Challenges and considerations for reproducibility of STARR-seq assays

Maitreya Das[1, 3, 5*], Ayaan Hossain[3, 5], Deepro Banerjee[4, 5], Craig Alan Praul[5] and Santhosh Girirajan[1, 2, 3, 4, 5*]

1. Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802
2. Department of Anthropology, Pennsylvania State University, University Park, PA 16802
3. Molecular and Cellular Integrative Biosciences Graduate program, Pennsylvania State University Park, PA 16802
4. Bioinformatics and Genomics Graduate program, Pennsylvania State University, University Park, PA 16802
5. Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802

Correspondence:

Santhosh Girirajan

205A Life Sciences Building

Pennsylvania State University

Email: sxg47@psu.edu

Maitreya Das

205 Life Sciences Building

Pennsylvania State University

Email: mud367@psu.edu

Abstract

High-throughput methods such as RNA-seq, ChIP-seq and ATAC-seq have well-established guidelines, commercial kits, and analysis pipelines that enable consistency and wider adoption for understanding genome function and regulation. STARR-seq, a popular assay for directly quantifying activity of thousands of enhancer sequences simultaneously, has seen limited standardization across studies. Further, the assay is long with >250 steps, and frequent customization of the protocol and variations in bioinformatics methods raise concerns for reproducibility of STARR-seq studies. Here, we assess each step of the protocol and analysis pipelines from published sources and in-house assays, and identify critical steps and quality control checkpoints necessary for reproducibility of the assay. We also provide guidelines for assay design, protocol scaling, customization, and analysis pipelines for better adoption of the assay. These resources will allow better optimization of STARR-seq for specific research needs, enable comparisons and integration across studies, and improve reproducibility of results.

Introduction

Enhancers are *cis*-acting DNA elements that regulate gene expression[1]. The ability of enhancers to recruit transcription factors through specific binding motifs to regulate the expression of target genes in a cell, tissue, and developmental stage-specific manner make them critical components of gene regulatory networks[2]. While enhancer-reporter assays[1] and comparative genomics[3] enabled initial discoveries, enhancers are usually mapped within nucleosome-free open chromatin regions that are hypersensitive to DNase I[4] or accessible to transposase[5], and to sequences bound by specific transcription factors[6] or histone modifications[7]. Such assays used to detect putative enhancer regions throughout the genome include DNase-seq[8], ATAC-seq[5], and ChIP-seq[9].

Self-Transcribing Active Regulatory Region sequencing or STARR-seq is a specialized form of an episomal massively parallel reporter assay (MPRA) that directly quantifies enhancer activity by relying on transcription factors within a host cell system, thereby removing any chromatin-associated biases[10]. However, predicted enhancer regions from such assays typically require further experimental validation to specifically detect regulation of expression by the particular genomic region. STARR-seq takes advantage of the property of enhancers to act bidirectionally and therefore candidate enhancer fragments cloned downstream of a minimal promoter sequence transcribe themselves in a host cell. A comparison between the final read count of self-transcribed fragments with the initial number of transfected fragments provides a quantifiable measure of enhancer activity. By assessing candidate enhancer libraries through massively parallel sequencing, researchers have built genome-wide enhancer activity maps[11], functionally validated enhancers identified by other methods such as ChIP-seq[12], ATAC-seq[13], and FAIRE-seq[14], tested the impact of non-coding variants within or near enhancer sequences[15,16,17,18], and assessed changes in enhancer activity due to external factors such as drug or hormone treatment[2,19]. Indeed, STARR-seq has emerged as a state-of-the-art method for functional interrogation of the non-coding genome.

While STARR-seq is powerful and versatile, there are several challenges associated with successfully carrying out this assay. *First*, the experimental protocol is laborious with over 250 steps, involving construction of a STARR-seq plasmid library, transfection of the library into host cells, sequencing of transfected (input) and self-transcribed (output) fragments to sufficient read depth, and bioinformatic analysis to identify peaks at enhancer sites[20]. Several of these steps

also require preliminary experiments, *for example*, optimizing library construction to achieve sufficient complexity, testing transfection efficiency, mitigating host cell line limitations, and assessing target coverage for reproducible discovery of enhancer location and activity. *Second*, variation across studies and a lack of benchmarking for selecting sequencing and data analysis parameters such as optimal read depth, choice of peak caller, cut-off scores for characterizing enhancer activity, and methods for data validation complicate comparisons across studies. *Third*, most published studies do not provide sufficient details for researchers to adapt or scale their protocols for specific needs such as modifying target library size and choosing enhancer fragment length. Furthermore, lack of quality control details for critical intermediate steps raises significant concerns for replicability and reproducibility of results from STARR-seq assays.

Several studies have addressed challenges associated with reproducibility in biology, including reports on best practice guidelines for various genomic analysis pipelines such as RNA-seq[21], ChIP-seq[22], and ATAC-seq[23] as well as the Reproducibility in Cancer Biology project (Center of Open Sciences and Science Exchange). *For example*, Errington and colleagues attempted 193 experiments from published reports, but were able to replicate only 50 due to failure of previous studies to report descriptive and inferential statistics required to assess effect sizes, as well as a lack of sufficient information on study design. Several of their experiments required modifications to the original protocols with results showing significant deviations from previous findings[24]. These studies illustrate the difficulties in successfully repeating published functional genomics experiments.

Here, we discuss four major aspects of a STARR-seq assay including *(i)* pre-experimental assay design, *(ii)* plasmid library preparation, *(iii)* enhancer screening, and *(iv)* data analysis and reporting. We also delineate several features specific to STARR-seq in the context of general reproducibility of biological experiments[25]. We demonstrate how each of these features vary across previous studies and score them based on details provided in the original publications. We also identify challenges, rate-limiting conditions, and critical checkpoints through a series of carefully designed STARR-seq assays, and propose best practice guidelines. We provide multiple resources for study design, library construction, sequencing, and data analysis to help future researchers conduct robust and reproducible STARR-seq assays.

Different assay goals result in varying experimental designs

The general strategy for STARR-seq consists of cloning a library of fragments selected from either a list of putative enhancers or the entire genome into a STARR-seq plasmid, which is transfected into a host cell line. The "input" library and the transcribed "output" fragments from the host cells are sequenced to high read depth, followed by bioinformatic analysis for quantification of enhancer peaks[10]. In our assessment of published STARR-seq studies, we found that the original protocol[10] was modified in myriad ways to fit various study goals and underlying biological contexts, including altering design features such as target library size, DNA source, fragment length, sequencing platform, choice of STARR-seq vector, and choice of host cell line (Fig. 1a). Reporting of these features are crucial for understanding the rationale behind downstream strategies such as protocol scaling, choice of quality control measures, and selection of optimal parameters for assay performance and data analysis. Here, we discuss STARR-seq experimental design features, how they vary across previous studies and affected study reproducibility, and suggest rationale for selecting different designs (Table 1).

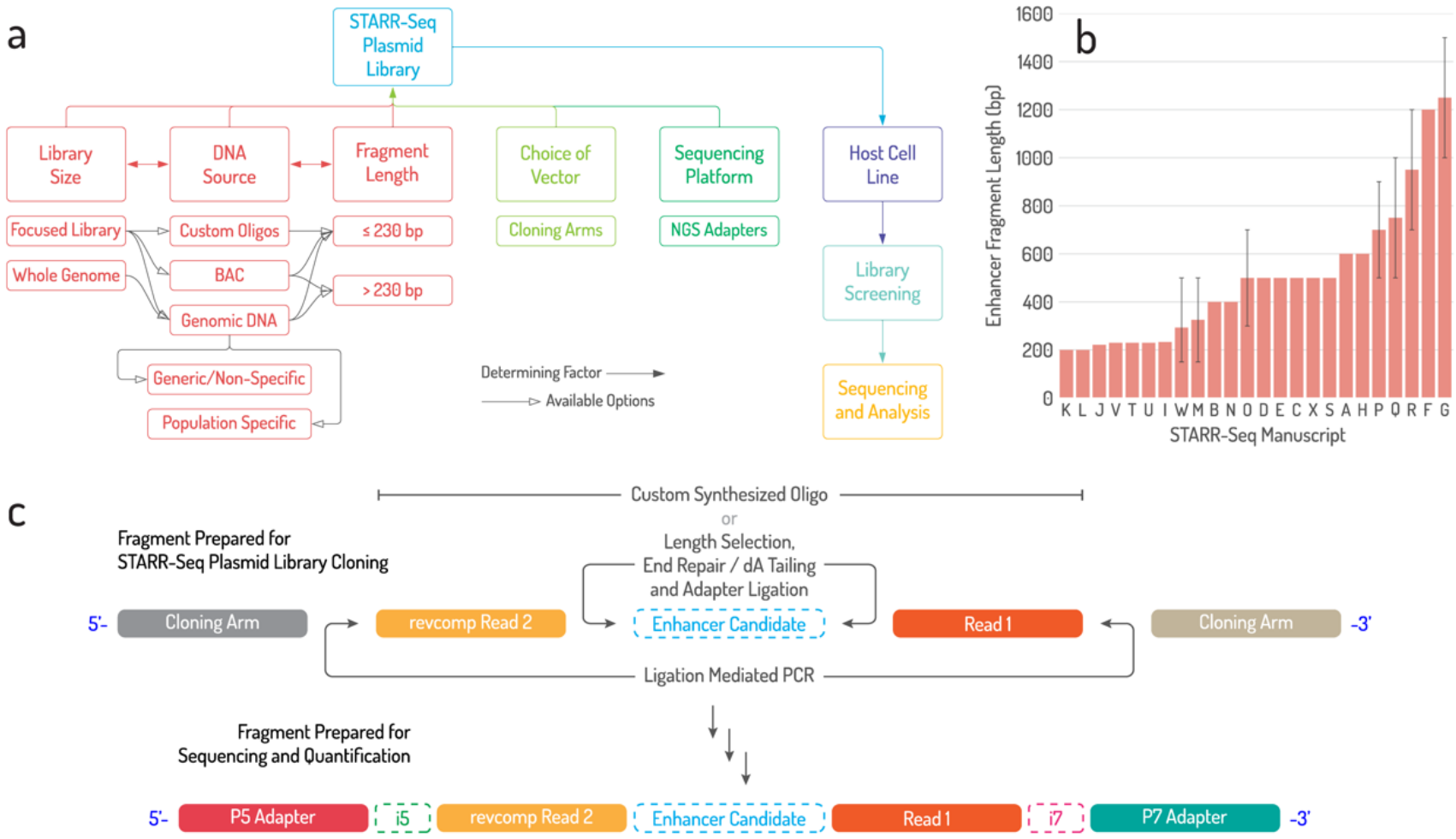

a
STARR-Seq Plasmid Library
Library Size
DNA Source
Fragment Length
Choice of Vector
Sequencing Platform
Host Cell Line
Focused Library
Whole Genome
Custom Oligos
BAC
Genomic DNA
≤ 230 bp
> 230 bp
Cloning Arms
NGS Adapters
Library Screening
Sequencing and Analysis
Generic/Non-Specific
Population Specific
Determining Factor
Available Options
b
Enhancer Fragment Length (bp)
1600
1400
1200
1000
800
600
400
200
0
K L J V T U I W M B N O D E C X S A H P Q R F G
STARR-Seq Manuscript
c
Custom Synthesized Oligo
or
Length Selection, End Repair / dA Tailing and Adapter Ligation
Fragment Prepared for STARR-Seq Plasmid Library Cloning
5'-
Cloning Arm
revcomp Read 2
Enhancer Candidate
Read 1
Cloning Arm
-3'
Ligation Mediated PCR
Fragment Prepared for Sequencing and Quantification
5'-
P5 Adapter
i5
revcomp Read 2
Enhancer Candidate
Read 1
i7
P7 Adapter
-3'

Figure 1: Experimental design features of a STARR-seq experiment. **(a)** Variations of pre-experimental design features and their impact on library preparation protocol are shown. **(b)** Fragment length distribution of published STARR-seq studies with de-identified papers on X axis and length on Y axis. Error bars indicate range of fragment lengths used by different studies. **(c)** Read architecture schematic of a STARR-seq plasmid library (top) and 'input' and 'output' sequencing libraries (bottom) are shown.

Target library size

The versatility of STARR-seq allows for enhancer screening across both whole genomes and targeted regions. Studies screening for enhancer activity across numerous targets typically require many replicates for each library preparation step and high cloning and transformation efficiency to achieve the required coverage of all fragments compared to studies with fewer targets. *For example*, a library generated by Johnson and colleagues to assess the effect of glucocorticoids on genome-wide enhancer activity contained over 560 million unique fragments with 59X coverage of the entire human genome[19]. The authors performed 60 replicates of sequencing adapter ligations and 72 transformations and grew the pooled libraries in 18-liter LB broth[19]. In contrast, Klein and colleagues synthesized a targeted library of 2,440 unique fragments along with customized adapter and cloning sequences in three replicates to compare context-specific differences among seven different MPRA designs, including two variations of STARR-seq[26]. Cloning and transformation of each tested design was performed in duplicates, and each reaction was grown in 100 ml of LB broth[26]. While adequately scaled protocols exist for whole genome STARR-seq in *Drosophila melanogaster* and humans[20,10,] scaling guidelines for focused libraries are not well reported since their sizes can vary from >7 million target sites[13] to a few thousands[26]. Therefore, it is necessary to report scaling guidelines to set uniform standards for differently sized libraries.

DNA source

A strong advantage of STARR-seq is its ability to screen random fragments of DNA from any source for enhancer activity. To this effect, DNA can be sourced from commercially available DNA repositories[11], from specific populations carrying non-coding mutations or SNPs to be assayed[15], cultured bacterial artificial chromosomes (BACs)[10], or custom oligo pools[17]. Different

sources have their own advantages and drawbacks that are important to consider for efficient library building (Table 1). For example, custom synthesized oligo pools are generally designed to be of uniform length and may include sequencing and cloning adapters as part of the design. These libraries may not require fragment length selection and adapter ligation and are also devoid of any representation bias caused by fragments of non-uniform lengths. However, current limitations for custom made fragments include a synthesis-based length restriction of 230 bp, higher costs, and a smaller library size as opposed to larger focused or whole genome libraries.

Fragment length

Length of each candidate enhancer fragment determines critical experimental parameters such as the extension temperature of ligation-mediated PCR (LM_PCR), sample volume-to-bead ratio for clean-up, insert-to-vector ligation ratio, and fragment molarity for cloning reactions. Most STARR-seq studies either use ~500 bp fragments sourced from sheared whole genome DNA or DNA isolated from BACs, or use <230 bp oligo pools (Fig. 1b). While fragments >500 bp may be more economical and better suited for genome-wide screens or enhancer discovery within larger targets, an important caveat is that larger fragments may result in a loss of resolution on the activity of individual enhancers due to a compounding effect of multiple closely located enhancers or the influence of flanking sequences on transcription factor binding [18,26]. In contrast, focused studies using shorter fragments allow for fine mapping the enhancer effects of individual TF binding motifs. *For example*, Klein and colleagues compared the effects of fragment length on enhancer activity by extending the flanking genomic sequences of the same candidate enhancer sites to create 192 bp, 354 bp, and 678 bp fragments[26]. The authors found less concordance in enhancer activity for the same candidate enhancers when compared between fragments of different lengths, with higher correlation between smaller fragments than larger fragments, further indicating that enhancers within longer fragments are subjected to additional biological effects. Further, library preparation from sheared DNA require accurate size selection methods as fragments of non-uniform length are more prone to uneven representation in the final library due to PCR and cloning biases. In fact, Neumayr and colleagues suggest keeping the fragment lengths within 100-300 bp of each other to avoid extreme biases[20].

Sequencing methodology and choice of vector

The final set of considerations for library design include the choice of sequencing platform and the compatible adapters and kits required for library preparation as well as the choice of the STARR-seq vector. Almost all STARR-seq libraries use Illumina-compatible designs, where the putative enhancer fragment is flanked by two adapters consisting of read 1 and read 2, followed by unique index sequences or barcodes (i5 and i7), and P5 and P7 sequences to facilitate the sequencing-by-synthesis method (Fig. 1c). The two adapters flanking the candidate fragment serve as "constant" sequences for primer annealing during library amplification as well as the primer recognition sites during Illumina sequencing (Fig. 2, Fig. S1). Cloning overhang arms are designed based on the selected STARR-seq vector and are ligated adjacent to the adapters through LM_PCR and used for library cloning. The overhangs are subsequently replaced by the index barcodes i5/i7 and adapters P5/P7 to facilitate sequencing (Fig. 1c). These adapter sequences are especially important for designing oligo libraries, LM_PCR primers, and blocking oligos during hybridization and capture of focused libraries (Fig. S1).

The choice of STARR-seq vector depends on the selected host cell line and the minimal promoter to be used. STARR-seq in *Drosophila*[10,27] and plant-derived cell lines[28,29] have various guidelines on the choice of vector and minimal promoter. For mammalian STARR-seq, Muerdter and colleagues observed that enhancer signals from the original minimal promoter, SCP1 (Super Core Promoter 1) were not consistent due to interference from the ORI promoter present on the same vector. They mitigated this issue by redesigning the vector to carry just the ORI promoter for a more robust assay[30]. In addition to these established vectors, studies have tested alternative promoters and built customized vectors to suit assay needs. For example, Sahu and colleagues used minimal δ1-crystallin gene (Sasaki) promoter for their experiments across multiple libraries. The authors further used a CpG-free vector backbone containing the Lucia reporter gene driven by EF1α promoter to assess the effect of CpG methylation on enhancer activity[31]. All of these features are important pre-experimental design considerations prior to building a robust STARR-seq plasmid library.

Roadmap to successful library construction

With all the design details in place, construction of a STARR-seq plasmid library includes several optimization and quality control steps with each step contributing to varying levels of

assay reproducibility. Here, we highlight key elements of the protocol that significantly alter the outcome and offer clarity and explanations behind various optimizations. The features that impact library building include insert preparation, vector linearization, library cloning, and library amplification (Fig. 2). Additionally, focused libraries also require a library enrichment step to capture the fragments of interest. This enrichment can be done prior to insert preparation using ATAC-seq[13], ChIP-seq[12], FAIRE-seq[14], fragments extracted from specific BACs, or fragments synthesized as custom oligo pools. Alternatively, library enrichment can also be carried out after insert preparation using techniques like capSTARR-seq, which uses special microarrays[32], or by using custom hybridization and capture probes[15].

While these steps typically follow kit-based protocols, there are limited scaling guidelines for differently sized libraries. To ensure replicability of the scaled protocols, details such as the number of replicate reactions performed, sample concentration and purity, details of the purification method used for intermediate product stages, and the experimental parameters must be reported. Additionally, reporting details of validations for intermediate steps such as length verification of fragments following length selection as well as adapter and cloning arm ligation steps provide intermediary checkpoints for researchers attempting similarly designed experiments. It is often difficult to understand whether the library being constructed is successful or not without sequencing the library, which is both time consuming and expensive, and it is also hard to detect and pinpoint experimental errors due to the length of the protocol. Therefore, having validation checkpoints at intermediate stages help address these problems. While performing STARR-seq, we shortlisted five such checkpoints in the protocol that require product assessment and validation before proceeding to the next step (Fig. 2). We also categorized intermediate assay steps as 'highly critical', 'critical' and 'slightly critical' based on the requirements of data reporting to replicate those steps (Fig. 2).

The final determinant of library coverage is the efficiency of molecular cloning and bacterial transformation. Cloning strategies such as In-fusion HD, Gibson assembly, and NEBuilder HiFi DNA Assembly allow for fast and one-step reactions that use complimentary overhang sequences on the inserts and the vector. While these methods have comparable accuracy and efficiency, the chosen method needs to be optimized prior to cloning the final library pool. The critical parameters to consider, optimize, and report include *(i)* purity and concentration of the insert and the vector, *(ii)* insert-to-vector ratio by mass or moles and the

maximum molarity per reaction supported by the cloning enzyme, *(iii)* the number of cloning reactions required and pooled, *(iv)* competent cell strain and volume per reaction, *(v)* the amount of cloned product transformed per reaction and the total number of reactions performed and pooled, *(vi)* the volume of culture used to grow the library, and *(vii)* the number of unique colonies observed in a dilution series for estimating transformation efficiency and library complexity. *For example*, Barakat and colleagues used genomic fragments obtained from chromatin immunoprecipitation to perform multiple PCR reactions and generate inserts for cloning via Gibson assembly. Each reaction was performed in duplicate, which were pooled prior to transformation in DH10B cells. The transformed product was grown in SOC medium, serially diluted, plated, and grown in LB to obtain between 8 and 31 million unique colonies[12]. In contrast, Kalita and colleagues used synthesized oligos for two rounds of nested PCRs to obtain the inserts for cloning using Infusion HD strategy, followed by multiple transformation reactions[17]. For each of these steps, the authors provided the exact volumes for each reaction, elution steps, and cultures, the number of replicates used, and the number of colonies obtained from the dilution series, enabling accurate replication of these experiments.

Complexity of the plasmid library is typically assessed by evaluating sequencing read representation of the target regions. *For example*, Johnson and colleagues performed MiSeq and used the Preseq tool to assess library complexity[19]. The verified STARR-seq plasmid library is then transfected into the host cell for screening. For sequencing, candidate fragments need to be re-amplified from the plasmid pool and fitted with complete P5 and P7 adapters and index barcodes replacing the cloning arms (Fig. 1c). To compare self-transcribed transcripts (output) with fragments present in the plasmid library (input), two sets of sequencing libraries (output and input) are prepared. While there is a general consensus on strategy for building the output library, the input library preparation varies across studies. *For example*, studies from Muerdter and colleagues and Neumayr and colleagues built the input library by directly amplifying the plasmid pool in a second round of LM_PCR, independent of library transfection[30,20]. On the other hand, several studies have isolated the input fragment DNA from a fraction of the transfected cells[10,33] or extracted DNA and RNA from the same cells to account for fragment loss during transfection[26].

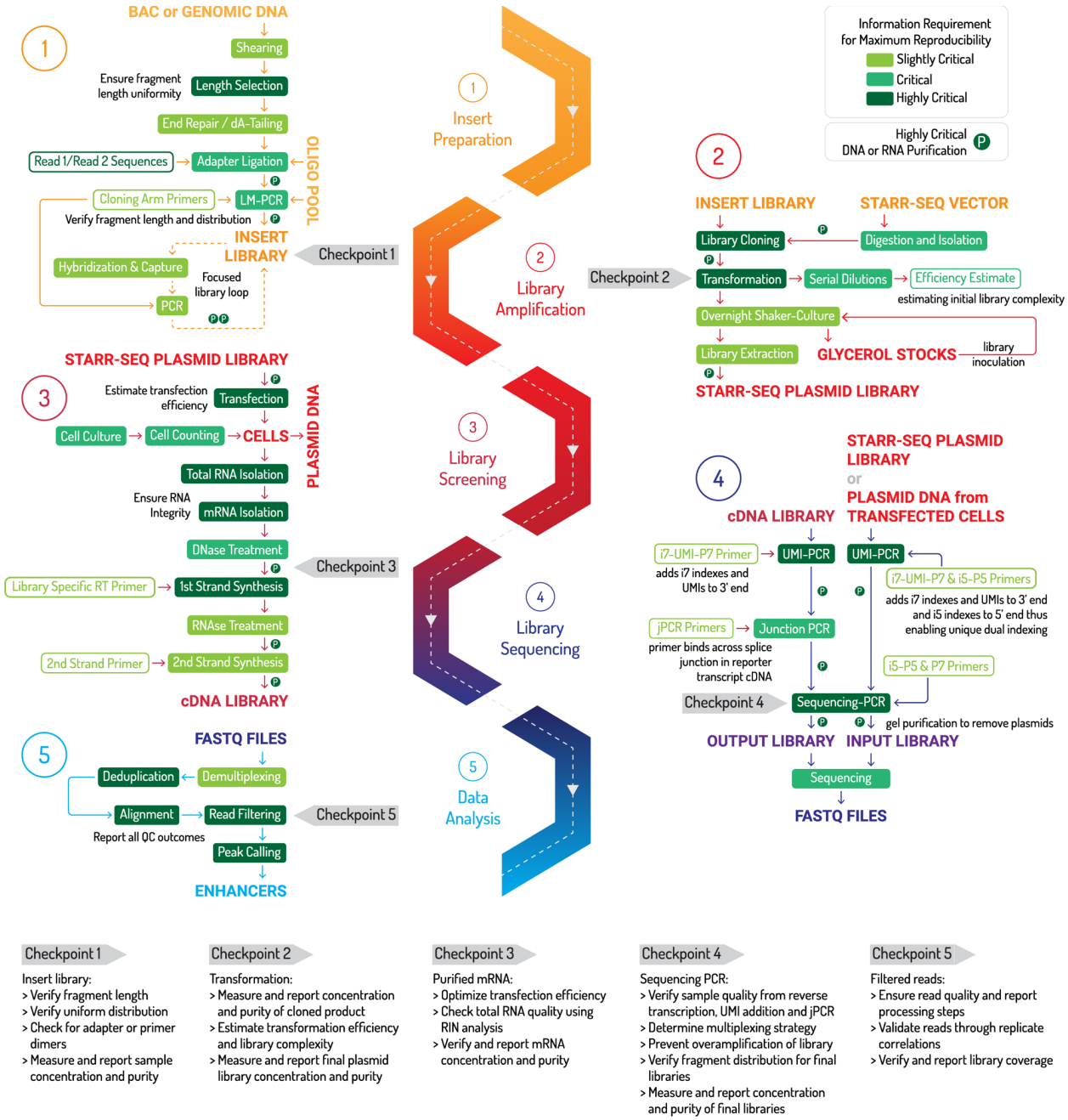


Figure 2: **Roadmap for a successful STARR-seq assay:** Schematic illustrates five major sections of STARR-seq. Individual steps in each section are categorized as **slightly critical (**light green**), critical (**green**), or highly critical (**dark green**)** based on the importance of reporting the methodological detail and intermediate results for a reproducible assay. Each section also has a checkpoint, which may serve as stopping points in the protocol to perform validation of previous steps prior to moving to the next step. Key quality control measures that may be carried out at these checkpoints are also provided in the bottom panel. Of note, section 4 illustrates a

methodology for adding Unique Molecular Identifiers (UMIs) and Unique Dual Indexes. A more detailed schematic along with sequence information is provided in Fig. S1 and Table S2. Alternative methods for UMI addition have also been reported in several studies.

Enhancer screening using STARR-seq

Enhancer screening involves preparation of the STARR-seq 'output' library followed by deep sequencing of the input and output libraries in parallel. The output library is derived from STARR-seq transcripts generated by functional enhancers when transfected into a host cell line. Here, the major design feature is selection of the host cell line based on the desired cellular environment or ease of handling and transfection efficiency. The cell line determines multiple transfection parameters such as *(i)* transfection methodology, *(ii)* amount of plasmid per transfection, *(iii)* the number of cells transfected per library, *(iv)* total number of replicates required for sufficient library complexity, and *(v)* incubation time of transfected cells. These details can vary significantly across studies and determine downstream experimental features of screening. *For example*, Wang and colleagues electroporated 120-130 million Lymphoblastoid Cells (LCLs) per replicate at 5 µg per one million cells and used five replicates for their assay comprising of >7 million unique 150 – 500 bp fragments[13]. Sahu and colleagues transfected 1 µg of each plasmid library of various target sizes per one million cells and used individually optimized reagents for different cell lines[31]. Due to these differences, it is important to consider the exact transfection method and its nuances. For instance, transfection through electroporation is more efficient for harder-to-transfect cells such as Embryonic Stem Cells (ESC) or LCLs and involves shorter incubation time of about 6 hours. However, this process requires higher amounts of plasmid and more cells per library due to the loss of cells during the electroporation process. Methods using lipofectamine may be recommended for various mammalian cell lines such as HEK293T, as they are more efficient and are gentler on the cells, but this procedure involves longer incubation times of 12 to 24 hours. Further, higher transfection efficiency reduces the need for multiple transfections for each biological replicate.

Another important consideration for selecting the cell line is to test whether there is a need for kinase inhibitors to prevent interferon response from the host cell line. Muerdter and colleagues demonstrated that cell lines such as HeLa S3 trigger a type I Interferon (INF-1)

response upon library transfection which disrupts enhancer signals and leads to false results. The authors suggest the use of kinase inhibitors TBK1 and PKR to mitigate the situation[30].

Following transfection, the next series of steps involve total RNA isolation, mRNA isolation and reverse transcription of STARR-seq transcript-specific mRNA using a reporter specific primer for first strand synthesis to build a cDNA library. A majority of studies have used well-described kits for these steps and follow standardized scaling guidelines. However, in order to replicate these steps, authors should report reaction parameters like sample concentrations, amount of starting material for each step, replicate information, and sample pooling details to ensure replicability of the steps. Additional contributors to reproducibility here include intermediate sample validations and quality control assessments including RNA integrity (RIN) analysis and RNA purification using TURBO DNase and RNAClean XP beads, all of which increase library quality. The cDNA library is then amplified using junction PCR primers to remove plasmid artifacts followed by ligation of sequencing adapters and index barcodes using LM_PCR, similar to the input sequencing library (Fig. 2). An important consideration here is to prevent overamplification of the library, which leads to sequencing bias, by adjusting the number of amplification cycles during the final PCR step. Sequencing libraries may be visualized on an agarose gel for a characteristic '*smear*' at the expected fragment length as opposed to a tight band, as demonstrated by Neumayr and colleagues[20].

A common occurrence in STARR-seq data is the presence of PCR duplicates due to the use of PCR-based sequencing library preparation methods. Self-transcribed mRNA may resemble PCR duplicates or vice-versa, leading to data anomalies. Previous studies have used various methods to remove these duplicates by filtering unique reads using computational tools such as PICARD[10,33]. However, for accurate removal of PCR duplicates, Kalita and colleagues added Unique Molecular Identifiers (UMIs) during the reverse transcription step such that PCR duplicate sequences possess the same UMI sequences, unlike self-transcribed mRNA[17]. There are now several published variations for the method of UMI addition. For instance, Neumayr and colleagues suggested adding UMIs following an extra second strand synthesis step for cDNA in a single cycle PCR reaction[20,17].

The final step is sequencing of the input and output libraries. Libraries are usually pooled and sub-sampled prior to sequencing. While these are standardized procedures, small variations in the steps can largely impact data reproducibility. *For example*, while running multiple

STARR-seq samples on the same lane of an Illumina-based sequencer, a phenomenon called 'index hopping' may occur, where index barcodes are assigned to the wrong libraries leading to inaccurate sequencing data[34]. For mitigation, each library should be assigned two indexes or unique dual indexes (Fig. 1c, Fig. 2) allowing for detection of these incidents and increased accuracy of demultiplexing individual libraries from a library pool[35].

STARR-seq analysis and reporting

Despite strictly following experimental design and protocols for reproducing an assay, lack of data guidelines can result in inconsistent findings. The main features that contribute to reproducibility here include sequencing depth, read processing, and read quality control, as well as choice of peak caller, cut-offs used for analysis, and data quality control and validations. While there are well described guidelines and data standards set for sequencing techniques such as RNA-seq[21], ChIP-seq[22], Hi-C[36] and ATAC-seq[23], data analysis for STARR-seq is not well standardized.

A major factor for any high-throughput study is sufficient coverage of the sequenced library. *For example*, the Sequence Quality Control project (SEQC) demonstrated that read depth and choice of analysis pipelines are key aspects for reproducibility of RNA-seq experiments[37]. Read depth and coverage requirements for techniques including ChIP-seq[22] and RNA-seq[38,39,21] have been thoroughly discussed and multiple data standards have been set by consortiums such as ENCODE. On the other hand, STARR-seq read depth requirements vary widely. In fact, even calculating library coverage of STARR-seq data using available RNA-seq guidelines result in inaccurate estimates and faulty inferences of current studies. Whole genome STARR-seq assays typically have lower coverage due to technical limitations, although these assays are highly useful for screening entire genomes for strong and distinct enhancer signals. *For example*, Johnson and colleagues obtained 59X coverage of the human genome in their library[19] while Liu and colleagues reported that 74.3% of their library were covered by at least 10 reads[11]. Smaller focused libraries, especially those aiming to quantify the effect of non-coding mutations and assays aiming to compare enhancer activity changes across multiple libraries aim for much higher coverage to reduce missing of potential enhancer signals. *For example*, Schone and colleagues obtained over 100X coverage of their targets[18]. There is large variation not only in the

sequence depth obtained by different studies but also in the way the results are reported, strongly affecting analysis parameters of the generated data. Hence, there needs to be an established minimum sequencing depth requirement for STARR-seq and reporting guidelines for read details such as read length, number of sequenced reads per library, and sequencing parameters used to generate reproducible data.

After completion of sequencing, raw reads generated from the sequencer are first demultiplexed and the correct sample labels are assigned. Reads are then mapped to a reference genome and filtered for unmapped reads, off-target reads, PCR duplicates, and reads with low mapping quality (based on assigned MAPQ scores). Read filtering results in the loss of a large percentage of reads and reduces the overall coverage of the library. Loss of coverage will reduce the strength of enhancer signals obtained and may result in incorrect quantification of target regions. *For example*, Johnson and colleagues lost up to 40% of their reads after quality control filtering[19]. In fact, in our own assay, which included a modified sequencing library design to include UMIs as well as unique dual indexing (Fig. 3a, Fig. S1), we filtered ~50% of all reads after quality control (Fig. 3b, supplementary methods). Therefore, not only is it important to consider read loss after sequencing during experimental design, but also to report read quality control steps such as filtering parameters and cut-offs used to enable replication of the analysis. The latter could be accomplished by posting the bioinformatics pipelines used for processing STARR-Seq experiments to a public repository such as GitHub. However, most studies overlook these details, while some only report the percentage of reads lost which make it difficult to reproduce their findings.

After ensuring read quality, studies often assessed data quality and reproducibility by calculating the Pearson's correlation coefficients between read counts within library replicates. Additionally, studies also calculated fold change in read counts of RNA from output compared to DNA from input libraries, and tested correlation of fold change between replicates to ensure consistency of data quality for each sample. Following data quality control assessment, several studies also combined replicates for further analysis to aggregate all possible enhancer activities.

To observe enhancer signals, read counts for a particular region of interest is compared between the output and input libraries to determine a 'peak' of concentrated reads indicative of enhancer activity. Previously, peak calling was carried out using the MACS2 algorithm originally developed for ChIP-seq[10] analysis or enhancer activity was demonstrated by

calculating statistically significant fold-change in reads per region between the input and output libraries. Recently, STARR-seq-specific algorithms such as STARRpeaker[40] and CRADLE[41] have been developed. However, a comparison of these methods and an assessment of peak reproducibility using these tools is absent. Here, we used all four methods on reads obtained from our in-house STARR-seq assay and compared the number of peaks and assessed overlaps between different callers (Fig 3c). While STARRpeaker-called peaks overlapped with all other peak callers (demonstrating peak calling robustness), the large variations observed for the number of peaks across different callers suggest a need for additional testing of these tools and guidelines on how to best utilize them, including the potential for considering peaks identified by multiple callers to be most robust for downstream analysis.

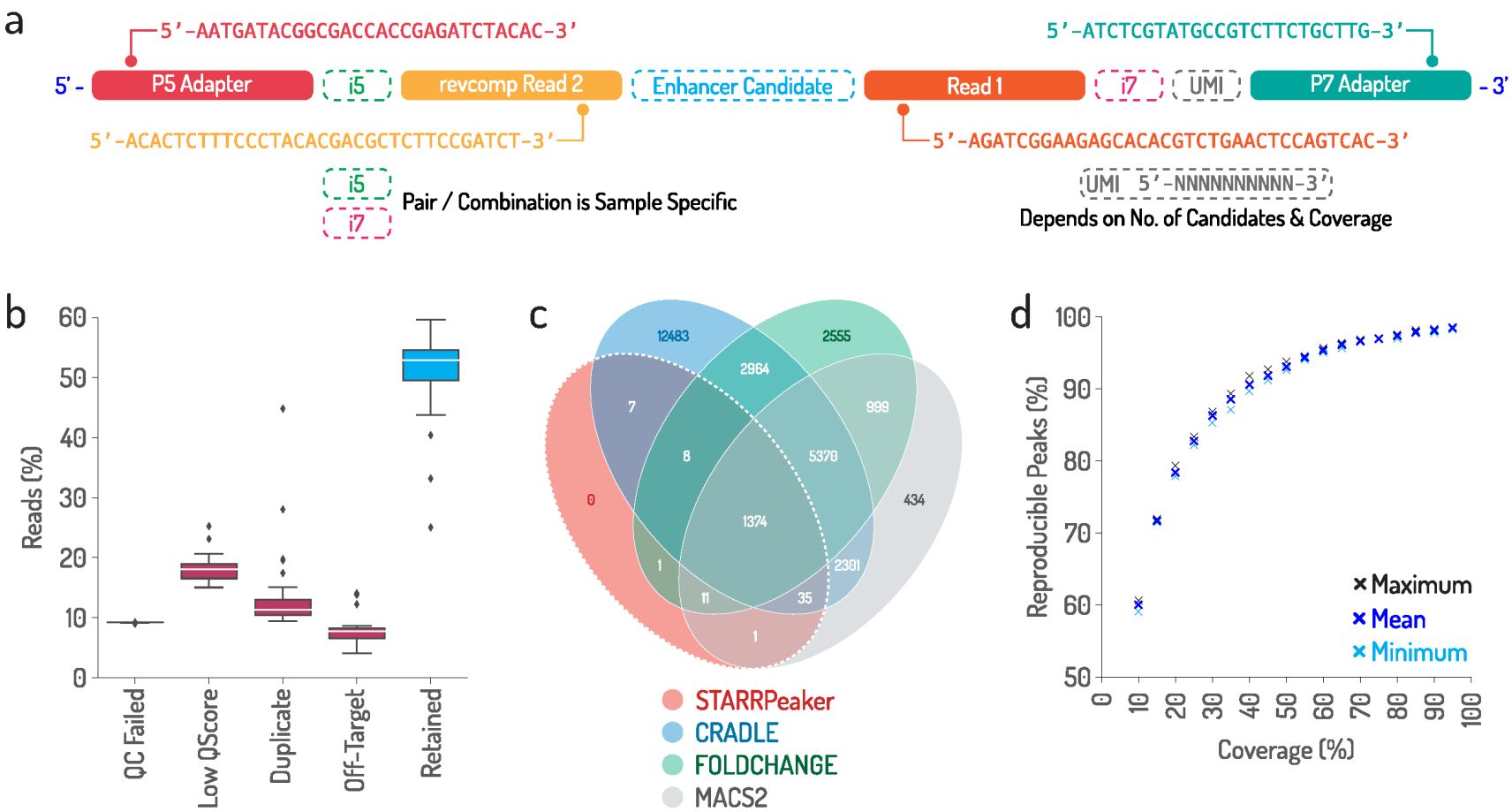

a
5'-AATGATACGGCGACCACCGAGATCTACAC-3'
5'-ATCTCGTATGCCGTCTTCTGCTTG-3'
5' -
P5 Adapter
i5
revcomp Read 2
Enhancer Candidate
Read 1
i7
UMI
P7 Adapter
- 3'
5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
5'-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3'
i5
i7
Pair / Combination is Sample Specific
UMI 5'-NNNNNNNNNN-3'
Depends on No. of Candidates & Coverage
b
Reads (%)
QC Failed
Low QScore
Duplicate
Off-Target
Retained
c
12483
2555
2964
7
999
8
5370
0
434
1374
1
2301
11
35
1
STARRPeaker
CRADLE
FOLDCHANGE
MACS2
d
Reproducible Peaks (%)
Coverage (%)
Maximum
Mean
Minimum

Figure 3: Read architecture, read depth loss, and peak calling comparisons (a) Schematic illustrating read architecture of final STARR-Seq sequencing library (input and output) prior to sequencing. (**b)** Box plots show the percentage of STARR-Seq reads retained and those removed due to various QC filters observed during analysis of in-house STARR-seq assays. (**c)** Overlap of STARR-Seq peaks identified from the same experiment by four different callers. (**d)** Percentage of peaks that were reproduced from in-house STARR-Seq dataset after down-sampling to various coverage thresholds. Analysis details provided in Supplementary Text.

To further investigate the importance of read coverage, we sub-sampled our reads and used STARRpeaker to call peaks to delineate the percentage of peaks retained at lower levels of sequencing coverage (Fig. 3d, Supplementary Methods). We observed that only 60% of the peaks were reproduced at ~56X coverage (~10% of our final coverage of 560X), while 90% of the peaks were reproduced at 224X coverage (~40% of our final coverage). We observed a steady increase in peaks with increase in coverage up to 560X demonstrating the importance of read depth for STARR-seq. Due to these limitations, there is an urgent need for benchmarking of analysis parameters and increased transparency of data reporting from STARR-seq studies into centralized repositories or web portals, as exemplified by the ENCODE consortium[42] for assays such as ChIP-seq, DNase-seq, and RNA-seq, to enable assay robustness and provide clarity for future researchers.

Conclusion

Successful functional genomic studies take years to design and perform, and each step goes through repeated iterations in multiple replicates to optimize parameters for successful outcomes. Even steps involving established kits and protocols need to be scaled according to the study design and often require customizations and intermediate step validations. While each of these factors contribute to overall assay reproducibility, most published studies focus solely on the final outcome and tend to under-explain the methods, optimizations, and intermediate validations, resulting in large gaps of knowledge for researchers attempting to reproduce the results or tailor the study for their own biological questions. Therefore, standardization of protocols and data reporting guidelines would strongly benefit researchers and significantly improve the quality of conducted assays and their publication. Here, we highlighted the different

challenges in performing STARR-seq, a particularly long and difficult assay with huge potential to identify detailed enhancer landscapes and validate enhancer function. We emphasize the importance of reporting details related to biological context, underlying hypothesis, and experimental design, and protocol and bioinformatic analysis parameters to ensure replicability of each step and outcome of the assay.

Box 1: Definitions of frequently used terminology in sequencing-based studies.

1. **Assay reproducibility:** Measure of how accurately and consistently assay results can be re-obtained to re-affirm previous conclusions.
2. **Bead clean-up:** Magnetic DNA or RNA purification beads designed to be selective for DNA or RNA fragments. In DNA purification, the ratio of volume of beads to the volume of sample is selective of the length of DNA being purified. Lower bead volumes fail to bind longer DNA fragments, thereby filtering them out. This is a method for fragment length selection.
3. **Cloning arms or overhangs:** Sequences added to adapter-ligated (read1/read2) inserts that are complementary to cloning sites of the STARR-seq vector. Ligation is accomplished using LM_PCR to facilitate cloning of the fragments into the vector.
4. **De-duplication:** Bioinformatic process for removing PCR duplicates from sequenced reads.
5. **De-multiplexing:** Bioinformatic process by which all NGS reads from pooled libraries are separated into sample-specific reads, based on sample-specific i5/i7 barcodes.
6. **Episomal assay:** Assays that transfect one or more DNA sequences into a host cell whereby the transfected sequences do not integrate within host cell genome, and remain independent of the influence from the host cell genome.
7. **Experimental replicability:** Measure of how accurately an experiment can be repeated with the exact parameters to obtain the same output data.
8. **i5/i7 barcodes or indexes:** These are 6-8 bp sequences ligated to sequencing libraries to provide a unique identity for each sample when pooled with libraries from other samples. Libraries can use single indexing (either i5 or i7 index) or dual indexing (both i5 and i7).
9. **Input library:** STARR-seq sequencing library constructed either directly from plasmid library pool or from plasmid DNA extracted from transfected cells. The number of reads from the input is used to normalize the number of 'output' reads to quantify enhancer activity.
10. **Library complexity:** Representation and coverage of all regions of interest in the library.
11. **LM_PCR: L**igation **M**ediated PCR is a type of PCR that uses primers carrying extended sequences (not having complementary sequence on the target strand) and ligates those extended sequences to the region being amplified.

12. **Massively parallel sequencing:** Next-generation sequencing (NGS) techniques that can sequence billions of reads in parallel based on a sequencing library.
13. **MPRA:** **M**assively **P**arallel **R**eporter **A**ssays are high-throughput assays that quantify the activity of a test fragment, typically cloned downstream of a minimal promoter, via massively parallel sequencing of barcoded reporter transcripts.
14. **Multiplexing:** A process where multiple libraries are pooled and sequenced on the same lane of the sequencer.
15. **Output library:** STARR-seq sequencing library built using self-transcribed reporter transcripts of candidate enhancers extracted from the host cell.
16. **P5/P7 Adapters:** Sequences flanking index (i5/i7) barcodes that help hybridize the fragment to the flow cell to enable sequencing.
17. **Peak calling:** Automated method by which sequenced STARR-seq reads from the output library are compared to reads from the input library to identify regions with significantly increased read counts or 'peaks'.
18. **Read 1/Read 2 adapters:** Sequences immediately flanking the fragment of interest that are sequenced as 'read 1' and 'read 2' on a standard Illumina flow cell. These regions serve as recognition sequences for Illumina sequencing primers.
19. **Sequencing library:** Libraries containing fragments of amplicon DNA or cDNA to be sequenced using NGS. Each fragment consists of a core region of interest flanked by read 1 and read 2 adapter sequences, index barcodes, and P5 and P7 flow cell adapters.
20. **STARR-seq plasmid library:** Library comprising of candidate enhancer fragments cloned downstream of a minimal promoter in a STARR-seq vector.
21. **STARR-seq:** **S**elf **T**ranscribing **A**ctive **R**egulatory **R**egion **Seq**uencing, is a specialized episomal MPRA that is used to directly measure enhancer activity by comparing output transcripts produced by a candidate enhancer sequence cloned downstream of a minimal promoter to the number of copies of the enhancer fragments used as input prior to transcription.
22. **UMIs:** **U**nique **M**olecular **I**dentifiers are randomly synthesized oligos (unknown sequences) of fixed length that can be added to fragments. If fragments undergo amplification via PCR, UMIs added to those fragments prior to PCR are also duplicated during PCR. This enables detection of PCR duplicates, as opposed to identical mRNA self-transcripts.

Box 2: Reproducibility across STARR-seq studies

To assess and quantify reproducibility for STARR-seq based studies, we systematically compared 24 studies[10,11,12,13,14,15,16,17,18,19,30,26,31,33,32,40,43,44,45,46,47,48,49,50] that performed STARR-seq experiments on mammalian or *Drosophila* cells (Fig. 4). We further identified 15 features critical for assay success and assessed their potential for reproducibility across the published studies. We scored each feature from 0 to 4 based on reporting of methodological details, design rationale, and quality control measures, with 0 indicating no detail and 4 indicating reporting of complete detail for each feature that would allow for replication of the assay. A detailed scoring rubric for each feature is provided in Table S1. Our approach provides a method for quantifying the extent of reproducibility of published studies. In particular, we observed that authors typically mentioned steps such as read filtering, mapping tools, peak callers and quality control experiments, but fail to provide associated intermediate data or parameters and cut-off thresholds used for their analyses, thereby reducing replicability of the steps. Furthermore, critical intermediate steps for sequencing such as library multiplexing and pooling that may be carried out at shared sequencing cores are often overlooked, leading to additional disparities in data quality. Therefore, standardization of these assay features and implementation of uniform data reporting guidelines will strongly improve reproducibility of STARR-seq assays.

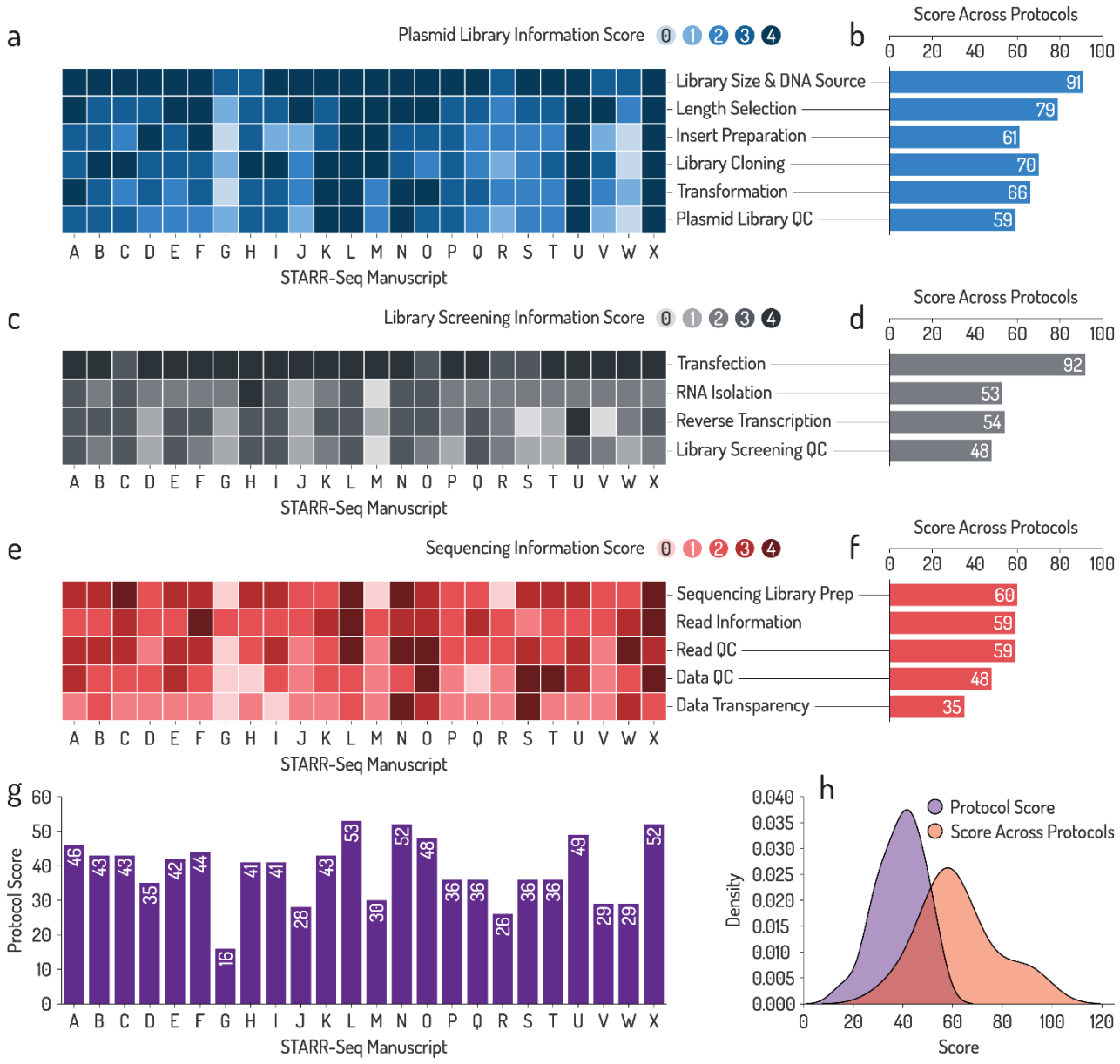


Figure 4: Reproducibility in current STARR seq studies. (a, c, e) Heatmap plot between feature scores for de-identified studies A-X illustrating scores for **(a)** assay design and plasmid library preparation, **(b)** library screening and **(c)** sequencing information. **(b, d, f)** Cumulative scores for each feature across studies illustrating reporting trends and highlighting features that are typically well explained or under-explained. **(g)** Cumulative scores for each assay (A-X) for all features enabling direct comparisons of reproducibility. **(h)** Kernel density estimate plot illustrating distribution of cumulative feature scores and cumulative assay reproducibility scores. We observed that critical design factors such as library scale and source and fragment length or

experimental features such as transfection scored ≥3 across 87.5% - 100% of publications, with studies missing only minor details and explanations. In contrast, features such as library screening QC, data QC and data transparency scored ≥3 across 16.6% - 37.5% of publications.

Table 1. Reproducibility features for biological assays with a focus on STARR-seq.

	Reproducibility factors for a biological assay	**Best practices for reporting reproducibility of a biological assay**	**Considerations and best practices for reporting STARR-seq specific features to maximize assay reproducibility**
1	**Experimental Design**	• Clear definition of the biological question.	• Biological question that would be addressed using STARR-seq such as genome-wide/targeted screening for enhancer activity, or comparison of enhancer activity due to changes in the host cell environment or due to mutations within or near enhancer sites.
		• Rationale for assay design to address the biological question.	• Rationale for choosing specific experimental design parameters including library size, library source, fragment length, vector, sequencing platform and host cell line.
2	**Protocol**	• Details of optimization of protocols, types of kits and reagents used. • Details of protocol scaling, number of replicates used, customizations carried out and data validations.	STARR-seq plasmid library construction • Selection and adaptation of all kits, reagents, and protocols used for insert preparation, library amplification, and confirming sufficient library complexity. Steps such as LM_PCR, library cloning and transformation may require additional optimization to achieve consistency across replicates. • Details of reaction parameters, such as PCR, primer details, number and details of utility of replicates, results of intermediate steps to ensure the protocol is working, and final sample library architecture (with enhancer fragment and sequencing adapters).
			STARR-seq screening • Optimization of host cell culture conditions, transfection methodology, transfection parameters, RNA isolation and reverse transcription kit protocols and associated modifications, and number of replicates required for each step for successful representation of all target sites, and report potential issues accounted for each step and mitigation strategies used (such as fragment loss). • Reaction details (such as reverse transcription), replicates used, primer details, and results for intermediate quality control steps. • For libraries using UMI's, details on UMI ligation strategy and illustration of sequence architecture with UMI position.
			Sequencing • Optimization of parameters for sequencing PCRs to prevent over-amplification, details on selection of multiplexing strategy used, and library pooling. • Details of sequencing parameters such as read length, single- or paired-end, and read depth based on library size and coverage, and read loss due to filtering was accounted.

			• Complete sequencing information including experimental and sequencing run parameters and read details.
3	**Quality Control**	• Sample clean-up and purity assessments. • Quality control and validation of intermediate steps. • Quality control and validation of results.	STARR-seq experimental protocol • Sample clean-up followed by concentration and purity after each experimental step for all replicates. • Assessment of fragment length distribution for insert and vector prior to cloning to check for primer or adapter dimers and fragments of irregular which may lead to PCR or cloning biases. • Assessment of transformation efficiency and library complexity through CFU assessment by plating of serially diluted transformant and control plasmids. • Optimization of transfection efficiency, host cell line immune responses and RIN analysis of isolated RNA. • Assessment of final fragment distribution of libraries prior to sequencing.
			STARR-seq data processing and analysis • Processing and filtering steps for raw reads, including sequence mapping and deduplication. • Assessment of read depth and library coverage as well as read quality across replicates. • All read and data QC parameters used to enable replication of the analysis.
4	**Data analysis**	• Analysis pipeline based on biological question and experimental parameters. • Thresholds and cut-offs for data validation and statistical inferences.	• Read processing steps used for sequence mapping, filtering of bad or unmapped reads, and read deduplication methodology (depending on the use of UMIs) • Peak caller to analyze enhancer activity. • Peak validation method for data QC. • Detailed pipelines for replication of complete analysis to ensure data reproducibility.
5	**Data reporting**	• Experimental design features, protocol, and analysis pipelines. • Intermediate steps to check if the protocol is working and final data.	• Details of library design and pre-experimental optimization steps. • Experimental method with scaling guidelines, protocol parameters, and analysis pipeline. • Details of quality assessment at critical stages of the protocol, and final raw and processed data.

Declarations

Consent for publication

All authors agree and consent for publication of the manuscript.

Competing interests

The authors declare that no competing interests exist in relation to this work.

Acknowledgements

We acknowledge Drs. Matthew Jensen, Yasuhiro Kyono, Istvan Albert, Aswathy Sebastian, Howard Salis, Ross Hardison, Corrine Smolen, and the Penn State Genomics Core Facility for technical support for this project. This work was supported by the NIH grants R01-GM121907, R21-NS122398, and resources from the Huck Institute of the Life Sciences to SG.

References

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

2. Shlyueva, D. *et al.* Hormone-Responsive Enhancer-Activity Maps Reveal Predictive Motifs, Indirect Repression, and Targeting of Closed Chromatin. *Mol. Cell* **54**, 180–192 (2014).

3. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).

4. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).

5. 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).

6. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).

7. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).

8. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–22 (2008).

9. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).

10. Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, A. S. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (80-.).* **339**, (2013).

11. Liu, Y. *et al.* Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol.* **18**, 1–13 (2017).

12. Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**, 276-288.e8 (2018).

13. Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9**, (2018).

14. Chaudhri, V. K., Dienger-Stambaugh, K., Wu, Z., Shrestha, M. & Singh, H. Charting the cis-regulome of activated B cells by coupling structural and functional genomics. *Nat. Immunol.* **21**, 210–220 (2020).

15. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 1–14 (2017).

16. Zhang, P. *et al.* High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat. Commun.* **9**, (2018).

17. Kalita, C. A. *et al.* High-throughput characterization of genetic effects on DNA-protein binding and gene transcription. *Genome Res.* **28**, 1701–1708 (2018).

18. Schöne, S. *et al.* Synthetic STARR-seq reveals how DNA shape and sequence modulate transcriptional output and noise. *PLoS Genet.* **14**, 1–24 (2018).

19. Johnson, G. D. *et al.* Human genome-wide measurement of drug-responsive regulatory activity. *Nat. Commun.* **9**, 1–9 (2018).

20. Neumayr, C., Pagani, M., Stark, A. & Arnold, C. D. STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries. *Curr. Protoc. Mol. Biol.* **128**, e105 (2019).

21. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 1–19 (2016).

22. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).

23. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. ATAC-seq data analysis. *Genome Biol.* **21**, 1–16 (2020).

24. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Challenges for assessing replicability in preclinical cancer biology. *Elife* **10**, 1–32 (2021).

25. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, 1–9 (2015).

26. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).

27. Zabidi, M. A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).

28. Sun, J. *et al.* Global Quantitative Mapping of Enhancers in Rice by STARR-seq. *Genomics, Proteomics Bioinforma.* **17**, 140–153 (2019).

29. Jores, T. *et al.* Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves. *Plant Cell* **32**, 2120–2131 (2020).

30. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).

31. Sahu, B. *et al.* Sequence determinants of human gene regulatory elements. *Nat. Genet.* **54**, 283–294 (2022).

32. Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* **6**, (2015).

33. Peng, T. *et al.* STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol.* **21**, 1–27 (2020).

34. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, 1–8 (2012).

35. MacConaill, L. E. *et al.* Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, 1–10 (2018).

36. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* **72**, 65–75 (2015).

37. Su, Z. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).

38. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Res.* **21**, 2213–2223 (2011).

39. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).

40. Lee, D. *et al.* STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol.* **21**, 1–24 (2020).

41. Kim, Y. S. *et al.* Correcting signal biases and detecting regulatory elements in STARR-seq data. *Genome Res.* **31**, 877–889 (2021).

42. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

43. Verfaillie, A. *et al.* Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. *Genome Res.* **26**, 882–895 (2016).

44. Brandt, M. M. *et al.* Chromatin Conformation Links Distal Target Genes to CKD Loci. *J. Am. Soc. Nephrol.* **29**, 462–476 (2018).

45. Klein, J. C., Keith, A., Agarwal, V., Durham, T. & Shendure, J. Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.* **19**, 1–13 (2018).

46. Van Ouwerkerk, A. F. *et al.* Identification of Functional Variant Enhancers Associated with Atrial Fibrillation. *Circ. Res.* **127**, 229–243 (2020).

47. van Weerd, J. H. *et al.* Trait-associated noncoding variant regions affect tbx3 regulation and cardiac conduction. *Elife* **9**, 1–26 (2020).

48. Selvarajan, I. *et al.* Integrative analysis of liver-specific non-coding regulatory SNPs associated with the risk of coronary artery disease. *Am. J. Hum. Genet.* **108**, 411–430 (2021).

49. Glaser, L. V. *et al.* Assessing genome-wide dynamic changes in enhancer activity during early mESC differentiation by FAIRE-STARR-seq. *Nucleic Acids Res.* **49**, 12178–12195 (2021).

50. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).

Supplementary Files

Challenges and considerations for reproducibility of STARR-seq assays

Maitreya Das[1, 3, 5*], Ayaan Hossain[3, 5], Deepro Banerjee[4, 5], Craig Alan Praul[5] and Santhosh Girirajan[1, 2, 3, 4, 5*]

1. Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802
2. Department of Anthropology, Pennsylvania State University, University Park, PA 16802
3. Molecular and Cellular Integrative Biosciences Graduate program, Pennsylvania State University Park, PA 16802
4. Bioinformatics and Genomics Graduate program, Pennsylvania State University, University Park, PA 16802
5. Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802

Table of Contents

1. Supplementary Text 3
2. Supplementary Methods 4
   Read deduplication, alignment and filtering 4
   Peak calling 4
   Comparing number of peaks by varying library coverage 4
3. Supplementary Figures 5
   Figure S1A 5
   Figure S1B 6
   Figure S1C 6
   Figure S1D 7
   Figure S1E 8
4. Supplementary Tables 9
   Table S1A 9
   Table S1B 10
   Table S1C 11
   Table S2 12
5. STARR-seq Protocol 14

Supplementary text

We conducted eight targeted STARR seq assays in HEK293T cells using three replicates per assay. While conducting the assay we came across various design and protocol inconsistencies that enabled us to generate a list of best practice guidelines and quality control checkpoints. We also evaluated 24 STARR-seq studies and scored important assay features for each study based on the number of details reported by the authors (see **Table S1** for scoring rubric). We also used our data to demonstrate the effects of read filtering, significance of read depth, and compared different peak callers. In the supplementary text, we provide details of our analysis and a working protocol used for our in-house STARR-seq experiments (see also **Table S2, Fig. S1**).

Supplementary methods

1. Read deduplication, alignment and filtering

We used custom scripts to remove all PCR duplicates from raw FASTQ files of STARR-seq input and output libraries obtained after demultiplexing. Then, we aligned the paired-end reads to GRCh38 human genome assembly using BWA-MEM[1] with default settings. Next, we removed reads that were (i) unaligned, (ii) low quality (mapping quality score<30), (iii) multi-mapped (reads that mapped to multiple locations with equal confidence), and (iv) off-target using SAMtools[2] with the following parameters: -F 2828 -f 2 -q 30.

2. Peak calling

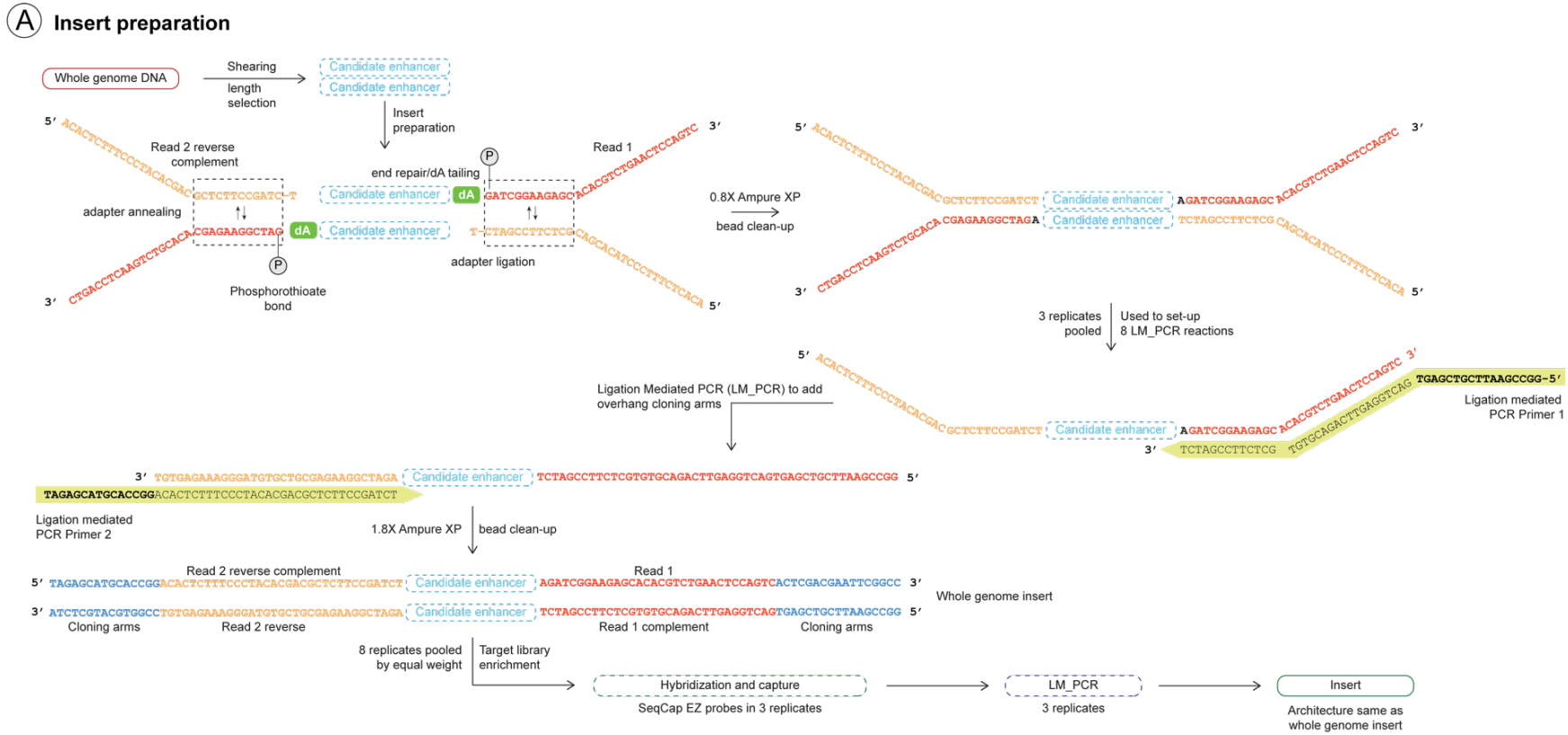
We called peaks using four methods: calculated fold change/enrichment using custom scripts, or using previously published tools such as MACS2[3], STARRPeaker[4], and CRADLE[5]. We first merged alignments from each replicate for each STARR-seq input and output library prior to calling peaks. For fold change-based peak calling method, we fragmented all regions in our libraries using a sliding window of size 500 bp and a stride of 50 bp. Then, we calculated the input and output library coverage for each of these windows using BEDTools[6]. After normalizing the library coverage by reads per million (RPM), we calculated the fold change of the output library compared to the input library, and windows with fold-change ≥1.5 were selected. Finally, we merged all overlapping windows to get a final set of enhancer regions.
We also performed peak calling using MACS2[3], STARRPeaker[4] and CRADLE[5] using the merged input library alignments as "controls" and merged output library alignments as "treatments" using default settings. We used BEDTools[6] to intersect CRADLE-called "active" peaks with peaks called by other callers.

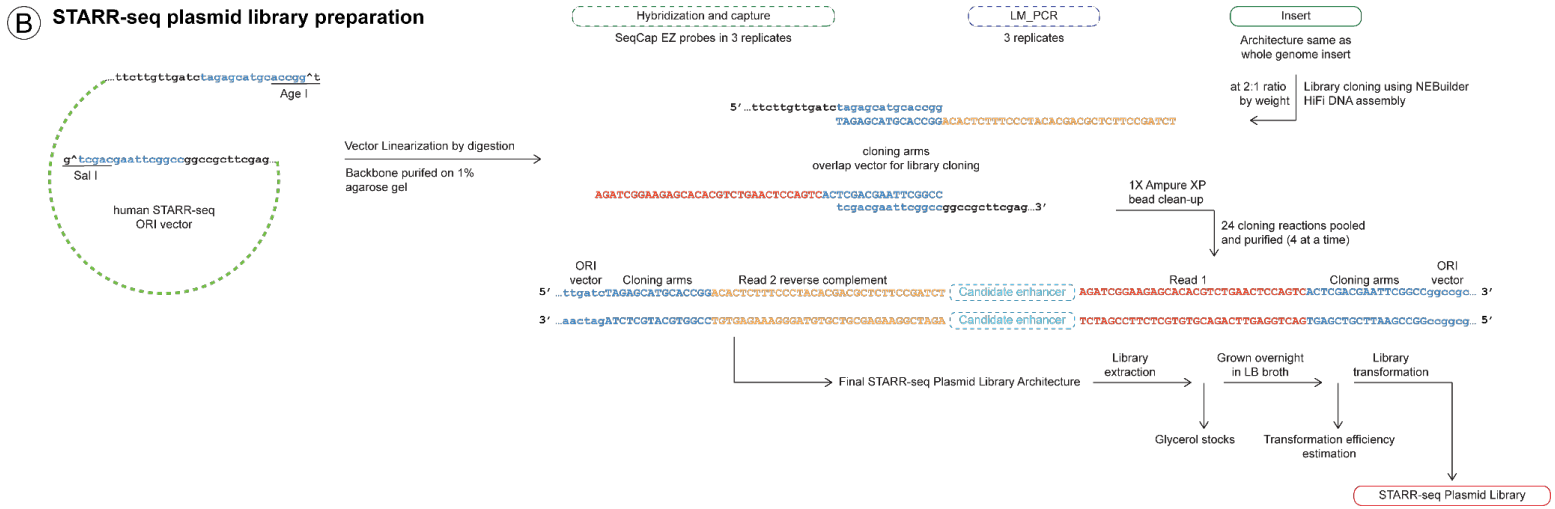
3. Comparing number of peaks by varying library coverage

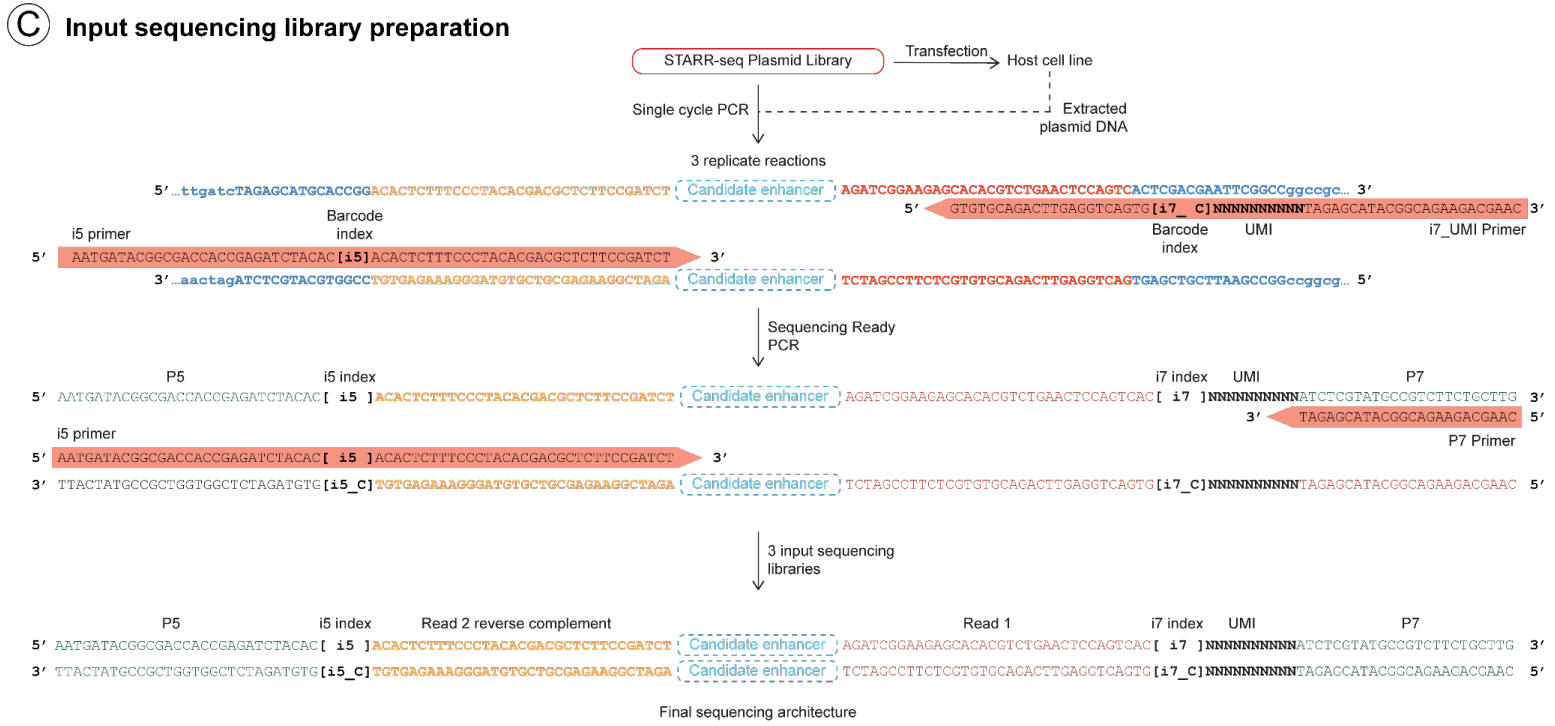
We randomly subsampled the input and output libraries from 10% of the original dataset to 90% with increments of 10% using SAMtools[2]. For each subsample, we also created three replicates by changing random seed parameter. We called peaks in the subsampled libraries using STARRPeaker[4] with the default settings.

Supplementary Figures

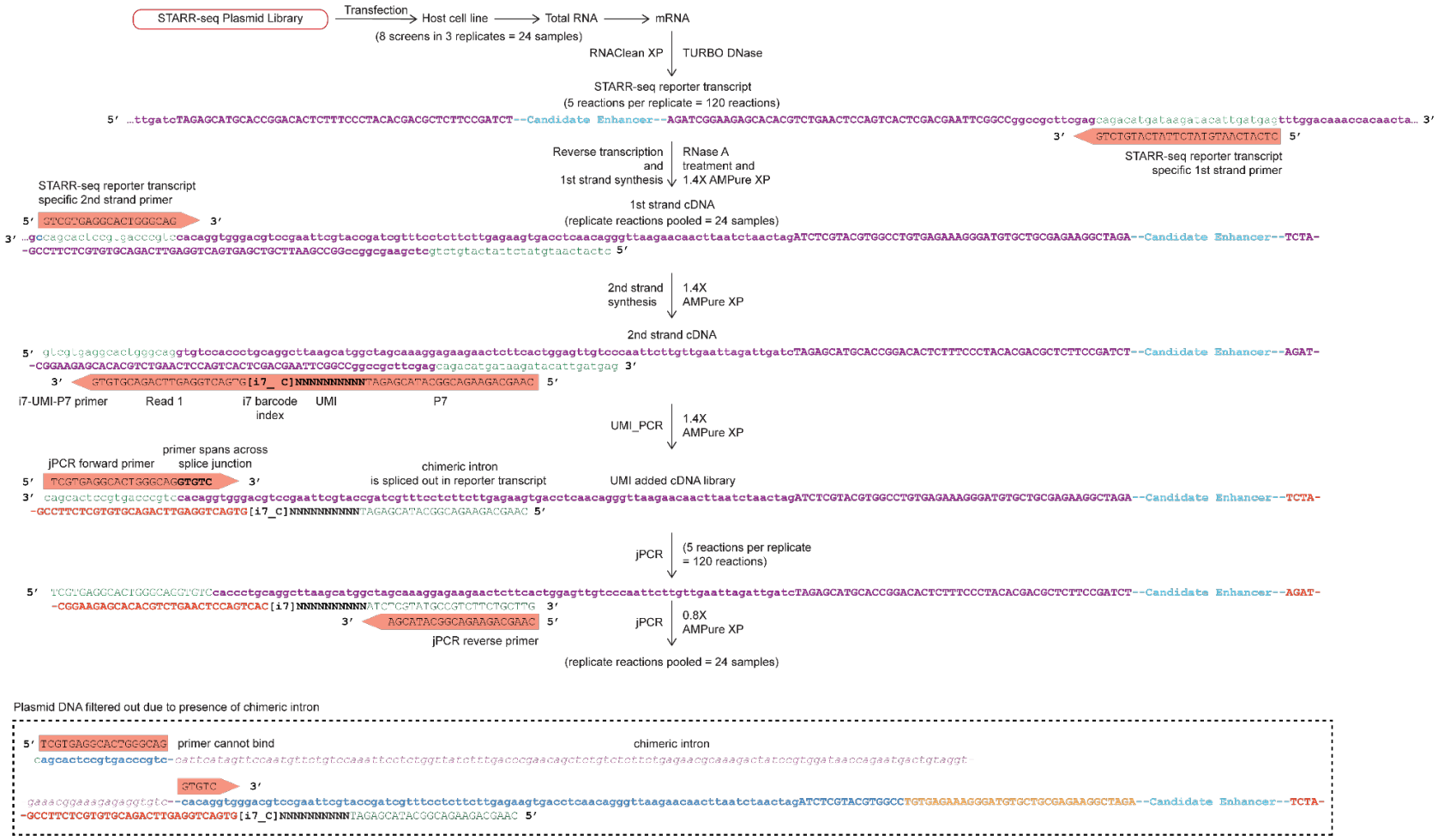
Figure S1 (A-E): Schematic representation of a STARR-seq protocol highlighting all sequence information.




B
STARR-seq plasmid library preparation
Hybridization and capture
SeqCap EZ probes in 3 replicates
LM_PCR
3 replicates
Insert
Architecture same as whole genome insert
at 2:1 ratio by weight
Library cloning using NEBuilder HiFi DNA assembly
Age I
Sal I
human STARR-seq ORI vector
Vector Linearization by digestion
Backbone purifed on 1% agarose gel
cloning arms overlap vector for library cloning
1X Ampure XP bead clean-up
24 cloning reactions pooled and purified (4 at a time)
ORI vector
Cloning arms
Read 2 reverse complement
Candidate enhancer
Read 1
Final STARR-seq Plasmid Library Architecture
Library extraction
Grown overnight in LB broth
Library transformation
Glycerol stocks
Transformation efficiency estimation
STARR-seq Plasmid Library


C
Input sequencing library preparation
STARR-seq Plasmid Library
Transfection
Host cell line
Single cycle PCR
Extracted plasmid DNA
3 replicate reactions
Candidate enhancer
i5 primer
Barcode index
Barcode index
UMI
i7_UMI Primer
Sequencing Ready PCR
P5
i5 index
i7 index
UMI
P7
P7 Primer
3 input sequencing libraries
Read 2 reverse complement
Read 1
Final sequencing architecture

D cDNA library generation and UMI addition



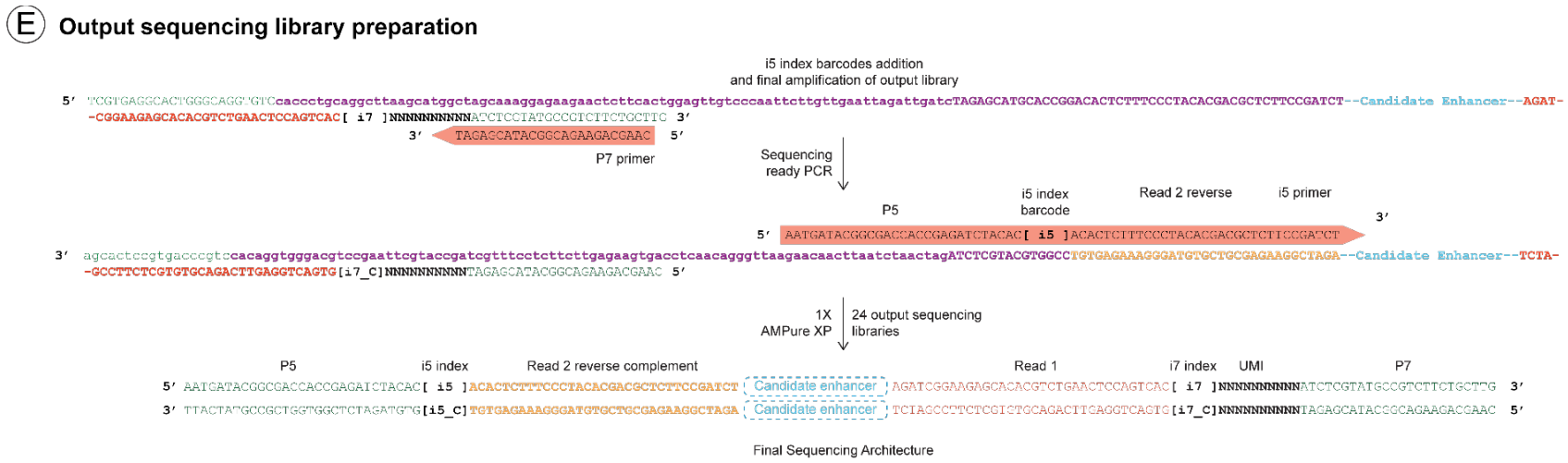


Figure S1: Schematic of STARR-seq experimental protocol and illustrating complete sequence information is shown. **(A)** Insert preparation including addition of sequencing adapters and cloning arms to library fragments to facilitate sequencing and library cloning. **(B)** STARR-seq plasmid library preparation including vector linearization and cloning of library inserts into human STARR-seq vector[7] followed by library amplification through transformation. **(C)** Preparation of 'input' sequencing library either directly from plasmid library or from DNA extracted from library-transfected host cells. This step adds on UMIs and index barcodes to the library prior to sequencing to sort for PCR duplicates and to sequence multiple libraries on the same sequencing lane (multiplexing). **(D)** cDNA library generation for STARR-seq screening includes transfection of the plasmid library into a host cell and reverse transcription of self-transcribed reporter transcripts. This step also involves adding UMIs for detecting and removing PCR duplicates. **(E)** Preparation of 'output' sequencing library involves adding sequencing barcode indexes to the screening library prior to sequencing to enable multiplexing. Both 'input' and 'output' libraries are pooled and sequenced in parallel for enhancer screening.

Supplementary Tables

1. **Table S1 (A-C):** Rubric used for scoring based on details provided for each feature

Steps	**Rubric for each score based on details provided**				
	0	**1**	**2**	**3**	**4**
Library size and DNA source	No detail	Only mentions genome wide/focused. Doesn't connect with goal of experiment	Library target justified and answers research question	Provides details on building target library including source of DNA and enrichment (if focused)	Explains scientific logic and provides complete detail on choosing library size, type, DNA source, enrichment.
Length selection	No detail	Fragment length	Fragment length and method of size selection	Justification of length and selection	Validation of selection
Insert Preparation	No detail	Mentions steps involved	Adapter ligation kit and LM_PCR reaction details including primers	Adapter ligation protocol parameters and number of reactions	Complete information on adapter sequences, explanation for choice
Library cloning	No detail	Insert source and vector name	Kit/reagent used	Total number of reactions, pooling information and purification steps	Insert and vector ratio with concentration and amount and optimization steps
Transformation	No detail	Transformation methodology	Competent cell name	Amount of competent cell and ligated product per reaction and number of reactions	Validation of transformation, complexity assessment, QC and optimization
Plasmid Library QC	No detail	Validation of any 1 step or QC checkpoint	Validation of any 2 steps or report both QC checkpoints	Validation of any 3 steps or provide intermediate reports for both QC checkpoints	Validation of all steps and report complete details for both checkpoints

Table S1A: Rubric used for scoring each feature for each study for assessing plasmid library information (Fig. 4 a, b).

Transfection	No detail	Cell type	Transfection methodology	Transfection protocol	Transfection efficiency, validation and optimization
RNA Isolation	No detail	Different sub-stages of RNA (total RNA, mRNA prior to RT	RNA extraction kits used for each sub-stage execution	Number of cells used, amount of final RNA, and description of protocol of kits used	Optimization and validation of RNA extracted at each sub-stage, purity and RNA integrity
Reverse Transcription	No detail	Reaction kit, primer and parameters	Amount of starting material and number of reactions.	Explanation for use of kit parameters	Validation and purification of cDNA obtained and optimizations.
Library screening QC	No detail	Explanation and validation of any 1 step	Explanation and validation of any 2 steps or report QC checkpoint	Explanation and validation of any 3 steps or report intermediate data for QC checkpoint	Explanation and validation of all steps and provide complete details for checkpoint

Table S1B: Rubric used for scoring each feature for each study for assessing library screening information (Fig. 4 c, d).

Sequencing Library preparation	No detail	Parameters of sequencing ready PCR reactions	Kit detail and protocol details for sequencing libraries (input and output)	Number of reactions, replicate information	Index information, lane and pooling information, final validation and amount sent for sequencing
Read Information	No detail	Mention sequencing platform and GEO	Read length, type of sequencing	Adapter type, total number of reads	Complete adapter sequence and reads per replicate
Read QC	No detail	Read QC steps reported in manuscript or supplementary document	Read QC tools used along with their parameters, raw data processing commands and post processing commands (after peak call) if applicable reported in manuscript or supplementary document	Final read depth or coverage per library reported in manuscript or supplementary document	Read loss due to mapping quality, N bases and bad/off-target reads reported in manuscript or supplementary document and enough information to calculate read depth or library coverage by a reader
Data QC	No detail	Correlation of replicates within input and/or output libraries reported in manuscript or supplementary document	Correlation across replicates within input and output libraries and details about transfection efficiency reported in manuscript or supplementary document	Correlation of enhancers reads between replicates and/or correlation of fold changes between replicates reported in manuscript or supplementary document	Number of reproducible peaks called across replicates and/or by sub-sampling filtered read file, reported in manuscript or supplementary document
Data transparency	No detail	Raw data and final enhancer activity data uploaded to public repository e.g. GEO	Filtered/Processed reads (after read QC) uploaded to public repo OR source code provided in public repo to reproduce filtered read file	Intermediate QC data uploaded to public repo OR source code provided in public repo to reproduce intermediate QC files	Complete analysis pipeline starting from raw data processing to analysis post peak calling uploaded to public repo. The source code should be sufficient to reproduce final results and QC data at all intermediate steps

Table S1C: Rubric used for scoring each feature for each study for assessing library sequencing information (Fig 4 e, f).

2. **Table S2:** List of oligo sequences used for in-house STARR-seq assays.

	Oligo name	Sequence (5' - 3')	Function
1	Adapter I	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTC ()	A' -- Read 1 sequence -- 'A'
2	Adapter II	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	Reverse Complement of Read 2
3	LM_PCR forward	**TAGAGCATGCACCGG**ACACTCTTTCCCTACACGACGCTCTTCCGATC*T	Cloning overhang in bold; *: phosphorothioate bond
4	LM_PCR reverse	**GGCCGAATTCGTCGAGT**GACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	Cloning overhang in bold
5	Blocking oligo I	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCCGGT	For hybridization and capture
6	Blocking oligo II	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-Amm	For hybridization and capture; Amm: modification
7	1st strand primer_ORI	CTCATCAATGTATCTTATCATGTCTG	Reverse transcription of STARR-seq reporter specific transcripts
8	2nd strand primer_ORI	GTCGTGAGGCACTGGGCA*G	Reverse transcription of STARR-seq reporter specific transcripts

9	i7_UMI_P7 primer	CAAGCAGAAGACGGCATACGAGAT**NNNNNNNNNN[i7_RC]**GTGACTGGAGTTCAGACGTGT*G	N: UMIs; i7: Unique Dual Index barcode (in bold) (NEB #E6440S manual); RC: reverse complement
10	jPCR forward	TCGTGAGGCACTGGGCA**G*G***T*G*T*C	Primer extends over spice junction (in bold)
11	jPCR reverse	CAAGCAGAAGACGGCATACG*A	Binds to P7 sequence
12	i5 primer	AATGATACGGCGACCACCGAGATCTACAC**[i5]**ACACTCTTTCCCTACACGACGCTCTTCCGATCT	i5: Unique Dual i5 Index barcode (NEB #E6440S manual)
13	P7 primer	CAAGCAGAAGACGGCATACGAGA*T	Binds to P7 sequence

Table S2: Sequence information for all sequences used for in-house STARR-seq assays. Sequences were ordered from Integrated DNA Technologies, Inc for use. Dilutions required for each oligo is provided in STARR-seq protocol.

STARR-SEQ PROTOCOL (Girirajan lab)

This mammalian (human) STARR-seq protocol is based on the protocol reported by Muerdter and colleagues[7] and Neumayr and colleagues[8] from Dr. Alexander Stark's lab, as well as the SeqCap EZ HyperCap Workflow by Roche (Roche Sequencing Solutions, Inc, CA 94588, USA) along with various modifications and adaptations. The major steps of the protocol are provided below, followed by detailed description of the steps.

(1) STARR-seq plasmid library preparation

(A) Insert preparation: End repair and dA tailing
(B) Insert preparation: Adapter ligation
(C) Insert preparation: LM_PCR
(D) Hybridization and Capture
(E) Vector preparation: Vector culturing
(F) Vector preparation: Vector linearization
(G) Library amplification: Library cloning
(H) Library amplification: Transformation
(I) Library amplification: Library storage and extraction

(2) STARR Seq screening

(A) Culturing of cells
(B) Transfection
(C) Total RNA isolation
(D) mRNA isolation
(E) TURBO DNase treatment
(F) RNAClean XP treatment
(G) cDNA library preparation: 1st strand synthesis
(H) RNase A treatment and AMPure XP clean-up
(I) UMI addition: 2nd strand synthesis
(J) UMI addition: UMI_PCR
(K) Junction PCR
(L) Sequencing Ready PCR: Output library preparation
(M) Sequencing Ready PCR: Input library preparation

(1A) Insert preparation: End repair and dA tailing

Before starting:

- The first section of this protocol follows the SeqCap EZ HyperCap workflow by Roche. This protocol replaces the NEBNext Illumina library preparation protocol reported by Neumayr and colleagues[8] to incorporate hybridization and capture of the target library using SeqCap EZ Prime Choice XL probes (Roche catalog # 08247510001).
- The required reagents for this protocol include KAPA Hyper Prep reaction kit (Roche catalog #KK8500) that consists of End repair and A-tailing buffer, End repair and A-tailing enzyme, ligation buffer, DNA ligase and KAPA HiFi HotStart ReadyMix (2X). This kit also contains primer mixes for standard illumina adapters however, use custom adapter sequences provided in supplementary table S2. Order additional polymerase for subsequent PCR steps KAPA HiFi HotStart ReadyMix (2X) (Roche catalog #07958935001)
- Prepare AMPure XP beads (Beckman Coulter catalog #A63881) for sample clean-up in 1.5ml Eppendorf tube (VWR catalog #87003-294) to equilibrate to room temperature for at least 30 mins.
- DNA is fragmented through sonication at core facility and then selected at specified length using blue pippin and fragment length distribution is verified on a bioanalyzer and provided in 5 tubes at ~ 33 ng/µl. Library is sheared and selected to ~ 500 bp.
- Take equal amounts (in ng) from each tube and pool to a total of 260ng DNA starting material for end repair and dA tailing.
- Starting DNA: 260 ng made up to 50 µl with ultra-water in 0.2 ml tubes (VWR catalog #20170-012) in 2 replicates.

End repair and dA tailing protocol:

1. Thaw reagents on ice. Assemble following reaction on ice. Mix thoroughly by pipetting and light tapping. Spin down and keep on ice.

	Reagent	Volume
1	KAPA End Repair & A-Tailing Buffer	7µl
2	KAPA End Repair & A-Tailing Enzyme mix	3µl
3	DNA + water	50µl
	Total	60µl

2. Incubate reactions on thermocycler (program: end_repair)

	Temperature	Time
1	20°C	30 mins
2	65°C	30 mins
3	4°C	hold

- Continue to adapter ligation without stopping.

(1B) Insert preparation: Adapter Ligation

1. Anneal adapter oligos: Resuspend adapter I and II sequences to 100 µM stock (All adapter sequences provided in supplementary table S2). Order fresh oligo sequences from IDT (Integrated DNA Technologies, Inc, IA-52241) for new library. Use NEBuffer2 (New England Biolabs catalog # B7002S).
2. Assemble following reaction in a 0.2 ml tube.

	Reagent/Oligo	Volume
1	Adapter I	10 µl
2	Adapter II	10 µl
3	NEBuffer 2	4 µl
4	Ultra-pure water	16 µl
	Total	40 µl (25µM)

3. Incubate reaction at 95°C for 5 mins on thermocycler. Keep on bench for 2 hours to cool slowly to room temperature.
4. Add 2 µl ultra-pure water + 3µl adapter mix for final conc at 15 µM for KAPA protocol. (5 µl adapter per reaction in the next step)
5. Thaw out adapter ligation reagents on ice and assemble following reaction in 0.2 ml tubes.

	Reagent/Oligo	Volume
1	Ultra-water	5 µl
2	KAPA Ligation Buffer	30 µl
3	KAPA DNA Ligase	10 µl
4	End Repair/dATail product **(Step A)**	60 µl
5	Adapters (15 µM)	5 µl
	Total	110 µl

6. Mix samples thoroughly by pipetting and spin down. Incubate in thermocycler (program: Adapter lig)

	Temperature	Time
1	20°C	15 mins

7. Clean-up reaction using 0.8X AMPureXP bead clean-up (protocol below).

AMPure XP Clean-up notes

- These are magnetic beads that bind to specific lengths of DNA depending on the ratio of bead volume and sample volume.
- Beads must be equilibrated to room temperature prior to use for at least 30 mins.
- Use only freshly prepared 80% ethanol (diluted in ultra-pure water) for wash steps.
- Ethanol wipe bench and magnetic rack prior to use.
- All further AMPure XP clean-up steps will follow the same protocol with different volume ratios and elution reagent type and volume.

AMPure XP Clean-up protocol:

1. Pipette appropriate amount of beads into sample tube. For 0.8X clean-up, pipette 88 µl beads into each 110 µl adapter ligation reaction and transfer mix to 1.5 ml Eppendorf tube.
2. Mix beads and sample thoroughly by vortexing for 10 secs and pipetting vigorously. Spin down the tube once finished.
3. Incubate mix for 5-10 mins at room temperature. Set-up magnetic rack during incubation and prepare 5 ml 80% ethanol.
4. Place tubes on magnetic rack and incubate for 5 mins (till a clear solution is observed). Carefully remove supernatant liquid without disturbing the beads. (Keep tube on the rack while discarding supernatant)
5. Add 200 µl 80% ethanol and incubate for 30 secs to 1 min and then remove promptly. Repeat this step. After 2 washes, leave tube on rack to dry.
6. If residue ethanol is scattered on the inner tube surface, spin down tube for 1-2 secs and then use P20 pipette to remove trace ethanol. Dry tube for 2-3 mins or till no liquid visible.
7. Do not over-dry. The beads should have a moist liquid coating and not crack up.
8. Remove tube from rack and add 40 µl 10mM Tris-HCl at pH 7.5 for elution. Mix thoroughly with beads and vortex lightly and spin down.
9. Incubate for 5 mins. Place tube back on magnetic rack. Wait till clear solution obtained and pipette solution into a fresh 1.5 ml Eppendorf tube.
10. Measure sample purity and concentration. Typical concentration observed is ~ 40 ng/µl but purity ratios will be high (>3 for both 260/280 and 260/230).
11. Use 3 µl adapter ligated DNA for subsequent LM_PCR step.

(1C) Insert preparation: Ligation Mediated PCR (LM_PCR)

- This step adds overhang arms to the adapter ligated fragments to facilitate library cloning.
- Use custom primers (LM_PCR forward and LM_PCR reverse) designed for human STARR-seq ORI vector[7] as provided in supplementary table S2.
- Order fresh primers from IDT for new library. Reconstitute oligos to 100 µM.

LM_PCR protocol:

1. Dilute primers and make primer pool according to the following:

	Reagent/Oligo	Volume
1	LM_PCR forward	2 µl
2	LM_PCR reverse	2 µl
3	Ultra-pure water	16 µl
4	Total	20 µl

2. Thaw out LM_PCR reagents on ice and assemble following reaction on ice in 0.2 ml PCR strip tubes.

	Reagent/Oligo	Volume
1	Adapter ligated DNA (Previous step)	3µl
2	Primer Mix	5µl
3	Ultra-water	17µl
4	KAPA HiFi HotStart Ready Mix	25µl
	Total	50µl

3. Gently mix and spin down and incubate in Thermocycler (Protocol: LM_PCR)

	Temperature	Time	Step
1	98°C	45 secs	Initial denaturation
2	98°C	15 secs	Denaturation
3	65°C	30 secs	Annealing
4	72°C	30 secs	Extension
5	72°C	60 secs	Final extension
6	4°C	Hold	

x 8

4. Clean-up reactions using 1.8X AMPure XP bead clean-up. (Use 90 µl beads for each 50 µl reaction. Elute samples in 30 µl of ultra-pure water.

5. Measure concentration and purity. Typical yield is ~ 10-20 ng/µl.

6. **Checkpoint 1:** Assess samples on bioanalyzer. Send 3 µl of each sample to assess purity and fragment length distribution. Fragments should show a steady increase in length based on the size of the adapter sequences and serves as the first validation of library preparation. This also shows presence of adapter dimers or primer dimers as well as potential PCR biases.

7. Samples can be stored at -20°C.

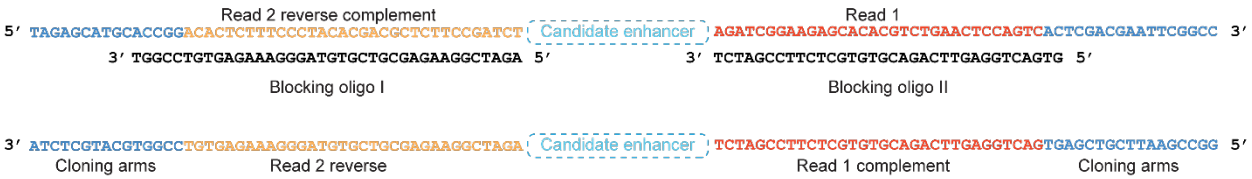
8. For whole genome libraries, repeat several replicates of LM_PCR as directed in protocol reported by Neumayr and colleagues (Neumayr et al 2019).

9. For focused libraries using hybridization and capture of target sites, perform at least 8 replicates. Assess each replicate for quality and length distribution.

10. Pool equal amounts from each replicate for hybridization and capture.

(1D) Hybridization and Capture

- This step allows for selectively capturing target library using custom designed hybridization and capture probes. The required kits needed for this include
  (1) SeqCap EZ Prime Choice XL probes (Roche catalog # 08247510001)
  (2) SeqCap EZ Hybridization and Wash kit (Roche catalog # 05634261001)
  (3) SeqCap EZ Pure Capture Bead Kit (Roche catalog # 06977952001)
  (4) Blocking Oligos custom designed (IDT), sequences provided in supplementary table S2
  (5) COT Human DNA (Sigma Aldrich catalog #11581074001)
- Upon receipt of probes, immediately aliquot into 4.5 µl aliquots in 0.2 ml tubes and store at -20°C. Each aliquot will be sufficient for 1 capture reaction.
- To prepare blocking oligos, reconstitute oligos to 400 µM. Use 2.5 µl (1000 pmol) of each for final reaction. Here is a schematic outline on the role of blocking oligos during hybridization and capture.



- Prepare pooled LM_PCR product from multiple replicates and measure final concentration and purity. Starting material should be ~ 1 µg of DNA.
- Prior to starting, thaw capture reaction aliquot on ice.
- Bring AMPure XP beads to room temperature for at least 30 mins for clean-up.

Hybridization protocol:

1. Prepare following reaction mix in a 1.5 ml Eppendorf tube.

	Component	Volume
1	COT Human DNA	5 µl
2	Blocking Oligos	2.5 + 2.5 µl
3	LM_PCR Pool	1 µg

2. Calculate total volume of the mixture and add 2X AMPure XP beads to mix and vortex for 10 seconds and incubate at room temperature for 10 mins.
3. Place tube onto magnetic rack and discard supernatant when it clears.
4. Wash beads with 190 µl 80% ethanol for 30 secs and remove and dry beads for 5 mins.

5. For each reaction, prepare a hybridization buffer mix:
   7.5 µl Hybridization Buffer + 3 µl Hybridization Component A = 10.5 µl
6. Add hybridization buffer mix to beads (per reaction) and vortex and incubate up to 2 mins and place on rack.
7. Pipette complete 10.5 µl solution from tube into fresh 200 µl tube containing 4.5 µl of SeqCap EZ Prime Choice XL probes and mix thoroughly.
8. Incubate on thermocycler (program: Hybridization)

	Temperature	Time
1	95°C	5 mins
2	47°C	20 hours

Wash and recover of captured library protocol:

- Follow SeqCap EZ HyperCap workflow (User's guide v2.3 pages 29 – 34) according to manufacturer's instructions completely for capturing target library from hybridized beads.
- Dilute all buffers according to the protocol.
- Prior to starting, allow capture beads and AMPure XP clean-up beads to equilibrate to room temperature for at least 30 mins.
- Proceed directly to post capture LM_PCR reaction. All reaction parameters and primers are same as previous LM_PCR.
- Use sample and beads as starting material for LM_PCR (~ 20 µl).
- Perform 1.8X AMPure XP clean-up of LM_PCR product. Elute in 50 µl ultra-pure water. Repeat AMPure XP clean-up step and elute in 40 µl. This significantly improves purity ratio of final product.
- Product now ready for library cloning.

(1E) Vector preparation: Vector culturing

Use human STARR-seq vector ADDgene #99296 as reported by Muerdter and colleagues[7]. The hSTARR-seq_ORI vector was a gift from Alexander Stark (Addgene plasmid # 99296 ; http://n2t.net/addgene:99296 ; RRID:Addgene_99296)

1. Use all standard bacteria culture protocols for preparation of LB broth, LB-agar plates and antibiotic stocks. Prepare fresh broth and plates for use.
2. Prepare LB plates along with both ampicillin (10mg/ml stock) and chloramphenicol (25 mg/ml in ethanol stock) according to standard lab protocol. (final ampicillin concentration is 100 µg/ml and chloramphenicol is 25 µg/ml in plate) (Vector carries both resistance markers)
3. Streak LB plate using a sterile pipette tip with inoculum from vector stab and incubated overnight at 37°C to grow isolated colonies.
4. Following day, pick multiple colonies using sterile pipette tips and inoculate 4 ml LB broth with ampicillin and chloramphenicol in 14 ml bacteria culture tube (VWR catalog #60819-524) (1 culture tube per colony) and shake at 300 rpm at 37°C overnight in shaker-incubator.
5. Use 1 ml of culture broth to make glycerol stock for each colony (Use standard procedure for glycerol stock preparation). Spin down rest of the cultures and extract vector DNA using ZymoPURE plasmid Miniprep kit (Zymo catalog #D4209).
6. Measure purity and concentration and send 2 samples for sanger sequencing at the genomics core facility using multiple primers designed for ORI vector sequence verification to verify sequence. (Sequences available on request)
7. Following sequence verification regrow sample with highest concentration and purity from the glycerol stock. In the morning, pick glycerol stock with a sterile pipette and inoculate 2 ml of LB broth with ampicillin and chloramphenicol and incubate at 300rpm at 37C for 8 hours as a starter culture. In the evening, use 1 ml of starter culture to incubate 150 ml of LB broth with ampicillin and chloramphenicol and incubate overnight at 300 rpm at 37°C.
8. Make fresh glycerol stocks using 1ml of culture. Spin down rest of the culture and extract vector DNA using ZymoPURE II plasmid Midiprep kit (Zymo catalog #D4200) and verify concentration and purity.

(1F) Vector preparation: Vector linearization

- Reagents required include enzymes SalI-HF (NEB catalog #R3138S) and AgeI-HF (NEB catalog #R3552S) along with supplied Cutsmart buffer and 6X purple loading dye.
- Perform gel extraction using Zymoclean Gel DNA Recovery Kit (Zymo catalog #D4001). Further PCR purification or sample concentration can be carried out using DNA clean and concentrator kit (Zymo #D4003).
- Bring AMPure XP beads to room temperature for at least 30 mins.
- Prepare 1% agarose gel.

Vector linearization protocol:

1. Setup following reaction for each digest. Set-up 8 such reactions in 0.2ml PCR strip tubes and incubate in thermocycler at 37°C for 2 hours. Approximately 1/3rd of the product will be recovered from the gel so scale reactions accordingly for larger libraries.

	Digest mix	Volume
1	Vector	~ 1µg
2	SalI	1µl
3	AgeI	1µl
4	Cutsmart	5µl
5	Water	Up to 50µl
	Total	50µl

2. Run on 1% agarose gel for 1 hour at 100V till bands separate and cut out heavier band (~2kb) and weigh each slice. Add 3 volumes of ADB buffer and melt gel at 55°C for 30 min. Column purify each gel slice separately and pool 4 samples together and measure concentration and purity. (Purity may be low).

3. Perform secondary clean-up with 1X AMPure XP beads for each pooled sample and confirm concentration. (Should be above 35ng/µl)

(1G) Library amplification: Library cloning

- For cloning, use the NEBuilder HiFi DNA Assembly Master Mix (NEB catalog #E2621L) with the 2-3 fragment assembly protocol for efficient cloning. Alternatives include In-fusion HD cloning and Gibson Assembly.
- Try multiple ligation ratios to optimize library cloning. Repeat cloning using multiple replicates to maximize library complexity.
- Use DNA clean and concentrator kit or AMPure XP beads for reaction clean-up prior to transformation.

Library cloning protocol:

1. Calculation of [insert: vector ratio] for ligation

Considerations:
a) Vector and insert mass and associated molarity
b) Maximum molar capacity per reaction (30 fmol – 200 fmol per reaction for ligating 2-3 fragments)
c) Final volume of reaction
d) Number technical replicates to be performed

For calculations:

- Insert length: Use average insert fragment length as observed on bioanalyzer in step C.
- Vector length: 2543 bp

Ratio calculations can be done using the NEBioCalculator (https://nebiocalculator.neb.com/#!/ligation) to calculate amount of DNA for insert and vector required for final reaction assembly. Example parameter calculation for cloning:

Starting material:

DNA	conc ng/ul	260/280	260/230
Insert_1	29.8	1.81	2.32
Insert_2	25.7	1.81	2.22
Vector	38.2	1.87	2.33

Cloning parameters:

Ratio	Vector (ng)	Vector (fmol)	Insert (ng)	Insert (fmol)	total fmol
1:2	100	63.64	51.91	127.3	190.94
1:2	50	31.82	25.95	63.62	95.44
1:2	70	44.54	36.34	89.1	133.64
1:7	39.29	25	71.38	175	200

2. Set up cloning reactions according to the following volumes. Perform at least 4 reactions for each condition.

	Sample	Amount (volume)
1	Insert	X (depends on ratio)
2	Vector	Y (depends on ratio)
3	NEBuilder HiFi Assembly mix	Up to 10ul
4	Water	10
5	total	20 ul

3. Incubate reactions at 50°C for 1 hour. Pool 4 reactions of same conditions and purify using 1X AMPure XP beads and measure concentration and purity.

Ligation reactions	conc ng/ul	260/280	260/230	Ratio	# Reactions
HiFi Assembly_1	24.1	1.92	1.5	1:7	4
HiFi Assembly_2	36	2	1.87	1:2	4
HiFi Assembly_3	33.4	1.87	1.48	1:2	4
HiFi Assembly_4	31.8	1.88	2.22	1:2	4
HiFi Assembly_5	42.5	1.91	2.21	1:2	4
HiFi Assembly_6	41.6	1.87	2.21	1:2	4
HiFi Assembly_7	42	1.89	2.27	1:2	4

(1H) Library amplification: Library transformation

- Use either NEB5 alpha (NEB #C2989, currently discontinued) and NEB10 beta (NEB #C3020K) electrocompetent cells. Following protocol has been tested with only NEB electrocompetent cells.
- Perform test transformation for each cloning ratio in 25 µl competent cells. Vary between 2 – 5 µl of cloned product per transformation reaction.
- For control, use supercoiled plasmid of choice, either plasmid provided with NEB electrocompetent cells (pUC19 Vector at 50 pg/µl or plasmid PX459[9] (ADDgene catalog #62988). Vector pSpCas9(BB)-2A-Puro (PX459) V2.0 was a gift from Feng Zhang (Addgene plasmid # 62988 ; http://n2t.net/addgene:62988 ; RRID:Addgene_62988) For PX459, dilute stock to 1:10 and use 2 µl per transformation.
- Prior to starting: Electroporation is a very time sensitive step and must be carried as quickly and efficiently as possible. Make sure everything is kept ready, labeled and within reach. After cells have thawed out, work as quickly as possible to proceed. If there are a large number of reactions, perform and thaw in batches of 4 or 8 reactions.
- Prepare adequate number of sterile LB-agar ampicillin plates according standard laboratory protocol for estimating cloning efficiency and CFU/µg. Prepare >20 plates from 500 ml LB-agar. Pre-warm to 37°C during transformation.
- Prepare adequate amount of sterile LB broth with ampicillin for inoculating and amplifying transformed library. Prepare 500 ml to 1 L LB broth depending on number of transformations performed. Use >600 ml for 100 µl (4 reactions) to 200 (8 reactions) µl of competent cells.
- Pre-chill electroporation cuvettes 200 µl and 20 µl tips **overnight** (or at least 1 hour) at 4°C.
- Tightly pack ice into an ice bucket and spray with ethanol to make a 'chilled ice sludge'. Move cuvettes into ice bucket.
- Pre-warm SOC media provided with electrocompetent cells to 37°C for 1 hour and place back on bench.
- Take out required number of sterile 1.5ml Eppendorf tubes (1 for each transformation reaction) and place on ice to chill.
- Thaw out electrocompetent cells on the ice. Gently flick tube to check thawing.
- Label each tube with name of transformation reaction.
- Label 14 ml round bottomed bacterial culture tubes with similar labels as tubes.
- Make sure to work inside a sterile environment (under a lamp or inside a biosafety cabinet). Wipe everything down with ethanol prior to starting transformation.
- Ensure electroporation machine is within reach and set to correct parameters depending on competent cells. Use 1700V for NEB5 alpha or NEB10 beta cells.

Transformation procedure:

1. After cells have completely thawed out, pipette out 25 µl of cells onto the pre-chilled and labeled Eppendorf tube (DO NOT pipette more than once or necessary). Use pre-chilled tips while pipetting.

2. Add appropriate amount of cloned product (2 – 5 µl) to cell. Gently mix by flicking tube from side by holding the tube around the cap a few times. Use pre-chilled tips. (High number of unique colonies observed for 5 µl of product in 25 µl of cells)
3. Transfer mix of cells and cloned product into electroporation cuvette without any bubbles. Wipe cuvette from the side and electroporate cells with 2 pulses. Record the time constants for each transformation. (Typical range is 4.4 – 4.6 milliseconds).
4. Immediately add 975 ml of pre-warmed SOC media to cuvette and pipette up and down to mix.
5. Transfer the transformant mixture into pre-labeled 14 ml bacterial culture tubes and incubate in a shaker-incubator at 300 rpm at 37°C for 1 hour.
6. Repeat this for each transformation reaction one by one.
7. During 1 hour incubation, prepare LB broth and plates for overnight culture. Label plates with name of cloning ratio being tested and serial dilution number.
8. Label flasks with cloning ratio being amplified. Keep separate flasks for different ratios. If ratios show similar efficiency, products can be pooled during extraction.
9. Prepare 100% ethanol and sterile spreader for plating. Prepare 1.5 ml Eppendorf tubes for serial dilution. Add 900 µl of LB broth (no antibiotics) to each tube.
10. After 1 hour incubation, take 14 ml tubes off the shaker-incubator. Transformant sample should be cloudy.
11. Use 1 reaction per ratio to test efficiency in dilution series. Take 100 µl of transformant and add to 900 µl of LB broth in the 1.5ml Eppendorf tube for the 1st dilution (labeled as [-2]) and mix thoroughly.
12. Change tips and take 100 µl of 1st dilution and add to another 900 µl LB broth in the next tube for the 2nd dilution (labeled as [-3]. Repeat this up to 5th or 6th dilutions (labeled [-6] and [-7] respectively). Repeat this for each cloning ratio and control plasmid.
13. From each tube, take 100 µl of diluted transformant and spread evenly onto the pre-warmed plates and incubate overnight at 37°C.
14. Add rest of the transformants to the LB broth in conical flasks with ampicillin. Make sure volume of broth is ~ ¼ of maximum volume of flask. Use 500 ml flasks for 150 ml cultures. Use 1 L flasks for 250 ml cultures. Pool replicates of same cloning ratios.
15. Incubate cultures overnight on shaker-incubator at 300 rpm and 37°C.

(1I) Library amplification: Library storage and extraction

1. After incubation between 12-16 hours check for colonies on LB plates. Depending on the cloning and transformation efficiency, there should be at least >1 colony on the [-6] plate for all successfully transformed libraries. DH10B should show increased number of colonies.
2. This can be used to compare efficiency of each cloning ratio and show overall estimates of library complexity. A single colony on [-6] is equivalent to at least 1 million unique colonies. Higher number of colonies indicates greater library complexity.
3. Control plasmid should show significantly more colonies than cloned products and can be used to calculate the Colony Forming Units (CFU)/ μg of plasmid transformed using the formular provided by NEB. This helps estimate the transformation efficiency of the competent cell being used.
4. After validation of transformation, check the optical density of the cultures (~ 2.6).

Library Storage

1. Pipette 5 ml of culture into a sterile round-bottomed 15 ml tube and spin down culture for 5 mins.
2. Pipette out LB broth and resuspend pellet in 750 ml of fresh LB broth. Mix with 750 ml of sterile glycerol and transfer to 2 ml screw-capped cryotube. Label tubes and transfer to -80°C for long term storage.
3. Make up to 5 glycerol stocks for each successfully transformed ratio.

Library extraction

1. Pipette remaining cultures into screw capped 500 ml centrifuge tubes and spin down. Aspirate media and use pellet to extract DNA using ZymoPURE II Plasmid Maxiprep kit (Zymo catalog #D4203) according to manufacturer's protocol.
2. Use up to 150 ml of culture per maxiprep reaction. Use centrifuge version of the protocol.
3. Perform optional endotoxin treatment on the extracted library using column provided in kit.
4. Verify concentration and purity of final library ready for transfection or input library preparation.

(2) STARR-seq Screening

(2A) Culturing of cells

- Follow standard cell culture techniques for all experiments. Conduct all cell culture inside biosafety cabinet. Use only cell culture grade labware including sterile serological pipettes (VWR catalog #89130-896 for 5 ml, catalog #89130-898 for 10 ml, catalog #89130-900 for 25 ml and catalog #414004-265 for 2 ml aspiration pipettes) and standard cell culture plates and culture dishes.
- Use HEK293T cells or cell line of choice for STARR-seq. Do appropriate literature review to see evidence of IFN response from cell line upon transfection. HEK293T cells have not shown to be highly susceptible.
- Make appropriate media for HEK293T cells using following recipes.

	Media	percentage	Vol (500 ml)	Catalog number
1	DMEM (High glucose)	87	435	Sigma Aldrich #D6429-500ML
2	PenStrep	1	5	Sigma Aldrich #P4333
3	FBS	10	50	Sigma Aldrich #F2442
4	Non-essential amino acids	1	5	Sigma Aldrich #M7145-100ML
5	HEPES	1	5	Sigma Aldrich #H0887-100ML

- For 500 ml solutions, take out 65 ml of DMEM from the bottle and prepare media inside DMEM bottle.
- Store FBS in 40-50 ml stocks in 50 ml tubes (VWR catalog #89039-656) and store at -20°C. Thaw out FBS overnight at 4°C.
- Store PenStrep at -20°C in 5-10 ml stocks in 15 ml tubes (VWR catalog #89039-666) and thaw out at room temperature.
- Scale recipe as needed.
- Thaw cells in to 6-welled plates (VWR catalog #10861-696) or a 6 cm dish (VWR catalog #25382-100) using standard lab protocol.
- Passage cells every 2-3 days or when ~80% confluent and track cell morphology. Pass cell through at least 3 passages prior to transfection.
- For transfection of STARR-seq library, use at least 30 to 50 million cells per replicate (1 x 15 cm dish (VWR catalog #430599). For efficiency, passage cells from 2 or 3, 80% confluent 10 cm dishes (VWR catalog #25382-166) per 15 cm dish.

(2B) Transfection

Use Lipofectamine 3000 reagent (Thermo Fisher catalog #L3000008) for HEK293T cells with OptiMEM (Thermo Fisher catalog #31985070). Transfection optimized and shown to be >70% efficient and up to 90%.

1. Ensure ~70-80% confluency and healthy morphology of cells prior to starting transfection.
2. Pre-warm optiMEM media to 37°C in water bath. Label 1.5 ml sterile Eppendorf tubes to be used for transfection. Keep everything inside biosafety cabinet.
3. Aspirate DMEM culture media from the dishes to be transfected, wash cells with 10 ml sterile 1X PBS (Sigma Aldrich catalog #806552-500ML), aspirate and add 30 ml of OptiMEM media to cells, 30 mins prior to transfection and place cells back into the incubator. (While aspirating, be careful not to detach cells from the surface. Add media dropwise or from the side of dish to prevent detachment.)
4. Scale lipofectamine 3000 protocol according to number of cells or size of culture dish/plate. For 15 cm dish, use following reagent volumes. Remaining protocol steps are unchanged.
5. Perform up to 3 transfections per cell line. To avoid batch effects, transfect replicates on separate days.

	Tube A	Vol
1	Lipofectamine 3000	57 µl
2	OptiMEM	1443 µl
3	total	1.5 ml

	Tube B	Vol
1	STARR-seq plasmid library	38 µg
2	P3000 enhancer reagent	76 µl
3	OptiMEM	1375 µl
4	Total	1.5 ml

6. Prepare tubes A and B and add mix from tube B into tube A dropwise, gently mix by tapping (DO NOT vortex) and incubate for 15 mins.
7. After incubation, gently pipette combined mix onto the cells and incubate for 24 hours.

(2C) Total RNA Isolation

- Any RNA isolation kit may be used and should be scaled according to number of cells and culture volume. Use Trizol plus RNA purification kit (Thermo Fisher Catalog #12183555) for HEK293T cells. This kit combines trizol and chloroform extraction with PureLink RNA Mini Kit.
- RNA isolation should be carried out inside fume hood. Wipe everything down with RNaseZap RNase Decontamination Solution (Thermo Fisher Catalog # AM9782) and use separate tips and pipettes for RNA. Eppendorf tubes should be clean, RNAse free and labeled for each replicate.
- Move tabletop centrifuge to cold room if temperature-controlled centrifuge not available at least 1 hour prior to starting and wipe down with RNAseZap.
- Harvest cells in trizol in the biosafety cabinet and then move to fume hood. Trizol and chloroform are biohazard chemicals and all associated waste should be discarded separately and disposed only in satellite waste area inside fume hood.

Total RNA Isolation Protocol:

1. For 15 cm dish, use up to 6 ml of trizol.
2. To collect cells, aspirate out all media from cells and wash the cells with 10 ml of PBS.
3. Add 6 ml trizol to cells directly and collect ~1 ml of cells in trizol per 1.5 ml Eppendorf tube (6 tubes per replicate) and incubate for 5 mins at room temperature.
4. Add 0.2 ml of chloroform to each 1 ml of cells and trizol mix vigorously by shaking the tube for 15 secs. Make sure tube cap is firmly closed and not leaking. Incubate at room temperature for 3 mins.
5. Centrifuge tubes at 12,000 g for 15 mins at 4°C. Gently place tubes back inside fume hood. Sample will form 2 phases inside tube. Lower phase will be red phase comprising of phenol-chloroform, an interphase and upper aqueous phase will comprise of RNA.
6. Carefully pipette 450 – 500 µl of aqueous phase into fresh Eppendorf tubes without disturbing other phases. Pipette 150 µl at a time for ease.
7. Add equal volume (450 – 500 µl) of 70% ethanol to each tube, vortex and spin down.
8. Add up to 700 µl of mix into spin column and centrifuge at 12,000 g for 30 secs at room temperature and discard flow-through. Repeat process until all sample has been processed using the same column for all tubes per replicate. (Use the same column for all 6 tubes per replicate).
9. After all sample has been processed per replicate add 700 µl wash buffer I and centrifuge at 12,000 g for 30 secs and discard flow-through.
10. Add 500 µl of wash buffer II and repeat centrifugation and discard flow-through. Repeat step.
11. Perform a final dry centrifuge spin at 12,000 g for 1 min to remove trace ethanol from the column.
12. Remove collection tube and place column into fresh tube for RNA collection. Add 100 µl of DEPC water to column and incubate sample for 1 min at room temperature.

13. Centrifuge tube at 12,000 g for 2 mins at room temperature and measure concentration and purity.
14. Add 100 µl DEPC water for 2nd elution and 3rd elution by repeating process and measure concentration and purity of each elution. If values are consistent, pool eluates and measure final concentration and purity. Purity ratio for 260/280 and 260/230 should be above 2.0.
15. Send samples for RIN analysis to genomics core facility for quality assessment.
16. RNA can be stored at -80°C or used for mRNA isolation.
17. Dilute total RNA to 750 ng/µl for mRNA isolation. Typical yield is ~1200 µl of dilute total RNA (~900 µg of total RNA).

(2D) mRNA Isolation

- Use dynabeads mRNA Purification Kit (Thermo Fisher catalog #61006) for mRNA isolation. Each kit contains reagents for 10 mRNA isolations.
- However, binding buffer and wash buffer B (Thermo fisher catalog #11900D) and elution buffer (Thermo Fisher catalog #A33566) will not be enough if beads are to be reused and can be ordered or prepared additionally. Binding buffer is not available separately and needs to be prepared or ordered with kit.
- Additional Lysis/Binding buffer (Thermo Fisher catalog #A33562) also needs to be ordered for reusing beads). Instructions are provided on manufacturer website as well as in STARR-seq protocol by Neumayr and colleagues (Neumayr et al 2019).
- Dilute total RNA sample to 750 ng/µl prior to starting mRNA isolation. (Beads can process up to 75 µg of RNA at a time and can be regenerated for reuse. Process 100 µl of diluted total RNA at a time.
- Use 200 µl of dynabeads per 100 µl of total RNA. Regenerate beads up to 6 times. Therefore, use ~ 600 µl of total RNA for 1 x 200 µl dynabead reaction. Reuse beads within same replicate to avoid cross contamination.
- Work inside RNA hood and wipe everything with RNAseZap.
- Set heat block A to 65°C and heat block B to 80°C.
- Set tabletop shaker incubator to 25°C (room temperature) and 60 rpm.
- Thaw out total RNA on ice before proceeding.

mRNA Isolation Protocol:

1. Separate the diluted total RNA into 100 µl batches in individual 1.5 ml Eppendorf tubes and keep on ice.
2. Transfer tube containing the sample being processed to 65°C for 2 mins (to remove secondary structures) and then immediately replace on ice.
3. To prepare beads, pipette 200 µl of the beads to a RNAse free 1.5 ml Eppendorf tube and place on magnetic rack for 30 secs. Carefully discard the supernatant without disturbing the beads adhering to the magnet.
4. Remove tubes from the rack and add 100 µl of binding buffer to the tube, lightly mix and place tubes back on rack. Remove supernatant. Add another 100 µl of binding buffer (or same volume of total RNA being processed) and mix. Beads are not ready for mRNA isolation.
5. Add 100 µl of total RNA to 100 µl of beads and mix by pipetting. Incubate on shaker incubator at room temperature and low rpm (~60 rpm) for 5 mins.
6. Transfer tubes to magnetic rack and remove supernatant. Add 200 µl of wash buffer B, mix and place tubes on magnetic rack. Remove supernatant and repeat wash step.
7. Remove the tubes from the rack and add 6-10 µl of elution buffer per 100 µl total RNA and incubate samples at 80°C for 2 mins and immediately place tubes back on rack.
8. Pipette supernatant (mRNA) into a fresh tube and measure concentration and purity.

9. To reuse the beads, remove the beads from the rack and add 300 µl of Lysis/Binding buffer to wash beads. Place back on the rack, remove the supernatant and add 100 µl binding buffer to proceed with the next round of mRNA isolation for the same replicate. (Note, this is only for reuse of beads and not regeneration which has a separate protocol provided on manufacturer's website).
10. Repeat until all total RNA has been processed to mRNA. Pool multiple batches for each replicate and measure final concentration and purity. Typical yield is between 45 – 70 µl of mRNA at ~ 400 ng/µl or 18 – 28 µg of mRNA and depends on starting amount of total RNA (around 1 – 5% of total RNA).
11. Store mRNA at -80°C or proceed with TURBO DNase treatment, RNA clean-up and reverse transcription.

(2E) TURBO DNase treatment

- This is required to remove any residual DNA from sample prior to reverse transcription. Use TURBO DNA-free Kit (Thermo Fisher catalog #AM1907).
- Starting material is around ~400 ng/µl or ~400 µg/ml and thus requires the rigorous protocol for TURBO DNase treatment and thus use 2 µl or 4 units of TURBO DNase enzyme per reaction.
- Process entire sample together per replicate. Scale protocol according to volume of mRNA.
- Set heat block to 37°C prior to starting.

TURBO DNase treatment protocol:

1. Add 0.1 volume of 10X TURBO DNase buffer to sample.
2. Add 2 µl (4 units) of TURBO DNase enzyme and mix gently (DO NOT vortex).
3. Incubate at 37°C for 60 mins.
4. Add 0.2 volumes of DNase inactivation reagent and mix gently (DO NOT vortex) and incubate at room temperature for 5 mins. Lightly flick tube to redistribute reagent every 1 min.
5. Centrifuge tubes at 10,000 g for 1.5 mins at room temperature and transfer supernatant to fresh 1.5 ml Eppendorf tube for the next step.

(2F) RNAClean XP treatment

- Perform clean-up of each TURBO DNase reaction using RNAClean XP beads (Beckman Coulter catalog #A63987)
- Warm beads to room temperature for at least 30 mins.
- Prepare fresh 80% ethanol for sample wash.
- Set tabletop shaker incubator to 37°C and 60 rpm (low rpm).
- Process 1 replicate per clean-up reaction. Scale volume of beads according to volume of mRNA.

RNAClean XP protocol:

1. Add 1.8 volume of bead per 1 volume of mRNA sample and mix by thoroughly pipetting >20 times and incubate at room temperature for 15 mins.
2. Transfer mix to magnetic rack and incubate for 10 mins or till when all beads have adhered to magnet.
3. Remove supernatant (should be clear liquid) and add 500 µl of 80% ethanol for 1st wash. Incubate for up to 2 mins, remove supernatant and add 500 µl of 80% ethanol for 2nd wash. Incubate for 2 mins and remove without disturbing beads.
4. Dry beads completely at room temperature for 5 mins. Check for any residual ethanol inside tube and remove.
5. Remove tubes from the rack, and add 20 µl DEPC water to beads for elution, mix by pipetting and vortexing and incubate in tabletop shaker incubator at 37°C at low rpm for 3 mins and place tubes back on magnetic rack.
6. Incubate for 1 min on the rack and pipette pure mRNA into a fresh 1.5 ml Eppendorf tube for further processing.
7. Assess final concentration and purity of mRNA. This is the final checkpoint for samples prior to cDNA library preparation.
8. Typical yield is ~12 µg of mRNA per replicate (600 – 900 ng/µl of mRNA in 19 µl of water (1 µl used for purity and concentration assessment)

(2G) cDNA library preparation: 1st Strand Synthesis

- For 1st strand synthesis, use Superscript III kit
- Divide total sample amount by 5 and round off to nearest multiple of 5 to determine number of reactions required per sample. Use 5 reactions using 2.4 µg of mRNA per reaction for ~12 µg samples.
- Adjust all samples to 12 µg in 20 µl (600 ng/µl) of DEPC water for uniformity across replicates. Use 4 µl (2.4 µg mRNA) per RT reaction.
- Though protocol suggests using 500 ng of mRNA at a time for 1st strand synthesis, we are performing reverse transcription only for fraction of mRNA that was self-transcribed due to enhancer activity and so taking larger amounts of mRNA is permissible.
- Dilute STARR-seq reporter specific primer to 2 µM. Dilute dNTP solution to 10 µM.

1st Strand Synthesis Protocol:

1. Assemble following reaction as reaction 1:

	Reagent/sample	Vol (µl)
1	mRNA	4
2	dNTP	1
3	Primer	1
4	Water	7
5	Total	13

2. Incubate at 65°C for 5 mins followed by 4°C for 1 min on a thermocycler (program RT_1)
3. Assemble following reaction as reaction II:

	Reagent	Vol (µl)
1	1st strand buffer	4
2	DTT	1
3	RNaseOUT	1
4	Superscript III	1
5	total	7

4. Add reaction II to reaction I for each RT reaction (final volume 20 µl)
5. Incubate according to following conditions on thermocycler (program RT_II):

1	50°C	60 mins
2	75°C	15 mins
3	4°C	hold

(2H) cDNA clean-up: RNase A treatment and AMPure XP clean-up

- Treat all reverse transcribed samples with RNase A to remove residual RNA.
- Pool 5 RT reactions (total = 100 µl per replicate) (5 reactions carried out per replicate)
- Move to DNA bench to avoid pipetting RNase A in RNA isolation bench
- Prepare 80% ethanol for AMPure XP bead clean-up wash steps
- Bring beads to room temperature for at least 30 mins prior to starting.

RNase A treatment and AMPure XP bead clean-up Protocol:

1. Add 1 µl of RNase A (10mg/ml) per 5 RT reactions.
2. Incubate at 37°C for 1 hour in thermocycler (program: RNase A treatment)
3. Perform **1.4X AMPure XP bead clean-up** protocol
4. Elute cDNA in 43 µl DEPC water for next step.
5. Measuring cDNA concentration is advised but is not accurate and hence it is also okay to proceed to either 2nd strand synthesis and UMI_PCR or directly to jPCR steps.

(2I) UMI addition: 2nd strand synthesis

- Unique Molecular Identifiers (UMIs) are essential to label and filter out PCR duplicates from the libraries that occur during library preparation.
- For addition of UMIs, follow UMI-STARR-Seq protocol reported by Neumayr and colleagues[8] including a 2nd strand synthesis reaction and UMI_PCR reaction prior to jPCR.
- Start 2nd strand synthesis using 2nd strand primer as shown by Neumayr and colleagues[8]. Dilute primer to 10 µM in DEPC water for use.
- Use KAPA 2X HiFi Hot Start Ready mix polymerase (Roche catalog #07958927001)

2nd strand synthesis protocol:

1. Process entire replicate cDNA (~43 µl) per reaction. Set-up following reaction for each replicate:

	Sample/reagent	Vol (µl)
1	cDNA	42.5
2	KAPA HiFi 2X Ready mix	50
3	2nd strand primer	7.5
4	Total	100

2. Incubate reactions in thermocycler (program: 2nd strand synth)

	Temperature	Time	Step
1	98°C	60 secs	Denaturation
2	65°C	30 secs	Annealing
3	72°C	30 secs	Extension

3. Reaction purification using 1.4X AMPure XP bead clean-up.
4. Elute library in 43 µl of DEPC water for UMI_PCR step.

(2J) UMI addition: UMI_PCR

- To add UMI, use custom designed i7-UMI-P7 primer to add i7 index sequence and UMI sequence simultaneously to cDNA. This helps retain unique dual indexing to mitigate index hopping as well as allow for PCR duplicate filtration as opposed to Neumayr and colleagues protocol where the i7 is replaced by the UMI.
- Dilute primers to 10 µM for use. Keep record of which index is being added to which sample and label all tubes.
- Process complete cDNA for each reaction.

UMI_PCR protocol:

1. Set-up following reaction for each replicate:

	Sample/reagent	Vol (µl)
1	cDNA	42.5
2	KAPA HiFi 2X Ready mix	50
3	i7-UMI-P7 primer	7.5
4	Total	100

2. Incubate reaction in thermocycler (program: 2nd strand synth) Same conditions as 2nd strand synthesis.

	Temperature	Time	Step
1	98°C	60 secs	Denaturation
2	65°C	30 secs	Annealing
3	72°C	30 secs	Extension

3. Clean-up reactions using 1.4X AMPure XP bead clean-up and elute in 50 µl of DEPC water per replicate. These samples are now ready for jPCR step.

(2K) Junction PCR (jPCR)

- Perform same number jPCR reactions as RT reactions. Use 10 µl of cDNA library per jPCR reaction.
- Use modified jPCR primers as provided in Neumayr et al 2019 protocol.
- jPCR allows amplification of only self-transcribed fragments and filters out plasmid transcripts

jPCR protocol:

1. Set-up following reactions for each replicate. (5 reactions per replicate)

	Sample/reagent	Vol (µl)
1	cDNA library	10
2	DEPC water	10
3	jPCR forward primer	2.5
4	jPCR reverse primer	2.5
5	KAPA 2X HiFi Ready Mix	25
	Total	50

2. Incubate reactions in thermocycler (program: jPCR)

	Temperature	Time	Step
1	+98°C	45 secs	Initial denaturation
2	+98°C	15 secs	Denaturation
3	+65°C	30 secs	Annealing
4	+72°C	30 secs	Extension
5	+72°C	60 secs	Final extension
6	+4°C	Hold	

X 16

3. Clean-up jPCR reactions by pooling all 5 reactions per replicate (250 µl per replicate) and perform a 0.8X AMPure XP clean-up using 200 µl of beads per replicate.
4. Elute in 50 µl DEPC water. Measure concentration and purity of final jPCR products prior to performing final sequencing ready PCR step.
5. Typical yield is around 50 – 130 ng/µl of 50 µl product. Yields may vary based on STARR-seq activity and starting amount for samples.

(2L) Sequencing ready PCR: Output library preparation

- This step is carried out to add i5 and P5 index adapters to the library before sending samples for sequencing.
- Use standard i5 primer as forward primer and P7 primer as shown in Neumayr et al 2019 and supplementary table S2.
- Dilute all primers to 10 µM prior to use.
- Bring AMPure XP beads to room temperature for at least 30 mins for PCR clean-up
- Prepare 1% agarose gel for test PCR run to determine number of cycles for PCR. Start with low cycle number. 5 cycles are sufficient for jPCR products obtained in the previous step.
- If performing test PCR, test on all samples since they may have different starting amounts.

Sequencing Ready PCR protocol:

1. Setup the following reaction for each replicate:

	Sample/reagent	Vol (µl)
1	jPCR product	10
2	DEPC water	10
3	i5 primer	2.5
4	P7 primer	2.5
5	KAPA 2X HiFi Ready Mix	25
	Total	50

2. Incubate reactions on thermocycler (program: SeqReady PCR)

	Temperature	Time	Step
1	+98°C	45 secs	Initial denaturation
2	+98°C	15 secs	Denaturation
3	+65°C	30 secs	Annealing
4	+72°C	30 secs	Extension
5	+72°C	60 secs	Final extension
6	+4°C	Hold	

X 5

3. Run 10 µl sample with 2 µl of 6X loading dye (for each replicate) on 1% agarose gel for 30 mins at 130V and observe band. Check for characteristic smear for each band at right size as shown by Neumayr and colleagues[8]. If band is concentrated, repeat PCR with lower cycles (if cycle >5) or try lower starting material (5 µl of jPCR).
4. Repeat PCR for correct starting amount and number of cycles as determined.
5. Clean up each reaction with 1X AMPure XP beads and elute in 20 µl. Typical yield can vary and may need to be diluted prior to sequencing.
6. Samples once sent to core facility will be reassessed for quality, checked for distribution on tapestation, further diluted and pooled into an equimolar pool for sequencing.

(2M) Sequencing ready PCR: Input library preparation

- To assess enhancer activity, plasmid library needs to be sequenced alongside output libraries for normalization.
- Library sequence architecture should be identical to enable pooling of library along with output libraries.
- Input library also goes through PCR steps and so also important to add UMIs to these libraries prior to sequencing similar to output libraries.
- Use custom i7-UMI-P7 primers designed for UMI addition. Use i5 primer to add i5 index and P5 adapter. Assign unique dual index pairs to each replicate of input library and use at least 3 replicates. Record indexes assigned. Dilute primers to 10 µM prior to use.
- Bring AMPureXP clean-up beads to room temperature for at least 30 mins for clean-up.
- Prepare 1% agarose gel for final library fragment length verification and purification.
- For starting material, use 1 µl of STARR-seq plasmid library (~750 ng/ µl) added with 39 µl of DEPC water.

UMI_PCR protocol for input library preparation:

1. Setup following reaction for each input replicate:

	Sample/reagent	Vol (µl)
1	STARR-seq plasmid library	40
2	KAPA HiFi 2X Ready mix	50
3	i7-UMI-P7 primer	5
4	i5 primer	5
5	Total	100

2. Incubate reaction in thermocycler (program: UMI_PCR)
3. Perform 1.4X AMPure XP clean-up of each replicate reaction. Elute in 20 µl DEPC water.

Sequencing Ready PCR protocol for input library preparation:

1. Setup following reaction for each input replicate: (use complete product)

	Sample/reagent	Vol (µl)
1	Input replicate DNA	20
2	i5 primer	2.5
3	P7 primer	2.5
4	KAPA 2X HiFi Ready Mix	25
	Total	50

2. Incubate reactions in thermocycler (program: SeqReady PCR)
3. Run entire sample (50 µl) with 10 µl 6X loading dye on a 1% agarose gel for 30 mins at 130V.
4. Visualize gel and verify fragment length. Slice out band of correct length and weigh gel slice.
5. Extract DNA form gel slice using Zymoclean Gel DNA recovery kit.
6. Clean-up extracted DNA using 1X AMPureXP bead clean-up and elute in 25 µl of DEPC water and measure concentration and purity. Typical yield is ~ 10 – 20 ng/µl.

- All input and output libraries are sent to sequencing core facility for sequencing.

References

1. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **00**, 1–3 (2013).

2. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

3. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, (2008).

4. Lee, D. *et al.* STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol.* **21**, 1–24 (2020).

5. Kim, Y. S. *et al.* Correcting signal biases and detecting regulatory elements in STARR-seq data. *Genome Res.* **31**, 877–889 (2021).

6. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* (2010) doi:10.1093/bioinformatics/btq033.

7. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).

8. Neumayr, C., Pagani, M., Stark, A. & Arnold, C. D. STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries. *Curr. Protoc. Mol. Biol.* **128**, e105 (2019).

9. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).