

MPRAnalyze - A statistical framework for Massively Parallel Reporter Assays

Tal Ashuach^{1,2,*}, David Sebastian Fischer^{3,4,*}, Anat Kreimer^{1,5,6},
Nadav Ahituv^{5,6}, Fabian Theis³, and Nir Yosef^{1,2,7,8,†}

¹Department of Electrical Engineering and Computer Sciences,
University of California, Berkeley, California USA.

²Center for Computational Biology, University of California,
Berkeley, California USA.

³Institute of Computational Biology, HelmholtzZentrum Munchen,
85764 Neuherberg, Germany

⁴TUM School of Life Sciences Weihenstephan, Technical University
of Munich, 85354 Freising, Germany

⁵Department of Bioengineering and Therapeutic Sciences,
University of California San Francisco, San Francisco, California,
USA

⁶Institute for Human Genetics, University of California San
Francisco, San Francisco, California, 94158, USA

⁷Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA,
USA

⁸Chan Zuckerberg BioHub, San Francisco, CA, USA

*These authors contributed equally to the work.

†Corresponding author. Email: niryosef@berkeley.edu

Abstract

Massively parallel reporter assays (MPRAs) are a technique that enables testing thousands of regulatory DNA sequences and their variants in a single, quantitative experiment. Despite growing popularity, there is lack of statistical methods that account for the different sources of uncertainty inherent to these assays, thus effectively leveraging their promise. Development of such methods could help enhance our ability to identify regulatory sequences in the genome, understand their function under various setting, and ultimately gain a better understanding of how the regulatory code and its alteration lead to phenotypic consequence.

Here we present MPRAnalyze: a statistical framework dedicated to analyzing MPRA count data. MPRAnalyze addresses the major ques-

tions that are posed in the context of MPRA experiments: estimating the magnitude of the effect of a regulatory sequence in a single condition setting, and comparing differential activity of regulatory sequences across multiple conditions. The framework uses a nested construction of generalized linear models to account for uncertainty in both DNA and RNA observations, controls for various sources of unwanted variation, and incorporates negative controls for robust hypothesis testing, thereby providing clear quantitative answers in complex experimental settings.

We demonstrate the robustness, accuracy and applicability of MPRA-Analyze on simulated data and published data sets and compare it against the existing analysis methodologies. MPRA-Analyze is implemented as an R package and is publicly available through Bioconductor [1].

Introduction

Enhancers are non-coding DNA sequences that contribute to the regulation of gene expression. Enhancers control the levels, timing and location of transcription, playing a crucial role in maintaining and determining cell identity and state. Sequence variants in enhancers can have significant consequences, as demonstrated by most disease-associated expression quantitative trait loci (eQTLs) identified through genome-wide association studies falling within non-coding regions of the genome [2].

Since enhancers regulate transcription by interacting with transcription factors and the transcriptional machinery, active enhancers tend to reside in areas of open chromatin. Additionally, enhancers have been shown to be marked by the histone modification H3K27ac [3, 4, 5, 6]. Assays that map these properties have been used extensively for genome-wide identification of active enhancers in various contexts [7, 8]. However, while these assays enable identification of candidate enhancers, they are limited to a binary view of enhancer activity, and are therefore insufficient to fully understand cis regulation of transcription. Functional assays are necessary to further our understanding of enhancers' role in gene expression regulation. Reporter assays have been used to functionally annotate enhancers, by introducing fluorescent reporter constructs regulated by the enhancer of interest, but these assays have limited throughput and don't scale to allow genome-wide functional annotations.

Recent advances in reporter assays address this issue, in a set of procedures denoted Massively Parallel Reporter Assays (MPRAs). These assays replace fluorescent reporters with sequence-based identifiers, denoted "barcodes". Broadly, a synthetic construct that contains a minimal transcriptional structure is introduced into cells. Each such construct is generally composed of an enhancer of interest, a minimal promoter and a unique barcode. The synthetic enhancer is assumed to regulate the transcription of the barcode sequence similarly to how the native enhancer regulates the transcription of its target gene. The cells then undergo RNA and DNA sequencing to measure both RNA transcript counts and DNA construct counts for normalization purposes, and the RNA/DNA ratio is

used to estimate the transcription rate. Relying on sequence-based identifiers allows using the vast combinatorial space of unique sequences instead of a limited set of fluorescent reporters, and leveraging next generation sequencing to measure the activity of thousands of enhancers in a given experiment.

MPRAs can be used to address several scientific questions. In mutagenesis experiments they are used to quantify the transcription rate of a variant, enabling a quantitative comparison to other variants, thereby measuring the effects of various mutations and alterations. Similarly, MPRAs can be used to classify enhancers as active: significantly affecting the native transcription rate of the promoter [9, 10]. In classification studies, control sequences are typically included to establish a baseline transcription rate for the minimal promoter used. Finally, MPRAs can be used for comparative analyses, comparing enhancer activity between different alleles [11, 12], tissues [13], or other conditions of interest. More complex experimental designs are also possible, for example measuring the interaction between alleles and condition [12], or measuring temporal behavior with time-series data [14].

Despite growing popularity of MPRAs, current studies have used various ad-hoc methods or methods that were not developed for MPRA data, such as DESeq2 [15], that relies on underlying assumptions that may not be true for MPRA data. Other MPRA analysis methods only address some of the types of questions MPRAs can address, such as QuASAR-MPRA[16] and mpralm [17] that only perform comparative analyses, and both rely on ratio-based summary statistics that limit the power of the analysis. In contrast, MPRAalyze provides a general statistical framework that allows all uses of MPRAs to be addressed using a single model, leverages the unique structure and characteristics of MPRA data, and avoids relying on limited statistics or over-reaching assumptions.

Results

MPRA data is produced from two parallel procedures: RNA-seq data from post-transduction cells measures the number of transcripts produced of each barcode, and DNA-seq data measures the number of construct copies of each barcode. Thus for each barcode in the experiment both DNA and RNA counts are observed, and the ratio RNA/DNA serves as a conceptual proxy for the transcription rate. However, both DNA and RNA measurements are products of sub-optimal and noisy procedures, an issue exacerbated by the unstable nature of a ratio: minor differences in the counts themselves can result in major shifts in the ratio, especially when dealing with small numbers. This problem can be handled by associating multiple barcodes with each enhancer, providing multiple replicates within a single experiment and a single sequencing library. This approach introduces an additional problem of how to properly summarize counts from multiple barcodes to a single transcription rate estimate of an enhancer, which is made difficult since transduction efficiency, while theoretically uniform across the different constructs, has a significant degree of variability

in practice (Figure 1A). Two methods of summarizing are commonly used: the aggregated ratio, which is the ratio of the sum of RNA counts across barcodes divided by the sum of DNA counts across barcodes; and the mean ratio, which is the mean of the observed RNA/DNA ratios across barcodes. Both of these summary statistics have inherent limitations. The aggregated ratio loses the statistical power that multiple barcodes provide, and the mean ratio is highly sensitive to noise, as demonstrated by Myint et al. [17].

1 MPRAalyze Model

We propose MPRAalyze, a dedicated model for the analysis of MPRA data that uses a graphical model to relate the DNA and RNA counts, model the uncertainty in both libraries and take advantage of the unique structure and opportunities presented by MPRA data (Figure 1B). Our model relies on the assumption of a linear relationship between the RNA counts and the DNA counts: $RNA = DNA \cdot \alpha$, similarly to ratio-based approaches, with ‘ α ’ denoting the transcription rate. To account for the variability of barcode abundances as well as other covariates (conditions of interest, batch effects, etc), our model constructs two generalized linear models (GLM). The first GLM is the DNA model, which estimates the latent construct counts from the external covariates and the observed DNA counts. The second GLM, the RNA model, estimates the rate of transcription from the external covariates, estimated construct counts obtained from the DNA model, and observed RNA counts. Formally, for a given enhancer, we have two vectors observations: DNA counts \vec{d} and RNA counts \vec{r} . Then the MPRAalyze models are:

$$\begin{aligned}\log(\vec{d}) &= \log(\vec{\hat{d}}) + \log(\vec{S}_D) \\ &= X_D \vec{\beta} + \log(\vec{S}_D) \\ \log(\vec{r}) &= \log(\vec{\hat{r}}) + \log(\vec{S}_R) \\ &= \log(\vec{\hat{d}} \cdot \vec{\alpha}) + \log(\vec{S}_R) \\ &= \log(\vec{\hat{d}}) + \log(\vec{\alpha}) + \log(\vec{S}_R) \\ &= X_D \vec{\beta} + X_R \vec{\gamma} + \log(\vec{S}_R)\end{aligned}$$

Where $\vec{\hat{d}}, \vec{\hat{r}}$ are the abundance estimates of the constructs and transcripts, respectively; S_D, S_R are external normalization factors used to correct technical effects such as library size; $\vec{\alpha}$ is the vector of transcription rate estimates, which can be either a single value if the analysis is looking for a single estimate (this would normally be the case for quantification and classification analysis), but can encode for multiple estimates for a single enhancer - for example if multiple

biological conditions are analyzed simultaneously, the model can compute the α estimate for transcription rate for each condition; β, γ are the model parameters and X_D, X_R are the design matrices, which encode the experimental design of the assay. Briefly, each column in the matrix corresponds to a single coefficient, and each row to a single sample. Numerical factors are incorporated as single columns, whereas for categorical factors (such as replicates or conditions of interest), each category has a separate coefficient and therefore a separate column, with one of the categories being absorbed as the reference (baseline) value, and the rest being treated as contrasts, and the values of the matrix being binary, determining the inclusion of each coefficient in the modelling of each sample (A simplified example is provided in Figure S1).

This formulation allows for straight-forward encoding of various covariates, and easily supports the common structure of MPRA experiments: multiple barcodes per enhancer, multiple replicates, and often multiple conditions analyzed simultaneously. This flexibility also allows for various covariates to only be modelled in one of the models, depending on the scientific question and experimental design. For instance, barcode-level effects should be incorporated into the DNA model to allow for proper normalization of the transcript counts, but should usually be excluded from the RNA model since we do not expect each barcode to result in a different transcription rate. Alternatively, in unpaired settings where the DNA sequencing was performed on pre-transduction libraries, there might not be separate DNA estimates for each condition being tested, in which case the conditions of interest would only be modelled in the RNA model, and excluded from the DNA model.

We optimize this model by maximizing the likelihood of the data using certain distributional assumptions. First, we assume that the latent construct counts, from which the observed DNA counts are sampled, are generated by a gamma distribution: $\vec{d} \sim \text{Gamma}(k, b)$. Second, we assume that the conditional distribution of the RNA counts follows a Poisson distribution: $\vec{r}|\vec{d} \sim \text{Poisson}(\alpha \cdot \vec{d})$, which results in a closed-form negative binomial likelihood for the RNA counts themselves: $\vec{r} \sim \text{NB}(\mu = \frac{\alpha \cdot k}{\beta}, \psi = k)$. The negative binomial distribution is a common approximation of sequencing data, and indeed all datasets we examined have the negative-binomial characteristic quadratic relationship between the mean and the variance. This relationship is also observed for the DNA libraries, which is expected of Gamma-distributed data if the distribution's shape parameter $k \approx 1$ [Figure S2].

Our framework accounts for barcode specific effects and leverages them for increased statistical power while simultaneously benefiting from the robustness of aggregating information across barcodes. Since a standard for MPRA experimental design has yet to be formed, the nested GLM construction is flexible and can be easily extended to changing experimental designs. Our model is also highly interpretable, easily allowing for quantitative estimates of enhancer activity to be extracted, as well as differential activity to be tested directly using established statistical tests. Our framework also explicitly leverages negative

controls when available, either to establish the null distribution in classification analyses or to correct for systemic bias in comparative analyses [see Methods].

To characterize the properties and evaluate the performance of the MPRA-Analyze model, we compared MPRA-Analyze’s performance and the properties of our model to other previously used and newly developed methods, using both simulated data and a set of four MPRA datasets detailed in Table 1. These datasets were chosen for representing a diversity of MPRA procedures (episomal or lentiviral integration, DNA sequencing from pre- or post-transduction), study focus (quantification, classification and comparative analyses), and experimental design (number of barcodes per enhancer). Note that only a subset of features of the Kwasnieski datasets are used (Only the weak and strong categories were used), and only a subset of samples in the Inoue-Kreimer dataset are used (only timepoint T0 in the quantification and classification analyses, and only T0 and T72 in the comparative analyses).

Dataset	Description	Integration	#Enhancers	#Controls	#Barcodes
Kwasnieski	Testing ENCODE putative weak/strong enhancers in K562 cells. [10]	Episomal	1200	568	4
Inoue (episomal)	Candidate liver enhancers in HepG2 cells without genomic integration. [9]	Episomal	2338	102	100
Inoue (chromosomal)	Candidate liver enhancers in HepG2 cells with lentivirus-based genomic integration [9]	Lentiviral	2338	102	100
Inoue-Kreimer	time-series MPRA during neural differentiation induction in human ESCs. [14]. For convenience, referred to as <i>Inoue-Kreimer</i> .	Lentiviral	2464	200	90

Table 1: MPRA datasets used for evaluation of MPRA-Analyze throughout the paper.

When fitted to MPRA data, we found that the MPRA-Analyze model is able to properly capture the characteristic of the data and provides a good fit across all datasets ($R^2 > 0.86$ for all datasets, figure 1C). To examine the validity of assuming the DNA counts are gamma-distributed, random data was generated from the fitted DNA GLM and the residuals were compared with the residuals of the observed counts. Quantile-based comparisons shows that the residuals are generated by similar distributions [Figure S3], indicating that this assumption does not significantly distort that distribution of the observed data.

2 Quantification

We set out to examine the properties of MPRA-Analyze’s estimate of transcription rate, denoted ‘alpha’, and compare it to the naïve ratio-based summary statistics. Overall the three estimates are largely in agreement (Pearson’s $r > 0.9$ across datasets, Figure S4), demonstrating that alpha is indeed capturing the correct signal.

To examine the accuracy of the estimates, we used the negative control sequences included in some of the datasets. These are assumed to have an identical transcription rate induced by the minimal promoter with no additional

enhancer activity. We examined the variance of the estimates on these sets. In the Kwasnieski dataset, which has a limited number of barcodes, the three estimates all had low variance. In the barcode-rich datasets, alpha was clearly more consistent across the negative controls than the other two estimates, with the aggregated ratio being the least consistent (Figure 2A).

We then set out to explore the effect of the number of available barcodes on the performance of the estimates. Using the barcode-rich datasets, barcodes were subsampled at various rates and estimates were recomputed for each enhancer (3 independent samples per enhancer per barcode rate). Using the full-data estimates as a baseline, we found that subsampling barcodes does not result in a systemic bias in any of the estimates (Figure 2B). Expectedly, all estimates showed reduced variance with increased barcodes, with the mean ratio underperforming the other two estimates, and alpha having a similar or lower variance than the aggregated ratio (Figure 2C).

In many cases the goal of quantifying enhancer activity is to rank and compare different enhancers, as in mutagenesis experiments. To compare the stability of the ordering of enhancers, the Pearson correlation was computed between the estimates in each sub-sample to the estimates of the full data. Alpha has either similar or higher correlation than both naive estimates across datasets and barcode abundance (Figure 2D).

Noting that these analyses are limited by a lack of ground truth, MPRA data was then simulated by generating random coefficients and using the same nested GLM construction as described above to generate samples. To avoid biasing the results, samples were generated with a log-normal noise model, instead of the default Gamma-Poisson convolutional model MPRAalyze uses [methods]. We generated 101 enhancers with gradually increasing transcription rates (from 2 to 3, in 0.01 steps). The analyses above were repeated with the simulated data. We found that while the measured bias was indeed not influenced by the number of barcodes, the mean ratio displayed a significant amount of bias compared with both alpha and the aggregated ratio (Figure 2E). Similar to the real data results, we found alpha has lower variance than both naive estimates, and higher correlation with the true transcription rates (Figure 2F-G).

Overall, alpha is as or more stable and accurate as the aggregated ratio when barcode information is limited, and is more consistent across similarly-behaving enhancers than both the aggregated and the mean ratio.

3 Classification

MPRA-based classification of active enhancers has previously been done by comparing the ratio-based estimates of candidate enhancers to the control set [9, 10], an approach that suffers from the limitations of the summary statistics demonstrated above. Other studies performed this analysis using DESeq2 [15], a differential expression analysis (DEA) method, by treating the DNA and RNA libraries as two conditions and looking for “differentially expressed” enhancers [11]. However, the method relies on an implicit assumption that the majority of

features do not display differential behavior, a valid assumption for DEA that does not hold for classification of MPRA data, in which the candidate enhancers are often explicitly selected as sequences that are likely to be active.

MPRAnalyze performs classification of active enhancers by testing each enhancer's alpha estimate against a null distribution describing the null transcription rate induced by the minimal promoter without enhancer activity. When negative controls are available, they are used to estimate the null distribution. When they are not available, MPRAnalyze relies on a conservative assumption that the mode of the distribution of transcription rate estimates is the center of the null distribution, and that values lower than the mode all belong to the null. Thus, MPRAnalyze estimates the null by locating the mode and using only the values lower than the mode to estimate the variance. Each enhancer's alpha values are then tested by computing Median-Absolute-Deviation (MAD) scores, median-based variants of Z-scores that are less sensitive to outliers.

To assess MPRAnalyze's performance in classification analyses we compared 6 methods: MPRAnalyze with and without controls; empirical p-values computed using the naive ratio estimates; and DESeq2 in either full mode (barcode-level data) or collapsed mode (summing across barcodes within each batch). DESeq2 hypothesis testing was performed using an asymmetric alternative hypothesis, only looking at enhancers that were more active in the RNA library than in the DNA library.

We examined the fraction of enhancers that were significantly active ($FDR < 0.05$) in each dataset, stratified by group: negative controls, candidate enhancers and positive controls when available. Expectedly, empirical p-values from the naive ratio estimates show a clear lack of power, only being comparably sensitive in the Inoue-Kreimer dataset. Both DESeq2-collapsed and MPRAnalyze without controls have inflated rates of activity in the Kwasnieski datasets, however overall both modes of DESeq2 and both modes of MPRAnalyze have reasonable results across datasets. Overall, MPRAnalyze and DESeq2 seem to find comparable numbers of active enhancers, with no clear advantage to either method [Figure 3A].

However, only looking at the fraction of enhancers that pass a threshold misses the overall statistical behavior of the method. We therefore examined the full p-value distribution of each method within each dataset, finding that despite comparable rates of enhancers found statistically significant, the MPRAnalyze model is far better calibrated to MPRA data. When negative controls are used, MPRAnalyze-generated p-values follow the theoretical behavior of the statistic: uniform on null-generated data, low values otherwise. When negative controls are not used, MPRAnalyze has some deviations from behavior, emphasizing the importance of including negative controls in classification experiments, to properly characterize the null behavior. Both MPRAnalyze modes have significantly higher statistical power, demonstrated in the distribution of p-values on the positive controls included in the Inoue datasets. Conversely, DESeq2 p-values do not follow the theoretical behavior of a well-calibrated test, instead mostly having a concentration of high and low p-values, both among candidates and controls [Figure S5].

We hypothesized that these calibration issues are partly explained by the asymmetric alternative hypothesis test we used when running DESeq2. When directly comparing p-values from MPRAalyze to the competing methods, we indeed found that overall the scores are correlated, and that the abundance of high-valued p-values are mostly among enhancers that MPRAalyze and the ratio-based p-values do not reject the null for, indicating that DESeq2 views these as “down regulated”, meaning enhancers whose expression in the RNA library is lower than DESeq2 model expects it to be based on the DNA library counts [Figure S6].

One of the major aspects of the DESeq2 model is the dispersion shrinkage mechanism that the model uses. This is common practice among differential expression methods, and includes pooling information across all features included in the dataset (genes for RNA-seq, candidate enhancers in MPRA). Since RNA-seq is a genome-wide assay, the set of measured features across different experiments can be assumed to remain stable, if not necessarily constant. This does not hold true for MPRA experiments, in which the number and composition of the measured regulatory sequences is curated according to the specific goals and context of the given experiment. We therefore hypothesized that DESeq2-based classification would be highly dependent upon the composition of enhancers included in the analysis. To demonstrate this, we re-ran the analysis on the Inoue-Kreimer dataset, but only included the 200 control enhancers and 685 enhancers that were classified as active by MPRAalyze, DESeq2-full and DESeq2-collapsed. This simulates a scenario in which an identical experiment was performed, producing the same data, but included fewer enhancers that were selected with higher degree of certainty of their activity. Since MPRAalyze only pools information across enhancers when correcting for library size, we expected it to recapitulate the original results and indeed MPRAalyze finds all candidates are significantly active. However, DESeq2-full only identifies 161 (23.5%) of the candidates as active, and DESeq2-collapsed finds no active enhancers at all [Figure 3C]. These results are not surprising, as the high abundance of activity would shift DESeq2’s estimate of the null behavior, whereas MPRAalyze avoids using the entire population to estimate the null. This reveals an inherent reproducibility issue in using differential expression analysis designed for RNAseq to perform MPRA classification.

4 Comparative Analysis

Another common use for MPRA is comparative studies, looking for differential behavior of enhancer induced transcription between conditions. These are often comparisons between different tissues, stimuli or comparing different alleles of an enhancer sequence [11, 13]. More complex experimental settings are also possible, e.g measuring temporal activity as in the Inoue-Kreimer data [14], or the interaction between differential allele activity and GATA1 presence in Ulirsch et al [12].

MPRAalyze’s model construction is based on generalized linear models, and

as such is highly flexible and extendable to various experimental designs. Performing differential activity analysis in the MPRAnalyze framework can be done in two straight-forward ways: first, since MPRAnalyze optimizes the model using likelihood maximization, any single hypothesis that can be encoded in a generalized linear model can be tested using the likelihood ratio test. This includes complex hypotheses that are not encoded in a single coefficient. Additionally, in simple two-condition designs, or in cases where multiple contrasts are compared to a single reference (e.g multiple different stimuli compared against the unstimulated behavior), the model coefficients can be extracted from the RNA model and tested using a Wald test. Both options are supported in the current implementation of MPRAnalyze, however results in this paper are based on the likelihood ratio test option.

When performing differential activity analysis in MPRA data, it can be important to account for possible systemic bias, for example if the cell types being compared have different inherent transcription rates. In RNA-seq experiments this issue is usually resolved via library size correction, but with MPRA this is not necessarily sufficient, and may actually introduce further bias. This is because for the library size to properly correspond to inherent bias in the data, either the vast majority of features must be non-differential, or the differential signal must be symmetric. Neither of these assumptions necessarily hold for MPRA data, since the features are curated and vary between different experiments. An MPRA can be designed with all features being up-activated (more active in the contrast condition than the reference), and in which the vast majority of features are indeed differentially active. To address this issue, MPRAnalyze utilizes negative controls in the data to define the null differential behavior. This is done by fitting a separate, joint model for the controls, in which each control enhancer has a distinct DNA model but they all share a single RNA model, essentially finding the common activity pattern across the conditions. The model coefficients of this joint model are then incorporated into the model fitted for each candidate enhancer as additional correction factors [see Methods].

Alternative methods have been developed to address this question. QuASAR-MPRA [16] was designed specifically for allelic-comparisons and uses a beta-binomial model and `mpralm` [17] which is a general differential-activity tool designed for MPRA which fits a linear model. Both methods use summary statistics and do not include barcode-level information in their model. `mpralm` enables using either the aggregated ratio or the mean ratio as the statistic, and therefore inherently suffers from the limitations of these statistics described above. QuASAR-MPRA, similar to MPRAnalyze, models the DNA and RNA separately, but it does so on the sum of counts across all barcodes in each condition, collapsing the data into a single measurement.

To compare the performance of MPRAnalyze to the above methods, we used the Inoue-Kreimer dataset and extended the subset of samples we used to include both T0 and T72 timepoints (0 and 72 hours into neural induction of hiPSCs). We then compared enhancer activity between the different timepoints, using the four methods: MPRAnalyze, `mpralm` in both aggregated ratio and

mean ratio modes, and QuASAR-MPRA.

The distribution of p-values [Figure 4A] shows that overall MPRAalyze and both modes of mpralm are well calibrated, following the expected mixture of uniform and low values among candidates, and showing slight inflation but overall uniform behavior among the controls. Conversely, QuASAR-MPRA seems poorly calibrated on both candidates and controls, recapitulating the results described by Myint et al [cite]. Indeed, QuASAR-MPRA only identified 2 candidates as significantly differential (BH-corrected p-values < 0.05).

We then examined the agreement among the methods by comparing the log Fold-Change estimates produced by each method. For QuASAR-MPRA we used the *betas.beta.binom* value, which is the logit transformation of the allelic skew, a proxy for biological effect size. Overall all four methods are well correlated (Pearson's $r \geq 0.84$ across all pairs), so similar biological effect is being observed by all methods. MPRAalyze is more conservative than both mpralm modes, and has overall a positive skew compared with QuASAR-MPRA [Figure 4B].

We then compared MPRAalyze's statistical effects to the other methods by looking at BH-corrected p-values [Figure 4B]. Both modes of mpralm are correlated with MPRAalyze among statistically significant candidates, but demonstrate reduced statistical power. QuASAR-MPRA is expectedly uncorrelated, since the model is not statistically well calibrated.

Further examination of the results excluded QuASAR-MPRA since it did not identify enough differentially activated candidates for the subsequent analyses. After filtering the results to only include candidate enhancers that are also classified as active in at least one of the conditions (BH-corrected $p < 0.05$, using MPRAalyze's classification method, since mpralm does not support classification analysis), we found that MPRAalyze identifies a higher number of significantly differentially active candidates than mpralm in either mode [Figure 4C]. Interestingly, mpralm in aggregate mode finds a roughly balanced number of enhancers that are increasing (99) and decreasing (91) in activity, and in mean mode finds more decreasing (89) than increasing (49), while MPRAalyze finds far more increasing (351) than decreasing (115) enhancers. However, the experimental design of the selected enhancers in the Inoue-Kreimer dataset explicitly favored sequences that corresponded to increase in gene expression over the course of differentiation, therefore MPRAalyze's up-skewed results are expected, and better correspond to the dataset design.

We then explored the set of candidates that were detected by each method. For each candidate sequence potential transcription factor binding sites (TFBS) were identified [see Methods]. The set of differentially active enhancers we divided to decreasing and increasing, then within each set each transcription factor was tested for enrichment (hypergeometric test, BH-corrected $p < 0.05$). To narrow down the results, we examined the union of top 15 most enriched transcription factors by each method [Figures 4D, full results at Supp.]. Additionally, we examined a *consensus+noise* option, wherein we took the consensus set of enhancers for which all methods agree and added artificial counts to match MPRAalyze number of significant enhancers. These counts were taken from the remaining population and are proportional to the fraction of enhancers

that contain a binding site to some factor, to simulate enrichment inflation that doesn't reflect true biological signal [See methods]. Among decreasing-activity enhancers, on which the three methods are largely in agreement, we get an expected similarity of enrichment scores across transcription factors, with MPRAalyze outperforming the other methods in two of the core pluripotency factors (NANOG, POU5F1). Among increasing-activity enhancers, in which the differences between the three methods are more profound, we find a higher variability of enrichment scores across the top-enriched factors. Overall, mpralm in *mean* mode fails to capture many of the enriched transcription factors captured by the two other methods, only finding 23 enriched transcription factors (compared with 106 and 195 found by mpralm *aggregate* and MPRAalyze, respectively) and having diminished statistical power. When comparing MPRAalyze to mpralm *aggregate*, we find overall agreement on the set of enriched factors, with MPRAalyze consistently having increased statistical power with few exceptions. This increase in power cannot be explained simply by MPRAalyze's detection of more enhancers, as can be seen by the *background + noise* results, demonstrating that MPRAalyze's results reflect true biological signal. Notable disagreements between the methods include TEAD2 and NRF1 that are enriched in MPRAalyze's results but not in either mode of mpralm. Both factors have been implicated in neurogenesis by previous studies [18, 19], and show increased expression in the later time frame [Figure S7]. In the other direction, MYF5 and GSX1 are enriched according to mpralm but not enriched according to MPRAalyze. However, when examining mRNA levels of these factors, both factors have low expression levels, below their characteristic expression levels in tissues they are known to be active in (based on GeneCard[20] reported levels), making them less attractive candidates of driving differential gene expression.

Overall, MPRAalyze identifies a similar biological signal to the competing methods, with increased statistical power which allows for more nuanced results.

Discussion

Massively Parallel Reporter Assays are a powerful technique for functional characterization of enhancer activity in a high-throughput manner. MPRAs can be used to directly quantify enhancer-induced activity and identify active enhancers, compare regulatory activity of different alleles, elucidate regulatory grammar via mutagenesis studies, and compare enhancer activity between conditions. Complex experimental design can include interaction studies, studying how sequence changes affect differential enhancer activity [12], or identifying temporal activity patterns in time-course MPRA data [14].

Since MPRAs are still a nascent technology, they often vary in experimental design. While MPRAalyze can analyze any MPRA dataset, the method benefits from certain experimental decisions that are generally recommended but not always leveraged in other analyses. First, pairing the DNA and RNA libraries

by extracting DNA from the same post-transduction libraries that the RNA libraries are extracted from, avoids introducing further experimental noise into the data, and enables MPRAalyze to better fit and relate the two models to increase accuracy of estimating nuisance factors. Additionally, as demonstrated in our results, increasing the number of available barcodes can greatly reduce the measured noise and increase performance of all methods. Finally, the inclusion of negative control sequences allows explicit modeling of the null behavior and avoids relying on assumptions that may bias the results and prevent proper interpretation of them. The curated nature of MPRA datasets makes negative control a valuable and often crucial aspect of properly interpreting the results.

MPRAalyze is the first method to emerge that offers a robust statistical framework that enables analyzing data for all major uses of MPRA in a unified model. MPRAalyze models noise in both DNA and RNA libraries and uses a powerful nested GLM design to control barcode-specific effects and leverage the multiplicity of barcodes for increased statistical power. The method is highly flexible and allows various complex study designs to be tested in a straightforward manner. Additionally, MPRAalyze avoids relying on population-level properties in the analysis, instead leveraging negative controls when available to establish null behaviors.

References

- [1] Wolfgang Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. en. In: *Nat. Methods* 12.2 (Feb. 2015), pp. 115–121. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3252. URL: <http://dx.doi.org/10.1038/nmeth.3252>.
- [2] Lucia A Hindorff et al. “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.23 (June 2009), pp. 9362–9367. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0903103106. URL: <http://dx.doi.org/10.1073/pnas.0903103106>.
- [3] Navneet Matharu and Nadav Ahituv. “Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease”. en. In: *PLoS Genet.* 11.12 (Dec. 2015), e1005640. ISSN: 1553-7390, 1553-7404. DOI: 10.1371/journal.pgen.1005640. URL: <http://dx.doi.org/10.1371/journal.pgen.1005640>.
- [4] Valer Gotea et al. “Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers”. en. In: *Genome Res.* 20.5 (May 2010), pp. 565–577. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.104471.109. URL: <http://dx.doi.org/10.1101/gr.104471.109>.

- [5] Menno P Creyghton et al. “Histone H3K27ac separates active from poised enhancers and predicts developmental state”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.50 (Dec. 2010), pp. 21931–21936. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.10160711107. URL: <http://dx.doi.org/10.1073/pnas.10160711107>.
- [6] Nathaniel D Heintzman et al. “Histone modifications at human enhancers reflect global cell-type-specific gene expression”. en. In: *Nature* 459.7243 (May 2009), pp. 108–112. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07829. URL: <http://dx.doi.org/10.1038/nature07829>.
- [7] Axel Visel et al. “ChIP-seq accurately predicts tissue-specific activity of enhancers”. en. In: *Nature* 457.7231 (Feb. 2009), pp. 854–858. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07730. URL: <http://dx.doi.org/10.1038/nature07730>.
- [8] Diego Villar et al. “Enhancer evolution across 20 mammalian species”. en. In: *Cell* 160.3 (Jan. 2015), pp. 554–566. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.01.006. URL: <http://dx.doi.org/10.1016/j.cell.2015.01.006>.
- [9] Fumitaka Inoue et al. “A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity”. en. In: *Genome Res.* 27.1 (Jan. 2017), pp. 38–52. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.212092.116. URL: <http://dx.doi.org/10.1101/gr.212092.116>.
- [10] Jamie C Kwasniewski et al. “High-throughput functional testing of ENCODE segmentation predictions”. en. In: *Genome Res.* 24.10 (Oct. 2014), pp. 1595–1602. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.173518.114. URL: <http://dx.doi.org/10.1101/gr.173518.114>.
- [11] Ryan Tewhey et al. “Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay”. en. In: *Cell* 165.6 (June 2016), pp. 1519–1529. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.04.027. URL: <http://dx.doi.org/10.1016/j.cell.2016.04.027>.
- [12] Jacob C Ulirsch et al. “Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits”. en. In: *Cell* 165.6 (June 2016), pp. 1530–1545. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.04.048. URL: <http://dx.doi.org/10.1016/j.cell.2016.04.048>.
- [13] Susan Q Shen et al. “Massively parallel cis-regulatory analysis in the mammalian central nervous system”. en. In: *Genome Res.* 26.2 (Feb. 2016), pp. 238–255. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.193789.115. URL: <http://dx.doi.org/10.1101/gr.193789.115>.
- [14] Fumitaka Inoue et al. “Massively parallel characterization of regulatory dynamics during neural induction”. en. July 2018. URL: <https://www.biorxiv.org/content/early/2018/07/16/370452>.

- [15] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. en. In: *Genome Biol.* 15.12 (2014), p. 550. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0550-8. URL: <http://dx.doi.org/10.1186/s13059-014-0550-8>.
- [16] Cynthia A Kalita et al. “QuASAR-MPRA: Accurate allele-specific analysis for massively parallel reporter assays”. en. In: *Bioinformatics* (Sept. 2017). ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btx598. URL: <http://dx.doi.org/10.1093/bioinformatics/btx598>.
- [17] Leslie Myint et al. “Linear models enable powerful differential activity analysis in massively parallel reporter assays”. en. Sept. 2017. URL: <https://www.biorxiv.org/content/early/2017/09/30/196394>.
- [18] Kotaro J Kaneko et al. “Transcription factor TEAD2 is involved in neural tube closure”. en. In: *Genesis* 45.9 (Sept. 2007), pp. 577–587. ISSN: 1526-954X. DOI: 10.1002/dvg.20330. URL: <http://dx.doi.org/10.1002/dvg.20330>.
- [19] Wen-Teng Chang et al. “A novel function of transcription factor alpha-Pal/NRF-1: increasing neurite outgrowth”. en. In: *Biochem. Biophys. Res. Commun.* 334.1 (Aug. 2005), pp. 199–206. ISSN: 0006-291X. DOI: 10.1016/j.bbrc.2005.06.079. URL: <http://dx.doi.org/10.1016/j.bbrc.2005.06.079>.
- [20] Gil Stelzer et al. “The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses”. en. In: *Curr. Protoc. Bioinformatics* 54 (June 2016), pp. 1.30.1–1.30.33. ISSN: 1934-3396, 1934-340X. DOI: 10.1002/cpbi.5. URL: <http://dx.doi.org/10.1002/cpbi.5>.
- [21] Charles E Grant, Timothy L Bailey, and William Stafford Noble. “FIMO: scanning for occurrences of a given motif”. en. In: *Bioinformatics* 27.7 (Apr. 2011), pp. 1017–1018. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btr064. URL: <http://dx.doi.org/10.1093/bioinformatics/btr064>.
- [22] Casey A Gifford et al. “Transcriptional and epigenetic dynamics during specification of human embryonic stem cells”. en. In: *Cell* 153.5 (May 2013), pp. 1149–1163. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2013.04.037. URL: <http://dx.doi.org/10.1016/j.cell.2013.04.037>.
- [23] Alexander M Tsankov et al. “Transcription factor binding dynamics during human ES cell differentiation”. en. In: *Nature* 518.7539 (Feb. 2015), pp. 344–349. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14233. URL: <http://dx.doi.org/10.1038/nature14233>.

Figures

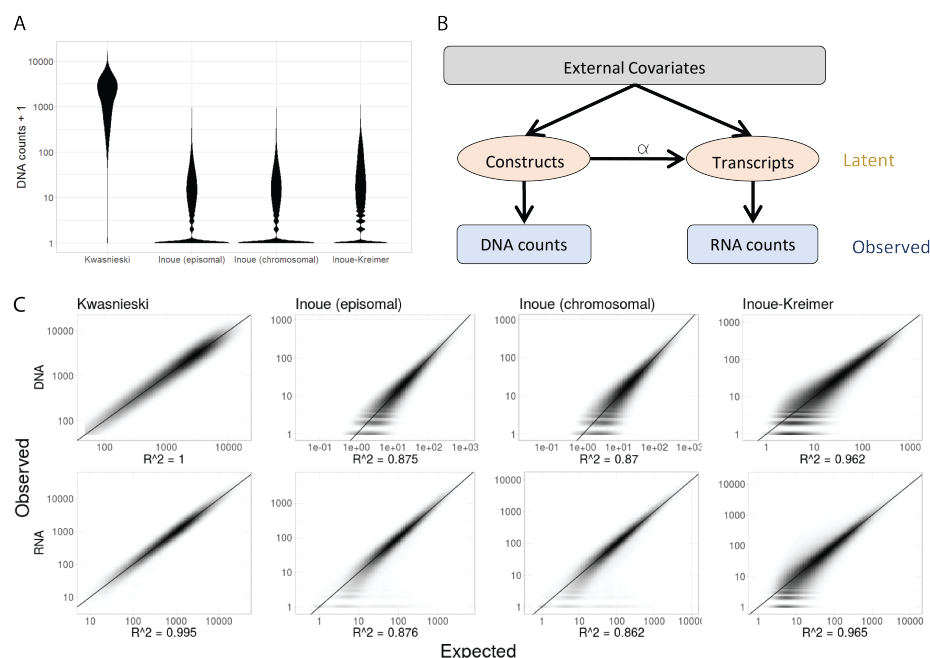


Figure 1: MPRAnalyze model properties and fit. **(A)** Distribution of construct abundances (DNA barcodes) across datasets, computed as the observed barcode count + 1 for visualization purposes. **(B)** A graphical representation of the MPRAnalyze model. External covariates (e.g conditions of interest, batch effects, barcode effects) are design-dependent; Latent construct and transcript counts are related by the transcription rate α . **(C)** Goodness of fit plots for both libraries across datasets. Expected counts were extracted from the fitted GLMs. MPRAnalyze's model fits MPRA data well, with $R^2 > 0.86$ across all datasets.

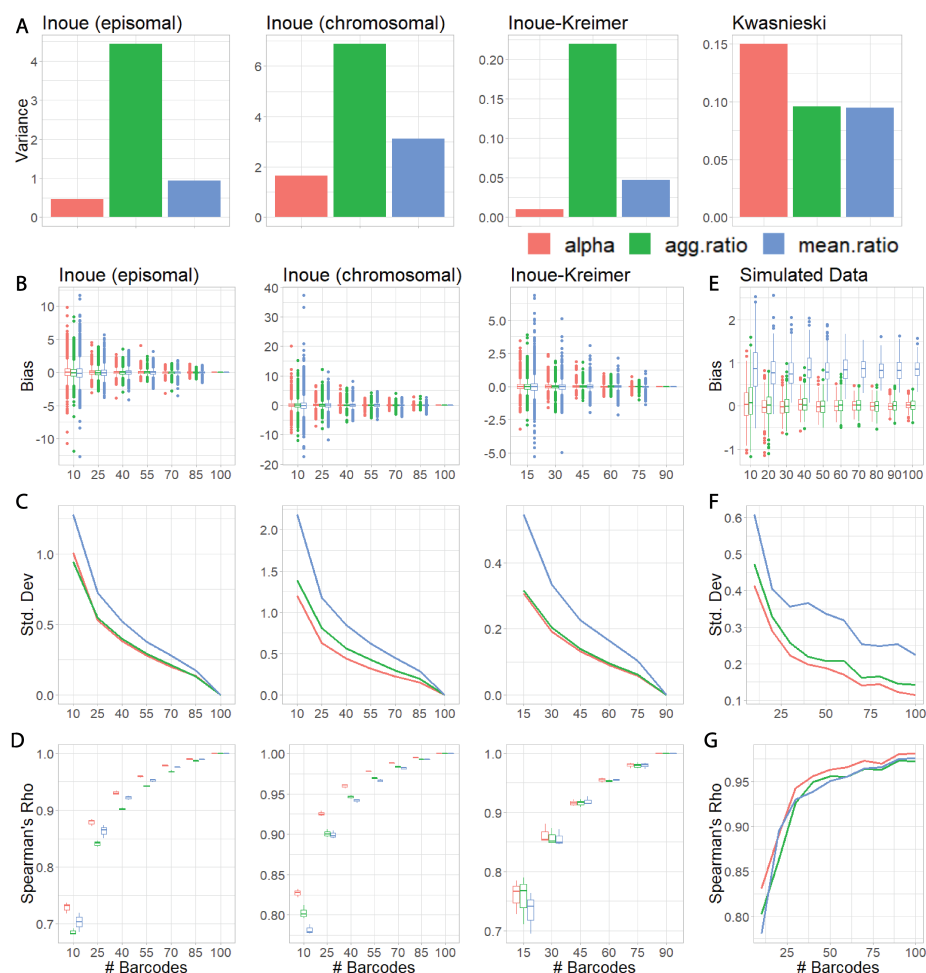


Figure 2: Comparison of MPRAnalyze's α estimate of transcription rate with the naive ratio-based estimates. **(A)** The variance measured among estimates of negative-control enhancers in each dataset (these are assumed to have an identical transcription rate). **(B-D)** Barcodes were subsampled and quantification was recomputed based on the partial data to measure the effect of barcode number on estimate performance [See methods for further subsampling details]. Analyses were performed using the full-data estimate as the ground truth. **(E-G)** MPRA data was simulated to provide an actual ground truth. In each case we measured the bias ($estimate - truth$) (**B,E**); the standard deviation ($\sqrt{Var(estimate - truth)}$) (**C,F**); and the Spearman correlation between the estimates and the ground truth (**D,G**).

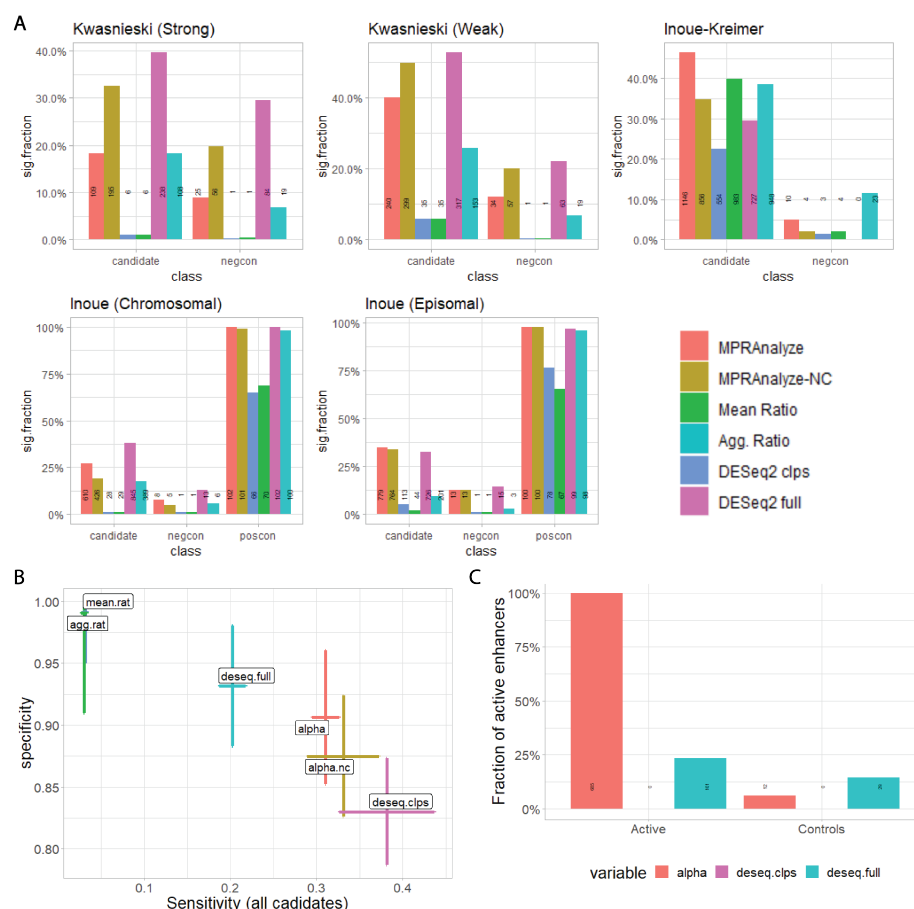


Figure 3: Classification analysis comparisons. **(A)** fraction of enhancers identified as significantly active (BH-corrected $P < 0.05$) by method and class of enhancer. MPRAnalyze results both in control-based (red) and no-controls (orange) modes; empirical p-values based on the mean ratio (blue) or aggregated ratio (green); DESeq2 results in collapsed mode (barcodes are summed within each batch, purple) or full mode (full data, light blue). Absolute number of active enhancers is displayed on the bars. **(B)** Precision-Recall curve. Precision is based on performance on the negative controls, Recall is based on the total population of enhancers, assuming all candidates are active. Error bars are \pm the standard deviation of these measures across datasets. **(C)** Fraction of active enhancers detected after re-running the analyses on 685 enhancers from the Inoue-Kreimer dataset that were identified as active by MPRAnalyze (regular mode) and both DESeq2 modes, and the 200 controls from the same dataset. MPRAnalyze recapitulates the same results, finding that 100% of the candidates are active, whereas DESeq2 full only identifies 161 (23.5%) and DESeq2 collapsed completely fails to identify any active enhancers.

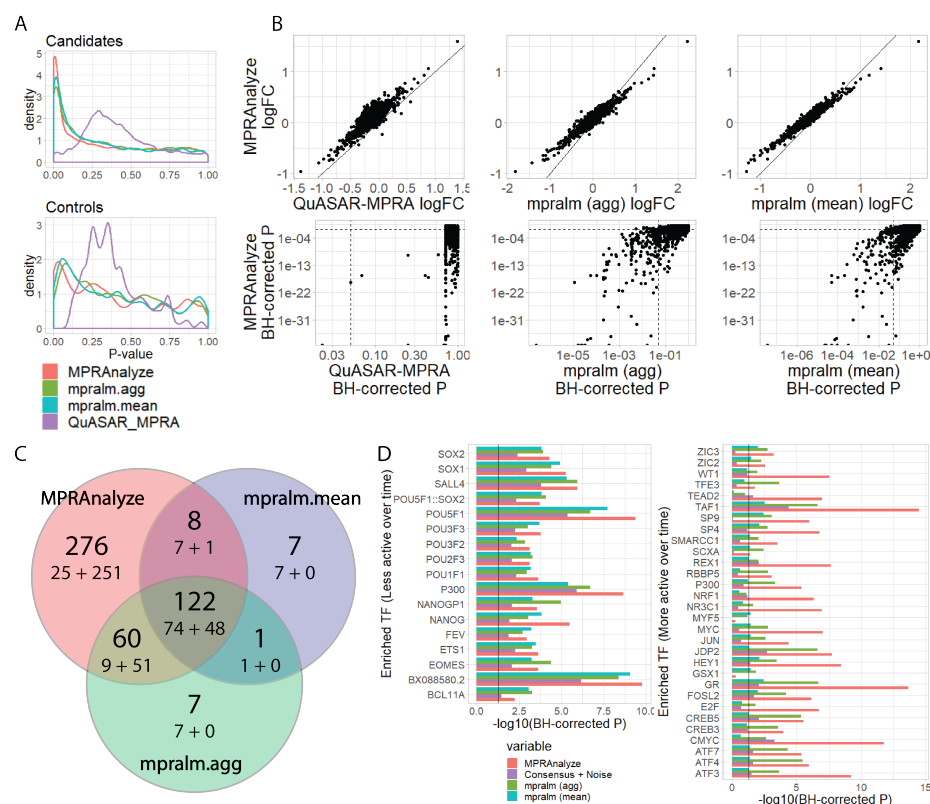
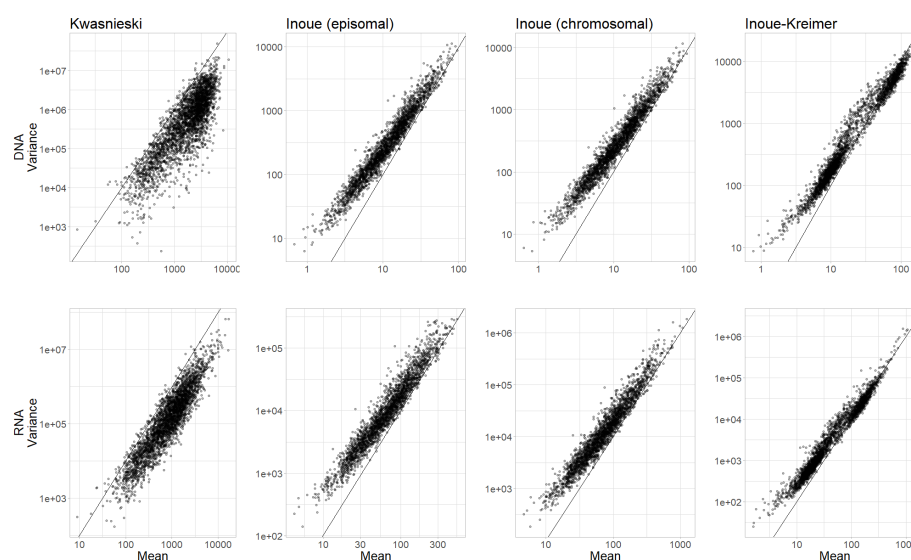


Figure 4: Comparative analysis results of comparing timepoint 0h to 72h in the Inoue-Kreimer dataset. **(A)** P-value distributions of candidates (top) and controls (bottom). QuASAR-MPRA is poorly calibrated, whereas MPRAnalyze and both mpralm modes follow the theoretical behavior (mixture of uniform and low values). **(B)** Direct comparison of MPRAnalyze to competing methods. Top panels show the biological effect size (log Fold-change); Bottom panels show the statistical significance (BH-corrected P; dotted lines are 0.05 threshold). **(C)** Venn diagram for MPRAnalyze and mpralm (both modes). The numbers in each area are (top) the total number of enhancers in the area, and (bottom) the number of decreasing-activity enhancers (left) + and increasing-activity enhancers (right). **(D)** Enrichment of transcription factor binding sites in differentially active enhancers as determined by each method. Solid line represents threshold of 0.05. (see Methods for further details).

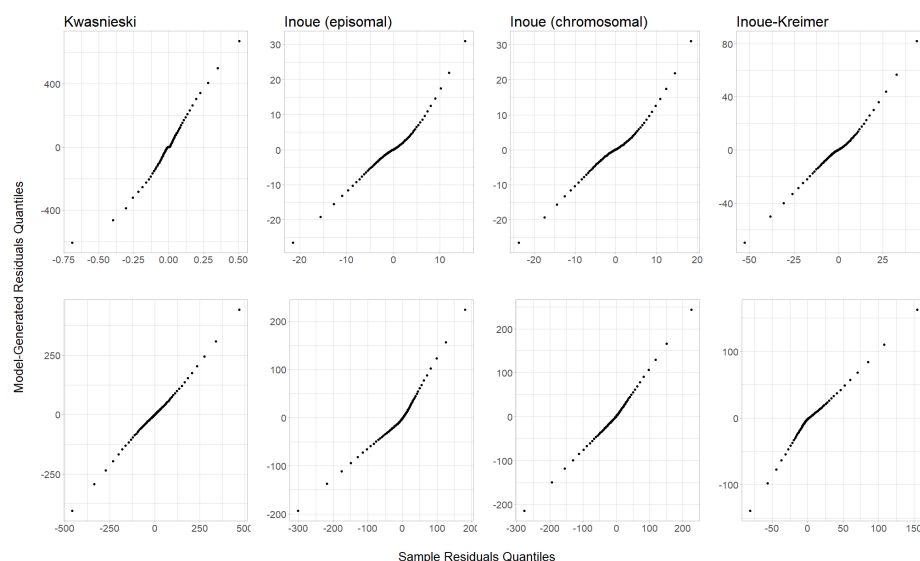
Sample	Condition	Barcode
A	Ref	1
B	Ref	2
C	Ref	3
D	Contrast	1
E	Contrast	2
F	Contrast	3

$$\begin{aligned}
 &\Rightarrow \log \begin{pmatrix} \hat{d}_A \\ \hat{d}_B \\ \hat{d}_C \\ \hat{d}_D \\ \hat{d}_E \\ \hat{d}_F \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_{contrast} \\ \beta_{BC_2} \\ \beta_{BC_3} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_0 + \beta_{BC_2} \\ \beta_0 + \beta_{BC_3} \\ \beta_0 + \beta_{contrast} \\ \beta_0 + \beta_{contrast} + \beta_{BC_2} \\ \beta_0 + \beta_{contrast} + \beta_{BC_3} \end{pmatrix} \\
 \\
 &\Rightarrow \log \begin{pmatrix} \hat{r}_A \\ \hat{r}_B \\ \hat{r}_C \\ \hat{r}_D \\ \hat{r}_E \\ \hat{r}_F \end{pmatrix} = \log(\hat{d}) + \log \begin{pmatrix} \alpha_{ref} \\ \alpha_{ref} \\ \alpha_{ref} \\ \alpha_{contrast} \\ \alpha_{contrast} \\ \alpha_{contrast} \end{pmatrix} = X_D \cdot \vec{\beta} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \gamma_0 \\ \gamma_{contrast} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_0 + \beta_{BC_2} \\ \beta_0 + \beta_{BC_3} \\ \beta_0 + \beta_{contrast} \\ \beta_0 + \beta_{contrast} + \beta_{BC_2} \\ \beta_0 + \beta_{contrast} + \beta_{BC_3} \end{pmatrix} + \begin{pmatrix} \gamma_0 \\ \gamma_0 \\ \gamma_0 \\ \gamma_0 + \gamma_{contrast} \\ \gamma_0 + \gamma_{contrast} \\ \gamma_0 + \gamma_{contrast} \end{pmatrix}
 \end{aligned}$$

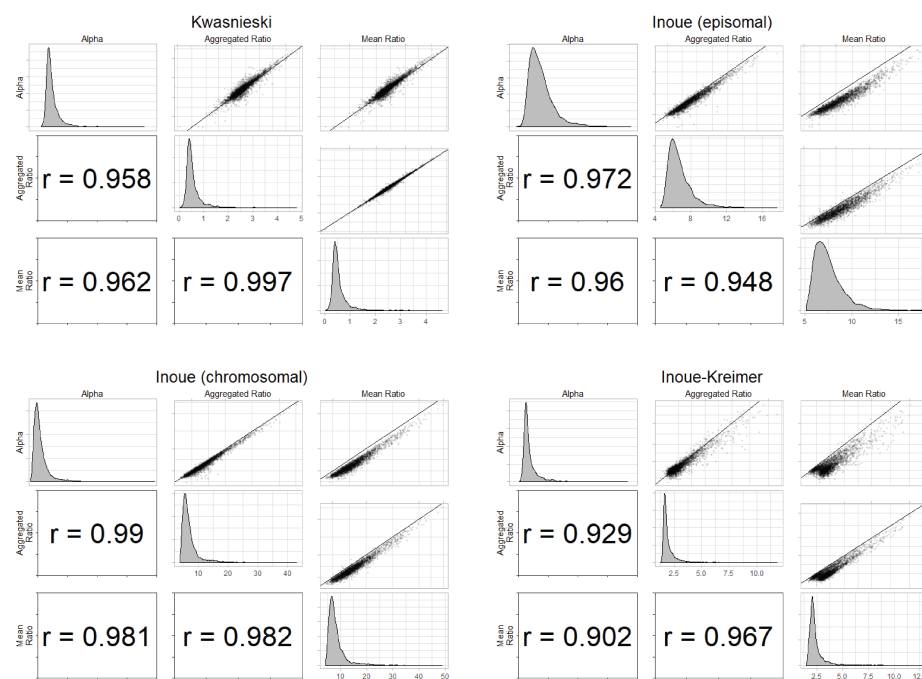
Supplemental Figure 1: A simplified example of the MPRAnalyze model: two conditions are tested with three barcodes in a paired experiment (each DNA observation has a corresponding RNA observation). No replicates or external normalization factors are included in this design to maintain simplicity. The DNA's model estimation of the latent DNA count, computed as $X_D \vec{\beta}$, is included in the RNA model. The α estimates of transcription rate can be extracted from the model as: $\alpha_{ref} = e^{\gamma_0}$, $\alpha_{ref} = e^{\gamma_0 + \gamma_{contrast}}$. Note that while modeling the barcodes in the RNA model is possible, the result will be a separate α estimator for each barcode, which is usually not desired. Barcode-level information is therefore only incorporated into the RNA model via the nested DNA model.



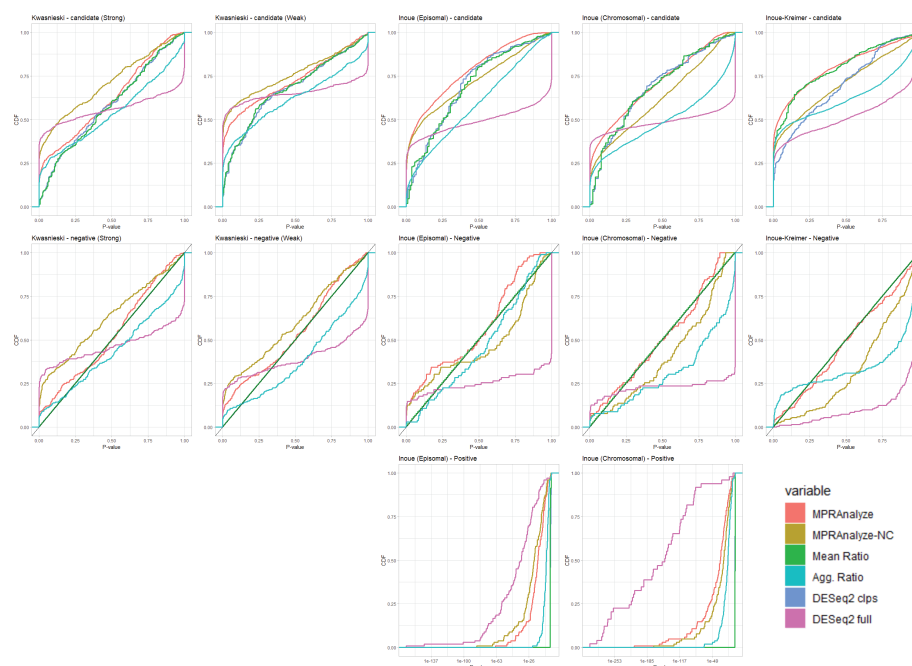
Supplemental Figure 2: Relationship between the mean and variance of the counts measured for each enhancer. Reference line (slope = 2) is a quadratic relationship.



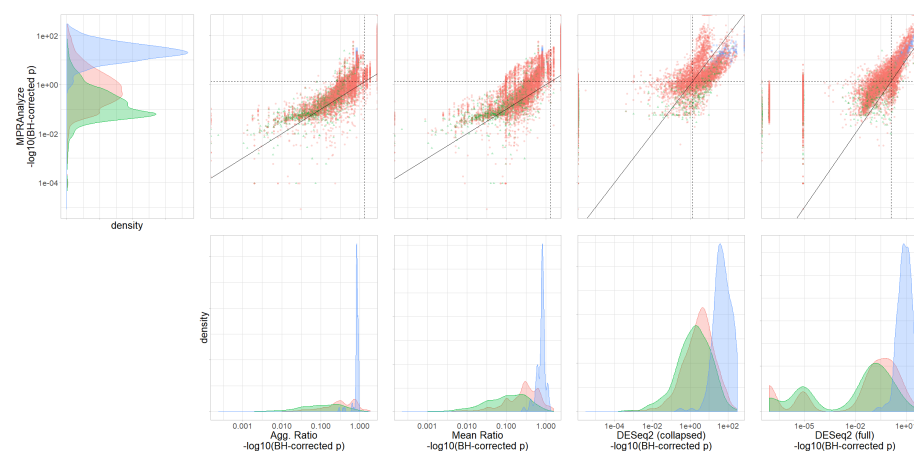
Supplemental Figure 3: Comparison between model residuals from the observed counts of each dataset, and residuals from data generated by Gamma (for DNA) and Negative Binomial (for RNA) using the model parameters. Quantile-quantile comparisons indicate that the observed noise and the generated noise follow similar distributions.



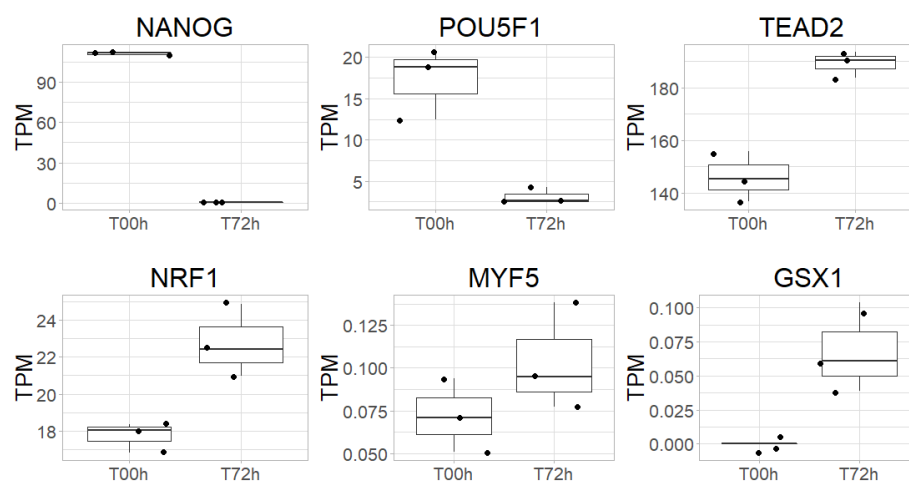
Supplemental Figure 4: Correlations of MPRAnalyze's *alpha* estimate with the naive ratio-based estimates. Correlations are Pearson's *r*.



Supplemental Figure 5: P-value CDF of classification analysis of enhancers for each dataset, stratified by enhancer type. top panels are candidate enhancers; middle are negative controls, with a reference line of the theoretical uniform CDF; bottom are positive controls, displayed in log-scale to make them more informative.



Supplemental Figure 6: Distribution of BH-corrected P-values of all methods. Analysis was performed separately for each dataset, then P-values were combined. Positive controls are from the Inoue episomal and chromosomal datasets only (blue). Candidates (red) and negative controls (green) are from the all datasets. Continuous line is a the identity line. Dotted lines are set at 0.05



Supplemental Figure 7: RNA expression levels for transcription factors with enriched binding sites in differentially active enhancers. NANOG and POU5F1, enriched among enhancers that reduce in activity over time, have an expected corresponding reduction in expression over time. TEAD2 and NRF1 are enriched in increasing-activity enhancers according to MPRAnalyze's results and indeed show a corresponding increase in expression. MYF5 and GSX1 are enriched in increasing-activity enhancers according to mpralm, but not according to MPRAnalyze.

Methods

Dataset Collection and Processing

For all datasets included in this paper, we relied on the processing and filtering performed by the authors of the original papers. This ensures that MPRAAnalyze’s performance isn’t reflecting any favorable processing steps we chose. **Kwasnieski:** The study[10] measured the activity of potential regulatory regions in K562 cells. Regions were selected according to ENCODE annotations of four groups: enhancers; weak enhancers; repressed enhancers; enhancers active in ESCs. The repressed and ESC-annotated enhancers were used as controls, and were excluded from the analysis after library size normalization factors were computed. In addition to control classes, each class had internal sets of scrambled sequences used as negative controls, which were used as controls in our analyses. **Inoue:** The study[9] compared activity in HepG2 cells of liver enhancers that were either episomal or chromosomally integrated using a lentivirus (lentiMPRA). While the study is comparative, the comparison is not between biological conditions and the results are therefore difficult to validate or interpret. We therefore decided to use the data as two separate quantification datasets. The datasets were analyzed together to better account for batch and barcode-specific effects, and α estimates were extracted from the joint model for each condition separately. **Inoue-Kreimer:** The study[14] identified enhancers with temporal activity over the first 72 hours after neural induction. lentiMPRA was performed in 7 timepoints (0, 3, 6, 12, 24, 48 and 72 hours after induction). For the purpose of our analysis, we used only the data from the first timepoint in the quantification and classification analyses, and timepoints 0 and 72 hours for the comparative analysis.

Computing Transcription Rate Estimates

All transcription rate estimates were computed for library size normalized MPRA data, using upper quartile normalization to compute size factors. **MPRAAnalyze’s** α was computed for each dataset using the quantification analysis (See supplemental methods). Across datasets, batch and barcode-level effects were modelled in the nested DNA model, but excluded from the RNA model design. This allows MPRAAnalyze to model nuisance effects but asserts that all barcodes associated with a single enhancer must share the same transcription rate. The **Mean Ratio** was computed using only pairs of observations that are both positive, so for each enhancer: $S = \{i \in [n] | R_i \neq 0, D_i \neq 0\}$, then: $Mean\ Ratio = \frac{1}{|S|} \sum_{i \in S} \frac{R_i}{D_i}$. Similarly, the **Aggregated Ratio** was computed using only positive observations, without requiring that both measurements of the pair are positive, so for each enhancer: $S_R = \{i | R_i \neq 0\}$, $S_D = \{i | D_i \neq 0\}$, then $Aggregated\ Ratio = \frac{\frac{1}{|S_R|} \sum_{i \in S_R} R_i}{\frac{1}{|S_D|} \sum_{j \in S_D} D_j}$.

Subsampling analysis

For the subsampling analysis, barcodes were sampled down to varying levels (for Inoue datasets: 15, 30, 45, 60, 75, 90 out of the total 100 barcodes; for Inoue-Kreimer: 15, 30, 45, 60, 75 of the total 90 barcodes). The analysis uses three independent replicates of this down-sampling process, so overall for each enhancer we get a set of $3 \times K$ estimates at various numbers of available barcodes, where $K = 6$ for the Inoue datasets and $K = 5$ for Inoue-Kreimer. The analyses were done on the entire down-sampled dataset in a single run and included the original data as well as the reduced-barcodes data, to neutralize any effect that the library size correction might have on the estimates.

Simulating MPRA data

MPRA data was simulated by generating random coefficients for the nested GLM construction that MPRAAnalyze uses. The *latent (true)* DNA and RNA counts were generated directly from the model, then log-normal noise was added to the latent counts to get the *observed* counts. Formally:

$$\begin{aligned}\vec{\beta} &= [\beta_0, \vec{\beta}_{batch}, \vec{\beta}_{BC}] \\ \beta_0 &\sim N(K, \sigma_0^2) \\ \vec{\beta}_{batch} &\sim N(0, \sigma_{batch}^2) \\ \vec{\beta}_{BC} &\sim N(0, \sigma_{BC}^2) \\ &\Downarrow \\ \vec{D}_{true} &= nint\left(\exp(X_d \vec{\beta})\right) \\ \vec{R}_{true} &= nint\left(\exp(\alpha \cdot X_d \vec{\beta})\right) \\ \vec{D}_{observed} &\sim nint\left(\log - Normal\left(\exp(X_d \vec{\beta}), \sigma_D^2\right)\right) \\ \vec{R}_{observed} &\sim nint\left(\log - Normal\left(\exp(\alpha \cdot X_d \vec{\beta}), \sigma_R^2\right)\right)\end{aligned}$$

where K controls the intercept term for the construct distribution, the variance of which is σ_0^2 ; $\sigma_{batch}^2, \sigma_{BC}^2$ control the size of batch and barcode effects, respectively; σ_D^2, σ_R^2 determine the noise levels added to the data; *nint* is the nearest-integer function, using base R's *round* function. An implementation of this simulation process is included in the MPRAAnalyze package.

Noise was generated using log-normal noise instead of Gamma/Negative Binomial to avoid generating data directly from MPRAAnalyze's model, which might bias the results.

Simulated data in this manuscript was generated with 3 batches, varying numbers of barcodes, $K = 5$, and $\sigma_0 = \sigma_{batch} = \sigma_{BC} = \sigma_D = \sigma_R = 0.5$.

Transcription Factor Binding Site enrichment analysis

The transcription factor binding site enrichment analysis was performed using the binary binding matrix computed by Inoue & Kreimer et al. [14], with each entry indicating the potential for binding (motif-based binding prediction using Fimo[21], $FDR < 10^{-4}$) or overlap with transcription factor ChIP-seq peaks from publicly available data[22, 23].