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

Jason Klein
memorial sloan kettering cancer center https://orcid.org/0000-0001-9566-6347
Vikram Agarwal
Calico Life Sciences LLC https://orcid.org/0000-0001-8148-952X
Fumitaka Inoue
UCSF https://orcid.org/0000-0003-0657-434X
Aidan Keith
University of Washington
Beth Martin
University of Washington
Martin Kircher
Berlin Institute of Health https://orcid.org/0000-0001-9278-5471
Nadav Ahituv
UCSF https://orcid.org/0000-0002-7434-8144
Jay Shendure (shendure@uw.edu)
University of Washington https://orcid.org/0000-0002-1516-1865

Method Article

Keywords: MPRA, enhancers, promoters, gene expression, gene regulation, reporter assay

Posted Date: October 12th, 2020

DOI: https://doi.org/10.21203/rs.3.pex-1065/v1

Abstract

Massively parallel reporter assays (MPRAs) functionally screen thousands of sequences for regulatory activity in parallel. Here, we further develop and apply a novel method to assemble and functionally test libraries of greater than 2000 of the same putative enhancers as 192-mers, 354-mers, and 678-mers. We achieved a yield of 95% for 354-mers and 84% for 678-mers. Importantly, we observe surprisingly large differences in functional activity. 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.

Introduction

The spatiotemporal control of gene expression is orchestrated in part by distally located DNA sequences known as enhancers. The first viral and cellular enhancers were identified by cloning fragments of DNA into a plasmid with a reporter gene and promoter 1–4 . Enhancement of transcription in such a reporter assay is a widely used approach 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 .

Massively parallel reporter assays (MPRAs) modify the in vitro reporter assays described above to facilitate simultaneous testing of thousands of putative regulatory elements or variants thereof 9–11 in a single experiment. Instead of relying on measurement of a conventional reporter, MPRAs characterize each element in a multiplex fashion, through sequence-based quantification of barcodes incorporated into the RNA, each associated with a different element 9–15 . MPRAs (a term we use broadly to encompass related methods including STARR-seq 16 and lentiMPRAs 17) have facilitated the scalable study of putative regulatory elements for goals ranging from functional annotation 16–18 to variant effect prediction 10–15,19 to evolutionary reconstruction 20,21 .

To date, several groups have implemented enhancer-focused MPRAs, but with diverse designs.

Some of the major differences include whether the enhancer is upstream 10,11 vs. within the 3' UTR of the reporter gene 16 , and whether the construct remains episomal vs. integrated 17 .

Additionally, most MPRAs test sequences cloned in one of two possible orientations, effectively assuming that enhancer activity is independent of orientation. Finally, while larger sheared genomic DNA fragments 16,22 , PCR amplicons 12 or captured sequences 23,24 have been used in MPRAs, most studies using MPRAs synthesize libraries of candidate enhancers on microarrays, and are therefore limited to testing shorter sequences (typically less than 200 bp).

Unfortunately, we have, as a field to date, largely failed to evaluate how these design choices impact or bias the results of MPRAs. 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.

Particularly as efforts to validate the >1 million putative human enhancers 5–8 , as well as the

growing number of disease-associated noncoding variants, begin to scale, a clear-eyed understanding of the biases and tradeoffs introduced by various MPRA experimental design choices is needed. To this end, we performed a systematic comparison, testing the same 2,440 sequences for regulatory activity using nine different MPRA strategies, including conventional episomal, STARR-seq, and lentiviral designs. Second, we tested the same sequences in both orientations relative to the promoter. Finally, we further developed our multiplex pairwise assembly protocol 27, and applied it to test short (192 bp), medium (354 bp), and long (678 bp) versions of the same enhancers. Our results quantify the impact of MPRA experimental design choices and also provide further insight into the nature of enhancers.

Reagents

KAPA HiFi 2x Readymix (Kapa Biosystems)

KAPA HiFi HotStart Uracil+ ReadyMix PCR Kit (Kapa Biosystems)

KAPA2G Robust HostStart ReadyMix (Kapa Biosystems)

Sybr Green (Thermo fisher scientific)

AMPure XP (Beckman coulter)

Qiagen Elution Buffer (EB)

QIAquick Gel Extraction Kit (Qiagen)

NextSeq Mid 300 cycle kit (Illumina)

Qubit DNA High Sensitivity Assay kit (Thermo Fisher Scientific)

300 Cycle NextSeq v2 High-Output kit (Illumina)

USER enzyme (NEB)

NEBNext End Repair Module (NEB)

DNA Clean and Concentrator 5 (Zymo Research)

Equipment

MiniOpticon Real-Time PCR system (Bio-Rad)

PacBioSequel System

Illumina Miseq

Illumina Nextseq

Procedure

1. All libraries were amplified off the array using the corresponding group specific primers in Supplemental Table 1 with KAPA HiFi HotStart Uracil+ ReadyMix PCR Kit (Kapa Biosystems).

2. During the first round of assembly, fragments A and B were assembled with HSSF-ATGC (5'-TCTAGAGCATGCACCGGATGC-3') and DO_31R_PU (5'-CCA/ideoxyU/GACCCT/ideoxyU/ACTGGG/ideoxyU//3deoxyU/-3') and fragments C and D were assembled with DO_8F_PU (5'-GCGACG/ideoxyU/CATGC/ideoxyU/GTTG/ideoxyU//3deoxyU/-3') and HSS_R (5'-CTTATCATGTCTGCTCGAAGC-3').

3. Assembled libraries were then purified with a 0.65x Ampure cleanup following the manufacturer's protocol, and eluted in 20 µl.

4. 2 µl of USER enzyme (NEB) was added to the purified assembly reactions and incubated at 37 °C for 15 minutes followed by 15 minutes 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.

5. All libraries were then quantified using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific) and eluted to 0.75 ng/ul.

6. Assemblies AB and CD were then assembled together following the multiplex pairwise assembly protocol[27].

7. After the second assembly, libraries were purified using a 0.6x AMPure cleanup and eluted in 30 µL EB.

8. We then amplified 1 uL of each assembly with HSS-F-ATGC-pu1F (5'-ACTTTATCAATCTCGCTCCAAACCTCTAGAGCATGCACCGGATGC-3') and HSS-R-clon-pu1R (5'-ACTTTATCAATCTCGCTCCAAACCCCATCATTCTGACCGGC-3') to add flow cell adapters and indexes.

9. 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.

Troubleshooting

Time Taken

6 hours

Anticipated Results

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. These 678 bp fragments had an 84% yield (1,887/2,236) and 27.9-fold IQR. We verified a subset of our long enhancers with PacBio sequencing (chimera rate of 16.5%).

References

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-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. The 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. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

18. 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).

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. *bioRxiv* 283168 (2018) doi:10.1101/283168.

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. 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).

26. 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).

27. Klein, J. C. *et al.* Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2016).

Acknowledgements

We thank Seungsoo 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.), and 1R21HG010683 (N.A.); National Institute of Mental Health grants 1R01MH109907 (N.A.) and 1U01MH116438 (N.A.), NRSA NIH fellowship 5T32HL007093 (V.A.) and the Uehara Memorial Foundation (F.I.). J.S. is an investigator of the Howard Hughes Medical Institute.

Figures

Figure S10

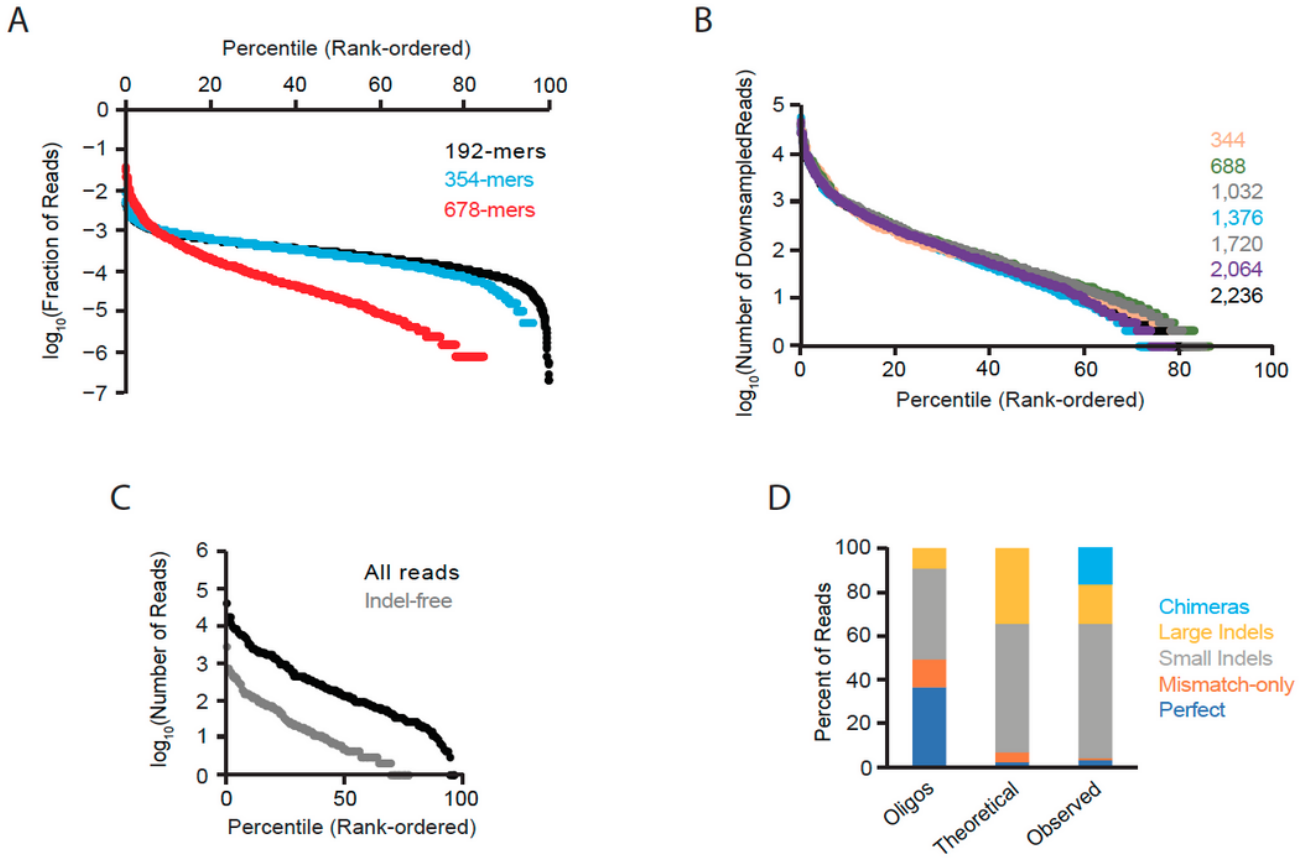

Figure 1

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 678ners). B) Uniformity for various HMPA reactions, with total number of target sequences ranging from 344 to 2,236. The sequencing reads were downsampled to normalize for the total number of targets (#Reads=581*#Targets). The Y axis is the number of downsampled 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.

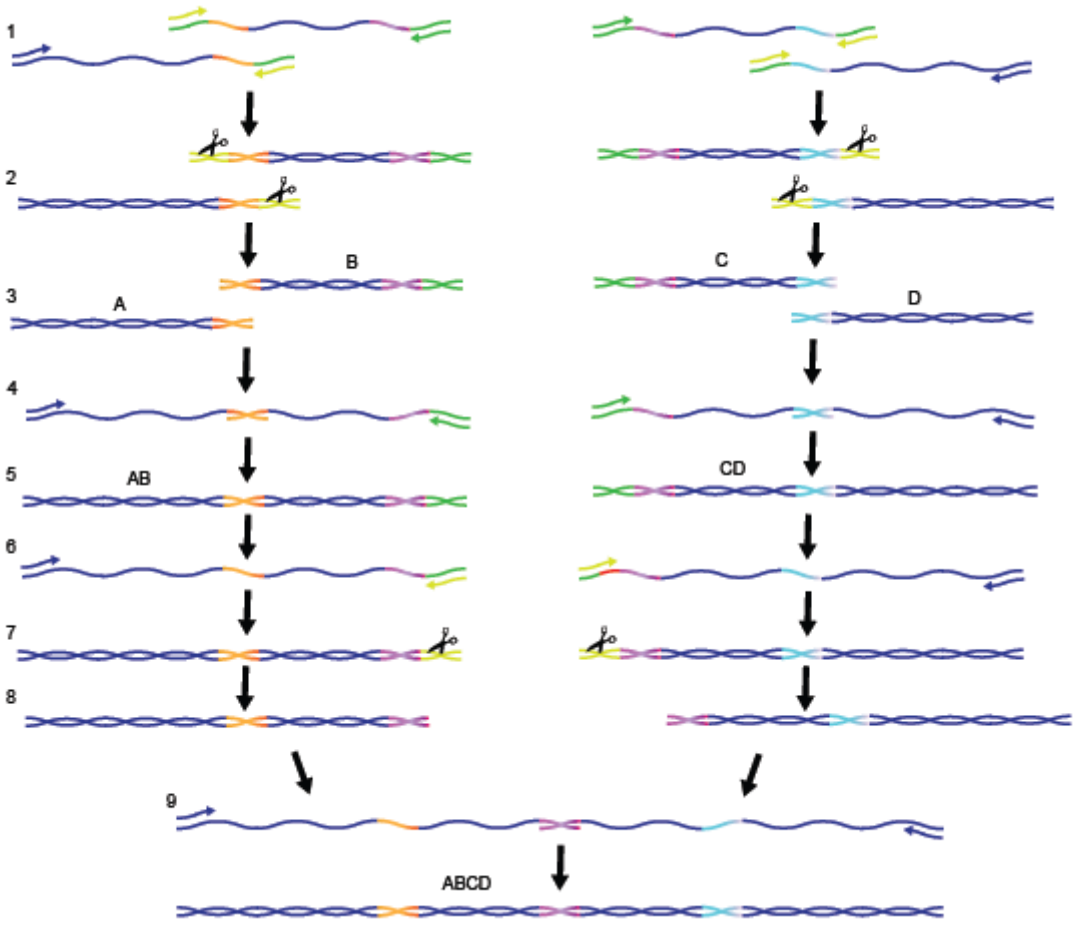

Figure 2

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.

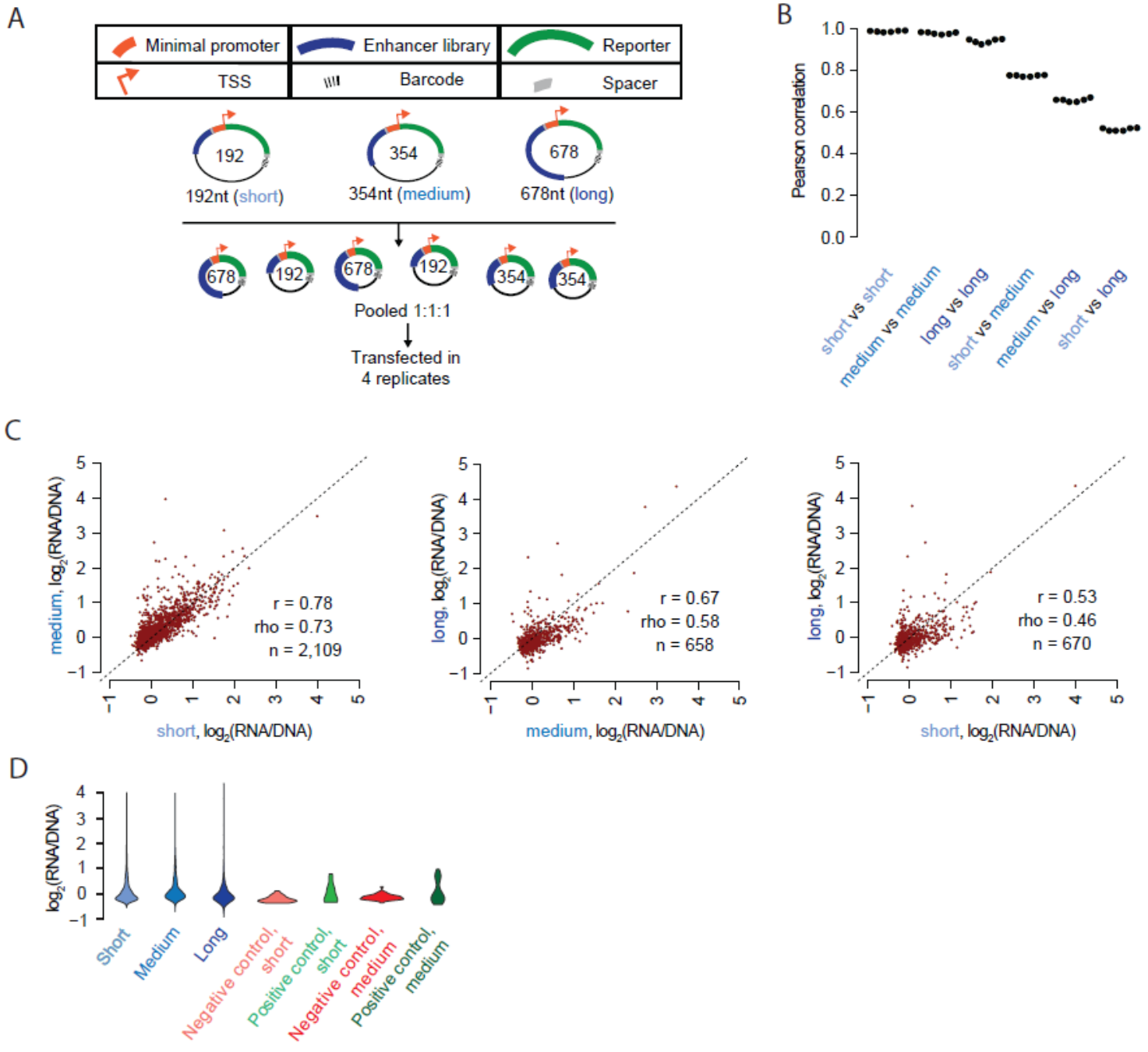

Figure 3

Including additional sequence context around tested elements leads to differences in the results of MPRAs. A) Experimental schematic. 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. 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 vs. medium, medium vs. long, and short vs. long versions of each element. D) Violin plot displaying the distribution of average log2(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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supptable.xlsx