

Benchmark of Tools and Methods used in the Analysis of Massively Parallel Reporter Assays

A thesis submitted by

Luke Parsley

in partial fulfillment of the requirements for the degree of

Master of Science

in

Mammalian Genetics

Tufts University

Graduate School of Biomedical Sciences

May 2022

Advisor: Ryan Tewhey, PhD

Abstract

Massively parallel reporter assays (MPRAs) are increasing in popularity, as they offer a means of measuring regulatory function of DNA sequences *en masse* and in a variety of contexts. MPRA has been utilized for determining transcription factor coordination (4,9), mRNA stability (10), variant effect (21,22), and more recently for evaluating the transcriptional pressures of chromatin accessibility (7,8). Each of these experimental questions requires slightly different demands for analysis, which has so far made it difficult to establish a unified analytical framework. In the last five years, a series of MPRA pipelines have been proposed to fill this void, including MPRAalyze (16), mpralm (17), QuASAR-MPRA (18), and MPRAmodel (21). Here we evaluated these different pipelines to determine the qualities of their modeling; to do this, we utilized a set of nine MPRA libraries designed for variant analysis across a variety of cell lines. These were designed around either UK and Japan Biobank GWAS or according to fine mapping in GTEx eQTL. Together, these provided a range of barcode and sequencing depth that enabled evaluations of model performance across a spectrum of data quality. We evaluated three primary criteria, including (1) the fitness of their models given a null hypothesis of no differential expression, (2) model accuracy, and (3) significance testing. We also emphasize the importance of normalization by assessing different methods for input to MPRAalyze and mpralm; in addition, we utilize MTSA barcode correction (15) to show how the effects of normalizing by barcode activity shapes downstream modeling. We believe this work can serve as a benchmark for determining the appropriate analytical model for a given MPRA design, as well as provide general recommendations for implementing MTSA in a way that is faithful to the design of these libraries.

Acknowledgements

This work would not have been possible without assistance from a variety of internal and external collaborators. Jacob Ulirsch and Layla Siraj at the Broad institute provide much needed context for the design of the MPRA libraries utilized in this study. Jacob additionally assisted by providing the PIP scores of the CRE-annotated variants that were calculated earlier in the conception of the libraries. Hannah Dewey, a computational scientist in the Tewhey lab and co-author of MPRAmodel, aided my understanding and usage of this analytical pipeline. She also made a variety of updates to the pipeline in the build-up to its first official version that provided it a more modular design. Drs. Ryan Tewhey and Steven Munger, serving as my PI and faculty reader, respectively, displayed great insight, compassion, and patience through this effort. I cannot thank them enough for this. Through the first two years of my graduate education, I was also blessed with a wonderful thesis committee, including Drs. Steven Munger, Steven Murray, and Alexander Poltorak. While the latter two were not directly involved in this Masters' thesis, they were all instrumental in promoting my research goals, providing unique insights, and assisting with grant applications over the years. Everyone in the Tewhey lab were supportive through this process and helped to create a comfortable and caring lab environment. Lastly, I would like to thank everyone in the JAX/Tufts' Mammalian Genetics program, including all students, administrators, and professors. This includes Meredith Theeman, who was instrumental in coordinating my leave of absence and helping me to navigate that time in my life.

Table of Contents

Title Page	i
Abstract	ii
Acknowledgements	iii
Table of Contents	iv
List of Tables	v
List of Figures	vi
List of Abbreviations	vii
Chapter 1: Introduction	1
1.1. Considerations of MPRA Design	1
1.2. Analytical Paradigms and Current MPRA Pipelines	9
1.3. Significance of the Work	15
1.4. Description of Contributions	16
Chapter 2: Materials and Methods	17
2.1. Data Description	17
2.2. MPRA Analysis	18
2.3. Normalization	19
2.4. Barcode Bias Correction	20
2.5. Computational Methods for Data Analysis	21
2.6. Description of Contributions	21
Chapter 3: Results	22
3.1. Common Analytical Considerations	22
3.2. Assessment of Model Fitness for Different Analytical Paradigms	34
3.3. Identification and Correction of Errant Barcode Effects	49
3.4. Description of Contributions	55
Chapter 4: Discussion	56
4.1. Common Considerations for Data Preparation and Design	57
4.2. MPRA Analytical Models Show Suitability for Different Data Constructs	63
4.3. Description of Contributions	68
Chapter 5: Bibliography	69

List of Tables

Table 4.1: Summary of modeling characteristics for each tested pipeline.....	63
--	----

List of Figures

Figure 1.1: MPRA Workflow.....	2
Figure 1.2: Breakdown of UKBB-derived MPRA libraries.....	7
Figure 3.1: Breakdown of barcode trends in UKBB libraries.....	24
Figure 3.2: Assessment of the effects of normalization on spread and model accuracy for OL27 and OL33 in A549.....	26
Figure 3.3: Assessment of the standard error in log2FC as a function of mean DNA reads.....	28
Figure 3.4: Linear modeling of normalized data according to the relationship between logit-transformed expression estimates and mean DNA reads in OL33-A549.....	29
Figure 3.5: Assessment of expression trends of UKBB controls across different libraries and cells.....	32
Figure 3.6: Assessment of sensitivity and specificity trends in UKBB data.....	33
Figure 3.7: Evaluation of p-value calibrations and distributions in OL28-GM12878 and OL30-SK-N-SH.....	36
Figure 3.8: Correlation of log2FC estimates between models and comparison of fitness.....	38
Figure 3.9: Assessment of log2FC as a modifier of sensitivity in different MPRA models for OL30-SK-N-SH.....	40
Figure 3.10: Breakdown of high-consensus log2FC variants in OL30-SK-N-SH for assessment of shared significance.....	42
Figure 3.11: Trends in p-value correlations between models and faceted by numbers of barcodes or reads.....	44
Figure 3.12: Evaluation of p-value calibrations and distributions in allelic skew analyses of OL28-GM12878 and OL30-SK-N-SH.....	46
Figure 3.13: Correlation of p-values and log2FC of allelic skew analyses for all tested MPRA models.....	48
Figure 3.14: Assessment of dispersion in log2FC estimates for each model using either the base or MTSA-corrected counts as input.....	50
Figure 3.15: Breakdown of model fitness and accuracy for base and MTSA-corrected counts of OL28-A549.....	52
Figure 3.16: Heatmap of barcode effect sizes and the expression of mapped RBPs across all cell lines of OL28.....	54

List of Abbreviations

bp – base pairs
CS – credible set
CRE – *cis*-regulatory element
MPRA – massively parallel reporter assay
MTSA – MPRA Tag Sequence Analysis
NGS – Next-Generation sequencing
PIP – posterior inclusion probability
RLE – relative log expression
SNP – single nucleotide polymorphism
TMM – trimmed mean of M values
UKBB – UK Biobank
UQ – upper quartile
UTR – untranslated region

Chapter 1: Introduction

The principle aim of the ENCODE project, as funded and laid out by NHGRI, is to expand our understanding of the human genome and the myriad ways in which it has evolved to regulate itself (1). The first three phases of ENCODE focused on whole-genome analyses of the mouse and human genomes, including the identification of an array of candidate regulatory elements that can serve as chromatin reorganizers, 2D or 3D genome remodelers, epigenetic modifiers, or *cis*-acting transcriptional regulators (2). Many of these elements have been experimentally tested for function and degree of effect, but an even larger portion remain putative due to their great number. To this end, ENCODE funded four rounds of tool development to establish new methods for identifying or evaluating important aspects of the genome *en masse*. One byproduct of this was the improvement and popularization of Massively Parallel Reporter Assays (MPRAs), a high-throughput method for profiling the transcriptional effect of the above-mentioned candidate elements in a variety of contexts. With expanding use of MPRA, the field has been populated by different analytical tools and paradigms, which oftentimes can vary between labs. This makes direct comparisons of datasets and methods difficult, especially when the MPRA libraries themselves are subject to their own design differences. To explore these differences, the following work seeks to evaluate the analytical tools that have been developed for MPRA and to determine their fitness for libraries with barcode multiplicity, an important factor in recent MPRA design.

1.1. Considerations of MPRA Design

MPRA is a tool for measuring the effect of *cis*-regulatory elements (CREs) *en masse*. The purpose of the assay is determined by the construction of an oligonucleotide

library, which can range from hundreds to over a hundred thousand respective sequences.

The general workflow is displayed in **Figure 1.1**. In brief, a library of elements is incorporated into a reporter vector and upstream of a reporter gene (GFP, *luc*, etc.). Each

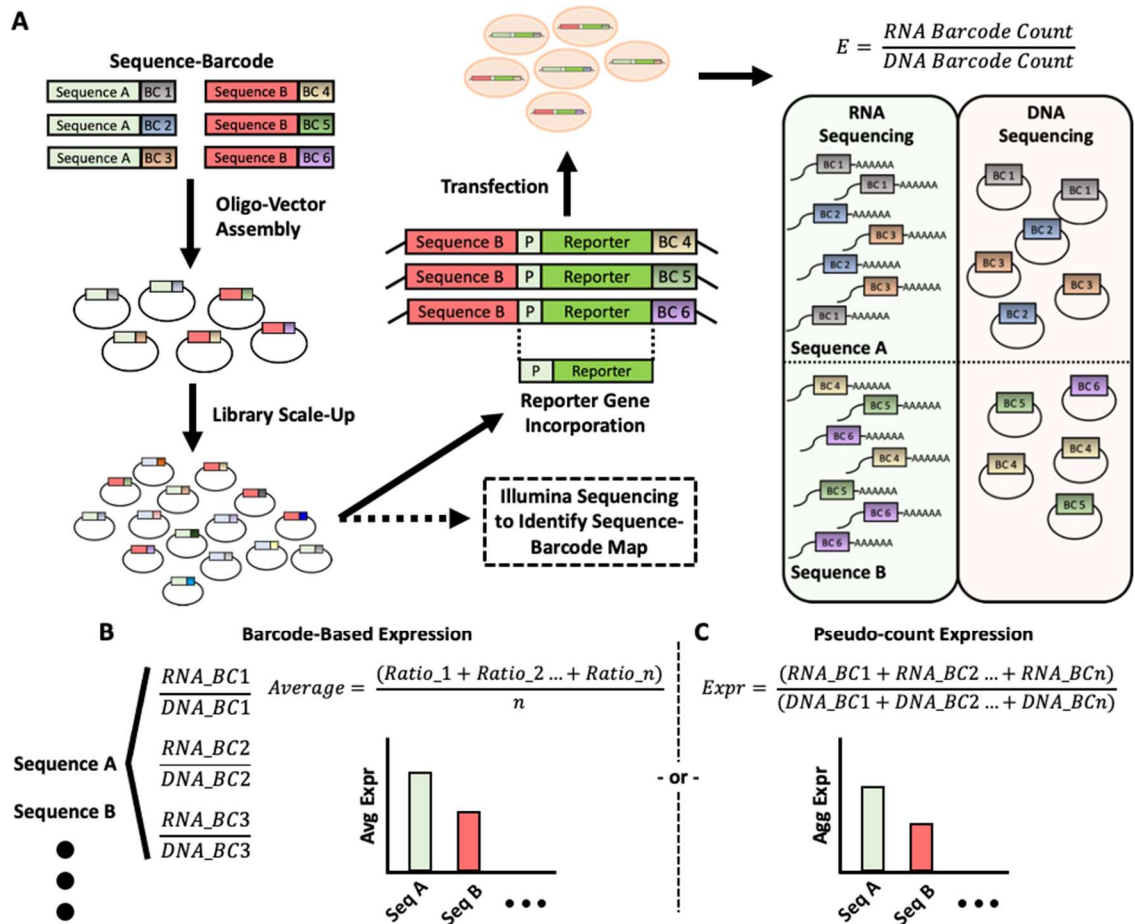


Figure 1.1 MPRA Workflow. (A) Visual depiction of the MPRA workflow, including library design, experimental setup, and basic ratio concept for expression evaluation. (B & C) Barcode-based and pseudo-count methods of organizing and evaluating counts data to determine expression effect of a given element.

plasmid contains a randomized barcode attached to the attendant gene, with the goal of matching every oligo to numerous unique barcodes. This serves a dual function, as it labels each oligo for measurements of mRNA expression while providing a redundancy to these measurements in the case that a barcode influences transcription. The assembled library is then transfected into relevant cell types and mRNA isolated 24 hours afterwards

for use in RNA-seq. Comparative expression values are generated by calculating ratios of plasmid count to mRNA count for each respective oligo-barcode pair.

This paradigm of library design is not universal, with variations made to barcode incorporation, element placement within the construct, and incorporation versus independence of the construct from the genome. The latter of these has become a concern when discussing the translatability of observations to their native contexts. To this end, many groups have introduced their MPRA libraries to cells in both an episomal and chromosomal context, making use of the lentiviral capacity to randomly incorporate into the genome (3-5). Inoue et al. performed a systematic comparison of enhancer element activity in the episomal and chromosomal context, finding that the existing ENCODE annotations and sequence-based models were more predictive in the chromosomal context, though these differences were only modest (3). While these *a priori* predictions may align better to genome-incorporated MPRA, it is important to characterize its cause. Kreimer et al., in what they called the Critical Assessment of Genome Interpretation eQTL challenge, asked participants to evaluate the contributions of different CREs within a given eQTL that had been profiled via MPRA (6). They found that a portion of active enhancers were poorly predicted by most participants; these elements mapped to regions with low DNase hypersensitivity and were enriched for restrictive histone modifiers. They concluded that the original episomal design permitted a false representation of activity, inflating the contributions of sequences that are natively masked in the genome (6). This effort did not include comparisons to lentiviral MPRA, so it is difficult to make concrete statements regarding the suitability of different MPRA models for evaluating enhancers. However, it underscores the importance of genomic profiling in the selection

of elements for MPRA and emphasizes that certain elements may be better suited to certain MPRA models.

For example, the Cohen lab has been specifically interested in the role of chromatin in determining CRE activity and so designed MPRA libraries possessing CREs with ‘strong’, ‘weak’, or ‘repressed’ profiles based on extensive chromatin profiling data from K562, their experimental cell line (8). This enabled them to design lentiviral ‘landing pads’ for different chromatin conditions in the K562 genome such that each CRE would be incorporated into active, inactive, or conditionally active regions. By designing their library and cell line in concert, they were able to control chromatin accessibility as a factor and better predict the effect size of CREs (8). Apart from other efforts by the Cohen group (7,9), one can observe this marriage of library design and experimental setup elsewhere, even inspiring new methods of MPRA that reinterpret its role as an *in vitro* methodology for mass production of mRNA (10). With newer experimental models being generated, it becomes essential to develop analytical tools that are equally modifiable.

Episomal MPRAs are the most common experimental design and are used to profile a wide array of genomic elements or regions. These often require that accommodations be made by both the designer of the library and the one evaluating transcriptional effect to ensure more accurate expression estimates. Central to this is the utilization of putative CREs in library design, especially when establishing sets of controls. By designing MPRA libraries around putative CREs (4) or with specific regulators in mind (miRNA binding sites; 11), one can increase the likelihood that an observed effect size is true; a true effect size is considered the transcriptional effect

exerted by a given element in its native context. Predictive models of miRNA repression by Vainberg et al. were able to accurately model both RNA levels ($R = 0.80$) and protein ($R = 0.84$) in their MPRA, which correlated well with native regulatory patterns ($R = 0.63$; 11). Davis et al., using a set of high-affinity c-AMP response elements from the JASPAR database were able to create an MPRA library that correlated well in the genomic and episomal context ($R = 0.92$; 4); they note, however, that the elements with poor overlap were non-canonical with few annotations, further underlining the importance of rigorous vetting of experimental sequences.

Equally important to episomal design is the orientation and size of the captured sequence. Most MPRA, because of technical limitations to mass oligonucleotide synthesis, have designed their sequences to be < 200 base pairs (bp) in length. Only one effort has been made to systematically determine the effect such a limitation has on transcriptional output. In the study, each tested enhancer was centered on an oligonucleotide with an adjacent sequence capture of 192-, 354-, or 678-bps, deemed short, medium, and long, respectively (12). While the short and medium sequences largely correlated in effect size ($R = 0.78$), this correlation was significantly weaker between the short and long sequences ($R = 0.53$). There are multiple ways to interpret this, ranging from coordinate interactions in the expanded sequence to the inclusion of more potent transcriptional modifiers that may not be immediately related to the tested element. For example, the study noted that the strongest coefficient separating effect size was inclusion of binding sites for RPC155, the catalytic subunit of RNA polymerase III (12). When the experimental question is related to enhancer activity, such a modifier can inflate the true transcriptional control exerted by an element. While it is generally

advisable to increase sequence size in a library, one must be aware of adjacent sequence features to ensure a mitigation of their potential effect.

Orientation relative to the reporter gene can present challenges as well, though this is dependent on the type of CRE being evaluated. Enhancers are known to be functional regardless of orientation within the genome, and this trend is observable in episomal MPRA as well (12); however, multiple MPRA studies of transcription factor (TF) spacing, density, and coordination have found that orientation is a key factor in modulating transcription (4, 9). This would suggest that elements like promoters, or others dominated by TFs, may require specific orientation to the reporter gene according to their native context. Cumulatively, controlling for sequence size, orientation (when appropriate), and filtering by genomic annotation can yield an MPRA library that maximizes information from expression estimates.

With this in mind, a brief discussion of the data sets used in this work is useful for setting the stage for a comparison of analytical pipelines, rather than a dissection of data relevancy. Nine MPRA libraries are used in this work, all generated by the Tewhey lab. Of these, seven are the primary drivers of the analysis performed, and represent a test case for the above-mentioned criteria. These seven are derived from GWAS of ~360,000 individuals in the UK Biobank (UKBB) repository with additional information from the Japan Biobank; the variants from these individuals were selected based on association with one of 96 complex traits which were further sorted into 10 primary domains (**Figure 1.2A**). By fine-mapping these variants, a posterior inclusion probability (PIP) could be generated from the model to determine whether a given variant is likely causal for a particular disease or effect. An additional step was taken to establish credible sets (CS)

by chromosomal position. A credible set is defined as the smallest set of variants for a single effect that cumulatively have > 95% probability of being causal. Only variants within small CSs were selected for inclusion in a library as this maximizes the number of true causal alleles tested. Additionally selecting variants by a PIP cutoff of 0.1 ensures

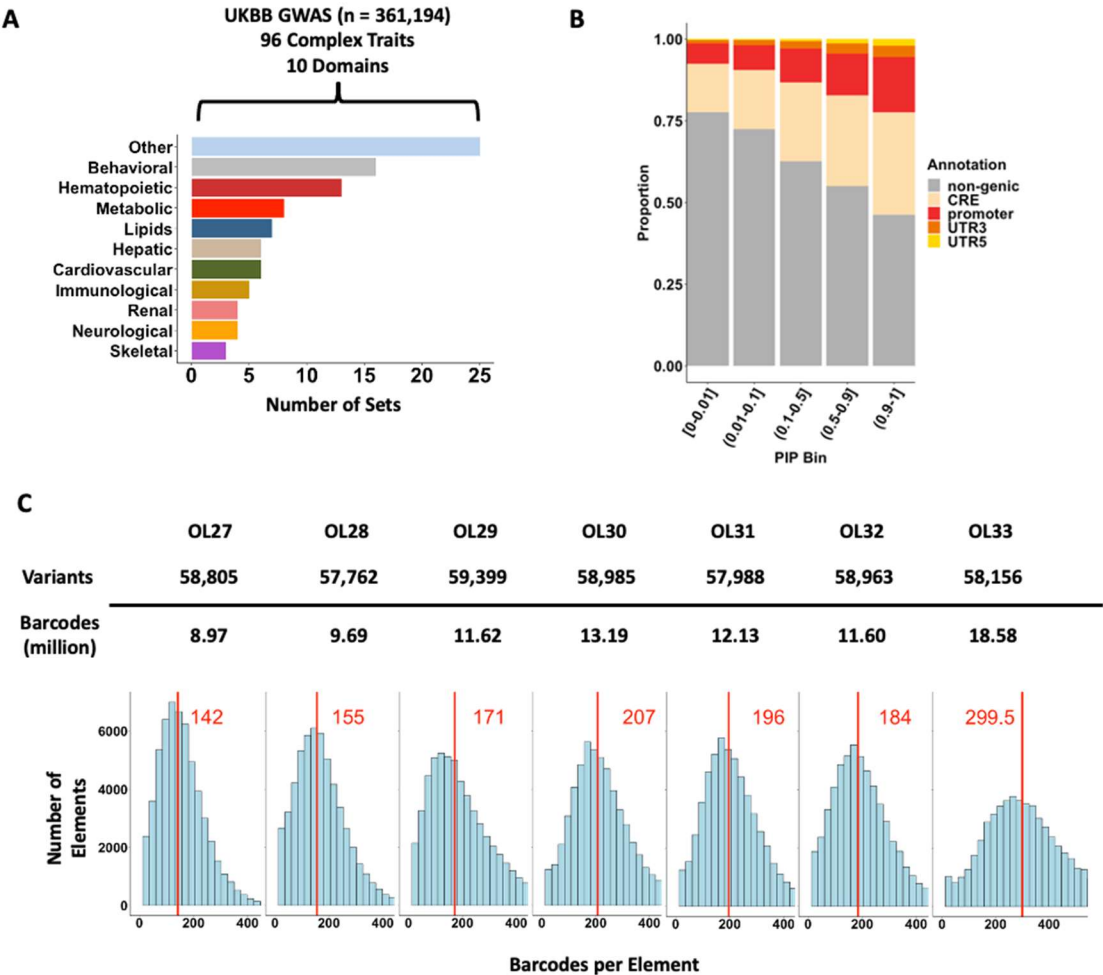


Figure 1.2 Breakdown of UKBB-derived MPRA libraries. (A) Variants identified from GWAS in the UKBB (n = 361,194) were selected based on association with one of 96 complex traits, and further sorted into one of 10 primary domains or an “Other” category. (B) The selected variants from UKBB are sorted into bins based on posterior inclusion probability (PIP). For each bin, variants are organized by type of mutation or by inclusion in a genomic element. (C) Summation of UKBB libraries (OL27-33) as sequenced after full complement of transfections. Variant and barcode numbers are described for each library (top). Histograms of associated barcodes by element are displayed below, with vertical red lines/numbers denoting the median number of barcodes per element within a library.

that a secondary pool of SNPs is similarly enriched for causal variants. We observe that increasing PIP scores correlates with coding regions or elements that harbor transcriptional regulators, like untranslated regions (UTRs), promoters, and CREs (**Figure 1.2B**). Together, these filtering steps provide a set of variants with high predicted effect size and clear association with annotated genomic elements.

A summation of these data sets as sequenced after all transfections is displayed in **Figure 1.2C**. Each has ~29,000 represented variants but vary widely in their respective barcode numbers and median number of barcodes per variant. This is advantageous for a deeper exploration of barcode effects on MPRA. Many previous studies have designed their libraries with a set number of barcodes per element (3,7,8,13,14); many will utilize ~10 barcodes by element, which creates greater individual variance and skew from low barcode depth. The use of randomized barcode incorporation establishes a greater median number of barcodes per element, though this increase to the diversity and size of the barcode pool comes at the cost of replicate consistency; each replicate will receive a slightly different pool of barcodes, which makes it difficult to perform barcode-based analyses. Additionally, randomized incorporation draws from a pool of synthetic tags, which can unknowingly possess motifs for transcriptional regulators that confound expression estimates (15). These issues with barcode read depth, representation, and transcriptional effect can be mitigated through the use of pseudo-counts, in which all reads for a given element are combined. This necessarily eliminates the finer resolution of expression offered in performing a barcode-based analysis (see **Figure 1.1B-C** for breakdown of barcode handling). To correct for barcode biases, Lee et al. have designed MPRA Tag Sequence Analysis (MTSA) as a tool for training and adjusting raw counts

data based on sequence-specific effects from barcodes or k-mers (15). Training can even be done by cell type, where the tool is sensitive enough to identify cell-specific regulators. The application of this tool to data sets of greater barcode complexity than its original study is one of the primary contributions of this work, as much of the initial profiling for MTSA involved fewer elements and set barcode pairings. The UKBB data is especially well-suited for this, with differing degrees of barcode multiplicity for similar numbers of variants. This is best exemplified by OL27 and OL33, possessing 8.97 and 18.58 million barcodes, respectively, and a median value of barcodes per element at 142 and 300 (**Figure 1.2C**).

While this data parity is ideal for evaluating barcode-based estimates of expression, there have been a variety of proposed models for performing MPRA analysis that can accommodate other data aspects, including cell-cell comparisons, allelic skew, counts normalization, shrinkage, and sequencing errors. Most of the project is concerned with comparing these models.

1.2. Analytical Paradigms and Current MPRA Pipelines

Multiple MPRA analytical models have been proposed for general use in recent years (16-19), in addition to the creation of web-based tools for designing MPRA libraries (20); all of this is in service of standardizing MPRA design and analysis within the discipline. At present, MPRA libraries encompass a wide array of potential designs suited to questions of allelic skew (21, 22), miRNA regulation (11), mRNA stability (10), chromatin accessibility (7,8), or transcription factor coordination (4,9). These are fundamentally reduced to quantitative and comparative assessments; however, library complexity between their designs have made it difficult to create a “one-size-fits-all”

approach to analytical modeling. An MPRA tool must be modifiable to accommodate tailor-made modeling and normalization. This work proposes to evaluate four such tools, MPRAalyze (16), mpralm (17), MPRAmodel (21), and QuASAR (18). QuASAR can perform skew analysis, but the foundation of its design is for RNA-sequencing experiments which do not fully translate to MPRA designs. For this reason, it has been previously noted as an outlier in these types of analyses (16,17) and is later dropped in the thesis to focus on more appropriate models.

These four pipelines operate with different analytical paradigms in mind. The paradigms can more neatly be organized into three categories, which will be explored throughout this work: count normalization, accounting for biases in input data, and model fitting. There are other published pipelines available that are not included in this work, namely MPRAscore and malacoda. MPRAscore is styled similarly to QuASAR and is only capable of performing allelic skew analyses (19). As such, it has been ignored in favor of tools that can accommodate multivariate designs. Malacoda presents a tool capable of analyzing most MPRA libraries and is uniquely able to integrate empirical priors in its calculations to estimate shrinkage (23). The authors note that genomic tracks (DNase hypersensitivity, gene scores, DNA methylation, etc.) can be used to train these priors, a powerful feature that no other modular pipeline offers; however, many features of the tool have yet to be implemented and so it too will be ignored in this work.

Common sources of bias in MPRA data are derived from differences in transcription rate between conditions (treatments, cell types, etc.) or from sequencing practices (batch effects, separate DNA/RNA sequencing, etc.). This requires normalization of counts data to control for these disparate factors. RNA-seq experiments,

a sister of MPRA analysis, provide a poor model for normalization as most resolve biases through library size correction (24). In MPRA, sequencing depth between conditions does not necessarily correspond to bias, because the elements tested often do not produce a symmetric effect across conditions. For example, a library of sequences may be designed for greater regulatory effect in one condition than in another, creating a natural parity in total RNA counts. This does not include differences in transfection efficiency between cell lines, or the fact that some elements, owing to poor representation within a library, will be more difficult to detect through the course of Next-Generation sequencing (NGS). Cumulatively, these effects present a challenge to normalization that is typically overcome through compartmentalization of the factors.

Each of the three modular pipelines used in this work offer the user the ability to define conditions for a given replicate. This enables a library size normalization to be performed for a set of replicates with shared conditions. Other methods of normalization that do not directly rely on interindividual variability in library size is employed by MPRAalyze. MPRAalyze performs an upper-quartile (UQ) normalization that adjusts counts for a sample by dividing its contents by the 75% quantile of its reads (16). The effect of this scales according to the most active elements within your sample, so any conditional effects are maintained while still generating a more consistent relative expression. In addition, elements of low read-depth are brought closer to zero, thereby reducing the potential contributions of these non-informative elements to your model. MPRAmodel utilizes DESeq2 for base normalization and performs a secondary normalization step based on the mode of expression within a sample. DESeq2 normalization is termed “median of ratios” and functions by dividing the counts of each

element within a sample by the median of that sample's size factors; these factors are calculated by a division of observed counts by the geometric mean for a given element (25). This method is more a reflection of read distributions than sequencing depth and is thus better able to account for sample differences. MPRAmodel will then take the DESeq2-normalized reads and calculate a logarithmic expression from the RNA/DNA ratio. All expression measures are subsequently subtracted by the mode of the expression distribution according to the assumption that the majority of elements do not produce differences in expression. The additional effort by MPRAmodel to adjust values by the median/mode of its sample estimates provides a strategy that is more resistant to potential skew in counts data, something an UQ normalization would be affected by. Lastly, while mpralm does default to sequencing depth normalization, it offers alternative methods in the form of relative log expression (RLE) and UQ methods, with RLE being more resistant to skew in counts data (17).

Perhaps the most important method for compartmentalization of biases is the separation of the DNA and RNA sequencing steps. This is a method that was proposed and improved upon by the Patwardhan and Mikkelsen groups, in which the plasmid library is sequenced pre-transduction and separate from any future transfections (26-28, 43; **Figure 1.1A**). The advantage is that one does not have to isolate RNA and DNA fractions after transfection, which necessarily leaves only a fraction of the transfected plasmids since these are not as greatly propagated as the RNA. Through repeated DNA sequencing of the plasmids, you can establish a prediction of element-barcode variety and density for a given transfection. While not a perfect reflection of the pool of transfected elements, it has already been shown that at sufficient read depth, the sequencing of these

two separately can produce well-correlated expression estimates between replicates (26). The UKBB data sets were gathered in this manner and this design is well-accounted for among the tested models; however, `mpralm` is alone in assuming any replicates between DNA and RNA correspond with one another.

Other common biases of input data, like batch effect, can be minimized through inclusion in the design matrix of the various models. `MPRAnalyze` and `mpralm` require matrices for both normalization and modeling, meaning that biases can be handled depending on what aspect they are most likely to affect. For instance, batch effects and replicates would be included in the normalization matrix to control for potential oversampling or other non-biological factors in experimentation. Conversely, barcodes might be a part of the model matrix, as fitting a separate, joint model to the expression effects of individual barcodes could correct barcode biases. However, these model matrices fit and adjust data based on sole categorical variables, being unable to account for multivariate biases. This means that barcode modeling would directly reflect the expression of their respective elements rather than deviations of expression within an element. In other words, only the barcodes corresponding to elements that are significantly differentially expressed would be adjusted by the model, thereby regressing measurements to the mean and reducing model sensitivity. It is for this reason that the proposed work is performing barcode bias correction separate from, and prior to, running the data through the various models.

Model fitting varies slightly between the different pipelines. In general, the pipelines model data with a negative binomial distribution due to its ability to accurately fit observed overdispersion, a common characteristic of RNA-seq data (25).

MPRAnalyze fits data to a Gamma-Poisson distribution, which with sufficiently high numbers of observations can approximate a negative binomial (16). Only mpralm employs a different modeling approach, using a variant of the voom methodology (29) that utilizes a linear model to fit the logarithmic relationship of expression ratio to DNA copy number (17). Linear models assume normality of data, something that RNA-seq and MPRA do not conform to; thus, mpralm asserts normality by smoothing a mean-variance relationship to create pseudo-likelihoods of differential expression (17).

Modeling serves other purposes too, often being used to facilitate comparisons or mitigate contributions of uninformative data. For example, MPRAnalyze uses negative controls to determine a base differential behavior between conditions (cell types, treatments, etc.), which is subsequently fitted to a separate, joint model that predicts activity for a given condition (16). These priors are then used to adjust expression estimates for different factors of a condition, increasing the sensitivity of the model in identifying uniquely regulated elements between these factors. MPRAmodel can similarly calculate cell type-specific dispersion, though this is not to facilitate comparisons but to further normalize counts according to their condition. As part of the DESeq2 normalization it performs, MPRAmodel also fits count data to a Bayesian model that “shrinks” observations with low read depth near to zero (25). These shrunken estimates are optional for downstream analysis, though the default is to utilize raw log2 fold-change values. Together, the diversity of these models and the myriad methods for bias correction provide a foundation for a thorough exploration of analytical paradigms in MPRA.

1.3. Significance of the Work

The expansion of MPRA usage across disciplines has led to the development of companion tools for analysis. While many were developed for specific MPRA library designs (18,19), some others can accommodate multivariate data from different MPRA sources (16,17). It is valuable to assess the qualities of these “one-size-fits-all” pipelines as a benchmark for those looking to perform their own MPRA and to assert a standard of best practices in analysis. This work profiles MPRAanalyze, mpralm, MPRAmodel, and QuASAR for model fitness, sensitivity, specificity, effect size distributions, and correlation of calls and activity measurements. Each of these pipelines is published or else has been used in prior works, as is the case for MPRAmodel (21,22). This work additionally compares standard analytical methods of normalization and the usage of controls in MPRA design. Most of these efforts are centered around seven MPRA libraries from the Tewhey lab, which were formed from UKBB GWAS. Each possesses ~29,000 variants but their respective barcode numbers range from ~9-18 million. This variability provides a unique opportunity to evaluate the effects of barcode bias in MPRA data from various cell types and barcode densities. To this end, a barcode bias correction tool (MTSA; 15) is utilized upstream of MPRAmodel to ascertain which barcode motifs produce the greatest effect in different cell types. Improvements to model fitness, sensitivity, and specificity are evaluated after barcode correction to promote the inclusion of this step to general MPRA analysis. This project critically assesses the currently available MPRA analysis pipelines to assert a series of best practices while illuminating the importance of barcode correction to improving analytical power.

1.4. Description of Contributions

All writing and background research in this chapter was performed by the student. The breakdown of the UKBB datasets was largely performed by the student, with the exception of the genomic annotation sets, which were provided as raw data from Jacob Ulirsch, a PhD candidate at the Broad Institute. The quantification of these annotations in **Figure 1.2B** was conducted by the student.

Chapter 2: Materials and Methods

2.1. Data Description

The data sets used in this work were designed as a collaboration between the Tewhey lab (The Jackson Laboratory), Finucane, and Sabeti labs (Broad Institute). They represent an effort to identify causal variants associated with different complex traits, utilizing UKBB GWAS (30), Japan Biobank GWAS (44), and Version 8 (V8) of GTEx eQTL. Variants were fine-mapped to establish a selection of SNPs with higher probability of being causal for a given trait. These were used to create posterior inclusion probability scores (PIP) for each variant. Separately, credible sets (CS) of variants were created to organize SNPs into smaller groupings with a 95% likelihood of containing a causal allele. After library construction, variants were overlaid with DNase hypersensitive regions and with H3K27-acetylation to identify any CRE annotations.

The UKBB libraries comprise ~170,000 variants, split into seven smaller libraries according to a shared association with a set of traits/diseases. A given test variant was selected based on PIP score > 0.10 across any complex traits (see **Figure 1.2A**) or inclusion within a smaller credible set (< 30 variants predicted for trait causality). The GTEx libraries comprise ~240,000 variants, similarly sorted into two major libraries according to shared trait associations. More stringent inclusion parameters were employed to select SNPs with > 0.50 PIP across any of the 49 tissues of the GTEx eQTL or with a 95% credible set containing fewer than 25 variants across three different tissues. Control variants were selected by location matching alternative variants that were either null GWAS or of a similar PIP score but excluded from GTEx eQTL. This produced

~500 negative controls and ~200 positive controls, with slight variations to these pools across libraries based on the inclusion of trait-specific controls.

The UKBB libraries were transfected into five ENCODE cell lines: A549, K562, SK-N-SH, GM12878, and HepG2. The GTEx libraries were transfected into four ENCODE cell lines: K562, SK-N-SH, HepG2, and HCT116. While the GTEx libraries were evaluated with the full complement of MPRA analytical tools, the bulk of data exploration was conducted with UKBB datasets.

2.2. MPRA Analysis

Analysis of MPRA data was performed using three published tools and a fourth in-house methodology. These include MPRAalyze (16), mpralm (17), QuASAR (18), and MPRAmodel, a package designed by the Tewhey lab and used in prior work (22). Optimization of these tools for similarly designed MPRA data yielded the following parameters:

MPRAmodel

MPRAmodel was run according to the default parameters of its package (v1.0.1.0). A summit shift normalization is performed on the log-fold change estimates to center the values at 0, altering this distribution relative to the other pipelines.

MPRAalyze

MPRAalyze was run according to the default parameters of its package, with few exceptions. The normalization method was set to upper quartile and the model matrices were set separately for the different types of analyses. The quantitative analysis was modeled by matrix for normalization but otherwise set to a null expression model. The negative controls were used to fit this model and estimate dispersion of test variants.

The comparative analysis was performed the same except its model matrix included allelic assignments for skew. In addition, the negative controls were not used for model fitting, as they were inappropriate for this design. The tool was set to extract standard errors from the fitted data to be used in a coefficient-based analysis.

mpralm

mpralm was run according to the default parameters of its package, with few exceptions. The tool does not remove zero counts data prior to fitting its model, so an initial filtering step was performed to exclude variants in which plasmid counts were not observed in the five replicates and RNA data was not observed in at least three of five replicates. These criteria were applied regardless of whether the analysis was barcode-based or pseudo-count. In addition, *mpralm* performs replicate matching, meaning it does not support varying numbers of replicates between its DNA and RNA data. As this is sometimes the case with the UKBB data, additional replicates beyond the standard five were removed to enable analysis.

QuASAR

QuASAR-MPRA was run according to the default parameters of its package. Its *betas.beta.binom* value and *betas_se* were used as proxies for log fold-change and its standard error, respectively.

2.3. Normalization

Variations of data normalization were created as input into *mpralm* and *MPRAnalyze*. Normalization of counts by RLE, UQ, trimmed mean of M values (TMM), and library size (called ‘totsum’ here) were performed using the `calcNormFactors()`

function of the edgeR package (31). When performing analysis via mpralm or MPRAnalyze, in-package normalization methods were silenced.

Establishment of Alternative Controls

Establishment of an expanded array of controls was done to create a pool of PIP-centric controls rather than the base location-matched ones. Positive controls constituted any variant with a CRE annotation and a PIP > 0.90. Negative controls constituted any variant with no CRE annotation and a PIP < 0.01.

2.4. Barcode Bias Correction

Barcode correction was performed according to the default parameters of the MTSA package, with a few exceptions (15). The tool was set to exclude variants with fewer than 100 barcodes, each of which must contain at least 5 total reads. Cutoffs were determined by evaluating barcode depth distributions for each library-cell combination and selecting criteria which provided at least 75,000 barcodes for training. These filtered barcodes were used for five-fold cross-validation and training of 8 k-mer effect sizes for correction of the full data. This was performed for eleven total libraries, including all OL28 transfected cell lines as well as all SK-N-SH transfected UKBB libraries.

RBP-Mapping to Barcodes

Effect sizes from the MTSA barcode correction were extracted and the absolute value taken to filter to the top 100 candidate barcodes. These barcodes were used as input to the CISBP-RNA database and its RNA scan tool to identify RNA binding proteins (RBPs) enriched in each tested library using a log PWM threshold of 6 (32). Motifs enriched across 3 or more cells were selected, yielding a pool of 28 total RBPs and their corresponding barcodes. To evaluate whether the patterning of barcode effect sizes

matched RBP expression in the different cell lines, RNA-seq expression profiles of the RBPs were extracted from whole-cell RNA-seq analysis published (33) to the Expression Atlas at EMBL-EBI. Where expression data was absent in whole cell analysis but present in nuclear or cytosolic fractions, these data were used instead. The TPM measurements were logit-transformed before analysis.

2.5. Computational Methods for Data Analysis

All computational analysis and MPRA models were performed in R, which forms the base for most of the tested elements. MTSA barcode correction was performed in an unmodified singularity docker image using compiled python scripts. Scripts for the different analyses are available on github (<https://github.com/c-parsll/MPRA-Benchmark>).

2.6. Description of Contributions

All research methods were optimized and performed by the student. This involved Bash, R, and python scripts written for use in designed Singularity containers on JAX's Sumner computing cluster. Parameters for MPRAmodel were actively discussed with Ryan Tewhey and Hannah Dewey, the two primary authors of this tool. The UKBB/Japan Biobank and GTEx datasets were provided to the student by the Tewhey lab as output from MPRAmatch, an in-house pipeline for counting associated tags in MPRA.

Chapter 3: Results

3.1. Common Analytical Considerations

Many different types of massively parallel sequencing methodologies have been developed in recent years, each presenting their own technical challenges. Two of the more important considerations, both in design and application, are the establishment of appropriate controls and of normalization methods that can accommodate individual sample biases. Typical RNA-sequencing methods will use library size as a scaling factor, or alternatively scale reads by depth within a given genomic frame (Reads Per Kilobase of exon model per Million mapped reads, RPKM) or by the length of its genes (Transcripts Per Million, TPM), though there is overlap of function between these two methods (34,35). RPKM first creates a “per million” scaling factor based on sample read depth which it subsequently applies to individual genes. TPM performs the inverse of this, creating an initial scaling factor based on the read counts contained within the kilobase length of each gene in the sample. The advantage here is that one can make direct comparisons of a gene between samples, as the TPM value in each sample tells you the proportion of total reads that map within it.

These different approaches highlight the central issue of RNA-seq normalization: between samples there exist proportional differences in the RNA population dictated by biological factors like treatment conditions or cell types, or by noise introduced through technical processes like RNA extraction and cDNA generation (34). This is further exacerbated by MPRA design, where expression scores are generated based on independent sequencing of DNA and RNA, and a subsequent ratio of cDNA reads by plasmid reads for a given element (36). The plasmid pool is often sequenced pre-

transduction to produce a prediction of the transfected pool, producing inherent differences in the depth and noise of the RNA and DNA libraries. The standard RNA-seq normalization methods are inadequate for this paradigm; normalizing read proportions by gene length is nullified by the fact that all elements in an MPRA library are the same length, which leads to a translation of the ‘per million’ scaling factor as a variation of library size scaling.

To accommodate comparisons of data between labs using different experimental designs, other normalization methods have been proposed that focus on adjusting counts data according to elements with the greatest observed expression or by expression of a set of assumed controls (35). These are explored here, including UQ, RLE, Trimmed Mean of M-values (TMM), and DESeq2, which performs its own median-of-ratios method that involves additional Bayesian shrinkage of elements with poor information content. DESeq2 is employed by MPRAmodel, so the two are used synonymously in this section. As a control, library size normalization was performed, hereafter referred to as ‘totsum’.

To more accurately describe the effects of normalization for MPRA, an initial assessment of library quality was performed to determine the number of informative barcodes (> 1 read per DNA replicate and > 1 read in at least three RNA replicates) and the total number of reads associated with this pool of barcodes (**Figure 3.1**). This provides a snapshot of data variability, with the expectation that libraries with few informative barcodes and reads possess inherently greater model dispersion. From this, OL27-A549 and OL33-A549 were selected for differential normalization, as this controls for cell-to-cell differences while still providing a library with minimal (~1 million

barcodes, ~19 million reads; OL27-A549) and maximal information (~7.5 million barcodes, ~84.5 million reads; OL33-A549).

The normalization methods can be sorted into two major categories based upon the population of elements used for normalization. UQ methods adjust counts based on

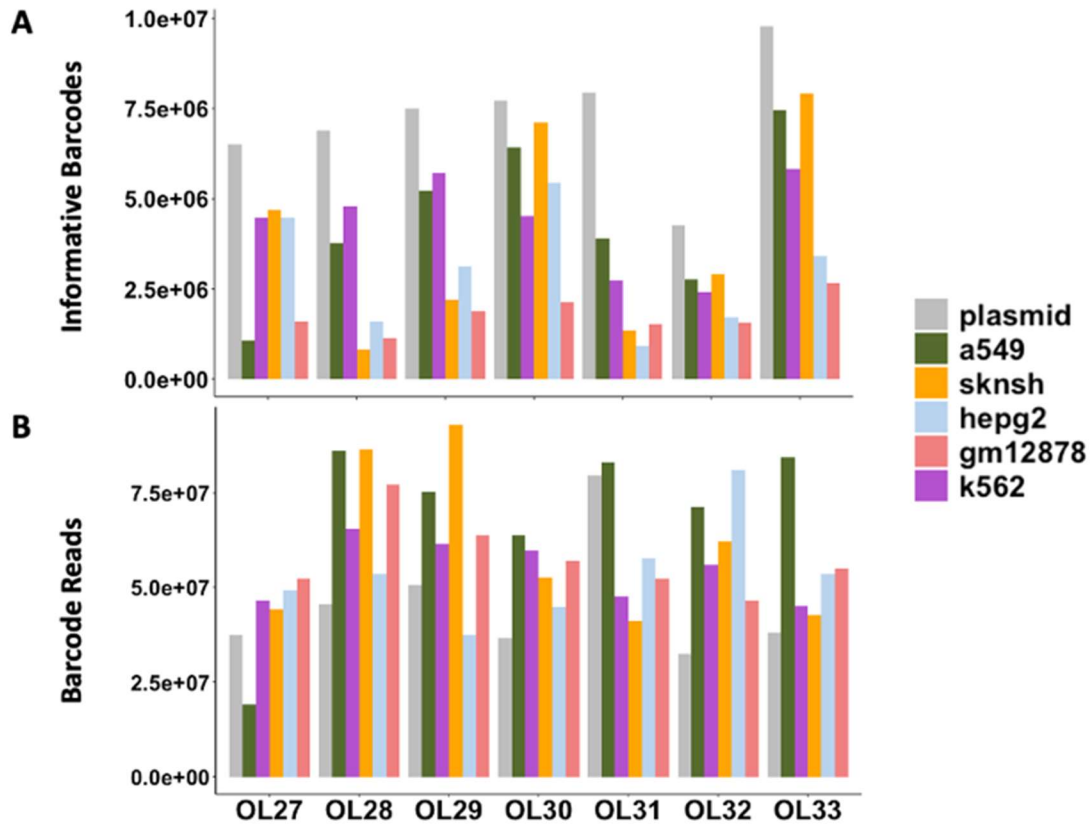


Figure 3.1. Breakdown of barcode trends in UKBB libraries. (A) Number of informative barcodes (> 1 read per plasmid replicate and > 1 read in at least three cDNA replicates) by library and cell. (B) Number of barcoded reads corresponding to informative barcode pools by library and cell.

the highest expressing elements, performing normalization by dividing counts with the upper quartile of reads and subsequently multiplying individual samples by the mean of these upper quartile estimates (24). This permits an adjustment of counts that retains differences in global expression between samples. RLE and TMM methods both assume that most elements are not differentially expressed and use this as a guide in slightly

different ways. TMM utilizes one sample as a reference and computes the weighted mean of log ratios between other samples with this reference, excluding all high-expressing elements. The mean of these log ratios is then used to adjust counts within a given sample. RLE calculates a geometric mean of counts for a given element and then independently divides the observed counts by this measurement. The median of these values within each sample is used as a correction factor. DESeq2 is simply an extension of this method but includes an additional shrinkage of elements with poor read depth (25). It bears mentioning that MPRA data does not always conform to these assumptions. Some MPRA are designed to assess promoters, enhancers, transcription factors, or other positive gene regulators, which means that there would be a prediction of activity. In these cases, RLE, TMM, and DESeq2 methodologies would be inappropriate owing to their assumption of no differential expression; however, the data explored here is enriched for variants with unknown transcriptional effects, meaning this null is sensible.

Since `mpralm` and `MPRAnalyze` provide the user with the option to silence tool-specific normalization, these were chosen for profiling. OL27-A549 and OL33-A549 were normalized according to RLE, UQ, TMM, and `totsum` methodologies, and afterwards used as input into `mpralm` and `MPRAnalyze`. The distribution of standard error for $\log_2FC$ estimates was calculated to assess how well normalization reduced sample-to-sample variation and improved consensus estimates (**Figure 3.2A**). Generally, the `totsum`, or library size normalization, was best at reducing the median of standard error for $\log_2FC$ but performed the poorest at reducing the overall dispersion of this estimate. The distribution and size of the standard error seemed to improve with higher quality data, as OL33-A549 displayed less dispersion and more precise measurements

regardless of the analytical tool used. There do not appear to be significant differences in the precision or distribution of standard error for log2FC between RLE, TMM, and UQ methods. DESeq2 seemed to perform similarly to RLE, though it is difficult to make

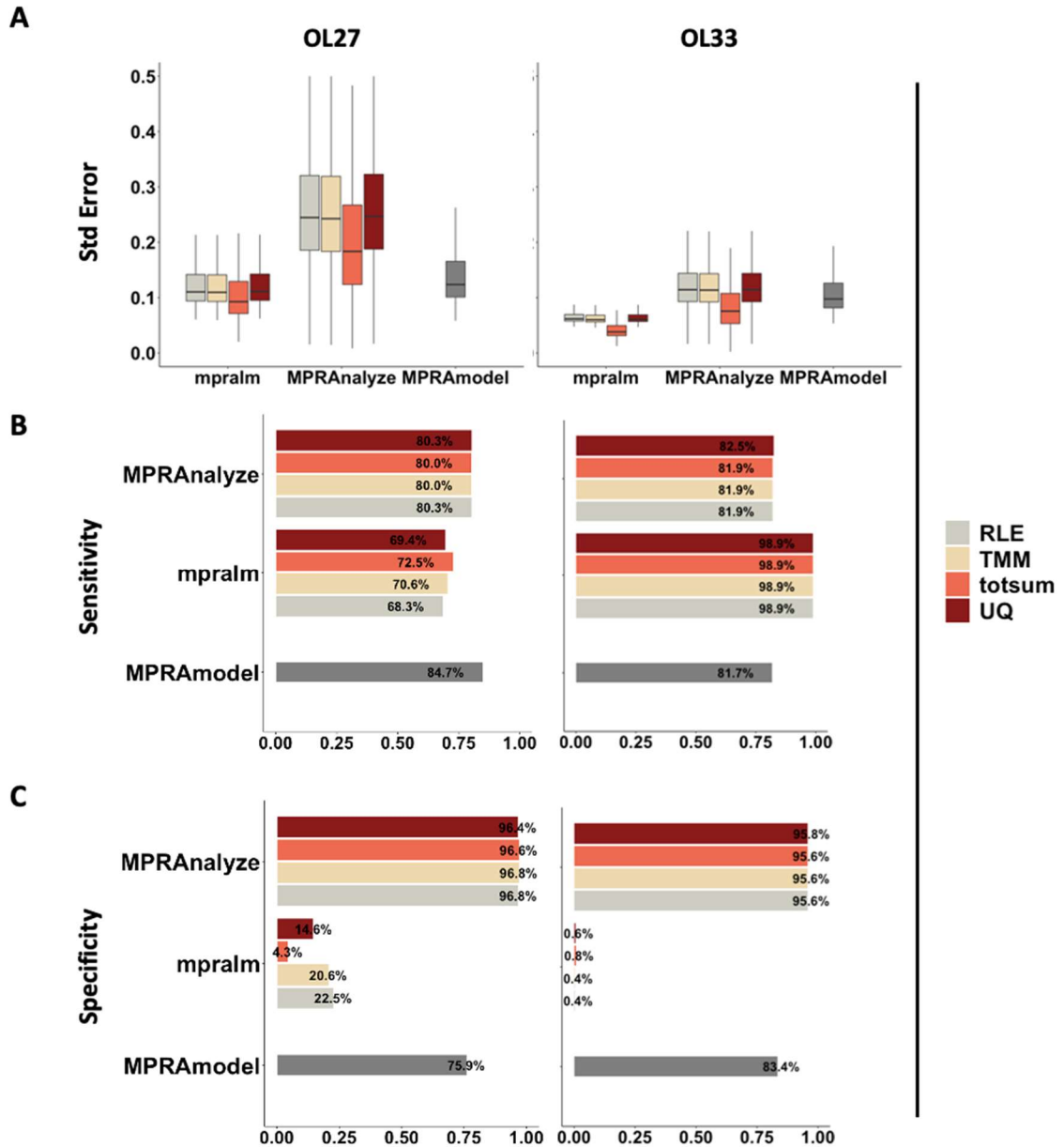


Figure 3.2. Assessment of the effects of normalization on spread and model accuracy for OL27 and OL33 in A549. (A) Distribution of standard error for log2FC estimates by MPRA tool and normalization method. (B) Measurement of MPRA tool sensitivity using library-designed positive controls, organizing by normalization method. (C) Measurement of MPRA tool specificity using library-designed negative controls, organizing by normalization method.

direct comparisons as MPRAmodel does not permit alternative normalization. To see if the normalization methods affected the power of mpralm and MPRAalyze to make more accurate predictions, the designed controls of each library were used to assess sensitivity (**Figure 3.2B**) and specificity (**Figure 3.2C**). These control variants include a variety of location or null GWAS matched SNPs that were agnostic to PIP scores. In OL27-A549, where there is less overall depth of reads and barcodes, we observe greater variability to model accuracy by normalization method. In most cases, sensitivity and specificity are inversely proportional, except in the case of MPRAalyze and MPRAmodel. This is assumed to be the result of their model parameters, which involve an initial fitting of fold-change and plasmid distributions with the negative controls; this would explain why their specificity is relatively unchanged through differential normalization or between OL33 and OL27 (**Figure 3.2C**). In mpralm, RLE and TMM provide a more optimal trade-off between higher sensitivity and specificity by regressing all expression estimates closer to the mean of expression in each sample. The UQ tendency to compensate counts data according to the most expressive elements appears to exacerbate noise in elements with low read counts, ultimately reducing specificity.

To see how spread is mitigated in high quality data, the standard error of expression estimates in OL33-A549 was plotted against the mean DNA reads of its individual elements according to the different normalization methods (**Figure 3.3**). When compared to MPRAalyze, DESeq2 appeared to perform the best at minimizing spread in less informative elements (**Figure 3.3A**), a likely result of the Bayesian shrinkage it performs when encountering elements with poor depth. Of the other tested methods, totsum consistently produced lower standard error regardless of the analytical tool or

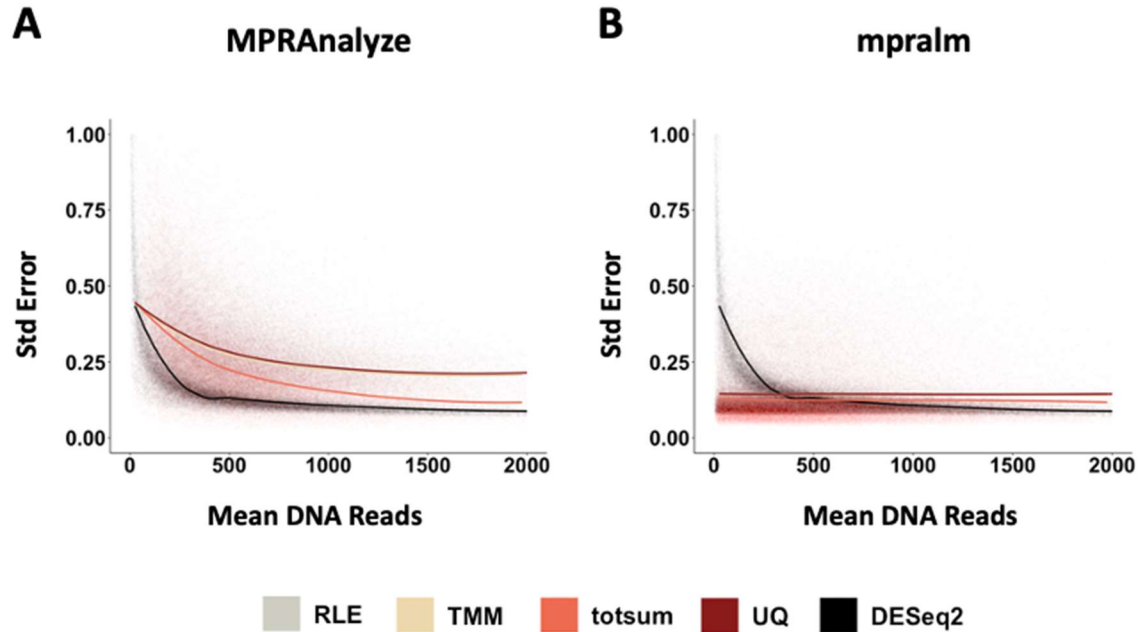


Figure 3.3. Assessment of the standard error in log₂FC as a function of mean DNA reads. Lowess fit of standard error in log₂FC estimates versus mean DNA reads of an element, utilizing different normalization methods in MPRAnalyze (A) and mpralm (B).

mean read depth of the plasmids. The other methods do not seem to vary across this spectrum, recapitulating what was observed in **Figure 3.2A** and further underlining that higher quality data is more robust to differences in normalization paradigms (i.e. normalizing by inter-individual differences in sample expression vs normalizing according to an aggregate of sample expression). mpralm consistently underestimated standard error of expression at all values of DNA read depth, a potential result of its unintentional shrinkage, which occurs because of the model adjusting counts to ensure all replicates have precisely 10 million reads (17). OL33-A549 possesses a total of > 84 million reads across its replicates, so such a step severely undercuts accurate expression estimates.

With this in mind, the normalization methods were evaluated for their effect on model fitness by plotting logit-transformed expression estimates against logit-

transformed DNA counts (Figure 3.4). This method of fitting is directly utilized by

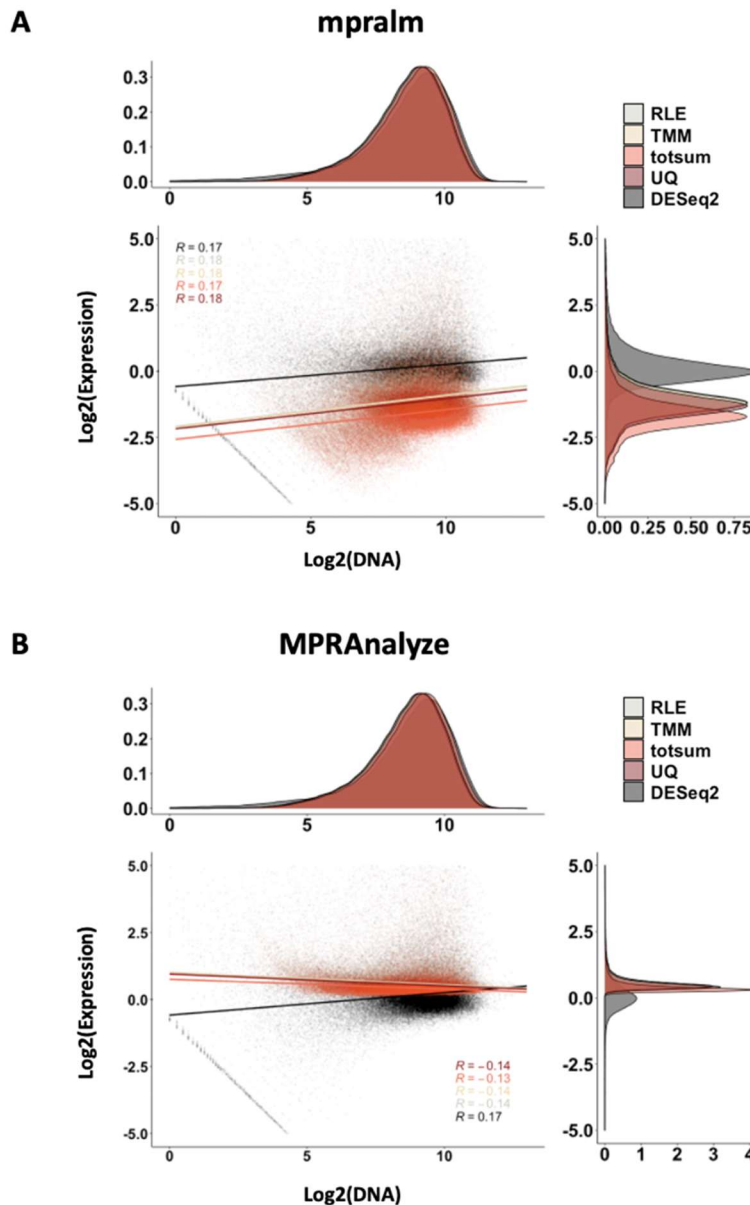


Figure 3.4. Linear modeling of normalized data according to the relationship between logit-transformed expression estimates and mean DNA reads in OL33-A549. Linear models for logit-transformed expression estimates and their corresponding mean DNA reads for each normalization method in mpralm (A) and MPRAnalyze (B). Density plots of $\log_2(\text{DNA})$ and $\log_2(\text{expression})$ are displayed above and to the right, respectively.

mpralm, which performs a linear modeling of the relationship of these factors to produce its expression estimates and their significance. While MPRAnalyze similarly models these factors, it does so with a generalized linear model, which makes this example a corollary measurement of fitness. DESeq2 is displayed as a proxy for MPRAnalyze. Fitness in this case is dependent on our underlying assumptions of the null hypothesis; since most variants are not expected to be active, a “better” fit would be one in which the linear model

does not exhibit a strong relationship between the plotted factors. In both models, the totsum method produces the poorest fit, though the differences are negligible.

Interestingly, the totsum method appears to underestimate expression in relation to observed DNA counts, suggesting the reduced standard error observed in **Figure 3.3** is not the result of improved fitness but of lower overall expression in its data. This is likely a result of totsum diluting global expression differences between replicates by adjusting counts to suit the proportion of reads in a single sample. TMM and RLE appear to slightly overestimate expression relative to UQ, a potential byproduct of the lower read counts generated by UQ from the division of samples by the upper 75% of expressing elements. This also accounts for the improved sensitivity and specificity of these methods observed in **Figure 3.2B-C**, as greater depth can improve the power of the model to determine true negatives.

The normalization methods discussed up to now have represented procedures that rely on the proportional expression of either the group mean/median (TMM/RLE) or on the upper quantile of expressed elements across a set of samples (UQ). Each of these approaches has previously proven inadequate for capturing the technical variability introduced by cDNA preparation, and each is prone to dilution of inter-individual differences in transcription rate (34,35,37). This ultimately reduces the power of the various MPRA analyses for modeling subtler instances of differential expression which can be typical of library designs for less potent transcriptional modifiers. An attractive alternative to these normalization methods is the use of either synthetic spike-ins, a set of control RNA sequences of varying length and GC content used to derive standard curves for transcript abundance (38), or newer methods like the removal unwanted variation

(RUV), which proposes the use of spike-in controls to directly adjust counts based on the performance of the genes (35). Spike-ins, however, are handled poorly by TMM and UQ methods and have been shown to produce poor global-scaling normalization factors (35,37). However, the application of control genes, variants, or samples for normalization remains an attractive option as it permits more direct comparisons of elements by utilizing precise proportional expression between replicates or conditions.

To this end, an evaluation of the controls in the UKBB libraries was conducted to determine their performance across libraries and cell lines. The log₂FC of all location and null-GWAS matched controls were extracted from MPRAModel for A549 in each library; a standard deviation of these expression estimates was calculated by variant to produce a filtering criterion. Thirty variants were selected from each pool to include the 15 most differentially expressed elements (High SD) and the fifteen with the best consensus expression (Low SD) (**Figure 3.5**). When viewing these by library in A549, we discover that the most differentially regulated positive or negative controls are those displaying the greatest expression (**Figure 3.5A**); this suggests that their incongruences are a result of degree of expression rather than exceptional outliers. The consensus variants appear to be expressed in a limited manner for both populations of controls. While this is appropriate for the negative controls, it does call into question the validity of a subset of the expression-modulating controls. These variants behave more variably across cell lines in OL33, even displaying enriched expression in SK-N-SH, K562, and HepG2 (**Figure 3.5B**), indicating a selective utility of controls by cell. While there exists variation in effect sizes of the negative controls, their expression largely correlates across libraries and cell lines.

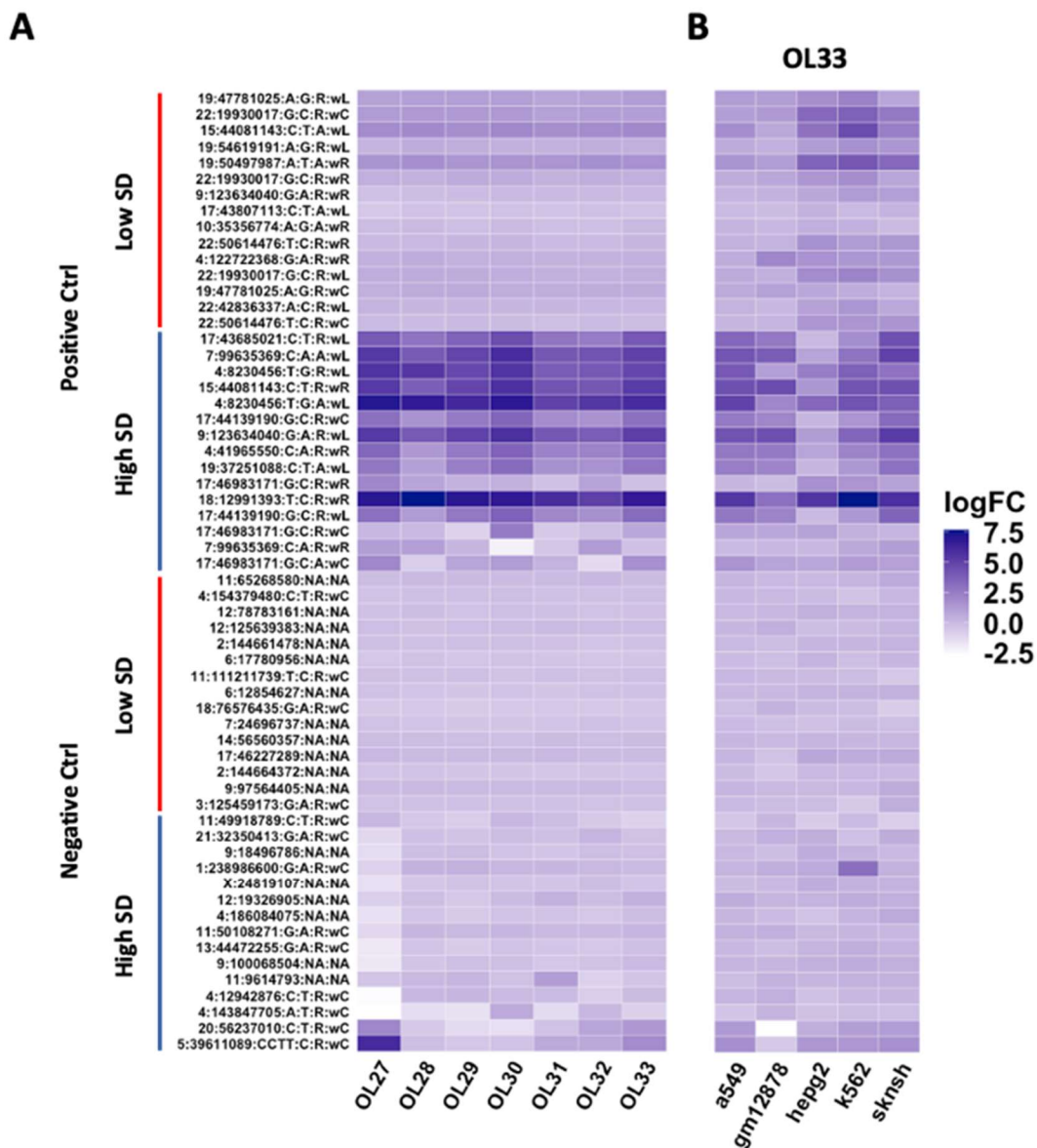


Figure 3.5. Assessment of expression trends of UKBB controls across different libraries and cells. 30 positive and negative controls were selected based on the standard deviation of their log₂FC values in A549 across the seven UKBB libraries. ‘High SD’ denotes the top 15 differentially expressed controls and ‘Low SD’ denotes those with greatest consensus expression. Expression is broken down as a heatmap by library in A549 (A) and across cell lines in OL33 (B).

To assess the suitability of the controls for measurements of sensitivity and specificity, these values were calculated for each individual library-cell combination and plotted against their corresponding barcode numbers or total read counts (**Figure 3.6**).

Sensitivity is directly proportional to barcode number and inversely proportional to read count in MPRAModel and mpralm; the opposite is true of specificity (Figure 3.6A-B).

This makes intuitive sense, as greater read depth would improve the precision of fold-

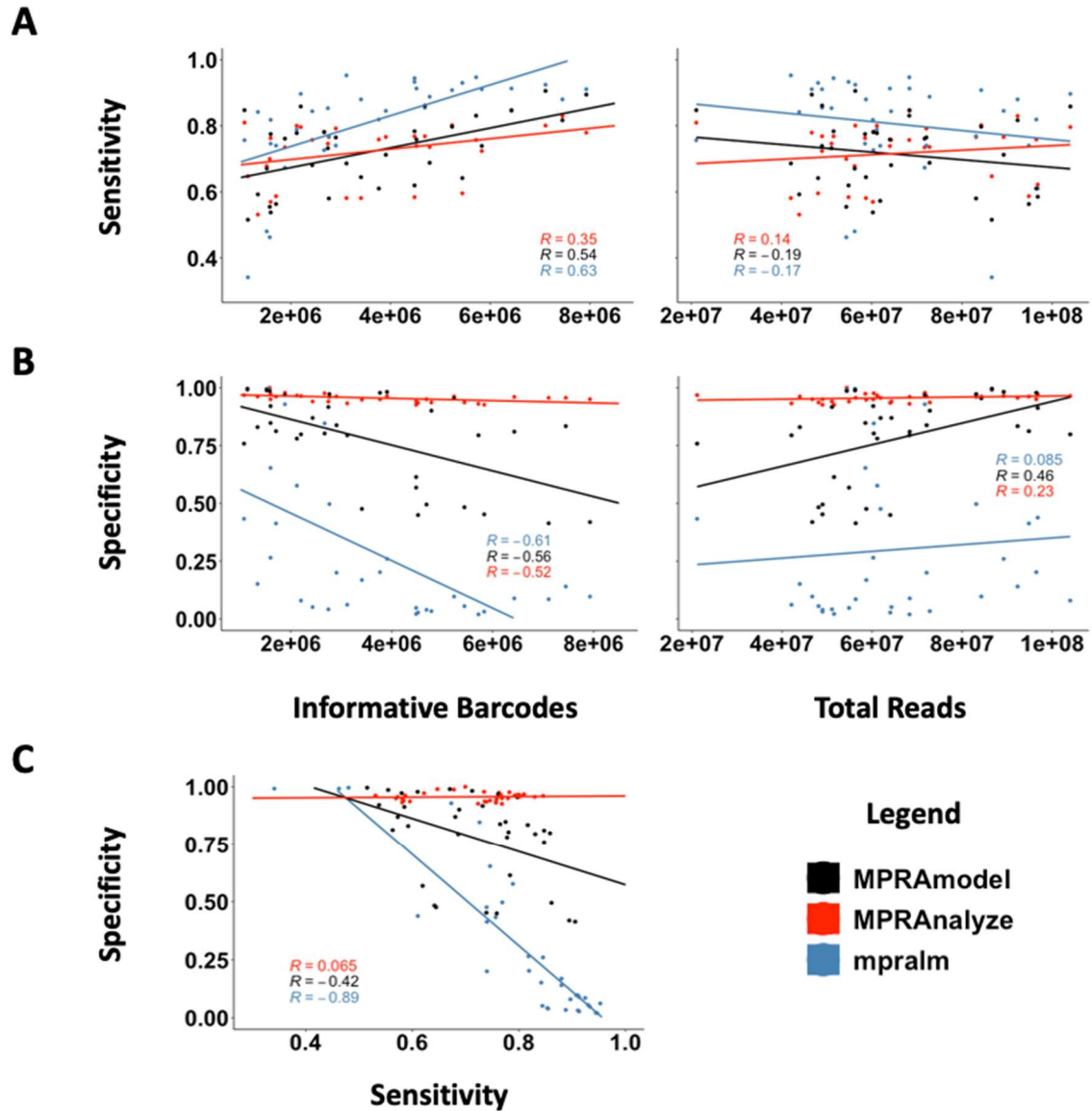


Figure 3.6. Assessment of sensitivity and specificity trends in UKBB data.

Sensitivity and specificity measurements were calculated for each library-cell pairing of the UKBB data in MPRAModel, MPRAnalyze, and mpralm, utilizing the UKBB-derived controls. Sensitivity (A) and specificity (B) were compared to informative barcode and total read counts per library to determine correlation of these factors. Sensitivity and specificity were compared against one another to determine the trade-off of the values in each model (C).

change estimates to discern true expression differences more accurately. Those controls which were behaving variably in **Figure 3.5** would likely be described better in datasets with greater depth, thus reducing sensitivity and improving specificity. In general, the inverse trend of specificity and sensitivity is observed in the analytical models, though the severity of this trade-off is exaggerated in *mpralm*, a byproduct of its inability to accommodate controls in its design matrix (**Figure 3.6C**). *MPRAnalyze* remains largely unaffected by changes to barcode number or read count in its measurement of control effects. While this model estimates factorial contributions to expression, its primary fitting is independently performed using negative controls, which would explain its consistent estimation of their effects. This makes *MPRAnalyze* an attractive option for MPRA data with less information, though the user must be confident of their controls for model fitting. The UKBB-designed controls appear to perform relatively well in the different models but their variability across cell lines and libraries suggests a poor fit for use in control-based normalization. Alternative controls are explored later for model fitness.

3.2. Assessment of Model Fitness for Different Analytical Paradigms

The three major MPRA tools evaluated in this work were exposed to a battery of tests to determine their performance for both quantitative and comparative analyses. *MPRAnalyze* and *MPRAmodel* were conducted as described by their default parameters, except in the case of comparative analysis, where *MPRAnalyze* was instructed not to use controls for modeling and to directly compare expression coefficients rather than perform a logarithm-of-the-odds analysis. *mpralm* can perform both pseudo-count and barcode-based analyses, so ‘*mpralm_barcode*’ will serve as representative for this alternative

MPRA analytical paradigm. The expression estimates from this method are derived from an averaging of expression across barcodes for a given element; these values are inherently more volatile owing to the poorer depth of reads in individual barcodes. Additionally, `mpralm` assumes direct translation of replicates in DNA and RNA data, which is not the case in the UKBB libraries. To assess the effects of this on modeling, and to control for differential plasmid representation across replicates, an alternative form of DNA counts was created in which the average of plasmid counts was used for each replicate. No significant differences in model fitness or accuracy were found (data not shown), so DNA counts were kept in their original format to better represent the biological noise of these data.

To assess the range of data quality for modeling, the ratio of barcodes to reads was evaluated for each UKBB library. From this, OL28-GM12878 and OL30-SK-N-SH were selected for representing low and high barcode:read ratios, respectively. This creates a parity in the precision of estimates for all models, which is informative for how each one handles data dispersion. This is especially important for `mpralm_barcode`, as a low barcode:read ratio should improve model performance due to higher quality observations. For quantitative analyses, a breakdown of significant calls and their distributions can be seen in **Figure 3.7**. Using Q-Q plots as a proxy for calibration of p-values, we find that most models display greater conservatism with a lower barcode:read ratio, with `mpralm_barcode` exhibiting the inverse relationship (**Figure 3.7A**). This may indicate a greater overall sensitivity of this model when provided with increased read depth, mirroring observations of sensitivity in `mpralm` from pseudocount read depth (**Figure 3.2B**). The calibration of p-values in `MPRAnalyze` is relatively unchanged by

this ratio, though the distribution of p-values in **Figure 3.7B** displays markedly different patterns that more closely align to total read depth than they do the barcode:read ratio.

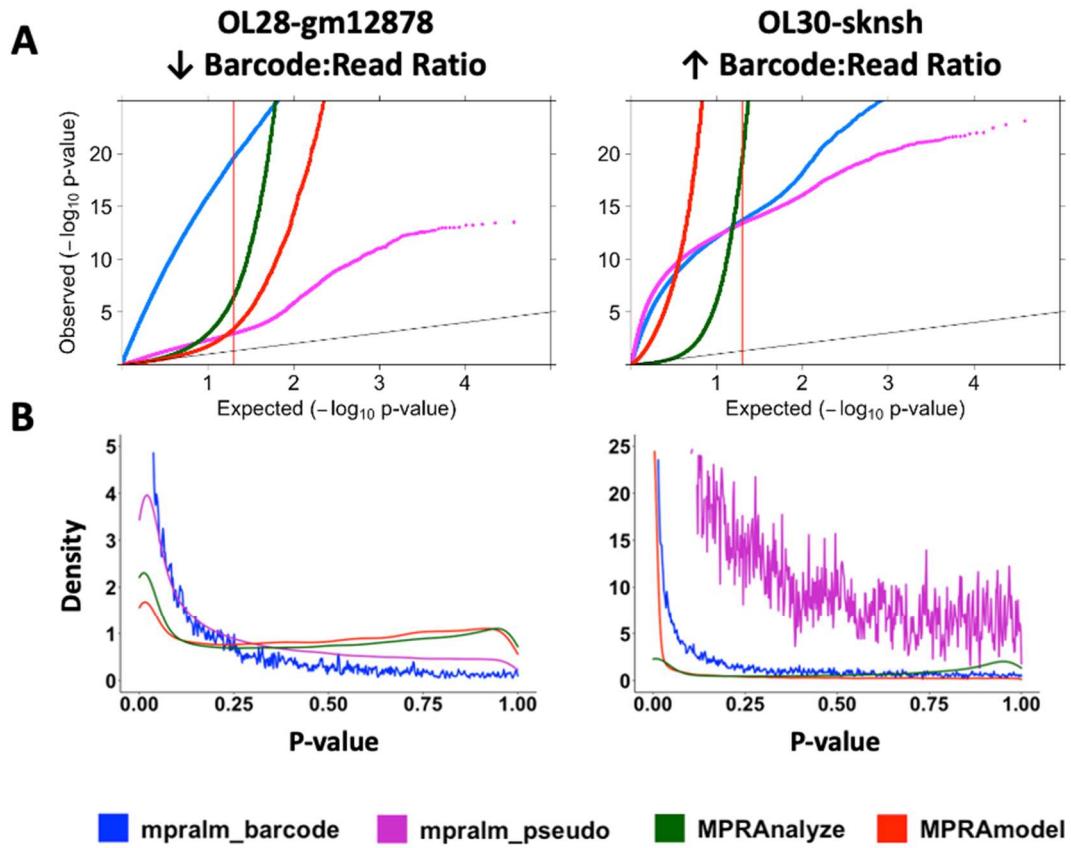
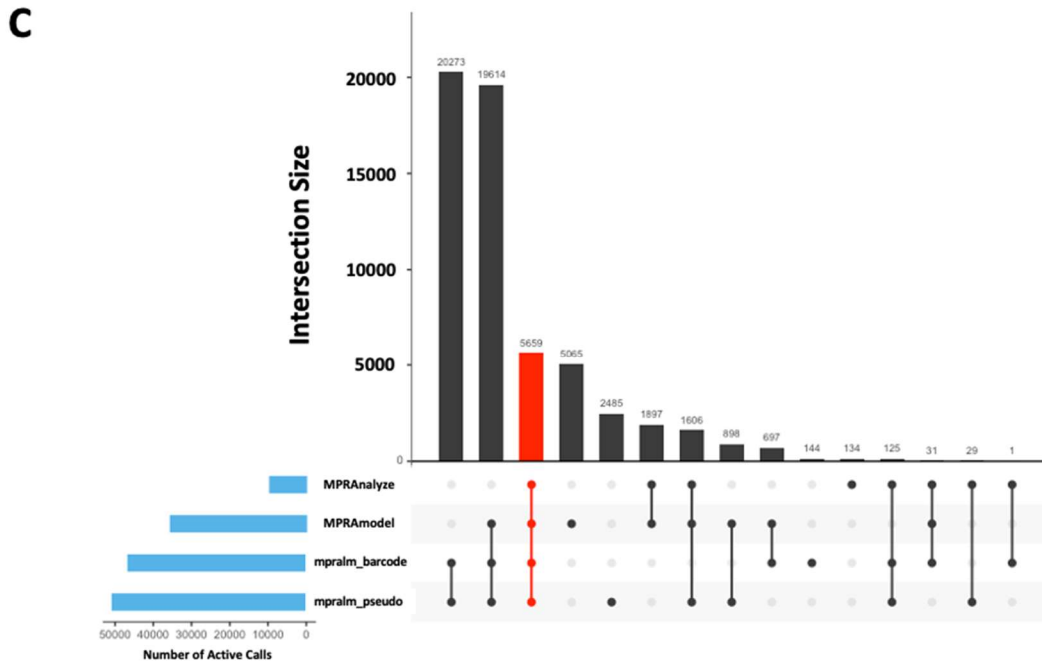


Figure 3.7. Evaluation of p-value calibrations and distributions in OL28-GM12878 and OL30-SK-N-SH. (A) Q-Q plots and (B) density plots of p-values from each model for data sets OL28-GM12878 and OL30-SK-N-SH. The vertical red line in Q-Q plots denotes $p\text{-value} = 0.05$. (C) Upset plot of significant calls (BH-adjusted $p\text{-value} < 0.01$) for OL30-SK-N-SH broken down by model. Bar plot in bottom left corner displays the number of calls made per model. Main bar plot displays the number of significant calls overlapping for different pairings of models. The red-highlighted bar indicates shared calls across all models.

OL28-GM12878 and OL30-SK-N-SH possess ~86 and ~56 million total reads, respectively. While the difference is not as great as those data sets observed in section 3.1, we observe more classical p-value distributions in OL28, with uniform spread except for an enrichment at low p-values. In both data sets, `mpralm_barcode` exhibits a divergent distribution enriched for p-values across a wider spectrum. Interestingly, `mpralm_pseudo` performs worst in the data set with lower total reads, remaining enriched for p-values across its entire spectrum. When peering at significant calls for each model in OL30-SK-N-SH (BH-adjusted $p\text{-value} < 0.01$), we observe `mpralm_pseudo` and `mpralm_barcode` with ~50,000 total calls each and `MPRAnalyze` with a more conservative ~5,000 calls; of these calls, only 134 are unique to `MPRAnalyze` as the rest are shared between the various pipelines. `MPRAnalyze` makes ~40,000 calls, of which half are shared with `mpralm` predictions.

The lack of concordance between the different models might suggest differential estimation of expression, except in the cases of extreme skew between the RNA and DNA data. To assess this, $\log_2\text{FC}$ values were extracted from each model and correlated with results from `MPRAnalyze` (**Figure 3.8A**). $\log_2\text{FC}$ values correlate well between `MPRAnalyze` and `MPRAnalyze` (0.99 and ~1 for OL28 and OL30, respectively) and between `mpralm_pseudo` and `MPRAnalyze` (0.94 and ~1 for OL28 and OL30, respectively). $\log_2\text{FC}$ values correlate variably between `mpralm_barcode` and

MPRAmodel, with correlation improving as the barcode:read ratio increases. The increased barcode:read ratio likely limits the contributions of expression outliers to the estimate and reduces the overall deviation of the data. It is important to note that the number of significant calls by mpralm_barcode was markedly higher in OL30-SK-N-SH as compared to OL28-GM12878, so the poor precision of log2FC estimates in OL28-GM12878 does not necessarily detract from modeling. Rather the increased variance seems to have deflated the significance of elements with poor read depth, providing a more conservative model. We see this mirrored across all models when approximating their fit for high (OL28-GM12878) and low barcode:read ratios (OL30-A549) (**Figure 3.8B**). All exhibit improved fitness in OL30-A549 as compared to OL28-GM12878, with MPRAmodel and mpralm showing the most significant improvements. The fitness of MPRAanalyze remained relatively unchanged regardless of data quality. This is perhaps a result of the log-additive method employed by MPRAanalyze for calculating expression, where it independently models expression contributions for a variety of standard (replicates, batches, barcodes) and user-designed factors (cell types, alleles, treatment conditions).

To determine how reliant significance testing is on base log2FC estimates for each model, and in particular for MPRAanalyze, variants were selected from OL30-SK-N-SH for which there are CRE annotations (UTRs, promoters, or enhancers). These variants were binned in a nested fashion according to their log2FC estimates and p-values to produce a heatmap that colors according to the fraction of CRE-positive variants (**Figure 3.9**). These CRE-annotated variants serve as an alternative approximation of a positive control population. In all models, we observe that the greatest fraction of CRE-positive

variants are contained within the highest-expressing bins. In MPRAmodel and mpralm-pseudo, these variants are largely contained in the upper-most row, corresponding to p-

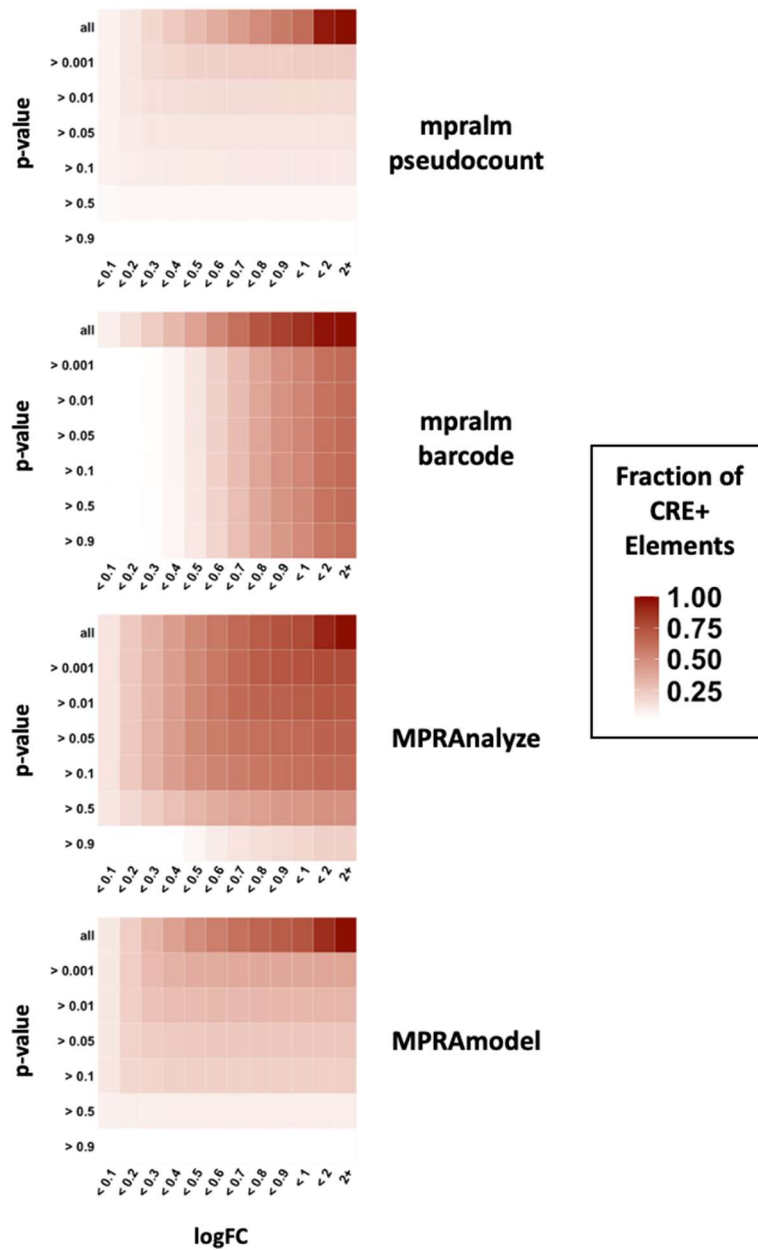


Figure 3.9. Assessment of log2FC as a modifier of sensitivity in different MPRA models for OL30-SK-N-SH. All variants in OL30-SK-N-SH for which there exist a CRE annotation (UTR, promoter, or enhancer) were selected. Variants were recursively binned according to their log2FC estimates and p-values in each MPRA model. Heatmaps were created displaying each bin, coloring in intensity the fraction of CRE-positive elements contained within their respective categories. Since all bins are recursive to those that came before them, this produces a gradation of sensitivity that is primarily modified by log2FC.

values < 0.001 , with an increasing gradation of inclusion as log2FC increases. This

suggests that base log2FC estimates are the primary drivers of their significance testing.

While mpralm_barcode contains most variants in the highest significance bin, a large

fraction are sorted across a range of p-values, especially in the higher log2FC bins. Given

that it performs the same analysis as `mpralm-pseudo`, we can assume that the divergence in p-values is a result of the model accounting for barcode number differences in the variants. `MPRAnalyze` displays a wide range of variant inclusion across p-value and $\log_2\text{FC}$ bins, underlining the design of the model to independently adjust expression estimates according to alternative factors. This could explain why the model produced 8-fold fewer significant calls as compared to `MPRAmodel` (**Figure 3.7C**). This recommends `MPRAnalyze` as the most conservative fit and lends credence to its multifaceted design as a more specific predictor of true expression differences.

To determine how significance varies across different $\log_2\text{FC}$ bins, all $\log_2\text{FC}$ values were extracted from each model for OL30-SK-N-SH. A standard deviation (SD) of expression estimates was taken to filter out all variants with > 0.1 SD, and then ten representative elements were selected from each of six fold-change groups, including 0.1, 0.5, 1, 1.5, 2, and $3+ \log_2\text{FC}$. These variants were plotted as a heatmap and their size faceted according to their p-value to observe the concordance of significance in variants with good consensus expression (**Figure 3.10**). In general, we find agreement between the models in determining the significance of variants across the $\log_2\text{FC}$ bins. While `MPRAmodel` appears to match $\log_2\text{FC}$ intensities, it assigns significance to nearly all variants regardless of this expression estimate. This could be interpreted as a byproduct of the low standard deviations of these groups improving the significance of their signal, something `MPRAmodel` would be especially susceptible to. As part of `MPRAmodel`'s normalization steps, `DESeq2` will perform a Bayesian shrinkage of poorly covered elements, bringing their activity closer to 0. This enables a model fitting that is agnostic to the volatility of variants with low read counts, permitting a detection of significance

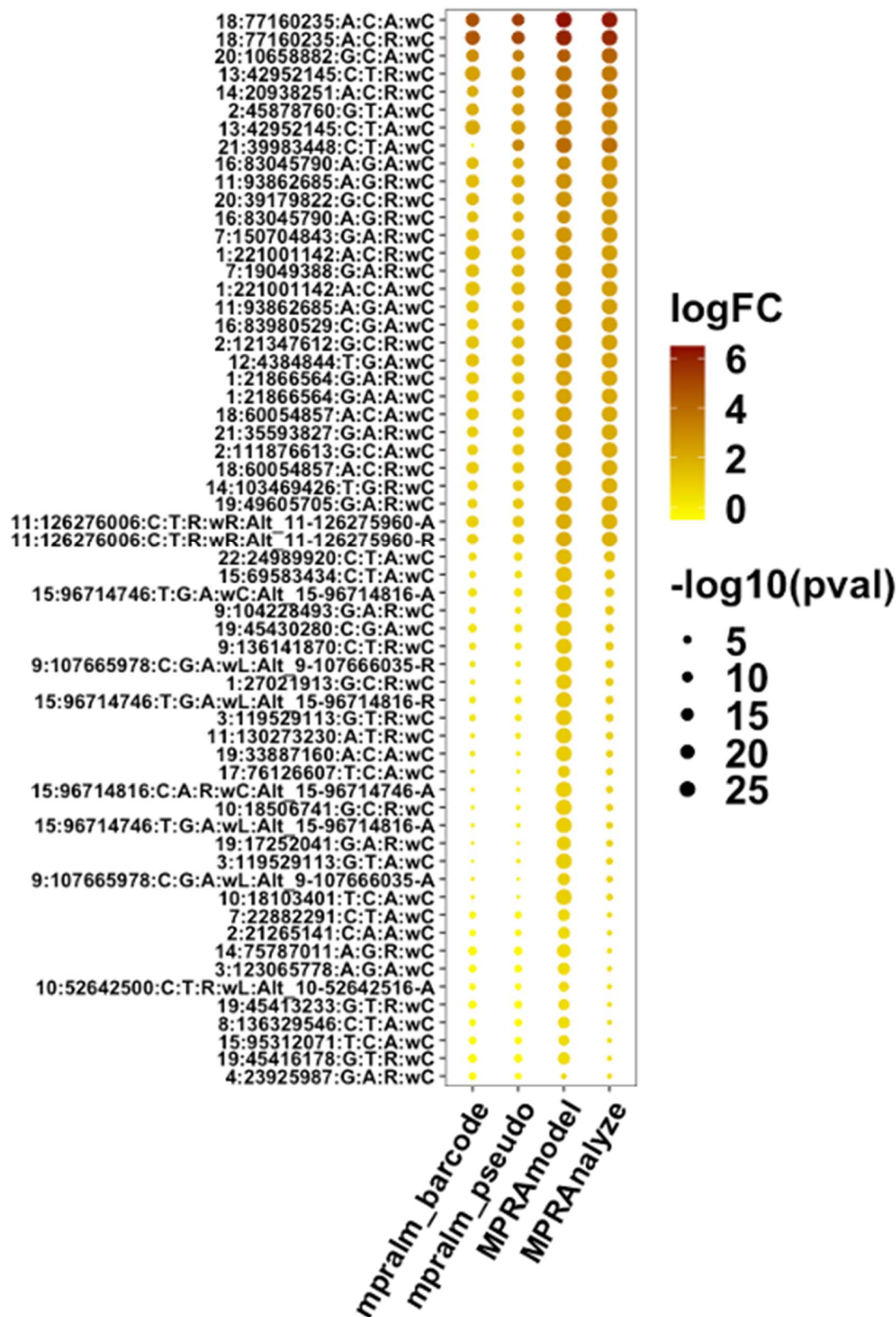


Figure 3.10. Breakdown of high-consensus log2FC variants in OL30-SK-N-SH for assessment of shared significance. Log2FC estimates and p-values were extracted from every model for each variant in OL30-SK-N-SH. A standard deviation of the log2FC values was calculated to select those variants with < 0.1 SD across the models. 10 variants were then selected from each of six expression groups (0.1, 0.5, 1, 1.5, 2, and 3 log2FC). These were plotted as a heatmap, and their size adjusted according to the negative logarithmic transformation of their p-values.

for more finely tuned expression estimates.

To see how well-correlated p-values are for all variants, a linear fit of these data was conducted using MPRAmodel as the basis for comparison (**Figure 3.11A**). It is again observed that read depth is the defining modifier of correlation between the models, specifically between mpralm-pseudo and MPRAmodel; the ~ 1.5 -fold increase in reads from OL30-SK-N-SH to OL28-GM12878 produces a correspondent improvement to correlation of 0.1 to 0.61, respectively. When plotting correlation of p-values as a factor of total barcodes and reads, we observe a direct relationship between total reads and improving correlation in mpralm-pseudo ($R = 0.32$; **Figure 3.11B**). There seems to be an inverse relationship of correlation with total numbers of barcodes for both mpralm-pseudo ($R = -0.45$) and mpralm-barcode ($R = -0.24$), though it is difficult to definitively state the nature of this relationship without understanding the barcode:read ratio, as average read depth is likely the more potent driver of this trend. MPRAalyze remains highly correlated with MPRAmodel regardless of total reads or barcodes within a data set, though this is not surprising given the strong concordance of their log2FC measurements and the similarity in their base modeling (negative binomial distributions informed by negative controls).

In addition to quantitative analyses, each of the models is capable of performing comparative analysis according to a designed factor, whether that be allele, cell type, or a

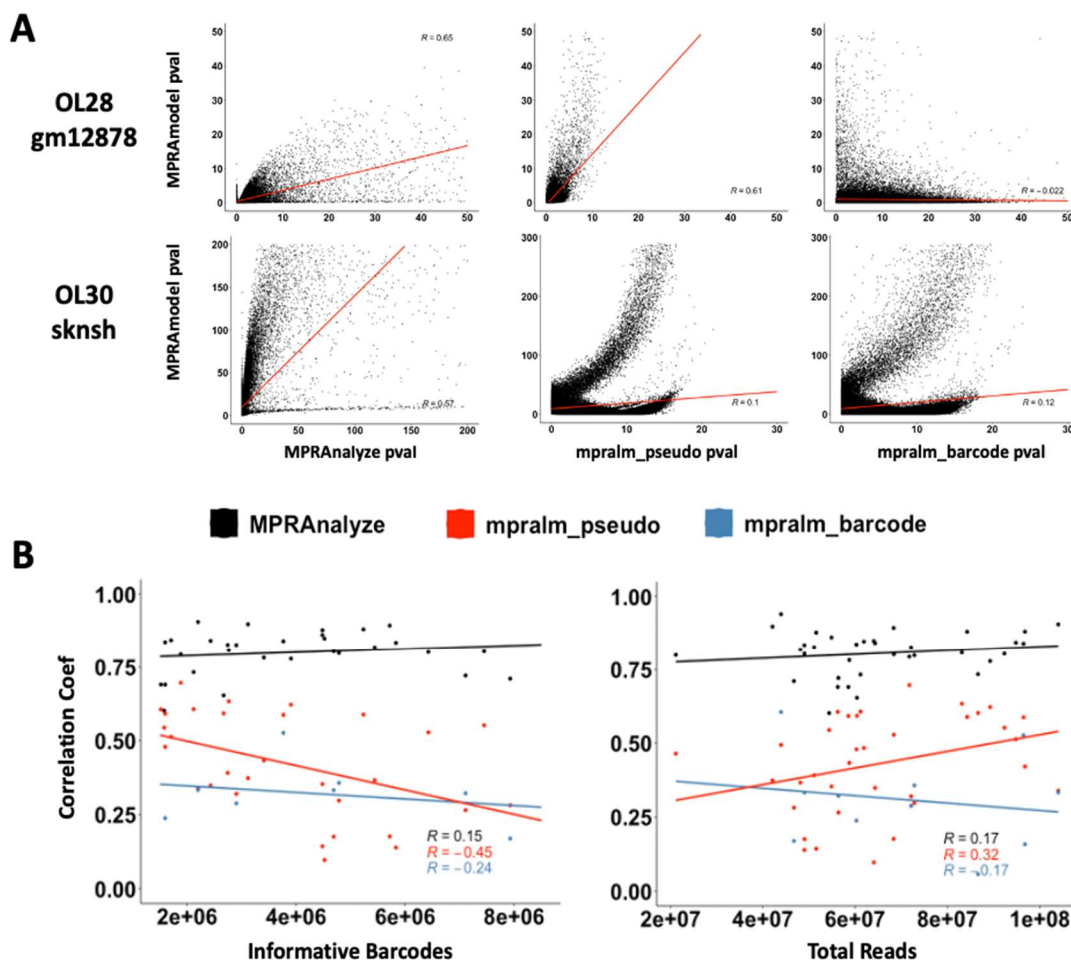


Figure 3.11. Trends in p-value correlations between models and faceted by numbers of barcodes or reads. (A) Correlation of p-values for each model as compared to MPRAmodel for OL28-GM12878 and OL30-SK-N-SH. Red line indicates linear fit of data. (B) Plotting of linear relationship between p-value correlation and number of barcodes (left) or number of reads (right) for all UKBB data sets.

given treatment condition. To assess the general fitness of the tools for this, an allelic skew analysis was performed for the variants in the UKBB data sets for which there existed reference and alternate alleles. QuASAR-MPRA, a retrofitted version of QuASAR designed specifically for allelic skew analysis (18), is used here as an outlier for model fitting. It was used similarly in the original MPRAAnalyze and mpralm papers (16,17); QuASAR-MPRA, hereafter referred to as QuASAR, reduces DNA and RNA

reads to pseudo-counts not just across barcodes but across replicates too. Not only does this eliminate the contributions of these replicates towards detection of skew, but it enables replicates of greater read quantity to dominate these predictions. To circumvent this issue and the absence of normalization, QuASAR directly compares allelic ratios in the DNA and RNA. MPRAmodel offers two methods for significance testing of allelic skew, a standard Welch's t-test and the ratio-based modeling of DESeq2. Both are employed here, with the DESeq2 method serving as an updated variation of QuASAR, as both perform significance testing of allelic ratios in RNA against DNA. To represent variations of fitness under different read depths, and to maintain thematic consistency, the results from OL28-GM12878 and OL30-SK-N-SH are displayed (**Figure 3.12**).

A Q-Q plot of p-values in OL28-GM12878 shows that increased read depth produces a more conservative estimate of allelic skew in mpralm, QuASAR, and the Welch's t-test variation of MPRAmodel (**Figure 3.12A**). MPRAnalyze and the DESeq2 methodology of MPRAmodel both display a lower threshold for assignment of significance, regardless of the library that is assessed. For MPRAnalyze, this is likely a result of the independent modeling it performs for each allele, which enables a fine-tuning of expression estimates. In the case of DESeq2, the comparison of allelic ratios within replicates and their respective conditions means that any variation in expression is a direct result of differences in expression. By extending the principles of QuASAR across replicates and groups, DESeq2 can produce a much more precise estimate of skew. This is recapitulated in p-value distributions, where we observe a classical shape in DESeq2 but an enrichment of p-values near 1 in QuASAR (**Figure 3.12B**). This advantage is reduced relative to the other models in OL30-SK-N-SH, as the lower depth

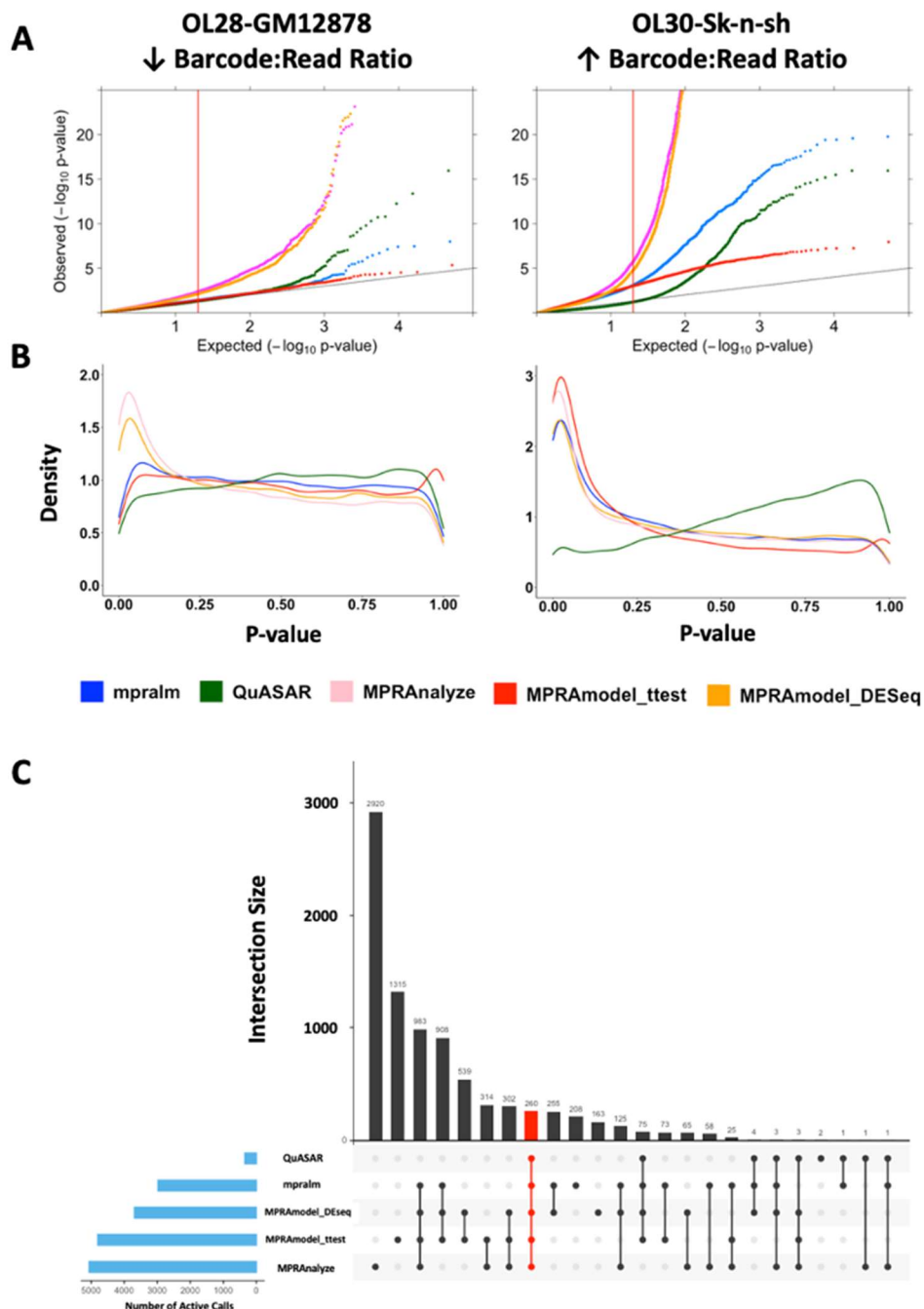


Figure 3.12. Evaluation of p-value calibrations and distributions in allelic skew analyses of OL28-GM12878 and OL30-SK-N-SH. (A) Q-Q plots and (B) density plots of p-values from each model for data sets OL28-GM12878 and OL30-SK-N-SH. The vertical red line in Q-Q plots denotes p-value = 0.05. (C) Upset plot of significant calls (BH-adjusted p-value < 0.1) for OL30-SK-N-SH broken down by model. Bar plot in bottom left corner displays the number of calls made per model. Main bar plot displays the number of significant calls overlapping for different pairings of models. The red-highlighted bar indicates shared calls across all models.

of the library appears to make ratio-based comparisons more volatile. When looking at number of significant calls in this library, we find QuASAR to be hyper-conservative in its estimate of ~300 skewed variants compared to the 3,000 + variants identified by the other models (**Figure 3.12C**). Only 963 skewed variants are shared between the mpralm, MPRAalyze, and different MPRAmodel protocols, with MPRAalyze uniquely identifying 2920 variants, or ~60% of its total calls. The lack of concordance between the various models underlines how differently each evaluates significance.

To assess whether there are broader similarities between skew estimates and their corresponding p-values, and whether these correlations match their shared methodologies (i.e. QuASAR and DESeq2, Welch's t-test and mpralm), each of these values was plotted against one another for all represented models (**Figure 3.13**). mpralm correlated best in both p-value ($R = 0.75$; **Figure 3.13A**) and log2FC ($R = 0.99$; **Figure 3.13B**) with the Welch's t-test methodology of MPRAmodel. These two are nearly identical in their assessment of skew, differing only in their base fold-change estimates for individual alleles, which may account for the discrepancies in their p-values. QuASAR displays the opposite tendencies with both mpralm and MPRAmodel, with poor correlation of skew estimates ($R = 0.41$ and $R = 0.39$, respectively) but fair correlation of p-values ($R = 0.74$ and $R = 0.52$, respectively). Interestingly, of these different models, QuASAR is least correlated with DESeq2 despite their similar methods of estimating skew. MPRAalyze is not especially well correlated with any other model as it pertains to skew or significance, though given its independent modeling of alleles it is less surprising to observe these differences. In fact, this may account for the large proportion of uniquely identified instances of skew.

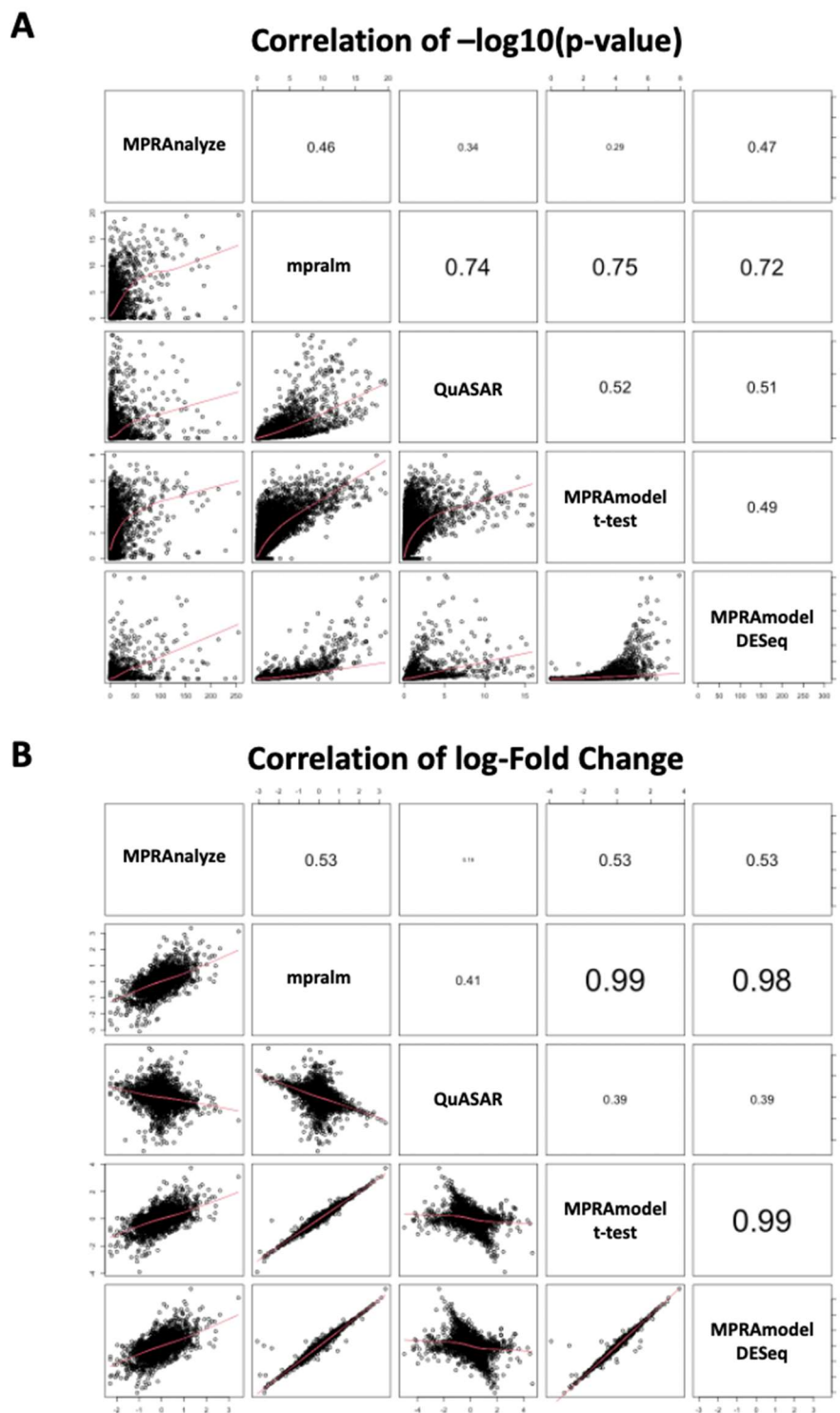


Figure 3.13. Correlation of p-values and log₂FC of allelic skew analyses for all tested MPRA models. Correlation matrices of $-\log_{10}(\text{p-values})$ (A) and log₂FC (B) of allelic skew analyses for all five MPRA models. Left-hand of matrices display the plots of their respective relationships. Right-hand of matrices display the correlation coefficient, faceted by size according to their strength of their correlation.

3.3. Identification and Correction of Errant Barcode Effects

Central to modeling of activity or skew is the elimination of confounding variables and their contributions to perceived expression. We have covered the utility of various normalization methods for controlling differences in the biological or technical noise of replicates, and have assessed the behavior of control populations of variants, a central facet of modeling for MPRAAnalyze and MPRAmodel. MPRA presents another oft-noted hurdle in that the barcodes used for massively parallel sequencing can also contain binding sites for transcriptional modifiers or RBPs (15,39). These have been shown to alter transcriptional activity for a given element or to alter the stability of the transcribed RNA, subsequently distorting its expression estimate or else increasing its dispersion and reducing significance (15,39). It is standard practice in MPRA design to create libraries with a high number of barcodes per element to ensure that any extreme behavior from a barcode does not appreciably shift average expression. However, this is unavoidable in a randomized barcode incorporation strategy, since one cannot control for the number of barcodes that will be associated to each element in the library. To this end, different groups have proposed tools for correcting barcode activity based on an assignment of effect size contributions to the aggregated expression estimate for a given element (15, 39). Here we evaluate one of these tools, MTSA, for its effects on the fitness of modeling in each MPRA analytical model. Rather than correct counts for all data sets, MTSA was performed for all transfected cells of OL28 and for all libraries transfected into SK-N-SH, as this permits an assessment of behavior within a library and cell subset.

MTSA works by evaluating the effect size contributions of individual barcodes to a given element; these effect sizes can be used to assess the finer control exerted by

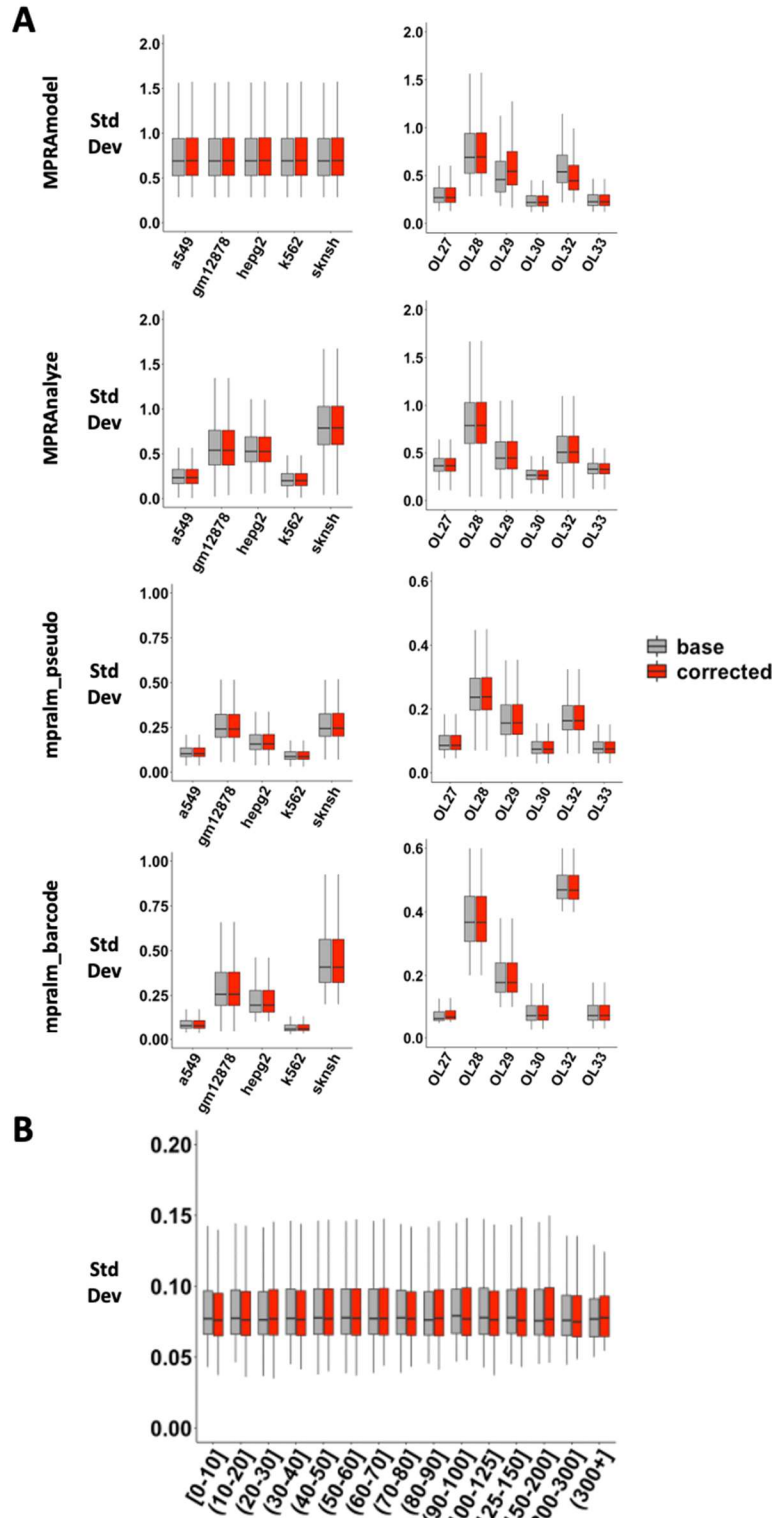


Figure 3.14.
Assessment of dispersion in log₂FC estimates for each model using either the base or MTSA-corrected counts as input. (A) Raw counts data from OL28 and all transfections of SK-N-SH were separately provided to MTSA for barcode correction. These corrected data were used as input to each of the assessed MPRA models and the standard deviation of their log₂FC estimates were extracted. A distribution of these values are sorted by model and for the transfected cells of OL28 (left) or the libraries of SK-N-SH (right). Distributions are colored according to their input, representing base (grey) or corrected counts (red). (B) Breakdown of standard deviation distributions across barcode bins for OL28-GM12878. Legend is the same as (A).

gapped k-mers across the barcodes of a data set, performing a 5-fold cross validation involving support vector regression (SVR) of scores (15). The expectation is that global dispersion in the data is reduced and that those barcodes requiring the greatest adjustment likely harbor regulatory motifs. To assess this first point, standard deviation of log2FC estimates were extracted from each model for all corrected data sets and their distributions compared against their original analyses (referred to here as ‘base’; **Figure 3.14**). We observe no significant changes to SD distributions for most of the models, except in the case of MPRAmodel, for which the distributions across SK-N-SH libraries can vary greatly (**Figure 3.14A**). While most libraries were unaffected, OL29 exhibited an expansion of dispersion and OL32 a reduction; no other model displayed these same trends despite utilizing the same data. The only notable difference in count normalization is the additional Bayesian shrinkage performed by MPRAmodel, so we can guess that a proportion of variants either acquired greater or lesser depth after barcode correction, creating correspondent changes to dispersion in OL29 and OL32, respectively. Given MTSA generates barcode effect sizes by element, it may be inappropriate to profile global SD distributions. In its place, we took OL28-GM12878 and binned variants according to their number of barcodes. When looking at dispersion across these bins, we observe a similarity in the base and corrected SD distributions but note a general trend of reduced median SD across nearly all groups (**Figure 3.14B**). Cumulatively, these trends suggest that MTSA is producing minute changes to counts data in most barcodes, which provides credence to the idea that the majority are inactive.

To assess the effects of MTSA-correction on model fitness and accuracy, the pseudo-count and barcode-based methodologies of mpralm were explored for OL28-

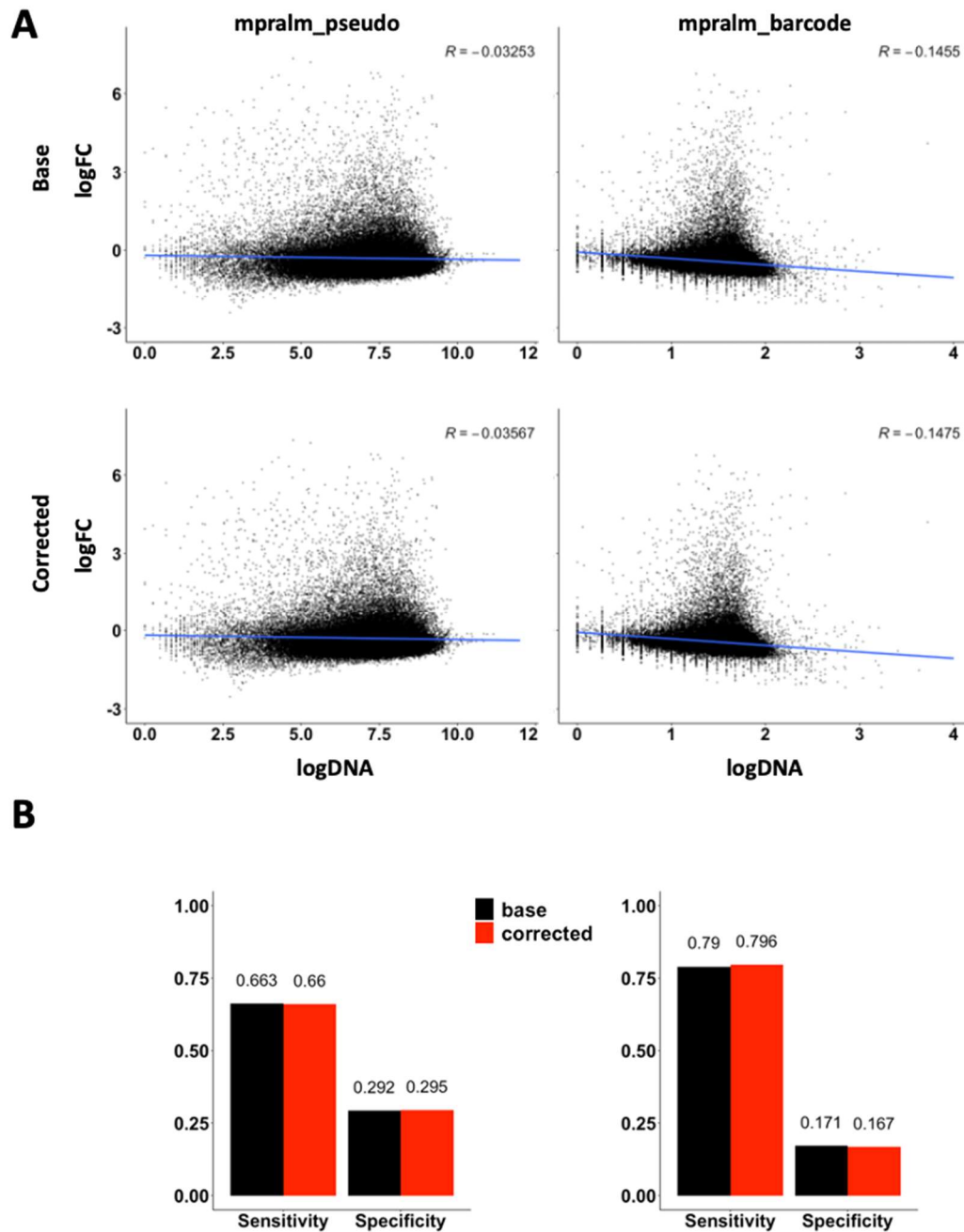


Figure 3.15. Breakdown of model fitness and accuracy for base and MTSA-corrected counts of OL28-A549. (A) Base and MTSA-corrected counts data of OL28-A549 were used as input for both pseudo-count and barcode-based modeling by mpralm. An approximation of model fitness is shown, plotting $\log_{10}$ of mean DNA counts against $\log_2 \text{FC}$ estimates. A linear fit and its corresponding correlation coefficient are displayed in each plot. (B) Assessment of sensitivity and specificity for mpralm models using the different counts data. Measurements of model accuracy were made using the UKBB-designed controls.

A549. This is because OL28-A549 possesses a relatively high barcode:read ratio and total read count. mpralm is being preferentially utilized for this because it is the only model capable of performing barcode-based analyses given the constraints of the UKBB libraries. It is also the only model that fits data according to a linear regression, making it simpler to visualize changes in fitness between the base and corrected data. We observe that model fitness is only slightly improved by the MTSA-correction in both mpralm methods (**Figure 3.15A**), and that the overall accuracy of the models is not significantly different between data inputs (**Figure 3.15B**). This is again suggestive of barcodes being labeled as largely inactive such that only a subset of variants is significantly altered by MTSA-correction.

To determine how this subset of barcodes is being altered and whether there is cell-type specific patterning to their effects, we extracted the top 100 barcodes from each cell line of OL28 according to their MTSA-predicted effect size. The original MTSA paper utilized multivariate linear regression models to estimate contributions to effect size of a variety of sequence factors. Through this they discovered that base frequency and RBPs were most predictive for effect size, and that together these factors could explain ~14-51% of variability (15). With this in mind, the list of barcodes produced by effect size filtering was mapped for RBP enrichment ($\log \text{PWM} > 6$ in at least three independent barcodes) using motifs from the CISBP-RNA database (32). This list of RBPs was then assessed for expression in the different cell lines using annotations from the Expression Atlas at EMBL-EBI (33). Dual heatmaps of barcode effect size and their corresponding RBP expression is displayed for OL28 in **Figure 3.16**. Most barcode-RBP pairs do not appear to be well correlated, though some display likely relationships, like

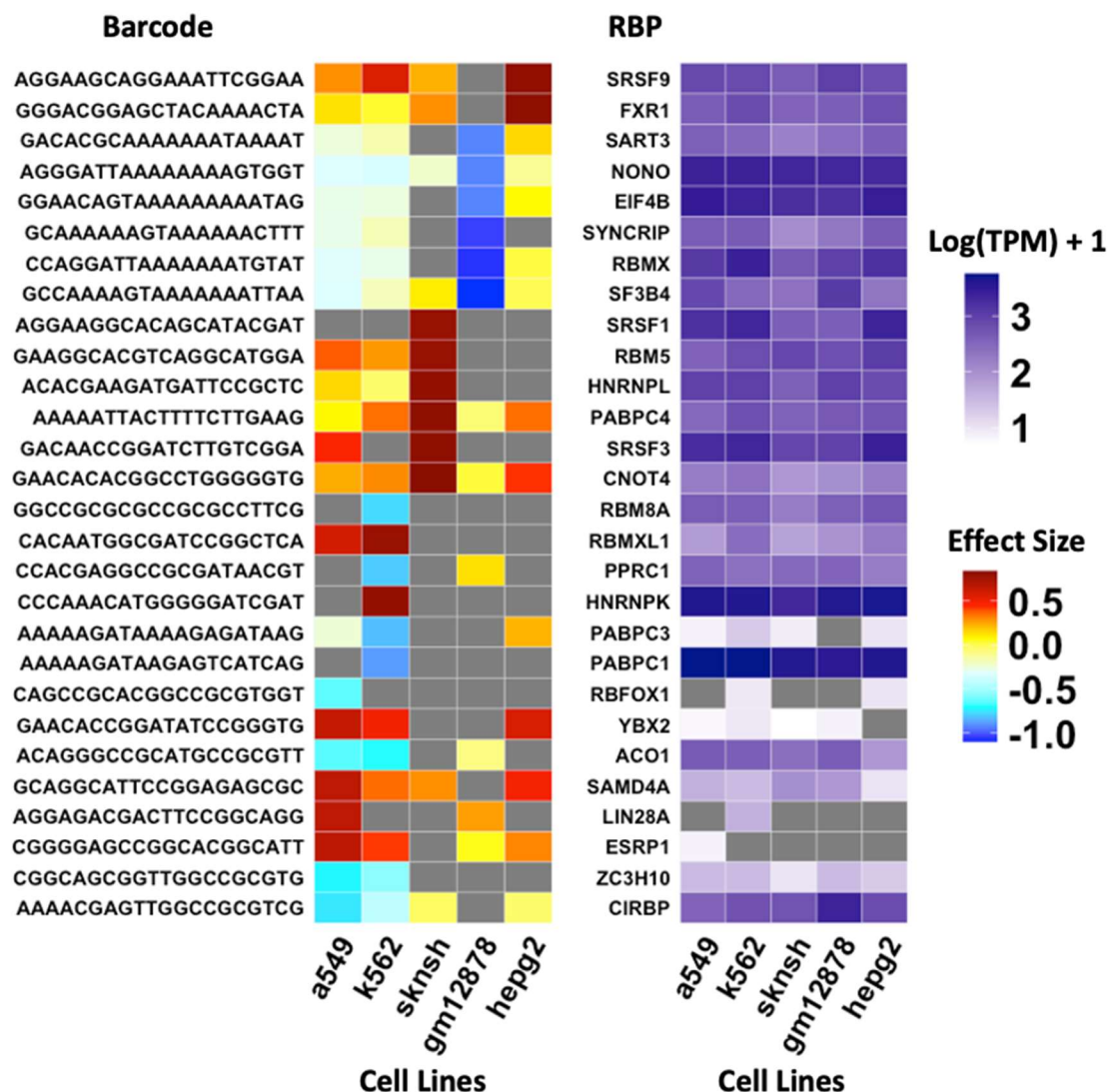


Figure 3.16. Heatmap of barcode effect sizes and the expression of mapped RBPs across all cell lines of OL28. Barcodes were selected from each library representing the top 100 effect sizes. These barcodes were used as input to the RNA scan tool of the CISBP-RNA database to identify sequences enriched for RBPs at a log PWM threshold of 6 (32). Expression of enriched RBPs was evaluated across the cell lines using RNA-seq expression profiles from previously published work (33) to the Expression Atlas at EMBL-EBI. Barcode-RBP pairs were created based on the highest PWM score associated with a given barcode. A heatmap of barcode effect sizes (left) and corresponding RBP expression (right) are displayed.

SF3b4. *SF3b4* encodes a subunit of the U2-type spliceosome important for pre-mRNA processing in eukaryotes (40). We observe a negative effect size in the barcode with

increasing expression of *SF3b4*; this is potentially indicative of higher frequency splicing events of the barcode's transcript from the binding of SF3B4. Altogether, these observations of MTSA portray a tool capable of capturing and mitigating the confounding effects of potent biological factors in barcode skew.

3.4. Description of Contributions

All research described herein was performed by the student. PIP scores for variants were provided to the student by Jacob Ulirsch, a PhD candidate from the Broad Institute whom contributed to prior research efforts with the UKBB/Japan Biobank and GTEx datasets.

Chapter 4: Discussion

Massively parallel reporter assays are a popular technique for evaluating the transcriptional contributions of a variety of *cis*-regulatory elements. These have come to include the functional characterization of distal regulatory elements, like enhancers (27,28,36,41), differential variant activity between cells and alleles (5,13,22), and more recent designs to assess the effects of chromatin accessibility in a randomized (7) and specific manner (8). This expansion in MPRA design has made it difficult to develop a unified framework for its analysis, leading to an *ad hoc* strategy tailored to lab-specific experimental designs. Within the last few years, different MPRA analytical models have been proposed to address multivariate designs (16,17) or for specific assessment of allelic skew (18,19). These vary greatly in the way they model data and determine activity, providing MPRA users with an assortment of analytical paradigms. Some of these pipelines, like MPRAmodel, even incorporate analytical redundancies, as MPRAmodel assesses allelic skew through both a Welch's t-test of activity-enriched variants and through the ratio-based summary statistics of DESeq2. This variety can make it difficult for MPRA users to decide which pipelines may best suit their library design or whether an *ad hoc* strategy is a better alternative.

The objective of this work is to profile these pipelines within a set of MPRA libraries less commonly utilized and accommodated by their designs. The original publications for many of these tools were primarily concerned with assessing libraries of even barcode assignment across tested elements, as this makes modeling of data matrices much simpler. The data assessed here is not only different in its use of degenerate barcode assignment but also in the average size of its libraries. To our knowledge, this is

the first concentrated effort to understand the fitness of different MPRA analytical models for libraries of this design and across this range of quality. The fact that the seven UKBB-GWAS libraries were transfected within five distinct cell types also presents an opportunity to evaluate behavior of shared elements across a range of conditions. The overall project can be divided into two primary areas. The first is a dissection of common issues with library design and analysis, including normalization methods for counts data, behavior of controls, and the application of barcode correction. The latter of these is especially important, as most barcode-based analyses are underpowered compared to their aggregated counterparts. The second phase is a dissection of different analytical tools for quantitative and comparative assessments as well as an evaluation of model priorities and accuracy. The benchmarking effort performed here may serve as a guide for deciding which of these models is best suited to a given MPRA design and how additional data preparation can improve the power of analysis.

4.1. Common Considerations for Data Preparation and Design

Normalization of RNA-seq data presents a unique challenge in MPRA. The standard methods of RPKM or TPM are unfit for this type of data, as the elements in MPRA are all the same length. As such, application of these methods is little different from a library size scaling of counts by replicate or condition, which can often dilute genuine differences in global transcription between samples (35,37). Most MPRA studies have utilized library size scaling for normalization (3-5,21,27,28), a method offered by both MPRAnalyze and mpralm in their modeling framework (16,17). However, there are alternatives to these in the form of UQ, RLE, and TMM normalization methods that have proven to be situationally appropriate in different RNA-seq studies, and in fact

MPRAnalyze offers UQ and RLE as options instead of library size scaling. We find that library size scaling produces the poorest model accuracy when evaluating sensitivity and specificity of controls in the UKBB data. This result is dependent on data quality, as libraries with less overall read depth displayed no difference in their model accuracy regardless of normalization method. In general, RLE and TMM performed the best in producing an overall better ratio of sensitivity to specificity, though UQ was not far behind. It should be noted that RLE and TMM methods are designed with a null hypothesis of no expression, meaning they are only appropriate for MPRA libraries in which the bulk of elements are unlikely to be active. For MPRA designed to assess direct transcriptional regulators, like promoters, enhancers, or transcription factor configurations, we advise the use of UQ normalization.

An important facet of normalization to MPRA is the reduction of dispersion in expression estimates. We find that standard error of these estimates behaves differently depending on the analytical model evaluated. Dispersion appears to be inherently less by element in `mpralm` as compared to MPRAnalyze, a likely result of its more straightforward measure of regulatory activity as a log-ratio of RNA-to-DNA. MPRAnalyze utilizes a log-additive fold-change estimate, termed α , which is adjusted according to the independent modeling of other technical and biological factors. This creates a more nuanced adjustment of counts post-normalization that contributes to greater overall variance between replicates. In a direct comparison to MPRAmodel, we find that none of the normalization methods is better at mitigating dispersion in low quality variants. We attribute this to MPRAmodel's regression of fold-change estimates to the mode of its distribution. While this is an artificial reduction to the dispersion of

fold-change, limiting this source of variance can be important for increasing the power of your analysis.

This relationship between reduced dispersion and improved model fitting is remarked upon in the original `mpralm` publication, where they show that improving sample correlations significantly improved the power of their analysis to detect differences (17). It is important in reducing dispersion to normalize according to the population of elements most likely to represent the average in each replicate. We see this manifest as an issue in library size scaling, which produced the poorest fit among the different normalization methods for both `mpralm` and `MPRAnalyze` despite performing best in reducing dispersion. TMM and RLE normalize according to either the second and third quantiles of data or by the median of count ratios, thus better representing the standard counts of a replicate. We observe that these two methods improved model fitting the best, with UQ a close second. UQ appeared to undervalue expression relative to RLE and TMM, which we suspect is a byproduct of its quantile threshold. The method can be susceptible to reducing global expression differences between replicates or samples in that it adjusts counts according to the upper 75% of expressing elements. Lowering the threshold may be appropriate in instances where differential expression is expected, to better represent the median of your data.

Alternatives to standard normalization methods have been proposed in the use of spike-ins (38) or via RUV (35), though these are traditionally used to mitigate batch effects or other latent covariates. However, spike-ins do offer a way of normalizing by element rather than scaling by sample-to-sample variations in read depth. To perform such a method in MPRA requires the establishment of a set of controls sequences for

which performance can be reasonably predicted. To this end, we evaluated the behavior of controls in the UKBB library to assess the method of control assignment through location-matching of variants that are null-GWAS. When looking at the controls across cell lines and libraries, we find that while they behaved largely the same across libraries, there was much less consensus across cell type; this trend was more prominent in the positive controls, with a subset displaying universally limited expression or else limited expression in A549 and GM12878. The consensus expression exhibited across libraries is promising, however, indicating that the control sequences overcome any technical noise introduced by library construction (variable barcode representation) or experimentation (transfection efficiencies). It is suggested to match controls specifically to your cell lines of interest, especially when considering the use of MPRAanalyze or MPRAmodel, as both use controls in the modeling of activity. In the event of variable performance of controls across cells, you introduce a biological bias that severely limits the capacity of these models to make cell-to-cell comparisons

Importantly, we also observe variable control behavior in different data constructs. An increase in the average number of barcodes per element tended to improve sensitivity at the expense of specificity. This is not, however, tied to read depth, as increases in coverage tended to improve sensitivity without altering specificity. These observations suggest that with sufficiently high barcode representation, read depth become the more potent modifier of model accuracy. Lastly, we established an alternative set of controls based on whether a variant mapped within a CRE annotation (UTRs, promoters, or enhancers). We find that almost all of these positive controls were predicted to have expression > 2 fold-change in every analytical model. While not all

were significant at this expression level, it represents a pool of elements which can be further filtered by cell-type specific genomic annotations to yield a set of sequences appropriate for sequence level normalization.

Normalization of barcodes according to potential activity is an increasingly recognized step in reducing biological bias. Multiple tools have been designed to address barcode biases in MPRA data, including MRPAdecoder (39) and MTSA (15). Both tools prepare raw MPRA data for analysis by normalizing RNA counts according to the abundance of barcodes in the plasmid DNA sample. MTSA was chosen for this work, as it permits a finer resolution of sequence effect through a five-fold cross-validation and training of gapped k-mers. This provides a more unbiased assessment of sequence contributions as the predicted effect for a barcode is produced via an aggregation of k-mers within its sequence. An expectation of this kind of correction would be a reduction to variance for individual elements. Correspondingly, we find little difference in dispersion when looking at different data sets and models in a global fashion. Instead, we find that MTSA-correction reduces the median of dispersion across most barcode bins, though the overall range of dispersion is maintained. This is expected since most barcodes do not alter transcriptional profiles. For more accurate modeling of effect, we find that filtering the initial MTSA training to barcodes present across all replicates is essential, otherwise you introduce technical bias into the modeling of sequence effect. MTSA will combine barcode counts across replicates for training, so this additional step ensures equal representation of barcodes for a given element. Barcodes not included in the training can be retroactively corrected using the predicted effect sizes from the k-mers.

Given the direct linear modeling of mpralm and its capacity to model barcode-based data, we chose to assess model accuracy and fitness using MTSA-corrected data. mpralm modeling was only slightly improved by barcode correction in both barcode and pseudo-count methods; this did not translate to any significant changes to sensitivity or specificity measurements. Together, these observations underscore that the majority of barcodes do not contribute to biological bias and that barcode correction is, first and foremost, for identifying active sequences. We also find that the barcode-based analyses are consistently underpowered compared to pseudo-count methods, as previously noted in other works (17), and that the addition of barcode bias correction does not improve these analyses enough to recommend over other methods.

Given that MTSA models sequence effects from expression data, there can be any number of sequence features contributing to effect, including RBPs, base frequencies, TFs, or secondary structures. Their original work found that ~14-51% of effect could be predicted by a combination of RBPs and base frequencies (15). When looking at RBP-barcode pairings of barcodes with greater sequence effect, we find that there are some pairings, like *SF3b4*, that display a potential functional relationship. Just because others do not display functional relationships does not mean they don't exist, but rather that the relationship could be caused by other transcriptional or post-transcriptional modifiers. Additionally, without knowing the exact profile of plasmids in the post-transduced sample, it can be difficult to assess replicate biases, where a given barcode may be more highly represented in a replicate relative to other barcodes for the same element. This can create an artificial bias to MTSA predictions, and so we cannot confidently recommend use of MTSA in data sets where DNA and RNA are sequenced from different

transduction pools. This includes the data sets tested here. There are other potential applications for this method towards MPRA of this design, which will be touched on further in the discussion.

4.2. MPRA Analytical Models Show Suitability for Different Data Constructs

In this work, we were able to profile MPRAalyze, mpralm, MPRAmodel, and QuASAR-MPRA for different types of analyses, utilizing the range of data quality inherent to the UKBB data sets to see how performance varies given different barcode and read depths. A summation of their base characteristics is displayed in **Table 4.1**.

Previous works have noted that MPRA with fewer than four replicates consistently

Table 4.1. Summary of modeling characteristics for each tested pipeline.

Characteristic	MPRAmodel	MPRAalyze	mpralm	QuASAR-MPRA
Quantitative	+	+	+	-
Comparative	+	+	+	+
Normalization Method	DESeq (“median of ratios”)	UQ (default), RLE, sequencing depth	Sequencing depth	None
Type of Model	GLM	GLM	LM	Beta-binomial
Significance Testing	Welsh’s T-test or DESeq2 (ratio-based summary statistics)	Likelihood ratio testing or Welsh’s T-test	Welsh’s T-test	Ratio-based summary statistics
Barcode Handling (pseudo-count/barcode-based)	+/-	+/+	+/+	+/-

produce underpowered assessments of regulatory activity (17), so the standard five replicates employed across the UKBB libraries ensure a higher quality assessment of model performance. All quantitative analyses, or measurements of activity, were performed using MPRAalyze, mpralm, and MPRAmodel, with mpralm utilized for both pseudo-count and barcode-based methodologies. Each of these tools employs a generalized linear model (GLM) for fitting data except for mpralm, which uses log-transformed expression ratios and plasmid counts to produce a linear regression model (LM). These two types of modeling tend to correspond well with one another, though linear models are noted to maintain control of type 1 error rates more reliably in situations where no differential count is expected (42). GLMs are typically recommended over LMs when type 1 error rate can be accounted for in your data. We have shown that the controls of the UKBB libraries behave appropriately and feel confident a direct comparison of MPRA models is appropriate.

When assessing model accuracy across the spectrum of UKBB libraries, we find that mpralm presents the most severe trade-off of sensitivity to specificity, generally decreasing in specificity with the more barcodes associated to a data set. The UKBB data sets appear to show a mildly inverted relationship between barcode number and total read count, which may account for this observation in mpralm. MPRAalyze and MPRAmodel only seem to improve in sensitivity with increasing total reads and are largely agnostic to differences in barcode numbers. This likely has to do with their foundational modeling, as both incorporate negative controls and so produce a consistently high specificity. MPRAalyze specifically builds its model from negative

controls and appears to perform the same regardless of data quality, making it an attractive option for MPRA data with poor coverage or barcode multiplicity.

Looking at expression estimates and significance testing, we find that MPRAmodel and MPRAanalyze correlate best, a likely byproduct of their shared modeling of negative controls. Despite these correlations, MPRAanalyze produces ~4-fold fewer significant calls than MPRAmodel, a likely byproduct of differences in p-value calibration. There is also the fact that MPRAanalyze performs independent modeling of other contributing factors to expression (replicates, batches, barcodes, etc.), which may similarly contribute to this discrepancy in variant calls. Indeed, when looking at expression estimates as a driver of significance, we find that only MPRAmodel and mpralm are primarily driven by this factor. For this reason, we recommend the use of MPRAanalyze when there is concern of confounding technical factors or if you wish to forcibly model a biological factor; while it will be a more conservative estimate of activity, the resulting calls will possess better overall properties. While we would generally recommend MPRAanalyze and MPRAmodel, this is contingent on the inclusion of quality negative controls to your MPRA design.

The most variable model performance was from mpralm, though the pseudo-count analyses was a consistently more powerful model than its barcode counterpart. Correlation in fold-change and p-values improved with increasing read depth across replicates. Interestingly, there was no discernible improvement to barcode-based modeling with increasing numbers of barcodes. The most notable improvements to mpralm modeling were provided by increased replicate correlation, underscoring the importance of mitigating RNA quantity biases between replicates. This also helps to

explain why mpralm utilizes a base library scale normalization over other more faithful normalization methods, as it tends to improve replicate-to-replicate correlations.

We also assessed the qualities of comparative analysis in the different models, including the DESeq and Welch's t-test methods of MPRAmodel as well as QuASAR-MPRA. Most of the modes of analysis follow their respective GLM or LM models, except QuASAR-MPRA and DESeq2, which both employ ratio-based summary statistics. QuASAR-MPRA was the most conservative of the models and diverged greatest in methodology, producing pseudo-counts across barcodes and replicates for different alleles. The reduction of data complexity severely limited the power of its analysis and so we do not recommend its use unless options for retaining replicates is added to the tool. MPRAalyze and DESeq2 proved to be the most sensitive of the tested models, but use of DESeq2 requires assumptions that may not necessarily hold for the UKBB data sets. Namely, DESeq2 directly compares allelic ratios in the RNA against those in the DNA, performing this by replicate. The UKBB data sets do not possess replicate-matched counts, so these are not exact comparisons. In addition, the DNA counts were sequenced pre-transduction, so the pool of plasmids transfected into a cell may comprise a different representative ratio than that observed in the originally sequenced pool. For similarly designed MPRA experiments, we would suggest the use of the base t-test analysis if MPRAmodel is your desired pipeline. This t-test methodology and mpralm are nearly perfectly matched in their fold-change estimates and significance, and so would offer a redundancy to allelic skew analyses that might help to identify an enriched pool of differentially expressed alleles. Lastly, we find that there is an inverse relationship between total read counts and relative sensitivity in the different models,

except for DESeq2 and MPRAalyze. This inverse relationship could be a product of the increased noise inherent to less informative data, permitting a higher proportion of false positives in the skew analysis.

As final recommendations, we make two suggestions for general MPRA data preparation that is beneficial to any library design or analytical model. First, we've found that location-matched controls do not behave the same across cell types and propose that choosing variants or elements based upon genomic annotations of specific cells is more appropriate. Additionally, the use of PIP scores proved a powerful tool for predicting general effect in MPRA. We recommend performing a variation of PIP scoring that is specific to the designed elements of a given MPRA library, whether the criteria be DHS sensitivity, histone modifications, ChIP-seq of transcriptional modifiers, DNA methylation patterns, or some other annotation type. Quality controls are generally important, but for MPRAmodel and MPRAalyze, which rely on controls for modeling, they are essential. Second, we propose the use of MTSA as a tool for filtering rather than adjusting barcodes. Without post-transduction RNA and DNA libraries as input, there is the potential for MTSA to interpret sampling bias as barcode bias and to generally underestimate true bias. We propose utilizing MTSA to identify barcodes with high individual effect sizes, as these are most likely to represent true bias, and to remove this pool of barcodes from analysis. While this does not involve a complete barcode correction of data, it removes problematic sequences from contributing to expression estimates, thereby improving the power of the model.

We believe this work can serve as a guide for assessing which MPRA analytical model is appropriate for a given data set and can inform on best practices in the event that *ad hoc* methodologies are required.

4.3. Description of Contributions

All writing and formatting, including construction of the summary table, was performed by the student.

Chapter 5: Bibliography

- 1 ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, 306(5696), 636–640.
- 2 The ENCODE Project Consortium, Snyder, M. P., Gingeras, T. R., Moore, J. E., Weng, Z., Gerstein, M. B., Ren, B., Hardison, R. C., Stamatoyannopoulos, J. A., Graveley, B. R., Feingold, E. A., Pazin, M. J., Pagan, M., Gilchrist, D. A., Hitz, B. C., Cherry, M., Bernstein, B. E., Mendenhall, E. M., Zerbino, D. R., Frankish, A., Flicek, P., & Myers, R. M. (2020). Perspectives on ENCODE. *Nature*, 583(7818), 693–698.
- 3 Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., Ahituv, N., & Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research*, 27(1), 38–52.
- 4 Davis, J. E., Insigne, K. D., Jones, E. M., Hastings, Q. A., Boldridge, W. C., & Kosuri, S. (2020). Dissection of c-AMP Response Element Architecture by Using Genomic and Episomal Massively Parallel Reporter Assays. *Cell Systems*, 11(1), 75–85.
- 5 Ray, J. P., de Boer, C. G., Fulco, C. P., Lareau, C. A., Kanai, M., Ulirsch, J. C., Tewhey, R., Ludwig, L. S., Reilly, S. K., Bergman, D. T., Engreitz, J. M., Issner, R., Finucane, H. K., Lander, E. S., Regev, A., & Hacohen, N. (2020). Prioritizing disease and trait causal variants at the TNFAIP3 locus using functional and genomic features. *Nature Communications 2020 11:1*, 11(1), 1–13.
- 6 Kreimer, A., Zeng, H., Edwards, M. D., Guo, Y., Tian, K., Shin, S., Welch, R., Wainberg, M., Mohan, R., Sinnott-Armstrong, N. A., Li, Y., Eraslan, G., Amin, T. bin, Tewhey, R., Sabeti, P. C., Goke, J., Mueller, N. S., Kellis, M., Kundaje, A., ... Yosef, N. (2017). Predicting gene expression in massively parallel reporter assays: A comparative study. *Human Mutation*, 38(9), 1240–1250.
- 7 Maricque, B. B., Dougherty, J. D., & Cohen, B. A. (2017). A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Research*, 45(4), e16–e16.
- 8 Maricque, B. B., Chaudhari, H. G., & Cohen, B. A. (2018). A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nature Biotechnology 2018 37:1*, 37(1), 90–95.
- 9 King, D. M., Hong, C. K. Y., Shepherdson, J. L., Granas, D. M., Maricque, B. B., & Cohen, B. A. (2020). Synthetic and genomic regulatory elements reveal aspects of Cis-regulatory grammar in mouse embryonic stem cells. *ELife*, 9, e41279.
- 10 Rabani, M., Pieper, L., Chew, G. L., & Schier, A. F. (2017). A Massively Parallel Reporter Assay of 3' UTR Sequences Identifies In Vivo Rules for mRNA Degradation. *Molecular Cell*, 68(6), 1083–1094.
- 11 Vainberg Slutskin, I., Weingarten-Gabbay, S., Nir, R., Weinberger, A., & Segal, E. (2018). Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. *Nature Communications 2018 9:1*, 9(1), 1–18.
- 12 Klein, J. C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., & Shendure, J. (2020). A systematic evaluation of the design and context dependencies

- of massively parallel reporter assays. *Nature Methods* 2020 17:11, 17(11), 1083–1091.
- 13 Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T. S., & Sankaran, V. G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell*, 165(6), 1530–1545.
 - 14 Shen, S. Q., Myers, C. A., Hughes, A. E. O., Byrne, L. C., Flannery, J. G., & Corbo, J. C. (2016). Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Research*, 26(2), 238–255.
 - 15 Lee, D., Kapoor, A., Lee, C., Mudgett, M., Beer, M. A., & Chakravarti, A. (2021). Sequence-based correction of barcode bias in massively parallel reporter assays. *Genome Research*, 31(9), 1638–1645.
 - 16 Ashuach, T., Fischer, D. S., Kreimer, A., Ahituv, N., Theis, F. J., & Yosef, N. (2019). MPRAnalyze: Statistical framework for massively parallel reporter assays. *Genome Biology*, 20(1), 1–17.
 - 17 Myint, L., Avramopoulos, D. G., Goff, L. A., & Hansen, K. D. (2019). Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics*, 20(1), 1–19.
 - 18 Kalita, C. A., Moyerbrailean, G. A., Brown, C., Wen, X., Luca, F., & Pique-Regi, R. (2018). QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics*, 34(5), 787–794.
 - 19 Niroula, A., Ajore, R., & Nilsson, B. (2019). MPRAScore: robust and non-parametric analysis of massively parallel reporter assays. *Bioinformatics*, 35(24), 5351–5353.
 - 20 Georgakopoulos-Soares, I., Jain, N., Gray, J. M., & Hemberg, M. (2017). MPRAator: a web-based tool for the design of massively parallel reporter assay experiments. *Bioinformatics*, 33(1), 137–138.
 - 21 Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, 165(6), 1519–1529.
 - 22 Khetan, S., Kales, S., Kursawe, R., Jillette, A., Ulirsch, J. C., Reilly, S. K., Ucar, D., Tewhey, R., & Stitzel, M. L. (2021). Functional characterization of T2D-associated SNP effects on baseline and ER stress-responsive β cell transcriptional activation. *Nature Communications* 2021 12:1, 12(1), 1–13.
 - 23 Ghazi, A. R., Kong, X., Chen, E. S., Edelstein, L. C., & Shaw, C. A. (2020). Bayesian modelling of high-throughput sequencing assays with malacoda. *PLOS Computational Biology*, 16(7), 1–18.
 - 24 Abbas-Aghababazadeh, F., Li, Q., & Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLOS ONE*, 13(10), e0206312.
 - 25 Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550.
 - 26 Melnikov, A., Zhang, X., Rogov, P., Wang, L., & Mikkelsen, T. S. (2014). Massively Parallel Reporter Assays in Cultured Mammalian Cells. *Journal of Visualized Experiments : JoVE*, 90, 51719.

- 27 Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Kinney, J. B., Kellis, M., Lander, E. S., & Mikkelsen, T. S. (2012). Rapid dissection and model-based optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3), 271-277.
- 28 Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S., & Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23(5), 800-811.
- 29 Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), 1-17.
- 30 Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203-209.
- 31 Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139.
- 32 Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecnas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172-177.
- 33 Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 489(7414), 101-108.
- 34 Liu, D., & Graber, J. H. (2006). Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. *BMC Bioinformatics*, 7, 77.
- 35 Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* 2014 32:9, 32(9), 896-902.
- 36 Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3), 159-164.
- 37 Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1), 1-13.
- 38 Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., & Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9), 1543-1551.
- 39 Letiagina, A. E., Omelina, E. S., Ivankin, A. v., & Pindyurin, A. v. (2021). MPRAdecoder: Processing of the Raw MPRA Data With a priori Unknown

- Sequences of the Region of Interest and Associated Barcodes. *Frontiers in Genetics*, 12, 618189.
- 40 Xiong, F., & Li, S. (2020). SF3b4: A Versatile Player in Eukaryotic Cells. *Frontiers in Cell and Developmental Biology*, 8, 14.
 - 41 Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., & Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, 24(10), 1595–1602.
 - 42 Warton, D. I., Lyons, M., Stoklosa, J., & Ives, A. R. (2016). Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, 7(8), 882–890.
 - 43 Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S. I., Cooper, G. M., Ahituv, N., Pennacchio, L. A., & Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, 30(3), 265-270.
 - 44 Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., Kubo, M., Shiono, M., ... Yanai, H. (2017). Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, 27(3 Suppl), S2-S8.