Published in final edited form as:

Genet Epidemiol. 2020 October ; 44(7): 785–794. doi:10.1002/gepi.22337.

Statistical considerations for the analysis of massively parallel reporter assays data

Dandi Qiao¹, Corwin Zigler², Michael H. Cho^{1,3}, Edwin K. Silverman^{1,3}, Xiaobo Zhou¹, Peter J. Castaldi^{1,4,*}, Nan H. Laird^{5,*}

¹Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

²Department of Statistics and Data Sciences, Department of Women's Health, University of Texas at Austin, Austin, TX, USA

³Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

⁴Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

⁵Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Abstract

Non-coding DNA contains gene regulatory elements that alter gene expression, and the function of these elements can be modified by genetic variation. Massively parallel reporter assays (MPRA) enable high-throughput identification and characterization of functional genetic variants, but the statistical methods to identify allelic effects in MPRA data have not been fully developed. In this work, we demonstrate how the baseline allelic imbalance in MPRA libraries can produce biased results, and we propose a novel, non-parametric, adaptive testing method that is robust to this bias. We compare the performance of this method with other commonly used methods, and we demonstrate that our novel adaptive method controls type I error in a wide range of scenarios while maintaining excellent power. We have implemented these tests along with routines for simulating MPRA data in the Analysis Tools for MPRA (@MPRA), an R package for the design and analyses of MPRA experiments. It is publicly available at <http://github.com/redaq/atMPRA>.

Keywords

MPRA; functional variants; gene expression; reporter assays

*These authors have made equal contributions to the work.

Data Availability Statement

The R package @MPRA is publicly available at <http://github.com/redaq/atMPRA>. The MPRA data that support the findings of this study are publicly available in the NCBI Gene Expression Omnibus (GEO) at <http://www.ncbi.nlm.nih.gov/geo/>, reference number GSE70531[Ulirsch and Sankaran 2015] and GSE109452[Zhou, et al. 2018].

Introduction

Most functional genetic variants associated with complex genetic diseases are located in noncoding regions and likely influence disease risk by altering gene expression levels [Cookson, et al. 2009; Edwards, et al. 2013; Hindorff, et al. 2009; Maurano, et al. 2012]. Given the rapidly expanding number of GWAS-identified loci, screening strategies are needed to identify the functional variants in these regions, and massively parallel reporter assays (MPRA) are one of the more promising and widely used methods for this purpose [Inoue and Ahituv 2015; Melnikov, et al. 2012; Shen, et al. 2016; Ulirsch, et al. 2016; White 2015].

MPRA leverages chip-based DNA synthesis to produce hundreds of thousands of synthetic DNA sequences. MPRA was initially described by Melnikov et al. [Melnikov, et al. 2012] in which the Mann-Whitney U test (MW) was used to quantify MPRA transcriptional effects by comparing DNA/RNA ratios. Subsequent publications have used a variety of other statistical approaches, some of which are designed specifically for MPRA analysis [Kalita, et al. 2018; Myint, et al. 2019; Shen, et al. 2016].

In this paper, we focus on statistical issues related to the use of MPRA to identify *allelic transcriptional effects* for single nucleotide polymorphisms (SNPs) by comparing pairs of DNA sequences that vary only by a single base-pair [Castaldi, et al. 2019; Myint, et al. 2019; Tewhey, et al. 2016; Ulirsch, et al. 2016]. However, the methods described here can be applied to compare MPRA data between any two conditions. We demonstrate how a common property of MPRA data, namely *allelic DNA imbalance* (i.e., the extent of difference in the number of input DNA molecules containing each allele of a SNP) can produce biased results. We propose a non-parametric, adaptive testing method that is robust to allelic DNA imbalance, and we compare the performance of this approach to existing methods in simulated and actual datasets. Finally, we describe an R package that implements the described tests. To facilitate the design and analysis of MPRA experiments, the package also includes simulation procedures to calculate power based on characteristics of specific MPRA datasets.

Materials and Methods

2.1 Brief overview of massively parallel reporter assays

A glossary of key terms is included in Table 1; a schematic overview of the generation of MPRA data is shown in Figure 1; and a detailed description of MPRA laboratory methods has been previously published [Melnikov, et al. 2014]. The goal of MPRA here is to compare the effects of different DNA sequences, represented by synthesized oligonucleotides or *oligos*, on gene expression [Kreimer, et al. 2017]. This is achieved by 1) synthesizing oligos, 2) inserting oligos into plasmids for *transfection* into cells, and 3) using sequencing to count the input DNA molecules and the transcribed RNA molecules from barcodes contained in each oligo.

The first step in an MPRA experiment is to design and construct an *MPRA library*, i.e. a collection of DNA oligos. For each synthesized oligo, a unique *barcode* sequence (usually

10–14 base pairs in length) is included within the sequence so that the oligo and its resulting RNA can be tracked throughout the experiment. For clarity, it is important to distinguish between the specific DNA sequences whose transcriptional effects we wish to quantify (i.e., *potential regulatory elements containing SNPs*) and the slightly longer oligos that contain these elements in addition to a barcode sequence and other sequence elements necessary for transcription. In the literature, the term oligo is often used to refer to both the regulatory element of interest and the synthesized sequence that contains that element. Barcodes are counted by sequencing both before (DNA count) and after transcription (RNA count). Barcoding allows for multiple *internal replicates* within the same transfection experiment, though in practice many studies collapse barcode counts into element-level counts prior to analysis. These internal replicates differ from *experimental replicates* (Supplementary Table S1), which refers to the generation of multiple sets of DNA and RNA counts by performing multiple transfections from the same MPRA library. For the rest of this paper, the term replicate refers only to experimental replicates.

To estimate DNA counts before transcription, a small sample is taken from the MPRA library and sequenced. A separate sample of the MPRA oligo library is transfected into cells, and the transcribed RNA is sequenced. At the end of this process, for each barcode we have a DNA count and a corresponding RNA count (Figure 1, Supplementary Table S1). Because the DNA count is not obtained from the same aliquot that undergoes transfection, it is important to note that there is no replicate-level correspondence between the DNA and RNA counts. To get more precise DNA count estimates for each barcode, an average across the DNA replicates is the best choice [Tewhey, et al. 2016].

2.2 Statistical approaches to identify allelic expression effects in MPRA

We are interested in identifying instances where two SNP alleles have significantly different effects on gene expression. Thus, for pre-defined pairs of oligos, we will use their barcode counts to quantify their *transcriptional efficiency*, i.e., the rate at which a DNA sequence is transcribed into RNA in a given experiment. In the ideal case where the starting DNA count is identical for all barcodes, the RNA count(s) of each oligo would be an accurate estimate of that oligo's transcriptional efficiency, and allelic effects could be determined by directly comparing the RNA counts of the oligos containing the two different SNP alleles. However, MPRA libraries typically have a skewed distribution of barcode counts, and accordingly the difference in DNA counts between two alleles of the same SNP can be many orders of magnitude (i.e., allelic DNA imbalance). This imbalance can lead to biased estimates of allelic transcriptional effect, as we describe below.

To demonstrate the extent of allelic DNA imbalance in published MPRA data, we examined the distribution of DNA counts from two published MPRA studies by Castaldi et al [Castaldi, et al. 2019] and Ulirsch et al [Ulirsch, et al. 2016]. In both studies, the DNA count distribution across oligos is heavily skewed to the right (Figure 2). We also plotted the difference in mean DNA counts between the two alleles across SNPs (Figure 2). Using the MW test on the DNA counts, we found that 11.9% and 5.7% of SNPs have a nominally significant allelic DNA imbalance (p-value < 0.05) in the Castaldi and Ulirsch data respectively. Among this set of SNPs, the median amount of the allelic DNA imbalance

(difference in mean DNA counts between two alleles) is 45.6 (IQR = 54.3) and 99.3 (IQR = 179.8) respectively. Thus, in these datasets, there is a small but important percentage of SNPs with allelic DNA imbalance in which the type 1 error rate of tests for allelic transcriptional effects can be inflated. In other words, even with completely random sampling of barcodes from well-mixed DNA samples, SNPs with observed DNA allelic imbalance would be more likely to be the top findings of a biased statistical method, as demonstrated in section 2.3 below.

When allelic DNA imbalance is present, adjustment for differences in DNA counts is necessary to obtain an unbiased estimate of allelic transcriptional effects. A common approach is to calculate the log of the ratio of RNA and DNA counts. However, when DNA or RNA counts vary greatly between SNP alleles, the variance of the numerator and denominator vary in ways that make it difficult to correctly model the variance of the ratio. For this reason, allelic DNA imbalance can be a major confounder in MPRA analyses despite ratio-based adjustments.

Table 2 describes the characteristics of the commonly used MPRA analysis methods that are evaluated in this paper. Many of them use the RNA/DNA ratio directly to estimate the difference in transcriptional efficiency, while others try to model the RNA count data while adjusting for the DNA count. The MW test (i.e. Wilcoxon rank-sum test) was one of the first and most widely used methods applied to MPRA data analysis [Melnikov, et al. 2012].

2.3 Potential bias due to allelic imbalance when using RNA/DNA ratios in real MPRA data

In this section, we address the relationship between allelic imbalance and the test for allele-specific transcriptional effect. Our hypothesis is that SNPs with allelic imbalance will be disproportionally represented among those testing positive for allele-specific transcriptional effects. We tested this hypothesis in the Castaldi and Ulirsch datasets using the Hypergeometric test.

Using a single replicate from K562 cells in the Ulirsch data, we identified SNPs with significant allelic transcriptional effects using the MW test on RNA/DNA ratios. To identify SNPs with allelic DNA imbalance, we also applied the MW test to the DNA counts only. Among the 7758 SNPs tested (counting instances of the same SNP against a different oligo background separately), 296 SNPs had nominally significant ($p < 0.05$) allelic transcriptional effects. Among these 296 SNPs, 35 of them had nominally significant allelic DNA imbalance (enrichment p-value via hypergeometric test = $3.7e-5$, Supplementary Table S2). We repeated this analysis in the Castaldi data and observed stronger enrichment of allelic DNA imbalanced SNPs among the SNPs with significant allelic transcriptional effects (enrichment p-value = $1.3e-51$). Thus, for single replicate data analyzed by the MW test, allelic DNA imbalance can induce an undesirable dependence between estimates of allelic transcriptional effects and allelic DNA imbalance.

We repeated this analysis using multiple replicates, analyzing two replicates from the Ulirsch data and expanding the set of methods to those that require at least two replicates. We observed that the MW test was still biased (enrichment p-value = $7.9e-11$). We also observed evidence of bias for QuASAR-MPRA [Kalita, et al. 2018], Fisher's Exact Test,

mprialm methods[Myint, et al. 2019], edgeR[Robinson, et al. 2010], and DESeq2[Love, et al. 2014] (enrichment p-values all <0.005) (Supplementary Table S2). The t-test did not show enrichment of allelic DNA imbalanced SNPs in the top results. As the number of replicates increases, the bias from allelic DNA imbalance may be expected to decrease since there is more data to model the variance of the ratio (or log of ratio). When we included five replicates from the Ulirsch data and tested for enrichment of allelic DNA-imbalanced SNPs, only the MW test and QuASAR-MPRA method had significant enrichment.

2.4 A novel, non-parametric, adaptive testing method for MPRA data

To develop an approach that is robust to allelic DNA imbalance, we propose an adaptive testing approach. This method selects between the MW test on RNA/DNA ratios and a stratified MW test across bins matched on DNA counts. When no allelic DNA imbalance is present, the MW test is used to compare the RNA/DNA ratios across the barcode between the two alleles; otherwise the stratified MW test is applied. The stratified MW test compares RNA counts between the two alleles after matching DNA counts into bins. Allelic DNA imbalance is tested by comparing the distribution of barcode counts in the MPRA library between the two alleles for each SNP using the MW test. The bins are determined using the subclassification method implemented in the MatchIt package[Ho, et al. 2011], and the bin size is adaptively reduced until the results are not biased by differences between the DNA distributions. Each bin should have approximately the same number of barcodes, and barcodes that do not have matching DNA counts are not analyzed. Then the stratified MW test implemented in the “coin” R package [Strasser and Weber 1999] is applied to obtain either the asymptotic p-value or the exact p-value if the sample size is small.

This approach preserves power when allelic DNA imbalance is absent while protecting against type 1 error when it is present. For data with multiple replicates, the DNA or RNA counts for each barcode are obtained by taking the mean across the replicates.

Results

4.1 Type I error rate as a function of allelic imbalance

The simulation procedure is described in detail in the Supplementary Material. Under the first simulation scenario relating allelic DNA imbalance to type I error, Figure 3 demonstrates that the type I error rate of multiple methods depends on the severity of allelic DNA imbalance. More specifically, the MW test, QuASAR-MPRA, mpralm mean method with the average estimator, and DESeq2 show increased inflation in type I error rate as the difference in mean DNA counts between the two alleles increases. The Fisher’s exact test has a highly inflated type I error rate (in this case around 0.7), so for visualization purposes it is not included in all the figures. The adaptive test maintains the type I error rate, even in the presence of severe allelic DNA imbalance. We have also explored other cutoffs (i.e., thresholds used to remove barcodes based on mean DNA count and mean RNA count) and demonstrated that similar inflation in type I error can be observed using other cutoffs (Supplementary Figure S1).

4.2 Overall type I error rate and power

For the methods that maintained the type I error rate even in the presence of severe allelic DNA imbalance, we further examined the type I error and power of these methods under a more realistic scenario where there is a distribution of allelic DNA imbalance and SNP transcriptional efficiencies, both of which were parameterized to closely match the estimated distributions from the Ulirsch data. DNA and RNA counts were generated for 3000 SNPs from the same negative binomial model used for the previous simulations. We varied the number of replicates and number of barcodes per oligo to compare the type I error and power of these methods, while keeping the mean depth per barcode constant. Figure 4 shows the overall type I error and power for the adaptive test, t-test, mpralm_sum, and edgeR for data with five replicates. We evaluated up to five replicates based on a previous recommendation that at least four independent replicates are necessary to optimize the performance of many MPRA analysis methods[Myint, et al. 2019]. We observed slight inflation for the mpralm sum method and t-test, and none for the other methods. Among these methods, edgeR and the mpralm_sum method had the best power when the number of barcodes is below 20, and the adaptive test was more powerful when the number of barcodes is over 20 (Figure 4). The improved performance of the adaptive approach is due to the larger number of barcodes available for comparison. These results are robust to the choice of different cutoffs for mean DNA or RNA count (Supplementary Figure S2).

4.3 Application of the adaptive testing approach to real MPRA data

We applied the adaptive approach to single replicate data from the Ulirsch and Castaldi datasets, and we observed no enrichment of allelic DNA-imbalanced SNPs in the significant results (Supplemental Table S2). Similarly, if we select two replicates from the Ulirsch data, no bias is observed (enrichment p-value 0.54). Note that for the application to a single replicate in the Ulirsch data, 296 SNPs were found to be significant by the MW tests with 35 of them having significant allelic imbalance, while 274 SNPs were found to be significant by the adaptive test (all of which are among the 296 significant SNPs using MW test), with 13 of them having significant allelic imbalance. Thus, the adaptive test was able to remove false positive SNPs with allelic imbalance, and the lack of enrichment was not due to sample size. More details can be found in Supplementary Table S2.

4.4 Implementation

We have collected functions for simulating an MPRA experiment and analyzing MPRA data in the @MPRA R package. For designing MPRA experiments, users can specify the number of barcodes per SNP, number of SNPs to be analyzed, number of replicates, total sequencing depth for each replicate, and the variance-to-mean relationship in the DNA and RNA distributions across the replicates. By default, @MPRA simulations use parameters estimated from the data in Ulirsch et al., but the software allows users to estimate key parameters from their own data for simulation studies. This package enables investigators to find the best set of experimental parameters to maximize the power of their MPRA experiment. For example, in Figure 4, we observed that with a fixed mean sequencing depth of 70 across barcodes and 5 replicates, the power of the adaptive approach increased as the number of barcodes for each SNP increased, but plateaus after 50. Therefore, in this

scenario, increasing the number of barcodes of SNP from 10 to 20 would greatly improve the power to detect allelic transcriptional effects. @MPRA also allows the user to observe the relationship between the number of replicates and the power of each analysis, which demonstrates that increasing the number of replicates often improves power dramatically (Supplementary Figure S3). @MPRA has implemented all of the methods analyzed in this paper.

Discussion

The allelic DNA imbalance that is ubiquitous in published MPRA data arises from technical limitations of the library construction process, including PCR amplification bias, which has been shown to produce highly skewed distributions when amplifying sequences present at low copy number[Kebschull and Zador 2015]. The resulting skewed count distribution in MPRA libraries leads to allelic imbalance in the input DNA, which strongly influences RNA counts and therefore is a strong confounder when estimating allele-specific transcriptional efficiency. MPRA methods that rely on ratio-based RNA/DNA representations are susceptible to bias resulting from allelic DNA imbalance, because the distribution of ratio (or log of ratio) of count data is difficult to correctly estimate[Curran-Everett 2013]. The inflation in type I error for most of the methods is likely due to the inadequate modeling of the variance of the ratio.

Since MPRA data are count-based, the application of established RNA-seq methodologies to MPRA analysis is a natural extension of these methods[Anders and Huber 2010; Love, et al. 2014; Robinson, et al. 2010]. QuASAR-MPRA applies beta-binomial models for RNA counts, and mpralm uses linear models and an empirically-derived inverse-variance weighting approach (Table 2). Our study finds that in many cases these approximations work well, but the performance of these models can be sub-optimal in both simulated and real data. The adaptive testing method that we propose adopts a non-parametric approach in which RNA counts are binned according to similar DNA counts then analyzed separately within bins. This more conservative approach has the advantage of maintaining type I error under any distribution of RNA and DNA counts, but it does this at the potential cost of reduced power. In order to address the issue of power, we implemented an adaptive testing procedure that applies the stratified MW only for SNPs with allelic DNA imbalance, thereby preserving power. This approach maintains the type I error since the choices of bins are not affected by the degree of association, but rather by the independent information of input allelic imbalance, as supported by the simulation results.

While our simulation studies suggest that the adaptive method is the most reliable approach with respect to controlling type 1 error, the edgeR and mpralm_sum methods have reasonable type 1 error rates when multiple replicates are available. In practice, the best choice for an MPRA analysis of allelic transcriptional effects depends on parameters specific for a given MPRA library. If allelic imbalance is severe and only one replicate is available, then the adaptive approach is clearly preferred. If allelic imbalance is not severe, or if multiple replicates are available, then edgeR or mpralm with an aggregate estimator may offer some power advantages when the MPRA design has less than 20 barcodes per

SNP allele. If there are more than 20 barcodes per oligo, the adaptive approach may be more powerful.

Many of our conclusions are consistent with a recent comparative analyses of MPRA methods by Myint et al [Myint, et al. 2019] that also examined the performance of multiple MPRA methods in real and simulated datasets. Both studies identified unacceptable inflation from the Fisher's exact test and suboptimal performance of QUASAR-MPRA and DESeq2, and there was agreement that the general performance of the t-test, edgeR, and mpralm with aggregate estimator was good. Our work is distinguished by our focus specifically on tests of allelic transcriptional effects and the effect of allelic DNA imbalance, and the adaptive testing approach presented here is novel and qualitatively different in its approach from the mpralm method presented in Myint et al.

A critical assumption of the adaptive method is that the measurement error for DNA counts is small compared to the measurement error in the RNA counts. Otherwise, barcodes will not be matched correctly into bins and the test will be anti-conservative. We did not encounter this scenario in any of the real MPRA datasets that we analyzed, and whenever possible we used simulation parameters estimated from real MPRA data. Another limitation is that while we presented the problem and compared the methods using real MPRA data from Ulirsche et al and Castaldi et al, more thorough study of the performance using additional MPRA data would be helpful for choosing the optimal methods in different situations. Also, since the power of a nonparametric test is usually lower than parametric tests, a possible approach to develop a more powerful test may be to fit a weighted linear regression on the log RNA counts and test the interaction term between DNA counts and alleles. This bypasses the problem of modeling the RNA-to-DNA ratios. Another alternative is to apply multiple methods to the same data, and to draw conclusions on the best methods to use based on a set of positive controls. Our @MPRA library provides convenient functions to perform such analyses. In addition, the library provides a useful function to evaluate the degree of bias of each method due to allelic imbalance. It also facilitates design of the experiment in determining how many replicates will be necessary for adequate detection power for a given library, and it provides a user-friendly tool to evaluate the performance of multiple MPRA analysis methods using simulations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge the helpful comments from Dr. Vincent Carey from the Channing Division of Network Medicine at Brigham and Women's Hospital, and from Dr. Alejandro Reyes at Dana Farber Cancer Institute.

Funding information

Funding of this project was provided by U.S. National Institutes of Health (NIH) grants R01 HL137927 and R01 HL147148 (to EKS, MHC, and XZ); K01 HL129039 (to DQ); R01 HL124233 and R01 HL126596 (to PJC); R01 HL113264 and R01 HL135142 (to MHC); R33 HL120794 (to EKS and XZ).

Conflict of Interest Statement

Dr. Silverman has received honoraria and consulting fees from Merck, grant support and consulting fees from GlaxoSmithKline, and honoraria from Novartis.

References

- Anders S, Huber W. 2010 Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
- Castaldi PJ, Guo F, Qiao D, Du F, Naing ZZC, Li Y, Pham B, Mikkelsen TS, Cho MH, Silverman EK and others. 2019 Identification of Functional Variants in the FAM13A Chronic Obstructive Pulmonary Disease Genome-Wide Association Study Locus by Massively Parallel Reporter Assays. *Am J Respir Crit Care Med* 199(1):52–61. [PubMed: 30079747]
- Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009 Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10(3):184–94. [PubMed: 19223927]
- Curran-Everett D 2013 Explorations in statistics: the analysis of ratios and normalized data. *Adv Physiol Educ* 37(3):213–9. [PubMed: 24022766]
- Edwards SL, Beesley J, French JD, Dunning AM. 2013 Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* 93(5):779–97. [PubMed: 24210251]
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106(23):9362–7. [PubMed: 19474294]
- Ho DE, Imai K, King G, Stuart E. 2011 MatchIt : Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 42(8):1–28.
- Inoue F, Ahituv N. 2015 Decoding enhancers using massively parallel reporter assays. *Genomics* 106(3):159–164. [PubMed: 26072433]
- Kalita CA, Moyerbrailean GA, 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. [PubMed: 29028988]
- Kebschull JM, Zador AM. 2015 Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* 43(21):e143.
- Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, Shin S, Welch R, Wainberg M, Mohan R, Sinnott-Armstrong NA and others. 2017 Predicting gene expression in massively parallel reporter assays: A comparative study. *Hum Mutat* 38(9):1240–1250. [PubMed: 28220625]
- Love MI, Huber W, Anders S. 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550. [PubMed: 25516281]
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J and others. 2012 Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190–5. [PubMed: 22955828]
- Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr., Kinney JB and others. 2012 Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30(3):271–7. [PubMed: 22371084]
- Melnikov A, Zhang X, Rogov P, Wang L, Mikkelsen TS. 2014 Massively parallel reporter assays in cultured mammalian cells. *J Vis Exp*(90).
- Myint L, Avramopoulos DG, Goff LA, Hansen KD. 2019 Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* 20(1):209. [PubMed: 30866806]
- Robinson MD, McCarthy DJ, Smyth GK. 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–40. [PubMed: 19910308]
- Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC. 2016 Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res* 26(2):238–55. [PubMed: 26576614]
- Strasser H, Weber C. 1999 On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics* 8(2):220–250.

- Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF and others. 2016 Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* 165(6):1519–1529. [PubMed: 27259153]
- Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS and others. 2016 Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* 165(6):1530–1545. [PubMed: 27259154]
- White MA. 2015 Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics* 106(3):165–170. [PubMed: 26072432]
- [dataset] Ulirsch JC, Sankaran VG. 2015 Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits [Microarray]. Gene Expression Omnibus; GSE70531.
- [dataset] Zhou X, Castaldi PJ, Guo F, Qiao D. 2018 Fine Mapping and Functional Characterization of Genetic Variants in the FAM13A Chronic Obstructive Pulmonary Disease GWAS locus using Massively Parallel Reporter Assays. Gene Expression Omnibus; GSE109452

Figure 1. Effects of allelic DNA imbalance in MPRA libraries.
This example illustrates MPRA data generation for a single experimental replicate of one SNP. Groups of oligos are synthesized that differ only in the SNP allele and the barcode sequence. After oligo synthesis, there is one copy of each unique sequence. Panel A illustrates that after PCR amplification, there is a skewed distribution of barcode counts due to irregularities of PCR, and the DNA counts from sequencing are obtained from a different aliquot than the one that undergoes transfection. Panel B illustrates how PCR amplification leads to allelic imbalance in DNA counts and how sampling variability between DNA aliquots reduces the accuracy of estimates of the DNA counts used for transfection.

Figure 2. Skewness in DNA count distribution across the barcodes.

The figures in the first row represent the distributions of DNA counts across barcodes in the Castaldi and Ulirsch data. The figures in the second row represent the distributions of the absolute difference in mean DNA counts between two alleles of each SNP in the Castaldi and Ulirsch data. The x-axes are log-scaled for clearer visualization.

Figure 4. Overall type I error rate and power of MPRA methods in simulated MPRA data with 5 replicates.

A cutoff of 5 on the mean DNA count was applied. The MW test, QuASAR-MPRA, DESeq2, and the mpralm method with average estimator were not included since they have already been shown to be biased by allelic DNA imbalance.

Table 1.

A glossary of key terms used here.

Terms	Definition
allelic DNA imbalance	The extent of difference in the number of DNA molecules between the two alleles of a SNP
allelic transcriptional effect	A measure of the extent to which one SNP allele alters gene expression relative to another SNP allele
barcodes	Very short sequence of typically 10–14 nucleotides included at the end of a synthesized oligo that allows that molecule to be tracked and quantified through sequencing
experimental replicate	A single transfection of an MPRA library
internal replicate	One copy of a functional element (oligo) that has a unique barcode sequence. When a single functional element (i.e. oligo) is synthesized multiple times with different barcode sequences, each individual full-length sequence constitute one internal replicate for the oligo.
MPRA library	Collection of synthesized sequences that are inserted into plasmids to be studied in an MPRA experiment
oligonucleotide (oligo)	Short sequence of DNA (typically 100–300 base pairs) that is synthesized as part of an MPRA library. Oligo is often used to refer to a specific regulatory element - thus in an MPRA experiment there can be multiple copies of an oligo that vary only in barcode sequence or SNP allele.
plasmid	A circular segment of DNA that contains functional elements necessary for independent transcription within cells. MPRA oligos are inserted into plasmids.
transfection	The process by which plasmids constituting an MPRA library are inserted into cells for subsequent transcription of oligos into RNA.

Table 2.

Commonly used approaches for analyzing MPRA data.

Test	Needs multiple replicates	Uses RNA/DNA Ratio	Description
MW test[Melnikov, et al. 2012]	No	Yes	Compares the RNA/DNA ratios between the two alleles
QuASAR-MPRA[Kalita, et al. 2018]	No	No	Sums over all barcodes and replicates for each allele, models the reference allele count conditional on the total allele count in the RNA using beta-binomial distribution. The null assumes reference allele/alternative allele count ratios are the same in the RNA and the DNA.
Fisher's exact test[Myint, et al. 2019]	No	No	2x2 table for total counts for each allele and DNA or RNA categories. Counts are summed over all barcodes and replicates.
t-test (paired)[Tewhey, et al. 2016]	Yes	Yes	Oligo counts are obtained by summing over all barcodes. RNA oligo counts/mean DNA counts are compared using a paired t-test between the two alleles.
Mpralm_mean[Myint, et al. 2019]	Yes	Yes	Linear mixed effect model where the average of $\log[(\text{RNA}+1)/(\text{DNA counts}+1)]$ across barcodes were used as outcome. The variance of the outcome is modeled as a smoothing function of the mean DNA count.
Mpralm_sum[Myint, et al. 2019]	Yes	Yes	Same as above, except the outcome is $\log[(\text{RNA oligo counts}+1)/(\text{DNA oligo counts}+1)]$.
DESeq2[Love, et al. 2014]	Yes	No	Assumes a negative binomial distribution for the RNA counts while fitting a model for the dispersion parameter. DNA counts are included as offset terms.
edgeR[Robinson, et al. 2010]	Yes	No	Negative binomial distribution for the output RNA counts. DNA counts are included as offset terms.