Uniform processing and analysis of IGVF massively parallel reporter assay data with MPRAsnakeflow

Jonathan D. Rosen, Arjun Devadas Vasanthakumari, Kilian Salomon, et al.

Genome Res. 2026 36: 1940-1951 originally published online August 11, 2026
Access the most recent version at doi:10.1101/gr.281462.125

Resource

Uniform processing and analysis of IGVF massively parallel reporter assay data with MPRAsnakeflow

Jonathan D. Rosen,[1] Arjun Devadas Vasanthakumari,[2] Kilian Salomon,[3] Nikola de Lange,[2] Pyaree Mohan Dash,[3] Pia Keukeleire,[2] Ali Hassan,[2] Alejandro Barrera,[4,5] Beniamin Krupkin,[6,7] Grace Oualline,[8] Martin Kircher,[2,3] Michael I. Love,[1,9] and Max Schubach[3]

[1]*Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA;* [2]*Institute of Human Genetics, University Hospital Schleswig-Holstein, University of Luebeck, 23562 Lübeck, Germany;* [3]*Berlin Institute of Health at Charité–Universitätsmedizin Berlin, 10117 Berlin, Germany;* [4]*Department of Biostatistics and Bioinformatics, Duke University Medical School, Durham, North Carolina 27710, USA;* [5]*Center for Advanced Genomic Technologies, Duke University, Durham, North Carolina 27708, USA;* [6]*Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94143, USA;* [7]*Institute for Human Genetics, University of California San Francisco, San Francisco, California 94143-0794, USA;* [8]*Department of Genetics, Yale School of Medicine, New Haven, Connecticut 06510, USA;* [9]*Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599-7420, USA*

As researchers and clinicians seek to identify human genomic alterations relevant to traits and disorders, identifying and aggregating evidence providing mechanistic support for associations between alterations and phenotypes remains challenging. In particular, the study of noncoding genomic variation remains a major challenge because of the lack of accurate functional annotation for activity in a given context and across alleles. Experimental evidence is critical for prioritizing and interpreting functional effects of genetic alterations. Massively parallel reporter assays (MPRAs) have emerged as a powerful high-throughput approach, enabling quantification of regulatory element activity and allelic effects, as well as systematic dissection of gene regulatory logic and variant effects across different contexts. However, the diversity of MPRA designs, lack of standardized formats, and many potential processing parameters hamper data integration, reproducibility, and meta-analyses across studies. To address these challenges, the Impact of Genomic Variation on Function (IGVF) Consortium established an MPRA focus group to develop community standards, including harmonized file formats, and robust analysis pipelines for a wide range of library types and experimental designs. Here, we present these formats and comprehensive computational tools, MPRAlib and MPRAsnakeflow, for uniform processing from raw sequencing reads to counts, processing, and visualization. Using diverse MPRA data sets, we investigated technical variability sources including barcode sequence bias, outlier barcodes, and delivery method (episomal vs. lentiviral). Our results establish best practices for MPRA data generation and analysis, facilitating robust, reproducible research and large-scale integration. The presented tools and standards are publicly available, providing a foundation for future collaborative efforts in regulatory genomics.

[Supplemental material is available for this article.]

Control of gene expression is a complex, dynamic process in which many regulatory elements determine when, where, and to what extent genes are active. Shifts in these regulatory programs drive development and shape evolutionary diversity, whereas disruptions can lead to disorders such as cancer, congenital malformations, and bleeding or metabolic disorders (Reijnen et al. 1992, 1993; Ludlow et al. 1996; De Castro-Orós et al. 2011; VanderMeer and Ahituv 2011; Huang et al. 2013). *Cis*-regulatory elements (CREs) such as enhancers and promoters modulate transcription through their recruitment of transcription factors (TFs) and the surrounding chromatin landscape. CREs can be difficult to recognize and interpret because they may lack clear sequence (e.g., TF motifs) or epigenetic hallmarks (e.g., open chromatin state or adjacent histone modifications), can influence genes over long genomic distances, and can have cell type–specific activity. Consequently, pinpointing functional variants that drive phenotypic differences within and between species remains a major challenge.

The National Human Genome Research Institute (NHGRI) launched the Impact of Genomic Variation on Function (IGVF) Consortium in 2021, aiming to systematically understand how genomic variation affects genome function and how these effects shape phenotypes (Engreitz et al. 2024). Massively parallel reporter assays (MPRAs) are a specific type of multiplex assays of variant effects (MAVEs) that can probe noncoding sequence effects and are a central approach within the IGVF efforts. MPRAs are high-throughput experiments that enable the simultaneous assessment of thousands to millions of genetic elements or variants in a single

Corresponding author: max.schubach@bih-charite.de
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.281462.125. Freely available online through the *Genome Research* Open Access option.

36:1940–1951 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/26; www.genome.org

experiment. By inserting elements or variants into reporter constructs, one can measure their effects on diverse cellular processes like transcriptional initiation, splicing, transcript stability, or protein function. To measure the effects of a candidate CRE (cCRE) on transcriptional activation, the test element is placed upstream of a minimal promoter and barcodes are uniquely mapped to that test element, allowing regulatory activity to be quantified by measuring the RNA levels of the transcribed barcode compared with the abundance of the barcode in DNA. Reporter plasmids can be tested episomally (episomalMPRA) (Melnikov et al. 2012) or integrated into the genome using lentiviral delivery (lentiMPRA) (Inoue et al. 2017), transposons, or recombination. Although these experiments enable testing thousands of elements in parallel, both episomal and genomic integrated MPRAs measure regulatory activity outside of the native genomic locus and thus may not fully recapitulate in vivo regulatory dynamics (Kosicki et al. 2025).

A related method, self-transcribing active regulatory region sequencing (STARR-seq), can be used simultaneously to study both cCREs and variants (Arnold et al. 2013; van Arensbergen et al. 2017). Unlike standard MPRAs, STARR-seq places cCREs downstream from a minimal promoter, allowing each element to serve as its own transcribed barcode and permitting direct quantification of enhancer activity based on the abundance of self-transcribed RNA. Although STARR-seq assays are also conducted within the IGVF consortium, differences in assay outputs and data processing necessitate different processing. In this work, the primary focus is on barcode-based MPRAs, with STARR-seq included or referenced where relevant.

In recent years, there has been a substantial increase in MPRA publications, with laboratories worldwide employing the technique to characterize cCREs or identify expression-altering variants (Kircher et al. 2019; Jiang et al. 2024; Agarwal et al. 2025). MPRAs are being used to study variants in genomic cohorts (Koesterich et al. 2023), to develop predictive tools (de Almeida et al. 2022; Gosai et al. 2024; Agarwal et al. 2025), to serve as benchmarks (Avsec et al. 2021, 2026; Jaganathan et al. 2025; Linder et al. 2025; Rafi et al. 2025), and to validate synthetic sequences designed for specific behaviors such as tissue-specific expression (de Almeida et al. 2022). These applications illustrate that MPRAs are not only interpreted in isolation but now form the basis for numerous meta-analyses. Accordingly, standards for MPRA experiments are essential to enable the reanalysis and integration of data from diverse studies as well as access for a growing community. To address this need, the IGVF consortium established an MPRA focus group in 2021 to develop community standards, file formats, and pipelines for consistent data processing and experimental coordination across the consortium (Engreitz et al. 2024).

Here, we present key products of this focus group, enabling researchers to conduct and analyze MPRA studies. We introduce standardized file formats defined for IGVF that comprehensively describe both experimental designs and results, facilitating reanalysis and downstream integration. To accommodate the variety of MPRA designs, we developed MPRAsnakeflow, a processing pipeline that supports uniform analysis of diverse MPRA assays from sequencing reads to count data, with particular optimization for variant-based designs. Complementing MPRAsnakeflow, we present MPRAlib, a Python library that facilitates downstream processing of DNA and RNA counts, filtering, outlier removal, and data visualization, providing researchers with a powerful tool for integrating MPRA data into Python notebooks and linking to other statistical tools (Ashuach et al. 2019; Myint et al. 2019; Keukeleire et al. 2025). Using multiple assays from IGVF and others, we systematically evaluate pipeline performance, assess sources of technical variation including barcode discrepancies and outliers across replicates, and compare different MPRA experimental designs. Together, these tools and analyses provide a comprehensive framework for robust and reproducible MPRA data processing and interpretation.

Results

MPRA design

Arbitrary sequences can in principle be tested via an MPRA, constrained only by technical limitations like construct or vector length. Typically, sequences are generated by DNA oligo synthesis, which currently is still most economical for sequence lengths of ~300 nucleotides (nt) and covers a majority of open chromatin peak lengths. Longer sequences can be derived either from combining shorter sequences or from selections or amplifications of genomic DNA fragments. When barcodes are not directly synthesized with each sequence but added randomly (e.g., by amplification with overhanging barcode primers), the resulting MPRA library must be sequenced to determine barcode-to-test sequence associations. Paired-end short-read sequencing is usually sufficient to cover the entire test sequence with a small overlap (we recommend >10 nt for high-complexity sequences). For example, 2 × 150 nt paired-end sequencing can cover a 200 nt test oligo + 20 nt barcode + 30 nt spacer with 50 nt overlap. Alternatively, long-read sequencing can be employed. However, we note that various MPRA designs have different requirements on the precision of these sequence read-outs to obtain an unambiguous association, specifically if sequence variants need to be distinguished.

Some MPRA experiments test synthetic or engineered sequences that possess particular activity properties, such as tissue-specific enhancer activity or optimal spacing between two elements (de Almeida et al. 2022; Georgakopoulos-Soares et al. 2023; Gosai et al. 2024). However, most MPRAs are used to test cCREs derived from a genomic or inferred (ancestral) genomic sequence, and identified through epigenetics maps, motif scans, or sequence models. For longer fragments, tiling might be necessary (Kosicki et al. 2025). These choices will impact the average element activity in the assay and thereby experimental sensitivity. Within IGVF activities, variants are one complementary focus of the consortium. Again, variants may be synthetic and prioritized using sequence models or selected from various sources like population data, ancestral alleles, or prior functional mapping from genome-wide association studies (GWAS) or expression quantitative trait loci (eQTL) studies. For variants, the reference and alternative sequence is designed. Usually one variant is centered within the oligo, but alternative positioning as well as the introduction of variant combinations is possible. We note that a sufficient reference activity is generally important to detect variant effects. Therefore, shifting the sequence toward known transcription factor binding sites (TFBSs) or using the cCRE is beneficial.

In addition to test sequences (e.g., variants or cCREs), control sequences are required to validate experiment performance. By allowing the quantification of baseline activity, they help to identify significantly activating or repressing sequences and establish that the activity distribution of tested sequences exceeds that of negative controls. The most common approach is to use (di-nucleotide) shuffled sequences as negative controls to estimate background activity. Positive element or variant controls are more challenging to

design. Ideally there exist previously tested sequences with high activity in the same cell type, or variants with known properties in the same system. If such sequences are not available, sequences with activity across many cell types or from a related cell type should be used. Generally, only a small number of positive controls (around hundred) might be required, whereas a larger set of several hundreds of negative controls is necessary for later statistical testing. However, if previously validated positive controls are unavailable, more putative positive controls are recommended.

We recommend avoiding the use of sequences with certain undesirable properties. Sequences containing long homopolymers or specific restriction sites (e.g., EcoRI or SbfI) used in MPRA library construction should be removed. Restriction sites may also be eliminated by sequence modification. Simple repeats and overlap with transposable elements may cause unwanted strong activity or construct silencing. The inclusion of sequences with binding sites for factors impacting chromatin structure (e.g., CTCF binding sites) should be excluded, because short MPRA constructs cannot model chromatin looping structures. Finally, sequences containing additional transcriptional start sites may interfere with the minimal promoter, resulting in inconsistent background activity of the core element.

To facilitate this process, we developed a workflow called MPRAOligoDesign, which can generate a design file from sequences, variants, genomic regions, and paired variants with genomic regions. It extends regions to the desired sequence length, tiles longer regions, designs reference and alternative sequences for variants, and can be configured to perform recommended property checks, removing sequences that do not pass the filters.

Standardized file formats

Within the past 10 years, more than 100 MPRA experiments have been published. Various efforts have been made to integrate different experiments into common databases, like MaveDB, MPRAVarDB, and MPRAbase (Jin et al. 2024; Rubin et al. 2025; Zhao et al. 2025). However, because there are no defined standards for metadata, such as the genomic region tested or sequence design, setting up similar data resources becomes very challenging. One major resource of IGVF will be a catalog of element and variant effects. With various different MPRAs being conducted, they must be integrated into a single catalog as well as a data portal for reuse by other projects. We see a tremendous need to harmonize metadata regarding the design and various stages of output files for MPRA data. We identified four categories for which standardized files are necessary—design, counts, effects, and genomic effects—and provide a detailed documentation of the file formats in Supplemental Note S1.

For "design," we defined a reporter sequence design file that contains the main metadata of designed oligos such as identifier, sequence, sequence origin, and class and provides key information about the purpose of the assay. Despite the "designed oligo" terminology, such design files should also be created for MPRA experiments not derived from oligo synthesis, and equivalent information should be collected. In the design file, each oligo has a unique name or ID and the nucleotide sequence. It further includes the category (variant, element, synthetic, scrambled), an assigned class (test, variant positive control, variant negative control, element active control, element inactive control), and additional information. Chromosome and start and end positions, along with the reference genome version, must be specified if applicable. For variants, the position within the tested sequence, a unique ID using the canonical SPDI notation (Holmes et al. 2020), and whether the reference or alternative allele was tested within the sequence must also be included. This file enables reanalysis of the data as well as the generation of element and/or variant effects from the data.

"Count" files report the number of observed reads (i.e., integer counts) counted per barcode or aggregated to elements per modality (i.e., DNA or RNA). Multiple counts and modalities can be described in the same file. Barcode-level files contain the barcode sequence and the associated oligo identifier. Element-aggregated files describe the replicate ID, oligo ID, aggregated and normalized DNA and RNA counts, the number of aggregated barcodes, and an inferred log-fold change (normalized RNA/DNA ratio). For IGVF, these are generated via the standardized pipeline MPRAsnakeflow (see below). However, existing published pipelines, like MPRAflow (Gordon et al. 2020), also generate similar types of data and could be converted to our standardized format.

"Effect" files report the results of differential activity analysis. This includes normalized counts, activity estimates (i.e., log-fold changes), and significance estimates (i.e., *P*- and q-values). This can be performed using tools like MPRAnalyze, BCalm, or mpralm (Ashuach et al. 2019; Myint et al. 2019; Keukeleire et al. 2025), with the count data and information from the sequence design file as input (e.g., to map oligos to reference or alternative alleles for a variant). Because differential activity analysis differs between elements and variants (e.g., statistical testing for elements always requires a control set of oligos), we separate them into element and variant effects files. In addition to the values listed above, element files contain aggregated modality counts. Variant files contain those for both the alternative and reference allele, a posterior probability of the regulatory effect, and a 95%-confidence interval of the effect.

Finally, "genomic effects" are those that can be mapped to a genomic location, and we have created standardized BED files for variants and elements. These contain information from effect files with genomic coordinates for visualization of MPRA results in genome browsers such as UCSC or Integrative Genomics Viewer (IGV) (Robinson et al. 2011; Perez et al. 2025). Tools for MPRA data from IGVF will support these standards. In our MPRAlib library (see below) a command set called validate-file is implemented to compare actual files with their schemas to ensure files are formatted correctly.

A unified MPRA pipeline: MPRAsnakeflow

MPRAsnakeflow is a further evolution of MPRAflow (Gordon et al. 2020), which established basic functionality for the analysis of MPRA data. We focus on six topics to enhance and extend its functionality: (1) speed and memory improvements to support large complex MPRA assays; (2) interoperability by containerization of software dependencies; (3) standardized outputs, quality reports and additional statistics to help uncover potential issues with the MPRA experiment; (4) support for multiple assays (incl. raw read structures); (5) optimized barcode assignment to support designs with small edit distances, like multiple variants for the same cCRE; and (6) new features like strand awareness, outlier removal, and support of multiple tools for statistical quantification. For a detailed comparison of MPRAflow with MPRAsnakeflow, see Supplemental Table S1. MPRAsnakeflow is implemented in the workflow language Snakemake (Mölder et al. 2025). It is organized into two subworkflows: the assignment workflow and the experiment workflow (see Fig. 1).

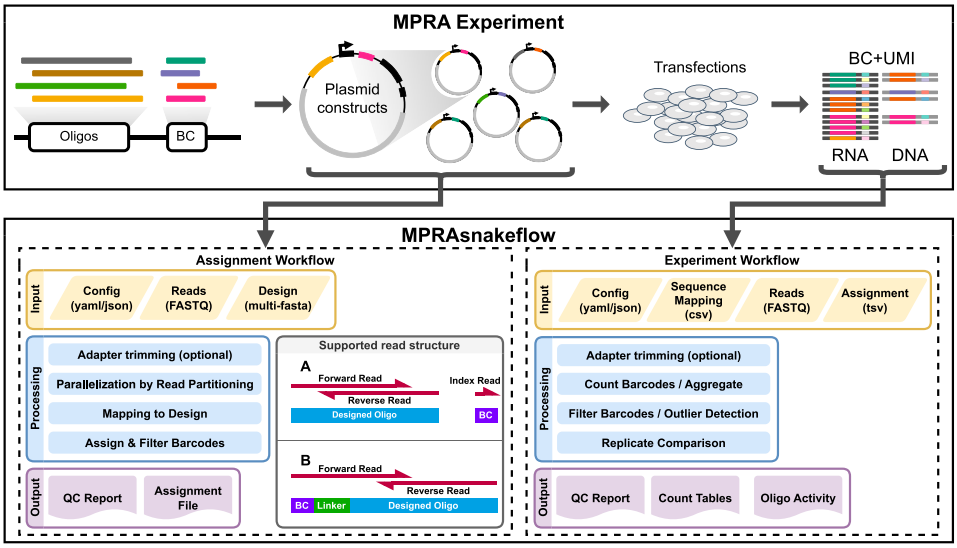

Figure 1. Overview of an MPRA experiment with UMIs (*top*) and the MPRAsnakeflow pipeline for analyzing MPRA experiments (*bottom*). MPRAsnakeflow consists of an assignment workflow (*left*), which maps barcodes to designed oligos, and an experiment workflow (*right*), which counts barcodes and computes oligo activity. Inputs are shown in yellow, main processing steps in blue, and outputs in purple. Notably, MPRAsnakeflow supports different input read structures, including separate reads for oligo and barcode (*A*), as well as combined structures in which the barcode and oligo are present within read pairs (*B*). Reads can also be single-end reads.

The assignment workflow aims to associate barcodes with designed oligos. It supports multiple read structures typical for certain assay types (Gordon et al. 2020; Gosai et al. 2024). For example, the barcode can either be sequenced in a separate index read or positioned at the beginning of the first paired-end read, separated from the start of the oligo by a linker. In brief, the assignment workflow first verifies that the design file (in multi-FASTA format) is appropriate, ensuring unique headers and no duplicate sequences. Next, paired-end reads of the sequenced oligos are merged using NGmerge (default) or fastq-join (Aronesty 2013; Gaspar 2018), with a minimum recommended overlap of 11 nt. The merged reads are then mapped to the design file using a read mapper (BBMap [default], BWA-MEM [Li and Durbin 2009; Bushnell 2014], or exact string matching). For long-read data, we implemented the Pacific Biosciences (PacBio) minimap wrapper pbmm2 (Li 2018). Barcodes are assigned based on the best unique sequence match. Finally, barcodes are aggregated and filtered based on a minimum number of required observations per oligo (default: three), and a majority of barcode occurrences must map to the same oligo (default: 0.75) to minimize ambiguity while allowing barcode collisions or best alignment errors. This workflow can be skipped if barcode associations are already known (e.g., by design or with custom assignment scripts).

The experiment workflow processes the sequenced barcodes extracted from DNA and RNA samples of the MPRA experiment in a specific tissue or cell type. It counts the observations of each barcode, merges DNA and RNA counts, and assigns barcodes to oligos using an assignment file, which can be generated by the assignment workflow. An optional step removes outlier barcodes. Barcode counts per oligo are then aggregated, and activity is calculated as the $\log_2$ ratio of normalized (counts per million reads) RNA over DNA counts for each oligo. When multiple replicates are available, the workflow computes correlations of DNA counts, RNA counts, and $\log_2$ ratios to assess data quality. Additionally, the experiment workflow supports the use of UMIs to disambiguate unique transcripts from PCR duplicates to reduce PCR artifacts.

Both subworkflows produce a variety of statistics and plots to help users inspect the data and identify potential errors. A unique feature of MPRAsnakeflow is the quality control report generated for both workflows, which summarizes key metrics and plots along with additional explanations. The quality report is provided as a single structured and image-embedded HTML file, enabling quick and efficient data inspection and making it easy to share results with collaborators. We have comprehensively documented MPRAsnakeflow; full documentation is available at https://mprasnakeflow.readthedocs.io. The documentation provides installation instructions, detailed descriptions of the available configuration options and workflow components, and a step-by-step tutorial. In addition, we present analyses of more than 10 example data sets, including recent, independently generated MPRA studies available from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) that were not part of the IGVF consortium, representing a broad range of library designs. Together, these examples illustrate the flexibility of the workflow and provide practical guidance for users seeking to adapt it to their own assays.

Mapping strategies for barcode associations

To mitigate the impact of sequencing errors, sequenced oligos are typically mapped (Alser et al. 2021) to the originally designed sequences of the MPRA experiment. Approaches relying solely on exact string matching, as implemented in some other pipelines like esMPRA (Li et al. 2025), fail to account for sequencing errors and consequently result in a reduced number of assigned reads and therefore less observed barcodes. The MPRAflow tool (Gordon et al. 2020) uses solely BWA-MEM and does not allow different mapping configurations without changing the software. Our observations indicate that the choice of mapping strategy can substantially influence downstream analyses, depending on the library design. For this reason, MPRAsnakeflow offers multiple configurable strategies: exact matching (requiring the full read sequence matching the designed oligo, in either the forward or

reverse-complement orientation), BWA-MEM, or BBMap. We observed that BWA-MEM reports MAPQ 0 values despite distinguishing oligos with minimal sequence differences in alignment (as apparent from alignment score differences reported in the output). This causes barcodes to be lost from associations in variant-rich MPRA libraries. To address this, we implemented an additional filtering step for BWA alignments, allowing users to still consider MAPQ 0 results based on the sequence similarity, number of mismatches, and alignment length.

Using different mapping strategies, we report in Figure 2 the assignable oligos, identified genomic regions without alternative alleles designed, the assigned variants (for which both reference and alternative oligo must be present), as well as the number of barcodes per oligo. In short, BBMap was able to assign more oligos on average (35,249) compared with BWA-MEM (33,736) and exact matching (33,471) across all MPRA sets used. The same holds for assigned regions and especially for variants. We see a drastic decrease in reported variants using the BWA mapping strategy especially for the 80K-neurons data set, confirming the outlined issues with MAPQ values when two oligos are in close edit distance (e.g., 1 nt for SNVs).

We further inspected the design on 80K-neurons because of the difference in variants for BWA-MEM and noticed that most reference alleles had at least three alternative alleles each (62%). Within this subset of oligos, the proportion of assignable variants dropped to <40% for BWA-MEM (see Supplemental Fig. S1). This shows that small edit distances play a crucial role in retaining variants and should be carefully considered for elements with many designed variants or for saturation mutagenesis (Kircher et al. 2019).

To compare the impact on downstream analyses, we utilized BCalm (Keukeleire et al. 2025) to determine significant variant effects for all four mapping strategies using two different MPRA experiments (8K-neurons [Koesterich et al. 2023], 80K-neurons). Supplemental Figure S2 shows the overlap in all variants and significant variants. BWA-MEM detected the lowest number of variants, followed by the exact matching, whereas most variants were detected by either the additional filtering applied to BWA output (8K-neurons) (Supplemental Fig. S2A) or BBMap (80K-neurons) (Supplemental Fig. S2B).

Variants detected by the default BWA-MEM were detected by all other mapping strategies, and only 0.1% ($n = 2$, 8K-neurons) or 0.2% ($n = 96$, 80K-neurons) of the variants were only in the exact matching strategy for the 8K-neurons and 80K-neurons data set, respectively. Overall, the concordance in identified variants is very high; for example, the proportion of variants detected across at least three out of four methods is 97.6% ($n = 3406$, 8K-neurons) and 94.7% ($n = 37{,}956$, 80K-neurons). When looking into the significant variants, concordance between the mapping strategies drops, with only 48.7% ($n = 416$, 8K-neurons) and 59.2% ($n = 1011$, 80K-neurons) of variants being significant for at least three out of four mapping strategies. We observe larger numbers of variants only significant for one of the mapping strategies. This occurs for 14.4% ($n = 123$, 8K-neurons) of variants (Supplemental Fig. S2C) and 24.3% ($n = 416$, 80K-neurons) of variants (Supplemental Fig. S2D), respectively. Notably for 8K-neurons, BWA-MEM with additional filtering ($n = 721$) and BBMap ($n = 710$) approaches yielded sets that are more than 1.5-fold larger than those of the other strategies (BWA $n = 456$, exact matching $n = 420$). When restricting to the two sets with the highest number of significant variants (BWA-MEM and BBMap), the concordance on the significant variants increased to >90%, which suggests that the higher number of barcodes assigned leads to a more stable variant effect quantification. Apart from the different number of mapped oligos, we could not identify large effects on replicate correlation ($n = 3$) across mapping strategies (see Supplemental Fig. S3).

In summary, the choice of the mapping strategy for the association significantly influences variant analysis outcomes on the count level, particularly for complex and large libraries. Exact matching approaches fail to account for sequencing errors, leading to fewer barcode and oligo assignments, whereas error-tolerant methods like BWA struggle with confident distinction of related sequences, necessitating custom scripts for mitigating this issue. MPRAsnakeflow implements BBMap as standard, because it showed an overall good performance (Fig. 2; Supplemental Fig. S1), but we recommend checking for other mapping strategies if the number of assigned barcodes or retained oligos is low.

Strand sensitivity in the association step

Most read mappers consider alignments for both strands and therefore do not distinguish between the two possible orientations of a test sequence. However, MPRA constructs can have orientation-specific effects so libraries can contain oligos designed in both orientations. Depending on the preparation step for association sequencing, sequence orientation might be lost or maintained. Some MPRA designs test effects of sequence orientation and therefore maintain orientation information in the sequencing library. MPRAsnakeflow provides a strand-sensitive mode, which appends unique flanking sequences to both the reference design and the sequencing reads in the association step. This enables strand-specific analysis and allows for the systematic testing of promoters and enhancers in both orientations (Agarwal et al. 2025).

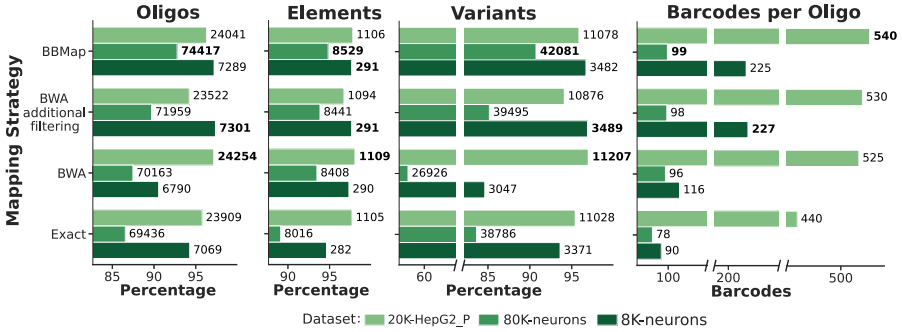

Figure 2. Comparison of four different mapping strategies for the association step for three different MPRA libraries (20K-HepG2_P, 80K-neurons, 8K-neurons). From *left* to *right*, we present the number of assigned oligos, the number of assigned elements (candidate *cis*-regulatory elements, reference sequence only), and the number of retrieved variants. A variant is only considered if both the reference and the alternative allele oligos are assigned. The *rightmost* bars show the number of barcodes per oligo (mean across replicates). The best result for each MPRA library is highlighted in bold in each panel.

Complexity and sequencing depth estimation

A common question is whether deeper sequencing of an MPRA library improves data quality, such as increasing

correlation across replicates. To address this, the pipeline implements two key indicators. First, we use pairs of replicates to estimate overall library complexity using the Lincoln–Petersen method (Petersen 1896; Lincoln 1930). The difference between the observed number of barcodes and the Lincoln index provides an estimate of the potential improvement achievable with additional sequencing. The Lincoln index can also be used to identify replicates or modalities with decreased barcode complexity. Second, the pipeline includes a built-in downsampling option to assess whether quality metrics (e.g., replicate correlation) already reach saturation. If not, this indicates the potential benefit of further sequencing. Downsampling can be performed either on individual barcode counts or on barcodes within the assignment file, using either a fixed number or a specified proportion. Multiple downsampling configurations can be applied within a single experiment run using named configurations, ensuring that only the necessary parts of the workflow are rerun, which saves disk space and computational resources.

For the small library experiment, 8K-neurons (Koesterich et al. 2023), we observe a median of 1,444,480 assigned barcodes across replicates after assignment and RNA and DNA merging, requiring barcodes to have at least one count of DNA and RNA in the full data set. The median of pairwise replicate Lincoln index values is 1,523,572. Consequently, we are missing ~5% of the barcodes from each replicate within this data set. We assume that this might be a common loss from transfections or transductions of MPRA libraries into cell populations.

The medium MPRA library (80K-neurons) (see Supplemental Table S2) presents a median of 5,459,247 assigned barcodes across replicates and a median Lincoln index of 6,243,618. Thus, we are missing ~12% of the library in each replicate. The Lincoln index for the large 240K-HepG2 library (see Supplemental Table S2) is 20,408,916, and we observe a median of 13,038,694 assigned barcodes across replicates. This is ~36% of missing barcodes.

Depending on the size of the designed library and the estimated complexity of the data, we see an increasing gap between missing barcodes in our libraries and a higher variability across replicates. For the medium and low complexity libraries, we see a good concordance between replicates, but the question arises whether more sequencing data could help for the higher complexity library. Therefore downsampling of RNA counts was performed for fixed proportions with MPRAsnakeflow (before assignment) and also for the final assigned count file (barcode reporter experiment) using MPRAlib. For all libraries (for 8K-neurons, see Supplemental Fig. S4A; for 80K-neurons, see Fig. 3A,B; for 240K-HepG2, see Supplemental Fig. S4B), we see a saturation in number of assigned barcodes as well as the observed oligos and the Pearson's correlation of the oligo activity (log_2-fold change). Both approaches show very similar trajectories and similar values for the observed number of barcodes. MPRAsnakeflow downsampling has a smoother curve and is more accurate but takes more time and computational resources, whereas using MPRAlib can be done within minutes, depending on the library size. Based on these results, 8K-neurons and 80K-neurons are already saturated, and new sequencing data will not substantially improve these metrics (Fig. 3; Supplemental Fig. S4A). For 240K-HepG2 (Supplemental Fig. S4B), we still see a slightly increasing line but with no improvement in the number of detected oligos and Pearson's correlation. Therefore, the benefits of resequencing will be neglectable.

A library for interactive MPRA count data analysis: MPRAlib

We developed a Python library, MPRAlib, to provide a simple and accessible framework for analyzing MPRA count data (Supplemental Fig. S5A). MPRAlib uses the AnnData object (Virshup et al. 2024) as its core data structure, with variables representing barcodes or oligos and observations corresponding to experimental replicates. The AnnData object is designed to support both

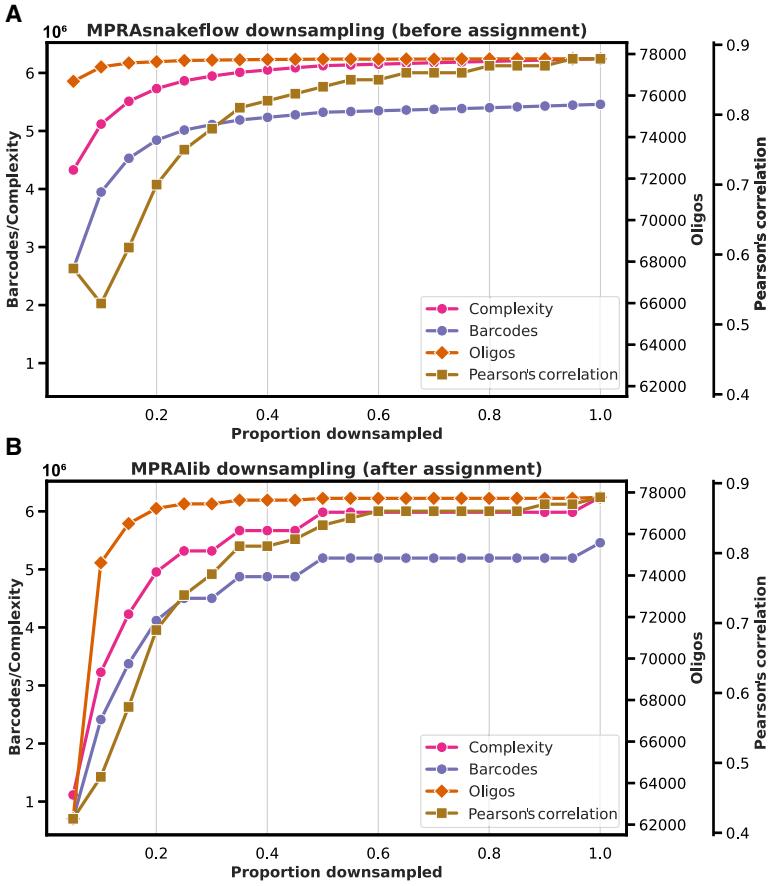

Figure 3. Downsampling analysis for 80K-neurons data set. RNA is downsampled in equal proportions to the overall number of counts. Complexity (Lincoln index), the number of barcodes, as well as retained oligos and Pearson's correlation of oligo activity across replicates (median values across replicates or replicate comparisons). (*A*) RNA count downsampling in MPRAsnakeflow. Sampling is done before the assignment on the raw counts. (*B*) Downsampling using the MPRAlib on the overall output (after the assignment).

barcode-level and oligo-level count data, enabling functionalities such as barcode outlier detection and data aggregation per oligo.

The main class, MPRAData, is an abstract base class that implements most core functions. Two subclasses, MPRABarcodeData and MPRAOligoData, inherit from MPRAData to support analyses at different data granularities. AnnData layers are used to store different modalities (e.g., RNA or DNA counts), normalized counts, and activity measures (such as $\log_2$ ratios). The flexible AnnData structure also allows the addition of metadata and pairwise data to variables or observations, which is leveraged to store information such as the number of barcodes per oligo, masking arrays for filtered data, and pairwise correlations across replicates.

Built-in functions are provided for reading MPRAsnakeflow outputs (at both barcode and oligo levels). We implemented a dynamic approach for adjusting the minimum required number of barcodes per oligo, allowing users to change this threshold at runtime without losing access to the original raw data. This flexibility enables users to recompute metrics, such as correlations, on different thresholds, without reloading the data. At the barcode level, MPRAlib offers various options for detecting outliers, setting minimum required counts per modality, and subsampling counts or barcodes. The library includes a comprehensive suite of plotting functions, such as replicate correlations, DNA versus RNA scatter plots, and histograms of barcodes per oligo (for example plots, see Supplemental Fig. S5B–D). To facilitate integration with differential analysis tools such as BCalm, mpralm, and MPRAnalyze (Ashuach et al. 2019; Myint et al. 2019; Keukeleire et al. 2025), MPRAlib supports a standardized metadata format and includes functions for exporting data in formats compatible with these tools. Additionally, MPRAlib can import results from these tools and generate further outputs, such as BED file tracks for genome browsers.

MPRAlib is extensively documented (https://mpralib.readthedocs.io) and is accompanied by several tutorials using Python notebooks, demonstrating its use in various scenarios. For convenience, we also provide a command-line interface (CLI) implementing commonly used workflows, such as correlation analyses and plotting routines. The CLI allows users to validate input file formats against standardized schemas, which are defined using JSON Schema (see Supplemental Note S1). The library is written in Python (version ≥3.10) and is available for installation via PyPI and Bioconda (Grüning et al. 2018).

Parallel processing and performance benchmarking

We used MPRAsnakeflow to analyze a very large experiment (240K-HepG2) (see Supplemental Table S2) consisting of approximately 240,000 oligos testing for element-level activity (60,000 oligos) and variant-level activity (170,000 oligos). Supplemental Table S3 contains a detailed list of MPRAsnakeflow rules with number of jobs; maximum, minimum, and average run-time; and memory usage on the 240K-HepG2 data set. In summary for the assignment workflow, the three primary input FASTQ files (paired oligo reads and one barcode read) were split into 30 partitions (assignment_fastq_split) resulting in an association pipeline workflow consisting of 164 jobs, with four steps accounting for 150 of these jobs: assignment_attach_idx, 60 jobs; assignment_mapping_bbmap, 30 jobs; assignment_mapping_bbmap_getBCs, 30 jobs; and assignment_merge, 30 jobs. With a maximum number of parallel jobs set to 24, the workflow completed in under 6 h with a maximum of 28 GB RAM required for a single job. The majority of jobs for the assignment step required <10 GB of RAM and completed in under an hour.

The experiment workflow consisted of 103 jobs and required significantly more time because of only six jobs, specifically the counts_umi_create_BAM jobs for the DNA- and RNA-seq files (three of each). The DNA-seq FASTQ files took ∼12 h to process, whereas the larger RNA-seq FASTQ files required ∼34 h; however, these were all run in parallel. The remaining jobs required only four additional hours in total.

We run the same data set with MPRAflow to compare memory and runtime with MPRAsnakeflow. MPRAflow required a substantial amount of more memory (maximum up to 116 GB for the "map_element_barcodes" step) with a slightly longer runtime of 10 h for the rate limiting BAM creation step of the experiment workflow (see Supplemental Table S3).

Barcode discrepancy detection and outlier removal

We applied three outlier detection methods to identify problematic barcodes across nine MPRA data sets, comprising six lentiviral-based assays and three episomal-based assays (indicated by "_P" suffix). The three outlier detection methods included (1) global outliers based on *z*-score thresholds across all barcodes, (2) oligo-specific outliers based on deviation from within-oligo expression patterns, and (3) expression outliers with extreme $\log_2$ expression ratios. Outlier detection patterns revealed noticeable differences between lentiviral and plasmid-based MPRA platforms (Fig. 4A). Global outlier detection showed platform-specific bias, with lentiviral assays exhibiting substantially higher rates (0.67%–1.56%) compared with plasmid-based assays (0.12%–0.47%). Oligo-specific outliers are quite consistent across all experiments, ranging from 0.85% to 1.84%. Notably, large expression outliers are very low in lenti-based assays (up to 0.01%) but had higher rates in plasmid-based assays, in particular 0.9% for 100K_GM12878_P (0.38% for 20K-HepG2_P, 0.03% for 60K-A549_P). This represents roughly fourfold to 90-fold increases over the lentiviral data sets. We also detected a difference in the barcode activity distribution ($\log_2$ ratio) between lenti and episomal assays. Global outliers for episomal showed a high activity, similar to large expression outliers (see Supplemental Fig. S6A).

To assess the stability of barcode-level outlier identification across replicates, we calculated the percentage of barcodes that were consistently identified as outliers across all three biological replicates using the global outlier detection method. Outlier concordance varied across data sets, ranging from 5.6% to 78.6% (Fig. 4B). Although most lentiviral assays demonstrated poor to moderate concordance (5.6%–49.0%), plasmid-based assays exhibited substantially higher consistency rates (56.9%–78.6%). This higher concordance in plasmid assays may reflect more uniform transfection conditions and reduced technical noise from integration-site effects that can vary between lentiviral replicates. We also notice an increased outlier concordance observed in differentiated cell types (cardiomyocytes vs. cardioprogenitors: 43.8% vs. 7.8%, $P<10^{-4}$; neurons vs. WTC11: 49.6% vs. 29.7%, $P<10^{-4}$; for *P*-value estimation, see Methods) despite expected greater cellular heterogeneity in the former from incomplete differentiation. This may reflect more synchronized transcriptional states and reduced stochastic gene expression noise after differentiation. Additionally, differentiated cells may exhibit more stable chromatin landscapes and consistent regulatory network activity compared with progenitor cells, which are inherently more plastic and variable in their transcriptional responses.

To investigate potential sequence-based biases in outlier detection, we analyzed the GC content distribution of barcodes

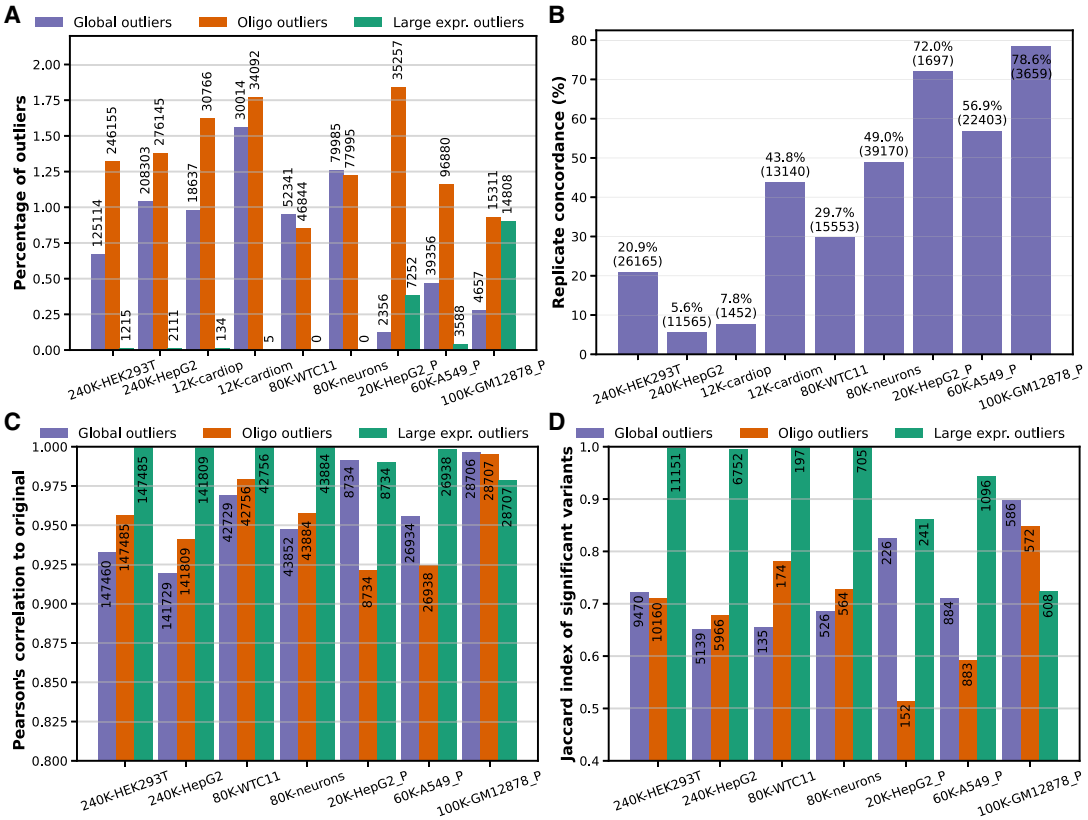

Figure 4. Comparison of different MPRA experiments (lenti and episomal) using different outlier detection strategies. Episomal MPRAs have an "_P" at the end of their ID. (*A*) Outlier detection rates on different data sets using three different methods: global outliers (blue), oligo outliers (orange), and expression outliers (green). The percentage shown is the average among the replicates for global and oligo outliers. (*B*) Concordance of the global outlier detection method across replicates. All lenti-based as well as the 100K episomal assays have three replicates, and the 20K and 60K episomal libraries have six and five replicates, respectively. (*C*) Pearson's correlation of variant coefficients between log-fold change estimates from original data sets versus data sets with outliers removed using three methods: global outliers, oligo-specific outliers (oligo outliers), and large expression outliers. Number of variants per comparison are within the bars. Only data sets with more than 1000 tested variants are shown. (*D*) Jaccard indices measuring overlap of significant variants detected in original versus outlier-filtered data sets. Number of overlapping significant variants are written within the bars. Only data sets with more than 1000 tested variants are shown.

identified as large expression outliers compared with background barcodes (those with no outliers detected by any method) (Supplemental Fig. S6B). Lentiviral-based assays (240K-HEK293T, 240K-HepG2) showed no significant difference in GC content between large expression outliers and background barcodes, with both populations displaying similar median GC content and overlapping distributions. In contrast, all tested plasmid-based assays showed a significant elevation in GC content among the large expression outliers compared with the background controls. These results suggest that GC content bias in large expression outliers is not universal but potentially depends on specific experimental conditions, library characteristics, or cell type–specific interactions with GC-rich sequences. The absence of GC bias in lentiviral assays indicates that integration-based delivery may be less susceptible to sequence composition effects compared with episomal approaches.

To assess whether outlier detection is driven by specific problematic sequences or random technical noise, we analyzed barcode sequence overlap between data sets and the concordance of outlier calls for shared sequences (see Supplemental Fig. S7). The 15-nt barcode space used in most data sets is enormous ($4^{15} \approx 1$ billion possible sequences), resulting in minimal sequence overlap between independent libraries. The data sets derived from the same underlying libraries showed substantial overlap: 240K-HEK293T versus 240K-HepG2 (16.2 million shared sequences), 12K-cardiop versus 12K-cardiom (1.9 million shared), and 80K-WTC11 versus 80K-neurons (5.4 million shared). Despite this substantial sequence overlap in related data sets, outlier concordance was low. For example, for 240K-HEK293T and 240K-HepG2, only three out of 16.2 million sequences were identified as large expression outliers in both data sets (Supplemental Fig. S7B). This discordance between sequence overlap and outlier concordance indicates that outliers are predominantly driven by data set–specific technical factors rather than inherent sequence properties.

To assess whether outlier removal improves the quality of MPRA variant effect analysis, we compared results from BCalm for original data sets and data sets with the outliers removed. We evaluated two key metrics across five data sets: Pearson's correlation of log-fold changes and Jaccard indices measuring the overlap of significant variants detected by each approach. The correlation analysis (Fig. 4C) showed consistently high correlations (>0.92) across data sets for all three outlier removal methods, indicating that outlier removal had minimal impact on overall effect size estimates. Similarly, the Jaccard index analysis (Fig. 4D)

demonstrated high concordance in significant variant calls between original and filtered data sets, with 15 of 21 comparisons (71%) exhibiting Jaccard indices above 0.7. Notably, an inverse relationship was observed between the number of detected outlier barcodes (Fig. 4A) and the Jaccard index, indicating that more extensive outlier removal leads to greater divergence in the set of significant variants identified.

Discussion

In this study, we present a comprehensive suite of tools and standards for the uniform processing, analysis, and sharing of MPRA data within the IGVF Consortium and for the MPRA scientific community. Our main contributions are the establishment of harmonized file formats and community standards to facilitate data integration and interoperability; the development of the MPRAsnakeflow workflow for robust, scalable, and reproducible MPRA data processing; and the MPRAlib Python library for interactive downstream analysis. Making use of such a unified pipeline, we showcase the processing of data sets with varying design strategies and from different experimental groups. The adoption of standardized file formats for design, counts, effects, and genomic mapping, together with schema validation tools in MPRAlib, represents a major advance for data sharing, reproducibility, and meta-analysis in the field. These standards enable seamless integration of data across experiments, laboratories, and consortia and lay the foundation for large-scale comparative studies and benchmarking of regulatory variant effects.

The benchmarking of mapping strategies across diverse MPRA libraries underscores the critical impact of mapping approaches on downstream analyses, particularly for variant-rich designs with small edit distances between sequences. Naive mapping approaches can lead to a substantial loss of variant assignments, reducing the quantifiable effects and the interpretability of the assays. We suggest BBMap as a sensible default for MPRAsnakeflow but enable the easy benchmarking of mapping strategies and suggest it as a best practice in MPRA data analysis.

Our systematic evaluation of barcode outlier detection and removal reveals that although outlier rates and replicate consistency vary significantly across assay platforms and cell types, the impact of outlier removal on downstream variant effect discovery was minimal. We found plasmid-based (episomal) assays to exhibit higher GC-content bias and greater concordance of outlier barcodes across replicates compared with lentiviral assays, which are more susceptible to integration site effects and technical noise. Differentiated cell types displayed higher replicate concordance for outlier detection despite greater cellular heterogeneity, suggesting that differentiation may stabilize reporter regulatory activity and reduce technical noise. Based on the overall small effects on activity estimation and differential effects, we suggest that explicit outlier removal is not necessary for variant effect analysis when using BCalm.

Although our standardization and implementation in MPRAsnakeflow and associated tools provide a major advance and support a wide range of experimental designs, they are not yet optimized for single-cell data, which is an emerging area in regulatory genomics. Certain less common designs also require user customization rather than running out of the box. In these cases, the modular structure of MPRAsnakeflow allows the necessary steps to be added, and we provide worked examples in the documentation. As a result, we are actively working to extend the pipeline and to further optimize it for new assay modalities. As MPRA designs become more complex and library sizes continue to grow, future work will be needed to reduce technical confounders and adapt to new settings. For example, the field will benefit from a comprehensive database of validated positive controls in multiple tissues and shared reference libraries to enable cross-laboratory benchmarking and longitudinal tracking of assay performance.

In conclusion, the tools and practices presented here provide a robust, reproducible, and scalable framework for MPRA data analysis, facilitating the functional interpretation of noncoding genetic variation. By enabling consistent data processing and integration, these resources empower researchers to uncover the regulatory architecture of the genome and to accelerate the discovery of causal variants underlying human disease and phenotypic diversity. We anticipate that broad adoption of these standards and pipelines will catalyze collaborative efforts and drive new insights in regulatory genomics.

Methods

Mapping strategies

We utilize the BWA maximal exact matches (MEM) algorithm (Li and Durbin 2009) with an increased clipping penalty (default: 80) to discourage soft-clipping of read-ends by the aligner, effectively requiring a global sequence alignment. The mapper calculates a mapping quality score (MAPQ) based on the difference between the best and the next-best alignment. In MPRAsnakeflow, BWA-MEM alignments are then filtered using a configurable MAPQ threshold (default: one) to remove ambiguous mappings. Further, we implemented an additional BWA step using MAPQ 0 and filter alignments based on sequence similarity, number of mismatches, and alignment length.

Additionally, MPRAsnakeflow provides support for BBMap (Bushnell 2014), an alternative *k*-mer-based aligner. BBMap reports higher MAPQ values, reflecting increased alignment confidence even for variant-rich libraries. For BBMap, a more stringent default MAPQ threshold (30) is applied to remove ambiguous assignments.

To compare the performance of these mapping strategies, we analyzed three distinct MPRA data sets of different library sizes and design choices: 8K-neurons, 20K-HepG2_P, and 80K-neurons (for a description of the data sets, see Supplemental Note S2; Supplemental Table S2). For each data set, we applied the different mapping routines and compared the number of assignable oligos, the identified genomic regions, and the number of variant sequences detected (see Fig. 2). Only oligos associated with at least 10 unique barcodes were considered for these analyses. Furthermore, we used BCalm (Keukeleire et al. 2025) to obtain a variant effect quantification ($\log_2$-fold change) and a Benjamini–Hochberg adjusted *P*-value per variant, omitting outlier barcodes; to identify significant variant effects (adjusted *P*-value < 0.1); and to assess the consistency of results across the tested mapping strategies.

Complexity and sequencing depth estimation

We randomly downsampled RNA counts from 5% to 95% of the original counts, in 5% increments. This analysis was performed on libraries of varying complexity: low (8K-neurons), medium (80K-neurons), and high (240K-HepG2) complexity (see Supplemental Table S2). After downsampling, we computed the Lincoln index, number of total assigned barcodes, and the total number of oligos retained with at least one barcode. To measure consistency of activity scores between replicates, we calculated Pearson's correlation across replicates at each downsampling level.

We compared these results to those obtained using the downsampling strategy implemented in MPRAlib to assess whether the metrics are comparable.

Parallel processing and performance benchmarking

240K-HepG2 was processed with MPRAsnakeflow (v0.4.4) on a high-performance computing cluster. Individual job execution was managed using the slurm executor plugin for Snakemake: https://snakemake.github.io/snakemake-plugin-catalog/plugins/executor/slurm.html. The number of jobs per Snakemake rule, as well as time and memory usage, was recorded in Supplemental Table S3.

MPRAflow (v2.3.5; Nextlow version 20.01.0.5264) was run on the same high-performance computing cluster as MPRAsnakeflow. Individual job execution was managed by setting the "process.executor" to "slurm" in a configuration file. The reported metrics in Supplemental Table S3 were collated using the "elapsed" and "maxRSS" format options to extract run time and memory usage, respectively, from slurm accounting via "sacct."

Barcode discrepancy detection and outlier removal

For the outlier analysis, seven IGVF data sets, one ENCODE data set, and one other MPRA data set were selected as follows: for lentiMPRAs, 240K-HEK293T, 240K-HepG2, 12K-cardiop, 12K-cardiom, 80K-WTC11, and 80K-neurons; for episomalMPRAs, 20K-HepG2_P, 60K-A549_P (Gosai et al. 2024), and 100K-GM12878_P (Abell et al. 2022). For more description on the data sets, see Supplemental Note S2 and Supplemental Table S2.

Raw count data were processed to calculate counts per million (CPM) for both DNA and RNA libraries in each replicate. We computed the sum of CPM values across replicates for DNA (dna_cpm_sum) and RNA (rna_cpm_sum), followed by the calculation of expression ratios as $\log_2$(rna_cpm_sum/dna_cpm_sum). Quality filtering was applied to remove low-confidence measurements: Barcodes with zero RNA CPM or DNA CPM below the 5th percentile were excluded. Additionally, we retained only oligos represented by at least 20 barcodes across replicates to ensure sufficient statistical power. We implemented three complementary outlier detection approaches:

1. **Global outliers**—For each replicate, we calculated *z*-scores for RNA counts across all barcodes and flagged barcodes with absolute *z*-scores exceeding three as global outliers.
2. **Oligo-specific outliers**—For each oligo, we computed the mean and standard deviation of RNA counts for each replicate. Barcodes deviating more than three standard deviations from their oligo-specific mean were classified as oligo outliers. Standard deviations of zero were replaced with one to handle oligos with invariant expression.
3. **Large expression outliers**—We calculated the median expression ratio for each oligo and identified barcodes with expression ratios exceeding five $\log_2$ units above their oligo-specific median as large expression outliers. These outlier detection methods are also implemented in MPRAlib.

Outlier barcode sequences were analyzed for GC content, using the outlier sequences as the test sequences and nonoutlier sequences identified from all replicates and all methods as a common background. Significant test between the two different distributions was done with the nonparametric Mann–Whitney *U* test. Finally, the variant testing results of the MPRA experiments generated using BCalm (Keukeleire et al. 2025) were compared before and after removing outliers to check for any differences in variant activity as well as different significant variants using the Jaccard index. A significant variant is defined with an adjusted *P*-value lower than 0.1 from BCalm.

Outlier concordance comparison

Given the lack of a suitable parametric distribution describing the outlier concordance metric (proportion $\in [0,1]$), we estimated the *P*-value of the one-sided test that the larger measured value was greater than the smaller by using random permutations to determine null distributions. For a specific null (the smaller concordance proportion), we assumed the ability to detect all outlier barcodes ("null barcodes") in all replicates that would yield the null proportion. We then assumed the additional observed outlier barcodes (n observed $-$ n null) represent random noise and randomly sampled these from the total number of barcodes in the experiment less the number of null barcodes. We then calculated the cardinality of the set consisting of the intersection of all sampled replicates. The *P*-value was reported as the proportion of these sampled consistency metrics that exceeded the observed larger value, namely, the proportion of samples for which we reject the null.

Code availability

All custom software developed for this study is provided as Supplemental Code. MPRAsnakeflow (v0.7.0) (Supplemental Code S1) is available at GitHub (https://github.com/kircherlab/MPRAsnakeflow) and archived at Zenodo (https://doi.org/10.5281/zenodo.18163777). MPRAlib (v0.10.5; Supplemental Code S2) is available at GitHub (https://github.com/kircherlab/MPRAlib) and archived at Zenodo (https://doi.org/10.5281/zenodo.18173084). MPRAOligoDesign (v0.1.2; Supplemental Code S3) is available at GitHub (https://github.com/kircherlab/MPRAOligoDesign) and archived at Zenodo (https://doi.org/10.5281/zenodo.18173304).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work has been supported by the Impact of Genomic Variation on Function (IGVF)/National Human Genome Research Institute (NHGRI) consortium (UM1 HG011966/UM1 HG012003) and by the Deutsche Forschungsgemeinschaft (DFG; 464313370). G.O. is supported by NHGRI grant R01HG012872 and a Pew fellowship award to Steven K. Reilly. Computation has been performed on the HPC for research cluster of the Berlin Institute of Health and the OMICS HPC cluster of the University of Lübeck. We thank all members of the IGVF MPRA focus group for their valuable discussions. We also thank all data producers, especially Chengyu Deng, Jessica D. West, Nadav Ahituv, Nicholas F. Page, Elizabeth Murray, Won Ma, Karen L. Mohlke, and Hyejung Won.

Author contributions: M.S. and M.I.L. conceived the project. K.S., J.D.R., and M.S. performed the mapping benchmarks. J.D.R. performed the performance benchmarks. N.d.L. and M.S. performed the complexity analysis. A.D.V. and M.S. performed the outlier detection analysis. M.S. developed MPRAlib. All authors developed or contributed to MPRAsnakeflow and standardized file formats. M.S., M.I.L., J.D.R., A.D.V., K.S., N.d.L., and M.K. wrote the manuscript.

References

Abell NS, DeGorter MK, Gloudemans MJ, Greenwald E, Smith KS, He Z, Montgomery SB. 2022. Multiple causal variants underlie genetic associations in humans. *Science* **375:** 1247–1254. doi:10.1126/science.abj5117

Agarwal V, Inoue F, Schubach M, Penzar D, Martin BK, Dash PM, Keukeleire P, Zhang Z, Sohota A, Zhao J, et al. 2025. Massively parallel characterization of transcriptional regulatory elements. *Nature* **639:** 411–420. doi:10.1038/s41586-024-08430-9

Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, Yang HT, Xue V, Knyazev S, Singer BD, et al. 2021. Technology dictates algorithms: recent developments in read alignment. *Genome Biol* **22:** 249. doi:10.1186/s13059-021-02443-7

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339:** 1074–1077. doi:10.1126/science.1232542

Aronesty E. 2013. Comparison of sequencing utility programs. *Open Bioinforma J* **7:** 1–8. doi:10.2174/1875036201307010001

Ashuach T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, Yosef N. 2019. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol* **20:** 183. doi:10.1186/s13059-019-1787-z

Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18:** 1196–1203. doi:10.1038/s41592-021-01252-x

Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, Bycroft C, Nicolaisen L, Arvaniti E, Pan J, et al. 2026. Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649:** 1206–1218. doi:10.1038/s41586-025-10014-0

Bushnell B. 2014. *BBMap: a fast, accurate, splice-aware aligner*. Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA.

de Almeida BP, Reiter F, Pagani M, Stark A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* **54:** 613–624. doi:10.1038/s41588-022-01048-5

De Castro-Orós I, Pampín S, Bolado-Carrancio A, De Cubas A, Palacios L, Plana N, Puzo J, Martorell E, Stef M, Masana L, et al. 2011. Functional analysis of LDLR promoter and 5′ UTR mutations in subjects with clinical diagnosis of familial hypercholesterolemia. *Hum Mutat* **32:** 868–872. doi:10.1002/humu.21520

Engreitz JM, Lawson HA, Singh H, Starita LM, Hon GC, Carter H, Sahni N, Reddy TE, Lin X, Li Y, et al. 2024. Deciphering the impact of genomic variation on function. *Nature* **633:** 47–57. doi:10.1038/s41586-024-07510-0

Gaspar JM. 2018. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19:** 536. doi:10.1186/s12859-018-2579-2

Georgakopoulos-Soares I, Deng C, Agarwal V, Chan CSY, Zhao J, Inoue F, Ahituv N. 2023. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nat Commun* **14:** 2333. doi:10.1038/s41467-023-37960-5

Gordon MG, Inoue F, Martin B, Schubach M, Agarwal V, Whalen S, Feng S, Zhao J, Ashuach T, Ziffra R, et al. 2020. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat Protoc* **15:** 2387–2412. doi:10.1038/s41596-020-0333-5

Gosai SJ, Castro RI, Fuentes N, Butts JC, Mouri K, Alasoadura M, Kales S, Nguyen TTL, Noche RR, Rao AS, et al. 2024. Machine-guided design of cell-type-targeting *cis*-regulatory elements. *Nature* **634:** 1211–1220. doi:10.1038/s41586-024-08070-z

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* **15:** 475–476. doi:10.1038/s41592-018-0046-7

Holmes JB, Moyer E, Phan L, Maglott D, Kattman B. 2020. SPDI: data model for variants and applications at NCBI. *Bioinformatics* **36:** 1902–1907. doi:10.1093/bioinformatics/btz856

Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. 2013. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339:** 957–959. doi:10.1126/science.1229259

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27:** 38–52. doi:10.1101/gr.212092.116

Jaganathan K, Ersaro N, Novakovsky G, Wang Y, James T, Schwartzentruber J, Fiziev P, Kassam I, Cao F, Hawe J, et al. 2025. Predicting expression-altering promoter mutations with deep learning. *Science* **389:** eads7373. doi:10.1126/science.ads7373

Jiang K, Liu T, Kales S, Tewhey R, Kim D, Park Y, Jarvis JN. 2024. A systematic strategy for identifying causal single nucleotide polymorphisms and their target genes on Juvenile arthritis risk haplotypes. *BMC Med Genomics* **17:** 185. doi:10.1186/s12920-024-01954-z

Jin W, Xia Y, Nizomov J, Liu Y, Li Z, Lu Q, Chen L. 2024. MPRAVarDB: an online database and web server for exploring regulatory effects of genetic variants. *Bioinformatics* **40:** btae578. doi:10.1093/bioinformatics/btae578

Keukeleire P, Rosen JD, Göbel-Knapp A, Salomon K, Schubach M, Kircher M. 2025. Using individual barcodes to increase quantification power of massively parallel reporter assays. *BMC Bioinformatics* **26:** 52. doi:10.1186/s12859-025-06065-9

Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, Costello JF, Shendure J, Ahituv N. 2019. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10:** 3583. doi:10.1038/s41467-019-11526-w

Koesterich J, An J-Y, Inoue F, Sohota A, Ahituv N, Sanders SJ, Kreimer A. 2023. Characterization of de novo promoter variants in autism spectrum disorder with massively parallel reporter assays. *Int J Mol Sci* **24:** 3509. doi:10.3390/ijms24043509

Kosicki M, Laboy Cintrón D, Keukeleire P, Schubach M, Page NF, Georgakopoulos-Soares I, Akiyama JA, Plajzer-Frick I, Novak CS, Kato M, et al. 2025. Massively parallel reporter assays and mouse transgenic assays provide correlated and complementary information about neuronal enhancer activity. *Nat Commun* **16:** 4786. doi:10.1038/s41467-025-60064-1

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760. doi:10.1093/bioinformatics/btp324

Li J, Zhang P, Xi X, Wang X. 2025. esMPRA: an easy-to-use systematic pipeline for MPRA experiment quality control and data analysis. *Bioinformatics* **41:** btaf315. doi:10.1093/bioinformatics/btaf315

Lincoln FC. 1930. *Calculating waterfowl abundance on the basis of banding returns*. U.S. Department of Agriculture. U.S. Government Printing Office, Washington, DC.

Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR. 2025. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat Genet* **57:** 949–961. doi:10.1038/s41588-024-02053-6

Ludlow LB, Schick BP, Budarf ML, Driscoll DA, Zackai EH, Cohen A, Konkle BA. 1996. Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibβ promoter resulting in the Bernard-Soulier syndrome. *J Biol Chem* **271:** 22076–22080. doi:10.1074/jbc.271.36.22076

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30:** 271–277. doi:10.1038/nbt.2137

Mölder F, Jablonski KP, Letcher B, Hall MB, van Dyken PC, Tomkins-Tinch CH, Sochat V, Forster J, Vieira FG, Meesters C, et al. 2025. Sustainable data analysis with Snakemake. *F1000Res* **10:** 33. doi:10.12688/f1000research.29032.3

Myint L, Avramopoulos DG, Goff LA, Hansen KD. 2019. Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* **20:** 209. doi:10.1186/s12864-019-5556-x

Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee CM, et al. 2025. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res* **53:** D1243–D1249. doi:10.1093/nar/gkae974

Petersen CGJ. 1896. The yearly immigration of young plaice into the Limfjord from the German Sea. In *Report of the Danish Biological Station (1895)* **6:** 5–84.

Rafi AM, Nogina D, Penzar D, Lee D, Lee D, Kim N, Kim S, Kim D, Shin Y, Kwak I-Y, et al. 2025. A community effort to optimize sequence-based deep learning models of gene regulation. *Nat Biotechnol* **43:** 1373–1383. doi:10.1038/s41587-024-02414-w

Reijnen MJ, Sladek FM, Bertina RM, Reitsma PH. 1992. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proc Natl Acad Sci* **89:** 6300–6303. doi:10.1073/pnas.89.14.6300

Reijnen M, Peerlinck K, Maasdam D, Bertina R, Reitsma P. 1993. Hemophilia B Leyden: substitution of thymine for guanine at position -21 results in a disruption of a hepatocyte nuclear factor 4 binding site in the factor IX promoter. *Blood* **82:** 151–158. doi:10.1182/blood.V82.1.151.bloodjournal821151

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26. doi:10.1038/nbt.1754

Rubin AF, Stone J, Bianchi AH, Capodanno BJ, Da EY, Dias M, Esposito D, Frazer J, Fu Y, Grindstaff SB, et al. 2025. MaveDB 2024: a curated

community database with over seven million variant effects from multiplexed functional assays. *Genome Biol* **26:** 13. doi:10.1186/s13059-025-03476-y

van Arensbergen J, FitzPatrick VD, de Haas M, Pagie L, Sluimer J, Bussemaker HJ, van Steensel B. 2017. Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* **35:** 145–153. doi:10.1038/nbt.3754

VanderMeer JE, Ahituv N. 2011. *cis*-Regulatory mutations are a genetic cause of human limb malformations. *Dev Dyn* **240:** 920–930. doi:10.1002/dvdy.22535

Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. 2024. anndata: access and store annotated data matrices. *J Open Source Softw* **9:** 4371. doi:10.21105/joss.04371

Zhao J, Baltoumas FA, Konnaris MA, Mouratidis I, Liu Z, Sims J, Agarwal V, Pavlopoulos GA, Georgakopoulos-Soares I, Ahituv N. 2025. MPRAbase a massively parallel reporter assay database. *Genome Res* gr.280387.124. doi:10.1101/gr.280387.124

Received September 25, 2025; accepted in revised form June 29, 2026.