DRAFT

Vol. 00 no. 00 2017
Pages 1–10

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 MPRAs. 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 regulatory function in LCLs. We also show that we can combine MPRA with QuASAR estimates to validate existing experimental and computational annotations of regulatory variants. Our study shows 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 (Melnikov *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

Table 1. Statistical methods for ASE and MPRA analysis.

	Type of test				
Conditions	T	Fisher	Bin	β-bin	QuASAR
Previously used in MPRA	X	X			
Previously used for ASE			X	X	X
Requires normally distributed data	X				
Underestimates the effect of biological variability	X	X	X		
Handles overdispersion				X	X
Accounts for base calling error					X

effects of genetic polymorphisms. Recent studies that used this technique to compare allelic variants of SNPs with the aim to dissect, at a large scale, the causal nucleotide in eQTL and Genome Wide Association Study (GWAS) signals. Specifically, (Vockley *et al.*, 2015) used a STARR-seq derived method (POP-STARR-seq) to measure allelic effects on gene expression for population based variation in 104 regulatory regions, and a more recent study by (Tewhey *et al.*, 2016) adapted MPRA to fine-map variants associated with gene expression in lymphoblastoid cell lines (LCLs) and HepG2.

The application of MPRA to quantify the allelic effects of regulatory variants is very similar to the challenge posed by allele-specific expression (ASE) in RNA-seq data. However, one key difference is the proportion of plasmids for each allelic construct may not be in a 1:1 ratio. Few off-the-shelf statistical methods have been used for processing and analyzing these large MPRA datasets (Table 1), but they do not consider several technical issues that can lead to false positives, such as base-calling error and over-dispersion. As demonstrated in RNA-seq ASE approaches,a binomial distribution fails to account for overdispersion and results in overly optimistic p-values, while a beta-binomial distribution is a more adequate choice (Kumasaka *et al.*, 2015; van de Geijn *et al.*, 2015). Compared to RNA-seq ASE methods that combine all reads across haplotypes, in MPRA we do not need to accommodate for the uncertainty in phasing or haplotyping as the sequence of each construct is known. Here we further extend QuASAR (Harvey *et al.*, 2014), an approach which considers both over-dispersion and base-calling errors, to test for allelic imbalance in MPRA constructs when the default proportions are not equal. The new method allows for independent estimates of the dispersion parameter depending on variant-specific read coverage, and produces summary statistics that are easy to incorporate in downstream analyses.

Here we tested our new method on MPRA data from Tewhey *et al.*. First we compared our new QuASAR-MPRA statistical test to other tests employed in MPRA and ASE analyses (Table 1). We then demonstrate that the QuASAR-MPRA test better calibrates the p-values under the null hypothesis, without sacrificing statistical power. Finally, we used the allelic effects identified by QuASAR-MPRA to investigate whether the genetic variants that fall within genomic annotations, such as TF binding motifs, are good predictors for allele-specific regulatory function. Our study shows the potential value of using robust allele-specific analysis in high throughput reporter assays, to improve fine mapping analysis of association signals and validate genomic annotations of regulatory variants.

2 METHODS

2.1 Data source and pre-processing

We downloaded processed read counts from GEO (GSE75661) `ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE75nnn/GSE75661/suppl/GSE75661_79k_collapsed_counts.txt.gz` (Tewhey *et al.*, 2016). This MPRA study was designed to look at ASE in 39,479 oligo pairs representing 3,642 eQTLs from the GEUVADIS RNA-seq dataset of lymphoblastoid cell lines (LCLs) from European and African individuals (Lappalainen *et al.*, 2013). It has a large number of experimental replicates (8 LCL replicates), and makes use of barcodes (an average of 73 unique 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 on the reference and alternate allele, and where the sum of two alleles was $\leq$ 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.

2.2 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 allelic 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 (Romanel *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 $\log_2$ 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 r_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 r_l and the proportion obtained from RNA reads ρ_l. To reject the null hypothesis $\rho_l = r_l$, 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:

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

$$\psi_l = [\rho_l(1-\epsilon)+(1-\rho_l)\epsilon] \quad (2)$$

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 prior belief that mean proportion is centered around ψ_l, which also incorporates base-calling error ϵ and the allelic ratio ρ_l parameter in the model. We can estimate ϵ using an EM procedure (Harvey *et al.*, 2014), but here for MPRA we fixed $\hat{\epsilon} = 0.001$ 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:

$$\hat{M}_b = \arg\max_{M_b}\left(\prod_{l=1}^{L}\Pr\left(R_l|N_l,\hat{\epsilon},\rho_l=r_l,M_b\right)\right) \quad (3)$$

We estimate $\hat{\rho}_l$ using (1) with $M_b = \hat{M}_b$ from (3) and a standard gradient method (L-BFGS-B) to maximize the log-likelihood function

$$l(\rho_l;\hat{M}_b,\hat{\epsilon}) = \log\Pr\left(R_l|N_l,\psi_l=\psi(\rho_l,\hat{\epsilon}_s),\hat{M}_b\right) \quad (4)$$

Finally, all parameters are used to calculate the LRT statistic, contrasting $H_1 : \rho_l = \hat{\rho}_l$ to $H_0 : \rho_l = r_l$ 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 ρ_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 logistic transformed parameter β_l as it provides a more robust fit and the second derivative is better behaved than that of ρ_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:

$$w_{n,l} = 1/\hat{\sigma}^2_{n,l} \qquad w^*_l = \sum_n w_n, l \quad (5)$$

$$\beta^*_l = \frac{1}{w^*_l}\sum_n \beta_{n,l} * w_{n,l} \qquad \sigma^*_l = \sqrt{1/w^*_l} \quad (6)$$

$$Z_l = \frac{\beta^*_l - \beta_0}{\sigma^*_l}, \beta_0 = \log\frac{r_l}{1-r_l}, \qquad p = 2\Phi(-|Z_l|) \quad (7)$$

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, λ, 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 quartile 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 wide 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 $\log(p_l/(1-p_l)) = \beta_0 + \beta_1 \times a_l$ using the QuASAR significant p-values ($p < 0.001$) as the observed binary outcome, and the genomic annotations a_l as the predictor.

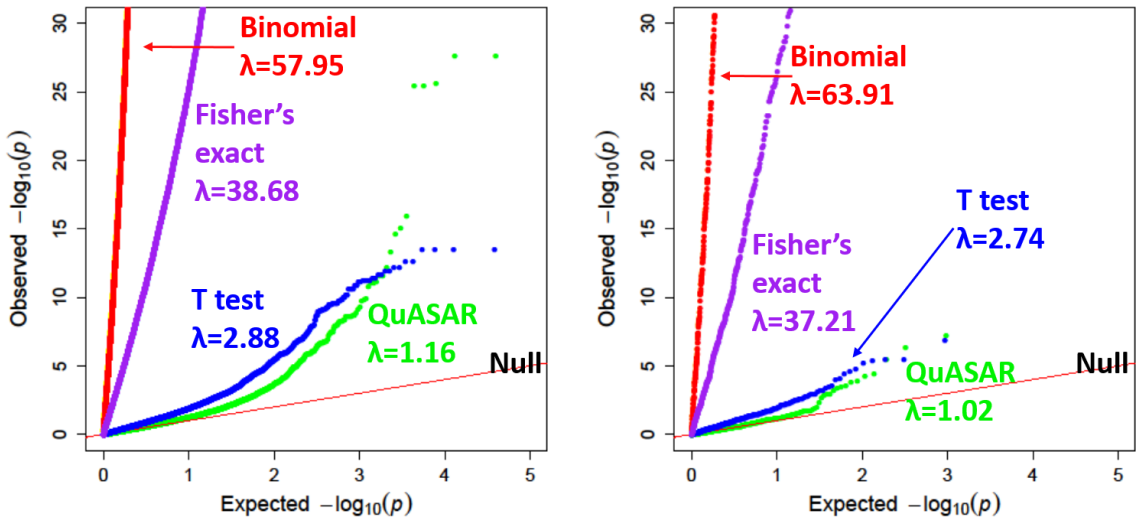

Fig. 1. Comparing ASE testing methods. QQplot depicting the p-value distributions from testing for ASE using four different methods in LCLs with all SNPs (Left) or SNPs predicted to not have any regulatory effect (non-effect SNPs, Right). λ measures genomic inflation deviation from the uniform.

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 $\lambda = 1.161$, while the binomial test produces the greatest inflation, with $\lambda = 57.953$. A paired t-test with independent estimation of variance and Welch's adjustment, as in Tewhey et al, results in $\lambda = 2.886$; while Fisher's exact test, as in (Vockley *et al.*, 2015) results in $\lambda = 38.680$. 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.*,

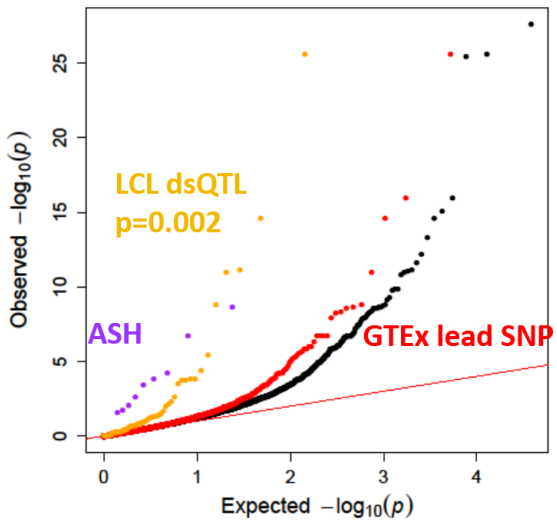

Fig. 2. Validating experimental annotations. QQ plot depicting the p value distributions from testing for ASE using QuASAR, overlapping with experimental genomic annotations in LCLs.

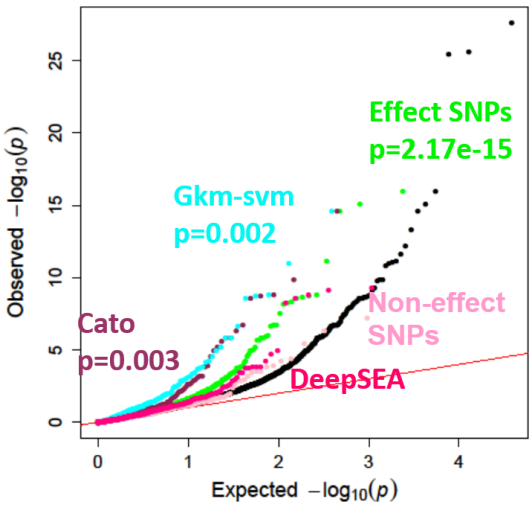

Fig. 3. Validating computational genomic annotations. QQ plot depicting the p-value distributions from testing for ASE using QuASAR, overlapping with computational genomic annotations in LCLs. Effect-SNP scores have a threshold of < -3 or > 3. CATO Maurano *et al.* (2015) prediction scores have a threshold of > 0.1. GKM-svm Lee *et al.* (2015) gapped kmer sequence-based computational method to predict the effect of regulatory variation has a threshold of < -6 or > 6. DeepSEA Zhou and Troyanskaya (2015) predicts genomic variant effects at the variant position using deep learning-based algorithmic framework. The functional significance predictions have a threshold of < 0.05.

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 approach). 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 ($p < 2 \times 10^{-16}$) and effect-SNPs ($p = 2.17 \times 10^{-15}$) in LCLs (Table S4). In HepG2, effect-SNPs were the greatest predictor ($p = 1.18 \times 10^{-10}$) (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 602 (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

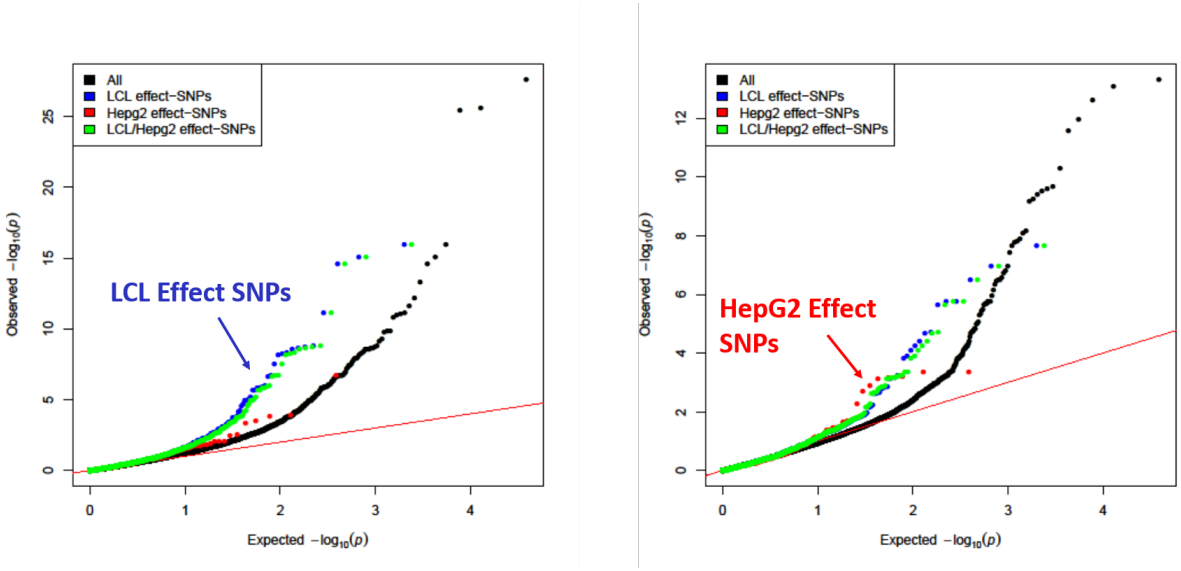

Fig. 4. Identifying cell type effects. QQ plot depicting the *p*-value distributions from testing for ASE using QuASAR, overlapping with effect-SNP annotations in LCLs (Left) and HepG2 (Right). LCL effect-SNPs (blue) are variants in TFs active in LCLs, HepG2 effect-SNPs (red) are variants in TFs active in HepG2, LCL/HepG2 effect-SNPs (green) are variants in TFs active in LCLs or HepG2.

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) (Vockley *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 wide-spread, 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)
AHA 14SDG20450118 (FL) and 17PRE33460295 (CK)

REFERENCES

Anders, S., Huber, W., Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., Mortazavi, A., Williams, *et al.* (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

Arnold, C., Gerlach, D., Stelzer, C., and Boryń, Ł. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (*.

Brem, R., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*.

Buil, A., Brown, A. A., Lappalainen, T., Viñuela, A., Davies, M. N., Zheng, H.-F., Richards, J. B., Glass, D., Small, K. S., Durbin, R., Spector, T. D., and Dermitzakis, E. T. (2014). Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nature Genetics*, **47**(1), 88–91.

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., Lappalainen, T., Adoue, V., Schiavi, A., Light, N., Almlöf, J., Lundmark, *et al.* (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, **16**(1), 195.

Cheung, V., Conlin, L., Weber, T., Arcaro, M., and Jen, K. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature*.

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*.

Consortium, G. *et al.* (2015). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**(6235), 648–660.

Degner, J. F., Pai, A. a., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., and Pritchard, J. K. (2012). DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**(7385), 390–4.

Dermitzakis, E. (2012). Cellular genomics for complex traits. *Nature Reviews Genetics*.

Gaffney, D., Veyrieras, J., and Degner, J. (2012). Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol*.

Gibbs, J., van der Brug, M., and Hernandez, D. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genetics*.

Harvey, C., Moyerbrailean, G., and Davis, G. (2014). QuASAR: Quantitative Allele Specific Analysis of Reads. *Bioinformatics*.

Hoffman, M., Ernst, J., Wilder, S., and Kundaje, A. (2012). Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, **10**(10), e1004722.

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., Core, L. J., Lis, J. T., Hernandez, N., Reymond, A., Deplancke, B., and Dermitzakis, E. T. (2013). Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science*, **342**(6159).

Kumasaka, N., Knights, A. J., and Gaffney, D. J. (2015). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, **48**(2), 206–213.

Kwasnieski, J., Mogno, I., and Myers, C. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the . . . .*

Kwasnieski, J., Fiore, C., Chaudhari, H., and Cohen, B. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome research*.

Lappalainen, T., Sammeth, M., Friedländer, M. R., t Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, *et al.* (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**(7468), 506–511.

Lee, D., Gorkin, D., Baker, M., Strober, B., and Asoni, A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature*.

Lee, S., Dudley, A., Drubin, D., and Silver, P. (2009). Learning a prior on regulatory potential from eQTL data. *PLoS Genetics*.

Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J. A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nature Genetics*, **47**(12), 1393–1401.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., The GTEx Consortium, Wright, F. a., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G., and Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science*, **348**(6235), 660–665.

Melnikov, A., Murugan, A., Zhang, X., and Tesileanu, T. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology*.

Melzer, D., Perry, J., Hernandez, D., and Corsi, A. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics*.

Moyerbrailean, G. A., Kalita, C. A., Harvey, C. T., Wen, X., Luca, F., and Pique-Regi, R. (2016). Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLOS Genetics*, **12**(2), e1005875.

Patwardhan, R., Hiatt, J., Witten, D., and Kim, M. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*.

Pickrell, J. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*.

Pique-Regi, R., Degner, J., Pai, A., and Gaffney, D. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*.

Romanel, A., Lago, S., Prandi, D., Sboner, A., Demichelis, F., Prandi, D., Baca, S., Romanel, A., Barbieri, C., Mosquera, *et al.* (2015). ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Medical Genomics*, **8**(1), 9.

Sharon, E., Kalma, Y., Sharp, A., and Raveh-Sadka, T. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology*.

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research*, **21**(10), 1728–37.

Stranger, B. E. (2007). Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.

Tewhey, R., Kotliar, D., Park, D., Liu, B., Winnicki, S., Reilly, S., Andersen, K., Mikkelsen, T., Lander, E., Schaffner, S., and Sabeti, P. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, **165**(6), 1519–1529.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, **12**(11), 1061–1063.

Veyrieras, J., Kudaravalli, S., and Kim, S. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics*.

Vockley, C., Guo, C., and Majoros, W. (2015). Massively parallel quantification of the regulatory effects of non-coding genetic variation in a human cohort. *Genome research*.

Wen, X., Luca, F., and Pique-Regi, R. (2015). Cross-population joint analysis of eqtls: fine mapping and functional annotation. *PLoS Genet*, **11**(4), e1005176.

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O 'connell, J. R., Mangino, M., Mägi, R., Madden, P. A., Heath, A. C., Nyholt, D. R., Martin, N. G., Montgomery, G. W., Frayling, T. M., and Hirschhorn, J. N. (2011). Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, **16**.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learningbased sequence model. *Nature Methods*, **12**(10), 931–934.