QuASAR-MPRA: Accurate allele-specific analysis for massively parallel reporter assays.

Cynthia A. Kalita 1, Gregory A. Moyerbrailean 1, Christopher Brown 2, Xiaoquan Wen 3, Francesca Luca 1,4* and Roger Pique-Regi 1,4*

1 Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI
2 Department of Genetics, University of Pennsylvania, Philadelphia, PA
3 Department of Biostatistics, University of Michigan, Ann Arbor, MI
4 Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI

ABSTRACT
Motivation: The majority of the human genome is composed of non-coding regions containing regulatory elements such as enhancers, which are crucial for controlling gene expression. Many variants associated with complex traits are in these regions, and may disrupt gene regulatory sequences. Consequently, it is important to not only identify true enhancers but also to test if a variant within an enhancer affects gene regulation. Recently, allele-specific analysis in high-throughput reporter assays, such as massively parallel reporter assays (MPRA), have been used to functionally validate non-coding variants. However, we are still missing high-quality and robust data analysis tools for these datasets.

Results: We have further developed our method for allele-specific analysis QuASAR (quantitative allele-specific analysis of reads) to analyze allele-specific signals in barcoded read counts data from MPRA. Using this approach, we can take into account the uncertainty on the original plasmid proportions, over-dispersion, and sequencing errors. The provided allelic skew estimate and its standard error also simplifies meta-analysis of replicate experiments. Additionally, we show that a beta-binomial distribution better models the variability present in the allelic imbalance of these synthetic reporters and results in a test that is statistically well calibrated under the null. Applying this approach to the MPRA data by Tewhey et al. (2016), we find 602 SNPs with significant (FDR 10%) allele-specific effects that by having the appropriate data analysis tools, we can greatly improve the power to detect allelic effects in high throughput reporter assays.

Availability: http://github.com/piquelab/QuASAR/mpra
Contact: fluca@wayne.edu; rpique@wayne.edu

1 INTRODUCTION

Genetic variants in non-coding regions are responsible for inter-individual differences in molecular and complex phenotypes. Quantitative trait loci (QTLs) for molecular and cellular phenotypes (Dermitzakis, 2012) have been crucial in providing stronger evidence and a better understanding of how genetic variants in regulatory sequences can affect gene expression levels (Stranger, 2007; Gibbs et al., 2010; Melzer et al., 2008; Cheung et al., 2003; Brem et al., 2002). However, eQTL studies have severe limitations in identifying the true causal variant, due to linkage disequilibrium (LD) limiting the resolution of analysis. The availability of extensive functional annotations (Consortium, 2012; Pique-Regi et al., 2011; Hoffman et al., 2012; Moyerbrailean et al., 2016) enables the integration of functional genomic information into eQTL analysis, which can be useful to dissect the causal variant and the functional basis of the observed associations (Gaffney et al., 2012; Veyrieras et al., 2008; Lee et al., 2009; Lappalainen et al., 2013; Kichaev et al., 2014; Wen et al., 2015; Pickrell, 2014). SNPs that fall within a transcription factor (TF) binding site (TFBS) represent a major mechanism underlying eQTLs (Degner et al., 2012). Recently, additional computational and experimental techniques have been developed to predict and detect allelic effects of SNPs in TFBS using DNase I footprinting and ChiP-seq data (from the ENCODE and Roadmap Epigenome projects) (Moyerbrailean et al., 2016; Lee et al., 2015; Maurano et al., 2015; Zhou and Troyanskaya, 2015). Still, it is a challenge to further validate if allelic effects in binding translate to effects on gene transcription. While all these existing computational annotations are useful for predicting the causal SNP in an eQTL, they do not prove the SNP is truly causal, nor do they properly quantify its effect on gene expression.

To dissect regulatory sequences and compare genetic effects on gene expression, different versions of high throughput reporter assays have emerged in the recent years. These include massively parallel reporter assays (MPRA) Melnikov et al. (2012); Kwasnieski et al. (2012) and self transcribing active regulatory regions sequencing (STARR-seq) Arnold et al. (2013) that can simultaneously measure the regulatory function of thousands of constructs at once. MPRAs utilize a multitude of unique synthesized DNA oligos that are associated with barcodes, cloned in a reporter plasmid and transfected into cells. The transcripts are then isolated for RNA-seq. The number of barcode reads in the RNA over the number of barcode reads from the plasmid DNA is used as a quantitative measure of expression driven by the synthetic enhancer region (Melmikov et al., 2012; Kwasnieski et al., 2012; Patwardhan et al., 2012; Sharon et al., 2012; Kwasnieski et al., 2014). MPRA and STARR-seq were originally created to identify and validate regulatory regions, but they can also be used to compare allelic

*to whom correspondence should be addressed
genomic annotations, such as TF binding motifs, are good predictors to investigate whether the genetic variants that fall within power. Finally, we used the allelic effects identified by QuASAR to then demonstrate that the QuASAR-MPRA test better calibrates the other tests employed in MPRA and ASE analyses (Table 1). We have used barcodes per oligo per replicate) to remove PCR duplicates, making this an ideal dataset to work with. We considered separately sequences in the forward and reverse strand direction in the library, as direction of the regulatory region could potentially affect reporter gene and therefore barcode expression. Tewhey et al. found that filtering the data to remove variants with low coverage greatly reduced the variability between replicates. Higher variance could then lead to falsely identifying ASE. We therefore began processing the dataset by applying a counts filter. For each direction we removed all cases with less than five reads in the reference and alternate allele, and where the sum of two alleles was ≤ 100. This gave us a total of 33,664 SNPs in the DNA library as input to the RNA library. For the RNA library, we first separated the library into forward and reverse directions, and then required that RNA constructs were in the DNA library. We used a counts filter of 5 for both reference and alternate alleles so that we were only looking at variants that had sufficient reads covering both alleles to test for allele-specific effects on expression. This left us with 19,173 SNPs in the forward library and 19,714 SNPs in the reverse library or 33,540 SNPs total represented.

## 2 METHODS

### 2.1 Data source and pre-processing

For the RNA library, we first separated the library into forward and reverse directions, and then required that RNA constructs were in the DNA library. We used a counts filter of 5 for both reference and alternate alleles so that we were only looking at variants that had sufficient reads covering both alleles to test for allele-specific effects on expression. This left us with 19,173 SNPs in the forward library and 19,714 SNPs in the reverse library or 33,540 SNPs total represented.

### 2.2 Baseline statistical methods for comparison

To test for ASE there are several different methods available (Table 1). The t-test, Fisher’s exact test and binomial test are classical tests remarkably appealing due to their simplicity. However, they have several limitations, as they cannot be tuned to the context of the experiment, such as levels of overdispersion (eg. from biological and technical variability) which are known to exist in ASE data (Castel et al., 2015; Skelly et al., 2011; Anders et al., 2010). A paired Student’s t-test for ASE can be used to test whether the mean expression of the reference allele is equal to the mean expression of the alternate allele. This test requires multiple replicates in order to calculate a mean for each allele expression group that has little variance, otherwise the test will not have the power to detect differences. Fisher’s exact test has been used previously to identify ASE (Romandel et al., 2015), by testing whether the reference and alternate allele counts’ proportions are the same. Rejection of the null hypothesis, however, only informs us that the difference between the average counts in the two samples is larger than one would expect between technical replicates. In the binomial test, the null hypothesis is that observed values for two categories do not deviate from the theoretically expected distribution of observations. In ASE, the binomial test is used to determine whether the ratio of the two alleles is significantly different from the expected proportion (e.g. 0.5). This is the classic test that has been employed previously to detect ASE in RNA-seq studies, and assumes that read counts within each gene are binomially distributed (Kilpinen et al., 2013; Consortium et al., 2015; Lappalainen et al., 2013; Buil et al., 2014). Even accounting for reference mapping bias in RNA-seq reads, p-values have been found to remain inflated, especially for very low (< 10) and very high (> 1000) coverage sites (Castel et al., 2015).

To reproduce the Student’s t-test performed by Tewhey et al., we calculated the log2 ratio for the reference and alternate allele constructs (RNA/DNA) for each replicate. These values were used as input for a
paired $t$-test in $R$. To perform the Fisher’s exact test on the MPRA counts data, we first added a pseudocount to each RNA and DNA reference and alternate allele counts and then used the `fisher.test` function in $R$.

To perform the binomial test on the MPRA counts data, we compared the reference and alternate allele counts to the DNA proportion (reference allele/reference allele + alternate allele). To combine the $p$-values for the two LCL individuals, we used Fisher’s method (Tewhey et al., 2016).

### 2.3 QuASAR Approach

QuASAR by default assumes that under the null hypothesis of no allelic imbalance the reference and alternate allele read counts should be at 1:1 ratio. However, in MPRA, the proportion $\tau_l$ of the reference reads is not necessarily 0.5 across all the $l$ genetic variants, due to differences in PCR amplification, as well as cloning and transformation efficiencies. Here, we have extended QuASAR to test for differences between the proportion of reference reads in DNA $\tau_l$ and the proportion obtained from RNA reads $\rho_l$.

To reject the null hypothesis $\rho_l = \tau_l$, we extend QuASAR’s beta-binomial model. The observed reference $R_l$ and alternate $A_l$ allele read counts at a given $l$ are modeled as:

$$
Pr \left( R_l | N_l, \psi_l, M_b \right) = \left( \frac{N_l}{R_l} \right) \frac{\Gamma(M_b + \psi_l) \Gamma(A_l + 1 - \psi_l) M_b}{\Gamma(N_l + M_b) \Gamma(\psi_l) \Gamma(A_l + 1 - \psi_l) M_b} (1)
$$

where $N_l = R_l + A_l$ is the total read count at $l$, and $M_b$ is a parameter that controls the effective number of samples supporting the base-calling error $\epsilon$ and the allelic ratio $\rho_l$ parameter in the model. We can estimate $\epsilon$ using an EM procedure (Harvey et al., 2014), but here for MPRA we fixed $\epsilon = 0.001$ as conservative estimate of the true error rate.

Another key difference with our previous implementation of QuASAR is that we use different $M_b$ parameters depending on the sequencing depth $N_l$. We bin $N_l$ into different quantiles (here deciles) and we estimate $M_b$ for each bin separately using a grid search:

$$
M_b = \arg \max_{M_b} \left\{ \prod_{l=1}^{L} Pr \left( R_l | N_l, \psi_l, \rho_l = \tau_l, M_b \right) \right\} (3)
$$

We estimate $\rho_l$ using (1) with $M_b = M_b$ from (3) and a standard gradient method (L-BFGS-B) to maximize the log-likelihood function

$$
l(\rho_l; M_b, \epsilon) = \log Pr \left( R_l | N_l, \psi_l = \psi(\rho_l, \epsilon), M_b \right) (4)
$$

Finally, all parameters are used to calculate the LRT statistic, contrasting $H_1: \rho_l = \rho_0$ to $H_0: \rho_l = \tau_l$ and the resulting $p$-value.

### 2.4 QuASAR meta-analysis

Using the QuASAR approach, we can generate summary statistics of the allelic imbalance that can be used for downstream analyses. For example, to compare DNA to RNA, or between RNA of different cell-types, or to perform meta-analysis of multiple MPRA libraries. Instead of using an estimate of the allelic proportion $\rho_l$, in the QuASAR approach we report the estimate of $\beta_l = \log(\rho_l/(1 - \rho_l))$ and its standard error using the second derivative (i.e. Hessian) of the log-likelihood function in (4). We prefer the transformed parameter $\beta_l$ as it provides a more robust fit and the second derivative is better behaved than that of $\rho_l$ on the edges.

To illustrate this for the Tewhey et al. data, we combine the summary statistics for the two LCL individuals using standard fixed effects meta-analysis of the two replicates:

$$
w_{n,l} = 1/\hat{a}_n^2 \quad w_l^2 = \sum_{n} w_{n,l} (5)
$$

$$
\beta_l^2 = \frac{1}{w_l} \sum_{n} \beta_{n,l} \times w_{n,l} \quad \sigma_l^2 = \frac{1}{w_l} \frac{\hat{a}_l^2}{\tau_l} (6)
$$

$$
Z_l = \frac{\beta_l^2 - \hat{b}_0}{\sigma_l^2}, \quad \beta_0 = \log \frac{\tau_l}{1 - \tau_l}, \quad p = 2\Phi(\pm |Z_l|) (7)
$$

Across all the paper, $p$-values were corrected for multiple testing using the Benjamini-Hochberg’s (BH) method (Benjamini and Hochberg, 1995). To compare the different approaches we quantify the genomic inflation parameter, $\lambda$, for a set of $p$-values (Yang et al., 2011). For this we calculated the ratio of the median of the $p$-value distribution to the expected median, thus quantifying the extent of the bulk inflation and the excess false positive rate.

### 2.5 Annotation Overlap

Table S1 reports the annotations we have considered with their sources. More specifically, we considered two major sets of annotations: experimentally and computationally derived. For experimental annotations we used molecular QTL and allele-specific hypersensitivity (ASH) data. For LCL dsQTLs (Lee et al., 2015) the score was required to be less than the 1st or greater than the 4th quartile. GTEx eQTLs (Melé et al., 2015) were required to be the lead SNP for the gene. ASH SNPs were from (Moyerbrailean et al., 2016).

A variety of different methods have been used recently to computationally predict the allelic effect of SNP on TF binding and chromatin accessibility. GKM-svm (Lee et al., 2015) uses gapped k-mer frequencies to predict the activity of larger functional genomic sequence elements, including the impact of a variant on DNase I sensitivity. It utilizes support vector machinery based on the structural risk minimization principle from statistical learning theory and kernel function which calculates the similarity between any two sequences. CATO (Maurano et al., 2015) quantifies the effect of SNPs on the energy of TF binding, through overlapping SNP DHS profiles with TF motifs and applying a logistic model which takes into account site dependent features and phylogenetic conservation. DeepSEA (Zhou and Troyanskaya, 2015) uses TF binding, DHS, and histone-mark profiles with genomic sequence information as input for training a deep learning-based algorithm and predict the effects that sequence alterations have on the chromatin. DeepSEA has three major features: integrating sequence information from a whole sequence context, learning sequence code at multiple spatial scales with a hierarchical architecture, and multitask joint learning of diverse chromatin factors sharing predictive features. For computational annotations we set the following thresholds. For effect-SNPs, the absolute motif score was required to be $> 3$. To run GKM-svm (Lee et al., 2015), we extracted sequences around MPRA variants (19bp total) and then ran the reference vs alternate allele sequences with either the GM12878 or HepG2 weights. We then used a threshold of $< -6$ or $> 6$ for the variant scores. DeepSEA (Zhou and Troyanskaya, 2015) variant scores were identified using the website tool with a vcf file input (containing the MPRA variants). The functional significance predictions have a threshold of $< 0.05$.

We overlapped SNPs from MPRA counts data with each annotation type. To identify particular annotations that predict the ASE found in the MPRA, we built logistic models $\log(\rho_l/(1 - \rho_l)) = \beta_0 + \beta_l \times n_l$ using the QuASAR significant $p$-values ($p < 0.001$) as the observed binary outcome, and the genomic annotations $n_l$ as the predictor.
3 RESULTS

3.1 Applying QuASAR-MPRA to identify ASE

We used the method proposed here, QuASAR-MPRA, to detect ASE in the MPRA data collected by Tewhey et al. In MPRA, ASE is defined as the departure in the RNA reads from the DNA proportion (the input allelic ratio). Because strand orientation may affect the enhancer function of the sequences tested, each SNP was tested for ASE in the two strand orientations separately (forward/reverse). The two LCL biological replicates were combined using meta-analysis (See Methods). The number of SNPs with significant ASE (10% FDR) were 309 (forward) and 293 (reverse) in LCLs (Table S2), 85 (forward) and 84 (reverse) in HepG2 (Table S3). We then compared these results to those obtained using other methods previously used for MPRA/ASE analysis (Figure 1) using the same input file with the same pre-processing filters (see Methods).

While some of the other methods seem to identify a larger number of SNPs with significant ASE, the distribution of p-values (Figure 1) shows that those methods have very skewed distributions. The majority of genetic variants tested are expected to have no impact and only those that were the truly causal eQTL SNP should have a significant p-value. We do not know a priori which variants have ASE, but in Figure 1 we would expect that the majority of p-values would follow the expected uniform distribution if the approach correctly models the data under the null hypothesis. In other words, only a fraction of MPRA constructs are expected to have significant allelic effects. To better quantify the departure from the expected distribution of p-values for each testing method we used the genomic inflation method. In this method, a greater departure from a lambda value of 1 corresponds to greater inflation in the test results (see Supplement for reverse oligo results). Based on the genomic inflation value λ, QuASAR-MPRA results in the lowest inflation, with λ = 1.161, while the binomial test produces the greatest inflation, with λ = 57.953. A paired t-test with independent estimation of variance and Welch’s adjustment, as in Tewhey et al, results in λ = 2.886; while Fisher’s exact test, as in (Vockley et al, 2015) results in λ = 38.680. Alternatively, we considered the p-value distributions only for the SNPs not predicted to affect TF binding (non-effect SNPs), as these SNPs are more likely to be true negatives. In Figure 1 (and S2) we see that the two methods with lowest lambda values show an even lower departure from the null, consistent with the computational method correctly predicting a large number of true positives. These results show that a beta-binomial distribution (generated using QuASAR-MPRA) better models the variability present in the allelic imbalance of these synthetic reporters and results in a test that produces p-values that are well calibrated under the null hypothesis.

3.2 Validation of experimental and computational annotations for functional non-coding variants

High-throughput reporter assays can be used not only to fine-map causal variants in both GWAS and eQTL studies, but also to validate SNP functional annotations (Kwasnieski et al., 2014). Here we take advantage that the p-values derived from QuASAR are well calibrated under the null hypothesis to examine enrichments for low p-values in both experimentally and computationally derived annotations for allele-specific effects on TF binding. The experimentally derived annotations included LCL dsQTLs (Degner et al., 2012), allele-specific hypersensitivity (ASH) SNPs (Moyerbrailean et al., 2016), and GTEx eQTLs (Melé et al., 2015). In both LCLs and HepG2, ASH SNPs had the greatest departure from the null, followed by LCL dsQTLs (Figure 2 and S1).

We then asked which computational annotations seem to be the most complete and accurate predictors of the effect of a sequence variant on binding as validated by MPRA. We considered effect-SNPs active in LCLs or HepG2 (Moyerbrailean et al.,...
Fig. 2. Validating experimental annotations. QQ plot depicting the p-value distributions from testing for ASE using QuASAR, overlapping with experimental genomic annotations in LCLs.

2016), non-effect SNPs (negative control) (Moyerbrailean et al., 2016), predicted functional SNPs from CATO (Maurano et al., 2015), GKM-svm (Lee et al., 2015) (a gapped kmer sequence-based computational method to predict the effect of regulatory variation), and DeepSEA (Zhou and Troyanskaya, 2015) (predicts genomic variant effects at the variant position using deep learning-based algorithmic framework). Each of the functional annotations show marked differences in p-value distribution. As expected, SNPs in active TF footprints, but not predicted to affect binding, show no departure from the overall distribution. In both LCLs and HepG2, CATO and GKM-svm SNPs had the greatest departure from the null, closely followed by effect-SNPs (Figure 3 and S3).

However, effect-SNPs annotated a considerably larger number of SNPs for both cell-types and were also able to predict cell type-specific effects. LCL effect-SNPs in LCLs had a p-value distribution with a greater departure from the null than the HepG2 effect-SNPs, whereas HepG2 effect-SNPs in HepG2 had a p-value distribution with a greater departure from the null than the LCL effect-SNPs (Figure 4). The differences found here in HepG2 however are minor, potentially due to fewer annotations.

Finally, to formally quantify which annotations are the best predictors of the ASE found in the MPRA, we used all experimental and computational annotations within a logistic model to predict which SNPs in the MPRA data have a nominally significant QuASAR p-value (p < 0.001). The top predictors were GKM-svm SNPs (p < 2 × 10⁻¹⁶) and effect-SNPs (p = 2.17 × 10⁻¹⁵) in LCLs (Table S4). In HepG2, effect-SNPs were the greatest predictor (p = 1.18 × 10⁻¹⁰) (Table S5).

4 DISCUSSION

High throughput reporter assays have proven extremely useful for the experimental validation of enhancer regions. The recent adaptation of MPRA to investigate ASE additionally allows for validation of regulatory variants in TF binding sites, which have been shown to be functionally relevant to fine map eQTLs and GWAS signals. These large datasets, however, require analysis methods to handle the intrinsic overdispersion resulting from the original plasmid proportions, variability in the allelic imbalance, and base-calling errors. Our QuASAR-MPRA approach identifies causal regulatory variants from high-throughput reporter assays by taking into account overdispersion present in the data. This results in a test we have shown to be well calibrated, with minimal inflation, as determined by lambda values close to 1. In addition to being a robust method to identify ASE in high throughput reporter assays, this method produces betas and standard errors for each SNP, which can be used in the fixed effects method to easily combine datasets. Additionally, we still retain a large number of discoveries (FDR 10%) compared to the original MPRA study (441 at 10%FDR) in LCLs.

Finally, we show that the allele-specific regulatory functions identified with QuASAR-MPRA can be used to validate genomic annotations as predictors for allele-specific effects. Knowing which annotations are the best predictors can aid in identifying true causal SNPs. Here we find that LCL dsQTLs and CATO, GKM-svm, and effect-SNPs are significantly predictive of ASE. Using genomic
annotations can additionally help us assign mechanism of action to these regulatory variants. If a variant impacts a TF binding site for example, this can lead to gene expression changes, and therefore phenotype.

Here we have used QuASAR-MPRA on an MPRA dataset, however this method can also be used for other high-throughput reporter assays, such as the ones derived from the STARR-seq protocol (e.g., POP-STARR-seq) (Wockley et al., 2015) and CRE-seq protocols (Kwasnieski et al., 2012), and in the context of high-throughput mutagenesis experiments. As the quest for functional validation of regulatory variants becomes more and more widespread, these high throughput reporter assays, when combined with a robust statistical test, represent a unique resource to functionally characterize genetic variants at an unprecedented and expandable scale.

ACKNOWLEDGEMENT

We would like to thank Wayne State University HPC Grid for computational resources and members of the Luca/Pique group for helpful comments and discussions.

Funding: NIH 1R01GM109215-01 (RPR, FL) and 17PRE33460295 (CK)

REFERENCES


