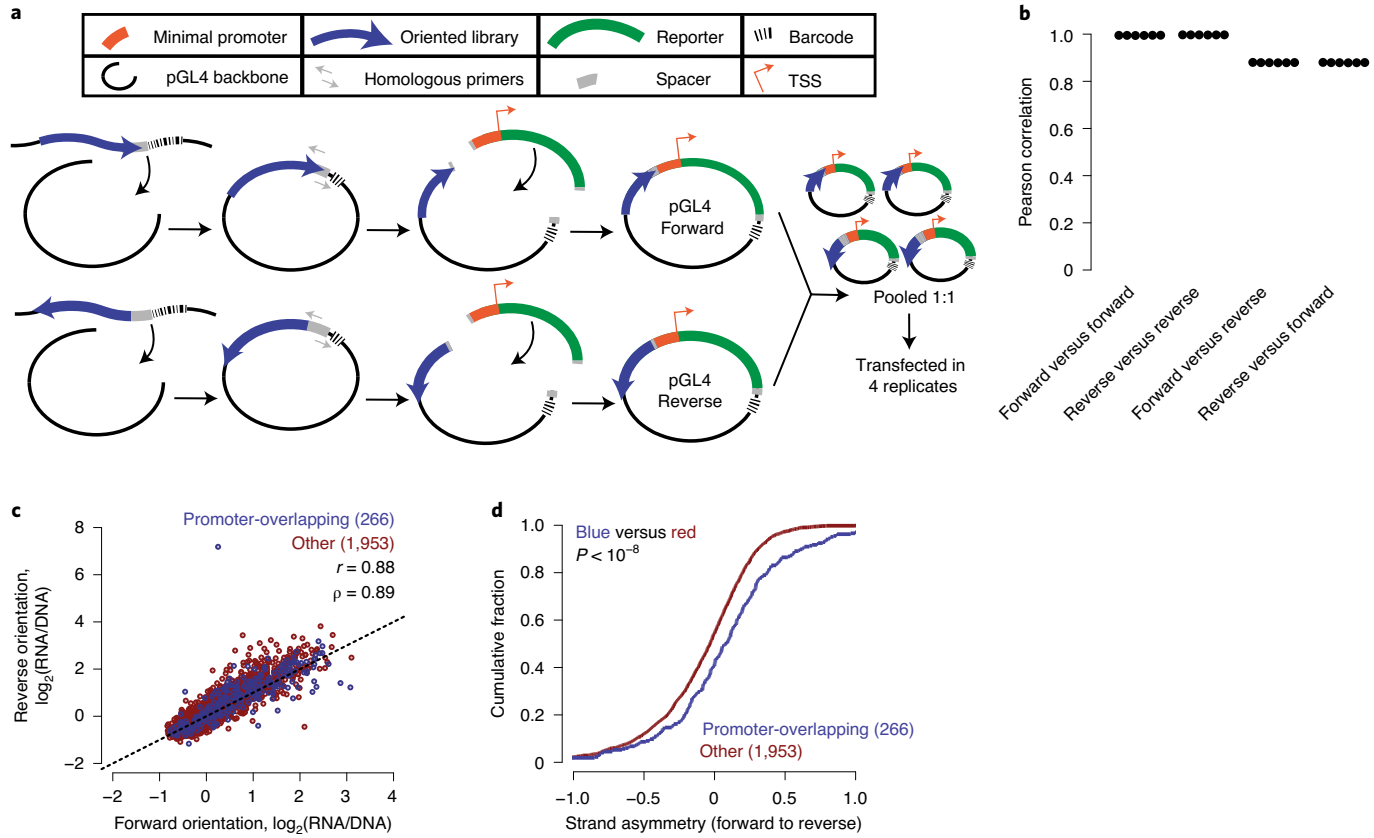

Fig. 4 | Enhancer activity is largely, but not completely, independent of sequence orientation. a, Workflow used to produce an MPRA library with each element in both orientations. The 2,336-element library was cloned into the pGL4 backbone in both orientations as two separate libraries. These were then pooled and transfected into HepG2 cells in quadruplicate. **b**, Beeswarm plot of the Pearson correlations corresponding to each of the six possible pairwise comparisons among the four replicates. The correlations are computed between observed enhancer activity values for elements positioned either in the same (forward versus forward and reverse versus reverse) or opposite (forward versus reverse and reverse versus forward) orientations. **c**, Scatter-plots of the average activity score of each element in the forward versus reverse orientation, split out by promoter-overlapping (blue; ±1 kb of the TSS of a protein-coding gene) and other (red) elements. **d**, Cumulative distributions measuring strand asymmetry between promoter-overlapping elements and other elements. Here, 'forward' and 'reverse' were defined as 'sense' and 'antisense', respectively, in relation to the orientation of the TSS for promoter-overlapping elements ($n = 266$); and were defined as 'plus' and 'minus'-stranded, respectively, in relation to the chromosome annotation for other elements ($n = 1{,}953$). Similarity of the blue distribution to that of the red was tested (one-sided Kolmogorov–Smirnov test, P value).

We trained lasso regression models to predict activities using features, which were re-computed for each of the three size classes (Fig. 6a, Supplementary Fig. 15 and Supplementary Table 4). The lower performance of the model for the long element library is possibly consequent to its fewer sequences, its lower technical reproducibility or an increase in the effect of nonlinear interactions between features that reduce predictive performance. Known predictors of enhancer activity were consistently present in the top coefficients, although their relative rankings differed depending on the size class being examined (Supplementary Fig. 15c and Supplementary Table 5). Next, we sought to explicitly model how differences in predicted factor binding might explain differences in enhancer activity, as measured by different pairs of size classes. For example, in attempting to explain observed activity differences in long versus short elements, we computed a set of features as the differences in predicted binding or measured ChIP-seq signal, between the long element and corresponding short element (for example, ΔARID3A = ARID3A$_{long}$ − ARID3A$_{short}$). Many of the top features originated from sequence-based differences in predicted binding in the extra genomic context surrounding the core element. Features consistently observed to explain activity differences in longer elements include RPC155, the catalytic core and largest component of RNA polymerase III; Jun and FOS, components of the AP-1 complex; ATF2, EZH2 and HDAC1/2, core histone-modifying enzymes; and the transcription factors ARID3A, DRAP1 and SP1/2/3 (Fig. 6b,c, Supplementary Fig. 16 and Supplementary Table 5).

Discussion

Over the past decade, MPRAs have enabled researchers to functionally test large numbers of DNA sequences for regulatory activity and in the process address numerous biological questions. While different groups utilize various backbones and assay designs, there has been no systematic comparison of how these different strategies influence results.

Here, we have sought to perform a systematic comparison of major MPRA strategies and to concurrently investigate the consequences of key design choices such as the assay, element orientation and element length. We generally observe concordance between different MPRA designs, albeit to varying degrees. Six of the nine assays exhibited both technical reproducibility as well as reasonable agreement with one another (pGL4, ORI, 5′/5′ WT, 5′/5′ MT, 5′/3′ WT and 5′/3′ MT). Furthermore, as we previously showed for the 5′/3′ WT and 5′/3′ MT assays, enhancer activities as measured by MPRAs[18] are reasonably well predicted by models based on primary sequence together with biochemical measurements at the corresponding genomic locations. Taken together, our results

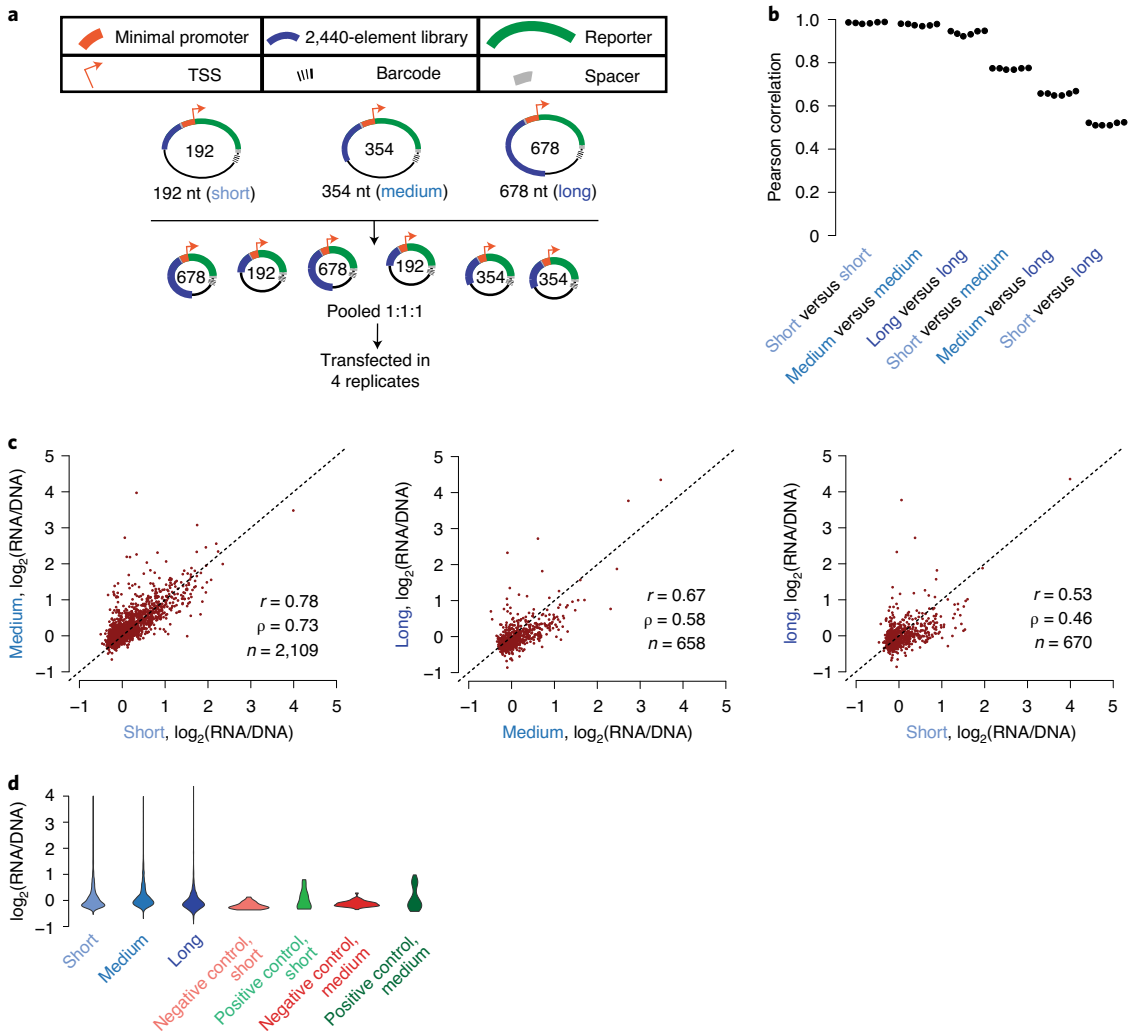

Fig. 5 | Including additional sequence context around tested elements leads to differences in the results of MPRAs. a, Experimental schematic. The 192-bp, 354-bp and 678-bp libraries were synthesized, assembled and cloned into the pGL4 backbone. These were pooled and transfected into HepG2 cells in quadruplicate. nt, nucleotide. **b**, Beeswarm plot of the Pearson correlation values corresponding to each of the six possible pairwise comparisons among the four replicates. The correlations are computed between observed enhancer activity values for elements measured in each of the three possible size classes. **c**, Scatter-plots of the average activity score of each element, comparing short versus medium, medium versus long, and short versus long versions of each element, and restricting to elements detected with at least ten unique barcodes at both lengths (*n*). **d**, Violin plot displaying the distribution of average $\log_2$(RNA/DNA) ratios for short, medium and long versions of the elements tested, as well as for positive and negative controls at short and medium lengths.

support a view wherein diverse MPRAs are all measuring enhancer activity, but design differences (such as integrated versus episomal; 5′ versus 3′ location of the enhancer) influence the results to a modest degree. For example, features influencing mRNA stability and splicing favor assays with the enhancer transcribed in the 3′ UTR (ORI and 3′/3′ WT), whereas promoter-binding transcription factors favor assays with the enhancer upstream of the promoter (pGL4 and 5′/5′ WT).

Overall, our results support a preference for three of the nine MPRA designs evaluated here (pGL4, ORI and 5′/5′ WT), which all had reasonable inter-assay correlations. The pGL4 assay has the advantage of representing the 'classic' enhancer reporter assay design, had the greatest dynamic range and was the most predictable with our lasso regression, but had the disadvantages of being episomal rather than integrated and of confounding enhancer activity with possible effects from promoter-binding proteins. The ORI assay (promoterless STARR-seq) has the advantage of eliminating the need to associate barcodes, potentially allowing for greater library complexities, and has a large dynamic range, but has the disadvantages of confounding enhancer activity with possible effects on messenger RNA splicing and/or stability, and also of being episomal rather than integrated. The 5′/5′ WT assay has the advantage of being integrated rather than episomal and, among lentiviral assays, mitigates the template switching issue by minimizing the distance between the enhancer and barcode. However, template switching still occurs to some degree, the assay exhibits a lower dynamic range than pGL4 or ORI assays and has similar potential for bias from promoter-binding proteins as pGL4.

A caveat of our HSS and ORI experiments is that by incorporating a barcode downstream of the enhancer, we introduced the possibility that barcode counts include short transcripts initiating within the candidate enhancer itself. Further exploration of this potential confounder, including additional experiments, is summarized in Supplementary Note 3.

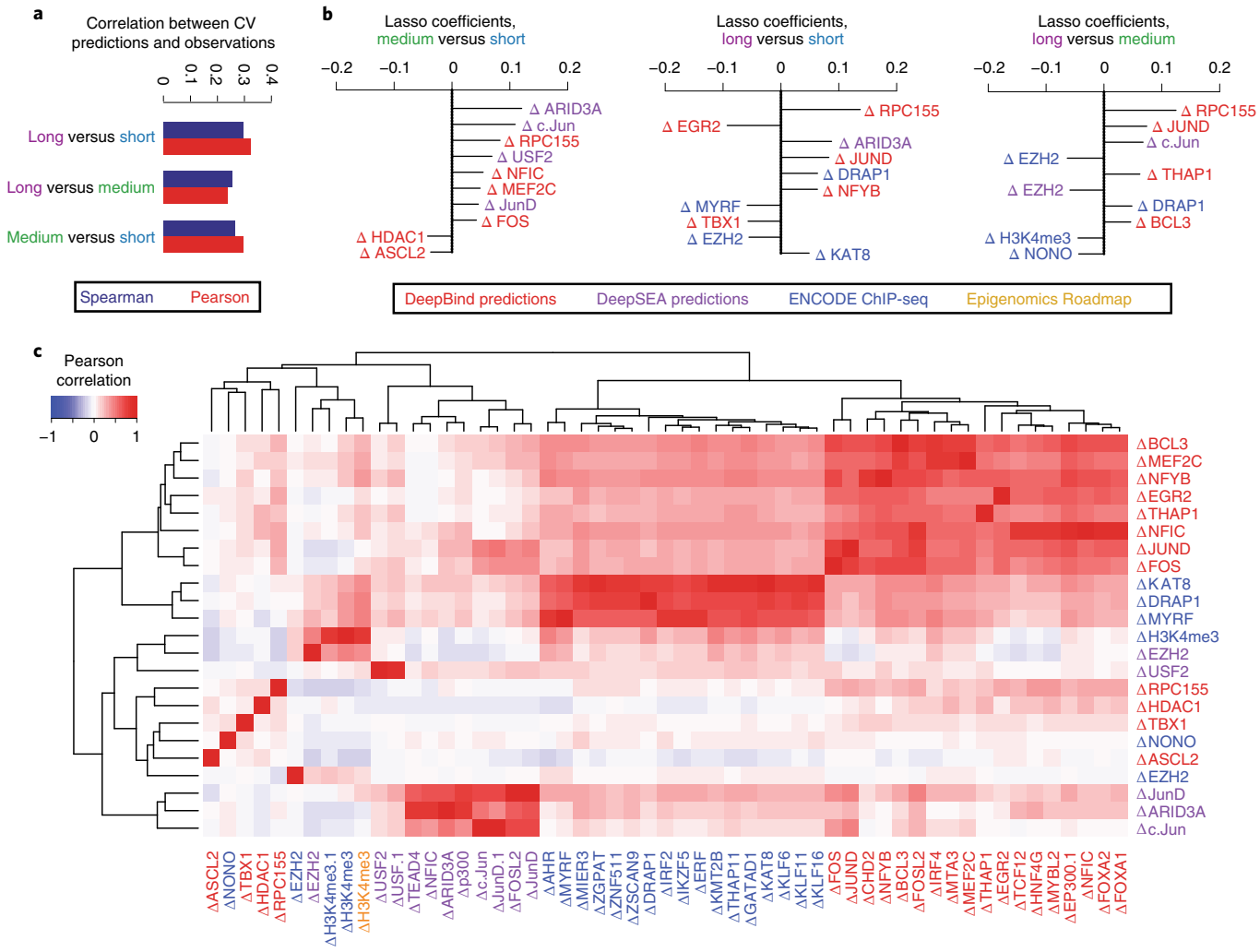

Fig. 6 | Predictive modeling of factors dependent on element size. a, Pearson and Spearman values between the tenfold CV predictions and observed values for each of the three size classes tested. **b**, The top ten coefficients derived from lasso regression models trained on the full dataset to predict observed values from the differential size comparisons indicated. Features with the extension '.1', '.2', etc. allude to redundant features or replicate samples. **c**, Pearson correlation matrix between the union of all top ten features from **b**, shown as rows and other features sharing a Pearson correlation either ≤−0.8 or ≥0.8, shown as columns. Feature names are colored according to the origin of the feature as shown in the boxed key. Hierarchical clustering was used to group features exhibiting similar correlation patterns.

Another key finding is our confirmation that the activity of enhancers, is largely, but not completely, independent of orientation, at least as measured for our subset of candidate enhancers tested using the pGL4 vector. This is of course part of the original definition of enhancers[1], but efforts to systematically test the validity of this assumption across a large number of sequences have been limited[16,36,37]. Previously, a subset of preinitiation-complex-bound enhancers were shown to have strong orientation-dependent activity, highlighting that these trends may be influenced by the choice of elements tested[38]. Candidate enhancer sequences derived from the vicinity of TSSs exhibited greater directionality, consistent with a subset of these bearing features of oriented promoters.

Finally, we developed improved methods to efficiently assemble longer DNA fragments from array-synthesized oligonucleotides and applied them to evaluate the extent to which including additional sequence context around tested elements impacts MPRA results. We successfully assembled 95% of 2,336×354-bp targets using MPA, compared to just 71% of 2,271×192–252-bp targets in our original description of the method[25]. Moreover, our HMPA is a protocol that in vitro assembles thousands of sequences, each over 600 bp, as a single library. In this manuscript, we synthesized >600 elements, each 678 bp, for 1–2% of what it would have cost from commercial vendors. Unlike potential alternatives, the method does not require specialized equipment, making it more widely accessible[39].

The sub-200-bp length of sub-sequences typically tested is a choice related to the technical limits of microarray-based synthesis. In the genome, there are no such limits and it remains unclear what the appropriate 'enhancer size' is to test in MPRAs and whether this choice matters. To evaluate this, we tested candidate enhancers at three different lengths. We observed correlations between the same elements tested at all lengths, but these correlations clearly drop off as a function of length difference. At the extreme, the activities of 678-bp versus 192-bp versions of the same candidate enhancers were more poorly correlated than nearly all of our inter-assay comparisons (Pearson $r = 0.53$, Spearman $\rho = 0.46$). Furthermore, these data suggest that the longer sequences are adding biologically relevant signal, as features corresponding to relevant transcription factors explain differences in activity of longer versus shorter sequences.

For example, a feature corresponding to RPC155, the catalytic subunit of RNA polymerase III, is the strongest coefficient separating the 678-bp constructs from the 192-bp and 354-bp constructs and also one of the stronger coefficients separating the 354-bp from 192-bp constructs. Although it is challenging to offer strict guidance in the absence of in vivo ground truth, we recommend testing longer sequences when possible.

In conclusion, we set out to rank the relative contribution of assay design, orientation and length on the results of MPRAs. Our results suggest a degree of caution in interpreting the results of all MPRAs, as they are all subject to influence by aspects of the assay design. We found that sequence length had the greatest effect, followed by assay design and finally orientation. Although MPRAs of high-complexity genome-wide fragment libraries are not length limited[16,22], MPRAs of designed libraries largely still are. For designed libraries in particular, further work is necessary to develop or improve methods such as HMPA to facilitate the construction of complex, uniform MPRA libraries of longer sequences, as well as to further explore the optimal parameters of element design (such as length and centering).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0965-y.

Received: 8 July 2020; Accepted: 27 August 2020;
Published online: 12 October 2020

References

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
2. Moreau, P. et al. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.* **9**, 6047–6068 (1981).
3. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729–740 (1983).
4. Neuberger, M. S. Expression and regulation of immunoglobulin heavy chain gene transfected into lymphoid cells. *EMBO J.* **2**, 1373–1378 (1983).
5. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
6. Kawaji, H., Kasukawa, T., Forrest, A., Carninci, P. & Hayashizaki, Y. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci. Data* **4**, 170113 (2017).
7. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
8. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
9. Patwardhan, R. P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
10. Patwardhan, R. P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
11. Melnikov, A. et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271 (2012).
12. Vockley, C. M. et al. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).
13. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **172**, 1132–1134 (2018).
14. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
15. Liu, S. et al. Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 194 (2017).
16. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
17. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
18. Inoue, F. et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).
19. Klein, J. C. et al. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 2434 (2019).
20. Arnold, C. D. et al. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during *cis*-regulatory evolution. *Nat. Genet.* **46**, 685–692 (2014).
21. Klein, J. C., Keith, A., Agarwal, V., Durham, T. & Shendure, J. Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.* **19**, 99 (2018).
22. Muerdter, F. et al. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).
23. Vanhille, L. et al. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* **6**, 6905 (2015).
24. Wang, X. et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9**, 5380 (2018).
25. Klein, J. C. et al. Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2016).
26. Kircher, M. et al. Saturation mutagenesis of disease-associated regulatory elements. *Nat. Commun.* **10**, 3583 (2019).
27. Hill, A. J. et al. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).
28. Sack, L. M., Davoli, T., Xu, Q., Li, M. Z. & Elledge, S. J. Sources of error in mammalian genetic screens. *G3* **6**, 2781–2790 (2016).
29. Smith, R. P. et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
30. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
31. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
32. FANTOM Consortium et al. Supplementary figures, tables and texts for FANTOM 5 phase 2. *Figshare* https://doi.org/10.6084/m9.figshare.1288777 (2015).
33. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
34. Engreitz, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
35. van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.3754 (2016).
36. Kvon, E. Z., Stampfel, G., Yáñez-Cuna, J. O., Dickson, B. J. & Stark, A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev.* **26**, 908–913 (2012).
37. Mikhaylichenko, O. et al. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev.* **32**, 42–57 (2018).
38. Weingarten-Gabbay, S. et al. Systematic interrogation of human promoters. *Genome Res.* **29**, 171–183 (2019).
39. Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343–347 (2018).

Methods

Design, barcoding and cloning of the enhancer library into the HSS vector. We used an existing array library from Inoue et al.[18]. This library consists of 2,440 unique 171-bp candidate enhancer sequences, based on ChIP-seq peaks in HepG2. Each sequence was flanked with a 15-bp sequence on the 5′ end (Original_Array_5adapter) and a 44-bp sequence on the 3′ end (Original_Array_3adapter) (Supplementary Table 8). More detail on enhancer design can be found in that manuscript[18]. We first amplified the library using the following primers: STARR-Seq-AG-f and spacer-AG-r (Supplementary Table 8). These amplify the library, excluding the previously designed barcodes, while adding homology to the STARR-seq vector (Addgene, 71509)[16] on the 5′ end and a spacer sequence on the 3′ end that we use for all subsequent libraries. We amplified 10 ng of array oligonucleotides with KAPA HiFi 2× Readymix (Kapa Biosystems) with a 65 °C annealing temperature and 30 s extension, following the manufacturer's protocol. We followed the reaction in real time using Sybr Green (Thermo Fisher Scientific) and stopped the reaction before plateauing, after ten cycles. We then purified the PCR product with a 1.8× AMPure XP (Beckman Coulter) cleanup and eluted in 50 µl of Qiagen Elution Buffer (EB), following the manufacturer's protocol. We took 1 µl of purified PCR product and amplified in triplicate a second reaction using KAPA HiFi 2× Readymix using primers STARR-Seq-AG-f and STARR-BC-spacer-r with a 35 s extension time and 65 °C annealing temperature for eight cycles (Supplementary Table 8). This round of PCR added a 15-bp degenerate barcode on the 3′ end of the spacer as well as homology arms to the 3′ end of the HSS vector. We then pooled the three reactions together, ran the products on a 1.5% agarose gel, and gel extracted the amplicon using the QIAquick Gel Extraction kit (Qiagen), following the manufacturer's protocol, eluting in 17.5 µl of Qiagen EB. We then cloned a 2:1 molar excess of our gel-extracted insert into 100 ng of the HSS vector (linearized with *Age*I and *Sal*I) with the NEBuilder HiFi DNA Assembly Cloning kit (NEB), following the manufacturer's protocol. We transformed 10-β electrocompetent cells (NEB C3020) with the plasmids in duplicate following the manufacturer's protocol, along with a no-insert negative control. We pooled the two transformations during recovery and plated 15 µl to estimate complexity. The following day, we estimated complexity as approximately 750,000 and grew a third of the transformation to represent a library of 250,000 in 100 ml of LB + ampicillin (Amp), so that each candidate enhancer is expected to associate on average with 100 different barcodes. We extracted the plasmid using the ZymoPURE II Plasmid Midiprep kit (Zymo Research).

Barcode association library for the nine MPRA assays. We amplified 5 ng of the HSS library with the following primers: P5-STARR-AG-ass-f and P7-STARR-ass-r (Supplementary Table 8). These primers add a sample-specific barcode and Illumina flow cell adaptors. We then spiked the library into a NextSeq Mid 300 Cycle kit with paired-end 149-bp reads and a 20-bp index read (which captured the 15-bp barcode as well as 5 bp of extra sequence to help filter for read quality), using the following custom primers: Read1 as STARR-AG-seq-R1, Read2 as spacer-seq-R2, Index1 as pLS-mP-ass-seq-ind1, and Index2 as STARR-AG-ind2 (Supplementary Table 8).

Library cloning. *From HSS to ORI vector.* We amplified 5 ng of the HSS library with the following primers: STARR-Seq-AG-f and STARR-Seq-AG-r (Supplementary Table 8) using KAPA HiFi 2× Readymix (Kapa Biosystems) with a 65 °C annealing temperature and 30 s extension. These primers amplify both candidate enhancers and previously assigned degenerate barcodes, and add homology arms to the ORI vector (Addgene, 99296)[22]. We followed the reaction in real time with Sybr Green (Thermo Fisher Scientific) and stopped the reaction before plateauing at 13 cycles. We gel extracted the amplicon on a 1.5% agarose gel as described above. We then cloned the library in a 2:1 molar excess into 100 ng of the hSTARR-seq_ORI vector (Addgene, 99296), linearized with *Age*I and *Sal*I, using the NEBuilder HiFi DNA Assembly Cloning kit (NEB), following the manufacturer's protocol. We then transformed 10-β electrocompetent cells (NEB C3020) with the plasmids in duplicate following the manufacturer's protocol, along with a no-insert negative control. We pooled the two positive transformations during recovery, plated 15 µl to estimate complexity and grew the remainder of the culture in 100 ml LB + Amp. The following day, we estimated the complexity as >500,000 and extracted the plasmid using the ZymoPURE II Plasmid Midiprep kit (Zymo Research).

From HSS to pGL4.23c MPRA vector. As described above, we amplified 5 ng of the HSS library with the following primers: pGL423c-AG-1f and pGL423c-AG-1r (Supplementary Table 8). These primers amplify both candidate enhancers and previously assigned degenerate barcodes, and add homology arms to the pGL4.23c MPRA vector (GenBank MK484105). We stopped the reaction before plateauing at 18 cycles. We linearized the pGL4.23c MPRA backbone, while removing the minimal promoter and reporter using *Hin*dIII and *Xba*I. We treated the linearized plasmid with Antarctic Phosphatase (NEB) following the manufacturer's protocol, and then gel extracted the plasmid on a 1% agarose gel, as described above. We then cloned our insert into the linearized backbone, transformed, and extracted DNA as described above. We then re-linearized the pGL4.23c backbone, containing our enhancer library with *Sbf*I and *Eco*RI, gel extracted, and inserted our minimal promoter + green fluorescent protein (GFP) cassette, which contains overlaps for *Sbf*I and *Eco*RI.

From HSS to lentiMPRA 5′/5′. We used similar methods as in the pGL4.23c library cloning with the following changes. The HSS library was amplified with pLS-mP-AG-2f and pLS-mP-AG-5r (Supplementary Table 8) for 17 cycles. After gel extraction, we cloned the insert into the pLS-mP (Addgene, 81225)[18], which had been linearized with *Sbf*I and *Age*I and treated with Antarctic Phosphatase. The resulting library was recut with *Sbf*I and *Age*I, residing between the designed candidate enhancer and barcode, and the minimal promoter was ligated in. We generated the minimal promoter with oligos minP_F and minP_R (Supplementary Table 8), which provide overlaps for *Sbf*I and *Age*I. The minimal promoter oligos were phosphorylated and annealed using T4 ligation buffer and T4 polynucleotide kinase (NEB) at 37 °C for 30 min, followed by 95 °C for 5 min, ramping down to 25 °C at 5 °C min^{-1}. We then diluted the annealed oligonucleotides at 1:200 and cloned into the linearized pLS-mP backbone with our enhancer library at a 2:1 molar excess.

From HSS to lentiMPRA 5′/3′. We used similar methods as for the pGL4.23c library with the following changes. The HSS library was amplified with pLS-mP-AG-2f and pLS-mP-AG-3r (Supplementary Table 8) for 17 cycles. After gel extraction, we cloned the insert into the pLS-mP backbone (Addgene, 81225)[18], which had been linearized with *Sbf*I and *Eco*RI and treated with Antarctic Phosphatase. Similarly to the pGL4.23c library, the resulting library was recut with *Sbf*I and *Eco*RI again, and our minimal promoter + GFP cassette inserted, containing overlaps for *Sbf*I and *Eco*RI.

From HSS to lentiMPRA 3′/3′. We used similar methods as for the pGL4.23c library with the following changes. The HSS library was amplified with pLS-mP-AG-3f and pLS-mP-AG-3R (Supplementary Table 8) for 13 cycles. After gel extraction, we cloned the insert into the pLS-mP backbone (Addgene, 81225)[18], which had been linearized with *Eco*RI only and treated with Antarctic Phosphatase.

Cell culture, lentivirus packaging and titration. HEK293T and HepG2 cell culture, lentivirus packaging and titration were performed as previously described with modifications[18]. Briefly, 12 million HEK293T cells were seeded in 15-cm dishes and cultured for 48 h. To generate WT lentiviral libraries (5′/5′ WT, 5′/3′ WT and 3′/3′ WT), the cells were co-transfected with 5.5 µg of lentiMPRA libraries, 1.85 µg of pMD2.G (Addgene, 12259) and 3.65 µg of psPAX2 (Addgene, 12260), which encodes a WT *pol*, using EndoFectin Lenti transfection reagent (GeneCopoeia) according to the manufacturer's instruction. To generate nonintegrating lentiviral libraries (5′/5′ MT, 5′/3′ MT and 3′/3′ MT), pLV-HELP (InvivoGen) that encodes a mutant *pol* was used instead of psPAX2. After 18 h, cell culture medium was refreshed and TiterBoost reagent (GeneCopoeia) was added. The transfected cells were cultured for 2 d and lentivirus collected and concentrated using the Lenti-X concentrator (Takara) according to the manufacturer's protocol. To measure DNA titer for the lentiviral libraries, HepG2 cells were plated at 1×10^5 cells per well in 24-well plates and incubated for 24 h. Serial volume (0, 4, 8 and 16 µl) of the lentivirus was added with 8 µg ml^{-1} polybrene to increase infection efficiency. The infected cells were cultured for 3 d and then washed with PBS three times. Genomic DNA was extracted using the Wizard SV genomic DNA purification kit (Promega). Multiplicity of infection was measured as relative amount of viral DNA (WPRE region, WPRE_F and WPRE_F) over that of genomic DNA (intronic region of *LIPC* gene, LIPC_F and LIPC_R; Supplementary Table 8) by qPCR using SsoFast EvaGreen Supermix (Bio-Rad), according to the manufacturer's protocol.

Transient transfections and lentiviral infections. HepG2 cells were seeded in 10-cm dishes (2.4 million cells per dish) and incubated for 24 h. For plasmid-based MPRA, cells were transfected with 10 µg of the plasmid libraries (HSS, ORI and pGL4) using X-tremeGENE HP (Roche) according to the manufacturer's protocol. The X-tremeGENE:DNA ratio was 2:1. For the lentiMPRA, the cells were infected with the lentiviral libraries (5′/5′ WT/MT, 5′/3′ WT/MT and 3′/3′ WT/MT) along with 8 µg ml^{-1} polybrene, with the estimated multiplicity of infection of 50 for WT and 100 for MT libraries. The cells were incubated for 3 d, washed with PBS three times and genomic DNA and total RNA was extracted using AllPrep DNA/RNA Mini kit (Qiagen). Messenger RNA was purified from total RNA using Oligotex mRNA Mini kit (Qiagen). All experiments for nine libraries were carried out simultaneously to minimize batch effect. Three independent replicate cultures were transfected or infected on different days.

RT–PCR, amplification and sequencing of RNA and DNA. DNA for all experiments was quantified using the Qubit dsDNA Broad Range Assay kit (Thermo Fisher Scientific). For all samples, a total of 12 µg of DNA was split into 24 × 50 µl PCR reactions (each with 500 ng of input DNA) with KAPA2G Robust HotStart ReadyMix (Kapa Biosystems) for three cycles with a 65 °C annealing and 40 s extension, using an indexed P5 primer and a unique molecular identifier (UMI)-containing P7 primer (Supplementary Table 9). After three cycles, reactions were pooled and purified with a 1.8× AMPure cleanup, following the

manufacturer's instructions and eluted in a total of 344 µl of Qiagen EB. The entire purified product was then used for a second round of PCR, split into 16 × 50 µl reactions each, with primers P5 and P7 (Supplementary Table 8). The reaction was followed in real time with Sybr Green (Thermo Fisher Scientific) and stopped before plateauing. PCRs were then pooled and 100 µl of the pooled PCR products was purified with a 0.9× AMPure cleanup and eluted in 30 µl for sequencing.

The mRNA for all experiments was treated with Turbo DNase (Thermo Fisher Scientific) following the manufacturer's instructions and then quantified using the Qubit RNA Assay kit (Thermo Fisher Scientific). For all samples, we performed three 20-µl RT reactions, each with one-third of the sample (up to 500 ng of mRNA). RT was performed using SuperScript IV (Thermo Fisher Scientific) and a gene-specific primer, which attached a UMI (Supplementary Table 9), following the manufacturer's instructions.

Complementary DNA for each sample was split into eight 50-µl PCRs using an indexed P5 primer and P7 (Supplementary Table 8) for three cycles. Reactions were then pooled together and purified with a 1.5× AMPure reaction and eluted in 129 µl of Qiagen EB. The purified PCR product was then split into six 50-µl PCRs with P5 and P7 following in real time with Sybr Green and stopped before plateauing. PCRs were then pooled and 100 µl of the pooled PCR products was purified with a 0.9–1.8× AMPure cleanup, depending on background banding and eluted in 30 µl for sequencing.

Two experiments at a time (each with three DNA replicates and three RNA replicates) were run on a 75-cycle NextSeq 550 v.2 High-Output kit with custom primers for each assay (Supplementary Table 8).

MPRA to evaluate the impact of enhancer orientation. To test enhancers in both orientations relative to the promoter (in the forward and reverse orientations), we synthesized the same 2,236 genomic sequences tested above[18], along with 100 controls previously tested in STARR-seq, which are described below[12]. These sequences were synthesized as 192-bp fragments with HSSF-ATGC and HSS-R (Supplementary Table 8). The forward orientation was amplified in a 50-µl PCR reaction using KAPA HiFi 2× Readymix (Kapa Biosystems) and primers HSS_pGL4_F and HSS_pGL4_R1; the PCR for the reverse orientation used the primers HSS_pGL4_F_orr2 and HSSpGL4_1_orr2 (Supplementary Table 8). PCRs were followed in real time with Sybr Green, stopped before plateauing (seven cycles) and purified in a 1× AMPure reaction, eluting in 25 µl of Qiagen EB. Overall, 1 µl of the purified products were put into a second PCR reaction, which added 15 bp of barcode sequence and homology to the pGL4.23c vector; the forward orientation used primers HSS_pGL4_F and HSS_pGL4_R2 and the reverse orientation used primers HSS_pGL4_F_orr2 and HSS_pGL4_R2 (Supplementary Table 8).

We linearized the pGL4.23c MPRA backbone with *Hin*dIII and *Xba*I (removing the minimal promoter and reporter) and gel extracted the backbone and insert PCR products. Inserts were cloned into the pGL4.23c plasmid using NEBuilder HiFi DNA Assembly Cloning kit (NEB), following the manufacturer's protocol. We transformed 10-β electrocompetent cells (NEB C3020) with the plasmids, grew up transformations in 100 ml of LB + Amp and extracted plasmid libraries using a ZymoPURE II Plasmid Midiprep kit (Zymo Research).

To clone in the minimal promoter and GFP for the forward orientation, 20 ng of the forward backbone was amplified with Len_lib_linF and Len_lib_linR (Supplementary Table 8) using NEBNext High-Fidelity 2× PCR Master Mix (NEB); the minimal promoter and GFP was amplified from 10 ng of the pLS-mP plasmid using minGFP_Len_HAF and minGFP_Len_HAR (Supplementary Table 8). For the reverse orientation, 20 ng of the backbone was linearized with Len_lib_linF and Rorr_R2_LinR (Supplementary Table 8); for the reverse orientation insert, previously gel extracted minimal promoter and GFP from pLS-mP was amplified using minPGFP_Revorr_Len_HA_F and Len_lib_linR (Supplementary Table 8). Both backbones were treated with Antarctic Phosphatase, following the manufacturer's protocol. All backbones and inserts were gel extracted, with the exception of the reverse orientation insert, which we purified in a 1.8× AMPure reaction. Plasmid libraries were cloned and extracted as previously described.

Transfections (four independent transfections), DNA/RNA extractions, RT of mRNA and qPCRs to amplify barcodes for sequencing were all performed as previously described for the enhancer-length experiments. The final PCRs for the DNA samples were purified in a 1.5× AMPure reaction, using 50 µl of PCR reaction and eluting in 15 µl of Qiagen EB; cDNA PCRs were gel purified. Libraries were separately denatured and pooled, pooling twice as much of the RNA samples as the DNA samples. Samples were loaded at a final concentration of 1.8 pM on a 75-Cycle NextSeq v.2 High-Output kit.

MPRA to evaluate the impact of including additional sequence context at tested elements. *Design of enhancer-length libraries for array synthesis.* We chose to synthesize the same 2,236 genomic sequences tested above[18]. We also included the top 50 and bottom 50 haplotypes, averaging 409 bp, from a screen conducted in the STARR-seq vector[12], and designed libraries of 192-bp and 354-bp sequences, centered at the position of the previously tested design. We also designed a library of 678-bp sequences for the 2,236 genomic sequences above. We extracted genomic sequence using bedtools getfasta[40]. To the 192-bp library, we added the HSSF-ATGC sequence to the 5′ end and the HSS-R-clon sequence to the 3′ end (Supplementary Table 8).

For the 354-bp library, we split each sequence into two overlapping fragments, A and B. Fragment A included positions 1–190 and fragment B included positions 161–354. To fragment A, we appended the HSSF-ATGC adaptor to the 5′ end and the DO_15R adapter to the 3′ end. To fragment B, we appended the DO_5F adapter to the 5′ end and the HSS-R-clon adaptor to the 3′ end (Supplementary Table 8).

For the 678-bp library, we only designed the 2,236 sequences from Inoue et al.[18]. We split the sequences into 13 different sets of 172 sequences each. We then split each sequence into four fragments. Fragment A included positions 1–190, fragment B included positions 161–352, fragment C included positions 323–514 and fragment D included positions 485–678. Adaptors and primers used for the 13 sets of HMPA are included in Supplementary Table 10.

Amplification of the 192-bp library. All 192-bp enhancers were amplified from the array using HSSF-ATGC and HSS-R-clon (Supplementary Table 8) with KAPA HiFi HotStart Uracil+ ReadyMix PCR kit (Kapa Biosystems) with Sybr Green (Thermo Fisher Scientific) on a MiniOpticon Real-Time PCR System (Bio-Rad) and stopped before plateauing.

Multiplex pairwise assembly for 354-bp library. All 5′ fragments were amplified off the array using HSSF-ATGC and DO_15R_PU (Supplementary Table 8) with KAPA HiFi HotStart Uracil+ ReadyMix PCR kit (Kapa Biosystems) and stopped before plateauing. All 3′ fragments were amplified off the array using DO_5F_PU and HSS-95R (Supplementary Table 8). Both were purified using a 1.8× AMPure cleanup and eluted in 20 µl Qiagen EB. Then, 2 µl of USER enzyme (NEB) was added directly to each purified PCR product and incubated for 15 min at 37 °C followed by 15 min at room temperature. Reactions were then treated with the NEBNext End Repair Module (NEB) following the manufacturer's protocol and purified using the DNA Clean and Concentrator 5 (Zymo Research) and eluted in 12 µl EB, following the manufacturer's protocol. We then quantified DNA concentrations for both treated samples using a Qubit and diluted samples to 0.75 ng µl^{-1}. We then assembled the 5′ and 3′ fragments as described previously[25]. Briefly, fragments were allowed to anneal and extend for five cycles with KAPA HiFi 2× HotStart Readymix (Kapa Biosystems) before primers HSSF-ATGC and DO_95R were added for amplification (Supplementary Table 8).

Hierarchical multiplex pairwise assembly for 678-bp library. All libraries were amplified off the array using the primers indicated in Supplementary Table 10 with KAPA HiFi HotStart Uracil+ ReadyMix PCR Kit (Kapa Biosystems) as described above. During the first round of assembly, fragments A and B were assembled with HSSF-ATGC and DO_31R_PU and fragments C and D were assembled with DO_8F_PU and HSS_R (Supplementary Table 10). Assembled libraries were then purified with a 0.65× Ampure cleanup following the manufacturer's protocol and eluted in 20 µl. Then, 2 µl of USER enzyme (NEB) was added to the purified assembly reactions and incubated at 37 °C for 15 min followed by 15 min at room temperature and then repaired using the NEBNext End Repair Module (NEB), following the manufacturer's protocol and purified using the DNA Clean and Concentrator 5 (Zymo Research) and eluted in 10 µl EB. All libraries were then quantified using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific) and eluted to 0.75 ng µl^{-1}. Assemblies AB and CD were then assembled together following the multiplex pairwise assembly protocol[25]. After the second assembly, libraries were purified using a 0.6× AMPure cleanup and eluted in 30 µl EB. We then amplified 1 µl of each assembly with HSSF-ATGC-pu1F and HSS-R-clon-pu1R to add flow cell adaptors and indexes (Supplementary Table 8). We performed the assembly for each set of 172 sequences separately, as well as for different combinations of sets, up to all 2,236 sequences at once[41].

Sequence validation of assembled libraries. Before cloning, we verified assembly and uniformity of our libraries. The multiplex pairwise assembly library (2,336 354-mers) was sequenced on a Miseq v.3 600 cycle kit with paired-end 305 bp reads. Reads were merged with PEAR v.0.9.5 (ref. [42]) and aligned to a reference fasta file with BWA mem v0.7.10-r789 (ref. [43]). Each of the 13 hierarchical pairwise assembly sub-libraries (172 678-mers) as well as different complexities (344, 688, 1,032, 1,376, 1,720, 2,064 and 2,236) were sequenced on a Miseq v.3 600 cycle kit with paired-end 300 bp reads. Paired-end reads were aligned to a reference fasta file with BWA mem v0.7.10-r789 (ref. [43]). As our HMPA library was longer than the maximum Illumina sequencing length (600 bp), we prepared our HMPA sub-library 3 (172 678-mers) for sequencing on the PacBioSequel System using V2.1 chemistry (Pacific Biosciences). The library was amplified with pu1L and pu1R and sent to the University of Washington PacBio Sequencing Services for library preparation and sequencing. We obtained 312,277 productive zero-mode waveguides with an average Pol Read length of 30,806 bp. After generating circular consensus sequences, we obtained 218,240 circular consensus sequence reads with a mean read length of 882 bp.

Barcoding and cloning of length libraries into pGL4.23c. We performed a two-step PCR to add barcodes and cloning adaptors for pGL4.23c onto our three different libraries. For the 192-mer and 354-mer library, we amplified 20 ng of the library with HSS-pGL4_F and HSS-pGL4_R1 (Supplementary Table 8) using NEBNext

High-Fidelity 2× PCR Master Mix (NEB) for 16 cycles. For the 678-mer libraries, we pooled all 13 sub-libraries at equal concentrations and then amplified 20 ng with the same primers and conditions above. All PCR products were purified with a 1.5× AMPure cleanup following the manufacturer's instructions and eluted in 50 µl. We then used 1 µl of each purified reaction for a second PCR to append the 15-bp degenerate barcodes and cloning adaptors. For the second reaction, we used HSS-pGL4_F and HSS_pGL4_R2 (Supplementary Table 8).

We linearized the pGL4.23c MPRA backbone, while removing the minimal promoter and reporter using *Hin*dIII and *Xba*I. We treated the linearized plasmid with Antarctic Phosphatase following the manufacturer's protocol and then gel extracted the plasmid on a 1% agarose gel. We then cloned all three libraries into the pGL4.23c plasmid using the NEBuilder HiFi DNA Assembly Cloning kit (NEB), following the manufacturer's protocol. The library was then transformed into 10-β electrocompetent cells (NEB C3020), grown in 100 ml of LB + Amp and extracted using the ZymoPURE II Plasmid Midiprep kit (Zymo Research). We then re-linearized each library with Len_lib_linF and Len_lib_linR and amplified the minimal promoter and GFP from 10 ng of the pLS-mP plasmid using minGFP_Len_HAF and minGFP_Len_HAR (Supplementary Table 8). We then gel extracted all linearized libraries and the minimal promoter + GFP insert on a 1% agarose gel. We inserted the minimal promoter and GFP using the NEBuilder HiFi DNA Assembly Cloning kit (NEB) as described above.

MPRA of all enhancer-length libraries. The day before transfection, we seeded HepG2 cells in five 10-cm dishes. On the day of transfection, we combined the 192, 354 and 678 pGL4.23c libraries at a 1:1:1 molar ratio and transfected 21 µg of pooled libraries into each 10-cm dish using Lipofectamine 3000 (Thermo Fisher Scientific), following the manufacturer's protocol. At 48 h after transfection, we extracted DNA and RNA from each replicate using the AllPrep DNA/RNA Mini kit (Qiagen), following the manufacturer's instructions.

We added UMIs to a total of 4 µg of DNA from each replicate split across eight reactions with KAPA2G Robust HotStart ReadyMix (Kapa Biosystems) for three cycles with a 65 °C annealing and 40 s extension, using P5-pLS-mP-5bc-idx and P7-pGL4.23c-UMI (Supplementary Table 8). After three cycles, reactions were pooled and purified with a 1.8× AMPure cleanup, following the manufacturer's instructions and eluted in a total of 87 µl of Qiagen EB. The entire purified product was then used for a second round of PCR, split into six 50-µl reactions each, with primers P5 and P7. The reaction was followed in real time with Sybr Green and stopped before plateauing. PCRs were then pooled and 100 µl of the pooled PCR products was purified with a 0.9× AMPure cleanup and eluted in 30 µl for sequencing.

RNA for each replicate was treated with Turbo DNase (Thermo Fisher Scientific) following the manufacturer's protocol and then quantified using the Qubit RNA Assay kit (Thermo Fisher Scientific). For all samples, we performed two 15-µl RT reactions, using a total of 15.75 µl RNA (1/2 total). RT was performed using Thermo Fisher SuperScript IV (Thermo Fisher Scientific) and a gene-specific primer (P7-pGL4.23c-UMI), which attached a UMI, following the manufacturer's instructions. The cDNA for each sample was split into four 50-µl PCRs using P5-pLS-mP-5bc-idx and P7 for three cycles. Reactions were then pooled together and purified with a 1.5× AMPure reaction and eluted in 64.5 µl of Qiagen EB. The purified PCR product was then split into three 50-µl PCRs with P5 and P7, followed in real time with Sybr Green (Thermo Fisher Scientific) and stopped before plateauing (11 cycles). PCR products were purified with a 1.5× AMPure reaction before sequencing on a 75-cycle NextSeq 550 v.2 High-Output kit.

For barcode associations, we amplified 5 ng of each library with P5_pGL4_Idx_assF and P7-pGL4-ass-R (Supplementary Table 8), following in real time with Sybr Green for 14–15 cycles. PCR products were purified with a 1× AMPure cleanup and eluted in 20 µl of Qiagen EB for sequencing. Libraries were separately denatured and pooled to account for part of the clustering bias on the NextSeq. We brought the 192 library to a final concentration of 1.65 pM, the 354 library to a final concentration of 2.15 pM and the 678 library to a final concentration of 2.9 pM. We then pooled an equal volume of each library and loaded on a 300 cycle NextSeq 550 v.2 Mid-Output kit with an 80 bp read 1 and 213 bp read 2 (to sequence part of contributing oligonucleotides A, C and D).

MPRA processing pipeline. *Reproducible MPRA analysis pipeline implementation.* We developed and utilized a fully reproducible processing pipeline to process the raw MPRA data. The sections below document the various components of the pipeline, which borrow heavily from our earlier work[18] and were implemented into a reproducible Nextflow-based codebase named MPRAflow[44].

Associating barcodes to designed elements. For each of the barcode association libraries, we generated Fastq files with bcl2fastq v.2.18 (Illumina), splitting the sequencing data into an index file delineating the barcode and two paired-end read files delineating the corresponding element linked to the barcode. If the paired-end reads overlapped in sequence, they were merged into one and aligned using BWA mem v0.7.10-r789 (ref. [43]) to a reference fasta file consisted of the designed elements (Supplementary Table 2). We carried forward the subset of merged reads whose mapped length corresponded to the expected length of the designed element ± 5 bp (171 ± 5, 192 ± 5, 354 ± 5 and 678 ± 5, depending on the element size), allowing indels or mismatches. To minimize the impact of sequencing errors, we associated a barcode to an element if: (1) the barcode–element pair was sequenced at least three independent times and (2) ≥90% of the barcode mapped to a single element. These barcode associations were then used as a dictionary to match barcodes detected in the RNA and DNA sequencing libraries in different MPRA designs.

Replicates, normalization and RNA/DNA activity scores. Barcodes were counted for RNA and DNA samples for each MPRA experiment, using UMIs to collapse barcodes derived from the same molecule during PCR and mapped to the element they were linked to, as identified by the dictionary of barcode–element associations. To normalize RNA and DNA for different sequencing depths in each sample, we followed a nearly identical scheme as one we had previously devised[18]. Briefly, for each replicate of each MPRA design, we first considered the subset of barcodes that were observed for both the RNA and DNA samples of the replicate. We then summed up the counts of all barcodes contributing to each element and computed the normalized counts as the counts per million (cpm) sequenced reads of that library. Finally, we computed enhancer activity scores as $\log_2$(RNA cpm/DNA cpm). To account for the differential scale among replicates of each experiment, we divided the RNA/DNA ratios by the median across the replicate value before averaging them. Due to low counts in the initial round of sequencing and poor sample quality, the three replicates from the 5′/3′ MT and 3′/3′ MT were re-sequenced and the data from each pair of technical replicates were pooled across the two independent sequencing runs. Even after pooling, the first replicates of these two assays exhibited poorer inter-replicate concordance than the other replicates (Fig. 2a and Supplementary Fig. 3) and thus were excluded during replicate averaging (Supplementary Table 2). In practice, this decision very modestly altered the numerical results and did not change the study's conclusions.

Modeling and analyses. *Features considered.* For each candidate enhancer, we computed a total of 915 features derived from either: (1) the sequence itself or (2) experimentally measured information, computed as a mean signal extracted from the corresponding region of the human genome (Supplementary Tables 3 and 4). The sequence-based features represent the conservation of the sequence, general G/C content, predicted chromatin state and likelihood of binding to an assortment of transcription factors and RNA-binding proteins. In contrast, the experimentally derived features represent empirical measurements of chromatin/epigenetic state, binding to transcription factors or transcriptional activity. The features were derived from custom Perl scripts, the UCSC genome browser[45], DeepSEA v.0.94 (ref. [46]), DeepBind v.0.11 (ref. [47]), with epigenomic data derived from the Epigenomics Roadmap Consortium[48], CAGE data from the FANTOM Consortium[33] and ChIP-seq data from the ENCODE Consortium[7].

Feature pre-processing. Right-skewed data such as ChIP-seq and CAGE signal were log-transformed to approximate a normal distribution, and each feature was then z score normalized to scale the features similarly. This enabled a direct comparison of coefficients among features derived from the resulting linear models.

Model training. As described before[18], we trained a lasso regression model on each of ten folds of the data, selecting enhancers that were measured with at least ten independent barcodes to reduce the impact of measurement noise in the assessment of model quality. A lasso regression model was chosen specifically because it employs an L1 regularization penalty, which leads to the selection of the fewest features that maximally explain the data. The strength of the regularization was controlled by a single λ parameter, which was optimized using tenfold CV on the entire dataset. To evaluate the most relevant features selected, we trained a lasso regression model on the full dataset and visualized the top 10–30 coefficients with the greatest magnitude. A full table of the selected features and their coefficients are provided (Supplementary Table 5). To compare differential enhancer activity between a pair of assays, we fit a loess ('locally estimated scatterplot smoothing') regression between one assay relative to the other and computed residuals from this fit, using the 'loess' function in R. We then fitted lasso regression models to explain these residuals, based upon the aforementioned procedure.

Luciferase assays. The 'medium', 'long' and 'deleted' versions of ten enhancers (total 30 sequences), APOE enhancer (positive control) and neg2 sequence (negative control) were synthesized along with minimal promoter and adaptor sequences (Supplementary Table 8) and cloned into the *Bgl*II and *Nco*I site of the pGL4.23c vector by Twist Bioscience. These were selected on the basis of highest differential activities, reproducibility and base balance (for synthesis). As two of them (chr2:106744003-106744357_medium and chr10:114391246-114391924_del) failed the cloning, these sequences were synthesized by Twist Bioscience and manually cloned into the *Bgl*II and *Nco*I site of the pGL4.23c vector using NEBuilder HiFi DNA Assembly Cloning kit (NEB). The plasmid sequences were confirmed by Sanger sequencing. All 32 plasmids and empty pGL4.23c were individually transfected along with pGL4.74 (Promega) into 1×10^4 HepG2 cells, as previously described[18]. Four independent replicate cultures were transfected. Firefly and *Renilla* luciferase activities were measured on a Glomax microplate reader (Promega) using the Dual-Luciferase Reporter Assay System (Promega).

Enhancer activity was measured as the fold change of each plasmid's firefly luciferase activity normalized to *Renilla* luciferase activity.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We developed a fully reproducible MPRA processing pipeline available to process the data into final enhancer activity scores. Raw and processed data have been deposited in the Gene Expression Omnibus at accession number GSE142696.

Code availability

A reproducible processing pipeline for MPRA data is available as a Nextflow-based MPRA processing pipeline named MPRAflow (https://github.com/shendurelab/MPRAflow)[44].

References

40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
41. Klein J. C. et al. A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. *Protoc. Exch.* https://doi.org/10.21203/rs.3.pex-1065/v1 (2020).
42. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
43. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* https://arxiv.org/abs/1303.3997 (2013).
44. Gordon, M. G. et al. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).
45. Karolchik, D. et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
46. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
47. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
48. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

Acknowledgements

We thank S. Kim and other members of the Shendure and Ahituv laboratories for general advice and critical feedback on the manuscript. This work was supported by the National Human Genome Research Institute grants 1UM1HG009408 (N.A. and J.S.), 5R01HG009136 (J.S.), 1R21HG010065 (N.A.), 1R21HG010683 (N.A.) and 5F30HG009479 (J.K.); National Institute of Mental Health grants 1R01MH109907 (N.A.) and 1U01MH116438 (N.A.); NRSA National Institutes of Health fellowship 5T32HL007093 (V.A.); and the Uehara Memorial Foundation (F.I.). J.S. is an investigator of the Howard Hughes Medical Institute.

Author contributions

J.K. and A.K. performed all cloning and sequencing for the nine assays and all experimental work for orientation and length sections. J.K. and J.S. conceived the HMPA protocol, and J.K. and A.K. developed and optimized it. A.K. produced schematic figures. M.K. developed the initial MPRA analysis pipeline. V.A. performed the computational analyses and generated all remaining figures and tables. F.I. performed the transfections and lentiviral transductions for the nine assays, carried out luciferase reporter experiments and wrote the associated methods sections. B.M. designed cloning steps and guided the development and testing of the MPRA assays. J.K., V.A., N.A. and J.S. wrote the remainder of the paper. N.A. and J.S. supervised the project.

Competing interests

V.A. is an employee of Calico Life Sciences LLC.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0965-y.

Correspondence and requests for materials should be addressed to N.A. or J.S.

Peer review information Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Nadav Ahituv, Jay Shendure

Last updated by author(s): Aug 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All software used for data processing are noted in the Methods section and appropriately cited. They are: PEAR v0.9.5, BWA mem v0.7.10-r789, bcl2fastq v2.18, DeepSEA v0.94, DeepBind v0.11, MPRAflow
Data analysis	The full MPRA analysis pipeline in this study can be found at http://github.com/shendurelab/MPRAflow

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are available for free access at the Gene Expression Omnibus (GEO) at the ID: GSE142696. The databases used in the study are cited in the text , and are included here: UCSC genome browser45, Epigenomics Roadmap Consortium48, CAGE data from the FANTOM Consortium33, and ChIP-seq data from the ENCODE Consortium7

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes are noted in each of the figures and resulted from the number of elements designed that were successfully measured after using our processing pipeline
Data exclusions	Two technical replicates (replicate 1 of 5/3MT and 3/3MT) were excluded during averaging due to poor sample quality, and these were directly noted in our Methods sections and text.
Replication	Each of the assays were performed with either three or four replicates to ensure faithful reproduction of the assay. All attempts at replication were successful.
Randomization	Not relevant, as we needed to know the identity of each sample prior to their analyses as each design backbone and therefore initial steps in processing are unique.
Blinding	Not relevant, as we needed to know the identity of each sample prior to their analyses as each design backbone and therefore initial steps in processing are unique.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HepG2 cells (ATCC HB-8065), HEK203T (ATCC-CRL-3216)
Authentication	HepG2 cells and Hek293T cells were not authenticated
Mycoplasma contamination	HepG2 cells and Hek293T cells were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No coomonly misidentified cell lines were used in this study