Title

lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements

Permalink

https://escholarship.org/uc/item/60h7c08s

Journal

Nature Protocols, 15(8)

ISSN

1754-2189

Authors

Gordon, M Grace
Inoue, Fumitaka
Martin, Beth
et al.

Publication Date

2020-08-01

DOI

10.1038/s41596-020-0333-5

Peer reviewed

nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-020-0333-5

Check for updates

lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements

M. Grace Gordon[1,2,3,16], Fumitaka Inoue[1,2,16] ✉, Beth Martin[4,16], Max Schubach[5,6,16], Vikram Agarwal[4,7], Sean Whalen[8], Shiyun Feng[1,2], Jingjing Zhao[1,2], Tal Ashuach[9], Ryan Ziffra[1,2], Anat Kreimer[1,2,9], Ilias Georgakopoulous-Soares[1,2], Nir Yosef[9,10], Chun Jimmie Ye[1,2,10,11,12], Katherine S. Pollard[2,8,10,13], Jay Shendure[4,14,15] ✉, Martin Kircher[5,6] ✉ and Nadav Ahituv[1,2] ✉

Massively parallel reporter assays (MPRAs) can simultaneously measure the function of thousands of candidate regulatory sequences (CRSs) in a quantitative manner. In this method, CRSs are cloned upstream of a minimal promoter and reporter gene, alongside a unique barcode, and introduced into cells. If the CRS is a functional regulatory element, it will lead to the transcription of the barcode sequence, which is measured via RNA sequencing and normalized for cellular integration via DNA sequencing of the barcode. This technology has been used to test thousands of sequences and their variants for regulatory activity, to decipher the regulatory code and its evolution, and to develop genetic switches. Lentivirus-based MPRA (lentiMPRA) produces 'in-genome' readouts and enables the use of this technique in hard-to-transfect cells. Here, we provide a detailed protocol for lentiMPRA, along with a user-friendly Nextflow-based computational pipeline—MPRAflow—for quantifying CRS activity from different MPRA designs. The lentiMPRA protocol takes ~2 months, which includes sequencing turnaround time and data processing with MPRAflow.

Introduction

Gene regulatory elements control a gene's transcription. These include sequences that activate transcription, such as promoters and enhancers; silencers that repress a gene; and insulators that restrict genes from interacting with certain regulatory elements. Nucleotide variation in these elements can have a major effect on phenotype. Mutations within them have been shown to be a major cause of human disease[1]. For example, >90% of all human disease genome-wide association studies (GWASs) have shown associations with noncoding variants[2], which colocalize with potential gene regulatory elements[3]. In addition, gene regulatory elements can be major drivers of evolutionary speciation, driving differences between species such as morphology, diet, and behavior[4]. These sequences can also be used as genetic switches to tune transgenes to specific levels in certain cell types or tissues.

In this protocol, we focus on gene activation associated regulatory elements, promoters and enhancers. These sequences can be identified in a genome-wide manner by biochemical methods such as chromatin immunoprecipitation followed by sequencing (ChIP-seq[5]), DNase I hypersensitive sites sequencing (DNase-seq[6,7]), assay for transposase-accessible chromatin using sequencing (ATAC-seq[8]), cleavage under targets and release using nuclease (CUT&RUN[9]), Hi-C[10] and others. However, these methods only help annotate CRSs, and additional experimental assays must be performed in

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. [2]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [3]Biological and Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, CA, USA. [4]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [5]Berlin Institute of Health (BIH), Berlin, Germany. [6]Charité–Universitätsmedizin Berlin, Berlin, Germany. [7]Calico Life Sciences LLC, South San Francisco, CA, USA. [8]Gladstone Institutes, San Francisco, CA, USA. [9]Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. [10]Chan-Zuckerberg Biohub, San Francisco, CA, USA. [11]Division of Rheumatology, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. [12]Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA. [13]Department of Epidemiology and Biostatistics and Institute of Computational Health Sciences, University of California, San Francisco, San Francisco, CA, USA. [14]Howard Hughes Medical Institute, Seattle, WA, USA. [15]Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA, USA. [16]These authors contributed equally: M. Grace Gordon, Fumitaka Inoue, Beth Martin and Max Schubach. ✉e-mail: fumitaka.inoue@ucsf.edu; shendure@uw.edu; martin.kircher@bihealth.de; nadav.ahituv@ucsf.edu

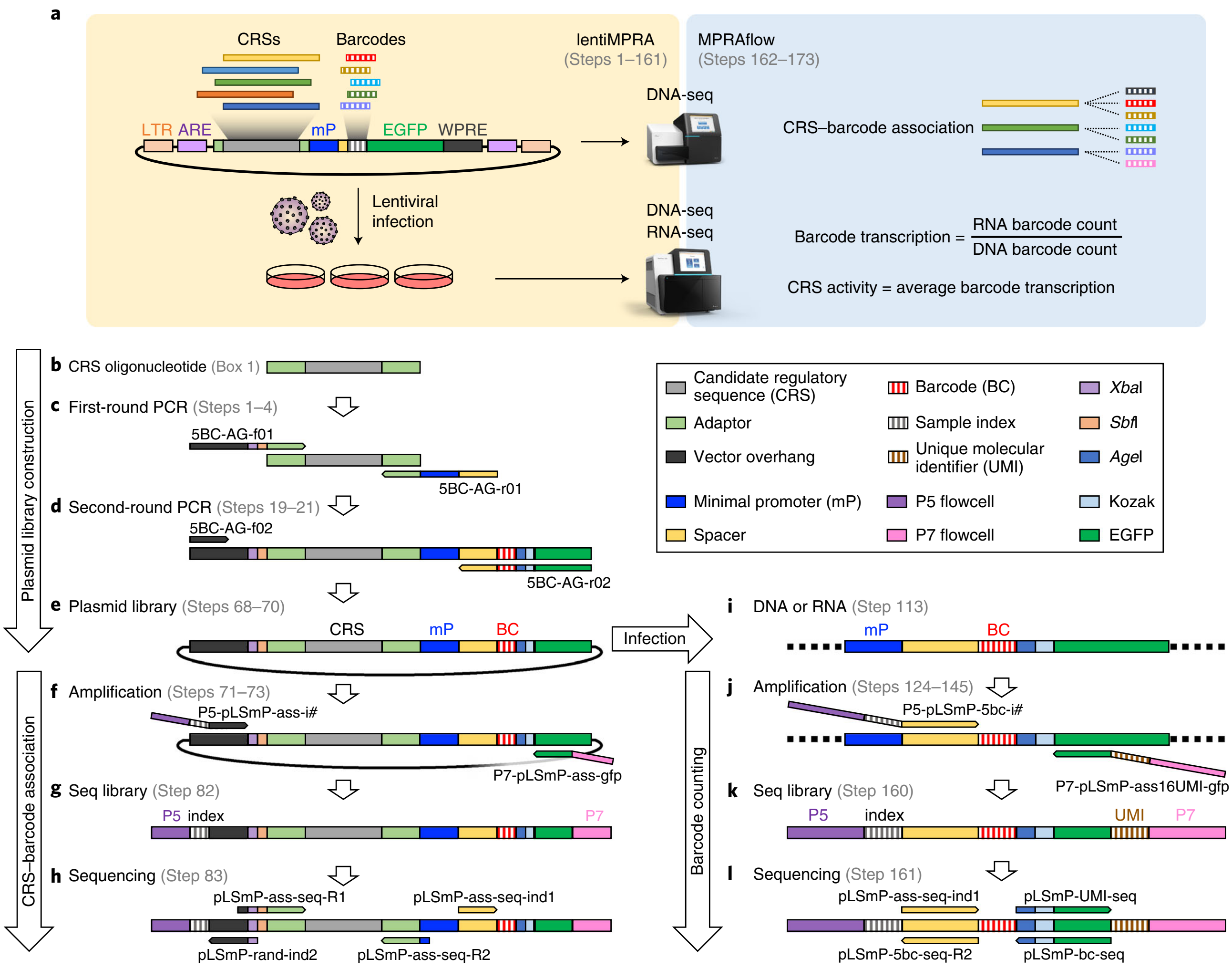


Fig. 1 | Schematics of lentiMPRA. a, Summary of lentiMPRA and MPRAflow. The lentiMPRA library is sequenced to associate CRSs and barcodes and to infect cells, using three replicates. DNA and RNA from the cells are sequenced to determine barcode transcription and CRS activity. **b**, CRS oligonucleotide. A 200-base CRS (gray) is flanked by PCR adaptor sequences (light green). **c**, First-round PCR. PCR primers add sequences that are complementary to the vector (black) to the upstream side, as well as minimal promoter (mP, blue) and spacer sequences (yellow) downstream of the CRS oligonucleotide. **d**, Second-round PCR. Reverse primer adds the barcodes (red-striped section) and GFP complementary sequences (green). **e**, Plasmid construct. **f**, Amplification for CRS–barcode association. Primers add P5 (purple) and sample index (gray-striped section) upstream and P7 (pink) downstream. **g**, Sequencing library structure. **h**, Sequencing reaction. Paired-end reads specify the CRS sequence, with index read 1 providing the barcode and index read 2 reading the sample index for multiplexing. **i**, Integrated DNA and expressed RNA in infected cells. **j**, Amplification for barcode counting. Primers add P5 and sample index upstream and P7 and UMI, brown stripe) downstream. **k**, Sequencing library structure. **l**, Sequencing reaction. Paired-end reads give barcode, index read 1 gives UMI, and index read 2 provides sample index for multiplexing. ARE, anti-repressor element; LTR, long terminal repeat; WPRE, Woodchuck hepatitis virus posttranscriptional regulatory element.

order to validate their predicted activity. Reporter assays are commonly used to characterize CRSs. In this assay, the CRS is placed either upstream of a reporter gene (i.e., in the case of testing promoters) or upstream of a minimal promoter followed by a reporter gene (i.e., in the case of testing enhancers). If the sequence is an activating regulatory element, it will turn on the reporter gene, providing a measurable output. However, these assays are primarily done on an individual basis and as such cannot assess the thousands of CRSs and their variants that have been identified via the aforementioned biochemical assays. Massively parallel reporter assays overcome this hurdle, providing the ability to test hundreds of thousands of sequences and their variants in parallel for their regulatory function[11]. This is done either by measuring RNA expression driven by the CRS by pairing it to a transcribed barcode, or by using the CRS itself as a barcode, as is done in the self-transcribing active regulatory region sequencing (STARR-seq) assay[12].

Here, we describe both a lentivirus-based MPRA (lentiMPRA) and MPRAflow, a computational tool for MPRA analysis that is based on the Nextflow framework[13] (Fig. 1a). lentiMPRA can be used in any cell type that can be efficiently infected via lentivirus, providing the ability to carry out MPRA in a broad range of cell types and tissues. In addition, owing to the viruses' inherent genomic integration, it provides an 'in-genome' readout, which we have shown provides more robust results that can be better predicted by both biochemical and sequence-based features as compared with episomal-based MPRA[14]. MPRAflow is a user-friendly computational pipeline that is compatible with a broad range of MPRA experiments.

Development of the protocol

We developed lentiMPRA to overcome the following limitations: (i) descriptive assays that detect potential regulatory elements (such as ChIP-seq, DNase-seq, ATAC-seq, CUT&RUN and Hi-C) identify candidate sequences within chromatin, yet most MPRAs analyze sequences in an episomal context; (ii) episomal-based MPRA is limited to cells that can be easily transfected. Lentivirus-based assays overcome both these limitations. Lentiviruses integrate into the genome, providing an in-genome readout. In addition, they can infect a large number of cells and tissue types, providing a more diverse range of cellular environments for MPRA. In this protocol, we further develop lentiMPRA by placing a barcode in the 5′ UTR of the reporter gene. This 5′ UTR barcoding method uses a shorter distance between the CRS and the barcode (102 bp) than previous 3′ UTR barcoding methods (801 bp), reducing the risk of CRS–barcode swapping[15]. In addition, unlike previous lentiMPRA, in which each CRS is synthesized together with multiple barcodes in a custom array, the 5′ UTR barcoding strategy adds barcodes via the PCR primer. This enables the ability to clone and test hundreds of thousands of CRSs using lentiMPRA.

To subsequently analyze MPRA results, several home-brewed MPRA computational analysis pipelines exist that are tailored to a specific lab and MPRA technique. However, these tools are not transferable between labs because of the large variability in MPRA designs, lack of documentation, complicated input files and the lack of parameterization of these tools. We thus developed MPRAflow, which provides a user-friendly, flexible, parallelized tool for quantifying CRS activity from a variety of MPRA experimental designs, including lentiMPRA, episomal-based MPRA and saturation mutagenesis designs, with easily interpretable visualizations that can be readily adopted by users regardless of their computational level. In addition to providing normalized fold change per CRS, MPRAflow can generate input files for MPRAnalyze[16], a tool that calculates a transcription rate for each tested CRS by fitting a generalized linear model with DNA and RNA counts. This pipeline enables the entire analysis to be completed with two commands on a terminal, greatly simplifying the computational tasks associated with MPRAs and therefore increasing the usability of this protocol.

Applications of the method

lentiMPRA can be used for numerous research purposes, such as analyzing hundreds of thousands of different candidate enhancers and their variants (e.g., rare and common GWAS-associated single-nucleotide polymorphisms (SNPs), evolutionary variants) in the genome, decoding the regulatory code, determining how it evolved in other species and generating specific genetic switches. It provides the ability to carry out these experiments in hard-to-transfect cells (e.g., primary cells, neurons, and many others) and integrates into the nucleus, providing an in-genome readout that we have shown is more reproducible and more predictive of functionality than either biochemical annotations or sequence-based models[14].

MPRAflow uses the pipelining tool Nextflow[13], which automatically runs MPRA processing code (written in Python, Bash, and R), manages all necessary packages and environments with Anaconda[17], and is compatible with a multitude of computational architectures, including a variety of high-performance computing (HPC) clusters and cloud computing systems. In addition, technical replicates and experimental conditions are parallelized through these HPC systems. Because MPRAflow is a package that allows non-bioinformatic researchers to easily analyze MPRA data, it can greatly increase the usability of this method in labs that do not have in-house bioinformaticians. In addition, MPRAflow provides easily interpretable graphics and produces files correctly formatted for readily available tools for further in-depth bioinformatic analysis such as MPRAnalyze[16].

Comparisons with other methods

Several different varieties of MPRA are available, such as episomal barcode–based MPRAs, STARR-seq, and others[11]. lentiMPRA differs from these methods because it provides an in-genome readout in a wider range of cell types. In STARR-seq, the CRS itself acts as the barcode. This attribute can potentially impact results because of the binding of RNA-associated factors and the RNA stability of the assayed sequence[15]. Using on average more than fifty 15-bp barcodes per CRS in lentiMPRA reduces this impediment. CRSs are usually generated via oligonucleotide synthesis but can also be produced by other processes, such as PCR or DNA capture–based methods. Barcodes can be added either as part of the synthesis or via PCR, providing flexibility in cloning design. Because lentiviruses integrate throughout the genome, we introduced anti-repressors on either side of the virus that, together with having >50 barcodes per assayed sequence, assist in overcoming differences due to varying genomic integration sites.

Previous MPRA processing tools have mainly focused on CRS library design or determination of CRS activity from count matrices, overlooking the computationally expensive task of processing sequencing data. MPRAflow is based on computational methods used in our previous MPRA work[14,15,18–20] and contains three utilities: association, count, and saturation mutagenesis. The association utility processes demultiplexed .fastq files and assigns barcodes to the CRS that are cloned with in the random pairing design. Sensitive alignment of merged paired-end reads provides robustness against sequencing and synthesis errors without strict read filters, even when CRS libraries contain sequences that differ by only one nucleotide. The count utility processes demultiplexed .fastq files to perform quality control (QC) across replicates, normalizes barcode count tables per CRS, and quantifies $\log_2$(RNA/DNA) ratios per CRS. MPRAnalyze inputs can also be produced using the count utility. Saturation mutagenesis dissolves multiple variants per CRS into single-variant ratios by applying a multivariate linear model, and it can be combined with the count utility. Each utility is executed with a single command on a terminal, and all utilities provide easily interpretable visualizations of all analyses performed.

Experimental design

Library design

CRSs can be identified using many of the aforementioned biochemical assays (ChIP-seq, DNase-seq, ATAC- seq, CUT&RUN, GWAS, Hi-C and others). Variants of interest within these CRSs can be identified via GWAS, GTEx (https://www.gtexportal.org/home/), various genomic websites such as Genome Aggregation Database (gnomAD[21]), comparative genomics and many other databases. The CRSs and variants tested ultimately depend on the goal of the study. Negative and positive controls should be included in the lentiMPRA library. For negative controls, sequences that could be used are those that are known not to be active in the assayed cell type, having silencing marks such as H3K27me3 within this tissue, or scrambled CRSs that are randomly selected from the library. For positive controls, sequences that are known to function as promoters/enhancers in this cell/type or tissue could be used. If such data do not exist, one can characterize CRSs from the cell where the lentiMPRA will be done via the aforementioned biochemical assays. These controls should be present within every technical and biological condition that will be tested. Tools such as MPRAnator[22] or MPRA Design Tools[23] can assist in choosing regions to test via MPRA and assembling the .fasta files required to order the libraries. Libraries can contain up to hundreds of thousands of sequences, depending on the infection efficiency of the cells (see Supplementary Table 1). The length of these sequences can also vary (as long as the combined length is not >10 kb, the optimal packaging capacity of lentivirus), depending on how the CRSs are generated (i.e., oligonucleotide synthesis, PCR, or capture). For more information on library design, see Box 1.

Library generation

For this protocol, we will focus on oligonucleotide synthesis because it is currently the most cost-effective way to generate fixed-length CRSs. Here, the synthesized oligonucleotide pool of the CRSs is amplified via two rounds of PCR, first to add the minimal promoter, and then to add the barcode. The amplified fragments are cloned via Gibson assembly into the *Sbf*I/*Age*I site of the pLS-SceI vector to construct the library. The resulting library is digested with I-*Sce*I to remove any vector that did not receive an insert. The recombination products are then electroporated into competent cells and plated onto ampicillin plates. Sanger sequencing of 16 colonies is then used to confirm the proper assembly of the library. The number of plates will dictate the number of barcodes each CRS will have on average. The number of colonies required for plasmid extraction will depend on the number of CRSs tested and the desired number of barcodes per CRS. Generally, it is ideal to have at least 50 barcodes per CRS, and the total number of colonies should roughly equal the desired library complexity. We recommend limiting the complexity of the library because of the finite nature of the multiplicity of

Box 1 | Library design criteria

1 Oligonucleotide synthesis length via Agilent is currently limited to 230 nucleotides (Fig. 1b, Extended Data Fig. 1a). Because the sequence should include 15-base common sequences (adaptors) on both sides of the CRS to amplify the library via PCR, the maximum length of the CRS is 200 bp. Twist Bioscience currently provides oligonucleotide synthesis service for oligonucleotides up to 300 bp and could increase CRS length. Three pairs of adaptor sequences are shown below. If two or three independent libraries are synthesized on the same array, use the second and third options. Asterisks represent the 200-base CRS sequence.

5′-AGGACCGGATCAACT**200base_CRS**CATTGCGTGAACCGA-3′
5′-AATGCTAGCGCATGG**200base_CRS**CTGCAACCTACGGAA-3′
5′-TTACGAGCCGTAGTC**200base_CRS**GCATCTCAACGTGGT-3′

2 The number of oligonucleotides that can be synthesized by Agilent is currently limited to 244,000 (Twist Bioscience's current limit is 696,000). The number of CRSs that can be tested with lentiMPRA is also limited by the infectability of the cells that are used (i.e., MOI is a limitation) and the number of cells available for the experiment (e.g., primary cells, hypo-proliferative cells), as explained in Box 2. For example, we have analyzed 164,000 CRSs with 50 barcodes per CRS using 15 million HepG2 cells at an MOI of 50. This can be simulated using Supplementary Table 1.

3 We recommend avoiding the use of homopolymers (>8 bases) in the design because they may cause a higher synthesis error rate.

infection (MOI) and the associated increase in sequencing costs. The complexity recommended in this protocol is 0.5–12 million total barcodes. The library should then be midi-prepped to extract the final plasmid library.

Association sequencing

To associate the barcode to the CRS, PCR is performed on the plasmid library to add flowcell sequences and sample indexes to the CRS–barcode pairs. The PCR product is then gel-extracted at the appropriate insert size (~471 bp for a 200-bp CRS) and sent for paired-end sequencing with an index read for barcode sequence, using custom primers provided in this protocol.

Lentiviral prep

The next step is to generate a lentivirus library. This is done by transfecting 293T cells with the plasmid library. Following 2 days in culture with titer boost reagent, the virus is collected and concentrated. To titrate the lentivirus, the cell type of interest is plated into 8 wells of a 24 well plate and infected with varying volumes of the virus (0, 1, 2, 4, 8, 16, 32, 64 μL) in each well. Cells are monitored for viability throughout this time in order to determine whether certain concentrations are toxic to them. Following a 3-d incubation (to reduce non- integrating lentivirus), genomic DNA is extracted from each well. qPCR is then carried out for each condition using primers against genomic DNA, integrated viral DNA, and plasmid backbone DNA. The MOI is calculated for each viral concentration (Supplementary Table 2). These values are then plotted against the viral volume to calculate the viral titer. Conditions need to be adjusted if cells are not viable.

Infection and sequencing

The lentiMPRA library is then infected into the cells of interest and incubated for 3 d. The number of cells required is determined on the basis of library complexity and the highest MOI that the cells can be infected with that is not toxic to the cells. We strongly recommend carrying out three technical replicates for each biological condition tested to assess reproducibility. The cells are then washed to reduce non-integrating lentivirus, and DNA and RNA are simultaneously extracted. RNA is treated with DNase and reverse transcription is done using construct-specific primers that contain P7 flowcell sequences and unique molecular identifiers (UMIs), to preserve the true counts of molecules through the amplification process. PCR is carried out on the DNA and RNA samples to amplify barcodes, adding P5 flowcell sequence and sample index upstream, and P7 flowcell sequence and UMI to the barcode. The sequencing libraries are then pooled and sent for paired-end sequencing with a UMI and sample index read.

Data processing

We built a computational tool, MPRAflow, to easily process demultiplexed .fastq data resulting from lentiMPRA and other MPRA experiments. If the barcodes are randomly paired with CRSs, the association utility can be run to assign barcodes to the appropriate CRS. We provide a workflow tailored to testing distinct CRSs, using Burrows–Wheeler Aligner (BWA[24]) to align sequences to the ordered oligonucleotide pool, a workflow for libraries containing single-nucleotide variants of the same CRS, using Bowtie2[25] and a list of the expected positions of the variants. The resulting pairing is then used in the count utility, which processes the barcode sequencing of the DNA and RNA to create normalized $\log_2$(RNA/DNA) ratios for the transcriptional activity of each CRS tested, along with easy-to-interpret visualizations. If more robust statistical analyses are desired, we provide the option to generate input files for MPRAnalyze[16], a generalized linear model approach. In addition, we provide an alternative workflow for quantifying expression of CRS libraries produced with saturation mutagenesis. It processes data into a matrix of RNA count, DNA count, and N binary columns indicating whether a specific sequence variant was associated with the barcode (T), which are used to fit a multiple linear regression model of $\log_2(\mathrm{RNA}_j) \sim \log_2(\mathrm{DNA}_j) + N + \text{offset}$ ($j \in \mathrm{T}$) and report the coefficients of N as effects for each variant. The utility processes multiple replicates and conditions in parallel if an HPC cluster is available but can also be run locally. This code is freely available on GitHub (https://github.com/shendurelab/MPRAflow).

Necessary expertise

Basic molecular biology and cell culture skills are required to perform lentiMPRA. For MPRAflow, a basic familiarity with command-line tools is needed.

Limitations

lentiMPRA has several limitations. These include a limitation in the number of CRSs that can be tested in cells that are not amenable to high lentivirus concentrations, although this can be ameliorated by using a larger number of cells. The use of oligonucleotide synthesis to generate the CRS library can also limit the number of sequences that can be tested, as well as their length. Improvements in DNA synthesis, as well as PCR or DNA capture–based methods, may ultimately overcome this limitation. Techniques that enable multiplex pairwise assembly of oligonucleotides[26] could also be used to increase CRS size by patching together specific oligonucleotides.

As for MPRAflow, although this tool is applicable to many types of MPRA, it does not support STARR-seq workflows because it does not include functionality for peak calling.

Materials

Biological materials

! CAUTION Cell lines should be regularly checked to ensure that they are authentic and that they are not infected with mycoplasma.

- 293T cells (ATCC, cat. no. CRL-3216, RRID: CVCL_0063)
- Cell lines of interest. All data shown in this protocol were generated from HepG2 cells (ATCC, cat. no. HB-8065, RRID: CVCL_0027)

Reagents

- DMEM (Life Technologies, cat. no. 11995-065)
- FBS (VWR International, cat. no. 89510-194)
- Penicillin–streptomycin (Life Technologies, cat. no. 15140-122)
- Trypsin-EDTA (0.05%; Life Technologies, cat. no. 25300-062)
- Polybrene (Sigma-Aldrich, cat. no. TR-1003-G)
- DPBS (Sigma-Aldrich, cat. no. D8537)
- Wizard SV Genomic DNA Purification System (Promega, cat. no. A2361)
- SsoFast EvaGreen Supermix (Bio-Rad, cat. no. 1725204)
- Primers and adaptors (custom-made by IDT with standard desalting; Supplementary Table 3)
- UltraPure DNase/RNase-free distilled water (Life Technologies, cat. no. 10977-023)
- SurePrint 244K Oligonucleotide Libraries (Agilent, cat. no. G7223A)
- TE buffer (Tris–EDTA; Teknova, cat. no. T0225)
- NEBNext High-Fidelity 2× PCR Master Mix (New England BioLabs, cat. no. M0541L)
- Ethyl alcohol (Sigma-Aldrich, cat. no. E7023-500ML)
- Buffer EB (Qiagen, cat. no. 19086)
- HighPrep PCR reagent (MagBio Genomics, cat. no. AC60050)
- Gel loading dye (6×; New England BioLabs, cat. no. B7025S)
- SeaKem LE agarose (Lonza, cat. no. 50004)
- TAE (Thermo Fisher Scientific, cat. no. BP13324)
- SYBR Safe DNA gel stain (Invitrogen, cat. no. S33102)
- QIAquick Gel Extraction Kit (Qiagen, cat. no. 28704)

- CutSmart buffer (10×; New England BioLabs, cat. no. B7204S)
- *Age*I-HF (New England BioLabs, cat. no. R3552L)
- *Sbf*I-HF (New England BioLabs, cat. no. R3642L)
- I-*Sce*I (New England BioLabs, cat. no. R0694S)
- NEBuilder HiFi DNA Assembly Master Mix (New England BioLabs, cat. no. E2621L)
- NEB 10-beta electrocompetent cells (New England BioLabs, cat. no. C3020K)
- LB base (Life Technologies, cat. no. 12780029)
- LB agar plates (15 cm; Teknova, cat. no. L5002)
- Carbenicillin (Teknova, cat. no. C2130)
- QIAprep Spin Miniprep Kit (Qiagen, cat. no. 27106)
- Qiagen Plasmid Plus Midi Kit (Qiagen, cat. no. 12945)
- DNA Ladder (1 kb; New England BioLabs, cat. no. N3232S)
- DNA ladder (100 bp; New England BioLabs, cat. no. N3231S)
- Qubit dsDNA HS Assay Kit (Life Technologies, cat. no. Q32851)
- Qubit RNA HS Assay Kit (Life Technologies, cat. no. Q32852)
- OPTI-MEM (Life Technologies, cat. no. 31985070)
- EndoFectin (Genecopoeia, cat. no. EFL1001-01)
- psPAX2 (Addgene, cat. no. 12260; RRID: Addgene_12260)
- pMD2.G (Addgene, cat. no.12259; RRID: Addgene_12259)
- pLS-SV40-mP-EGFP (Addgene, cat. no. 137724; RRID: Addgene_137724)
- pLS-SceI (Addgene, cat. no. 137725; RRID: Addgene_137725)
- ViralBoost reagent (Alstem, cat. no. VB100)
- Lenti-X Concentrator (Takara, cat. no. 631232)
- AllPrep DNA/RNA Mini Kit (Qiagen, cat. no. 80204)
- RNase-free DNase Set (Qiagen, cat. no. 79256)
- 2-Mercaptoethanol (Bio-Rad, cat. no. 1610710) **! CAUTION** 2-Mercaptoethanol is toxic, so it should be handled in a hood while wearing disposable gloves.
- TURBO DNA-free Kit (Life Technologies, cat. no. AM1907)
- SuperScript II Reverse Transcriptase (Life Technologies, cat. no. 18064-071)
- SYBR Green I nucleic acid gel stain (10,000×; Invitrogen, cat. no. S7563)

Equipment

- Pipettes (20 μL, 200 μL and 1,000 μL; Rainin, cat. nos. 17014392, 17014391 and 17014382)
- Filter tips (20 μL, 200 μL and 1,000 μL; Rainin, cat. nos. 17005860, 17005859 and 17007081)
- Serological pipettes (5 mL, 10 mL and 25 mL; Genesee Scientific, cat. nos. 12-102, 12-104 and 12-106)
- Pipet-Aid XP (Drummond, cat. no. 4-000-101)
- Cell culture plates (24 well, 10 cm, and 15 cm; Genesee Scientific, cat. nos. 25-107, 25-202, and 25-203)
- Inverted fluorescence microscope (Leica, DMIL LED)
- CO_2 incubator (Thermo Fisher Scientific, Thermo Forma Series II water jacketed)
- Hemocytometer (Hausser Scientific, cat. no. 3200)
- DNA LoBind tubes (Eppendorf, cat. no. 022431021)
- PCR tubes (8-strip; Axygen, cat. no. PCR-0208-FCP-C)
- Vortex mixer (Thermo Fisher Scientific, cat. no. 88880017)
- Spectrophotometer (NanoDrop 8000; Thermo Fisher Scientific, cat. no. ND-8000-GL)
- Qubit fluorometer (Life Technologies, cat. no. Q32857)
- qPCR instrument (QuantStudio v. 6 Flex Real-Time PCR System; Applied Biosystems, cat. no. 4485699)
- Thermal cycler (ProFlex PCR system; Applied Biosystems, cat. no. 4484073)
- DynaMag-2 magnet (Thermo Fisher Scientific, cat. no. 12321D)
- Tabletop centrifuge (Myspin 6; Thermo Fisher Scientific, cat. no. 75004061)
- Gel electrophoresis system (Mupid-2plus; Takara, cat. no. AD110)
- Gel casting set (Takara, cat. no. AD216)
- Gel combs (Takara, cat. no. AD214)
- Safe Imager 2.0 Blue-Light Transilluminator (Life Technologies, cat. no. G6600)
- Heating dry bath (Thermo Fisher Scientific, cat. no. 88880027) **! CAUTION** The temperature displayed by the digital thermometer in the heat bath may not be accurate. To calibrate the instrument, we recommend using an alcohol thermometer to measure the temperature of water in a tube placed on the instrument.

- Alcohol thermometer
- Cuvettes (1-mm gap; BTX Harvard Apparatus, cat. no. 450124)
- Gemini X2 Electroporation System (BTX Harvard Apparatus, cat. no. 452007)
- 37 °C shaker (New Brunswick Scientific, Excella E24)
- Round-bottom tubes (14 mL; Corning, cat. no. 352059)
- Tubes (50 mL; Corning, cat. no. 352070)
- 37 °C incubator (Boekel scientific, cat. no. 133001)
- T225 flasks (Corning, cat. no. 431082)
- Centrifuge (Eppendorf, cat. no. 022625501)
- Polyethersulfone (PES) filter units (0.45 μm; Thermo Fisher Scientific, cat. no. 165-0045)
- Cell lifters (Corning, cat. no. 3008)
- Luer-Lok syringes (3 mL; BD, cat. no. 309657)
- Needles (20-gauge; BD, cat. no. 305179)
- Parafilm (Heathrow Scientific, cat. no. HS234526B)

Software

- Conda (https://docs.conda.io/en/latest/miniconda.html)
- Linux (https://www.linux.org/pages/download/)

Reagent setup

DMEM (with 10% (vol/vol) heat-inactivated FBS)

Incubate FBS at 55 °C for 40 min. Supplement DMEM with 10% (vol/vol) heat-inactivated FBS and 1% (vol/vol) penicillin–streptomycin. Store at 4 °C for up to 3 months.

DMEM (with 5% (vol/vol) heat-inactivated FBS)

Supplement DMEM with 5% (vol/vol) heat-inactivated FBS and 1% (vol/vol) penicillin–streptomycin. Store at 4 °C for up to 3 months.

80% Ethanol

Dilute 8 mL of ethyl alcohol with 2 mL of UltraPure distilled H_2O. Store at room temperature (RT; 22–25 °C) for up to 2 weeks.

LB medium

Suspend 20 g of LB base in 1 L of distilled water and sterilize by autoclaving. Store at RT for up to 4 months.

TAE–agarose gels

Dissolve SeaKem LE agarose in TAE by boiling. Either 0.3 mg, 0.45 mg or 0.54 mg of agarose per 30 mL TAE should be used to obtain 1%, 1.5% or 1.8% (wt/vol) gel, respectively. Add 3 μL of SYBR safe DNA gel stain and cast the gel using an appropriate gel casting set and comb, as described below.

Procedure

Library amplification ● Timing 3 h

1 Dissolve the Agilent oligonucleotide (10 pmol) (Fig. 1b and Box 1) in 100 μL TE buffer to obtain a 100 nM solution.

2 Set up the first-round PCR reaction. This reaction adds a vector overhang sequence upstream and minimal promoter and adaptor sequences downstream of the CRSs (Fig. 1c, Extended Data Fig. 1b).

Reagent	Volume (μL)	Final conc.
Agilent oligonucleotide (100 nM)	2	1 nM
NEBNext High-Fidelity 2× PCR Master Mix	100	1×
5BC-AG-f01 (100 μM)	1	0.5 μM
5BC-AG-r01 (100 μM)	1	0.5 μM
Ultrapure distilled H_2O	96	
Total volume	200	

3 Split the premixture into five PCR tubes (40 μL per tube).
▲ CRITICAL STEP Splitting the PCR reaction into multiple tubes is important to reduce the risk of PCR 'jackpotting' (errors that occur during the early PCR cycles and get amplified exponentially) or amplification bias accidentally occurring during PCR.

4 Run the PCR reaction as follows:

Cycle no.	Denature	Anneal	Extend
1	98 °C, 2 min		
2–6 (5 cycles)	98 °C, 15 s	60 °C, 20 s	72 °C, 30 s
7			72 °C, 5 min

5 Combine the PCR products in a DNA LoBind tube.

6 Bring the HighPrep PCR reagent to RT for at least 30 min before use. Shake thoroughly to fully resuspend the magnetic beads. Add 1 volume (200 μL) of HighPrep PCR reagent and mix thoroughly by pipetting up and down 6–8 times.

7 Incubate the mixture for 5 min at RT.

8 Place the tube on the magnet for 2–3 min until the solution clears and beads pull to the side of the wall.

9 Carefully remove the supernatant by pipetting without disturbing the beads. A small amount (10–20 μL) of supernatant can be left in the tube.

10 With the tube on the magnet, add 500 μL of 80% (vol/vol) ethanol and incubate for 30 s.

11 Remove the ethanol by pipetting while the tube is still on the magnet.

12 Repeat the 80% (vol/vol) ethanol washing (Steps 10 and 11).

13 Flash-spin (200–1,000*g*, 22–25 °C, 3 s) the tube, immediately place it back on the magnet and remove the supernatant.

14 Dry the bead pellet for 2–3 min. Do not overdry the beads.

15 Add 50 μL of Buffer EB to the beads and mix by pipetting and vortexing.

16 Place the tube on the magnet for 1–2 min until the solution is clear.

17 Transfer 45–50 μL of the eluate to a DNA LoBind tube.

18 Measure the DNA concentration using a NanoDrop spectrophotometer. The expected concentration is 5–20 ng/μL.

19 Set up the second-round PCR reaction. This reaction adds a 15-bp barcode and vector overhang sequence downstream of the first-round PCR fragment (Fig. 1d, Extended Data Fig. 1b).

Reagent	Volume (μL)	Final conc.
First-round PCR product	Variable (100 ng)	
NEBNext High-Fidelity 2× PCR Master Mix	200	1×
5BC-AG-f02 (100 μM)	2	0.5 μM
5BC-AG-r02 (100 μM)	2	0.5 μM
Ultrapure distilled H_2O	Make up to 400 μL	
Total volume	400	

20 Split the premixture into 10 PCR tubes (40 μL per tube).

21 Run the PCR reaction as follows:

Cycle no.	Denature	Anneal	Extend
1	98 °C, 2 min		
2–13 (12 cycles)	98 °C, 15 s	60 °C, 20 s	72 °C, 30 s
14			72 °C, 5 min

22 Combine the PCR products into a DNA LoBind tube.

23 Add 200 μL of 6× gel loading dye (final conc. 2×) and mix the solution by vortexing.

24 Run the sample on two 1.5% (wt/vol) TAE–agarose gels (30 mL of 5 × 6-cm mini gels with 3-cm-width wells) and visualize the DNA using SYBR Safe DNA gel stain.

25 Cut the DNA band (419 bp) using a blue-light Safe Imager.

▲**CRITICAL STEP** We highly recommend using a blue-light Safe Imager, because the UV transilluminator markedly decreases the recombination efficiency.

26 Purify the DNA from the gel slice, using QIAquick Gel Extraction Kit according to the manufacturer's protocol.

27 Elute the DNA in 50 μL of Buffer EB per column. If multiple columns are used, combine the eluate.

28 Purify the DNA using 1.2 volumes of HighPrep PCR reagent and following Steps 6–17.

29 Measure the DNA concentration using a NanoDrop spectrophotometer. The expected concentration is ~25 ng/μL.

■**PAUSE POINT** Purified DNA can be stored at −20 °C for months.

? TROUBLESHOOTING

Vector linearization ● Timing 7 h to overnight

30 Set up the vector digestion reaction as follows:

Reagent	Volume (μL)	Final conc.
pLS-SceI	Variable (10 μg)	
CutSmart buffer (10×)	20	1×
*Age*I-HF (20 U/μL)	5 (100 U)	0.5 U/μL
*Sbf*I-HF (20 U/μL)	5 (100 U)	0.5 U/μL
Ultrapure distilled H_2O	Make up to 200 μL	
Total volume	200	

31 Incubate the reaction at 37 °C for 3 h to overnight.

32 To complete the plasmid digestion, add 5 μL of *Age*I-HF (20 U/μL) and 5 μL of *Sbf*I-HF (20 U/μL) to the reaction.

33 Incubate the reaction at 37 °C for 3 h to overnight.

34 Vortex for 30 s and incubate at 80 °C for 20 min.

35 Purify the DNA using 0.65 volume (136.5 μL) of HighPrep PCR reagent and following Steps 6–17.

36 Measure the DNA concentration using a NanoDrop spectrophotometer.

37 To check the DNA size and quality, run 100–200 ng of the linearized vector and purified insert DNA (from Step 29) on a 1% (wt/vol) gel along with a 1-kb DNA ladder. Make sure that specific single bands (7.8-kb linearized vector and 419-bp insert DNA) appear, but that no other bands appear on the gel.

■**PAUSE POINT** Linearized vector can be stored at −20 °C for a year.

? TROUBLESHOOTING

Recombination and electroporation ● Timing 3 d (5 h hands-on time)

38 Set up the recombination reaction as follows:

Reagent	Volume (μL)	Final conc.
Linearized pLS-SceI (from Step 36)	Variable (1 μg)	
Purified insert DNA (from Step 29)	Variable (250 ng)	
NEBuilder HiFi DNA Assembly Master Mix	100	1×
Ultrapure distilled H_2O	Make up to 200 μL	
Total volume	200	

39 Incubate the reaction at 50 °C for 60 min in a heating dry bath.

40 Place the tube on ice.

41 Purify the DNA using 0.65 volume (136.5 μL) of HighPrep PCR reagent and following Steps 6–17.

42 Set up the digestion reaction to get rid of undigested vectors as follows:

Reagent	Volume (μL)	Final conc.
Recombination product	44	
CutSmart buffer (10×)	5	1×
I-*Sce*I (20 U/μL)	1 (20 U)	0.4 U/μL
Total volume	50	

43 Incubate the reaction at 37 °C for 1 h.

44 Purify the DNA using 1.8 volume (90 μL) of HighPrep PCR reagent and following Steps 6–14.

45 To elute the DNA, add 20 μL of Buffer EB to the beads and mix by pipetting and vortexing.

46 Place the tube on the magnet for 1–2 min and transfer 18 μL of the eluate to a DNA LoBind tube.

47 Measure the DNA concentration using a NanoDrop spectrophotometer. Make sure the DNA concentration of the recombination product (the eluate) is >25 ng/μL, so as not to need to add >4 μL to 100 μL of competent cells in Step 50.

■ **PAUSE POINT** The recombinant product can be stored at −20 °C for at least a month.

48 Prewarm 4–5 mL of Stable Outgrowth Medium (from the NEB 10-beta electrocompetent cells) at 37 °C for at least 30 min.

49 Thaw NEB 10-beta electrocompetent cells on ice. We usually use 100 μL (one tube) of the competent cells for low-complexity libraries (0.5–2 million total barcodes) and 400 μL (four tubes) for high-complexity libraries (8–12 million total barcodes).

▲ **CRITICAL STEP** Competent cells and cuvettes should be kept on ice during the following procedure.

50 Add 100 ng of the recombination product per 100 μL competent cells. The volume of DNA should be <4 μL per 100 μL competent cells.

51 Gently transfer 50 μL of the competent cells to a 1-mm-gap cuvette without creating bubbles. Two cuvettes are prepared for 100 μL of competent cells.

52 Gently tap the cuvettes on the counter to move the cells to the bottom.

53 Place the cuvettes in a Gemini X2 electroporator and shock the cells with the following settings: voltage, 2.0 kV; resistance, 200 ohms; capacitance, 25 μF, number of pulses, 1; gap width, 1 mm.

54 Immediately add 450 μL of prewarmed Stable Outgrowth Medium to the cuvettes, thoroughly mix by pipetting up and down, and transfer to a 14-mL conical tube.

55 Repeat the electroporation for all cuvettes, combining the electroporated bacteria in a single tube (total 1 mL culture per 100 μL competent cells).

56 Add fresh Stable Outgrowth Medium and scale up to 4 mL in total. If 400 μL competent cells were used, there is no need to add more.

57 Incubate the cells at 37 °C for 1 h with agitation (200 r.p.m.). Prewarm ten 15-cm LB agar plates at 37 °C.

▲ **CRITICAL STEP** We recommend using 15-cm plates rather than larger plates because these enable fine-tuning of the colony numbers collected.

58 Dilute 2 μL of the bacteria in 400 μL of fresh LB medium in a 1.5-mL tube and plate the entire tube of diluted bacteria (402 μL) in a prewarmed 15-cm LB agar plate along with 20 μL of 100 mg/mL carbenicillin. This plate will be used for colony counting and plasmid mini prep.

59 Plate undiluted bacteria onto the other nine prewarmed 15-cm LB agar plates (400 μL/plate), along with 100 μL/plate of 100 mg/mL carbenicillin. A higher amount of carbenicillin than usual is added because the dense culture conditions increase the risk of non-transformed bacteria growth.

60 Incubate the plates at 37 °C overnight.

61 To check the plasmid sequence of individual colonies, pick 16 colonies from the diluted-bacteria plate, purify the plasmids using a QIAprep Spin Miniprep Kit, and send them for Sanger sequencing using n40.dn.F and EGFP.up.R primers (Supplementary Table 3). Confirm that the sequence structure corresponds to the design (Fig. 1e, Extended Data Fig. 1c).

■ **PAUSE POINT** The plates can be stored at 4 °C for a month.

? TROUBLESHOOTING

Colony counting and plasmid library prep ● Timing 3 h

62 Count the number of colonies on the diluted-bacteria plate. If there are too many colonies, count colonies in a quarter of the area and multiply by four to estimate the total number of colonies on the plate.

63 Estimate the total number of colonies per undiluted-bacteria plate by multiplying the colony count in the diluted-bacteria plate by 200 (Supplementary Table 1).

64 Determine the number of undiluted-bacteria plates to be used for the following plasmid preps. The total number of colonies needed can be determined by multiplying the number of designed CRSs by the desired number of barcodes per CRS (Supplementary Table 1).

▲**CRITICAL STEP** The ideal number of barcodes per CRS is between 50 and 200. Fewer barcodes per CRS may reduce reproducibility. More barcodes per CRS requires more cells, more virus and deeper sequencing reads, which increase costs. In addition, associating >200 barcodes per CRS does not increase reproducibility.

65 Add 5–6 mL of LB medium to each bacterial plate and gently scrape the colonies, using a cell lifter without disturbing solid agarose.

66 Collect the bacterial suspension and combine into a few 50-mL tubes.

67 Add 5–6 mL of fresh LB medium again to the plates and collect as much leftover bacteria as possible into the tubes.

68 Purify the plasmids using a Qiagen Plasmid Plus Midi Kit, following the 'standard protocol' in the manufacturer's protocol. The number of columns to be used depends on the amount of bacteria. We usually use four columns of Qiagen Plasmid Plus Midi Kit per undiluted-bacteria plate.

69 Measure the plasmid concentration using a NanoDrop spectrophotometer. The expected concentration is 0.5–2 µg/µL.

70 To check the DNA size and quality, run 100–200 ng of the plasmid on a 1% (wt/vol) agarose gel along with a 1-kb DNA ladder.

■**PAUSE POINT** The purified plasmid library can be stored at −20 °C for years.

Sequencing for CRS–barcode association ● Timing 2–4 weeks (4 h hands-on time plus sequencing turnaround time)

71 Set up the PCR reaction. This reaction adds a P5 flowcell sequence and the sample index sequence upstream and a P7 flowcell sequence downstream of the CRS–barcode fragment (Fig. 1f,g, Extended Data Fig. 1d). Use different sample index sequences for pLSmP-ass-i# if multiple libraries are generated and multiplexed (Supplementary Table 3).

Reagent	Volume (µL)	Final conc.
Plasmid library	Variable (40 ng)	
NEBNext High-Fidelity 2× PCR Master Mix	100	1×
pLSmP-ass-i# (100 µM)	1	0.5 µM
pLSmP-ass-gfp (100 µM)	1	0.5 µM
Ultrapure distilled H_2O	Make up to 200 µL	
Total volume	200	

72 Split the premixture into five PCR tubes (40 µL per tube).

73 Run the PCR reaction as follows:

Cycle no.	Denature	Anneal	Extend
1	98 °C, 1 min		
2–16 (15 cycles)	98 °C, 15 s	60 °C, 20 s	72 °C, 3 min
17			72 °C, 5 min

▲**CRITICAL STEP** Incomplete DNA elongation may create chimeric DNA annealing in the next cycle and can cause CRS–barcode swapping. A longer extension time (3 min) can help to reduce this risk.

74 Combine the PCR products in a DNA LoBind tube.

75 Add 100 µL of 6× gel loading dye (final conc. 2×) and mix the solution by vortexing.

76 Run the sample on a 1.5% (wt/vol) agarose gel (30 mL of 5 cm × 6-cm mini gel with 3-cm-width well).

77 Cut the DNA band (470 bp) using a blue-light Safe Imager.

78 Purify the DNA from the gel slice using the QIAquick Gel Extraction Kit according to the manufacturer's protocol.
79 Elute the DNA in 50 µL Buffer EB per column. If multiple columns are used, combine the eluate.
80 Purify the DNA using 1.8 volumes of HighPrep PCR reagent, following Steps 6–17.
81 Measure the DNA concentration using Qubit dsDNA HS Assay Kit according to the manufacturer's protocol.
82 To check the DNA size and quality, run 50–100 ng of the DNA on a 1.5% (wt/vol) agarose gel along with a 100-bp DNA ladder.

■ **PAUSE POINT** Purified DNA can be stored at −20 °C for months.

83 Send the purified DNA and custom primers for sequencing (Fig. 1h, Extended Data Fig. 1d, Supplementary Table 3). The sequencing should be done using paired-end reads covering the full CRS with some overlap (here 146 bp each), with 15 cycles for index read 1 and 10 cycles for index read 2. Index read 1 provides the barcode sequence, and index read 2 provides the sample index (let the sequencing facility know that index read 2 should be used for demultiplexing and that short reads should not be masked, bcl2fastq parameters: `--minimum-trimmed-read-length 0 --mask-short-adapter-reads 0`). A minimum 10× coverage of sequencing reads based on the total number of barcodes is required. For example, we use an Illumina MiSeq v.2 run (15 million reads) for a 0.5-million-barcode library or an Illumina NextSeq mid-output run (120 million reads) for an 8- to 12-million-barcode library (Supplementary Table 1).

Read	Cycles	Primer	Output
Read 1	146	pLSmP-ass-seq-R1	CRS (upstream, forward)
Read 2	146	pLSmP-ass-seq-R2	CRS (downstream, reverse)
Index read 1	15	pLSmP-ass-seq-ind1	Barcode (forward)
Index read 2	10	pLSmP-rand-ind2	Sample index

Lentivirus packaging ● Timing >1 week (>5.5 h hands-on time)

84 Culture 293T cells in DMEM (with 10% heat-inactivated FBS).
85 Seed 10–12 million 293T cells per T225 flask. The number of flasks depends on the library complexity and the infectability of the cells. For example, we use one flask for the 0.5-million-barcode library or six T225 flasks for the 8- to 12-million-barcode library when carrying out lentiMPRA in HepG2 cells.
86 Incubate the cells for 2 d.
87 Prepare premixtures A and B as follows:
- Premix A: 800 µL/flask OPTI-MEM and 60 µL/flask EndoFectin.
- Premix B: 800 µL/flask OPTI-MEM, 10 µg/flask plasmid library, 6.5 µg/flask psPAX2 and 3.5 µg/flask pMD2.G.

88 Add premix A to premix B by pipetting and mix the mixture by inverting the tube.
89 Incubate the tube at RT for 15 min.
90 Add 1.6 mL/flask of the mixture to 293T cells (from Step 86).
91 Incubate the cells for 8–14 h.

! CAUTION During the following procedure (Steps 92–103), the liquid and plastic waste should be discarded into 10% (vol/vol) bleach, because they are contaminated with lentivirus. Culture plates and tubes for storage should be clearly labeled as lentiviral contaminants.

92 Replace the media with 30 mL/flask DMEM (with 5% heat-inactivated FBS) supplemented with 1× ViralBoost reagent (60 µL reagent per 30 mL medium).
93 Incubate the cells for 36–48 h.
94 Check GFP expression using a fluorescence microscope. The majority of cells are expected to express strong GFP, because viral RNA, including the GFP gene, is transcribed via the 5′ long terminal repeat (LTR).
95 Filter the supernatant through a 0.45-µm PES filter system. Use multiple filters (one filter for up to three T225 flasks), so as not to get clogged.
96 Transfer the flow-through into multiple 50-mL tubes (30 mL per tube).
97 Add 1/3 volume (10 mL per tube) of Lenti-X concentrator reagent, close the lid tightly and mix gently by inverting the tubes.

98 Seal the lid with Parafilm and place the tube in a refrigerator at 4 °C for at least 4 h.

■ **PAUSE POINT** The tubes can be stored at 4 °C up to 1 week.

99 Centrifuge the tubes at 1,500*g* for 45 min at 4 °C.

100 Discard the supernatant into 10% (vol/vol) bleach by gentle decanting.

101 Discard the remainder of the supernatant into 10% bleach by pipetting without disturbing the pellet.

102 Gently resuspend the pellet in cold DPBS. We usually use 600 µL DPBS per T225 flask.

103 Store the lentivirus at 4 °C.

▲ **CRITICAL STEP** We do not recommend freezing the virus, especially in the case of high-complexity libraries, because freeze–thaw cycles substantially decreases the viral titer. We did not see marked loss of the viral titer when stored at 4 °C for up to 3 weeks. The following infection experiments should be done within 3 weeks.

■ **PAUSE POINT** The virus can be stored at 4 °C up to 3 weeks.

Lentivirus titration ● Timing >1 week (3 h hands-on time)

104 Infect the lentivirus library (0, 1, 2, 4, 8, 16, 32, 64 µL) into the cells to be used, extract genomic DNA from the cells, and perform qPCR as described in Box 2 (steps 1–14).

105 Plot the MOI for each condition and draw a linear approximation with the virus volume on the *x* axis and the MOI on the *y* axis (Supplementary Table 2).

106 On the basis of its slope and the number of cells seeded, calculate the virus titer (in transducing units per microliter), using Supplementary Table 1.

Lentivirus infection and DNA/RNA extraction ● Timing >1 week (>3.5 h hands-on time)

107 Determine the number of integrations per barcode (i.e., total number of a particular barcode existing in an entire cell population), using the spreadsheet provided in Supplementary Table 1. We recommend a range between 50 and 500 integrations per barcode. Fewer numbers increase the risk of barcode loss during the downstream procedure. Higher numbers are better but increase the cost.

108 Seed an appropriate number of cells in 10-cm or 15-cm dishes. The number of cells required is determined as total barcode integrations (total number of any barcodes existing in the entire cell population) divided by the MOI of the cells (Supplementary Table 1). Three independent biological replicates should be performed, and each replicate sample should be treated separately during the following procedures (Steps 109–148).

109 Incubate the cells overnight.

110 Refresh the culture media (culture conditions depend on the cell type used) and add Polybrene at the appropriate concentration (Box 2).

111 Add an appropriate amount of the lentivirus library. The volume of virus required is given as total barcode integrations divided by virus titer (Supplementary Table 1).

112 Refresh the culture media (culture conditions depend on the cell type used) with no Polybrene the following day.

113 After 2 d (3 d of culture in total), check GFP fluorescence to confirm proper lentiviral integration and expression (Fig. 1i). GFP expression depends on the library design. If the majority of the CRSs in the library are active enhancers, the cells should have strong GFP expression.

114 Remove the culture media and wash the cells with DPBS three times. Remove the DPBS completely.

115 Add RLT Plus lysis buffer (from the AllPrep DNA/RNA Mini Kit) supplemented with 2-mercaptoethanol (10 µL of 2-mercaptoethanol per 1 mL of RLT Plus). We usually use 1,200 µL or 2,400 µL of lysis buffer per 10-cm dish or 15-cm dish, respectively.

116 Scrape the cells using a cell lifter and homogenize the cell lysis, using a 3-mL syringe and 20-gauge needle.

■ **PAUSE POINT** The cell lysate can be frozen and stored at −80 °C for months.

117 Transfer the lysate to DNA columns. Use two or four columns per 10-cm dish or 15-cm dish, respectively.

118 Extract genomic DNA and total RNA simultaneously using the AllPrep DNA/RNA Mini Kit according to the manufacturer's protocol. For RNA samples, perform DNase treatment between two of the 350-µL RW1 washes using Qiagen's RNase-free DNase Set, according to the manufacturer's protocol.

119 Elute DNA in 30 µL/column of Buffer EB and combine the eluates of each replicate in a single tube. Keep each of the three replicates separate.

120 Elute RNA in 30 µL/column of RNase-free H_2O and combine the eluates of each replicate in a single tube. Keep each of the three replicates separate.

Box 2 | Test infection of the cells to be used for lentiMPRA ● Timing >1 week (3 h hands-on time)

1 Culture cells that will be used for lentiMPRA. Culture conditions depend on the cell type.
2 Trypsinize (if adherent cells) and count the cells.
3 Seed the cells into eight wells of a 24-well plate at a number that will provide 70–80% confluency the next day. The number of cells per well depends on cell type. For example, in the case of HepG2 and K562 cells, we seed 0.1 million cells per well.
4 Incubate the cells overnight.
5 Refresh the culture media and add Polybrene at a final concentration of 8 µg/mL.
? TROUBLESHOOTING
6 Add 0, 1, 2, 4, 8, 16, 32, or 64 µL of control lentivirus (e.g., pLS-SV40-mP-EGFP) to the wells. Lentivirus can be generated according to Steps 84–103 of the main procedure.
! CAUTION During steps 6–9, the liquid and plastic waste needs to be discarded into 10% bleach, as it is contaminated with lentivirus.
7 The next day, refresh the culture media without Polybrene. Roughly check the cell survival as compared with that the uninfected well.
8 After 2 d (3 d of culture in total), check the cell survival and GFP expression under a fluorescence microscope. If there is substantial cell death in certain wells, do not proceed with them for the following genomic DNA extraction step.
9 Remove the culture media and wash the cells three times with 500 µL of DPBS each time.
10 Extract genomic DNA from each well, using a Wizard SV Genomic DNA Purification System according to the manufacturer's protocol.
11 Measure the DNA concentration using a NanoDrop spectrophotometer and dilute the DNA to 10 ng/µL in 8-strip PCR tubes.
12 Set up qPCR using primers that amplify viral DNA (WPRE.F and WPRE.R, Supplementary Table 3), plasmid backbone DNA (BB.F and BB.R, Supplementary Table 3) and genomic DNA (LP34.F and LP34.R, Supplementary Table 3) for each sample. We usually use SsoFast EvaGreen Supermix according to the following settings.

Reagent	Volume (µL)	Final conc.
Template DNA (10 ng/µL)	2.5 (25 ng)	
SsoFast EvaGreen Supermix	5	1×
Forward primer (100 µM)	0.1	1 µM
Reverse primer (100 µM)	0.1	1 µM
UltraPure distilled H_2O	2.3	
Total volume	10	

13 Run the qPCR as follows:

Cycle no.	Denature	Anneal and extend	Gradient increase
1	95 °C, 1 min		
2–36 (35 cycles)	95 °C, 10 s	60 °C, 30 s	
37			60–95 °C in 15 min

14 Determine the MOI by calculating the relative amount of viral DNA over genomic DNA with the subtraction of the relative amount of backbone DNA, using Supplementary Table 2.
15 The resultant highest MOI is considered to be the maximum MOI for the cells to be used, because this is the viral amount that gives high GFP expression without substantial cell death.
▲ CRITICAL STEP If the maximum MOI of the cells is not high enough or the cell material is limited, the number of CRSs to be synthesized should be reduced accordingly. See also Box 1, step 2.
▲ CRITICAL STEP We do not recommend an MOI of >100, even if cell death was not observed, because adding a high amount of virus increases non-integrating virus left in the cells after 3 d of culture, which can increase background noise in DNA barcode amplification.

121 Measure the concentrations of the DNA and RNA samples using a NanoDrop spectrophotometer. At least 12 µg DNA and 60 µg RNA per replicate are required.
■ PAUSE POINT DNA can be stored at −20 °C for years. RNA can be stored at −80 °C for months.

Reverse transcription ● Timing 4 h

122 Treat RNA samples with DNase, using the TURBO DNA-free Kit and following the manufacturer's protocol for 'Rigorous DNase treatment'. Because the RNA sample has already been treated with DNase during the AllPrep procedure (Step 118), TURBO DNase treatment can be done using a high-concentration condition (without any dilution).
123 Measure the RNA concentration using the Qubit RNA HS Assay Kit. At least 60 µg or 240 µg total RNA per replicate is required for a low- (0.5–2 million total barcodes) or high- (8–12 million total barcodes) complexity library, respectively.
■ PAUSE POINT DNase-treated RNA can be stored at −80 °C for months.
124 For RNA samples, perform a reverse-transcription reaction in 8-strip PCR tubes. This reaction adds a 16-bp UMI and a P7 flowcell sequence downstream of the barcode (Extended Data Fig. 1e). For a

low-complexity library (0.5–2 million total barcodes), use the amounts given in the table below. For a high-complexity library (8–12 million total barcodes), we recommend multiplying all amounts in the following table by 4.

Reagent	Volume (µL)	Final conc.
RNA	Variable (60 µg total RNA)	
P7-pLSmP-ass16UMI-gfp (100 µM)	0.25	0.25 µM
dNTP mix (10 mM, from SuperScript II Reverse Transcriptase)	5	0.5 mM
UltraPure distilled H_2O	Make up to 65 µL	
Total volume	65	

125 Incubate the reaction at 65 °C for 5 min using a thermal cycler.
126 Place the tubes on ice.
127 Add 20 µL 5× First Strand buffer (from SuperScript II Reverse Transcriptase) and 10 µL 0.1 M DTT (from SuperScript II Reverse Transcriptase).
128 Incubate the tubes at 42 °C for 2 min using a thermal cycler.
129 Add 5 µL of Superscript II.
130 Incubate the tubes at 42 °C for 50 min, followed by incubation at 70 °C for 15 min using a thermal cycler.
■ **PAUSE POINT** cDNA can be stored at −20 °C for months.

Library prep and sequencing for RNA and DNA barcode counts ● Timing 2–4 weeks (6 h hands-on time plus sequencing turnaround time)

131 Dilute DNA samples (from Step 121) to a final concentration of 120 ng/µL. For RNA samples (from Step 130; we refer to RT products as RNA samples in the downstream steps to distinguish them from samples derived from genomic DNA), use all 100 µL of RT products for the following first-round PCR reaction.
132 Perform a first-round PCR reaction with all three replicates of both DNA and RNA samples. This reaction adds the P5 flowcell sequence and sample index sequence upstream and a 16-bp UMI and P7 flowcell sequence downstream of the barcode (Fig. 1j,k, Extended Data Fig. 1e). Use different sample index sequences for each sample (Supplementary Table 3). For a low-complexity library (0.5–2 million total barcodes), use the amounts given in the table below. For a high-complexity library (8–12 million total barcodes), we recommend multiplying all amounts in the following table by 4.

Reagent	Volume (µL)	Final conc.
DNA or cDNA	100 (12 µg DNA or entire RT product)	
NEBNext High-Fidelity 2× PCR Master Mix	200	1×
P7-pLSmP-ass16UMI-gfp (100 µM)	2	0.5 µM
P5-pLSmP-5bc-i# (100 µM)	2	0.5 µM
UltraPure distilled H_2O	96	
Total volume	400	

133 Split the premixture into 8 PCR tubes (50 µL per tube).
134 Run the PCR reaction as follows.

Cycle no.	Denature	Anneal	Extend
1	98 °C, 1 min		
2–4 (3 cycles)	98 °C, 10 s	60°C, 30 s	72 °C, 1 min
5			72 °C, 5 min

135 After the PCR reaction, combine all eight tubes for each sample in a DNA LoBind tube.
136 Purify the DNA, using 1.8 volumes (700 µL) of HighPrep PCR reagent and following Steps 6–14.
137 Add 60 µL of Buffer EB to the beads and elute the DNA by pipetting and vortexing.
138 Place the tube on the magnet for 1–2 min and transfer 55–58 µL of the eluate to a LoBind tube. Store the tubes on ice.

139 Set up the preliminary PCR reaction for each sample in a 96-well qPCR plate. This run finds the number of PCR cycles required for the following second-round PCR reaction.

Reagent	Volume (µL)	Final conc.
First-round PCR product	5	
NEBNext High-Fidelity 2× PCR Master Mix	10	1×
P7 (100 µM)	0.1	0.5 µM
P5 (100 µM)	0.1	0.5 µM
SYBR Green I nucleic acid gel stain (100×)	0.1	1×
Ultrapure distilled H_2O	4.7	
Total volume	20	

140 Run the qPCR reaction using a qPCR instrument as follows.

Cycle no.	Denature	Anneal	Extend
1	98 °C, 1 min		
2–31 (30 cycles)	98 °C, 10 s	60 °C, 30 s	72 °C, 1 min

141 On the basis of the raw amplification curve of each sample, determine the number of cycles at which the amplification nearly plateaus for each sample.

142 Set up the second-round PCR reaction for each sample as follows. Because it is expected that the numbers of cycles required for DNA and RNA samples will be different, we recommend running the PCRs for them separately. For a low-complexity library (0.5–2 million total barcodes), use the amounts given in the table below. For a high-complexity library (8–12 million total barcodes), we recommend multiplying all amounts in the following table by 4.

Reagent	Volume (µL)	Final conc.
First-round PCR product	50	
NEBNext High-Fidelity 2× PCR Master Mix	100	1×
P7 (100 µM)	1	0.5 µM
P5 (100 µM)	1	0.5 µM
Ultrapure distilled H_2O	48	
Total volume	200	

143 Split the premixture into 5 PCR tubes (40 µL per well).

144 Run the PCR reaction with the cycle number determined by the preliminary run (Step 141) as follows:

Cycle no.	Denature	Anneal	Extend
1	98 °C, 1 min		
2–X (X cycles)	98 °C, 10 s	60 °C, 30 s	72 °C, 1 min

145 Combine each sample into a DNA LoBind tube.

146 Purify the DNA using 1.8 volume (360 µL) of HighPrep PCR reagent following Steps 6–14.

147 Elute the DNA in 20 µL EB.

148 Measure the DNA concentration using a NanoDrop spectrophotometer.

149 Pool three replicates of DNA samples or RNA samples using an equal amount (1 µg) from each replicate, using Supplementary Table 4. Keep DNA and RNA samples separated until these are pooled at the later step (Step 160).

■ **PAUSE POINT** DNA can be stored at −20 °C for months.

150 Add equal volume of 6x gel loading dye (final conc. 3×) and mix the solution by vortexing.

151 Run the pooled sample on a 1.8% (wt/vol) agarose gel (30 mL of 5 cm ×6 cm mini gels with 1.3 cm-width well).

152 Cut the DNA bands (162 bp) using a blue-light Safe Imager.
153 Purify the DNA from the gel slices using QIAquick Gel Extraction Kit according to the manufacturer's protocol. Use one column for each of DNA or RNA samples.
154 Elute the DNA in 50 µL Buffer EB per column.
155 Purify the DNA using 1.8 volume (90 µL) of HighPrep PCR reagent following Steps 6–14.
156 Add 20 µL of Buffer EB to the beads and elute the DNA by pipetting and vortexing.
157 Place the tube on the magnet for 1-2 min and transfer 18 µL of the eluate to a LoBind tube.
158 Measure the DNA concentration using a Qubit dsDNA HS Assay Kit according to manufacturer's protocol.
159 To check the DNA size and quality, run 20-30 ng of the DNA on a 1.8% (wt/vol) gel along with 100-bp DNA ladder.
160 Pool the DNA and RNA samples in a single LoBind tube with 1:3 ratio to obtain 100 µL mixture at the final concentration of 10 nM (1 ng/µL), using Supplementary Table 4.

■ **PAUSE POINT** The pooled DNA can be stored at −20 °C for months.

161 Send the sequencing library and custom primers for sequencing (Fig. 1l, Extended Data Fig. 1e, Supplementary Table 3). The sequencing should be done with paired-end 15 bp (barcode length), 16 cycles for index read 1 and 10 cycles for index read 2. Index read 1 provides the UMI sequence, and index read 2 provides the sample index (let the sequencing facility know that the index read 2 should be used for demultiplexing and that short reads should not be masked, bcl2fastq parameters: `--minimum-trimmed-read- length 0 --mask-short-adapter-reads 0`). An average of 10×(DNA) and 30×(RNA) coverage of the library (based on number of barcodes) via sequencing reads is required. For example, we use an Illumina NextSeq high-output run (400M reads over 3 replicates) for the 0.5M barcode library or three runs (1.2B reads over three replicates) for the 8–12 million barcode library (Supplementary Table 1). Sequencing cycles, primers, and expected output for each read is shown below.

Read	Cycle	Primer	Output
Read 1	15	pLSmP-ass-seq-ind1	Barcode (forward)
Read 2	15	pLSmP-bc-seq	Barcode (reverse)
Index read 1	16	pLSmP-UMI-seq	UMI
Index read 2	10	pLSmP-5bc-seq-R2	Sample index

Data processing ● Timing total process time 4 h–4 d, depending on read depth; around 1 h hands-on time

162 Ensure that your computer is running Linux; we conducted testing on centOS 7.
163 Install conda (https://docs.conda.io/en/latest/miniconda.html) if it is not already on your system.
164 Clone the repository by typing the following command into a terminal window:

```
git clone https://github.com/shendurelab/MPRAflow.git
```

165 Once the repository is cloned, change directory into the configuration folder named 'conf' in the repository and create the conda environment in your system; it will automatically install all packages needed to run the pipeline. Use the following commands:

```
cd MPRAflow
conda env create -n MPRAflow -f environment.yml
```

▲ **CRITICAL STEP** Ensure that the Nextflow config file is set up correctly for your system. This pipeline comes with a nextflow.config file set up to run in local mode. Note that running the pipeline in local mode will not take advantage of certain parallelization steps that have been implemented with Nextflow. We recommend running this software on HPC clusters, allowing each process to be run as a separate 'qsub' command. The config file contains example code for Sun Grid Engine (SGE)/Univa Grid Engine (UGE) scheduler; further documentation on supported executors can be found on the Nextflow website (https://www.nextflow.io/docs/latest/executor.html). Remove the '\\' from the architecture you would like to use and place '\\' in front of any architectures not currently in use. A '\\' on every line of the configuration file runs the pipeline on your local

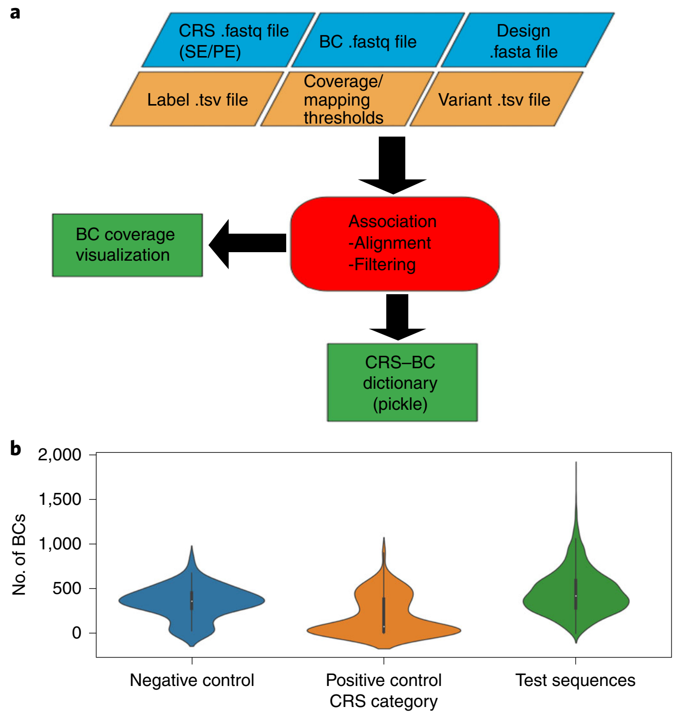


Fig. 2 | Overview of MPRAflow association utility. a, Mandatory inputs (blue), optional flags (orange), output files (green) and utility (red). The program requires .fastq files for the insert, either single-end (SE) or paired-end (PE) reads, and a design file, which is a .fasta file containing the synthesized oligonucleotides. The user can also specify a tab-delimited file with a mapping of CRS names given in the design file and a grouping, such as a control category (e.g., positive or negative control), a .tsv file of variants in the ordered oligonucleotide pool to be used for a tailored alignment strategy, and can accept various parameters for filtering the pairing based on mapping qualities and number of observed barcodes mapping to the CRS. The program outputs a Python dictionary in pickle format, mapping barcodes to their CRS. **b**, A violin plot of barcode coverage for each enhancer, grouped by labels provided in the label .tsv file. The violin plot features a kernel density (blue, yellow, and green), showing the underlying distribution of the data, and a boxplot. In the boxplot, the white dot is the median, the box represents the interquartile range (IQR), and the whiskers are 1.5 × IQR. Outliers are represented as points. BC, barcode.

machine. Consult your cluster's Wiki page for cluster-specific commands and change ‘clusterOptions’ to reflect these specifications. In addition, for large libraries, more memory can be specified under ‘clusterOptions’ using your system's equivalent of the ‘-mem_free’ option.

166 Run the association utility. We will use the data in GSE142696 (ref. [15]). The user must specify the barcode, CRS read 1 and CRS read 2 (if applicable) .fastq files, and a .fasta file containing a synthesized CRS oligonucleotide pool. (Fig. 2a) The most basic command for single-end sequencing of the CRSs is shown below. However, there are many optional flags that enable customization of the pipeline as outlined in Table 1. The program will return a dictionary of CRS-barcode pairs and a visualization of the number of barcodes associated with each enhancer (Fig. 2b). If an HPC cluster with a queuing system (e.g., SGE/UGE) is available, we recommend enabling Nextflow to submit commands to the queue. Program runtimes will scale with the size of data (Extended Data Fig. 2a). The command to download the data and run the program is shown below:

```
mkdir -p Assoc_Basic/data
cd Assoc_Basic/data
# download data
wget  ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM4237nnn/GSM4237954/
suppl/GSM4237954_9MPRA_elements.fa.gz
# trim PRC handle
zcat GSM4237954_9MPRA_elements.fa.gz |awk ‘{ count+=1; if (count == 1)
{ print } else { print substr($1,1,171)}; if (count == 2) { count=0 } }' >
design.fa
cd <path/to/MPRAflow>/MPRAflow
# run MPRAflow association
```

Table 1 | Association utility options

Option	Description
--fastq-insert	Full path to library association .fastq file for insert (must be surrounded with quotes)
--variants	Tab-separated values (.tsv) file with reference name, variant positions, ref bases, alt bases; only input for variant analysis workflow
--fastq-bc	Full path to library association .fastq file for barcode (must be surrounded with quotes)
--design	Full path to .fasta file of ordered oligonucleotide sequences (must be surrounded with quotes)
--name	Name of the association. Files will be named after this
--fastq-insertPE	Full path to library association .fastq file for read 2 if the library is a paired-end library (must be surrounded with quotes)
--min-cov	Minimum coverage of barcode to count it (default = 3)
--min-frac	Minimum fraction of barcodes to map to single insert (default = 0.5)
--mapq	Map quality (default = 30)
--baseq	Base quality (default = 30)
--cigar	Require exact match, for example, 200 million (default = none)
--outdir	The output directory where the results will be saved and what will be used as a prefix (default = outs)
--w	Specific name for work directory (default = work)
--with-timeline	Creates HTML file showing processing times
--split	Number of read entries per .fastq chunk for faster processing (default = 2000000)
--labels	.tsv file with the oligonucleotide pool .fasta file and a group label (for example, positive_control); if no labels are desired, a file will be automatically generated
--h, --help	Help message

Boldface options are mandatory; the others are optional.

```
conda activate MPRAflow
nextflow run --w <path/to/Basic>/Assoc_Basic/work association.nf
--name SRR10800986 --fastq-insert "<path/to/Basic>/Assoc_Basic/data/
SRR10800986_1.fastq.gz" --fastq-insertPE "<path/to/Basic>/Assoc_Basic/
data/SRR10800986_3.fastq.gz" --fastq-bc "<path/to/Basic>/Assoc_
Basic/data/SRR10800986_2.fastq.gz" --design "<path/to/Basic>/Assoc_
Basic/data/design.fa"
```

167 If the user creates a library focused on quantifying differential activity of single-nucleotide variants, the user can specify the option `--variants`, which allows the user to specify a tab-separated values (.tsv) file containing the reference name, variant position, reference and alternative bases for sensitive mapping. Below is an example of an expected .tsv format:

```
reference_name variant_positions ref_bases alt_bases
ref_1          130               A         T
ref_2          108               G         A
ref_3          67,99             A,C       C,T
```

168 Create an experiment .csv file. This file will contain the relevant information for the experiment, such as condition, replicates, .fastq file names, and an output prefix. The appropriate file for GSE142696 (ref. [15]) is:

```
Condition,Replicate,DNA_R1,DNA_R2,DNA_R3,RNA_R1,RNA_R2,RNA_R3
HEPG2,1,SRR10800881_1.fastq.gz,SRR10800881_2.fastq.gz,
SRR10800881_3.fastq.gz,SRR10800882_1.fastq.gz,SRR10800882_2.fastq.gz,
SRR10800882_3.fastq.gz
HEPG2,2,SRR10800883_1.fastq.gz,SRR10800883_2.fastq.gz,
SRR10800883_3.fastq.gz,SRR10800884_1.fastq.gz,SRR10800884_2.fastq.gz,
SRR10800884_3.fastq.gz
HEPG2,3,SRR10800885_1.fastq.gz,SRR10800885_2.fastq.gz,
SRR10800885_3.fastq.gz,SRR10800886_1.fastq.gz,SRR10800886_2.fastq.gz,
SRR10800886_3.fastq.gz
```

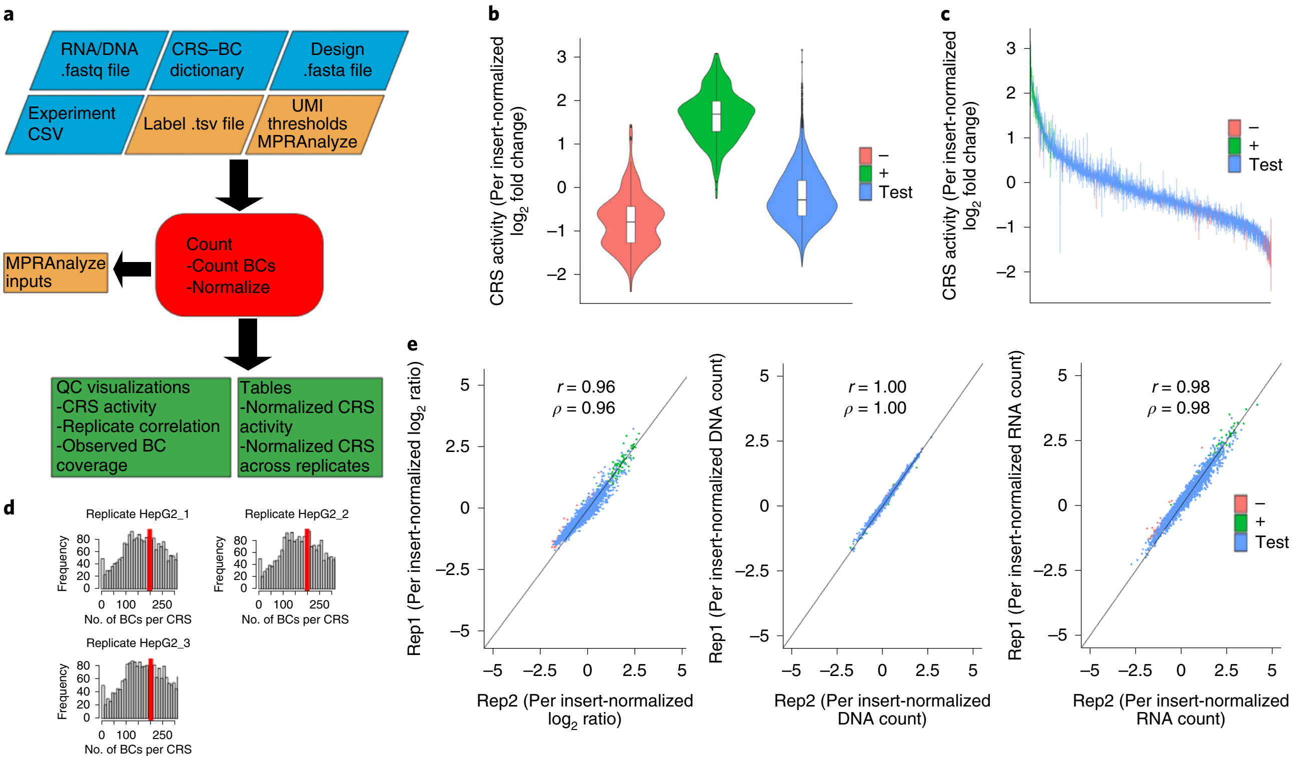


Fig. 3 | Overview of count utility. a, Mandatory inputs (blue), optional flags and outputs (orange), output files (green) and utility (red). The user must specify the directory containing all .fastq files for the RNA and DNA sequencing, the CRS-barcode dictionary from the association utility, a design file (.fasta file containing the synthesized oligonucleotides), and an experimental comma-separated file (CSV) outlining the number of replicates and conditions used. The user can also specify a tab-delimited file with a mapping of CRS names given in the design file and a grouping, such as control category (e.g., positive or negative control), and tune parameters, for example, to specify whether a UMI was used or whether the user would like to generate the input files for MPRAnalyze. **b**–**e**, The program will produce normalized activity of each CRS from each replicate, as well as across replicates, along with several visualizations. **b**, CRS activity normalized by insert and grouped by label determined in the label file. The violin plot features a kernel density, showing the underlying distribution of the data and a boxplot. In the boxplot, the center line is the median, the box represents the interquartile range (IQR), and the whiskers are 1.5 × IQR. Outliers are represented as points. **c**, Normalized activity of each CRS across replicates, colored by label and represented as a boxplot across replicates, where the box represents the IQR, and the whiskers are 1.5 × IQR. Outliers are represented as points. **d**, Distribution of observed barcode coverage per CRS in each replicate. The mean number of barcodes tagging each CRS is shown in red. **e**, Correlation of normalized log$_2$(RNA/DNA), DNA counts and RNA counts. BC, barcode.

169 Run the count utility for GSE142696 (ref. [15]). The user must specify the following: the directory containing the DNA and RNA sequencing data (with names specified in the experiment.csv file), a Python dictionary in pickle format from the association utility, the .fasta file containing synthesized the CRS oligonucleotide pool, and the experiment .csv file (Fig. 3a). There is some flexibility provided for the user at this stage, as outlined in Table 2. Below is an example of the minimal command. If an HPC cluster with a queuing system is available, we recommend submitting this command to the queue. This command will output normalized activity tables as well as QC visualizations (Fig. 3b–e). Program runtimes will scale with the size of data (Extended Data Fig. 2b).

```
# download data
conda install sra-tools
mkdir -p Count_Basic/data
cd Count_Basic/data
prefetch SRR10800881 SRR10800882 SRR10800883 SRR10800884 SRR10800885 SRR10800886
fastq-dump --gzip --split-files SRR10800881 SRR10800882 SRR10800883 SRR10800884 SRR10800885 SRR10800886
cd <path/to/MPRAflow>/MPRAflow
# run count.nf
conda activate MPRAflow
```

Table 2 | Count utility options

Option	Description
--dir	.fasta directory (must be surrounded with quotes)
--association	pickle dictionary from library association process
--design	.fasta file of ordered insert sequences
--e, --experiment	Experiment .csv file
--labels	.tsv file with the oligonucleotide pool .fasta file and a group label (e.g., positive_control); a single label will be applied if a file is not specified
--outdir	The output directory where the results will be saved (default = outs)
--bc-length	Length of barcode (default = 5)
--umi-length	Length of UMI when given (default = 10)
--no-umi	Flag if no UMI was used in the experiment
--merge-intersect	Only retain barcodes in RNA and DNA fractions (TRUE/FALSE, default = FALSE)
--mpranalyze	Flag to generate only MPRAnalyze outputs
--thresh	Minimum number of observed barcodes to retain insert (default = 10)
--w	Specific name for work directory (default = work)
--with-timeline	Create HTML file showing processing times
--h, --help	Help message

Boldface options are mandatory; the others are optional.

```
nextflow run --w <path/to/Basic>/Count_Basic/work count.nf --experiment-file "<path/to/Basic>/Count_Basic/data/experiment.csv" --dir "<path/to/Basic>/Count_Basic/data" --outdir "<path/to/Basic>/Count_Basic/output"
```

170 Add the flag `--mpranalyze` to the command above to produce the inputs for MPRAnalyze, a tool that determines MPRA activity using generalized linear models[16].

171 If the library was created using a saturation mutagenesis approach, first create an assignment file, which is a .tsv file of barcodes followed by the variants in the CRS in the format `CRS_name:position:reference>alternative`. An example is found below:

```
AAAAAACTAATACCA IRF6:104:A>G IRF6:408:G>A
AAAAAAGCAGGAACA IRF6:66:A>G IRF6:373:T>A
AAAAAAGCATTCTGT IRF6:371:G>T IRF6:510:G>A IRF6:560:C>T
AAAAACACTACTGGT IRF6:326:C>T IRF6:509:T>A
```

172 Create a specific saturation mutagenesis experiment.csv file for conditions, replicates and counts. The following is an example:

```
Condition,Replicate,COUNTS
condition1,1,cond1_1_counts.tsv.gz
condition1,2,cond1_2_counts.tsv.gz
condition2,1,cond2_1_counts.tsv.gz
condition2,2,cond2_2_counts.tsv.gz
```

173 To run the saturation mutagenesis utility, specify the directory containing the count tables. These tables can be created with the count workflow, called <condition>_<replicate>_counts.tsv.gz, and are saved in the folder <outdir>/<condition>/<replicate>/. The assignment file is generated in Step 171 and a specific saturation mutagenesis experiment.csv file is needed (Fig. 4a).

174 There is some flexibility provided for the user at this stage, as outlined in Table 3. Below is an example of the minimal command using example data for the telomerase reverse transcriptase (*TERT*) promoter from Kircher et al.[20], already present in the MPRAflow git repository. If an HPC cluster with a queuing system is available, we recommend submitting this command to the queue.

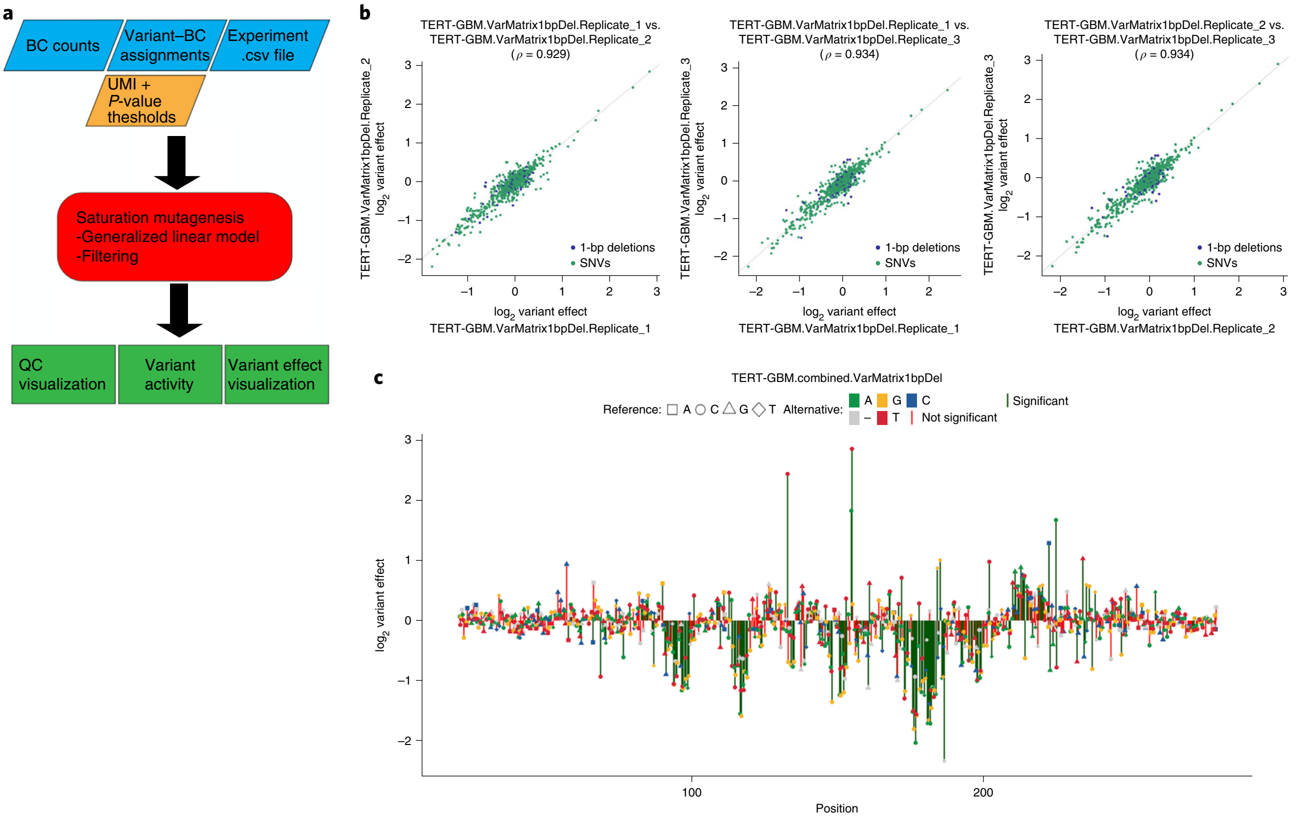


Fig. 4 | Overview of saturation mutagenesis utility. a, Mandatory inputs (blue), optional flags and outputs (orange), output files (green), and utility (red). The user must specify the directory containing all barcode count files, including DNA and RNA counts, the variant to barcode assignment file, and an experimental comma-separated file outlining the number of replicates and conditions used. The user can also set UMI and *P*-value thresholds to be used for filtering variants and distinguishing between significant and not-significant variant effects. The program will produce $\log_2$ variant effects, *P* values and a visual output of correlation, as well as a saturation mutagenesis variant effect plot of the region. **b**, Correlation between replicates. Here, we show the correlation between three replicates of the *TERT* promoter in a glioblastoma cell line from Kircher et al. 2019 (ref. [20]). ρ is the Pearson correlation between two samples (model with 1-bp indels). Only variants with ≥10 barcodes are shown. **c**, Saturation mutagenesis effect plot of the combined model from three replicates of the *TERT* promoter in a glioblastoma cell line from Kircher et al. 2019 (including 1-bp indels). 'Position' refers to the variant position of the original target insert. Only variants with ≥10 barcodes are shown. Significance level is $P < 1 \times 10^{-5}$.

Table 3 | Saturation mutagenesis utility options

Option	Description
--dir	Directory of count files (must be surrounded with quotes)
--assignment	Variant assignment file
--e,--experiment	Experiment .csv file
--outdir	The output directory where the results will be saved (default = outs)
--thresh	Minimum number of observed barcodes to retain insert (default = 10)
--pvalue	Minimum *P* value for significant variant effects (default = 1e-5)
--w	Specific name for work directory (default = work)
--with-timeline	Create HTML file showing processing times
--h, --help	Help message

Boldface options are mandatory; the others are optional.

```
conda activate MPRAflow
nextflow run saturationMutagenesis.nf --dir "examples/saturation
Mutagenesis/" --assignment "examples/saturationMutagenesis/TERT.
variants.txt.gz" --e "examples/saturationMutagenesis/experiment.csv"
-outdir "examples/saturationMutagenesis/output"
```

175 The outputs of the saturation mutagenesis utility are $\log_2$ variant effects for each replicate, as well as a joint estimate of activity across all replicates. In addition, correlation plots between replicates and variant effect profile plots spanning the whole region are generated (Fig. 4b,c). Further documentation with detailed descriptions of input and output files, as well as more examples, can be found at https://mpraflow.readthedocs.io/.

Troubleshooting

Troubleshooting advice can be found in Table 4.

Table 4 | Troubleshooting table

Step	Problem	Possible reason	Solution
Step 29	Low DNA yield. At least 250 ng of insert DNA is required for the recombination reaction	DNA amplification was not enough. DNA loss during gel extraction	Multiply the PCR reaction or increase the number of cycles in the second-round PCR to 15. More cycles (>15 cycles) can decrease the library complexity
Step 37	Uncut vector DNA appears on the gel	Insufficient restriction enzyme reaction	Perform restriction digestion twice or three times (Steps 30–36)
Step 61	Contamination with empty vectors	Vector linearization and/or I-*Sce*I digestion were not sufficient	Lower rates of empty vector contamination (<10%, one or two out of 16 colonies) are acceptable. In this case, proceed with the protocol. If the rate is >10%, redo vector linearization with a longer incubation time and make sure you have achieved complete linearization using an agarose gel. Perform I-*Sce*I digestion with a longer incubation time
	Mutation and indels observed in the plasmids	These can be derived from synthesis/PCR/sequencing errors	As these errors are unavoidable, we usually observe that >50% of sequences contain mutations and/or deletions. Proceed with the protocol; these erroneous sequences should be ruled out during the analysis step. Synthesis error rates might be improved by ordering oligonucleotides that are synthesized with high-fidelity from the manufacturer
Box 2, step 5	Low infection efficiency	Polybrene concentration may not be appropriate	Optimization of Polybrene concentration may be required. Seed cells in a 24-well plate, infect with control virus, along with different amounts of Polybrene (e.g., 0, 2, 4, 8, 16, 32 μg/mL at final concentration), and observe cell death and GFP expression. In our experience, 8 μg/mL works well for most cell types, including HepG2 cells, K562 cells, H1 hESCs (human embryonic stem cells), and WTC11 iPSCs (induced pluripotent stem cells). Polybrene kills neural cell types, including neural progenitors, and should be avoided when using those types of cells

Timing

The timing for Steps 71–83 and 131–161 includes sequencing time, which can vary from x x to y y.
Steps 1–29, library amplification: 3 h
Steps 30–37, vector linearization: 7 h to overnight
Steps 38–61, recombination and electroporation: 3 d (5 h hands-on time)
Steps 62–70, colony count and plasmid library prep: 3 h
Steps 71–83, sequencing for CRS–barcode association: 2–4 weeks (3 h hands-on time plus sequencing turnaround time)
Steps 84–103, lentivirus packaging: >1 week (>5.5 h hands-on time)
Steps 104–106, lentivirus titration: >1 week (3 h hands-on time)
Steps 107–121, lentivirus infection and DNA/RNA extraction: >1 week (>3.5 h hands-on time)
Steps 122–130, reverse transcription: 4 h
Steps 131–161, library prep and sequencing for RNA and DNA barcode count: 2–4 weeks (6 h hands-on time plus sequencing turnaround time)
Steps 162–175, data processing: 4 h–4 d, depending on read depth (Extended Data Fig. 2)
Box 2, test infection: >1 week (3 h hands-on time)

Anticipated results

The output of a lentiMPRA experiment will consist of two sets of data: association sequencing and DNA/RNA barcode sequencing. Success of association sequencing preparation can be assessed by the size of the band (419 bp) observed during library preparation. The association sequencing should contain paired-end reads that cover the CRS (200 bp) and an index read to cover the barcode (15 bp). The recommended sequencing depth will vary significantly with the complexity of the library being tested, but we generally suggest 10 reads per unique barcode expected. MPRAflow's association utility should be run on this dataset to determine the number of barcodes per CRS (Fig. 2b). Generally, we aim for 50–200 barcodes per CRS; libraries with >600 barcodes per CRS should be cloned again because integration and sequencing will limit the coverage of the library and the sensitivity of the experiment.

The quality of the preparation of the DNA and RNA barcode sequencing library can be assessed by the size of the band (162 bp). The sequencing results should contain paired-end reads that cover the barcode (15 bp) and an index read for a UMI (16 bp). These files should be demultiplexed and run through MPRAflow's count utility. This will return normalized count tables for all experimental conditions and replicates tested, as well as a final table of activity of each CRS normalized across replicates. A broad overview of activity can be seen by user-defined categories (Fig. 3b,c), allowing for assessment of control sequences. Averaged observed barcodes per CRS can be checked through histograms to verify coverage of the barcodes (Fig. 3d). In addition, the correlations between technical replicates are shown for DNA count, RNA count and $\log_2$(RNA/DNA) (Fig. 3e). A successful experiment will allow the user to determine which CRSs increase transcriptional activity and which do not. To determine the active sequences, we can compare our test sequences with scrambled controls. These scrambled sequences provide a null distribution that can be used for robust statistical testing.

Reporting Summary

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

Data availability

A 5′ lentiMPRA dataset conducted in HepG2 cells[15] has been deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession no. GSE142696.

Code availability

The source code is freely available at https://github.com/shendurelab/MPRAflow.

References

1. Chatterjee, S. & Ahituv, N. Gene regulatory elements, major drivers of human disease. *Annu. Rev. Genomics Hum. Genet* **18**, 45–63 (2017).
2. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
3. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
4. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol.* **3**, e245 (2005).
5. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein- DNA interactions. *Science* **316**, 1497–1502 (2007).
6. Crawford, G. E. et al. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl Acad. Sci. USA* **101**, 992–997 (2004).
7. Sabo, P. J. et al. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl Acad. Sci. USA* **101**, 4537–4542 (2004).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
9. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.* **13**, 1006–1019 (2018).
10. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
11. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **10**, 159–164 (2015).

12. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
13. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
14. Inoue, F. et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).
15. Klein, J. et al. A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. Preprint at *bioRxiv* https://doi.org/10.1101/576405 (2019).
16. Ashuach, T. et al. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183 (2019).
17. Anaconda software distribution v.2–2.4.0 (Anaconda, 2016).
18. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* **25**, 713–727.e710 (2019).
19. Ryu, H. et al. Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. Preprint at *bioRxiv* https://doi.org/10.1101/256313 (2018).
20. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
21. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
22. Georgakopoulos-Soares, I., Jain, N., Gray, J. M. & Hemberg, M. MPRAnator: a web-based tool for the design of massively parallel reporter assay experiments. *Bioinformatics* **33**, 137–138 (2017).
23. Ghazi, A. R. et al. Design tools for MPRA experiments. *Bioinformatics* **34**, 2682–2683 (2018).
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
25. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
26. Klein, J. C. et al. Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2015).

Acknowledgements

This work was supported by National Human Genome Research Institute grants 1UM1HG009408 (N.A. and J.S.) and 1R21HG010065 and 1R21HG010683 (N.A.), as well as a Ruth L. Kirschstein Predoctoral Individual National Research Service Award 1F31HG011007 (M.G.G.), an NRSA NIH fellowship 5T32HL007093 (V.A.), National Institute of Mental Health grants 1R01MH109907 and 1U01MH116438 (N.A. and K.S.P.), and the Uehara Memorial Foundation (F.I.). J.S. is an investigator of the Howard Hughes Medical Institute.

Author contributions

F.I. and B.M. developed lentiMPRA; R.Z. assisted in developing lentiMPRA; M.G.G., M.S., V.A., S.W., S.F., J.Z., T.A., A.K., I.G.-S., N.Y., C.J.Y., K.S.P., M.K., J.S. and N.A. assisted in developing MPRAflow; and all authors contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41596-020-0333-5

Correspondence and requests for materials should be addressed to F.I. or J.S. or M.K. or N.A.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 8 January 2020; Accepted: 17 April 2020;
Published online: 08 July 2020

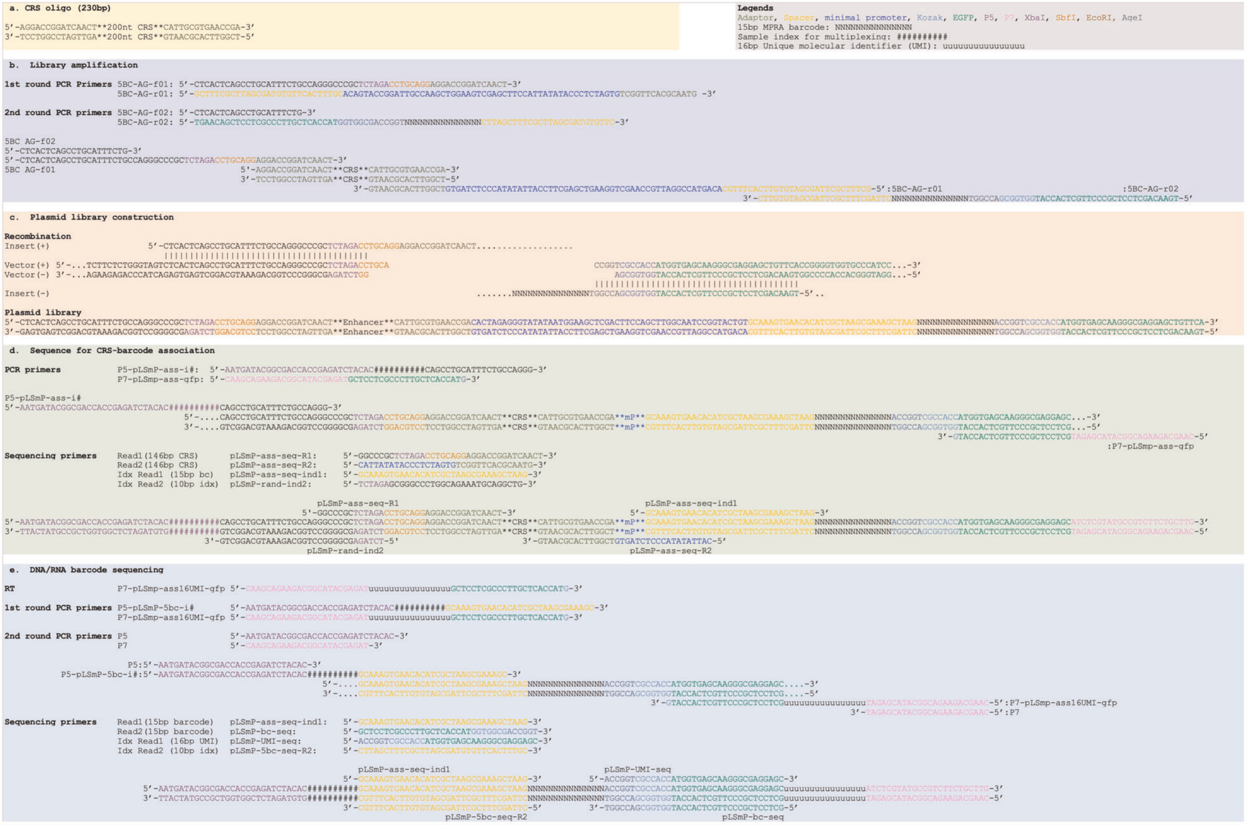
Related links

Key references using this protocol

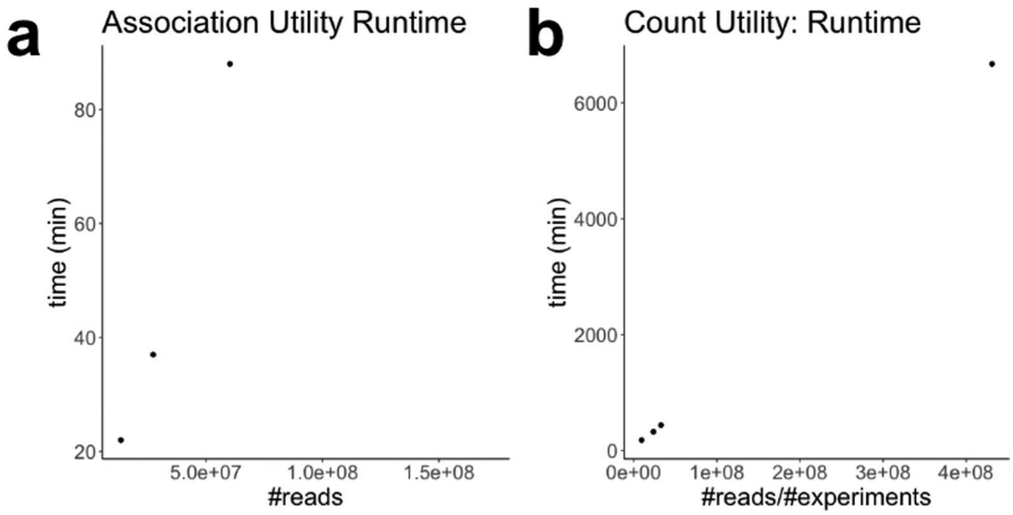
Inoue, F. et al. *Genome Res.* **27**, 38–52 (2017): https://doi.org/10.1101/gr.212092.116
Klein, J. et al. Preprint at *bioRxiv* 576405 (2019): https://doi.org/10.1101/576405
Kircher, M. et al. *Nat. Commun.* **10**, 3583 (2019): https://doi.org/10.1038/s41467-019-11526-w

Key data used in this protocol

Klein, J. et al. Preprint at *bioRxiv* 576405 (2019): https://doi.org/10.1101/576405


a. CRS oligo (230bp)
5′-AGGACCGGATCAACT**200nt CRS**CATTGCGTGAACCGA-3′
3′-TCCTGGCCTAGTTGA**200nt CRS**GTAACGCACTTGGCT-5′
Legends
Adaptor, Spacer, minimal promoter, Kozak, EGFP, P5, P7, XbaI, SbfI, EcoRI, AgeI
15bp MPRA barcode: NNNNNNNNNNNNNNN
Sample index for multiplexing: ##########
16bp Unique molecular identifier (UMI): uuuuuuuuuuuuuuuu
b. Library amplification
1st round PCR Primers
5BC-AG-f01: 5′-CTCACTCAGCCTGCATTTCTGCCAGGGCCCGCTCTAGACCTGCAGGAGGACCGGATCAACT-3′
5BC-AG-r01: 5′-GCTTTCGCTTAGCGATGTGTTCACTTTGCACAGTACCGGATTGCCAAGCTGGAAGTCGAGCTTCCATTATATACCCTCTAGTGTCGGTTCACGCAATG -3′
2nd round PCR primers
5BC-AG-f02: 5′-CTCACTCAGCCTGCATTTCTG-3′
5BC-AG-r02: 5′-TGAACAGCTCCTCGCCCTTGCTCACCATGGTGGCGACCGGTNNNNNNNNNNNNNNNCTTAGCTTTCGCTTAGCGATGTGTTC-3′
5BC AG-f02
5BC AG-f01
:5BC-AG-r01
:5BC-AG-r02
c. Plasmid library construction
Recombination
Insert(+)
Vector(+)
Vector(-)
Insert(-)
Plasmid library
d. Sequence for CRS-barcode association
PCR primers
P5-pLSmP-ass-i#
P7-pLSmp-ass-gfp
Sequencing primers
Read1 (146bp CRS) pLSmP-ass-seq-R1
Read2 (146bp CRS) pLSmP-ass-seq-R2
Idx Read1 (15bp bc) pLSmP-ass-seq-ind1
Idx Read2 (10bp idx) pLSmP-rand-ind2
:P7-pLSmp-ass-gfp
e. DNA/RNA barcode sequencing
RT
P7-pLSmp-ass16UMI-gfp
1st round PCR primers
P5-pLSmP-5bc-i#
2nd round PCR primers
P5
P7
Sequencing primers
Read1 (15bp barcode) pLSmP-ass-seq-ind1
Read2 (15bp barcode) pLSmP-bc-seq
Idx Read1 (16bp UMI) pLSmP-UMI-seq
Idx Read2 (10bp idx) pLSmP-5bc-seq-R2
:P7-pLSmp-ass16UMI-gfp
:P7
pLSmP-UMI-seq
pLSmP-5bc-seq-R2
pLSmP-bc-seq

Extended Data Fig. 1 | Sequence scheme of lentiMPRA. a, Synthesized CRS oligo sequence. **b**, Primers and their binding in 1st and 2nd round PCR for library amplification. **c**, Recombination and plasmid library sequence. **d**, Primers and their binding in library amplification and sequencing for CRS–barcode association. **e**, Primers and their binding in reverse transcription, library amplification and sequencing for barcode counting.



Extended Data Fig. 2 | Time complexity study of MPRAflow. a, The Association Utility run time scales with number of reads when holding the number of FASTQ chunks at 2M reads. As this is an alignment the memory requirements are not trivial, requiring approximately 1GB of memory per 3M reads. **b**, The Count Utility run time scales with number of reads divided by the number of experiments running in parallel. This step does not require much memory, where 500M reads can be processed in <0.5GB.

natureresearch

Corresponding author(s): Fumitaka Inoue, Martin Kircher, Jay Shendure, Nadav Ahituv

Last updated by author(s): Feb 28, 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 Authors & Referees 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	Raw BCL files were processed with BCL2FASTQ to generate FASTQ files we analyzed in this study
Data analysis	All sequencing data was analyzed with v2.1 of MPRAflow, a pipeline developed in our lab for processing MPRA datasets. This code is freely available

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/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 sequencing data analyzed for the count and association utilities are available in the "Gene Expression Omnibus" accession code: GSE142696 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE142696), used to generate figure 2 and 3. The sequencing data analyzed for the saturation mutagenesis utility is available in the "Gene Expression Omnibus" accession code: GSE126550 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126550), used to generate figure 4. Both datasets are publicly available.

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	All MPRA experiments are performed with 3 technical replicates
Data exclusions	No data were excluded.
Replication	Each of the assays was performed with three replicates to ensure faithful reproduction of the assay.
Randomization	Not relevant, as we needed to know the identity of each sample prior to the analyses.
Blinding	Not relevant, as we needed to know the identity of each sample prior to the analyses.

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
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

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 (ATCC, HB-8065), 293T (ATCC, CRL-3216)
Authentication	Cells were not authenticated
Mycoplasma contamination	Cells were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	Not applicable