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

Gracie Gordon1,2,3,\*, Fumitaka Inoue1,2,\*,^, Beth Martin4,\*, Max Schubach5,6,\*, Vikram Agarwal4,5, Sean Whalen7, Shiyun Feng1,2, Jingjing Zhao1,2, Tal Ashuach8, Ryan Ziffra1,2,3, Anat Kreimer8, Ilias Georgakopoulous-Soares1,2, Nir Yosef8,9, Chun Ye1,2,10,11, Katherine S Pollard2,7,9,10, Jay Shendure4,12,13,^, Martin Kircher5,6,^, Nadav Ahituv1,2,^

1Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158, USA

2Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94158, USA

3Biological and Medical Informatics Graduate Program, University of California San Francisco, San Francisco, CA, USA.

4Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

5Calico Life Sciences LLC, South San Francisco, CA 94080, USA

5Berlin Institute of Health (BIH), 10178, Berlin, Germany

6Charité - Universitätsmedizin Berlin, 10117, Berlin, Germany

7Gladstone Institutes, San Francisco, CA 94158, USA

8Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California Berkeley, Berkeley, CA, USA

9Chan-Zuckerberg Biohub, San Francisco, CA 94158, USA

10Department of Epidemiology and Biostatistics and Institute of Computational Health Sciences, University of California San Francisco, San Francisco, CA, USA

11Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA

12Howard Hughes Medical Institute, Seattle, WA 98195, USA

13Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA 98195, USA

\* These authors contributed equally to the work.

^Correspondence to: F.I. ([fumitaka.inoue@ucsf.edu](mailto:fumitaka.inoue@ucsf.edu)), J.S. ([shendure@uw.edu](mailto:shendure@uw.edu)), M.K. ([martin.kircher@bihealth.de](mailto:martin.kircher@bihealth.de)), N.A. ([nadav.ahituv@ucsf.edu](mailto:nadav.ahituv@ucsf.edu))

**ABSTRACT**

Massively Parallel Reporter Assays (MPRAs) can simultaneously measure the function of thousands of candidate regulatory sequences (CRS) in a quantitative manner. In this method, CRS 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-seq and normalized for cellular integration via DNA sequencing of the barcode. This technology has been used to test hundreds of thousands of sequences and their variants for regulatory activity, decipher the regulatory code and its evolution and for the development of genetic switches. Lentivirus-based MPRA (lentiMPRA) allows the use of this technique in hard to transfect cells and also obtain ‘in genome’ readouts. Here, we provide a detailed protocol for lentiMPRA along with a user friendly and flexible computational pipeline, MPRAflow, for general MPRA analyses.

**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, or 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 disease1. For example, over 90% of all human disease genome-wide association studies (GWAS) have shown associations with noncoding variants2 and colocalize with potential gene regulatory elements3. In addition, gene regulatory elements can be major drivers of evolutionary speciation, driving differences between species such as morphology, diet, and behavior4. 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-seq5), assay for transposase-accessible chromatin using sequencing (ATAC-seq6), cleavage under targets and release using nuclease (CUT&RUN7), Hi-C8 and others. However, these methods only help annotate candidate regulatory sequences (CRSs), and additional experimental assays must be performed in order to functionally characterize them. Reporter assays are commonly used to characterize these sequences. 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 (MPRAs) overcome this hurdle, providing the ability to test hundreds of thousands of sequences and their variants in parallel for their regulatory function9. This is done either by pairing the CRS to a barcode, or by using the CRS itself as a barcode, as done in the self-transcribing active regulatory region sequencing (STARR-seq10) assay. The RNA is then measured under the premise that if the CRS is an active regulatory element then it will lead to transcription.

Here, we describe both a lentivirus-based MPRA (lentiMPRA) and MPRAflow, a computational tool for MPRA analysis based upon the Nextflow framework (**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, due to the viruses’ inherent genomic integration, it provides an ‘in genome’ readout, which we have shown to provide more robust results that can be better predicted by both biochemical and sequence-based features compared to episomal-based MPRA11. Finally, to analyze the results, we describe MPRAflow, 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: 1) Descriptive assays that detect potential regulatory elements (such as ChIP-seq, DNase, ATAC-seq, CUT&RUN, and Hi-C) identify candidate sequences within chromatin, yet most MPRAs analyze sequences in an episomal context; 2) episomal-based MPRA is limited to cells that can easily be transfected. Lentivirus-based assays overcome both these limitations. Lentiviruses integrate into the genome, providing an ‘in genome’ readout. In addition, they can infect a multitude 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 minimizes the distance between the CRS and barcode (102 bp) than previous 3′ UTR barcoding method (801 bp), reducing the risk of CRS-barcode swapping12. In addition, unlike previous lentiMPRA where each CRS is synthesized together with multiple barcodes in a custom array, the 5′ UTR barcoding strategy adds barcodes via the PCR primer. This allows the ability to clone and test hundreds of thousands of CRSs using lentiMPRA.

To subsequently analyze MPRA results, there are several home-brewed MPRA computational analyses pipelines tailored to each lab and MPRA technique. We thus developed MPRAflow, which provides a user friendly, flexible, parallelized tool for 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 MPRAnalyze13, a tool that calculates a transcription rate for each tested CRS by fitting a generalized linear model with DNA and RNA counts.

**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 SNPs, evolutionary variants) in the genome, decoding the regulatory code, 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 which we have shown is more reproducible and is more predictive of functionality than both biochemical annotations and sequence-based models10.

MPRAflow is written using nextflow, Python, Bash, and R; and can be parallelized with multiple computational architectures including High Performance Compute (HPC) clusters. As 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. Additionally, MPRAflow provides easily interpretable graphics and produces files correctly formatted for readily available tools for further in-depth bioinformatic analysis such as MPRAnalyze13.

**Comparisons with other methods**

There are several different varieties of MPRA, such as regular episomal-based MPRA, STARR-seq, and others9. lentiMPRA differs from these methods as it provides an ‘in-genome’ readout in a wider range of cell types. In STARR-seq, the candidate sequence itself acts as the barcode. This attribute can potentially impact results due to binding of RNA-associated factors and RNA stability of the assayed sequence12. Using over thirty (i.e., on average) 15-20nt barcodes per assayed sequence in lentiMPRA reduces this impediment. CRSs are usually generated via oligo synthesis, but can also be produced by other processes, such as PCR or DNA-capture based methods. Barcodes can be either added as part of the synthesis or via PCR, providing a flexibility in cloning design. As lentiviruses integrate throughout the genome, we introduced anti-repressors on either side of the virus that together with having over 30 barcodes per assayed sequences assist in overcoming differences due to varying genomic integration sites.

Previous MPRA processing tools have mainly focused on CRS library design and 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 work11,12,14,15 and contains two utilities: association and count. The association utility processes demultiplexed FASTQ files and assigns CRS-barcode pairs in the random pairing design. The count utility processes demultiplexed FASTQ files to perform QC across replicates, normalizes barcode count tables per CRS, and quantifies log2(RNA/DNA) ratios per CRS. MPRAnalyze inputs can also be produced using the count utility. Each utility is executed with a single command on a terminal and both 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, ATAC-seq, CUT&RUN, GWAS, Hi-C and others). Variants of interest within these CRS can be identified via GWAS, GTEx various genomic websites such as Genome Aggregation Database (gnomAD16), comparative genomics and many other databases. The CRSs and variants tested ultimately depends 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, 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 these data does 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 MPRAnator17 or MPRA Design Tools18 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, depending on how the CRS are generated (i.e., oligosynthesis, PCR, or capture).

Library generation

For this protocol, we will focus on oligosynthesis as it is currently the most cost effective way to generate fixed-length CRSs. Here, the synthesized oligo 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 pLS-SceI vector (Addgene 137725) 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 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 be 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 due to the finite nature of the multiplicity of infection (MOI) and the associated increase in sequencing costs. 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 index to the CRS-barcode pairs. The PCR product is then gel extracted at the appropriate insert size (~471 bp) and sent for PE146 sequencing with a 15 bp 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 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 three day incubation (to reduce for 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 then calculated for each viral concentration (**Supplementary Table 2**). These values are then plotted against the viral volume to calculate the viral titer. Conditions should not be used if cells are not viable.

Infection and sequencing

The lentiMPRA library is then infected into the cells of interest and incubated for three days. The number of cells required is determined based on the library complexity and the highest MOI that the cells can be infected with that isn’t toxic to the cells. It is highly recommended to carry out three technical replicates for each biological condition tested in order to assess reproducibility. The cells are then washed to reduce for non-integrating lentivirus and DNA and RNA are simultaneously extracted. For RNA, mRNA is purified (if necessary), treated with DNase and reverse transcription is done using barcode-specific primer that contains 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 and UMI to the barcode. The sequencing libraries are then pooled and sent for PE15 sequencing with a 16 bp UMI and 10 bp sample index.

Data processing

We built a computational tool, MPRAflow, to easily process demultiplexed FASTA data resulting from lentiMPRA and other MPRA experiments. If the barcodes are randomly paired with the CRS, 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 (BWA19) to align sequences to the ordered oligo pool and a workflow for libraries containing single nucleotide variants of the same CRS, using Bowtie220 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 log2(RNA/DNA) ratios for 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 MPRAnalyze13, a generalized linear model approach. Additionally, 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 tag, which are used to fit a multiple linear regression model of log2(RNAj) ~ log2(DNAj) + *N* + offset (j ∈ T) and report the coefficients of *N* as effects for each variant. The utility processes multiple replicates and conditions in parallel if a high-performance computing (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, only a basic familiarity with command line tools is needed.

**Limitations**

There are several limitations for lentiMPRA. This includes a limitation in the number of CRSs that can be tested in cells that are not amenable to high lentivirus concentrations, though that can be amended by using an increased number of cells. The use of oligosynthesis to generate the CRS library can also limit the number of sequences that can be tested including their length. Improvements in DNA synthesis can ultimately overcome this limitation as well as PCR or DNA capture-based methods. Techniques that allow for multiplex pairwise assembly of oligos21 could also be a way to increase CRS size by patching together specific oligonucleotides. As for MPRAflow, while this tool is applicable to many types of MPRA, it does not support STARR-seq workflows as this tool does not include functionality for peak calling.

**Materials**

Biological materials

* 293T cells (ATTC, CRL-3216)
* Cell lines of interest. All data shown in this protocol were generated from HepG2 cells (ATTC, HB-8065).

Reagents

* DMEM (Life Technologies, cat. no. 11995-065)
* FBS (VWR international, cat. no. 89510-194)
* Penicillin-Streptomycin (Life Technologies, cat. no. 15140-122)
* 0.05% Trypsin-EDTA (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 Laboratories, cat. no. 1725204)
* UltraPure DNase/RNase-Free Distilled Water (Life Technologies, cat. no. 10977-023)
* SurePrint 244K Oligonucleotide Libraries (Agilent, cat. no. G7223A)
* TE (Teknova, cat. no. T0225)
* NEBNext High-Fidelity 2X PCR Master Mix (New England Biolabs, cat. no. M0541L)
* Ethyl alcohol (Sigma-Aldrich, cat. no. E7023-500ML)
* EB (Qiagen, cat. no. 19086)
* HighPrep PCR reagent (MagBio Genomics, cat. no. AC60050)
* 6x gel loading dye (New England Biolabs, cat. no. B7025S)
* SeaKem LE agarose (Lonza, cat. no. 50004)
* Qiaquick Gel Extraction kit (Qiagen, cat. no. 28704)
* 10x Cutsmart (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 Broth Base (Life Technologies, cat. no. 12780029)
* 15 cm LB agar plates (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)
* 1kb ladder marker (New England Biolabs, cat. no. N3232S)
* 100 bp ladder marker (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)
* EndoFection (Genecopoeia, cat. no. EFL1001-01)
* psPAX2 (Addgene, cat. no. 12260)
* pMD2.G (Addgene, cat. no.12259)
* pLS-SV40-mP-EGFP (Addgene, cat. no. 137724)
* pLS-SceI (Addgene, cat. no. 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 Laboratories, cat. no. 1610710)
* Oligotex RNA mini kit (Qiagen, cat. no. 70022)
* TURBO DNA-free kit (Life Technologies, cat. no. AM1907)
* SuperScript II Reverse Transcriptase (Life Technologies, cat. no. 18064-071)

Equipment

* Pipettes
* Filter tips
* Serological pipettes
* Pipet-aid XP (Drummond, cat. no. 4-000-101)
* Cell culture plates (24-well, Genesee Scientific, cat. no. 25-107; 10 cm, Genesee Scientific, cat. no. 25-202; 15 cm, Genesee Scientific, cat. no. 25-203)
* Fluorescent inverted microscope
* CO2 incubator
* Hemocytometer
* DNA LoBind tube (Eppendorf, cat. no. 022431021)
* 8-strip PCR tubes (Axygen, cat. no. PCR-0208-FCP-C)
* Vortex mixer (Thermo Fisher Scientific, cat. no. 88880017)
* Nanodrop 8000 Spectrophotometer (Thermo Fisher Scientific, cat. no. ND-8000-GL)
* Qubit fluorometer (Life Technologies, cat. no. Q32857)
* qPCR instrument, QuantStudio 6 Flex Real-Time PCR Systems (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 drybath (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 measuring the temperature of water in a tube placed on the instrument using an alcohol thermometer.

* 1mm-gap cuvettes (Btx Harvard Apparatus, cat. no. 450124)
* Gemini X2 Electroporation System (Btx Harvard Apparatus, cat. no. 452007)
* 37°C shaker
* 14 mL round bottom tubes (Corning, cat. no. 352059)
* 50 mL tubes (Corning, cat. no. 352070)
* 37°C incubator
* T225 flasks (Corning, cat. no. 431082)
* Centrifuge (Eppendorf, cat. no. 022625501)
* 0.45 μm PES filter units (Thermo Fisher Scientific, cat. no. 165-0045)
* Cell lifters (Corning, cat. no. 3008)
* 3 mL luer lock syringes (BD, cat. no. 309657)
* 20G needles (BD, cat. no. 305179)

Software

* Conda (<https://docs.conda.io/en/latest/miniconda.html>)
* Linux

Reagent setup

DMEM (10% heat-inactivated FBS)

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

DMEM (5% heat-inactivated FBS)

Supplement DMEM with the heat-inactivated FBS at a final concentration of 5% (vol/vol) and 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 H2O. Store at room temperature (RT; 22–25 °C) for up to two weeks.

LB broth medium

Suspend 20 g of LB broth base in one liter of distilled water and sterilize in autoclave. Store at room temperature.

**Procedure**

**Box 1. Test infection in the cells to be used for lentiMPRA TIMING >4 d cell culture and transfection plus 3 h hands-on time**

1. Culture cells that will be used for lentiMPRA.
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 0.1M cells per well.
4. Incubate the cells overnight.
5. Refresh culture media by adding polybrene at a final concentration of 8 μg/mL.

**TROUBLESHOOTING**

**CAUTION** During the following procedure, the liquid and plastic waste needs to be discarded into 10% bleach, as they are contaminated with lentivirus.

1. Add 0, 1, 2, 4, 8, 16, 32, 64 μL of control lentivirus (e.g. pLS-SV40-mP-EGFP) to the wells. Lentivirus can be generated according to the protocol described below (steps 84-103).
2. The next day, refresh the culture media without polybrene. Roughly check the cell survival compared to the well without infection.
3. After two days (three days culture in total), check the cell survival and GFP expression under a fluorescent microscope. If there is substantial cell death in certain wells, do not proceed with them for the following genomic DNA extraction step.
4. Remove the culture media and wash the cells with 500 μL/well DPBS three times.
5. Extract genomic DNA from each well using Wizard SV Genomic DNA Purification System according to manufacturer’s protocol.
6. Measure the DNA concentration using a Nanodrop and dilute the DNA to 10 ng/μL in 8-strip PCR tubes.
7. Setup 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 | 1x |
| Forward primer (100 μM) | 0.1 | 1 μM |
| Reverse primer (100 μM) | 0.1 | 1 μM |
| Ultrapure distilled H2O | 2.3 |  |
| Total volume | 10 |  |

1. 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 sec | 60°C, 30 sec |  |
| 37 |  |  | 60°C-95°C in 15 min |

1. Determine the MOI by calculating the relative amount of viral DNA over genomic DNA with the subtraction of relative amount of backbone DNA, using **Supplementary Table 2.**
2. The resultant highest MOI is considered as the max MOI for the cells to be used, because this is the viral amount that gives high GFP expression without substantial cell death.

**CRITICAL** If the max 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 **Box2**, step 2.

**CRITICAL** We do not recommend an MOI of more than 100, even if cell death was not observed, because adding high amount of virus increases non-integrating virus left in the cells after three days of culture, which can increase background noise in DNA barcode amplification.

**Box 2. Library design criteria**

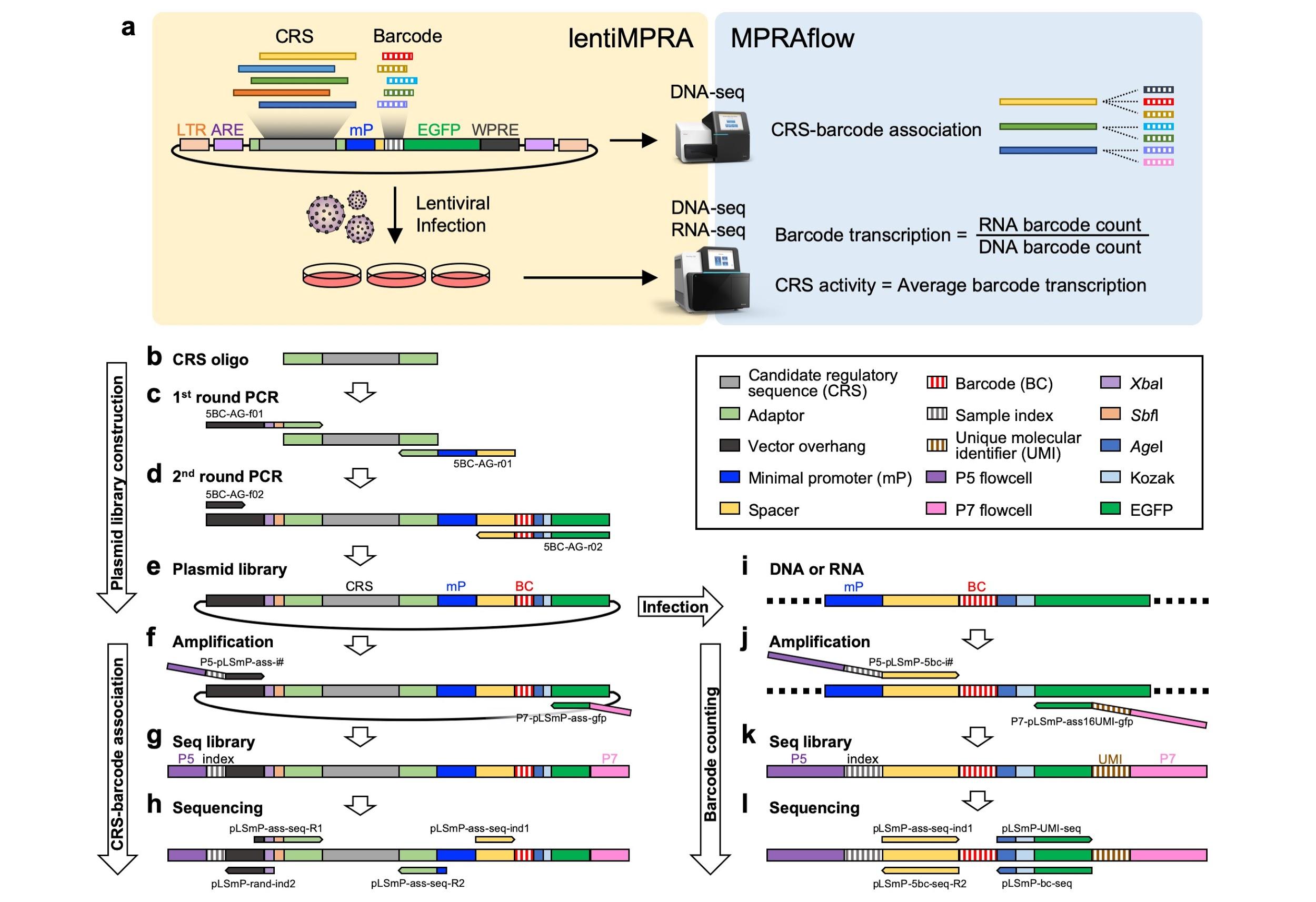
1. Oligosynthesis length via Agilent is currently limited in length to 230 nucleotides (**Fig. 1b, Supplementary Fig. 1a**). As the sequence should include 15-base common sequences (adaptors) on both sides of CRS to amplify the library via PCR, the length of the CRS should be 200 bp. Twist Bioscience currently provides oligosynthesis service of oligos up to 300 bp and could increase CRS length. Three pairs of the adaptor sequences are shown below. If two or three independent libraries are synthesized on the same array, use the second and third options.

5′-AGGACCGGATCAACT\*\*200base\_CRS\*\*CATTGCGTGAACCGA-3′

5′-AATGCTAGCGCATGG\*\*200base\_CRS\*\*CTGCAACCTACGGAA-3′

5′-TTACGAGCCGTAGTC\*\*200base\_CRS\*\*GCATCTCAACGTGGT-3′

1. The number of oligos that can be synthesized by Agilent is limited to 244,000 (Twist Bioscience 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. limited MOIs) and the number of cells available for the experiment (e.g. primary cells, hypoproliferative cells etc.), as explained in **Box 1**. 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.**
2. We recommend avoiding homopolymers (>8 base) from the design, as they may cause a higher synthesis error rate.



**Fig. 1 | Schematics of lentiMPRA.** **a**, Summary of lentiMPRA and MPRAflow. The lentiMPRA library is sequenced to associate between CRS and barcodes and infected into cells using three replicates. DNA and RNA from the cells is sequenced to obtain barcode transcription and CRS activity. LTR, long terminal repeat; ARE, anti-repressor element; WPRE, Woodchuck hepatitis virus posttranscriptional regulatory element. **b**, CRS oligo. 200-base CRS (grey) 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 and minimal promoter (mP, blue) and spacer sequences (yellow) downstream of the CRS oligo. **d**, Second round PCR. Reverse primer adds the barcodes (red stripe) and GFP complementary sequences (green). **e**, Plasmid construct. **f**, Amplification for CRS-barcode association. Primers adding P5 (purple) and sample index (grey stripe) 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 15 bp barcode and index read 2 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 16-bp unique molecular identifier (UMI, brown stripe) downstream. **k**, Sequencing library structure. **l**, Sequencing reaction. Paired-end 15 bp reads give barcode, index read 1 gives UMI and index read 2 provides sample index for multiplexing.

**Library amplification TIMING 3 h**

1. Dissolve the Agilent oligo (10 pmol) in 100 μL TE to obtain 100 nM solution.
2. Setup the first round PCR reaction. This reaction adds vector-overhang sequence upstream and minimal promoter and adaptor sequence downstream of the CRSs (**Fig. 1c, Supplementary Fig. 1b**).

|  |  |  |
| --- | --- | --- |
| Reagent | Volume (μL) | Final conc. |
| Agilent oligo (100 nM) | 2 | 1 nM |
| NEBNext High-Fidelity 2X PCR Master Mix | 100 | 1x |
| 5BC-AG-f01 (100 μM) | 1 | 0.5 μM |
| 5BC-AG-r01 (100 μM) | 1 | 0.5 μM |
| Ultrapure distilled H2O | 96 |  |
| Total volume | 200 |  |

1. Split the premixture into 5 PCR tubes (40 μL per tube).

**CRITICAL** Splitting the PCR reaction into multiple tubes is important to reduce the risk of PCR jackpotting.

1. Run the PCR reaction as follows:

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

1. Combine the PCR products into a DNA LoBind tube.
2. Add 1x volume (200 μL) of HighPrep PCR reagent and mix thoroughly by pipetting up and down 6-8 times.
3. Incubate the mixture for 5 min at room temperature.
4. Place the tube on the magnet device for 2-3 min.
5. Carefully remove the supernatant by pipetting. 10-20 μL of supernatant can be left in the tube.
6. With the tube placed on the magnet, add 200 μL of 80% ethanol.
7. Remove the ethanol by pipetting.
8. Repeat the 80% ethanol washing (Steps 10 and 11).
9. Flash spin the tube, immediately place the tube on the magnet and remove the supernatant.
10. Dry the bead pellet for 2-3 min. Do not over dry the beads.
11. Add 50 μL of EB to the beads and mix by pipetting and vortexing.
12. Place the tube on the magnet device for 1-2 min.
13. Transfer 45-50 μL of the eluate to a DNA LoBind tube.
14. Measure the DNA concentration using a Nanodrop.
15. Setup the second round PCR reaction. This reaction adds 15-bp barcode and vector-overhang sequence downstream of the first round PCR fragment (**Fig. 1d, Supplementary Fig. 1b**).

|  |  |  |
| --- | --- | --- |
| Reagent | Volume (μL) | Final conc. |
| 1st round PCR product | Variable (100 ng) |  |
| NEBNext High-Fidelity 2X PCR Master Mix | 200 | 1x |
| 5BC-AG-f02 (100 μM) | 2 | 0.5 μM |
| 5BC-AG-r02 (100 μM) | 2 | 0.5 μM |
| Ultrapure distilled H2O | Make up to 400 μL |  |
| Total volume | 400 |  |

1. Split the premixture into 10 PCR tubes (40 μL per tube).
2. Run the PCR reaction as follows:

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

1. Combine the PCR products into a DNA LoBind tube.
2. Add 200 μL of 6x gel loading dye (final conc. 2x) and mix the solution by vortexing.
3. Run the sample on two 1.5% agarose gels (30 mL of 5 cm x 6 cm mini gels with 3 cm-width well)
4. Cut the DNA band (419 bp) using a blue light safe imager.

**CRITICAL** We highly recommend using a blue light safe imager, because the UV transilluminator drastically decreases the recombination efficiency.

1. Purify the DNA from the gel slice using Qiaquick Gel Extraction kit according to manufacturer’s protocol.
2. Elute the DNA in 50 μL EB per column. If multiple columns are used, combine the eluate.
3. Purify the DNA using 1.2x volume of HighPrep PCR reagent following steps 6-17.
4. Measure the DNA concentration using a Nanodrop.

**TROUBLESHOOTING**

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

**Vector linearization TIMING 7 h - overnight**

1. Setup the vector digestion reaction as follows:

|  |  |  |
| --- | --- | --- |
| Reagent | Volume (μL) | Final conc. |
| pLS-SceI | Variable (10 μg) |  |
| 10x cutsmart buffer (NEB) | 20 | 1x |
| *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 H2O | Make up to 200 μL |  |
| Total volume | 200 |  |

1. Incubate the reaction at 37°C for 3 h – overnight.
2. Add 5 μL of *Age*I-HF (20 U/μL) and 5 μL of *Sbf*I-HF (20 U/μL) to the reaction.
3. Incubate the reaction at 37°C for 3 h – overnight.
4. Vortex for 30 sec and incubate at 80°C for 20 min.
5. Purify the DNA using 0.65x volume (136.5 μL) of HighPrep PCR reagent following steps 6-17.
6. Measure the DNA concentration using a Nanodrop.
7. To check the DNA size and quality, run 100-200 ng of the linearized vector and purified insert DNA (step 29) on a 1% gel along with a 1 kb DNA size marker. Make sure that specific single bands (7.8 kb linearized vector and 419 bp insert DNA) but no other bands appear on the gel.

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

**Recombination and electroporation TIMING 5 h plus two overnights for bacterial plate culture and suspension culture for mini prep**

1. Setup the recombination reaction as follows.

|  |  |  |
| --- | --- | --- |
| Reagent | Volume (μL) | Final conc. |
| Linearized pLS-SceI (step 36) | Variable (1 μg) |  |
| Purified insert DNA (step 29) | Variable (250 ng) |  |
| NEBuilder HiFi DNA Assembly Master Mix | 100 | 1x |
| Ultrapure distilled H2O | Make up to 200 μL |  |
| Total volume | 200 |  |

1. Incubate the reaction at 50°C for 60 min in a heating dry bath.
2. Place the tube on ice.
3. Purify the DNA using 0.65x volume (136.5 μL) of HighPrep PCR reagent following steps 6-17.
4. Setup the digestion reaction to get rid of undigested vectors as follows:

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

1. Incubate the reaction at 37°C for 1 h.
2. Purify the DNA using 1.8x volume (90 μL) of HighPrep PCR reagent following steps 6-14.
3. To elute the DNA, add 20 μL of EB to the beads and mix by pipetting and vortexing.
4. Place the tube on the magnet device for 1-2 min and transfer 18 μL of the eluate to a DNA LoBind tube.
5. Measure the DNA concentration using a Nanodrop. Make sure the DNA concentration is higher than 25 ng/μL, so as not to add more than 4 μL to 100 μL competent cells.

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

1. Prewarm 4-5 mL of NEB 10-beta/Stable Outgrowth Medium reagent at 37°C for at least 30 min.
2. Thaw NEB 10 beta electrocompetent cells on ice. We usually use 100 μL (1 tube) of the competent cells for low complexity libraries (less than 2 million) and 400 μL (4 tubes) for high complexity libraries (8-12 million).

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

1. Add 100 ng of the recombination product per 100 μL competent cells. The volume of DNA should be less than 4 μL per 100 μL competent cells.
2. Gently transfer 50 μL of the competent cells into a 1mm-gap cuvette without creating bubbles. Two cuvettes are prepared for 100 μL competent cells.
3. Gently tap the cuvettes on the counter to move the cells to the bottom.
4. Place the cuvettes in a Gemini X2 electroporator and shock the cells with the following settings:

Voltage, 2.0 kV; Resistance, 200 ohm; Capacitance, 25 uF, # of pulses, 1; Gap width, 1 mm.

1. Immediately add 450 μL of prewarmed NEB 10-beta/Stable Outgrowth Medium into the cuvettes, thoroughly mix by pipetting up and down, and transfer into a 14 mL conical tube.
2. Repeat the electroporation for all cuvettes, combining the electroporated bacteria into the single tube (total 1 mL culture per 100 μL competent cells).
3. Add fresh Outgrowth Medium and scale up to 4 mL in total. If 400 μL competent cells were used, no need to add.
4. Incubate the cells at 37°C for 1 h with agitation. Prewarm ten 15 cm LB agar plates in 37°C.

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

1. Plate the bacteria onto nine of the 15 cm LB agar plates (400 μL/plate), along with 100 μL/plate of 100 mg/mL Carbenicillin.Higher amount of Carbenicillin than general is added because the dense culture condition increases the risk of non-transformed bacteria growth.
2. Dilute 2 μL of the bacteria in 400 μL of fresh LB medium in a 1.5 mL tube and plate the entire diluted bacteria (402 μL) in a 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.
3. Incubate the plates at 37°C overnight.
4. To check the plasmid sequence of individual colonies, pick 16 colonies from the 2 μL-bacteria plate, purify the plasmids using QIAprep spin miniprep kit, and send them for Sanger sequencing using n40.dn.F and EGFP.up.R primers (**Supplementary Table 3**). Confirm the sequence structure is corresponding to the design (**Fig. 1e, Supplementary Fig. 1c**).

**TROUBLESHOOTING**

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

**Colony counting and plasmid library prep TIMING 3 h**

1. Count the number of colonies on the 2 μL-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.
2. Estimate the total number of colonies per 400 μL-bacteria plate by multiplying 200 to the colony count in the 2 μL-bacteria plate (**Supplementary Table 1)**.
3. Determine the number of 400 μL-bacteria plates to be used for the following plasmid preps. Total number of colonies needed can be determined by multiplying designed CRSs and desired number of barcodes per CRS (**Supplementary Table 1**).

**CRITICAL** 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 more than 200 barcodes per CRS does not increase reproducibility.

1. Add 5-6 mL of LB medium to each bacterial plate and gently scrape colonies without disturbing solid agarose.
2. Collect the bacteria suspension and combine into a few 50 mL tubes.
3. Add 5-6 mL of fresh LB medium again to the plates and collect as much bacteria leftover as possible into the tubes.
4. Purify the plasmids using 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 three columns of QIAGEN plasmid plus midi kit per 400 μL-bacteria plate.
5. Measure the plasmid concentration using a Nanodrop.
6. To check the DNA size and quality, run 100-200 ng of the plasmid on a 1% gel along with a 1 kb ladder marker.

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

**Sequencing for CRS-barcode association TIMING 4 h hands-on time plus sequencing turnaround time**

1. Setup the PCR reaction. This reaction adds P5 flowcell sequence and sample index sequence upstream and P7 flowcell sequence downstream of the CRS-barcode fragment **(Fig. 1f-g, Supplementary 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 2X PCR Master Mix | 100 | 1x |
| pLSmP-ass-i# (100 μM) | 1 | 0.5 μM |
| pLSmP-ass-gfp (100 μM) | 1 | 0.5 μM |
| Ultrapure distilled H2O | Make up to 200 μL |  |
| Total volume | 200 |  |

1. Split the premixture into 5 PCR tubes (40 μL per tube).
2. Run the PCR reaction as follows:

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

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

1. Combine the PCR products into a DNA LoBind tube.
2. Add 100 μL of 6x gel loading dye (final conc. 2x) and mix the solution by vortexing.
3. Run the sample on a 1.5% agarose gel (30 mL of 5 cm x 6 cm mini gel with 3 cm-width well)
4. Cut the DNA band (471 bp) using a blue light safe imager.
5. Purify the DNA from the gel slice using the Qiaquick Gel Extraction kit according to manufacturer’s protocol.
6. Elute the DNA in 50 μL EB per column. If multiple columns are used, combine the eluate.
7. Purify the DNA using 1.8x volume of HighPrep PCR reagent following steps 6-17.
8. Measure the DNA concentration using Qubit dsDNA HS according to the manufacturer’s protocol.
9. To check the DNA size and quality, run 50-100 ng of the DNA on a 1.5% gel along with a 100 bp ladder marker.

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

1. Send the purified DNA and custom primers for sequencing **(Fig. 1h, Supplementary Fig. 1d, Supplementary Table 3**). The sequencing should be done using paired-end 146 bp reads 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 sample index (let the sequencing facility know that the index read 2 should be used for demultiplexing). 10x coverage of sequencing reads than the total barcodes is required. For example, we use an Illumina MiSeq v2 run (15M reads) for a 0.5M barcode library or an Illumina NextSeq mid-output run (120M reads) for a 8-12M barcode library **(Supplementary Table 1**).

|  |  |  |  |
| --- | --- | --- | --- |
| Read | Cycle | Primer | Output |
| Read1 | 146 | pLSmP-ass-seq-R1 | CRS (upstream, forward) |
| Read2 | 146 | pLSmP-ass-seq-R2 | CRS (downstream, reverse) |
| Index read1 | 15 | pLSmP-ass-seq-ind1 | Barcode (forward) |
| Index read2 | 10 | pLSmP-rand-ind2 | Sample index |

**Lentivirus packaging TIMING >4 d cell culture and transfection plus >5.5 h hands-on time**

1. Culture 293T cells in DMEM (10% heat-inactivated FBS).
2. 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 0.5M barcode library or six T225 flasks for the 8-12M barcode library when carrying out lentiMPRA in HepG2 cells.
3. Incubate the cells for 2 days.
4. Prepare premixtures A and B as follows;

Premix A; 700 μL/flask OPTI-MEM and 60 μL/flask EndoFection.

Premix B; 700 μL/flask OPTI-MEM, 10 μg/flask plasmid library, 6.5 μg/flask psPAX2, and 3.5 μg/flask pMD2.G.

1. Transfer premix A into premix B by pipetting and mix the mixture by inverting the tube.
2. Incubate the tube at room temperature for 15 min.
3. Add 1.4 mL/flask of the mixture into 293T cells.
4. Incubate the cells for 8-14 h.

**CAUTION** During the following procedure, the liquid and plastic waste should be discarded into 10% bleach, as they are contaminated with lentivirus. Culture plates and tubes for storage should be clearly labeled as lentiviral contaminants.

1. Replace the media by 30 mL/flask of DMEM (5% heat-inactivated FBS) supplemented with 1x ViralBoost reagent (60 μL reagent per 30 mL medium).
2. Incubate the cells for 36-48 h.
3. Check GFP expression using fluorescent microscope. 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.
4. Filter the supernatant through 0.45 μm PES filter system. Use multiple filters (one filter up to three T225 flasks), so as not to get clogged.
5. Transfer the flow-through into multiple 50 mL tubes (30 mL per tube).
6. Add 1/3 volume (10 mL per tube) of Lenti-X concentrator reagent, close the lid tightly and mix gently by inverting the tubes.
7. 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 one week.

1. Centrifuge the tubes at 1,500 g for 45 mins at 4°C
2. Discard supernatant into 10% bleach by gentle decanting.
3. Discard the remainder of supernatant into 10% bleach by pipetting without disturbing the pellet.
4. Gently resuspend the pellet in cold DPBS. We usually use 600 μL DPBS per T225 flask.
5. Store the lentivirus at 4°C.

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

**CRITICAL** We do not recommend freezing the virus, especially in case of high complexity libraries, because freeze-thawing significantly decreases the viral titer. We did not see drastic loss of the viral titer when stored at 4°C up to three weeks. The following infection experiments should be done within three weeks.

**Lentivirus titration TIMING >4 d cell culture and transfection plus 3 h hands-on time**

1. 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 1** (steps 1-14).
2. 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**).
3. Based on its slope and the number of cells seeded, calculate the virus titer (TU/μL) using **Supplementary Table 1**.

**Lentivirus infection and DNA/RNA extraction TIMING >4 d cell culture and transfection plus >3.5 h hands-on time**

1. Determine the number of integrations per barcode (i.e. total number of a particular barcode existing in entire cell population). 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 (**Supplementary Table 1**).
2. Seed 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 MOI of the cells (**Supplementary Table 1**). Three independent biological replicates should be performed.
3. Incubate the cells overnight.
4. Refresh culture media with adding polybrene at the appropriate concentration (**Box 1**).
5. Add appropriate amount of the lentivirus library. The volume of virus required is given as total barcode integrations divided by virus titer (**Supplementary Table 1**).
6. Refresh the culture media with no polybrene the following day.
7. After two days (three days culture in total), check GFP fluorescence to confirm proper lentiviral integration and expression (**Fig. 1i**). GFP expression depends on the library design.
8. Remove the culture media and wash the cells with DPBS three times. Remove DPBS completely.
9. Add RLT plus lysis buffer supplemented with 2-mercaptoethanol. We usually use 1200 μL or 2400 μL lysis buffer per 10 cm dish or 15 cm dish, respectively.
10. Scrape the cells using a cell lifter and homogenize the cell lysis using a 3 mL syringe and 20G needle.

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

1. Transfer the lysate to DNA columns. Use two or four columns per 10 cm dish or 15 cm dish, respectively.
2. Extract genomic DNA and total RNA simultaneously using the AllPrep DNA/RNA mini kit according to manufacturer’s protocol.For RNA samples, perform DNase treatment inbetween two of the 350 μL RW1 washes using Qiagen’s RNase-free DNase set, according to the manufacturer’s protocol.
3. Elute DNA in 30 μL/column of EB and combine each replicate in a single tube. Keep each of the three replicates separated.
4. Elute RNA in 30 μL/column of RNase-free H2O and combine each replicate in a single tube. Keep each of the three replicates separated.
5. Measure the concentration of DNA and RNA samples using a Nanodrop. 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**

1. (Optional) Extract mRNA from the total RNA samples using Oligotex mRNA mini kit according to the manufacturer’s protocol. mRNA should be eluted from the oligotex beads in 90 μL OEB.
2. Treat RNA samples with DNase using the TURBO DNA-free kit following “Rigorous DNase treatment” following the manufacturer’s protocol. As the RNA sample is already treated with DNase during the AllPrep procedure (step 118), TURBO DNase treatment can be done using a high concentration condition without any dilution.
3. Measure the RNA concentration using the Qubit RNA HS. At least 1.2 μg mRNA or 60 μg total RNA per replicate is required for a low complexity library (0.5-2M total barcodes).

**PAUSE POINT** DNase-treated RNA can be stored at -80°C for months.

1. For three replicates of RNA samples, perform a reverse-transcription reaction in 8-strip PCR tubes. This reaction adds 16-bp unique molecular identifier (UMI) and P7 flowcell sequence downstream of the barcode (**Supplementary Fig. 1e**). For a high complexity library (8-12M total barcodes), we recommend to multiply by 4X the reaction.

|  |  |  |
| --- | --- | --- |
| Reagent | Volume (μL) | Final conc. |
| mRNA or total RNA | Variable (1.2 μg mRNA or 60 μg total RNA) |  |
| P7-pLSmP-ass16UMI-gfp (100 μM) | 0.25 | 0.25 μM |
| 10 mM dNTP mix | 5 | 0.5 mM |
| Ultrapure distilled H2O | Make up to 65 μL |  |
| Total volume | 65 |  |

1. Incubate the reaction at 65°C for 5 min using a thermal cycler.
2. Place the tubes on ice.
3. Add 20 μL 5x First strand buffer and 10 μL 0.1 M DTT.
4. Incubate the tubes at 42°C for 2 min using a thermal cycler.
5. Add 5 μL of Superscript II.
6. Incubate the tubes at 42°C for 50 min followed by 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 6 h hands-on time plus sequencing turnaround time**

1. Dilute DNA samples (step 121) at a final concentration of 120 ng/μL. For RNA samples (step 131), use all 100 μL RT products for the following first round PCR reaction.
2. For three replicates of both DNA and RNA samples, perform first round PCR reaction. This reaction adds P5 flowcell sequence and sample index sequence upstream and 16-bp UMI and P7 flowcell sequence downstream of the barcode (**Fig. 1j-k, Supplementary Fig. 1e**). Use different sample index sequences for each sample (**Supplementary Table 3**). For a high complexity library (8-12M total barcodes), we recommend to multiply by 4X the reaction.

|  |  |  |
| --- | --- | --- |
| Reagent | Volume (μL) | Final conc. |
| DNA or cDNA | 100 |  |
| NEBNext High-Fidelity 2X PCR Master Mix | 200 | 1x |
| P7-pLSmP-ass16UMI-gfp (100 μM) | 2 | 0.5 μM |
| P5-pLSmP-5bc-i# (100 μM) | 2 | 0.5 μM |
| Ultrapure distilled H2O | 96 |  |
| Total volume | 400 |  |

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

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

1. Combine each sample into a DNA LoBind tube.
2. Add 1.8x volume (720 μL) HighPrep PCR reagent and mix thoroughly by mix pipetting up and down 6-8 times.
3. Purify the DNA using 1.8x volume (700 μL) of HighPrep PCR reagent following steps 6-14.
4. Add 60 μL of EB to the beads and elute the DNA by pipetting and vortexing.
5. Place the tube on the magnet device for 1-2 min and transfer 55-58 μL of the eluate to a LoBind tube. Store the tubes on ice.
6. Setup 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. |
| 1st round PCR product | 5 |  |
| NEBNext High-Fidelity 2X PCR Master Mix | 10 | 1x |
| P7 (100 μM) | 0.1 | 0.5 μM |
| P5 (100 μM) | 0.1 | 0.5 μM |
| 100x SYBR | 0.1 | 1x |
| Ultrapure distilled H2O | 4.7 |  |
| Total volume | 20 |  |

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

|  |  |  |  |
| --- | --- | --- | --- |
| Cycle no. | Denature | Anneal | Extend |
| 1 | 98°C, 1 min |  |  |
| 2-26 (25 cycles) | 98°C, 10 sec | 60°C, 30 sec | 72°C, 1 min |

1. Based on the raw amplification curve of each sample, determine the number of cycles where the amplification nearly plateaus for each sample.
2. Setup the second round PCR reaction for each sample as follows. As it is expected that the number of cycles required for DNA and RNA samples are different, we recommend running the PCR for them separately.

|  |  |  |
| --- | --- | --- |
| Reagent | Volume (μL) | Final conc. |
| 1st round PCR product | 50 |  |
| NEBNext High-Fidelity 2X PCR Master Mix | 100 | 1x |
| P7 (100 μM) | 1 | 0.5 μM |
| P5 (100 μM) | 1 | 0.5 μM |
| Ultrapure distilled H2O | 48 |  |
| Total volume | 200 |  |

1. Split the premixture into 5 PCR tubes (40 μL per well).
2. Run the PCR reaction with the cycle number that were determined by preliminary run (step 143) as follows:

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

1. Combine each sample into a DNA LoBind tube.
2. Purify the DNA using 1.8x volume (360 μL) of HighPrep PCR reagent following steps 6-14.
3. Elute the DNA in 20 μL EB.
4. Measure the DNA concentration using a Nanodrop.
5. Pool three replicates of DNA samples or RNA samples with each replicate equal amount (1 μg), using **Supplementary Table 4**. Keep DNA and RNA samples separated until these are pooled at the later step (step 162).

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

1. Add equal volume of 6x gel loading dye (final conc. 3x) and mix the solution by vortexing.
2. Run the pooled sample on a 1.8% agarose gel (30 mL of 5 cm x 6 cm mini gels with 1.3 cm-width well).
3. Cut the DNA bands (156 bp) using a blue light safe imager.
4. Purify the DNA from the gel slices using Qiaquick Gel Extraction kit according to manufacturer’s protocol. Use one column for each of DNA or RNA samples.
5. Elute the DNA in 50 μL EB per column.
6. Purify the DNA using 1.8x volume (90 μL) of HighPrep PCR reagent following steps 6-14.
7. Add 20 μL of EB to the beads and elute the DNA by pipetting and vortexing.
8. Place the tube on the magnet device for 1-2 min and transfer 18 μL of the eluate to a LoBind tube.
9. Measure the DNA concentration using a Qubit dsDNA HS according to manufacturer’s protocol.
10. To check the DNA size and quality, run 20-30 ng of the DNA on a 1.8% gel along with 100 bp ladder marker.
11. Pool the DNA and RNA samples in a single LoBind tube with 1:3 ratio to obtain 100 uL 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.

1. Send the sequencing library and custom primers for sequencing **(Fig. 1l, Supplementary Fig. 1e, Supplementary Table 3**). The sequencing should be done with paired-end 15 bp, 16-cycle for index read 1 and 10-cycles for index read 2. Index read 1 provides the UMI sequence, and index read 2 provides sample index (let the sequencing facility know that the index read 2 should be used for demultiplexing). 10x (DNA) and 30x (RNA) coverage of the library via sequencing reads is required. For example, we use an Illumina NextSeq high-output run (400M reads) for the 0.5M barcode library or three runs (1.2B reads) for the 8-12M barcode library **(Supplementary Table 1**).

|  |  |  |  |
| --- | --- | --- | --- |
| Read | Cycle | Primer | Output |
| Read1 | 15 | pLSmP-ass-seq-ind1 | barcode (forward) |
| Read2 | 15 | pLSmP-bc-seq | barcode (reverse) |
| Index read1 | 16 | pLSmP-UMI-seq | UMI |
| Index read2 | 10 | pLSmP-5bc-seq-R2 | Sample index |

**Data processing TIMING depends on read depth 4h-4d (Fig. 2a-b)**

164. Ensure your computer is running a Linux distribution (add more details about requirements)

165. Install conda (https://docs.conda.io/en/latest/miniconda.html) if not already on your system.

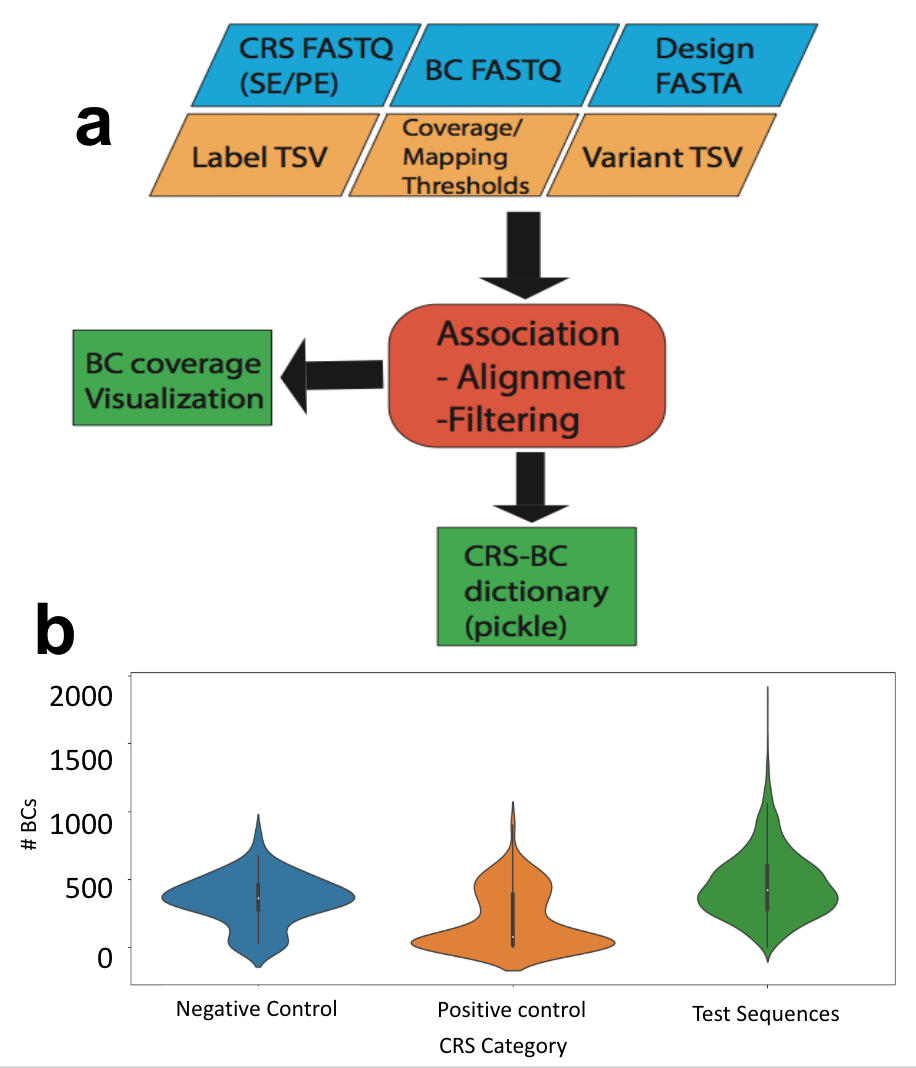
166. Clone repository git clone https://github.com/shendurelab/MPRAflow.git

167. Once cloned, change directory into the conf folder in the repository and create the conda environment in our system.

cd MPRAflow

conda env create -n MPRAflow -f environment.yml

**CRITICAL** Ensure the nextflow config file is set up correctly for your system. This pipeline comes with a nextflow.config file setup to run on HPC clusters, allowing each process to be run as a separate 'qsub' command. The config contains example code for Sun Grid Engine (SGE) scheduler, further documentation on supported executors can be found on Nextflow’s website (<https://www.nextflow.io/docs/latest/executor.html>). Please remove the \\ for the architecture you would like to use and place \\ in front of any architectures not currently in use. A '\' in front of all of them runs the pipeline on your local machine. Please consult your cluster's Wiki page for cluster-specific commands and change clusterOptions = to reflect these specifications. Additionally, for large libraries, more memory can be specified in this location.

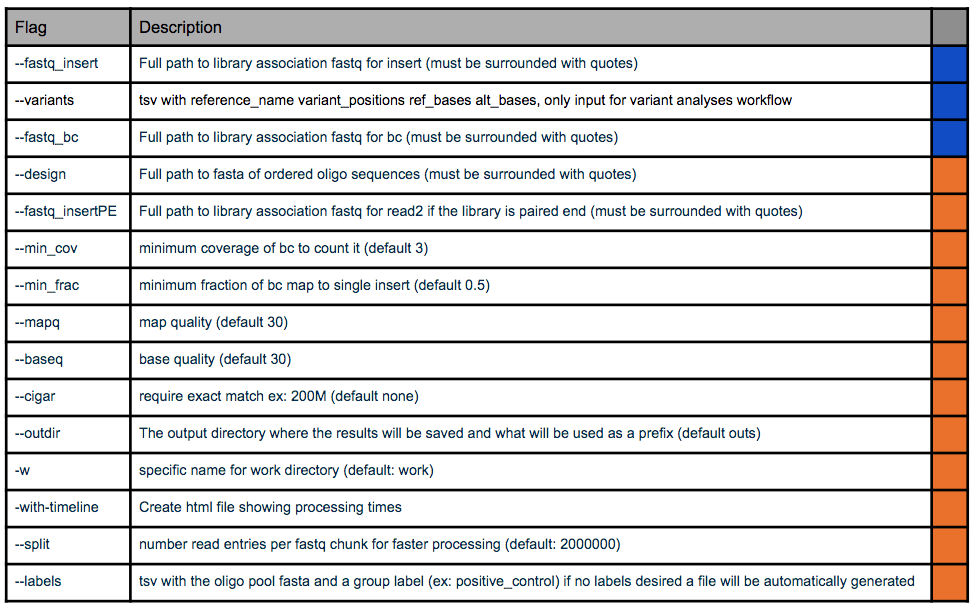


**Fig. 2 | Overview of MPRAflow association utility.** **a**, Mandatory inputs (blue), optional flags (orange), output files (green) and utility (red). The program requires FASTQ files for the insert, either single-end (SE) or paired-end (PE) reads and a design file, which is a FASTA file containing the synthesized oligos. 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, a tab-separated values file (TSV) of variants in the ordered oligo 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 visualization of barcode coverage for each enhancer, grouped by labels provided in the label TSV.

168. Run the association utility. The user must specify the barcode, CRS read 1 and CRS read 2 (if applicable) FASTQ files and a FASTA file containing synthesized CRS oligo pool. (**Fig. 2a**) The most basic command for single-end sequencing of the CRS can be seen below. However, there are many optional flags that allows for customization of the pipeline as outlined in **Table 1**. If a HPC cluster is available with a queuing system such as SGE is available, we recommend submitting this command to the queue.

conda activate MPRAflow

nextflow run association.nf --fastq\_insert "${fastq\_prefix}\_R1\_001.fastq.gz" --design "ordered\_candidate\_sequences.fa" --fastq\_bc "${fastq\_prefix}\_R2\_001.fastq.gz"



**Table 1 | Association Utility options.** Blue rows are mandatory and orange are optional.

169. Create an experiment.csv file. This file contains the relevant information for the experiment, such as the condition, replicate, FASTQ file names, and an output prefix. For example:

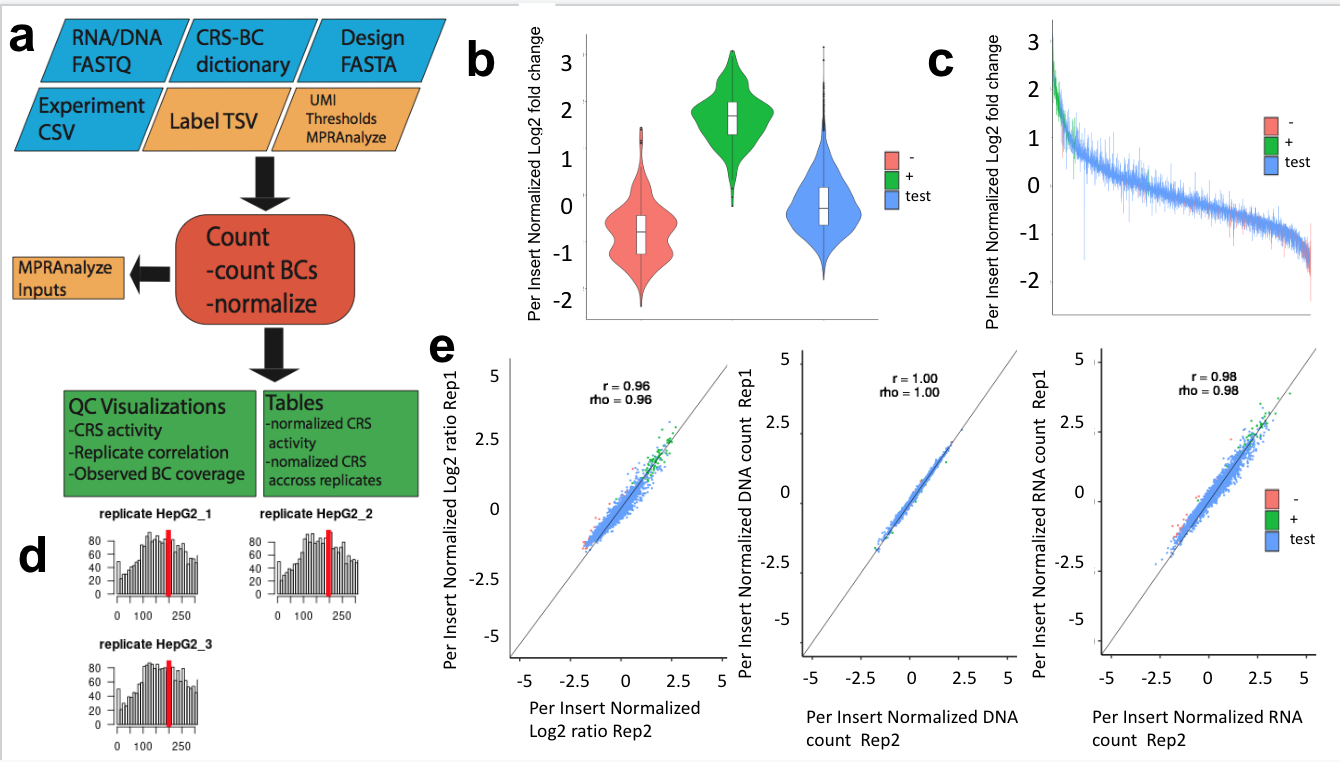
Condition,Replicate,DNA\_BC\_F,DNA\_UMI,DNA\_BC\_R,RNA\_BC\_F,RNA\_UMI,RNA\_BC\_R,Name

condition1,1,cond1\_rep1\_DNA\_FWD\_reads.fastq.gz,cond1\_rep1\_DNA\_IDX\_reads.fastq.gz,cond1\_rep1\_DNA\_REV\_reads.fastq.gz,cond1\_rep1\_RNA\_FWD\_reads.fastq.gz,cond1\_rep1\_RNA\_IDX\_reads.fastq.gz,cond1\_rep1\_RNA\_REV\_reads.fastq.gz,desired name

condition1,2,cond1\_rep2\_DNA\_FWD\_reads.fastq.gz,cond1\_rep2\_DNA\_IDX\_reads.fastq.gz,cond1\_rep2\_DNA\_REV\_reads.fastq.gz,cond1\_rep2\_RNA\_FWD\_reads.fastq.gz,cond1\_rep2\_RNA\_IDX\_reads.fastq.gz,cond1\_rep2\_RNA\_REV\_reads.fastq.gz,desired name

condition2,1,cond2\_rep1\_DNA\_FWD\_reads.fastq.gz,cond2\_rep1\_DNA\_IDX\_reads.fastq.gz,cond2\_rep1\_DNA\_REV\_reads.fastq.gz,cond2\_rep1\_RNA\_FWD\_reads.fastq.gz,cond2\_rep1\_RNA\_IDX\_reads.fastq.gz,cond2\_rep1\_RNA\_REV\_reads.fastq.gz,desired name

condition2,2,cond2\_rep2\_DNA\_FWD\_reads.fastq.gz,cond2\_rep2\_DNA\_IDX\_reads.fastq.gz,cond2\_rep2\_DNA\_REV\_reads.fastq.gz,cond2\_rep2\_RNA\_FWD\_reads.fastq.gz,cond2\_rep2\_RNA\_IDX\_reads.fastq.gz,cond2\_rep2\_RNA\_REV\_reads.fastq.gz,desired name

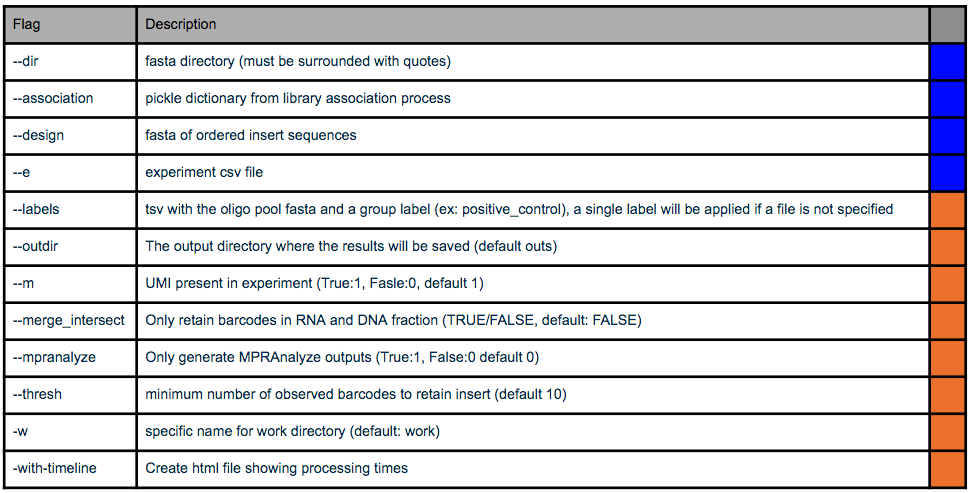


**Fig. 3 | Overview of count utility.** **a**, Mandatory inputs (blue), optional flags and outputs (orange), output files (green) and utility (red). The user must specify the directory containing all FASTQ files for the RNA and DNA sequencing, the CRS-barcode dictionary from the Association Utility, a design file (FASTA file containing the synthesized oligos), 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 and tune parameters such as specifying if a unique molecular identifier (UMI) was used, or if the user would like to generate the input files for MPRAnalyze. The program will produce normalized activity of each CRS from each replicate as well as across replicates along with several visualizations (**b**-**e**). **b**, CRS activity normalized by insert and grouped by label determined in the label file. **c**, Normalized activity of each CRS across replicates colored by label. **d**, Distribution of observed barcode coverage per replicate. **e**, Correlation of normalized log2(RNA/DNA), DNA counts, and RNA counts colored by label.

170. Run the count utility. The user must specify the directory containing the DNA and RNA sequencing data. A Python dictionary in pickle format from the association utility, the FASTA file containing synthesized CRS oligo pool, and the experiment.csv file (**Fig. 3a**). There is some flexibility provided for the user at this stage outlined in **Table 2**. Below is an example of the minimal command. If a HPC cluster is available with a queuing system such as SGE is available, we recommend submitting this command to the queue.

conda activate MPRAflow

nextflow run count.nf --dir "bulk\_FASTQ\_directory" --e "experiment.csv" --design "ordered\_candidate\_sequences.fa" --association "dictionary\_of\_candidate\_sequences\_to\_barcodes.p"



**Table 2 |** **Count utility options.** Blue rows are mandatory and orange are optional.

171. By setting the flag -mpranalyze to 1 in the command above the user can produce the inputs for that analysis pipeline.

172. If the user has created their library using a saturation mutagenesis approach, they must first create an assignment file in the format below:

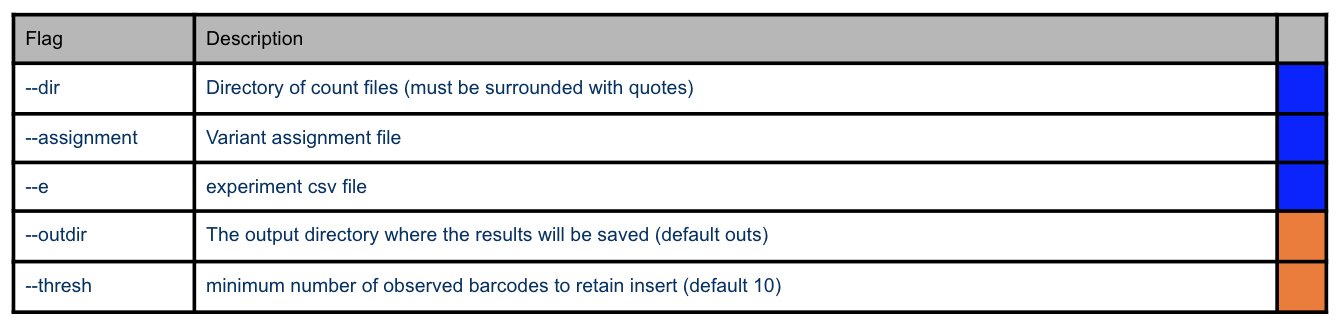
AAAAAACTAATACCA IRF6:104:A>G IRF6:408:G>A

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

173. To run the saturation mutagenesis utility run the user must specify the directory containing the Count tables, the assignment file shown above, and the experiment.csv file. There is some flexibility provided for the user at this stage outlined in **Table 3**. Below is an example of the minimal command. If a HPC cluster is available with a queuing system such as SGE is available, we recommend submitting this command to the queue.

conda activate MPRAflow

conda run saturationMutagenesis.nf –dir “count\_table\_directory” –assignment “var\_assign.txt” –e “experiment.csv”



**Table 3 |** **Saturation Mutagenesis utility options.** Blue rows are mandatory and orange are optional.

**Troubleshooting**

**Table 3 | Troubleshooting.**

|  |  |  |  |
| --- | --- | --- | --- |
| Step | Problem | Possible reason | Solution |
| Box1, step5 | Low infection efficiency | Polybrene concentration may not be appropriate | Optimization of polybrene concentration may be required. Seed cells in a 24-well plate and infect control virus along with different amount of polybrene (e.g. 0, 2, 4, 8, 16, 32 μg/mL at a final concentration), and observe cell death and GFP expression. In our experience, 8 μg/mL works good for most cell types including HepG2, K562, H1 hESCs, and WTC11 iPSCs. Polybrene kills neural cell types, including neural progenitors and should be avoided when using those cells. |
| Step 29 | Low DNA yield. At least 250 ng 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 of the second round PCR up to 15 cycles. More cycles (>15 cycles) can decrease the library complexity. |
| Step 37 | Uncut vector DNA appear on the gel. | Insufficient restriction enzyme reaction. | Perform restriction digestion twice or three times (step 31-36). |
| Step 61 | Contamination of empty vectors. | Vector linearization and/or I-*Sce*I digestion were not sufficient. | Lower rate (less than 10%, one or two out of 16 colonies) is acceptable. Proceed with the protocol. If higher rate, redo vector linearization with longer incubation time and make sure you have complete linearization using an agarose gel. Perform I-*Sce*I digestion with longer incubation time. |
| Step 61 | Mutation and indels observed in the plasmids. | These can be derived from synthesis/PCR/sequencing errors. | As these errors are insolvable. We usually observe >50% of sequences contain mutations and/or deletions. Proceed with the protocol, and these erroneous sequences should be ruled out during the analysis step. Synthesis error rates might be improved by ordering oligos that are high-fidelity synthesized from the manufacturer. |

**Timing**

Box 1, Step 1-15, test infection: >4 d cell culture and transfection plus 3 h hands-on time

Steps 1–29, library amplification: 3 h

Steps 30–37, vector linearization: 7 h - overnight

Steps 38–61, recombination and electroporation: 5 h hands-on time plus two overnights for bacterial plate culture and suspension culture for mini prep

Steps 62-70, colony count and plasmid library prep: 3 h

Steps 71–83, sequencing for CRS-barcode association: 3 h hands-on time plus sequencing turnaround time

Steps 84-103, lentivirus packaging: >4 d cell culture and transfection plus >5.5 h hands-on time

Steps 104-106, lentivirus titration: >4 d cell culture and transfection plus 3 h hands-on time

Steps 107-121, lentivirus infection and DNA/RNA extraction: >4 d cell culture and transfection plus >3.5 h hands-on time

Steps 122-131, reverse transcription: 4 h

Steps 122-131, library prep and sequencing for RNA and DNA barcode count: 6 h

Steps 164-171, data processing: depends on read depth (**Supplementary Fig. 2**)

**Anticipated results**

The output of a lentiMPRA experiment will consist of two sets of data: association sequencing, and DNA/RNA sequencing. Success of association sequencing preparation can be assessed by the size of the band (~417 bp) observed during library preparation. The association sequencing should contain paired-end reads that cover the CRS (230 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 more than 600 barcodes per CRS should be cloned again since 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 sequencing library can be assessed by the size of the band (~156 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**). Additionally, the correlation between technical replicates are shown for DNA count, RNA count, and log2(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 which 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 has been deposited into the NCBI Sequence Read Archive (SRA) under accession no. ?????.

**Code availability**

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

**Author contributions statements**

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

**Acknowledgments**

This work was supported by the National Human Genome Research Institute grant number 1UM1HG009408 (N.A. and J.S.) and 1R21HG010065 and 1R21HG010683 (N.A.), NRSA NIH fellowship 5T32HL007093 (V.A.), National Institute of Mental Health grant numbers 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.

**Competing interests**

The authors declare no competing interests.

**REFERENCES**

1 Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu Rev Genomics Hum Genet.* **18:45-63.**, 10.1146/annurev-genom-091416-035537. Epub 092017 Apr 091417. (2017).

2 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature.* **461**, 747-753. doi: 710.1038/nature08494. (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 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).

7 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. doi: 1010.1038/nprot.2018.1015. Epub 2018 Apr 1012. (2018).

8 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* **326**, 289-293. (2009).

9 Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **10**, 30008-30002 (2015).

10 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science.* **339**, 1074-1077. doi: 1010.1126/science.1232542. Epub 1232013 Jan 1232517. (2013).

11 Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38-52. doi: 10.1101/gr.212092.212116. Epub 212016 Nov 212099. (2017).

12 Klein, J. *et al.* A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. *bioRxiv*, 576405 (2019).

13 Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183. doi: 110.1186/s13059-13019-11787-z. (2019).

14 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. doi: 710.1016/j.stem.2019.1009.1010. Epub 2019 Oct 1017. (2019).

15 Ryu, H. *et al.* Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. *bioRxiv*, 256313 (2018).

16 Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210 (2019).

17 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. doi: 110.1093/bioinformatics/btw1584. Epub 2016 Sep 1096. (2017).

18 Ghazi, A. R. *et al.* Design tools for MPRA experiments. *Bioinformatics.* **34**, 2682-2683. doi: 2610.1093/bioinformatics/bty2150. (2018).

19 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

20 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

21 Klein, J. C. *et al.* Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res* **8** (2015).