Supplemental Note 1

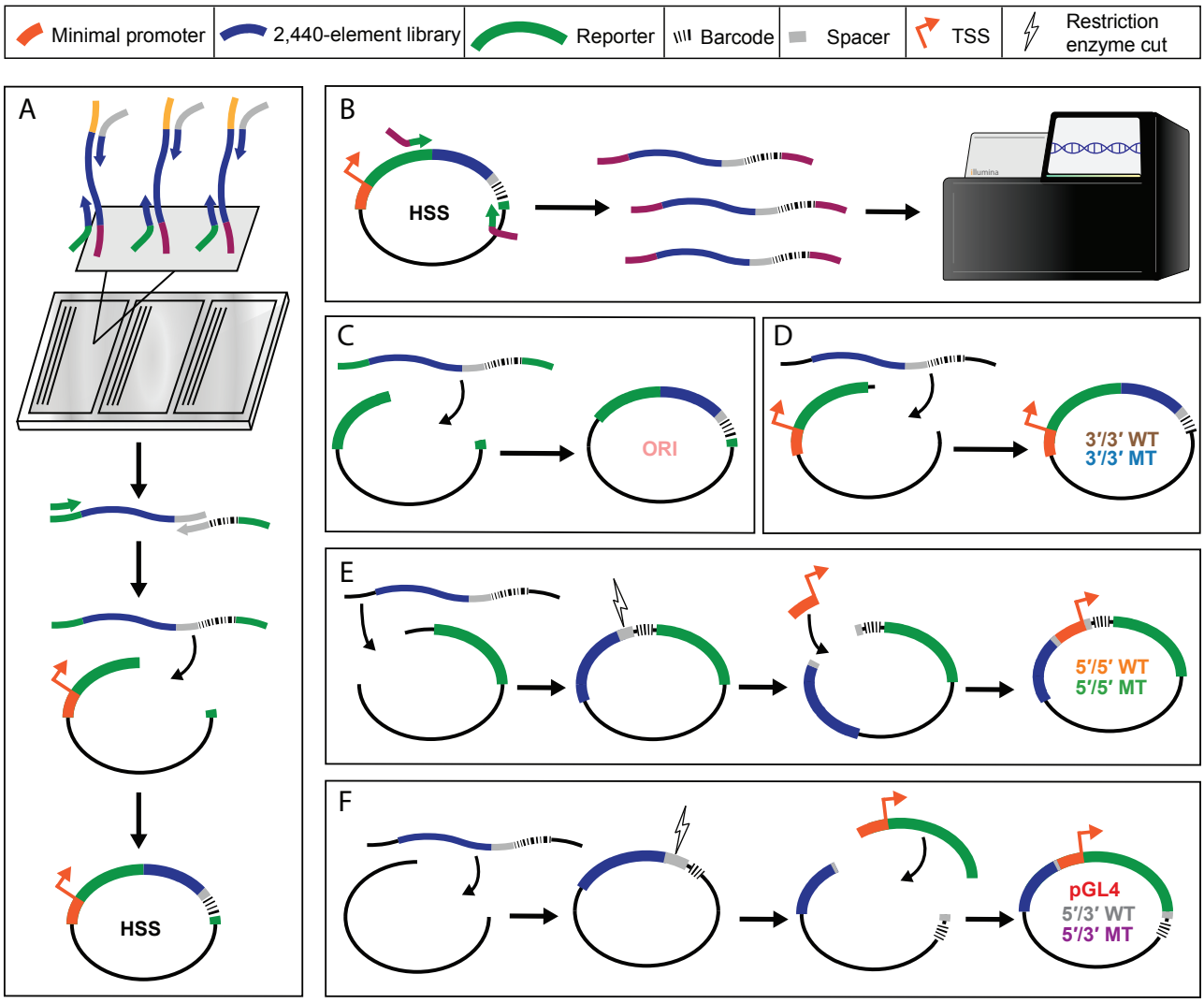
First, although assays like STARR-seq wherein the enhancer serves as the barcode are more straightforward to implement, our understanding of how position (3′ of the promoter, rather than 5′ as in a more conventional reporter assay vector) or the fact that the sequence is serving as the barcode, influences results, remains incomplete. Arnold *et al*. notably benchmarked STARR-seq against 142 conventional luciferase assays (r = 0.83), but STARR-seq has yet to be systematically compared to other MPRAs[16]. Second, although we previously showed differences between episomal vs. integrated MPRAs[17], it is not clear how these differences rank relative to those resulting from other design choices. Third, although the orientation-independence of enhancers has been evaluated in *Drosophila*[16,25,26], to our knowledge the robustness of this assumption has not previously been systematically tested in a mammalian system. Finally, the typical choice to test <200 bp fragments, each corresponding to a putative enhancer, is entirely based on technical limitations of massively parallel DNA synthesis, rather than on any principled understanding of the actual size of enhancers. The consequences of this choice for the results obtained remain largely unquantified.

Supplemental Note 2

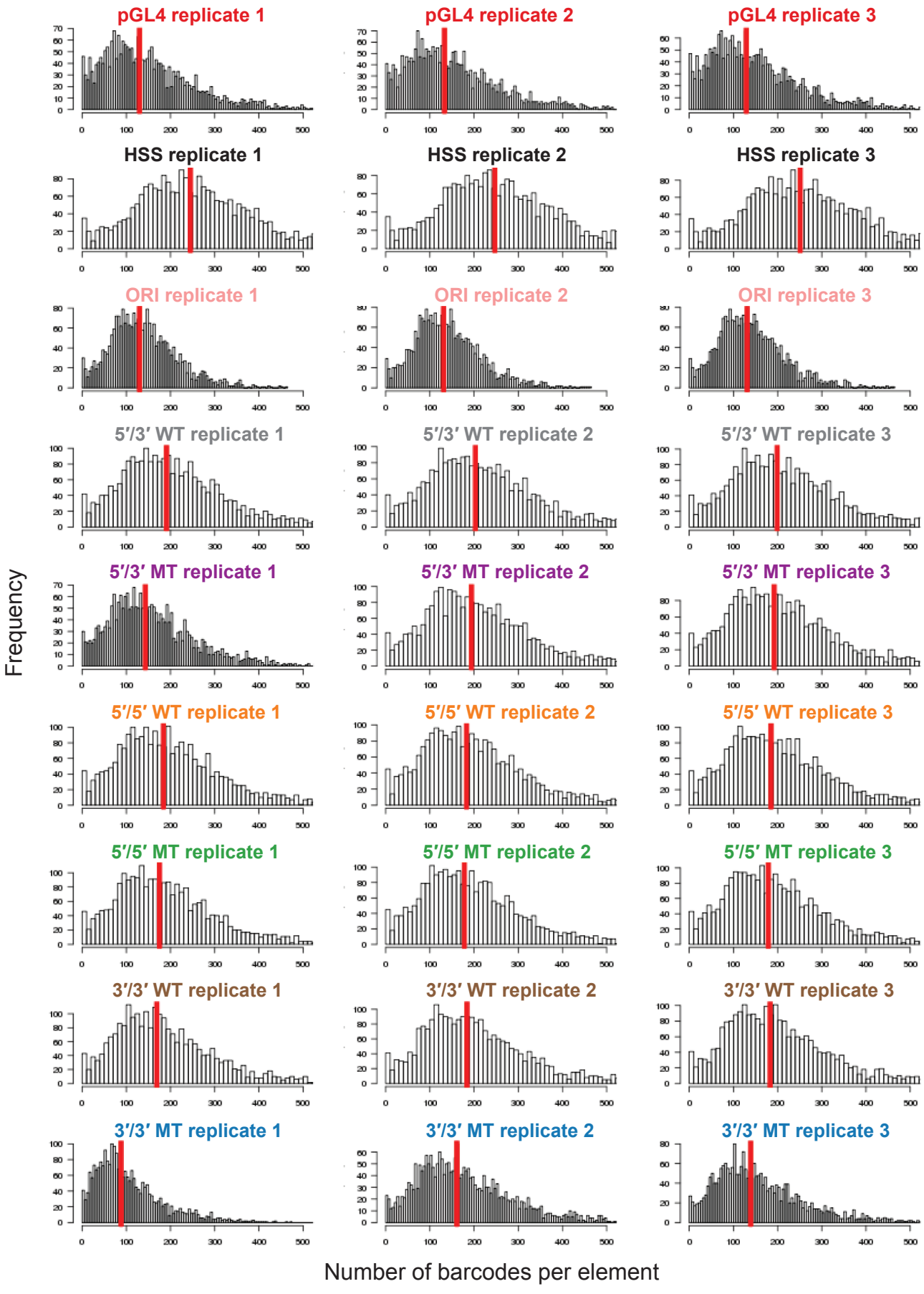
Potential reasons for inter-feature correlations include: i) several paralogous factors may bind to nearly identical motifs, leading to overlapping sequence-based binding predictions, ii) several factors may bind in a complex, leading to a redundant ChIP-seq signal, and iii) an upstream factor might lead to the deposition or removal of multiple epigenetic marks simultaneously, leading to a correlated or anti-correlated signal. Of note, another potential confounder is that the criteria used to select elements relied on features including FOXA1/2, HNF4A, EP300, H3K27ac, CHD2, RAD21, and SMC3 ChIP peaks[17], which could exaggerate the contributions of these factors, or bias the contributions of other features, in models attempting to predict general enhancer activity in HepG2 cells.

Supplemental Note 3

To test the possibility that barcode counts include short transcripts initiating within the candidate enhancer itself, we repeated the HSS and ORI experiments using two different 5′ primers -- our original primer, which hybridizes immediately upstream of the barcode, or a new primer, which hybridizes upstream of the candidate enhancer. The median number of barcode counts for all replicates was greater than 100 (**Supplemental Figure 17A**). Estimated activities in HSS and ORI experiments with these distinct primer sets were correlated, which is modestly reassuring but does not completely rule out this potential confounder of this assay design, particularly as these are preselected enhancer candidates that might initiate transcription poorly (**Supplemental Figure 17A-B**; HSS: Pearson r = 0.89; Spearman rho = 0.90; ORI: Pearson r = 0.84; Spearman rho = 0.83; for comparison, Pearson r for biological replicates ranged from 0.92-1.0). Of note, there was a significant batch effect between these vs. our original experiments, which were performed months earlier with different reagent kits and at a later cell passage, highlighting the importance of controlling and documenting technical variables whenever possible (**Supplemental Figure 17C-D**).



Supplemental Figure 1. Cloning of MPRA vector libraries.
A) We amplified an existing array library from Inoue *et al.*[18], adding homology to the STARR-seq vector on the 5′ end and a spacer sequence for cloning on the 3′ end. The library was put into a second PCR reaction adding barcodes and homology to STARR-seq to the 3′ end, and cloned sequences into the STARR-seq vector.
B) We amplified enhancers from the STARR-seq vector library, adding flow cell adapters to the sequences, and sequenced the library on a NextSeq Mid 300 cycle kit to associate enhancer sequences to their corresponding barcodes.
C) The STARR-seq ORI assay was cloned by amplifying enhancers from the STARR-seq library while adding homology to the STARR-seq ORI vector, and cloned into the linearized ORI backbone.
D) We cloned the 3′/3′ lenti- and mutant lentiMPRA libraries by first linearizing the pLSmP backbone without removing the minimal promoter and GFP. The enhancer library and barcodes were then amplified from the STARR-seq vector adding homology to the linearized backbone, and cloned into the vector.
E) To clone the 5′/5′ lenti- and mutant lentiMPRA libraries, we linearized the pLSmP vector such that the minimal promoter was removed.. We then linearized plasmids at the spacer sequence and cloned in the minimal promoter such that it was between the enhancer and barcode sequences.
F) The pGL4.23c episomal assay and the 5′/3′ lenti- and mutant lentiMPRA libraries were cloned in the same manner as the 5′/5′ vector library, except the pGL4.23c and pLSmP backbones were initially linearized to remove both the minimal promoter and reporter gene; after cloning in our library, the minimal promoter and reporter were inserted into the spacer sequence such that enhancers were upstream and barcodes downstream of the reporter.


pGL4 replicate 1
pGL4 replicate 2
pGL4 replicate 3
HSS replicate 1
HSS replicate 2
HSS replicate 3
ORI replicate 1
ORI replicate 2
ORI replicate 3
5′/3′ WT replicate 1
5′/3′ WT replicate 2
5′/3′ WT replicate 3
5′/3′ MT replicate 1
5′/3′ MT replicate 2
5′/3′ MT replicate 3
5′/5′ WT replicate 1
5′/5′ WT replicate 2
5′/5′ WT replicate 3
5′/5′ MT replicate 1
5′/5′ MT replicate 2
5′/5′ MT replicate 3
3′/3′ WT replicate 1
3′/3′ WT replicate 2
3′/3′ WT replicate 3
3′/3′ MT replicate 1
3′/3′ MT replicate 2
3′/3′ MT replicate 3
Frequency
Number of barcodes per element

Supplemental Figure 2. Histograms of barcodes per element for each replicate of the 9 MPRA methods. Shown are histograms indicating the number of observed barcodes per element, for each of the 2,440 elements tested in each of the 3 replicates for each of the 9 MPRA methods. Shown with a vertical red line is the median number of barcodes per element.

Figure S3

A

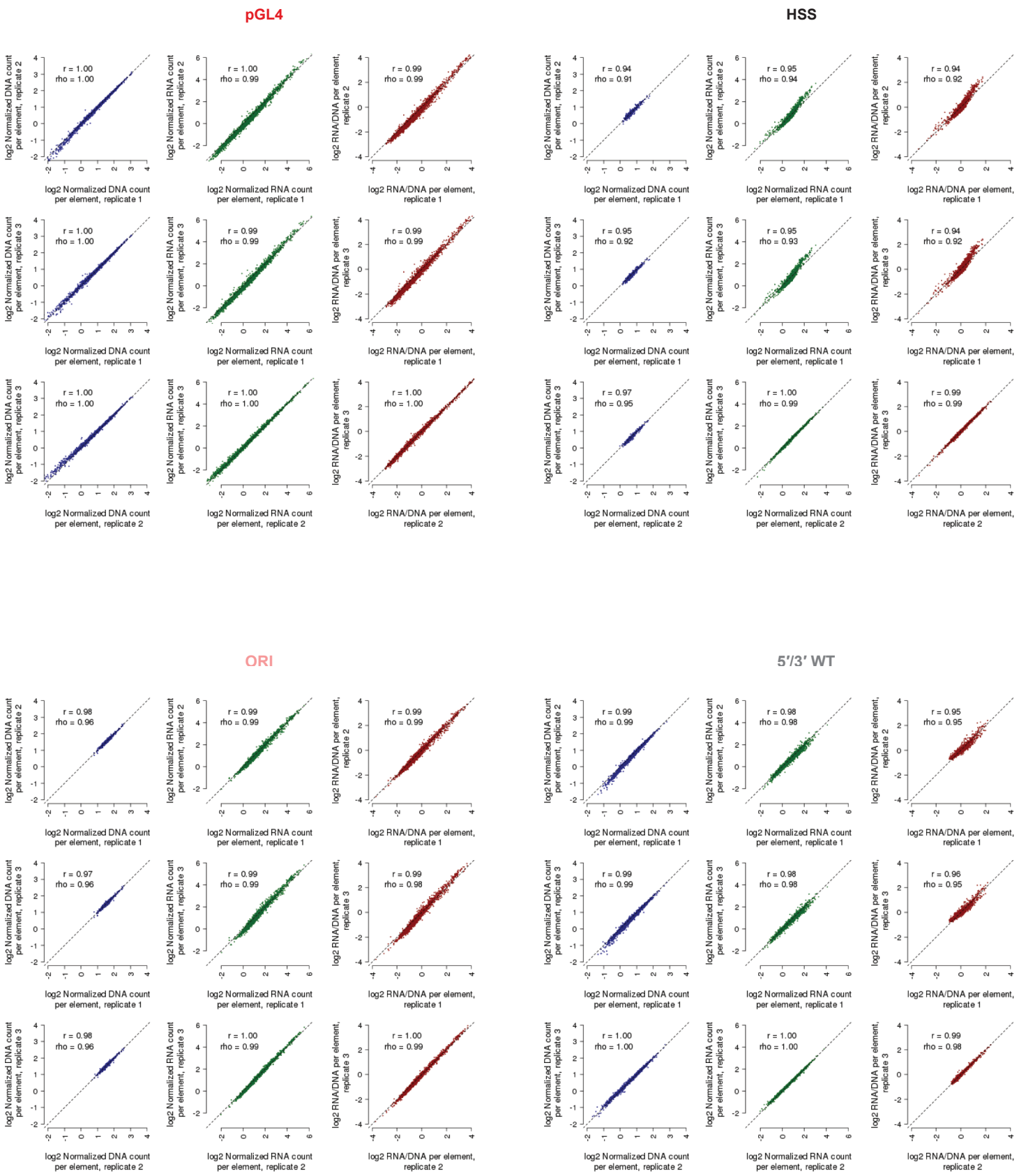
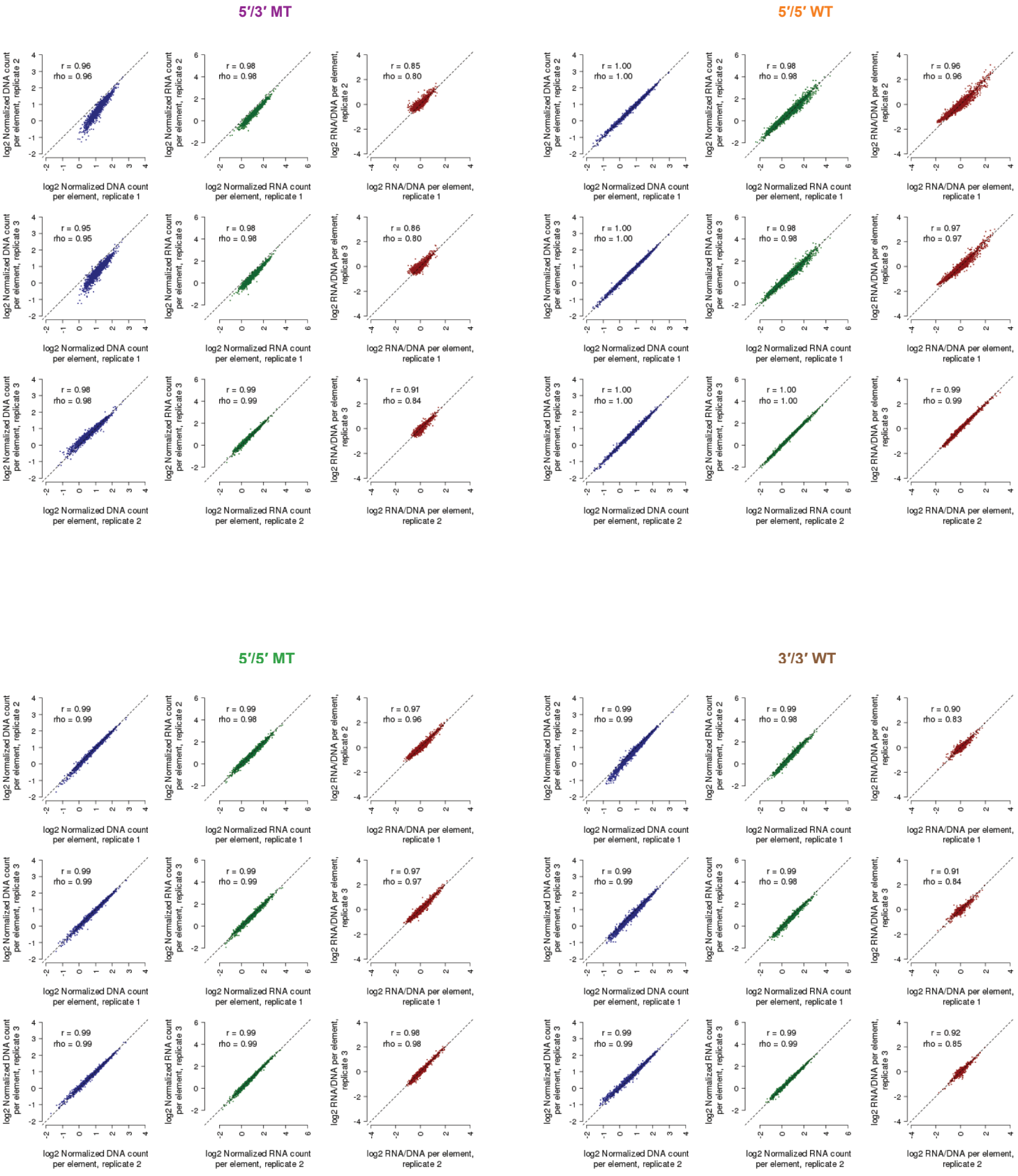
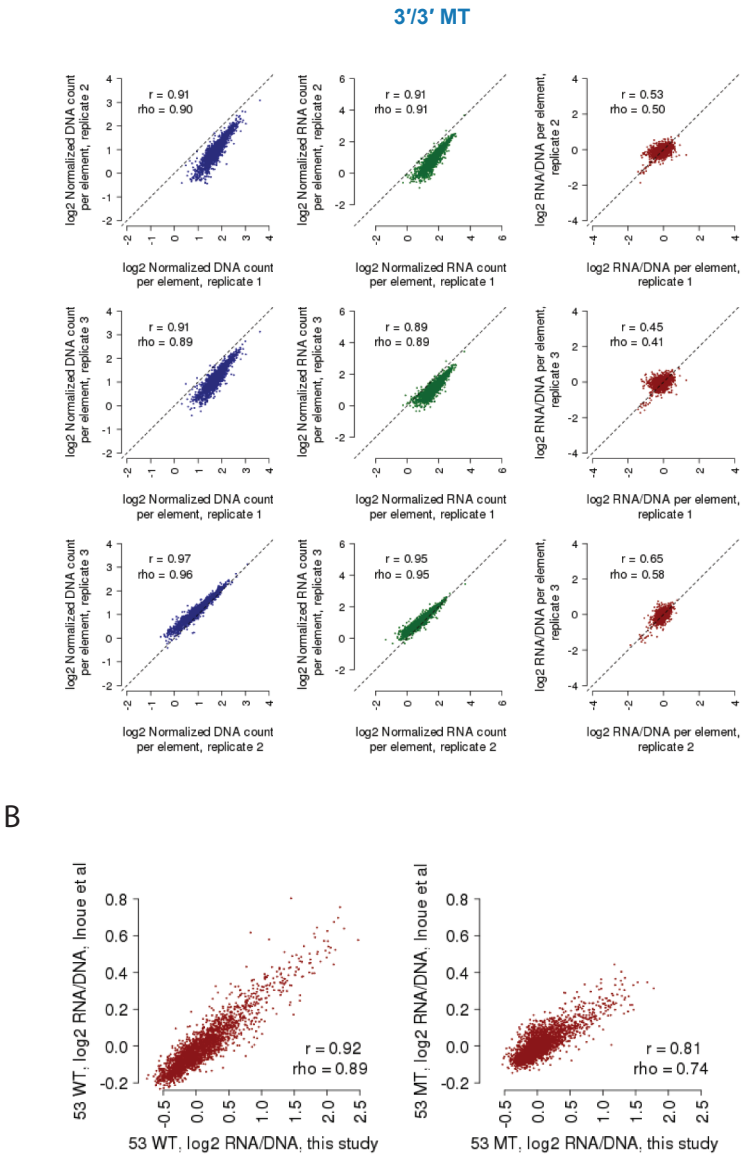
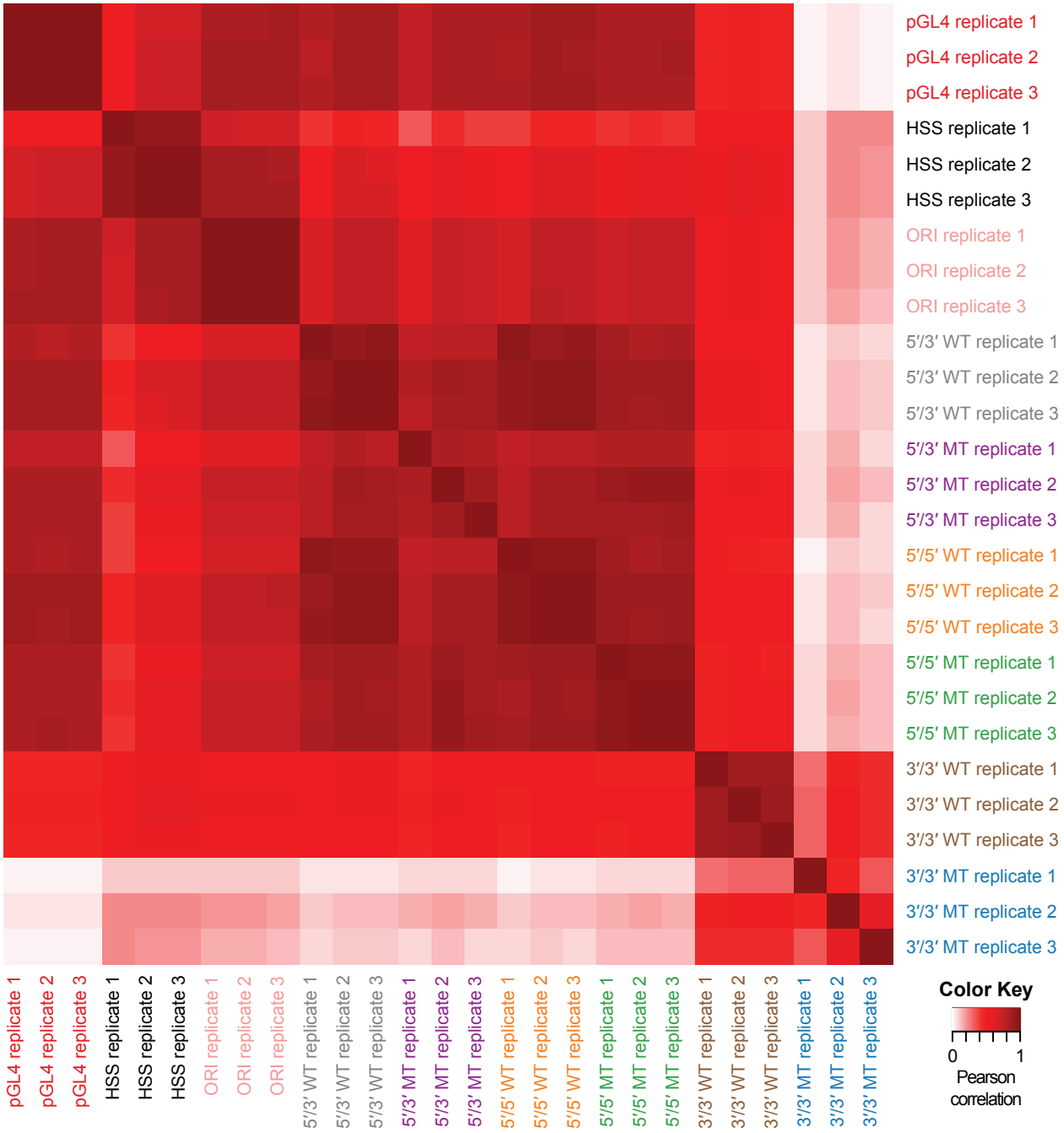

pGL4
HSS
ORI
5′/3′ WT
log2 Normalized DNA count per element, replicate 1
log2 Normalized RNA count per element, replicate 1
log2 RNA/DNA per element, replicate 1
log2 Normalized DNA count per element, replicate 2
log2 Normalized RNA count per element, replicate 2
log2 RNA/DNA per element, replicate 2
log2 Normalized DNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 3
log2 RNA/DNA per element, replicate 3
r = 1.00 rho = 1.00
r = 1.00 rho = 0.99
r = 0.99 rho = 0.99
r = 1.00 rho = 1.00
r = 0.99 rho = 0.99
r = 0.99 rho = 0.99
r = 1.00 rho = 1.00
r = 1.00 rho = 1.00
r = 1.00 rho = 1.00
r = 0.94 rho = 0.91
r = 0.95 rho = 0.94
r = 0.94 rho = 0.92
r = 0.95 rho = 0.92
r = 0.95 rho = 0.93
r = 0.94 rho = 0.92
r = 0.97 rho = 0.95
r = 1.00 rho = 0.99
r = 0.99 rho = 0.99
r = 0.98 rho = 0.96
r = 0.99 rho = 0.99
r = 0.99 rho = 0.99
r = 0.97 rho = 0.96
r = 0.99 rho = 0.99
r = 0.99 rho = 0.98
r = 0.98 rho = 0.96
r = 1.00 rho = 0.99
r = 1.00 rho = 0.99
r = 0.99 rho = 0.99
r = 0.98 rho = 0.98
r = 0.95 rho = 0.95
r = 0.99 rho = 0.99
r = 0.98 rho = 0.98
r = 0.96 rho = 0.95
r = 1.00 rho = 1.00
r = 1.00 rho = 1.00
r = 0.99 rho = 0.98

Figure S3

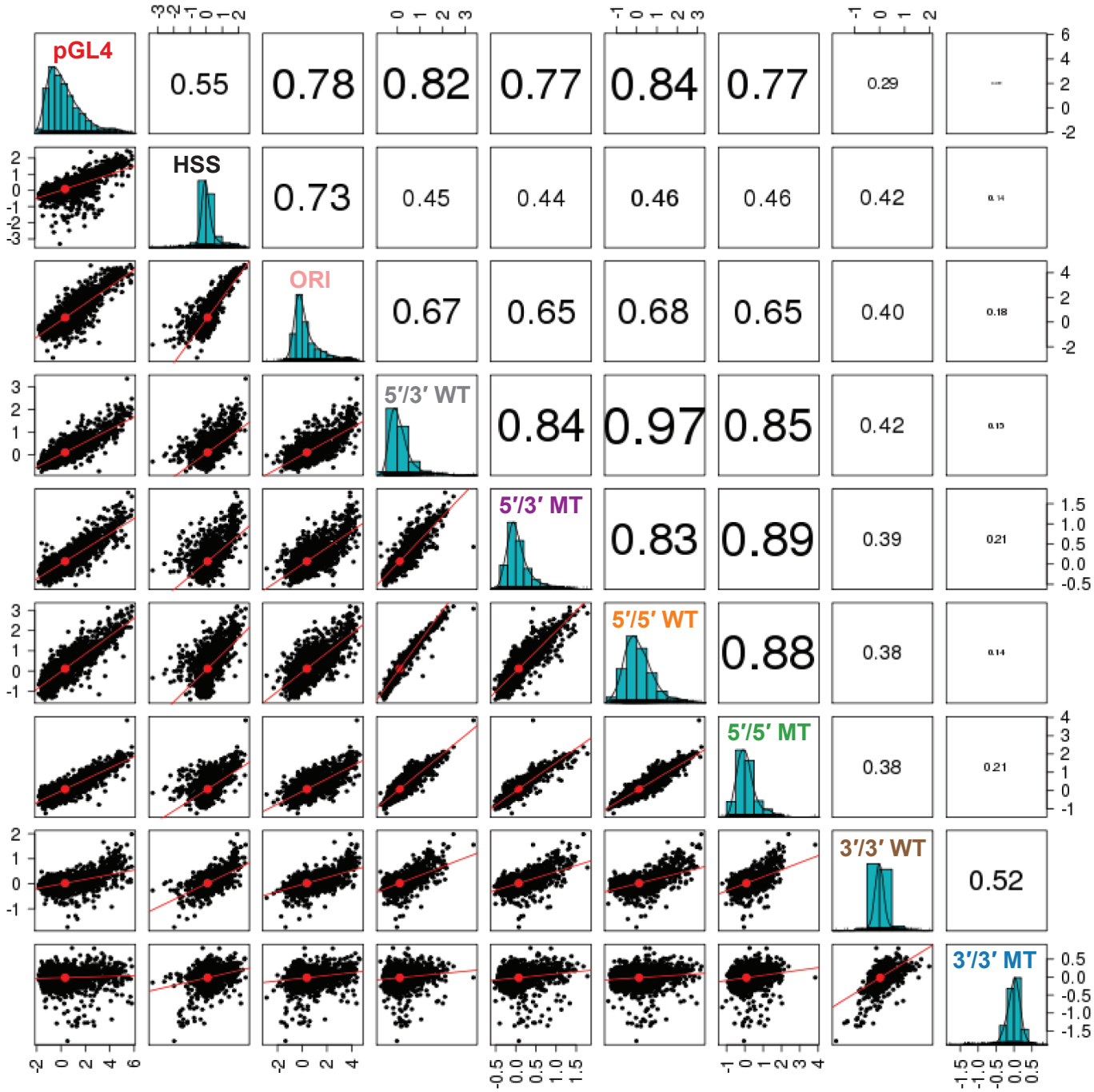

5′/3′ MT
5′/5′ WT
5′/5′ MT
3′/3′ WT
log2 Normalized DNA count per element, replicate 1
log2 Normalized DNA count per element, replicate 2
log2 Normalized RNA count per element, replicate 1
log2 Normalized RNA count per element, replicate 2
log2 RNA/DNA per element, replicate 1
log2 RNA/DNA per element, replicate 2
log2 Normalized DNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 3
log2 RNA/DNA per element, replicate 3
r = 0.96 rho = 0.96
r = 0.98 rho = 0.98
r = 0.85 rho = 0.80
r = 0.95 rho = 0.95
r = 0.98 rho = 0.98
r = 0.86 rho = 0.80
r = 0.98 rho = 0.98
r = 0.99 rho = 0.99
r = 0.91 rho = 0.84
r = 1.00 rho = 1.00
r = 0.98 rho = 0.98
r = 0.96 rho = 0.96
r = 1.00 rho = 1.00
r = 0.98 rho = 0.98
r = 0.97 rho = 0.97
r = 1.00 rho = 1.00
r = 1.00 rho = 1.00
r = 0.99 rho = 0.99
r = 0.99 rho = 0.99
r = 0.99 rho = 0.98
r = 0.97 rho = 0.96
r = 0.99 rho = 0.99
r = 0.99 rho = 0.98
r = 0.97 rho = 0.97
r = 0.99 rho = 0.99
r = 0.99 rho = 0.99
r = 0.98 rho = 0.98
r = 0.99 rho = 0.99
r = 0.99 rho = 0.98
r = 0.90 rho = 0.83
r = 0.99 rho = 0.99
r = 0.99 rho = 0.98
r = 0.91 rho = 0.84
r = 0.99 rho = 0.99
r = 0.99 rho = 0.99
r = 0.92 rho = 0.85



Supplemental Figure 3. Reproducibility among replicates for each of the 9 MPRA methods.
A) 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 each of the 9 MPRA methods tested. Also indicated is the Pearson (*r*) and Spearman (rho) correlation values. Candidate enhancers supported by fewer than 10 barcodes were filtered out prior to this analysis to reduce the impact of technical noise.
B) Scatter plots displaying the relationship between 5′/3′ WT and 5′/3′ MT ratios from this study and the indicated datasets from Inoue *et al.*[18]. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values.

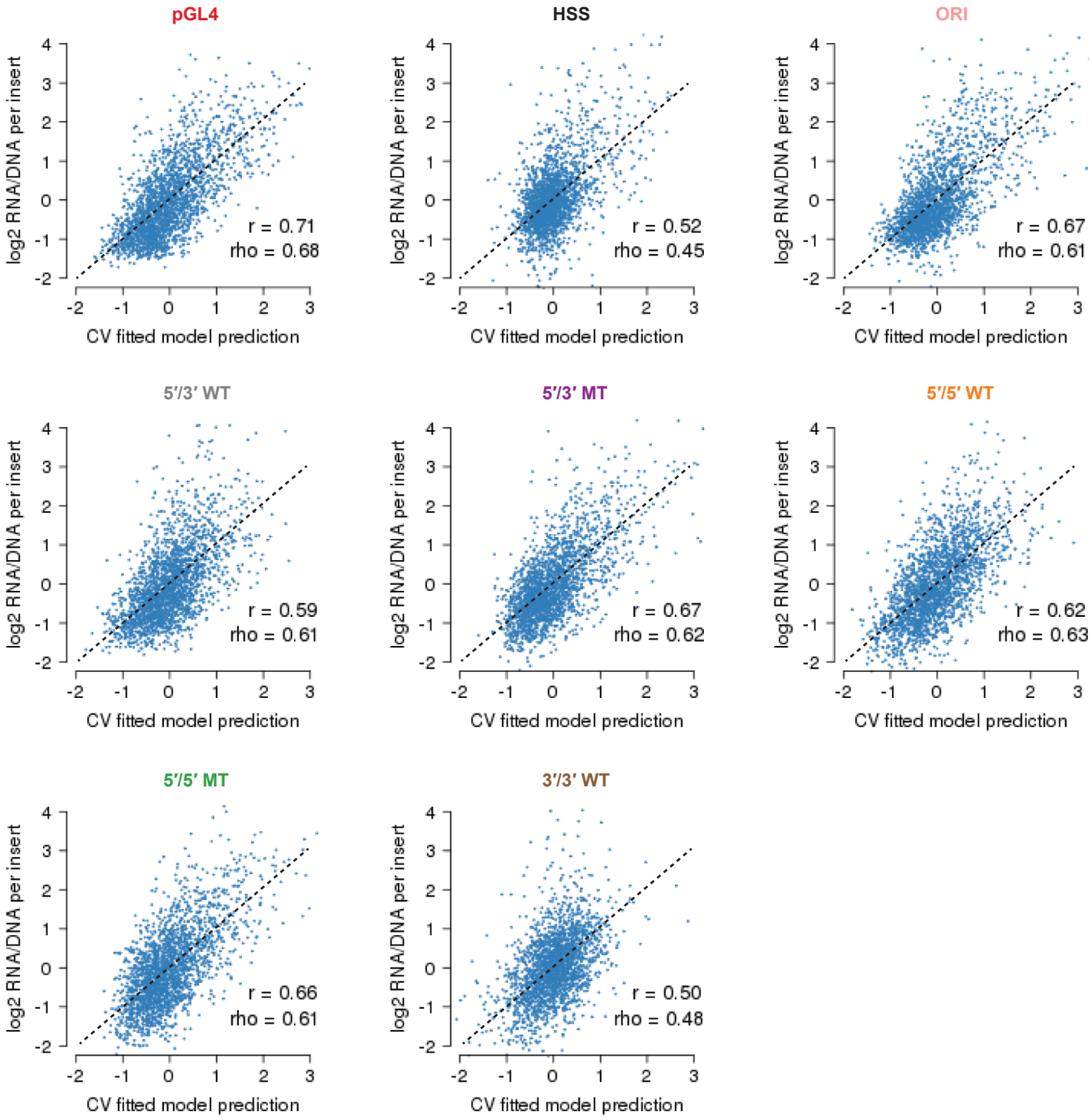


Supplemental Figure 4. Reproducibility among RNA/DNA ratios among individual replicates for each of the 9 MPRA methods.
Shown is a heatmap displaying the pairwise Pearson correlation among individual replicates, using RNA/DNA ratios from each of the 3 replicates for each of the 9 MPRA methods tested.

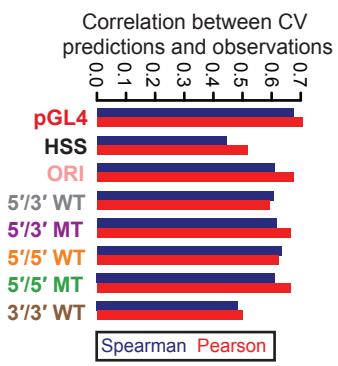


Supplemental Figure 5. Reproducibility of averaged RNA/DNA ratios.
This figure is equivalent to **Figure 2B**, except it displays the scatter matrix values using Spearman correlation values instead of Pearson correlation values.

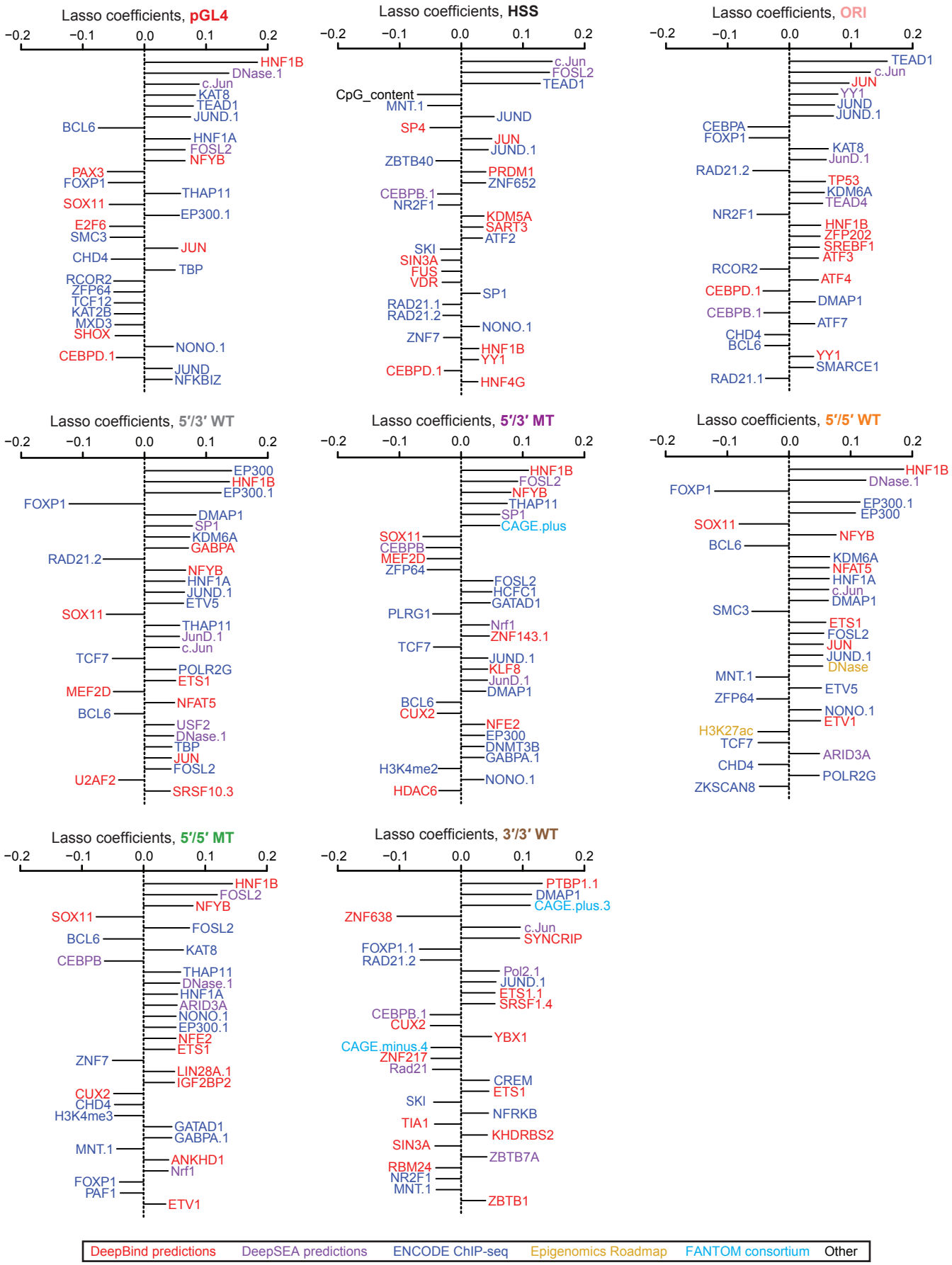
A

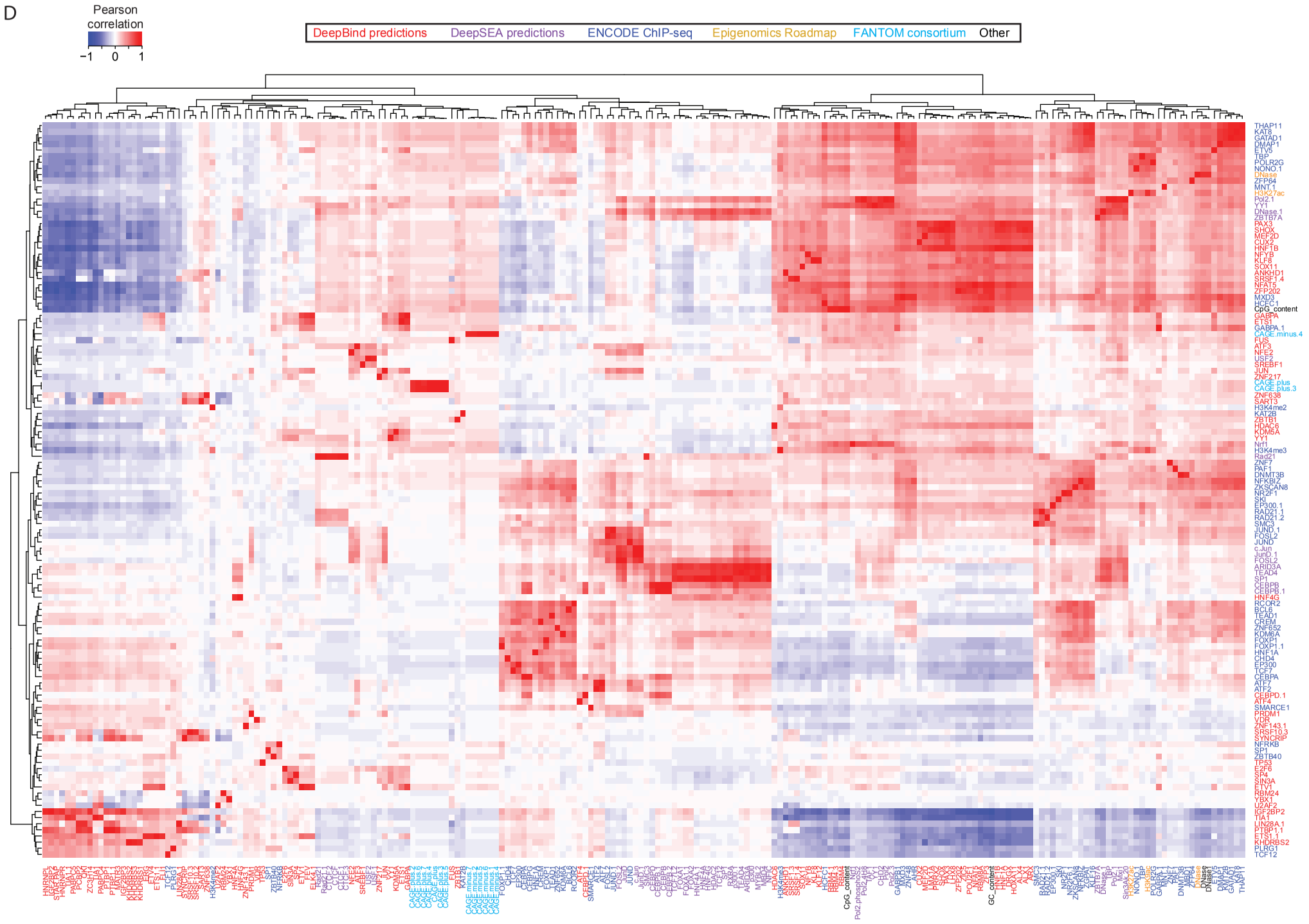

pGL4
r = 0.71
rho = 0.68
HSS
r = 0.52
rho = 0.45
ORI
r = 0.67
rho = 0.61
5′/3′ WT
r = 0.59
rho = 0.61
5′/3′ MT
r = 0.67
rho = 0.62
5′/5′ WT
r = 0.62
rho = 0.63
5′/5′ MT
r = 0.66
rho = 0.61
3′/3′ WT
r = 0.50
rho = 0.48
log2 RNA/DNA per insert
CV fitted model prediction

B


Correlation between CV predictions and observations
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
pGL4
HSS
ORI
5′/3′ WT
5′/3′ MT
5′/5′ WT
5′/5′ MT
3′/3′ WT
Spearman
Pearson

C


Lasso coefficients, pGL4
−0.2 −0.1 0.0 0.1 0.2
HNF1B
DNase.1
c.Jun
KAT8
TEAD1
JUND.1
BCL6
HNF1A
FOSL2
NFYB
PAX3
FOXP1
THAP11
SOX11
EP300.1
E2F6
SMC3
JUN
CHD4
TBP
RCOR2
ZFP64
TCF12
KAT2B
MXD3
SHOX
NONO.1
CEBPD.1
JUND
NFKBIZ
Lasso coefficients, HSS
−0.2 −0.1 0.0 0.1 0.2
c.Jun
FOSL2
TEAD1
CpG_content
MNT.1
JUND
SP4
JUN
JUND.1
ZBTB40
PRDM1
ZNF652
CEBPB.1
NR2F1
KDM5A
SART3
ATF2
SKI
SIN3A
FUS
VDR
SP1
RAD21.1
RAD21.2
NONO.1
ZNF7
HNF1B
YY1
CEBPD.1
HNF4G
Lasso coefficients, ORI
−0.2 −0.1 0.0 0.1 0.2
TEAD1
c.Jun
JUN
YY1
JUND
JUND.1
CEBPA
FOXP1
KAT8
JunD.1
RAD21.2
TP53
KDM6A
TEAD4
NR2F1
HNF1B
ZFP202
SREBF1
ATF3
RCOR2
ATF4
CEBPD.1
DMAP1
CEBPB.1
ATF7
CHD4
BCL6
YY1
SMARCE1
RAD21.1
Lasso coefficients, 5′/3′ WT
−0.2 −0.1 0.0 0.1 0.2
EP300
HNF1B
EP300.1
FOXP1
DMAP1
SP1
KDM6A
GABPA
RAD21.2
NFYB
HNF1A
JUND.1
ETV5
SOX11
THAP11
JunD.1
c.Jun
TCF7
POLR2G
ETS1
MEF2D
NFAT5
BCL6
USF2
DNase.1
TBP
JUN
FOSL2
U2AF2
SRSF10.3
Lasso coefficients, 5′/3′ MT
−0.2 −0.1 0.0 0.1 0.2
HNF1B
FOSL2
NFYB
THAP11
SP1
CAGE.plus
SOX11
CEBPB
MEF2D
ZFP64
FOSL2
HCFC1
GATAD1
PLRG1
Nrf1
ZNF143.1
TCF7
JUND.1
KLF8
JunD.1
DMAP1
BCL6
CUX2
NFE2
EP300
DNMT3B
GABPA.1
H3K4me2
NONO.1
HDAC6
Lasso coefficients, 5′/5′ WT
−0.2 −0.1 0.0 0.1 0.2
HNF1B
DNase.1
FOXP1
EP300.1
EP300
SOX11
NFYB
BCL6
KDM6A
NFAT5
HNF1A
c.Jun
DMAP1
SMC3
ETS1
FOSL2
JUN
JUND.1
DNase
MNT.1
ETV5
ZFP64
NONO.1
ETV1
H3K27ac
TCF7
ARID3A
CHD4
POLR2G
ZKSCAN8
Lasso coefficients, 5′/5′ MT
−0.2 −0.1 0.0 0.1 0.2
HNF1B
FOSL2
NFYB
SOX11
FOSL2
BCL6
KAT8
CEBPB
THAP11
DNase.1
HNF1A
ARID3A
NONO.1
EP300.1
NFE2
ETS1
ZNF7
LIN28A.1
IGF2BP2
CUX2
CHD4
H3K4me3
GATAD1
GABPA.1
MNT.1
ANKHD1
Nrf1
FOXP1
PAF1
ETV1
Lasso coefficients, 3′/3′ WT
−0.2 −0.1 0.0 0.1 0.2
PTBP1.1
DMAP1
CAGE.plus.3
ZNF638
c.Jun
SYNCRIP
FOXP1.1
RAD21.2
Pol2.1
JUND.1
ETS1.1
SRSF1.4
CEBPB.1
CUX2
YBX1
CAGE.minus.4
ZNF217
Rad21
CREM
ETS1
SKI
NFRKB
TIA1
KHDRBS2
SIN3A
ZBTB7A
RBM24
NR2F1
MNT.1
ZBTB1
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
Epigenomics Roadmap
FANTOM consortium
Other


D
Pearson correlation
−1 0 1
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
Epigenomics Roadmap
FANTOM consortium
Other

Supplemental Figure 6. Coefficients and performance of the models to predict element efficacy among the MPRA methods.

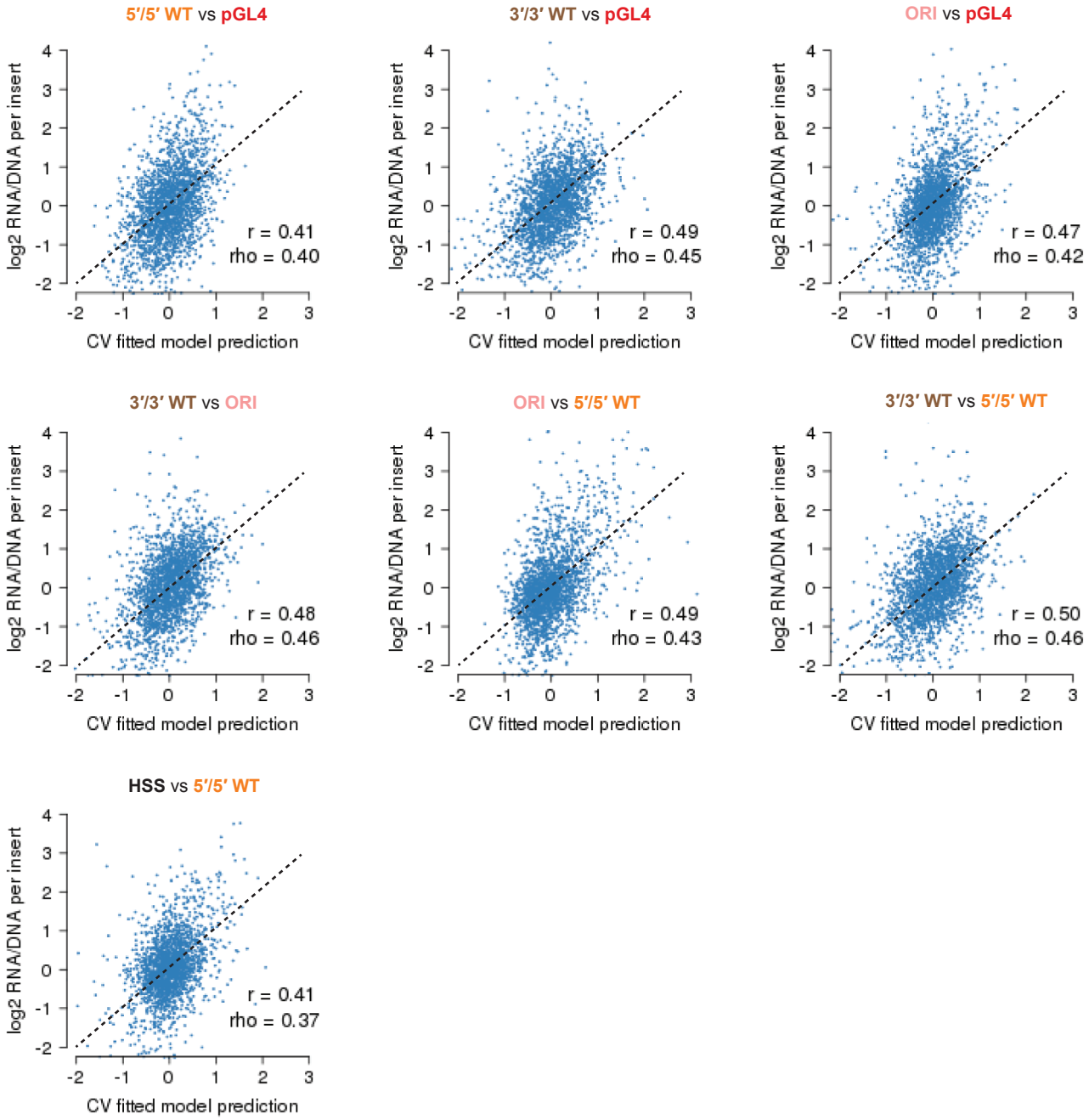
A) Scatter plots displaying the relationship between the 10-fold cross-validated predictions derived from lasso regression models and the observed RNA/DNA ratios, for each of 8 MPRA methods tested (excluding 3′/3′ MT due to poor inter-replicate reproducibility). Also indicated are the Pearson (*r*) and Spearman (rho) correlation values.

B) Pearson and Spearman correlation coefficients for 10-fold cross-validated predictions derived from lasso regression models and the observed RNA/DNA ratios, for each of the 8 MPRA methods tested.

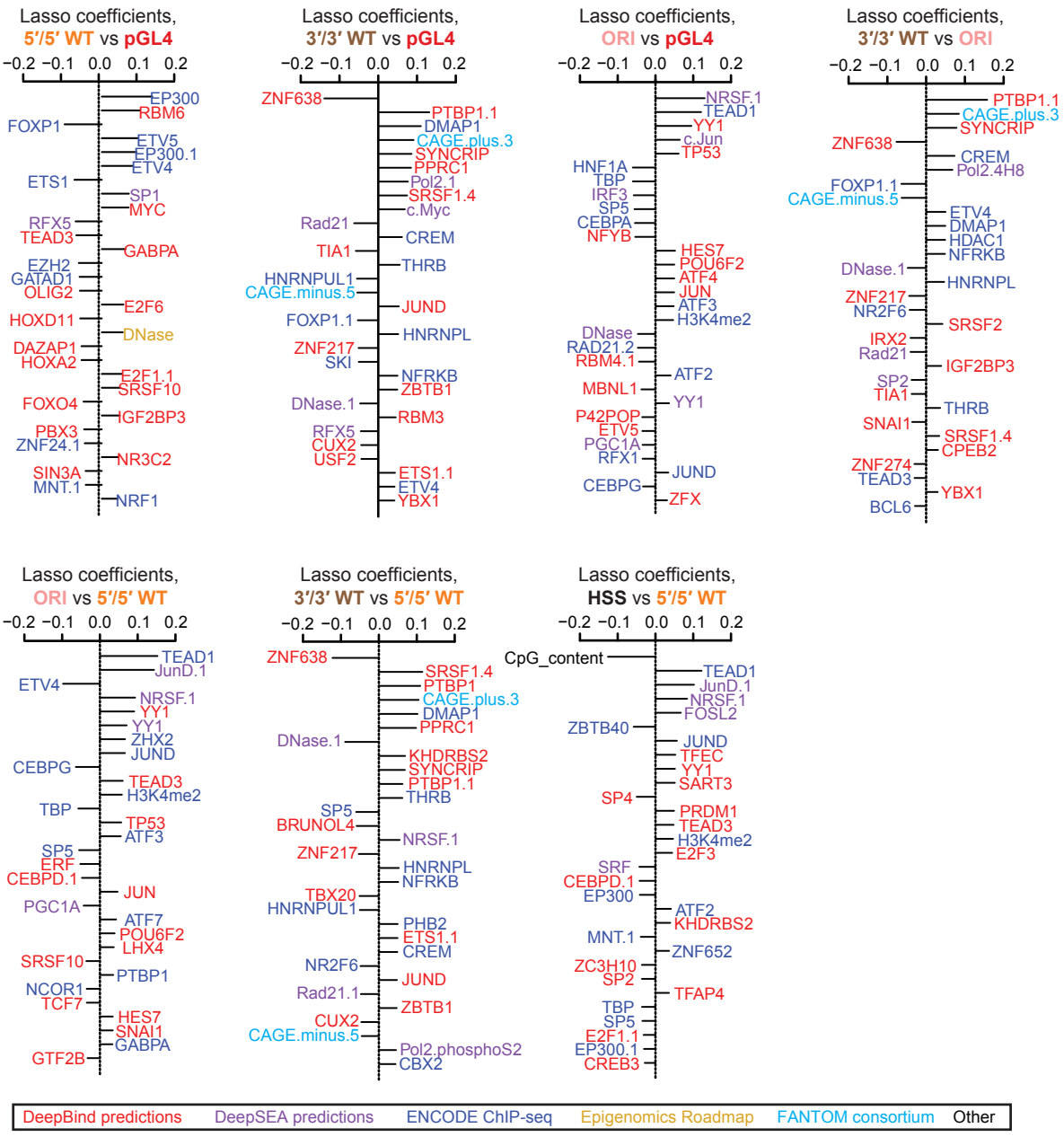
C) The top 30 coefficients derived from a lasso regression model trained on the full dataset derived from the eight indicated MPRA methods. Features with the extension ".1", ".2", etc allude to redundant features or replicate samples.

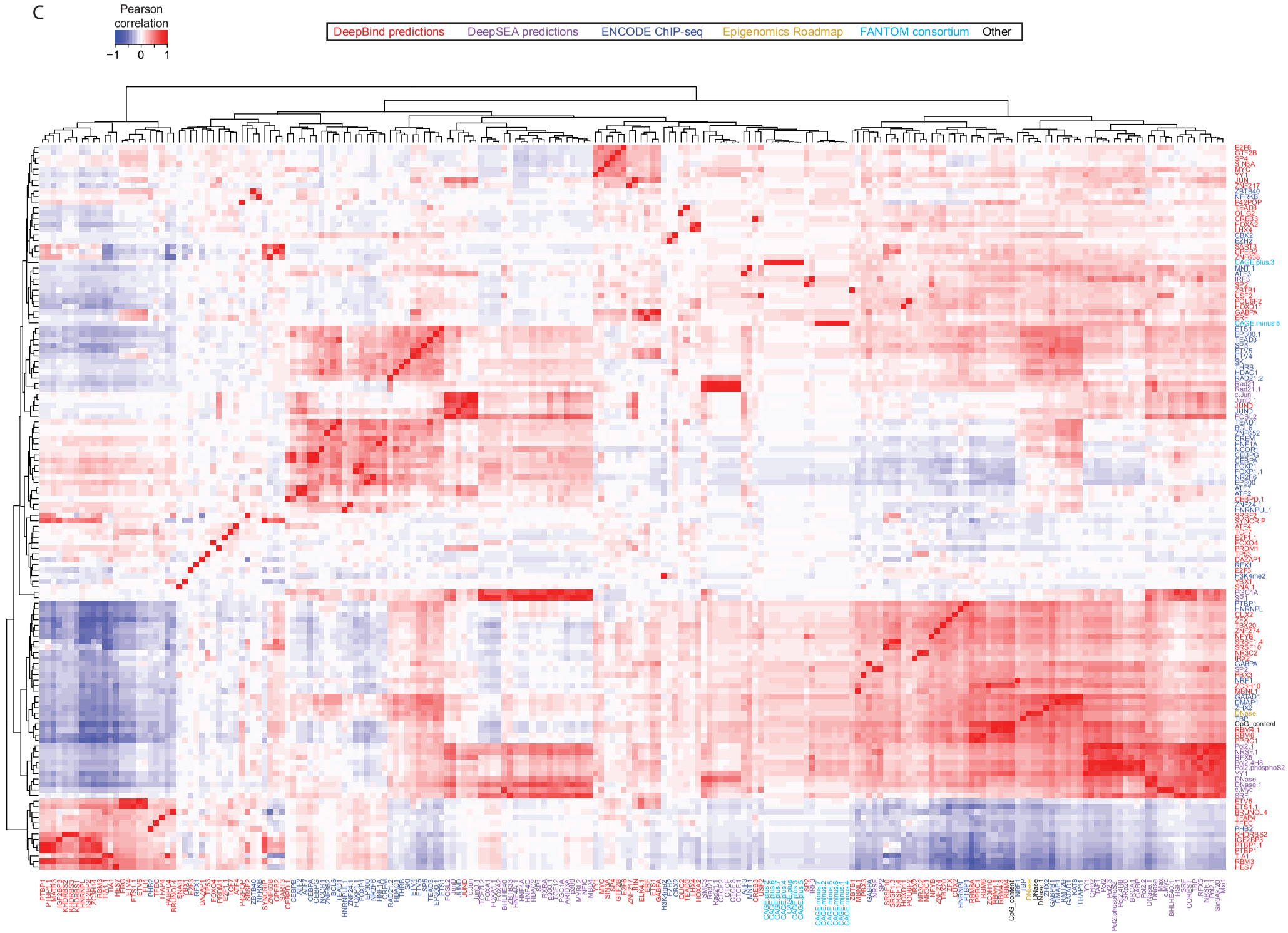
D) Figure has been rotated to fit to sheet. Pearson correlation matrix between the union of all top 30 features from (C), shown as rows, and other features sharing a Pearson correlation either ≤ –0.8 or ≥ 0.8, shown as columns. Feature names are colored according to the origin of the feature as shown in the boxed key above. Hierarchical clustering was used to group features exhibiting similar correlation patterns.

A

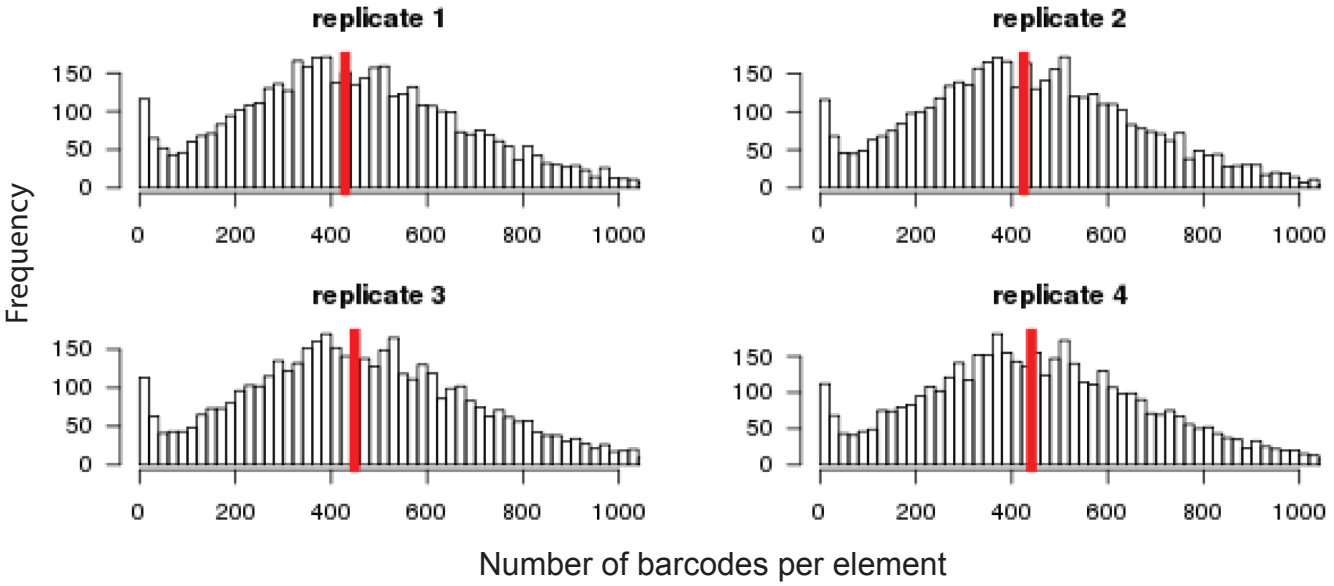

5′/5′ WT vs pGL4
r = 0.41
rho = 0.40
3′/3′ WT vs pGL4
r = 0.49
rho = 0.45
ORI vs pGL4
r = 0.47
rho = 0.42
3′/3′ WT vs ORI
r = 0.48
rho = 0.46
ORI vs 5′/5′ WT
r = 0.49
rho = 0.43
3′/3′ WT vs 5′/5′ WT
r = 0.50
rho = 0.46
HSS vs 5′/5′ WT
r = 0.41
rho = 0.37
log2 RNA/DNA per insert
CV fitted model prediction

B


Lasso coefficients,
5′/5′ WT vs pGL4
−0.2 −0.1 0.0 0.1 0.2
EP300
RBM6
FOXP1
ETV5
EP300.1
ETV4
ETS1
SP1
MYC
RFX5
TEAD3
GABPA
EZH2
GATAD1
OLIG2
E2F6
HOXD11
DNase
DAZAP1
HOXA2
E2F1.1
SRSF10
FOXO4
IGF2BP3
PBX3
ZNF24.1
NR3C2
SIN3A
MNT.1
NRF1
Lasso coefficients,
3′/3′ WT vs pGL4
−0.2 −0.1 0.0 0.1 0.2
ZNF638
PTBP1.1
DMAP1
CAGE.plus.3
SYNCRIP
PPRC1
Pol2.1
SRSF1.4
c.Myc
Rad21
CREM
TIA1
THRB
HNRNPUL1
CAGE.minus.5
JUND
FOXP1.1
HNRNPL
ZNF217
SKI
NFRKB
ZBTB1
DNase.1
RBM3
RFX5
CUX2
USF2
ETS1.1
ETV4
YBX1
Lasso coefficients,
ORI vs pGL4
−0.2 −0.1 0.0 0.1 0.2
NRSF.1
TEAD1
YY1
c.Jun
TP53
HNF1A
TBP
IRF3
SP5
CEBPA
NFYB
HES7
POU6F2
ATF4
JUN
ATF3
H3K4me2
DNase
RAD21.2
RBM4.1
ATF2
MBNL1
YY1
P42POP
ETV5
PGC1A
RFX1
JUND
CEBPG
ZFX
Lasso coefficients,
3′/3′ WT vs ORI
−0.2 −0.1 0.0 0.1 0.2
PTBP1.1
CAGE.plus.3
SYNCRIP
ZNF638
CREM
Pol2.4H8
FOXP1.1
CAGE.minus.5
ETV4
DMAP1
HDAC1
NFRKB
DNase.1
HNRNPL
ZNF217
NR2F6
SRSF2
IRX2
Rad21
IGF2BP3
SP2
TIA1
THRB
SNAI1
SRSF1.4
CPEB2
ZNF274
TEAD3
YBX1
BCL6
Lasso coefficients,
ORI vs 5′/5′ WT
−0.2 −0.1 0.0 0.1 0.2
TEAD1
JunD.1
ETV4
NRSF.1
YY1
YY1
ZHX2
JUND
CEBPG
TEAD3
H3K4me2
TBP
TP53
ATF3
SP5
ERF
CEBPD.1
JUN
PGC1A
ATF7
POU6F2
LHX4
SRSF10
PTBP1
NCOR1
TCF7
HES7
SNAI1
GABPA
GTF2B
Lasso coefficients,
3′/3′ WT vs 5′/5′ WT
−0.2 −0.1 0.0 0.1 0.2
ZNF638
SRSF1.4
PTBP1
CAGE.plus.3
DMAP1
PPRC1
DNase.1
KHDRBS2
SYNCRIP
PTBP1.1
THRB
SP5
BRUNOL4
NRSF.1
ZNF217
HNRNPL
NFRKB
TBX20
HNRNPUL1
PHB2
ETS1.1
CREM
NR2F6
JUND
Rad21.1
ZBTB1
CUX2
CAGE.minus.5
Pol2.phosphoS2
CBX2
Lasso coefficients,
HSS vs 5′/5′ WT
−0.2 −0.1 0.0 0.1 0.2
CpG_content
TEAD1
JunD.1
NRSF.1
FOSL2
ZBTB40
JUND
TFEC
YY1
SART3
SP4
PRDM1
TEAD3
H3K4me2
E2F3
SRF
CEBPD.1
EP300
ATF2
KHDRBS2
MNT.1
ZNF652
ZC3H10
SP2
TFAP4
TBP
SP5
E2F1.1
EP300.1
CREB3
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
Epigenomics Roadmap
FANTOM consortium
Other

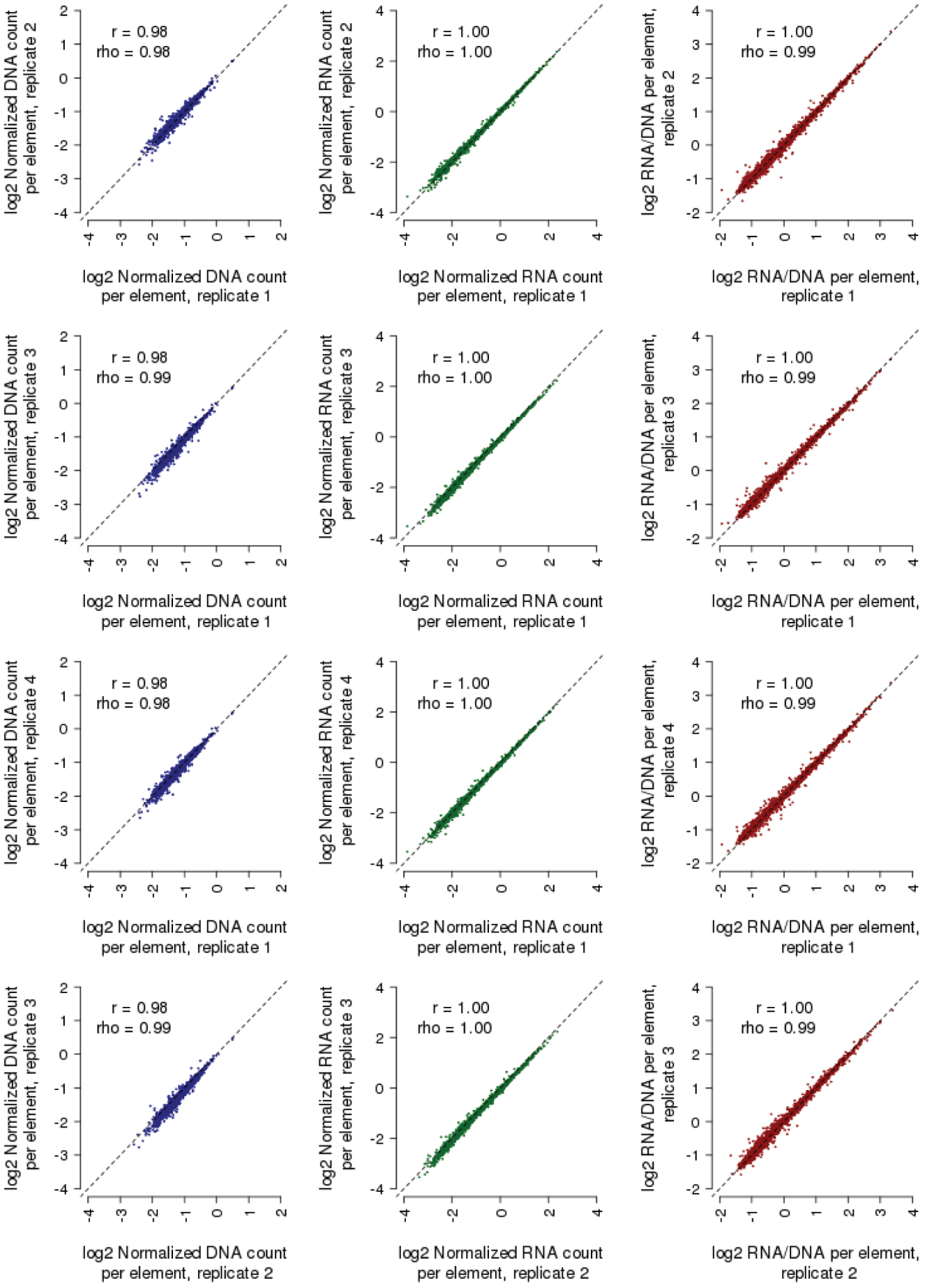

C
Pearson correlation
−1
0
1
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
Epigenomics Roadmap
FANTOM consortium
Other

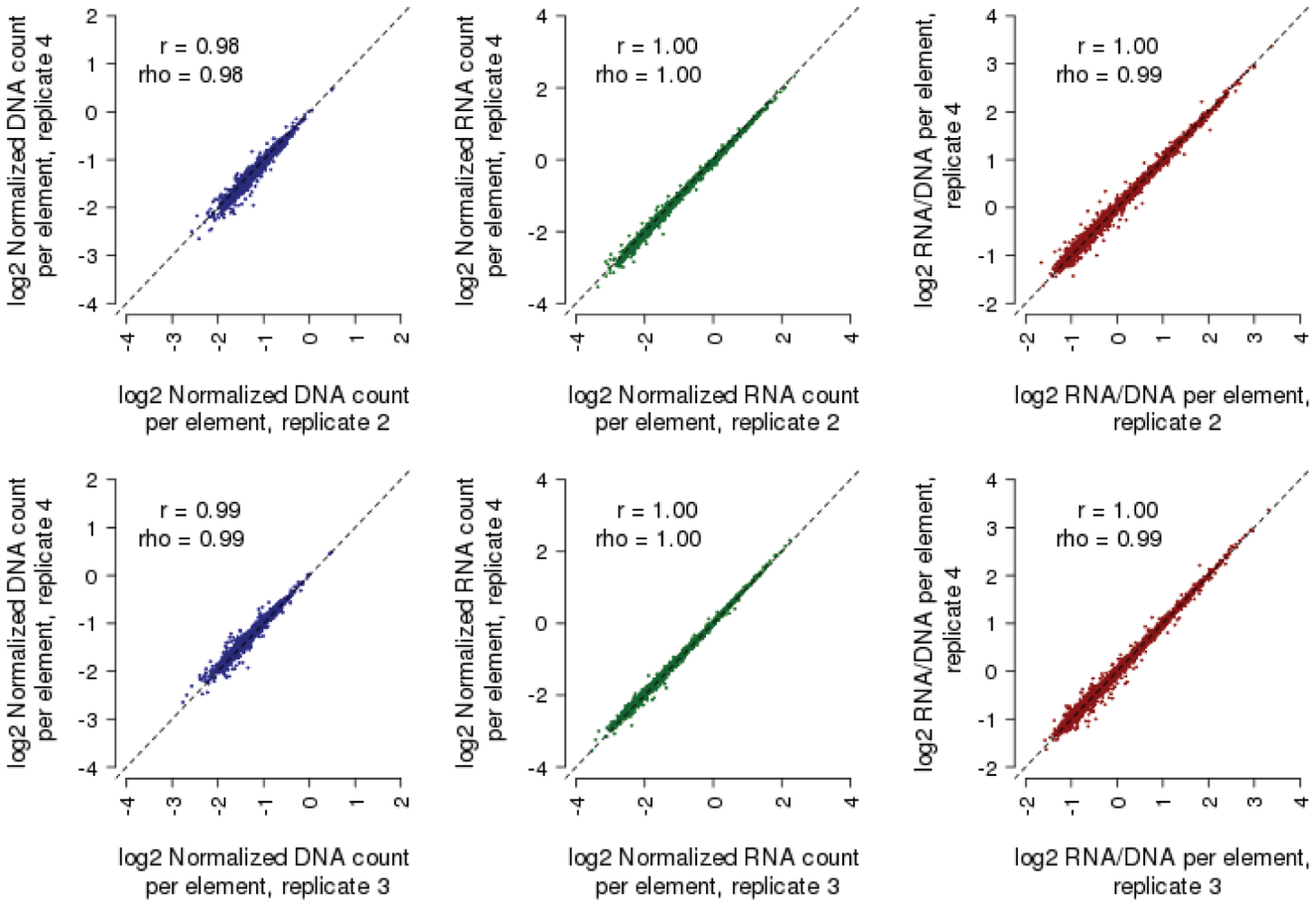
Supplemental Figure 7. Coefficients and performance of the models to predict differential element efficacy among the MPRA methods.
A) Scatter plots displaying the relationship between the 10-fold cross-validated predictions derived from lasso regression models and the observed RNA/DNA ratios, for each of the 7 indicated differential comparisons tested. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values.
B) The top 30 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 extension ".1", ".2", etc. allude to redundant features or replicate samples.
C) Figure has been rotated to fit to sheet. Pearson correlation matrix between the union of all top 30 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 above. Hierarchical clustering was used to group features exhibiting similar correlation patterns.



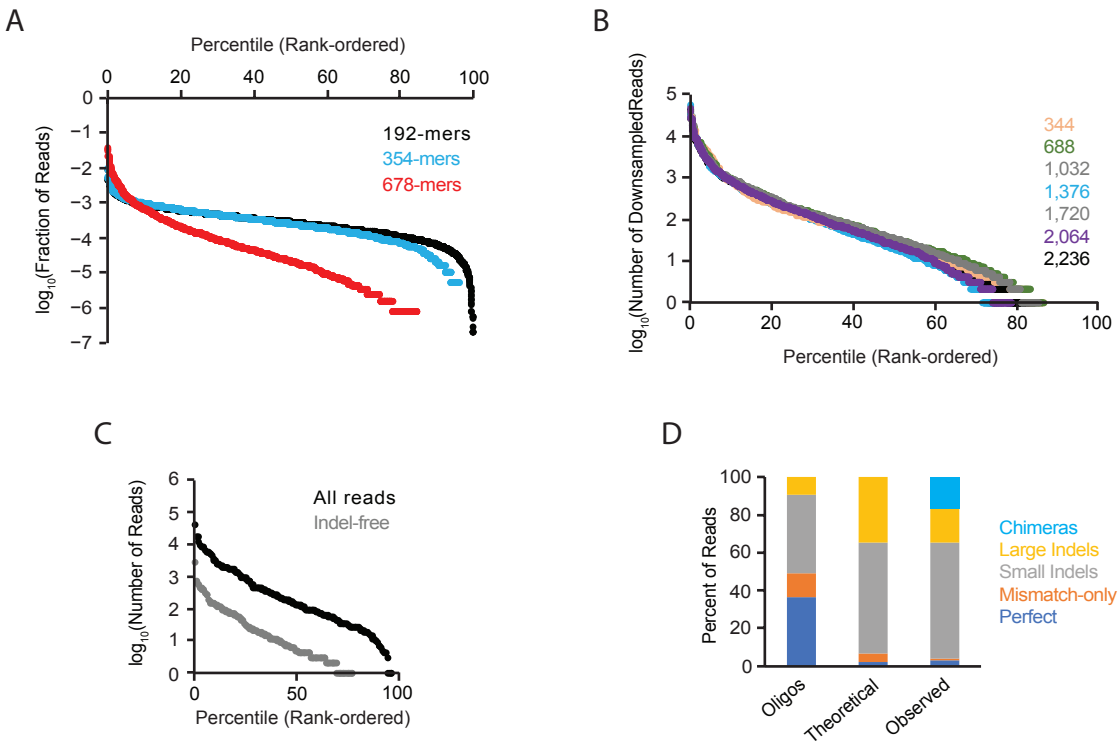
Supplemental Figure 8. Histograms of barcodes per element for each replicate of the the library testing orientation.
Shown are histograms indicating the number of observed barcodes per element, for each of the 2,336 elements tested in each of the 4 replicates for the library testing orientation (i.e., forward and reverse). Shown with a vertical red line is the median number of barcodes per element.

Figure S9


r = 0.98
rho = 0.98
log2 Normalized DNA count per element, replicate 2
log2 Normalized DNA count per element, replicate 1
r = 1.00
rho = 1.00
log2 Normalized RNA count per element, replicate 2
log2 Normalized RNA count per element, replicate 1
r = 1.00
rho = 0.99
log2 RNA/DNA per element, replicate 2
log2 RNA/DNA per element, replicate 1
r = 0.98
rho = 0.99
log2 Normalized DNA count per element, replicate 3
log2 Normalized DNA count per element, replicate 1
r = 1.00
rho = 1.00
log2 Normalized RNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 1
r = 1.00
rho = 0.99
log2 RNA/DNA per element, replicate 3
log2 RNA/DNA per element, replicate 1
r = 0.98
rho = 0.98
log2 Normalized DNA count per element, replicate 4
log2 Normalized DNA count per element, replicate 1
r = 1.00
rho = 1.00
log2 Normalized RNA count per element, replicate 4
log2 Normalized RNA count per element, replicate 1
r = 1.00
rho = 0.99
log2 RNA/DNA per element, replicate 4
log2 RNA/DNA per element, replicate 1
r = 0.98
rho = 0.99
log2 Normalized DNA count per element, replicate 3
log2 Normalized DNA count per element, replicate 2
r = 1.00
rho = 1.00
log2 Normalized RNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 2
r = 1.00
rho = 0.99
log2 RNA/DNA per element, replicate 3
log2 RNA/DNA per element, replicate 2



Supplemental Figure 9. Reproducibility among replicates for the library testing orientation.
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 the four replicates, including data from both orientations (i.e., forward and reverse) tested. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values. Candidate enhancers supported by fewer than 10 barcodes were filtered out prior to this analysis to reduce the impact of technical noise.



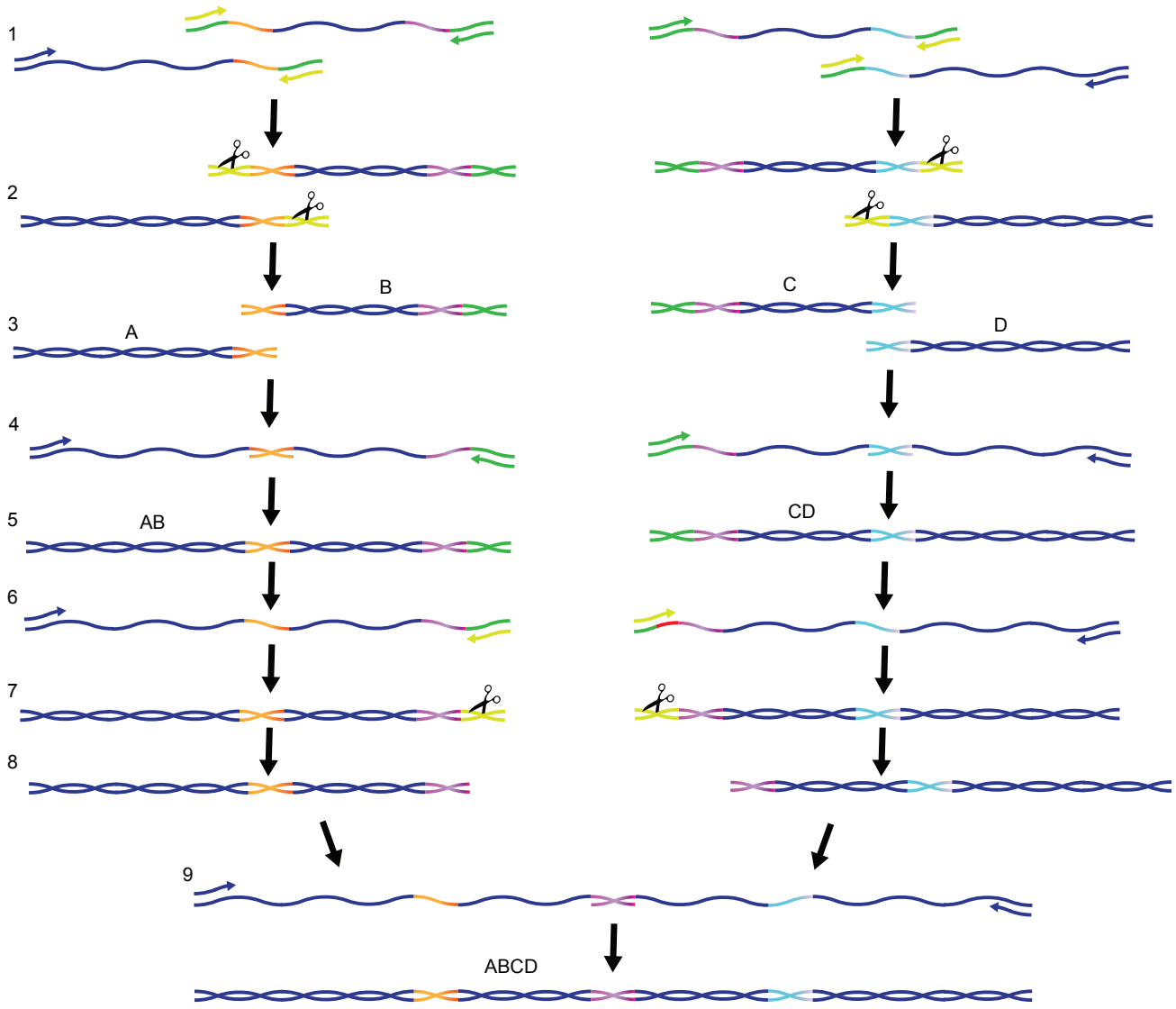
Supplemental Figure 10. Evaluation of Multiplex Pairwise Assembly (MPA) and Hierarchical Multiplex Pairwise Assembly (HMPA) library quality.

A) Plot of the uniformity of indel-free reads for 2,336 x 192mers (amplified off Agilent 230mer array), 2,336 x 354mers (after Multiplex Pairwise Assembly), and 2,236 x 678mers (after a single Hierarchical Multiplex Pairwise Assembly). The x-axis is rank-ordered according to the most to least abundant from left to right. The y-axis is the fraction of either indel-free reads (for 192mers and 354mers) or all reads (for 678mers).

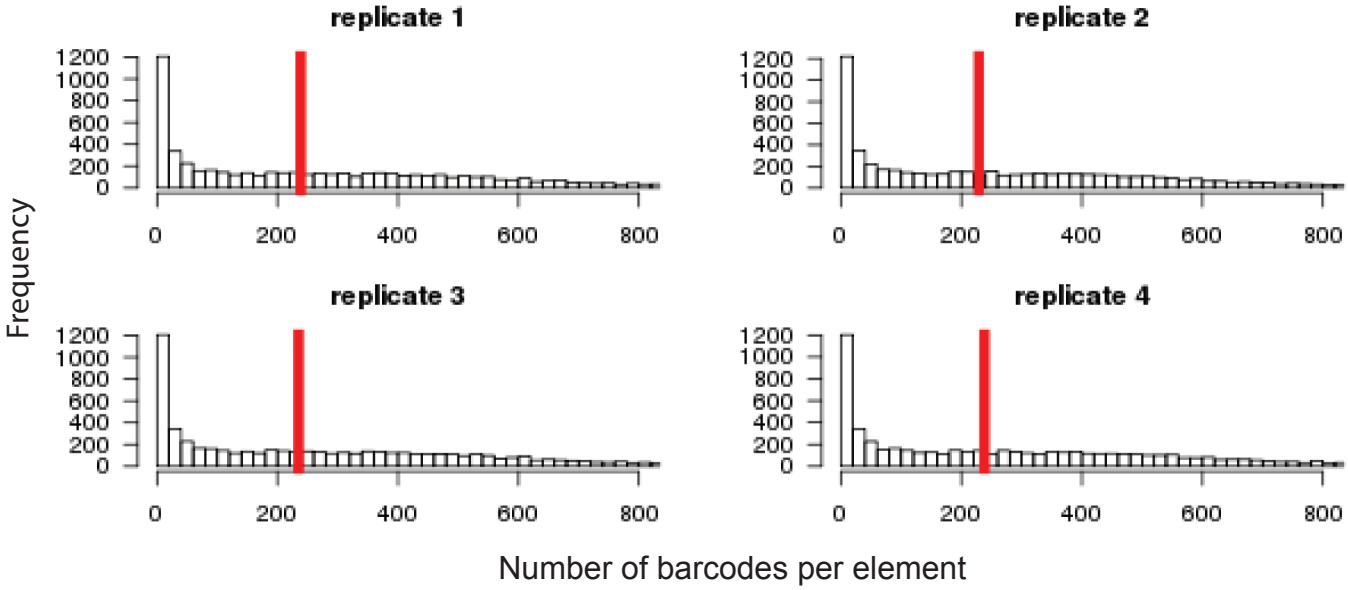
B) Uniformity for various HMPA reactions, with total number of target sequences ranging from 344 to 2,236. The sequencing reads were down-sampled to normalize for the total number of targets (# reads = 581 * # targets). The Y-axis is the number of down-sampled reads from a given target sequence.

C) One sub-library (of 172 targets), was sequenced on a PacBio Sequel. The plot shows the uniformity for all aligning reads (black) and indel-free reads (grey).

D) Composition of the sub-library of 172 targets. The first column shows the breakdown of oligos by Illumina sequencing. The second column shows the theoretical breakdown of 678mers based on each target consisting of four independent oligos. The third column shows the breakdown of 678mers based on PacBio sequencing.

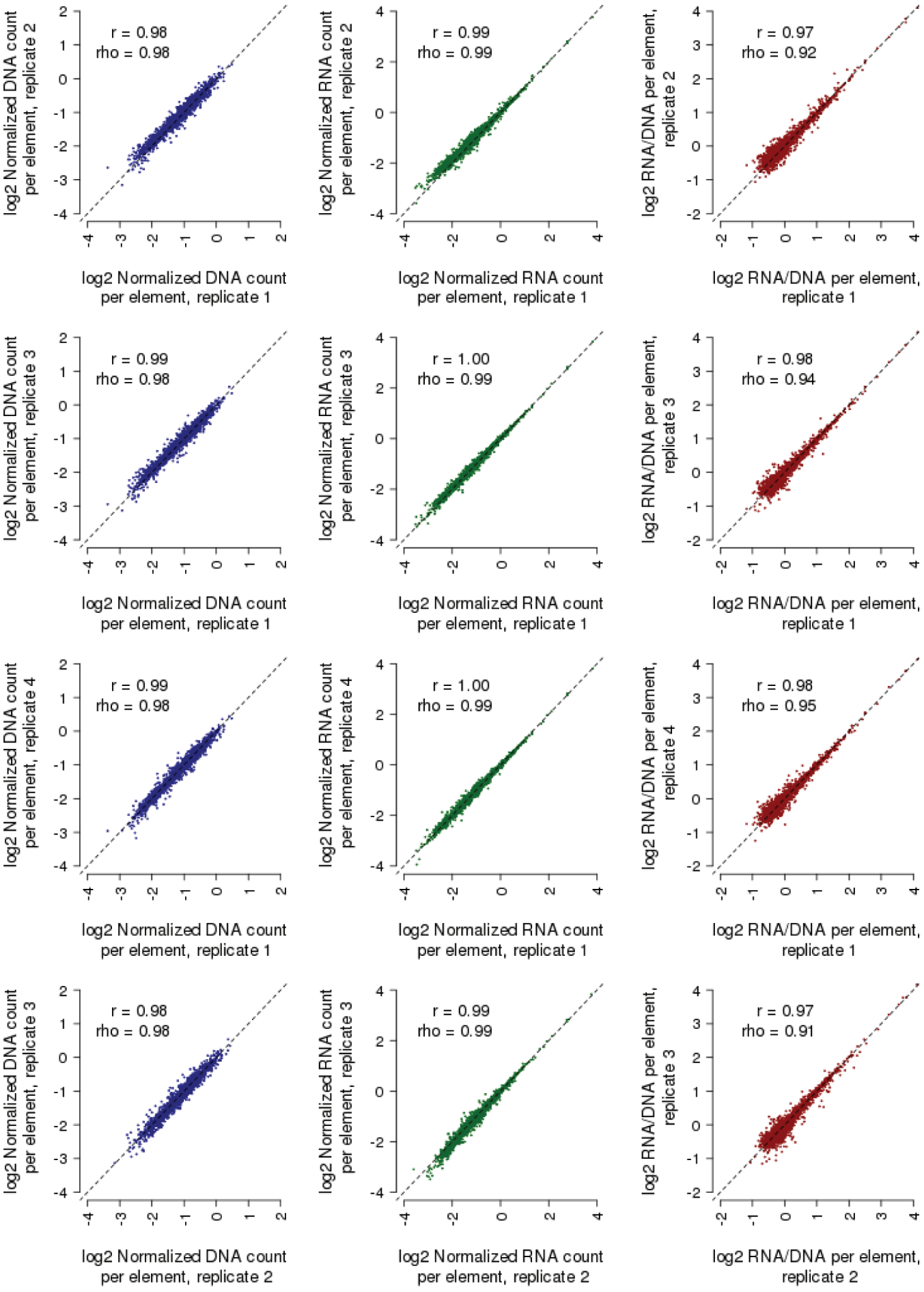


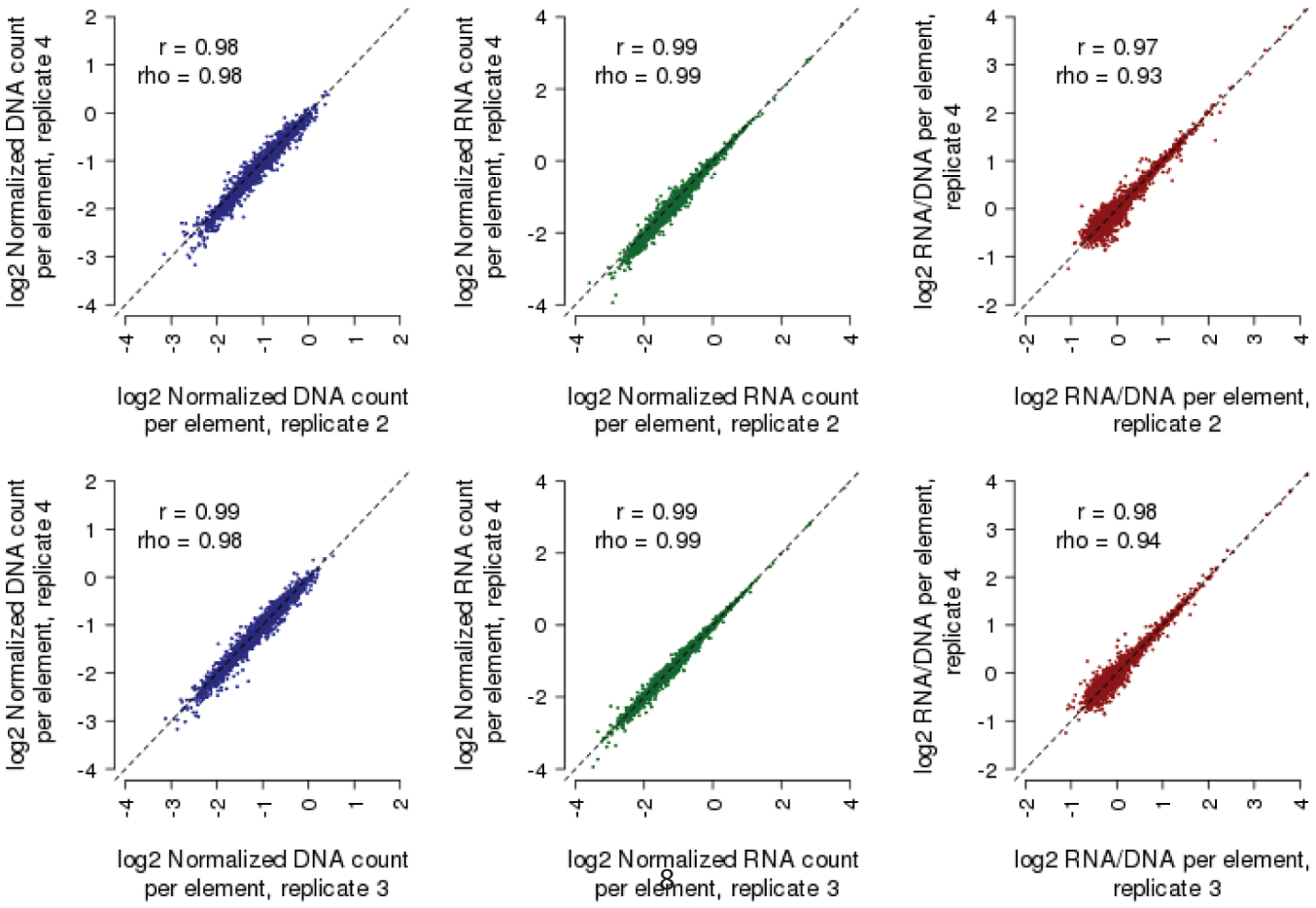
Supplemental Figure 11. Hierarchical Multiplex Pairwise Assembly (HMPA) strategy.
1. To generate a library of 678 bp enhancers, we ordered each enhancer as four oligonucleotides to be assembled (fragments “A”, “B”, “C”, and “D”). To assemble fragments “AB” and CD”, sequences were designed such that the 3′ end of fragments A and C had 30 bps of homology to the 5′ ends of fragments B and D, respectively (shown in red and orange). To remove adapter sequences from these ends of the fragments, uracil primers (shown in yellow) were used to incorporate uracils into the adapters during qPCR amplification.
2. The resulting fragments were treated with USER enzyme (scissors) and put into an end-repair reaction, **3.** effectively removing the adapters.
4. Fragments were assembled in a qPCR reaction by allowing the fragments to first anneal to one another for 5 cycles of PCR without primers, and then adding primers targeting the 5′ end of fragments A and C, and the 3′ ends of fragments B and D, **5.** resulting in fragments “AB” and “CD”.
6-9. Adapter sequences were removed from the 3′ end of AB fragments and the 5′ end of CD fragments, and the final 678 bp “ABCD” enhancer sequences were assembled using the aforementioned assembly reaction.



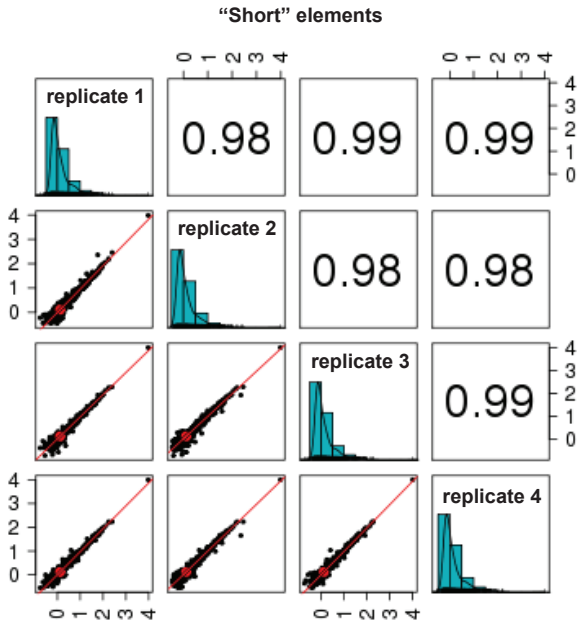
Supplemental Figure 12. Histograms of barcodes per element for each replicate of the the library testing element size.
Shown are histograms indicating the number of observed barcodes per element, for each of the 2,336 elements tested in each of the 4 replicates for the library testing size classes (i.e., short, medium, and long). Shown with a vertical red line is the median number of barcodes per element.

A

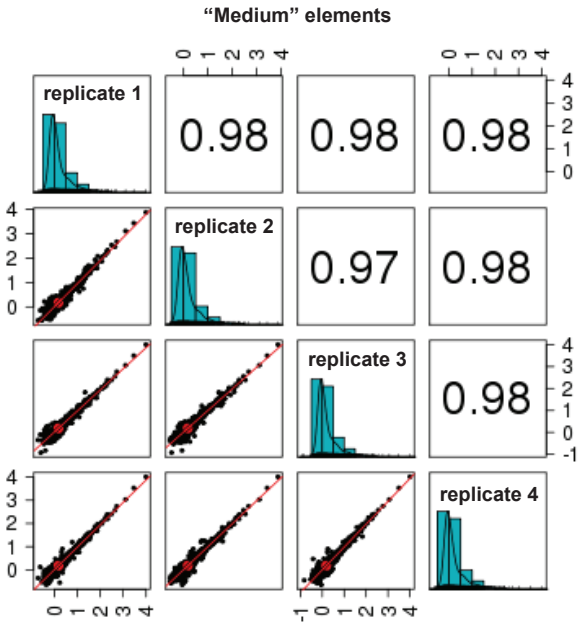

r = 0.98
rho = 0.98
log2 Normalized DNA count per element, replicate 2
log2 Normalized DNA count per element, replicate 1
r = 0.99
rho = 0.99
log2 Normalized RNA count per element, replicate 2
log2 Normalized RNA count per element, replicate 1
r = 0.97
rho = 0.92
log2 RNA/DNA per element, replicate 2
log2 RNA/DNA per element, replicate 1
r = 0.99
rho = 0.98
log2 Normalized DNA count per element, replicate 3
log2 Normalized DNA count per element, replicate 1
r = 1.00
rho = 0.99
log2 Normalized RNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 1
r = 0.98
rho = 0.94
log2 RNA/DNA per element, replicate 3
log2 RNA/DNA per element, replicate 1
r = 0.99
rho = 0.98
log2 Normalized DNA count per element, replicate 4
log2 Normalized DNA count per element, replicate 1
r = 1.00
rho = 0.99
log2 Normalized RNA count per element, replicate 4
log2 Normalized RNA count per element, replicate 1
r = 0.98
rho = 0.95
log2 RNA/DNA per element, replicate 4
log2 RNA/DNA per element, replicate 1
r = 0.98
rho = 0.98
log2 Normalized DNA count per element, replicate 3
log2 Normalized DNA count per element, replicate 2
r = 0.99
rho = 0.99
log2 Normalized RNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 2
r = 0.97
rho = 0.91
log2 RNA/DNA per element, replicate 3
log2 RNA/DNA per element, replicate 2


r = 0.98
rho = 0.98
log2 Normalized DNA count per element, replicate 4
log2 Normalized DNA count per element, replicate 2
r = 0.99
rho = 0.99
log2 Normalized RNA count per element, replicate 4
log2 Normalized RNA count per element, replicate 2
r = 0.97
rho = 0.93
log2 RNA/DNA per element, replicate 4
log2 RNA/DNA per element, replicate 2
r = 0.99
rho = 0.98
log2 Normalized DNA count per element, replicate 3
r = 0.99
rho = 0.99
log2 Normalized RNA count per element, replicate 3
r = 0.98
rho = 0.94
log2 RNA/DNA per element, replicate 3

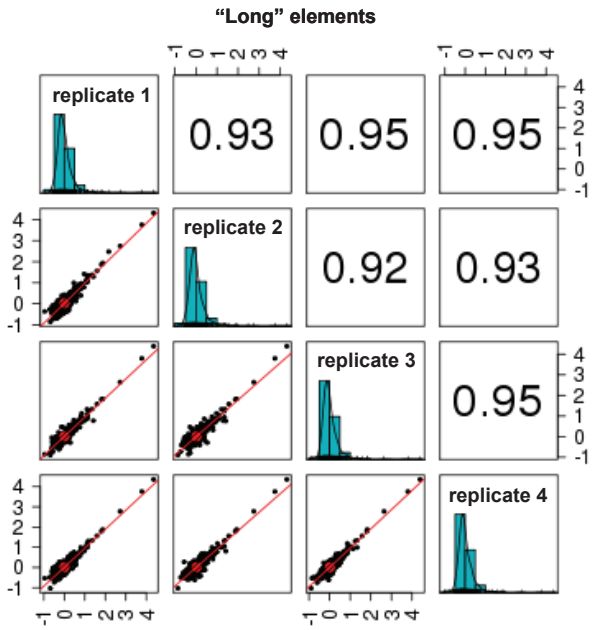
B


"Short" elements
replicate 1
replicate 2
replicate 3
replicate 4
0.98
0.99
0.99
0.98
0.98
0.99

C


"Medium" elements
replicate 1
replicate 2
replicate 3
replicate 4
0.98
0.98
0.98
0.97
0.98
0.98

D

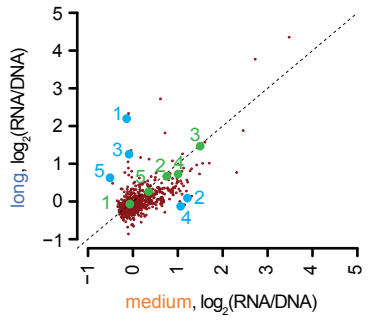


Supplemental Figure 13. Reproducibility among replicates for testing of libraries of the same enhancers with different amounts of sequence context.

A) 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 the four replicates, including data from all size classes (i.e., short, medium, and long) tested. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values. Candidate enhancers supported by fewer than 10 barcodes were filtered out prior to this analysis to reduce the impact of technical noise.

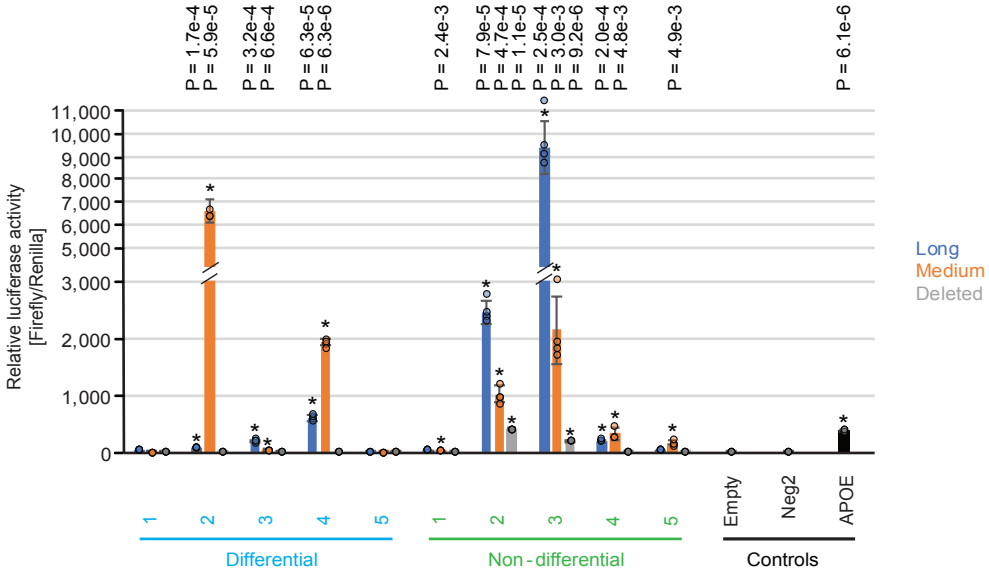
B-D) Pearson correlation between pairs of replicates for the subset of elements tested in the "short" (B), "medium" (C), or "long" (D) size class.

A

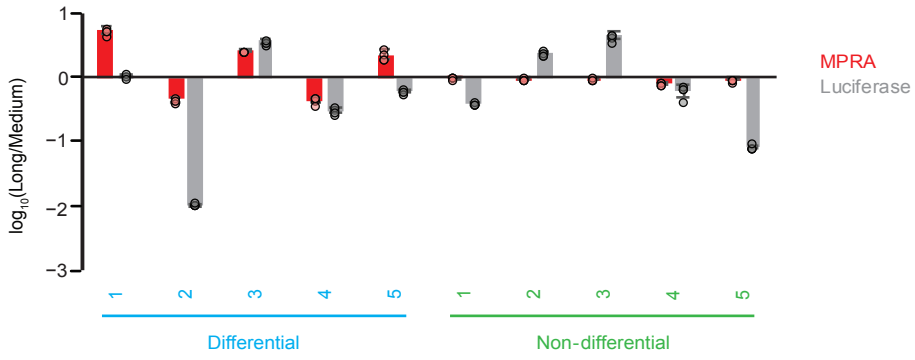


	#	Medium chr coordinates	Long chr coordinates
Differential	1	4:184031803-184032157	4:184031641-184032319
	2	1:120359023-120359377	1:120358861-120359539
	3	19:18490967-18491321	19:18490805-18491483
	4	9:119329149-119329503	9:119328987-119329665
	5	2:106744003-106744357	2:106743841-106744519
Non-differential	1	9:116149915-116150269	9:116149753-116150431
	2	15:66124997-66125351	15:66124835-66125513
	3	14:81769663-81770017	14:81769501-81770179
	4	2:3469702-3470056	2:3469540-3470218
	5	10:114391408-114391762	10:114391246-114391924

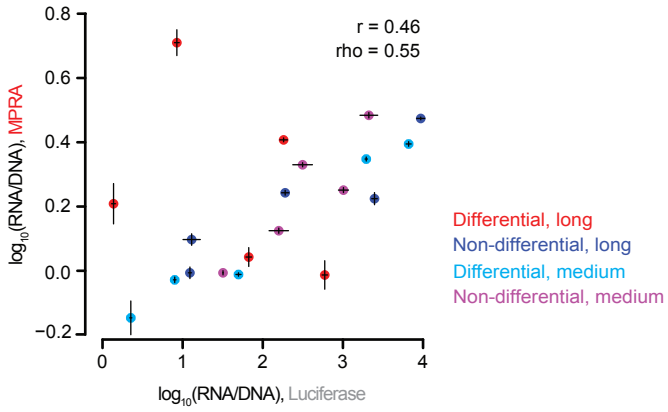
B



C

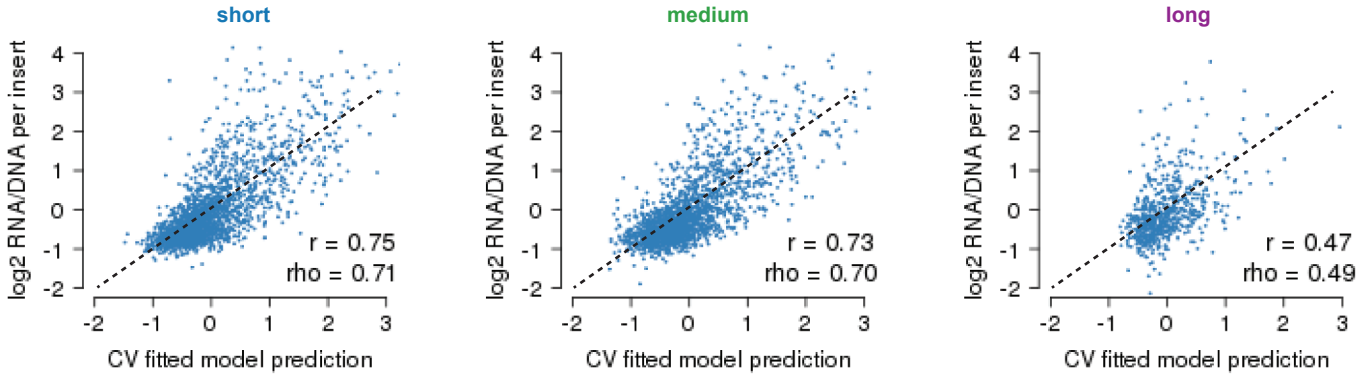


D

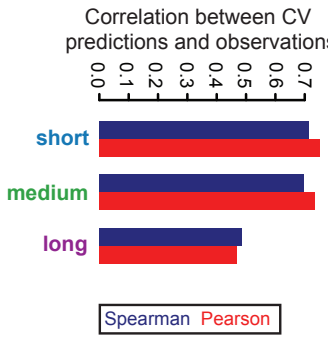


Supplemental Figure 14. Experimental evaluation of differential size activity with luciferase assays.
A) Scatter plots of the average activity score of each element, comparing medium vs. long versions of each element (as shown in **Figure 5C**). Five colored elements were chosen that exhibited non-differential (green) or differential (cyan) activity between the two size classes. The chromosomal coordinates are shown (human genome build "hg19") for each element.
B) The five non-differential and differential elements were cloned into the pGL4.23 vector and measured with a luciferase assay. A "deleted" version, in which the "long" sequence was tested after deleting its internal "medium" sequence, was also tested, along with an empty vector, negative control (Neg2), and positive control (APOE). n = 4 biologically independent cell cultures. Error bars indicate the standard error around the mean for four replicates, with significance assessed using a one-sided t-test comparing against Neg2 activity (*$p < 0.01$).
C) Summary of differential activity between long and medium sizes, as assessed by both the MPRA data shown in (A) and Luciferase data shown in (B). n = 4 biologically independent cell cultures. Error bars indicate the standard error around the mean for four replicates.
D) Scatter plot of the relationship between Luciferase and MPRA measurements. Error bars indicate the standard deviation among four independent replicates.

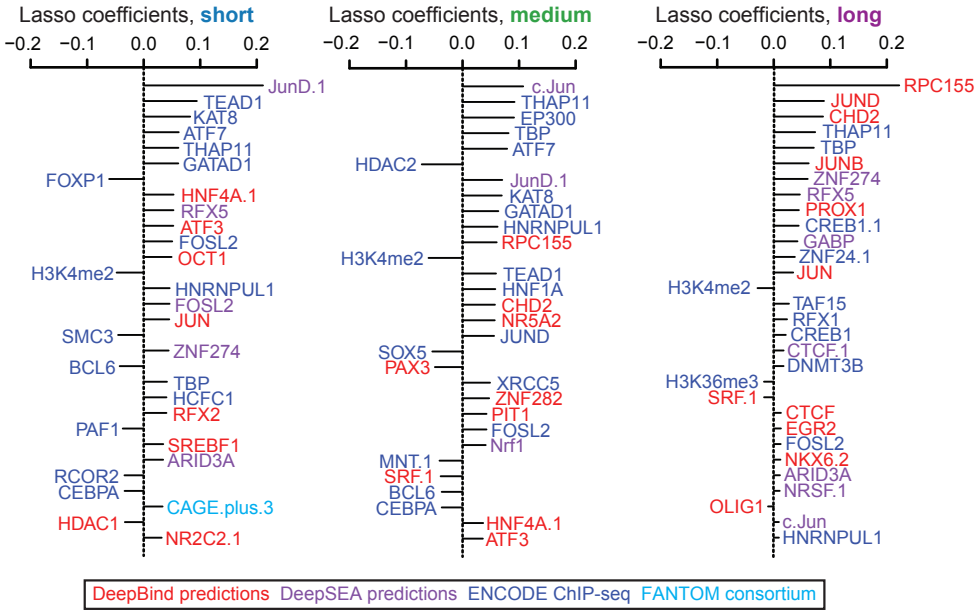
A

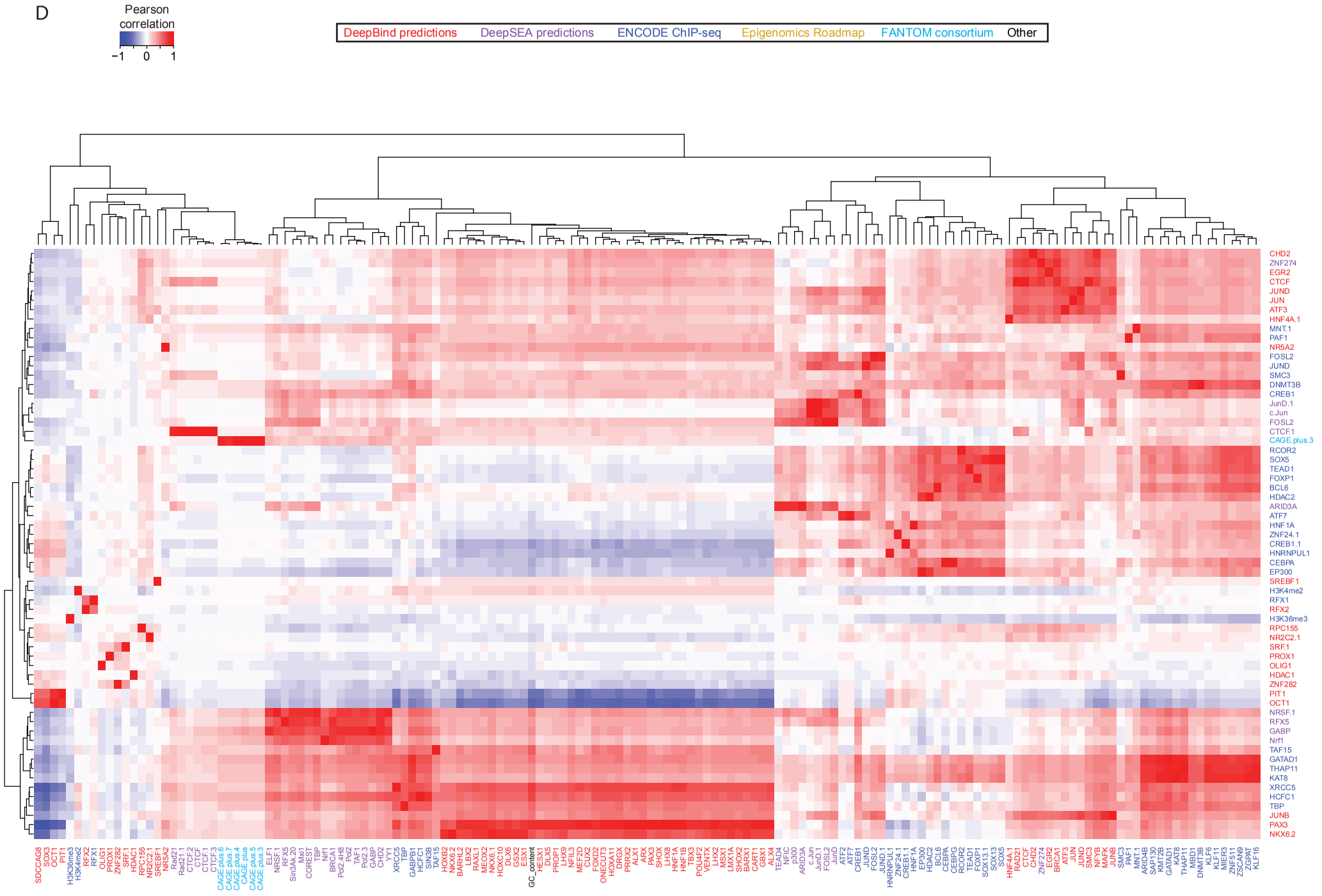

short
medium
long
log2 RNA/DNA per insert
CV fitted model prediction
r = 0.75
rho = 0.71
r = 0.73
rho = 0.70
r = 0.47
rho = 0.49

B


Correlation between CV predictions and observations
short
medium
long
Spearman Pearson

C


Lasso coefficients, short
Lasso coefficients, medium
Lasso coefficients, long
JunD.1
TEAD1
KAT8
ATF7
THAP11
GATAD1
FOXP1
HNF4A.1
RFX5
ATF3
FOSL2
OCT1
H3K4me2
HNRNPUL1
FOSL2
JUN
SMC3
ZNF274
BCL6
TBP
HCFC1
RFX2
PAF1
SREBF1
ARID3A
RCOR2
CEBPA
CAGE.plus.3
HDAC1
NR2C2.1
c.Jun
THAP11
EP300
TBP
ATF7
HDAC2
JunD.1
KAT8
GATAD1
HNRNPUL1
RPC155
H3K4me2
TEAD1
HNF1A
CHD2
NR5A2
JUND
SOX5
PAX3
XRCC5
ZNF282
PIT1
FOSL2
Nrf1
MNT.1
SRF.1
BCL6
CEBPA
HNF4A.1
ATF3
RPC155
JUND
CHD2
THAP11
TBP
JUNB
ZNF274
RFX5
PROX1
CREB1.1
GABP
ZNF24.1
JUN
H3K4me2
TAF15
RFX1
CREB1
CTCF.1
DNMT3B
H3K36me3
SRF.1
CTCF
EGR2
FOSL2
NKX6.2
ARID3A
NRSF.1
OLIG1
c.Jun
HNRNPUL1
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
FANTOM consortium


D
Pearson correlation
−1
0
1
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
Epigenomics Roadmap
FANTOM consortium
Other
CHD2
ZNF274
EGR2
CTCF
JUND
JUN
ATF3
HNF4A.1
MNT.1
PAF1
NR5A2
FOSL2
JUND
SMC3
DNMT3B
CREB1
JunD.1
c.Jun
FOSL2
CTCF.1
CAGE.plus.3
RCOR2
SOX5
TEAD1
FOXP1
BCL6
HDAC2
ARID3A
ATF7
HNF1A
ZNF24.1
CREB1.1
HNRNPUL1
CEBPA
EP300
SREBF1
H3K4me2
RFX1
RFX2
H3K36me3
RPC155
NR2C2.1
SRF.1
PROX1
OLIG1
HDAC1
ZNF282
PIT1
OCT1
NRSF.1
RFX5
GABP
Nrf1
TAF15
GATAD1
THAP11
KAT8
XRCC5
HCFC1
TBP
JUNB
PAX3
NKX6.2
SDCCAG8
SOX3
OCT1
PIT1
H3K36me3
H3K4me2
RFX2
RFX1
OLIG1
PROX1
ZNF282
SRF.1
HDAC1
RPC155
NR2C2.1
SREBF1
NR5A2
Rad21
Rad21.1
CTCF2
CTCF
CTCF.1
CTCF3
CAGE.plus.6
CAGE.plus.1
CAGE.plus.4
CAGE.plus
CAGE.plus.5
CAGE.plus.3
ELF1
NRSF.1
RFX5
Sin3AK.20
MAT
COREST
TBP
Nrf1
BRCA1
Pol2.4H8
Pol2
TAF1
Pol2.3
CHD2
YY1
XRCC5
TBP
GABPB1
HCFC1
SIN3B
TAF15
HOXA13
NKX6.2
BARHL2
LBX2
RAXL1
MEOX2
NKX6.1
HOXC10
DLX6
ESX1
ESX1
GC_content
HESX1
DLX5
PROP1
LHX9
NFIL3
MEF2D
FOXJ2
FOXD2
ONECUT3
HOXA11
DRGX
PRRX2
ALX1
ARX
PAX3
SHOX
LHX8
HNF1A
HNF1B
TBX3
POU4F2
VENTX
LHX2
MSX1
LMX1A
SHOX2
BARX1
CART1
GBX1
ALX4
TEAD4
NFIC
FOXI1
ARID3A
c.Jun
JunD.1
FOSL2
JunD
ATF2
ATF7
CREB1
JUND
FOSL2
JUND.1
HNRNPUL1
ZNF24.1
CREB1.1
HNF1A
EP300
HDAC2
BCL6
CEBPA
CEBPG
RCOR2
TEAD1
FOXP1
SOX13.1
SOX13
SOX5
HNF4A.1
RAD21
CTCF
CHD2
ZNF274
EGR2
BRCA1
ATF3
JUN
JUND
SMC3
NFYB
MAFK
JUNB
SMC3
PAF1
MNT.1
ARID4B
SAP130
KMT2B
GATAD1
KAT8
THAP11
MBD1
DNMT3B
KLF6
KLF11
MEF2A
ZNF511
ZSCAN9
ZGPAT
KLF16

Supplemental Figure 15. Coefficients and performance of the models to predict element efficacy among the three size classes.

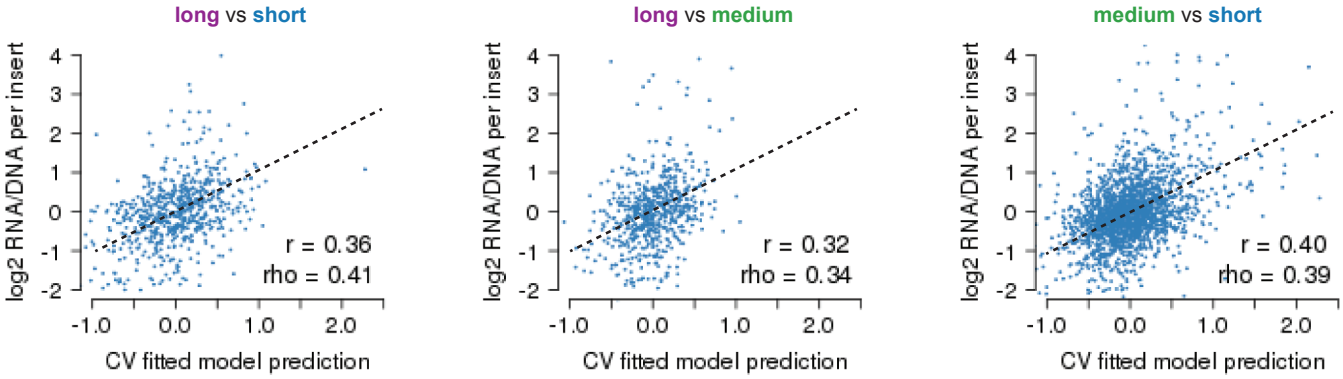
A) Scatter plots displaying the relationship between the 10-fold cross-validated predictions derived from lasso regression models and the observed RNA/DNA ratios, for each of the 3 size classes tested. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values.

B) Summary of these r and rho values for each of the comparisons shown in (A).

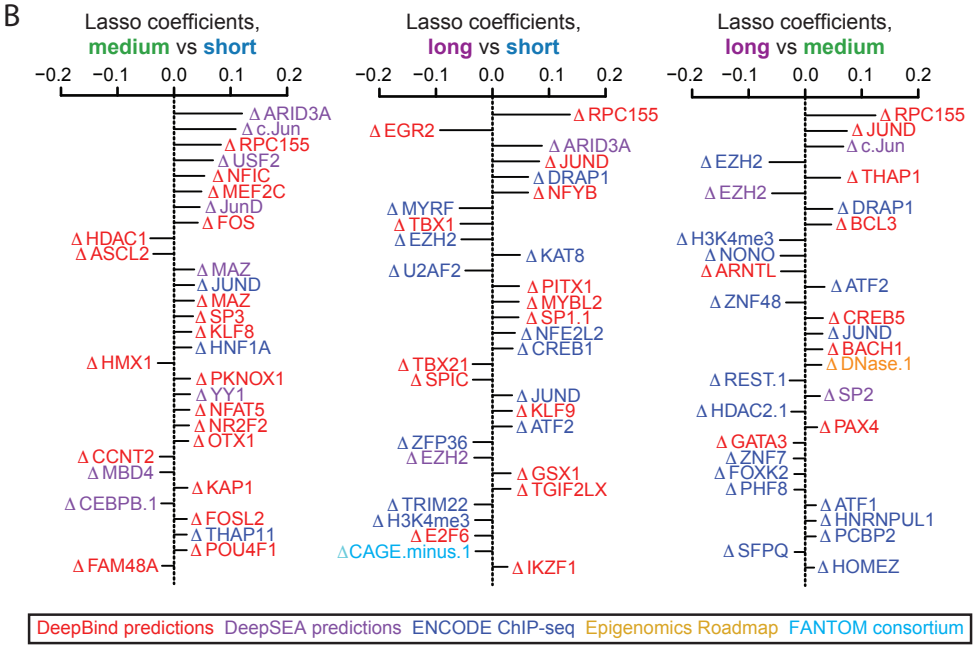
C) The top 30 coefficients derived from lasso regression models trained on the full dataset to predict observed values from the indicated size classes. Features with the extension ".1", ".2", etc allude to redundant features or replicate samples.

D) Figure has been rotated to fit to sheet. Pearson correlation matrix between the union of all top 30 features from (C), 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 above. Hierarchical clustering was used to group features exhibiting similar correlation patterns.

A


long vs short
log2 RNA/DNA per insert
r = 0.36
rho = 0.41
CV fitted model prediction
long vs medium
r = 0.32
rho = 0.34
medium vs short
r = 0.40
rho = 0.39

B


Lasso coefficients, medium vs short
Lasso coefficients, long vs short
Lasso coefficients, long vs medium
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
Epigenomics Roadmap
FANTOM consortium


C
Pearson correlation
−1 0 1
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
Epigenomics Roadmap
FANTOM consortium
Other
ΔNR2F2
ΔGATA3
ΔTHAP1
ΔMEF2C
ΔFOSL2
ΔNFIC
ΔNFYB
ΔBCL3
ΔBACH1
ΔTHAP11
ΔKAT8
ΔDRAP1
ΔPITX1
ΔKLF8
ΔCREB5
ΔU2AF2
ΔZFP36
ΔDNase.1
ΔMYRF
ΔZNF48
ΔMAZ
ΔYY1
ΔNFAT5
ΔGSX1
ΔEGR2
ΔKAP1
ΔMYBL2
ΔE2F6
ΔCCNT2
ΔIKZF1
ΔFOS
ΔJUND
ΔFAM48A
ΔSP1. 1
ΔFOXK2
ΔHDAC2. 1
ΔHNF1A
ΔHOMEZ
ΔJUND
Δc. Jun
ΔJunD
ΔARID3A
ΔMBD4
ΔSFPQ
ΔPCBP2
ΔOTX1
ΔREST. 1
ΔHMX1
ΔPAX4
ΔTBX21
ΔZNF7
ΔNFE2L2
ΔCREB1
ΔH3K4me3
ΔEZH2
ΔSP2
ΔPHF8
ΔPOU4F1
ΔPKNOX1
ΔEZH2
ΔUSF2
ΔCAGE. minus
ΔMAZ
ΔKLF9
ΔTBX1
ΔRPC155
ΔHDAC1
ΔATF2
ΔNONO
ΔCEBPB. 1
ΔATF1
ΔHNRNPUL1
ΔTRIM22
ΔARNTL
ΔTGIF2LX
ΔSP3
ΔSPIC
ΔASCL2
ΔKLF13
ΔGLIS2
ΔTRIM22
ΔSPIC
ΔASCL2
ΔTGIF2LX
ΔHDAC1
ΔRPC155
ΔZBTB7A
ΔMAZ
ΔSIN3AK20
ΔHOMEZ
ΔATF1
HNRNPUL1
ΔHNF1A
ΔATF2
ΔNONO
ΔHDAC2. 1
ΔFOXK2
ΔFOSL2
ΔJUND
Δc. Jun
ΔJunD. 1
ΔCEBPB. 1
ΔCEBPB. 2
ΔMYBL2
Δp300.1
ΔARID3A
ΔSP1. 1
ΔTCF7L2
ΔMBD4
ΔTEAD4
ΔMEF2C
ΔTCF12
ΔTBX1
ΔE2F6
ΔREST. 1
ΔPCBP2
ΔSFPQ
ΔOTX1
ΔHMX1
ΔPOU4F1
ΔPAX4
ΔPKNOX1
ΔUSF2
ΔUSF1
ΔSP2
Δ CAGE minus.7
Δ CAGE minus.2
Δ CAGE minus.3
Δ CAGE minus.6
Δ CAGE minus.5
Δ CAGE minus.4
Δ CAGE minus.1
ΔMEF2B
ΔIRF4
ΔMAFK
ΔBATF3
ΔMAFK
ΔMEF2C
ΔBCL3
ΔEP300.2
ΔEP300.1
ΔSP1.1
ΔFAM48A
ΔE2F6
ΔHMGN3
ΔCCNT2
ΔHDAC2
ΔHNF4G
ΔTCF12
ΔMEF2A
ΔFOS
ΔJUND
ΔKAP1
ΔE2F6
ΔCHD2
ΔTHAP1
ΔGATA3
ΔNR2F2
ΔNFIC
ΔFOXA1
ΔFOXA2
ΔEZH2
ΔH3K4me3.1
ΔPHF8
ΔH3K4me3
ΔH3K27ac.3
ΔNRSF.1
ΔMAZ
ΔCTCF
ΔYY1
ΔCOREST
ΔSin3Ak.20
ΔMax.1
ΔPol2.4H8
ΔPol2
ΔTAF1
ΔRXRA.3
ΔCREB1
ΔNFE2L2
ΔZNF7
ΔJUND
ΔMER3
ΔZGPAT
ΔZNF511
ΔZNF143
ΔDRAP1
ΔKAT8
ΔKLF16
ΔKLF12
ΔKMT2B
ΔKAT2A
ΔGATAD1
ΔDNase
ΔDNase
ΔDNase.1
ΔNFE2
ΔZNF580
ΔIKZF5
ΔATF3
ΔZFP36
ΔU2AF2
ΔKLF8
ΔCREB5
ΔPITX1
ΔPITX3
ΔMYRF
ΔMXD3
ΔZNF48
ΔMAZ
ΔTHAP1
ΔTBX3
ΔMGA
ΔYY1
ΔPOU2F1
ΔNFAT5
ΔDUXA
ΔPOU2F2
ΔMSX2
ΔLMX1B
ΔGSX1
ΔESX1
ΔGC_content
ΔZFP3
ΔZFP2
ΔZFP36
ΔPOU3F4
ΔPOU3F1
ΔKMT2C
ΔFOXC2
ΔHNF1A
ΔHNF1B
ΔONECUT3
ΔALX1
ΔARX
ΔHOXA11
ΔLMX1A
ΔBARX1
ΔDLX1
ΔLHX6
ΔMEOX2
ΔPHOX2B
ΔNKX6.1
ΔPHOX2A
ΔGBX1
ΔALX4

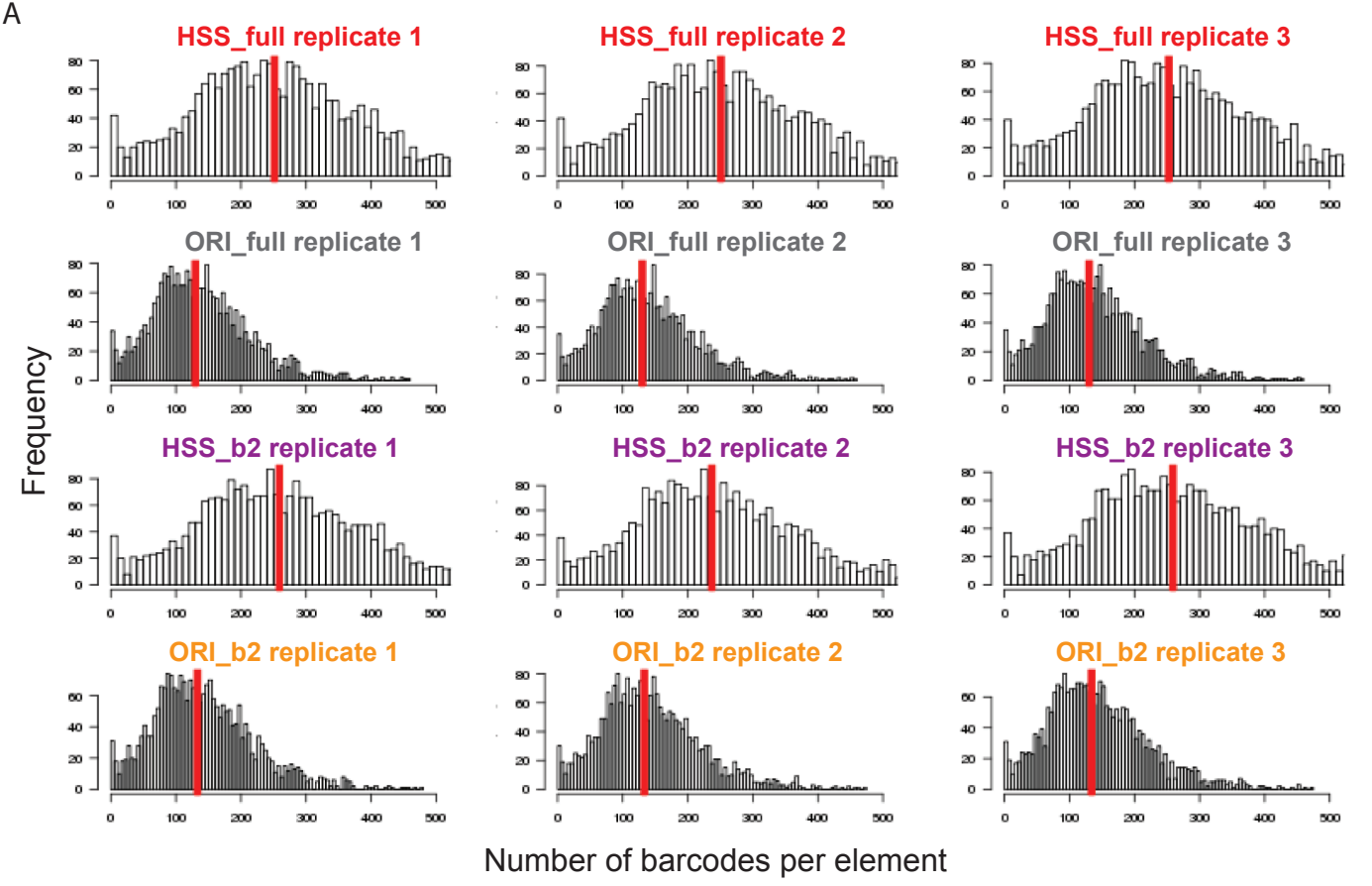
Supplemental Figure 16. Coefficients and performance of the models to predict differential element efficacy among the three size classes.

A) Scatter plots displaying the relationship between the 10-fold cross-validated predictions derived from lasso regression models and the observed RNA/DNA ratios, for each of differential size comparisons displayed. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values.

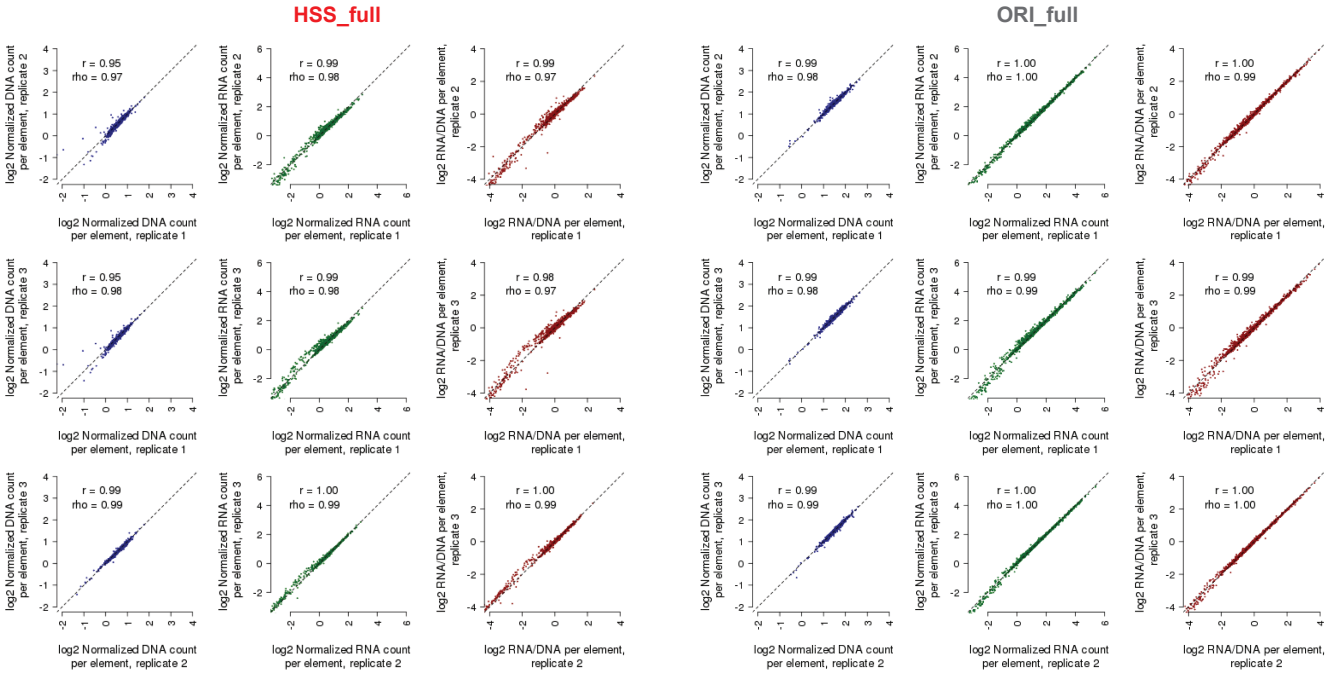
B) The top 30 coefficients derived from lasso regression models trained on the full dataset to predict observed values from the differential size comparisons indicated. Features with the extension ".1", ".2", etc allude to redundant features or replicate samples.

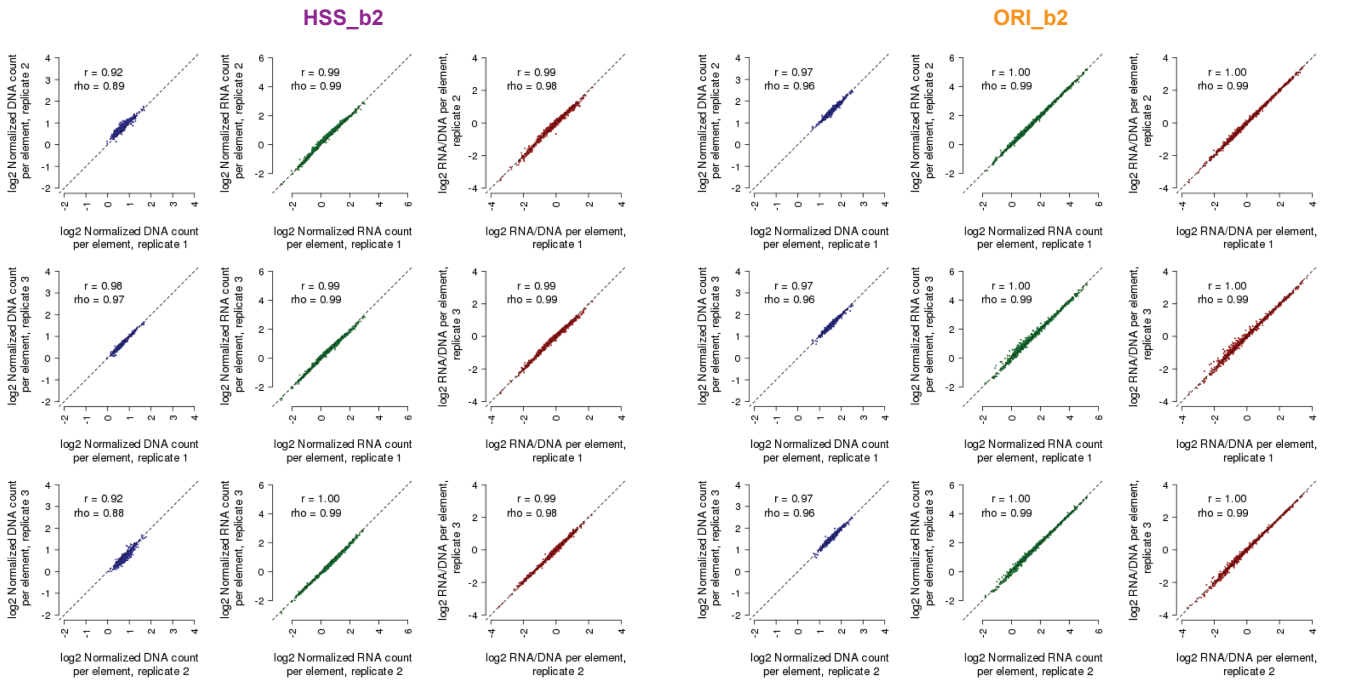
C) Figure has been rotated to fit to sheet. Pearson correlation matrix between the union of all top 30 features from (B), shown as rows, and other features sharing a Pearson correlation either ≤ –0.8 or ≥ 0.8, shown as columns. Feature names are colored according to the origin of the feature as shown in the boxed key above. Hierarchical clustering was used to group features exhibiting similar correlation patterns.

A

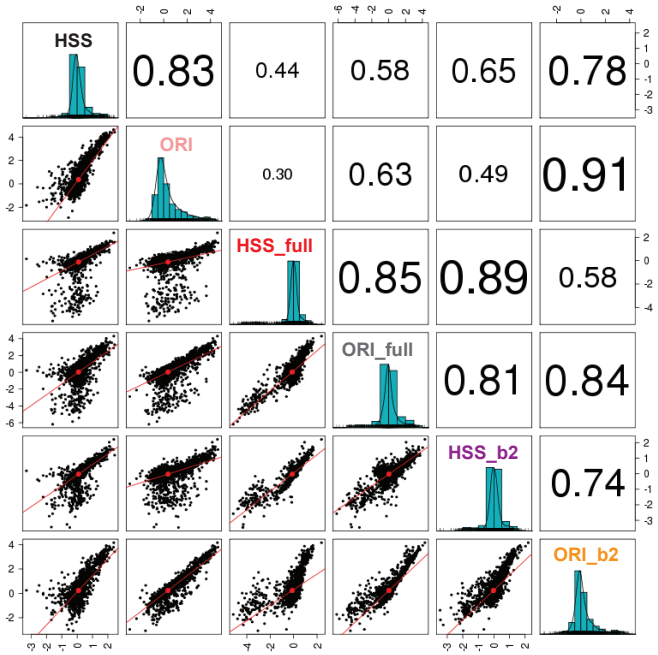

HSS_full replicate 1
HSS_full replicate 2
HSS_full replicate 3
ORI_full replicate 1
ORI_full replicate 2
ORI_full replicate 3
HSS_b2 replicate 1
HSS_b2 replicate 2
HSS_b2 replicate 3
ORI_b2 replicate 1
ORI_b2 replicate 2
ORI_b2 replicate 3
Frequency
Number of barcodes per element

B

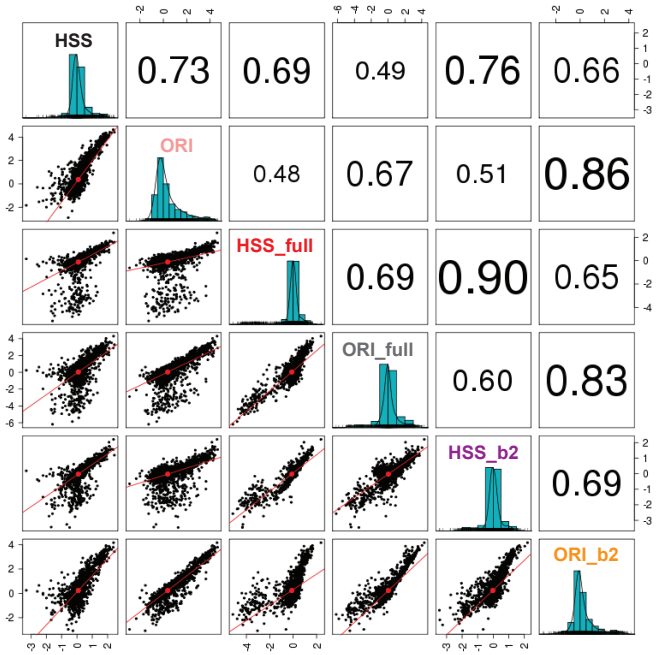

HSS_full
ORI_full
log2 Normalized DNA count per element, replicate 1
log2 Normalized RNA count per element, replicate 1
log2 RNA/DNA per element, replicate 1
log2 Normalized DNA count per element, replicate 2
log2 Normalized RNA count per element, replicate 2
log2 RNA/DNA per element, replicate 2
log2 Normalized DNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 3
log2 RNA/DNA per element, replicate 3

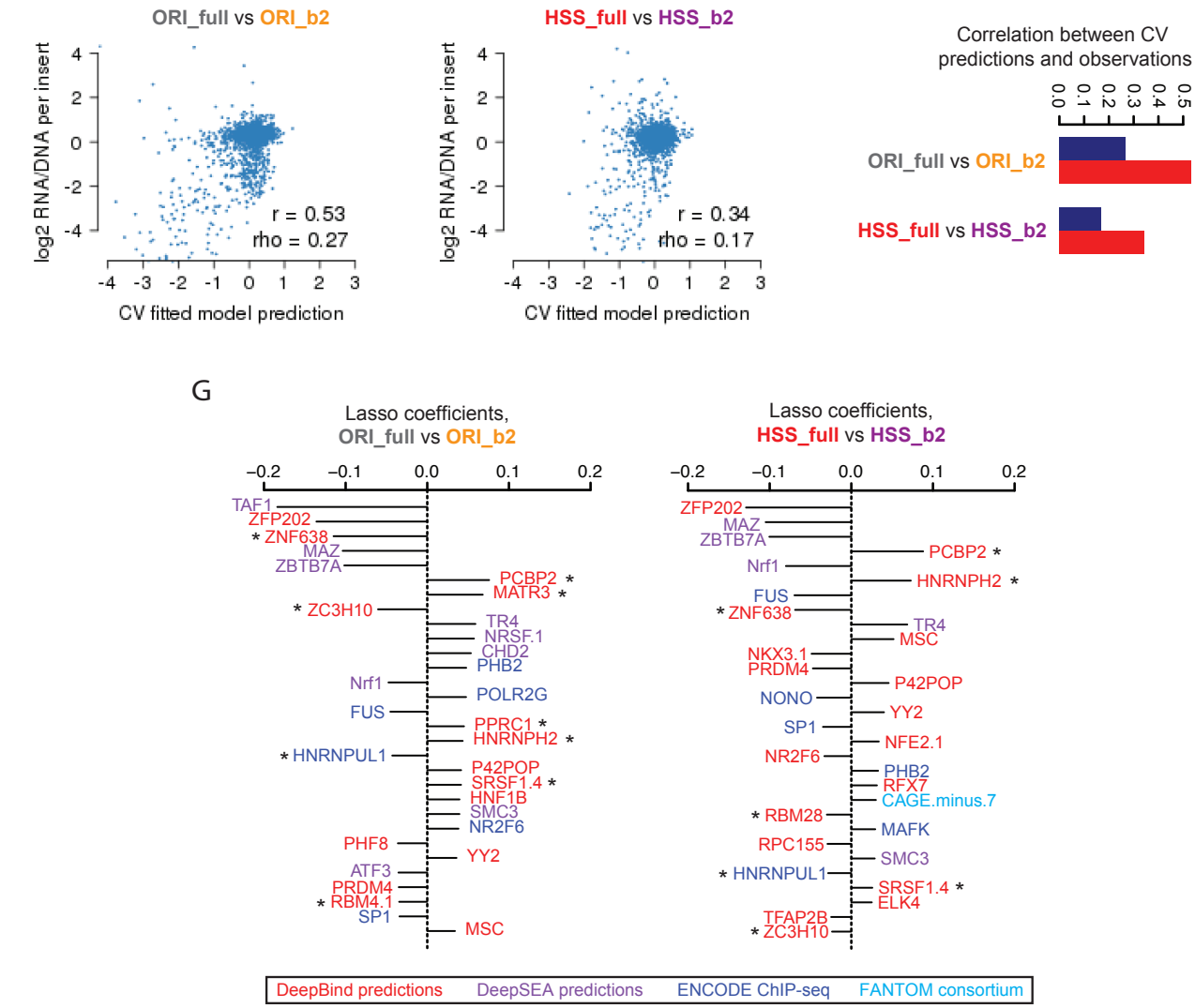

HSS_b2
ORI_b2
r = 0.92
rho = 0.89
r = 0.99
rho = 0.99
r = 0.99
rho = 0.98
r = 0.97
rho = 0.96
r = 1.00
rho = 0.99
r = 1.00
rho = 0.99
r = 0.98
rho = 0.97
r = 0.99
rho = 0.99
r = 0.99
rho = 0.99
r = 0.97
rho = 0.96
r = 1.00
rho = 0.99
r = 1.00
rho = 0.99
r = 0.92
rho = 0.88
r = 1.00
rho = 0.99
r = 0.99
rho = 0.98
r = 0.97
rho = 0.96
r = 1.00
rho = 0.99
r = 1.00
rho = 0.99
log2 Normalized DNA count per element, replicate 1
log2 Normalized RNA count per element, replicate 1
log2 RNA/DNA per element, replicate 1
log2 Normalized DNA count per element, replicate 2
log2 Normalized RNA count per element, replicate 2
log2 RNA/DNA per element, replicate 2
log2 Normalized DNA count per element, replicate 3
log2 Normalized RNA count per element, replicate 3
log2 RNA/DNA per element, replicate 3

C

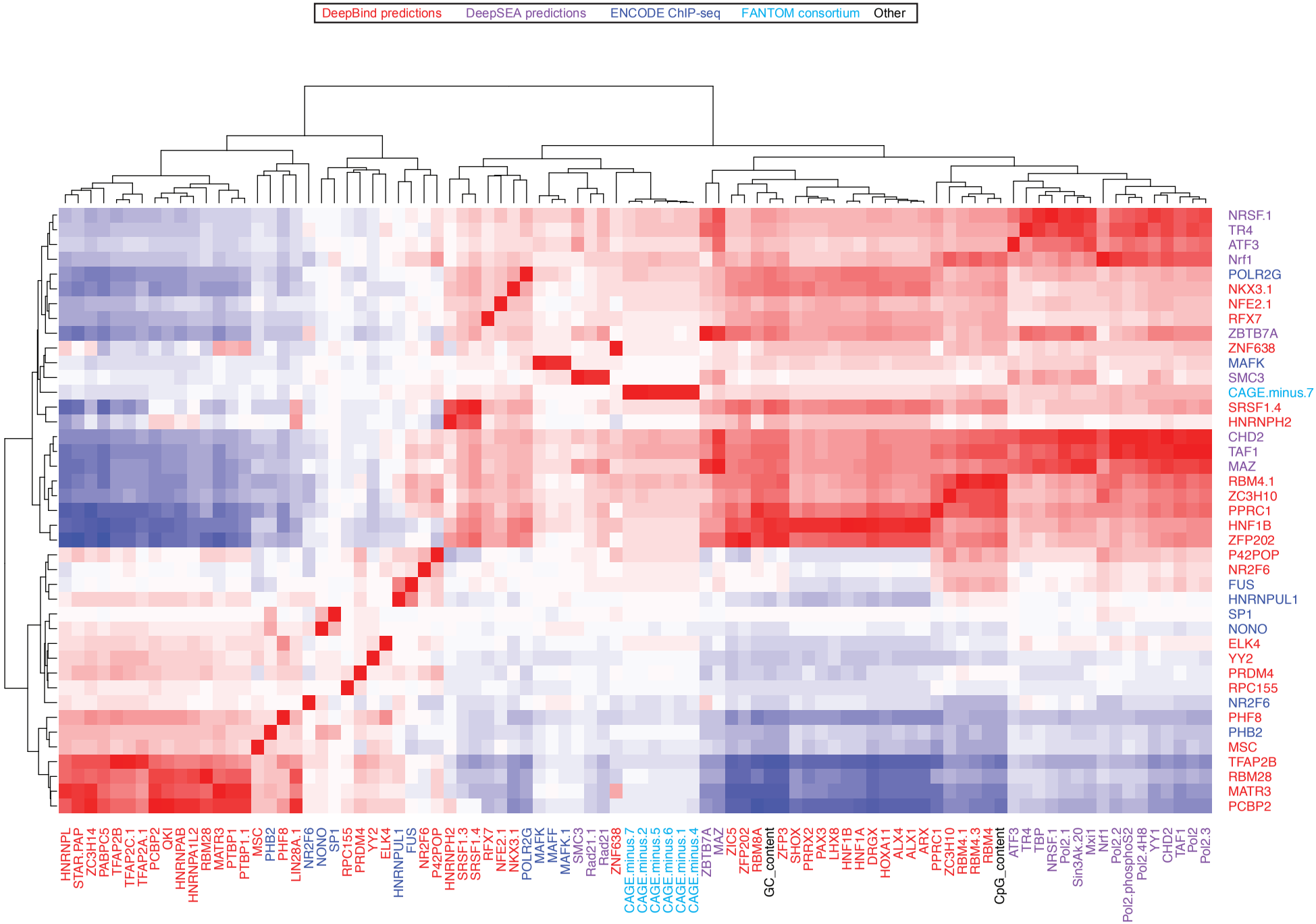

HSS
0.83
0.44
0.58
0.65
0.78
ORI
0.30
0.63
0.49
0.91
HSS_full
0.85
0.89
0.58
ORI_full
0.81
0.84
HSS_b2
0.74
ORI_b2

D


HSS
0.73
0.69
0.49
0.76
0.66
ORI
0.48
0.67
0.51
0.86
HSS_full
0.69
0.90
0.65
ORI_full
0.60
0.83
HSS_b2
0.69
ORI_b2


E
ORI_full vs ORI_b2
log2 RNA/DNA per insert
CV fitted model prediction
r = 0.53
rho = 0.27
HSS_full vs HSS_b2
r = 0.34
rho = 0.17
F
Correlation between CV predictions and observations
ORI_full vs ORI_b2
HSS_full vs HSS_b2
G
Lasso coefficients, ORI_full vs ORI_b2
TAF1
ZFP202
* ZNF638
MAZ
ZBTB7A
PCBP2 *
MATR3 *
* ZC3H10
TR4
NRSF.1
CHD2
PHB2
Nrf1
POLR2G
FUS
PPRC1 *
HNRNPH2 *
* HNRNPUL1
P42POP
SRSF1.4 *
HNF1B
SMC3
NR2F6
PHF8
YY2
ATF3
PRDM4
* RBM4.1
SP1
MSC
Lasso coefficients, HSS_full vs HSS_b2
ZFP202
MAZ
ZBTB7A
Nrf1
FUS
* ZNF638
NKX3.1
PRDM4
NONO
SP1
NR2F6
* RBM28
RPC155
* HNRNPUL1
TFAP2B
* ZC3H10
PCBP2 *
HNRNPH2 *
TR4
MSC
P42POP
YY2
NFE2.1
PHB2
RFX7
CAGE.minus.7
MAFK
SMC3
SRSF1.4 *
ELK4
DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
FANTOM consortium

H


DeepBind predictions
DeepSEA predictions
ENCODE ChIP-seq
FANTOM consortium
Other
NRSF.1
TR4
ATF3
Nrf1
POLR2G
NKX3.1
NFE2.1
RFX7
ZBTB7A
ZNF638
MAFK
SMC3
CAGE.minus.7
SRSF1.4
HNRNPH2
CHD2
TAF1
MAZ
RBM4.1
ZC3H10
PPRC1
HNF1B
ZFP202
P42POP
NR2F6
FUS
HNRNPUL1
SP1
NONO
ELK4
YY2
PRDM4
RPC155
NR2F6
PHF8
PHB2
MSC
TFAP2B
RBM28
MATR3
PCBP2
HNRNPL
STAR.PAP
ZC3H14
PABPC5
TFAP2B
TFAP2C.1
TFAP2A.1
PCBP2
QKI
HNRNPAB
HNRNPA1L2
RBM28
MATR3
PTBP1
PTBP1.1
MSC
PHB2
PHF8
LIN28A.1
NR2F6
NONO
SP1
RPC155
PRDM4
YY2
ELK4
HNRNPUL1
FUS
NR2F6
P42POP
HNRNPH2
SRSF1.3
SRSF1.4
RFX7
NFE2.1
NKX3.1
POLR2G
MAFK
MAFF
MAFK.1
SMC3
Rad21.1
Rad21
ZNF638
CAGE.minus.7
CAGE.minus.2
CAGE.minus.5
CAGE.minus.6
CAGE.minus.1
CAGE.minus.4
ZBTB7A
MAZ
ZIC5
ZFP202
RBM8A
GC_content
ZFP3
SHOX
PRRX2
PAX3
LHX8
HNF1B
HNF1A
DRGX
HOXA11
ALX4
ALX1
ARX
PPRC1
ZC3H10
RBM4.1
RBM4.3
RBM4
CpG_content
ATF3
TR4
TBP
NRSF.1
Pol2.1
Sin3Ak.20
Mxi1
Nrf1
Pol2.2
Pol2.phosphoS2
Pol2.4H8
YY1
CHD2
TAF1
Pol2
Pol2.3

Supplemental Figure 17. Evaluation of potential confounding artifact of internal transcriptional initiation in STARR-seq assays.

A) Shown are histograms indicating the number of observed barcodes per element, for each of the 2,440 elements tested in each of the 3 replicates for each of the 4 MPRA methods indicated. HSS_full and ORI_full allude to assays in which full length mRNA (i.e., not arising from internal transcriptional initiation within the regulatory element tested) was measured. HSS_b2 and ORI_b2 (b2 for "batch 2") allude to assays that were performed identically to HSS and ORI (as shown in **Figure 1**) but serve as an internal control for HSS_full and ORI_full, respectively, to mitigate batch effects. Shown with a vertical red line is the median number of barcodes per element.

B) 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 each of the 4 additional MPRA methods tested. Also indicated is the Pearson (*r*) and Spearman (rho) correlation values. Candidate enhancers supported by fewer than 10 barcodes were filtered out prior to this analysis to reduce the impact of technical noise.

C-D) Scatter matrix displaying scatter plots corresponding to each of the 15 pairs of possible inter-assay comparisons (lower diagonal elements). Shown on the diagonal is a histogram of the log2(RNA/DNA) ratios, averaged among replicate samples, for the original HSS and ORI (as shown in **Figure 2B**) as well as HSS_full, ORI_full, HSS_b2, and ORI_b2. Also shown are Pearson (C) and Spearman (D) correlation values among each pair of comparisons, with the size of the text proportional to the magnitude of the correlation coefficient (upper diagonal elements).

E) Scatter plots displaying the relationship between the 10-fold cross-validated predictions derived from lasso regression models and the observed RNA/DNA ratios, for each of differential comparisons displayed. Also indicated are the Pearson (*r*) and Spearman (rho) correlation values.

F) Summary of these r and rho values for each of the comparisons shown in (E).

G) The top 30 coefficients derived from lasso regression models trained on the full dataset to predict observed values from the differential comparisons indicated. Features with the extension ".1", ".2", etc allude to redundant features or replicate samples, and asterisks (*) allude to annotated RNA-binding proteins.

H) Figure has been rotated to fit to sheet. Pearson correlation matrix between the union of all top 30 features from (G), 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 above. Hierarchical clustering was used to group features exhibiting similar correlation patterns.