DRAFT

Vol. 00 no. 00 2014
Pages 1–7

QuASAR: Quantitative Allele Specific Analysis of Reads

Chris T. Harvey [1], Gregory A. Moyerbrailean [1], Gordon O. Davis [1], Xiaoquan Wen [2], Francesca Luca [1*] and Roger Pique-Regi [1*]

[1]Center for Molecular Medicine and Genetics, Wayne State University, 540 E Canfield, Scott Hall, Detroit, MI 48201, USA
[2]Department of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

Motivation: Expression quantitative trait loci (eQTL) studies have discovered thousands of genetic variants that regulate gene expression and have enabled a better understanding of the functional role of non-coding sequences. However, eQTL studies are generally quite expensive, requiring large sample sizes and genome-wide genotyping of each sample. On the other hand, allele specific expression (ASE) is becoming a very popular approach to detect the effect of genetic variation on gene expression, even within a single individual. This is typically achieved by counting the number of RNA-seq reads matching each allele at heterozygous sites and rejecting the null hypothesis of 1:1 allelic ratio. In principle, when genotype information is not readily available it could be inferred from the RNA-seq reads directly. However, there are currently no methods that jointly infer genotype and test for ASE or that include the uncertainty in the genotype calls within the ASE inference step.

Results: Here, we present QuASAR, Quantitative Allele Specific Analysis of Reads, a novel statistical learning method for jointly detecting heterozygous genotypes and inferring ASE. The proposed ASE inference step takes into consideration the uncertainty in the genotype calls while including parameters that model base-call errors in sequencing and allelic over-dispersion. We validated our method with experimental data for which high quality genotypes are available. Results for an additional dataset with multiple replicates at different sequencing depths demonstrate that QuASAR is a powerful tool for ASE analysis when genotypes are not available.

Availability: `http://github.com/piquelab/QuASAR`

Contact: `fluca@wayne.edu;rpique@wayne.edu`

1 INTRODUCTION

Quantitative trait loci (QTLs) for molecular cellular phenotypes (as defined by Dermitzakis, 2012), such as gene expression (eQTL) (e.g. Stranger *et al.*, 2007), transcription factor (TF) binding (Kasowski *et al.*, 2010), and DNase I sensitivity (Degner *et al.*, 2012) have begun to provide a better understanding of how genetic variants in regulatory sequences can affect gene expression levels (see also Stranger *et al.*, 2007; Gibbs *et al.*, 2010; Melzer *et al.*, 2008; Gieger *et al.*, 2008). eQTL studies in particular have been successful in identifying genomic regions associated with gene expression in various tissues and conditions (e.g., Maranville *et al.*, 2011; Barreiro *et al.*, 2012; Nica *et al.*, 2011; Smirnov *et al.*, 2009; Dimas *et al.*, 2009; Ding *et al.*, 2010; Grundberg *et al.*, 2011; Lee *et al.*, 2014; Fairfax *et al.*, 2014). While previous studies have shown an enrichment for GWAS hits among regulatory variants in lymphoblastoid cell lines (LCLs) (Nica *et al.*, 2010; Nicolae *et al.*, 2010), a full understanding of the molecular mechanisms underlying GWAS hits requires a functional characterization of each variant in the tissue and environmental conditions relevant for the trait under study (e.g. estrogen level for genetic risk to breast cancer Cowper-Sal·lari *et al.*, 2012).

The ongoing GTEx project will significantly increase the number of surveyed tissues for which eQTL data are available and will represent a useful resource to functionally annotate genetic variants. However, the number of cell-types and environments explored will still be a very small fraction compared to a presumably large number of regulatory variants that mediate specific GxE interactions. eQTL studies are generally quite expensive, requiring large sample sizes ($n > 70$) which may be difficult to achieve for tissues that are obtained by surgical procedures or are difficult to culture *i*n vitro. Even if biospecimens are readily available at no cost, eQTL studies require large amounts of experimental work to obtain genotypes and gene expression levels. As the measurement of gene expression using high-throughput sequencing (RNA-seq) is becoming more popular than microarrays, RNA-seq library preparation is also becoming less expensive (46$ per sample) while costs of sequencing are also very rapidly decreasing (for example, 16M reads per sample would cost 49$ using a multiplexing strategy). Additionally, the sequence information provided by RNA-seq can be used to call genotypes (Shah *et al.*, 2009; Duitama *et al.*, 2012; Piskol *et al.*, 2013), detect and quantify isoforms (Trapnell *et al.*, 2010; Katz *et al.*, 2010) and to measure allele specific expression (ASE), if enough sequencing depth is available (Degner *et al.*, 2009; Pastinen, 2010a).

ASE approaches currently represent the most effective way to assay the effect of a cis-regulatory variant within a defined cellular environment, while controlling for any trans-acting modifiers of gene expression, such as the genotype at other loci (Pastinen, 2010b; Kasowski *et al.*, 2010; McDaniell *et al.*, 2010; Cowper-Sal·lari *et al.*, 2012; Reddy *et al.*, 2012; McVicker *et al.*, 2013). As such, ASE studies have greater statistical power to detect genetic effects in cis and can be performed using a smaller sample size than a traditional eQTL mapping approach.

In the absence of ASE, the two alleles for a heterozygous genotype at a single nucleotide polymorphism (SNP) in a gene transcript are represented in a 1:1 ratio in RNA-seq reads. To reject the null hypothesis and infer ASE it is necessary to identify

*to whom correspondence should be addressed

© Wayne State University 2014.

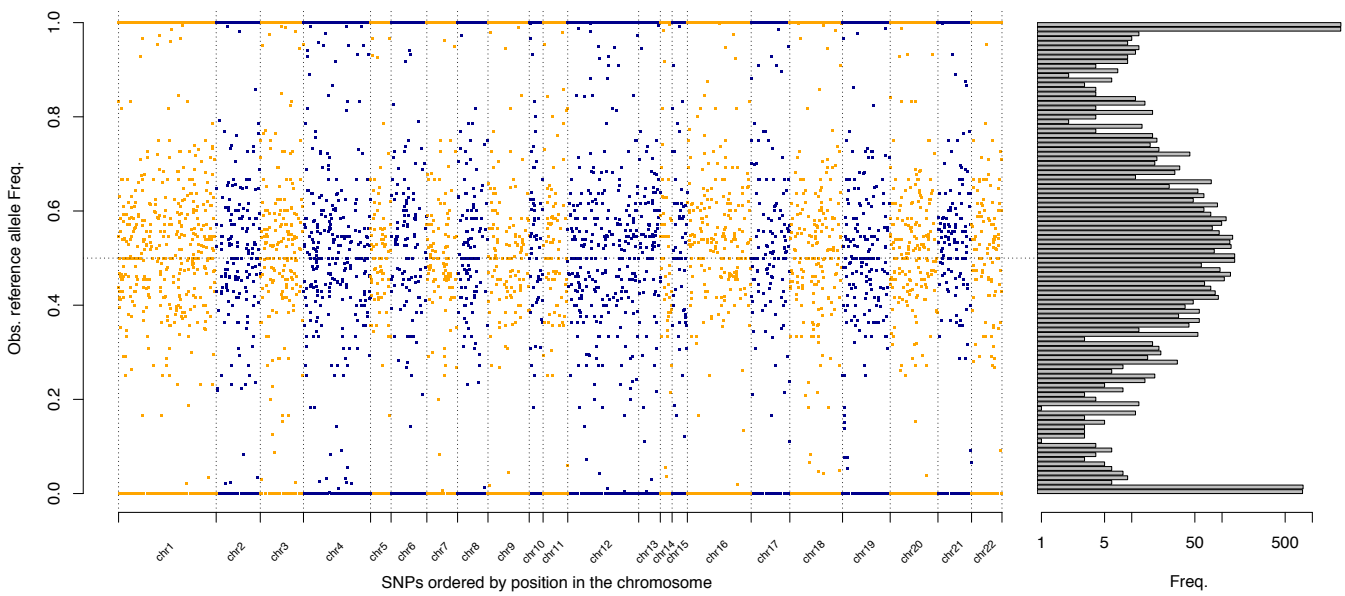

Fig. 1. Reference allele frequency from reads overlapping SNPs. (Left) Each dot represents a SNP covered by at least 15 RNA-seq reads. The y-axis represents the fraction of RNA-seq reads that match the reference allele (observed $\hat{\rho}_l$). The x-axis represents the order of the SNP position in a chromosome. (Right) Histogram showing the distribution of $\hat{\rho}_l$ values across the genome. The three modes ($\rho \in \{1, 0.5, 0\}$) correspond respectively to the three possible genotypes: homozygous reference (RR), heterozygous under no ASE (RA), and homozygous alternate (AA).

heterozygous SNPs with high confidence and a significant departure of the 50% allelic ratio. While genotyping and ASE are usually considered two separate problems, miss-calling an homozygous SNP as heterozygous will also likely induce an error in rejecting the ASE null hypothesis; thus, we argue that the two problems should be solved together.

While it is possible to obtain genotype information from RNA-seq (Shah *et al.*, 2009; Duitama *et al.*, 2012; Piskol *et al.*, 2013), to the best of our knowledge, all existing methods for detecting ASE consider that the genotypes are known and usually the error probabilities associated with genotyping are not taken into account for the ASE step. While overall genotyping quality can also be modeled within the ASE model (McVicker *et al.*, 2013), there is currently no method that for each SNP can jointly genotype and analyze ASE. Here, we propose a novel framework for quantitative allele specific analysis of reads (QuASAR) that starts from a single or multiple RNA-seq experiments from the same individual and can directly identify heterozygous SNPs and assess ASE accurately by taking into account base-calling errors and overdispersion in the ASE ratio. QuASAR is then evaluated in two different datasets that demonstrate the accuracy and the importance of incorporating the genotype uncertainty in determining ASE.

2 APPROACH

QuASAR starts with experimental high-throughput sequencing data. Here we focus on RNA-seq, but the same or similar pipeline can be applied to DNase-seq, ChIP-seq, ATAC-seq or other types of functional genomics library preparation. Figure 1 illustrates the underlying problem: detecting SNPs covered by reads with high allelic imbalance for which homozygosity (in the presence of base-calling errors) can also be rejected.

We focus our attention on sites that are known to be variable in human populations, specifically we consider all SNPs from the 1000 Genomes project (1KG) with a minor allele frequency MAF > 0.02. We index each SNP with $l \in \{1, ..., L\}$, and each sample by $s \in \{1, ..., S\}$. Samples are all from the same individual and may represent different experimental conditions or replicates. We only consider SNPs represented in at least 15 reads across all the samples. At each site l, three alternative genotypes are possible $g_l \in \{0, 1, 2\}$ being homozygous reference (RR), heterozygous (RA), or homozygous alternate (AA) respectively. For each sample s and site l, $N_{s,l}$ represents the total number of reads and $\boldsymbol{r}_{sl} = \{r_{slk}\}_{k=1}^{N_{sl}}$ take the value 1 if read k matches the reference allele, and 0 if it matches the alternate allele. We can then model the data $\mathcal{D} = \{\{\boldsymbol{r}_{sl}\}_{s=1}^{S}\}_{l=1}^{L}$as a mixture model

$$\Pr(\mathcal{D}) = \prod_{s=1}^{S}\prod_{l=1}^{L}\sum_{g_l \in \{0,1,2\}} \Pr(r_{sl}|g_l)\Pr(g_l) \tag{1}$$

where $\Pr(g_l)$ represents the prior probability associated with each genotype. $\Pr(r_{sl}|g_l)$ depends on the genotype, for $G_l = 0$:

$$\Pr(\boldsymbol{r}_{s,l}|g_l = 0; \epsilon_s) = \prod_{k=1}^{N_{sl}}(1-\epsilon_s)^{r_{slk}}\epsilon_s^{1-r_{slk}} \tag{2}$$

where we will only observe reads matching the alternate allele if those are base-calling errors, here modeled by the parameter ϵ_s. Conversely, for $G_l = 2$ we have the following:

$$\Pr(\boldsymbol{r}_{s,l}|g_l = 2; \epsilon_s) = \prod_{k=1}^{N_{sl}}(1-\epsilon_s)^{1-r_{slk}}\epsilon_s^{r_{slk}} \tag{3}$$

If the genotype is heterozygous $G_l = 1$, we may preferentially observe the reference allele with probability ρ_l (or the alternate allele with probability $1 - \rho_l$). This results in the following model:

$$\Pr\left(\boldsymbol{r}_{s,l}|g_l = 1; \epsilon_s\right) = \prod_{k=1}^{N_{sl}} \left((1-\rho_l)(1-\epsilon_s) + \rho_l\epsilon_s\right)^{1-r_{slk}} \cdot \left(\rho_l(1-\epsilon_s) + (1-\rho_l)\epsilon_s\right)^{r_{slk}} \quad (4)$$

These expressions can be simplified by considering that $R_{sl} = \sum_{k=1}^{N_l} r_{slk}$ and $A_{sl} = N_{sl} - R_{sl}$ are respectively the number of reads from sample s observed at site l matching the reference allele and the alternate allele:

$$\Pr\left(\boldsymbol{r}_{sl}|g_l = 0; \epsilon_s\right) = \psi\left(1, \epsilon_s\right)^{R_{sl}} \left[1 - \psi\left(1, \epsilon_s\right)\right]^{A_{sl}} \quad (5)$$

$$\Pr\left(\boldsymbol{r}_{sl}|g_l = 2; \epsilon_s\right) = \psi\left(0, \epsilon_s\right)^{R_{sl}} \left[1 - \psi\left(0, \epsilon_s\right)\right]^{A_{sl}} \quad (6)$$

$$\Pr\left(\boldsymbol{r}_{sl}|g_l = 1; \epsilon_s, \rho_{sl}\right) = \psi\left(\rho_{sl}, \epsilon_s\right)^{R_{sl}} \left[1 - \psi\left(\rho_{sl}, \epsilon_s\right)\right]^{A_{sl}} \quad (7)$$

where $\psi(\rho, \epsilon) = [\rho(1-\epsilon) + (1-\rho)\epsilon]$ and makes explicit that $g_l = 2$ (or $g_l = 0$) is indistinguishable from $\rho_{ls} = 0$ (or $\rho_{ls} = 1$) when $g_l = 1$. In QuASAR we resolve this identifiably problem by assuming that those cases with extreme ASE imbalance across all replicates are more likely to be homozygous genotypes.

To fit the mixture model we use an EM algorithm (see Methods for more details) in which we estimate sample specific base-calling error rates $\hat{\epsilon}_s$ (ρ is fixed to 0.5) and we are also able to provide a posterior probability for the genotype. For the ASE inference step, we wish to reject the null hypothesis $\rho_{sl} = 0.5$. We additionally consider that ψ in (5-7) is a random variable Ψ_{sl} sampled from a $\sim$Beta$(\alpha_{sl}, \beta_{sl})$ distribution with:

$$\alpha_{sl} = \psi_{sl} M_s \quad \beta_{sl} = \left[1 - \psi_{sl}\right] M_s \quad \psi_{sl} = \psi\left(\rho_{sl}, \epsilon_{ls}\right) \quad (8)$$

where M_s hyper parameter controls for over-dispersion and results in a better calibrated test as shown in the Results section. The resulting distribution on the number of reads after combining (7) and (8) is known as Beta-binomial distribution (13). Finally, the inference step takes into account the over-dispersion and genotype uncertainty, and can be formalized as a likelihood ratio test (LRT):

$$\Lambda_{sl} = \frac{\sup_{\rho_{sl} \in \{0, 0.5, 1\}} \left\{\Pr\left(\boldsymbol{r}_{sl}|\rho_{sl}, \hat{\epsilon}_s, \hat{M}_s\right)\right\}}{\sup_{\rho_{sl}} \left\{\Pr\left(\boldsymbol{r}_{sl}|\rho_{sl}, \hat{\epsilon}_s, \hat{M}_s\right)\right\}} \quad (9)$$

where the set of parameters $\{\hat{\epsilon}_s, \hat{M}_s\}_{s=1}^S$ are maximum likelihood estimates (MLE) under the null hypothesis $\rho_{ls} = 0.5$ (see Methods section). To calculate a p-value we use the property that $-2\log(\Lambda_{sl})$ is asymptotically distributed as χ_1^2.

3 METHODS

3.1 Experimental data

Lymphoblastoid cell lines (LCLs: GM18507 and GM18508) were purchased from Coriell Cell Repository and human umbilical vein endothelial cells (HUVECs) from Lonza. LCLs were cultured and starved according to (Maranville *et al.*, 2011). Cryopreserved HUVECs were thawed and cultured according to the manufacturer protocol (Lonza), with the exception that 48 hour prior to collection the medium was changed to a starvation medium, composed of phenol-red free EGM-2, without Hydrocortisone and supplemented with 2% charcoal stripped-FBS. Cells were washed 2X using ice cold PBS, lysed on the plate, using Lysis/Binding Buffer (Ambion), and frozen at -80C. mRNA was isolated using the Ambion Dynabeads mRNA Direct kit (Life Technologies). We then prepared libraries for Illumina sequencing using a modified version of the NEBNext Ultra Directional RNA-seq Library preparation kit (New England Biolabs). Briefly, each mRNA sample was fragmented (300 nt) and converted to double-stranded cDNA, onto which we ligated barcoded DNA adapters (NEXTflex-96 RNA-seq Barcodes, BIOO Scientific). Double-sided SPRI size selection (SPRISelect Beads, Beckman Coulter) was then performed to select 350-500 bp fragments. The libraries were then amplified by PCR and pooled together for sequencing on the Illumina HiSeq 2500 at the University of Chicago Genomics Core.

For each LCL sample, libraries from 9 replicates were pooled for a total of 42.3M and 34.9M 50bp PE reads, respectively. A total of 18 replicates across 6 time points were performed for the HUVEC samples (267M reads total), to capture a wide range of basal physiologic conditions.

3.2 Pre-processing

To create a core set of SNPs for ASE analysis, we first removed rare (MAF $< 2\%$) variants from all 1KG SNPs. We also removed SNPs within 25 bases up- or downstream of another SNP or short InDels as well as those SNPs in regions of annotated copy numbers or other blacklisted regions (Degner *et al.*, 2012). Sequencing reads were aligned to the reference human genome hg19 using `bwa mem` (Li and Durbin, 2009 `http://bio-bwa.sourceforge.net`). Reads with quality <10 and duplicate reads were removed using `samtools rmdup` (`http://github.com/samtools/`). Using a mappability filter (Degner *et al.*, 2012), we removed reads that do not map uniquely to the reference genome and to alternate genome sequences built considering all 1KG variants. Aligned reads were then piled up on these SNPs using `samtools mpileup` command. Reads with a SNP at the beginning or at the end of the read were also removed to avoid any potential experimental bias. Finally, the pileups were re-formatted so that each SNP has a count for reads containing the reference allele, and a count for those containing the alternate allele.

3.3 Model fitting and parameter estimation

In order to use the expectation maximization (EM) technique (McLachlan and Krishnan, 2007), we first convert (1) to a "complete" likelihood, as if we knew the underlying genotypes $\mathcal{G} = \{G_l\}_{l=1}^L$:

$$L(\Theta) = \Pr\left(\mathcal{D}, \mathcal{G}|\Theta\right) = \Pr\left(\mathcal{D}|\mathcal{G}; \Theta\right) \Pr\left(\mathcal{G}|\Theta\right) = \quad (10)$$

$$= \prod_{l=1}^{L} \prod_{g=0}^{2} \left\{\Pr(G_l = g) \prod_{s=1}^{S} \Pr\left(\boldsymbol{r}_{sl}|g_l = g; \epsilon_s, \rho_{sl}\right)\right\}^{G_l^g}$$

where $G_l^g \equiv \mathbf{1}(G_l = g)$ are binary indicator variables and Θ represents the set of all parameters of the model. In log-likelihood form we have:

$$l(\Theta) = \log L(\Theta) = \sum_{l=1}^{L} \sum_{g=0}^{2} G_l^g \ln(\Pr(G_l = g)) + \sum_{s=1}^{S} \sum_{l=1}^{L} \{ \quad (11)$$

$$G_l^0 \left[R_{sl} \ln \psi(1, \epsilon_s) + A_{sl} \ln(1 - \psi(1, \epsilon_s))\right]$$

$$+ G_l^1 \left[R_{sl} \ln \psi(\rho_{sl}, \epsilon_s) + A_{sl} \ln(1 - \psi(\rho_{sl}, \epsilon_s))\right]$$

$$+ G_l^2 \left[R_{sl} \ln \psi(0, \epsilon_s) + A_{sl} \ln(1 - \psi(0, \epsilon_s))\right]\}$$

During the genotyping step, in order to maintain identifiability of the model, we fix $\rho_{sl} = 0.5$ for all loci. Although M_s could also be estimated within the EM procedure, we only consider overdispersion on the ASE step. These two choices lead to a much simpler EM procedure and a slightly conservative estimate of ϵ_s.

E-Step: From the complete likelihood function (11) we derive the expected values for the unknown genotype indicator variables

$\mathbb{E}(G_l^g|\mathcal{D},\Theta) = \langle G_l^g \rangle$ given the observed data and the current estimates for the model parameters. We are also interested in $\langle G_l^g \rangle = \Pr(G_l^g = g|\mathcal{D},\hat{\Theta})$, because they are the posterior probability of each genotype given the data:

$$C = (\langle G_l^0 \rangle + \langle G_l^0 \rangle + \langle G_l^0 \rangle)^{-1} \tag{12}$$

$$\langle G_l^0 \rangle = C \Pr(G_l = 0) \exp\left(\sum_{s=1}^{S} [R_{ls} \ln(1-\hat{\epsilon}_s) + A_{ls} \ln(\hat{\epsilon}_s)]\right)$$

$$\langle G_l^1 \rangle = C \Pr(G_l = 1) \exp\left(\ln(0.5) \sum_{s=1}^{S} [R_{ls} + A_{ls}]\right)$$

$$\langle G_l^2 \rangle = C \Pr(G_l = 2) \exp\left(\sum_{s=1}^{S} [R_{ls} \ln(\hat{\epsilon}_s) + A_{ls} \ln(1-\hat{\epsilon}_s)]\right)$$

The prior genotype probabilities $Pr(G_l = g)$ are obtained from the 1KG allele frequencies assuming Hardy-Weinberg equilibrium, but the user can change this.

M-Step: Using the expected values from the E-step, the complete likelihood is now a function of ϵ_s that is easily maximized

$$\hat{\epsilon}_s = \text{logit}^{-1}\left[\ln \frac{\sum_{l=1}^{L} \left(\langle G_l^0 \rangle A_{sl} + \langle G_l^2 \rangle R_{sl}\right)}{\sum_{l=1}^{L} \left(\langle G_l^0 \rangle R_{sl} + \langle G_l^2 \rangle A_{sl}\right)}\right]$$

After we run QuASAR to infer genotypes across samples from the same individual, for each site we have a posterior probability for each genotype $\langle G_l^g \rangle$, and a base-calling error $\hat{\epsilon}_s$ estimated for each sample. From these posteriors, discrete genotypes are called by using the genotype with the highest posterior probability; the maximum a posteriori (MAP) estimate.

ASE-Inference: To detect ASE we only consider sites with an heterozygous MAP higher than a given threshold (e.g., $\langle G_l^1 \rangle > 0.99$). We then test the possibility that ρ_{sl} is different than 0.5 while also taking into account overdispersion using a beta-binomial model (by combining (7) and (8)):

$$= \Pr\left(R_{sl}|N_{sl},\psi_{sl},M_s\right) =$$

$$= \binom{N_{sl}}{R_{sl}} \frac{\Gamma\left(M_s\right)\Gamma\left(R_{sl}+\psi_{sl}M_s\right)\Gamma\left(A_{sl}+(1-\psi_{sl})M_s\right)}{\Gamma\left(N_{sl}+M_s\right)\Gamma\left(\psi_{sl}M_s\right)\Gamma\left((1-\psi_{sl})M_s\right)} \tag{13}$$

where M_s controls the effective number of samples supporting the prior belief that $\rho = .5$ and is estimated using grid search:

$$\hat{M}_s = \arg\max_{M_s} \left(\prod_{l=1}^{L} \Pr\left(R_{sl}|N_{sl},\hat{\epsilon}_s,\rho_{sl}=0.5,M_s\right)\right) \tag{14}$$

Then, we estimate $\hat{\rho}_{sl}$ using (13) and $\hat{M}_s$ from (14) using a standard gradient method (L-BFGS-B) to maximize the following log-likelihood function

$$l(\rho_{sl};\hat{M}_s,\hat{\epsilon}_s) = \Pr\left(R_{sl}|N_{sl},\psi_{sl}=\psi(\rho_{sl},\hat{\epsilon}_s),\hat{M}_s\right) \tag{15}$$

Finally, all these parameters are used to calculate the LRT in (9) to get a p-value.

Additionally, we can provide an estimate of the standard error associated with the parameter ρ_{sl} using the second derivative of the log-likelihood function (15):

$$\hat{\sigma}_{\hat{\rho}_{sl}} = \left|\frac{\partial^2}{\partial \rho_{sl}^2} l(\rho_{sl};\hat{M}_s,\hat{\epsilon}_s)\right|_{\rho_{sl}=\hat{\rho}_{sl}}^{-\frac{1}{2}} \tag{16}$$

alternatively we can also recover a standard error from (9) (as is asymptotically distributed as $\chi^2_{df=1}$), allowing the p-value to be used to back solve for the standard error:

$$\hat{\sigma}_{\hat{\rho}_{sl}} = \left|\frac{\hat{\rho}_{ls}}{Q\left(\frac{p_{ls}}{2}\right)}\right| \tag{17}$$

where $Q()$ is the quantile function for a standard normal distribution and p_{ls} is the p-value from (9). We use the first form (16) when $\hat{\rho} \sim 0.5$ and (17) otherwise as they give a better approximation at those ranges, respectively. Alternatively, if we do not need $\hat{\sigma}_{\hat{\rho}_{sl}}$ we can use (15) to obtain a profile likelihood confidence interval for ρ_{sl}.

4 RESULTS

We implemented the QuASAR approach as detailed in the Methods section in an `R` package available at `http://github.com/piquelab/QuASAR`. We first sought to evaluate QuASAR genotyping accuracy using RNA-seq reads obtained from two lymphoblastoid cell-lines (LCLs) that already have very high quality genotypes calls from the 1KG project (GM18507 and GM18508). As illustrated in Table 1, we are able to accurately genotype thousands of loci, and genotyping error rates decrease with an increase of the MAP threshold. The method by construction is more conservative in making heterozygous calls. In case of doubt, between a heterozygous genotype with extreme allelic imbalance, or homozygous genotype with base call errors, our model would lean in favor of the homozygote call. This is a crucial feature for accurate inference of ASE as we will discuss in more detail in this section.

Sample	**QuASAR Performance**	**MAP $\langle G_l^g \rangle$**		
		>0.5	**>0.9**	**>0.99**
NA18507	Heterozygotes	*2891*	*2864*	***2828***
	False Discoveries	*1.25%*	*1.01%*	*0.88%*
	Homozygotes	*8799*	*8752*	*8497*
	False Discoveries	*4.28%*	*4.15%*	*4.14%*
NA18508	Heterozygotes	*2447*	*2418*	***2386***
	False Discoveries	*1.06%*	*0.70%*	*0.46%*
	Homozygotes	*7462*	*7427*	*7224*
	False Discoveries	*4.50%*	*4.36%*	*4.33%*

Table 1. QuASAR accuracy in genotyping. Each row reports the number of heterozygous and homozygous SNPs identified by QuASAR and the percentage of false discoveries when compared to 1KG genotypes. Each column uses a different MAP threshold to define high confidence genotype calls.

We next sought to characterize QuASAR performance in genotyping and in calling ASE from RNA-seq experiments sequenced at different depths. In total we analyzed 18 samples (3 replicates across 6 different time-points) for an individual for which genotypes were not previously available. We combined the 18 fastq files in different ways as input for QuASAR, to obtain an empirical power curve (see Figure 2). As expected, we observed that our ability to detect heterozygotes (MAP> 0.99) increases with the sequencing depth. The number of heterozygous sites detected seems to start to plateau at $10^{4.5} = 31,623$ when the total number of RNAs-seq reads exceeds 100 million. At a more modest sequencing depth of 16 million, we can still detect more than 3,000 heterozygous sites that likely correspond to genes with higher gene expression.

After obtaining the genotypes, we then assessed whether there is evidence of ASE on any of the SNPs determined to be heterozygous. Using the LRT (9) we determined ASE and obtained p-values. We controlled the FDR using the q-value procedure (Storey, 2002). As shown in Figure 3, our ability to detect ASE greatly increases with the number of SNPs we are able to genotype, which in turn is a function of coverage (Figure 2). Using 100 million reads we achieve detecting more than 100 SNPs with ASE at 10%FDR.

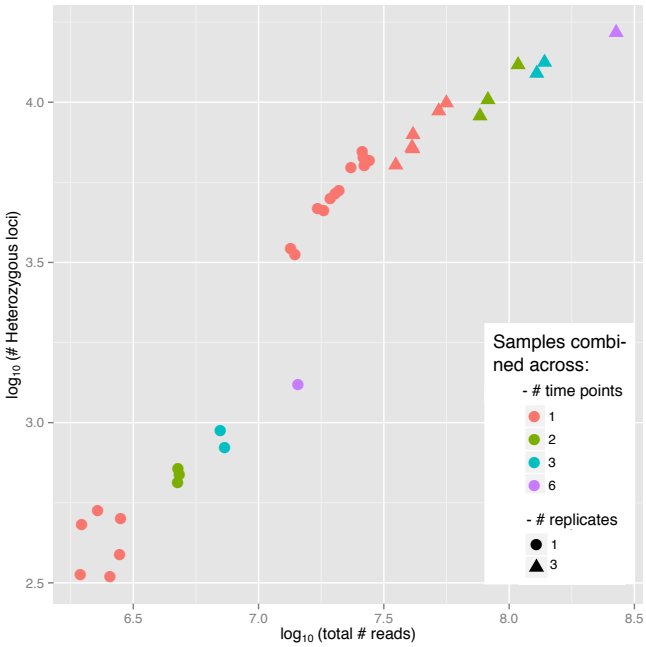

Fig. 2. Empirical power in detecting heterozygous SNPs as a function of sequencing depth. Each point represents a single input dataset to QuASAR: either as a single experiment replicate and time point (red dot), combining multiple time points (2 = green, 3 = blue, 6 = purple), or combining replicates (1 = dot, 3 = triangle). The x-axis represents the $\log_{10}$ of the total number of RNA-seq reads in the fastq input files. The y-axis represents the $\log_{10}$ of the total number of SNPs that are determined to be heterozygous.

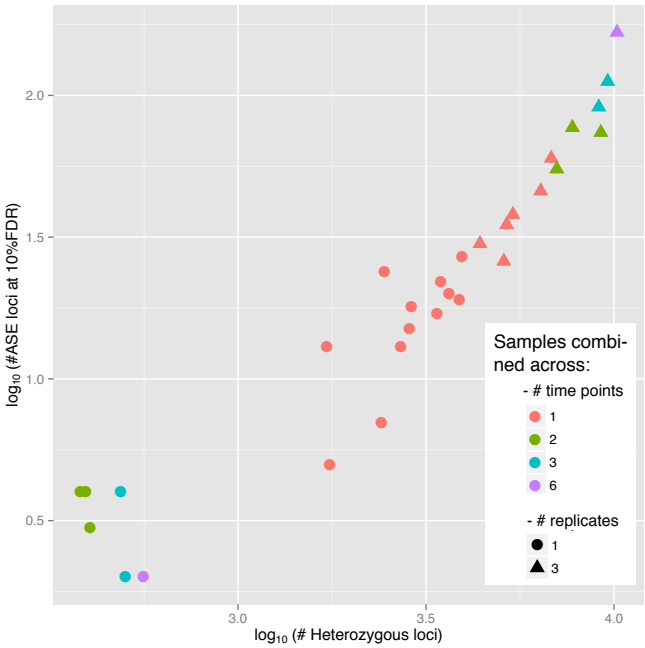

Fig. 3. Empirical power in detecting ASE as a function of the number of heterozygous SNPs detected. Each point represents a single input dataset to QuASAR as in Figure 2. The x-axis represents the $\log_{10}$ of the total number of SNPs that are determined to be heterozygous. The y-axis represents the $\log_{10}$ of the number of SNPs that have a significant p-value for ASE at 10% FDR.

In order to determine if the ASE inference in QuASAR is well calibrated and to compare QuASAR to other ASE tests we used QQ-plots, see Figure 4. A QQ-plot compares the quantiles observed from a test statistic to those that are expected under a null distribution (e.g., p-values are uniformly distributed between 0 and 1). The shape of the QQ-plot curve is useful to judge how well the p-values are calibrated when we expect that a large number of the tests conducted are sampled from the null distribution. In this latter scenario, we expect that the QQ-plot curve would follow the $y = x$ line for the bulk of the higher range of p-values. For small p-values, we expect that the curve starts to depart from the 1:1 line representing the small proportion of tests that are not sampled from the null distribution. Figure 4 clearly shows that the Binomial test is too optimistic, and will likely lead to many false discoveries. The Beta-binomial model is well calibrated, but if we are not certain about the genotype being a true heterozygote it can lead to very small p-values that are false positives. In QuASAR, we use a Beta-binomial model and we also consider uncertainty on the genotype, which results in the most conservative of the three different tests, while likely avoiding the most common causes of false positives in ASE analysis.

5 DISCUSSION

QuASAR is the first approach that detects genotypes and infers ASE from the same sequencing data. In this work, we focused on RNA-seq, but QuASAR can be applied to other data types (ChIP-seq, DNase-seq, ATAC-seq, and others). Indeed, the more experimental data we have from the same underlying individual across many experimental samples (data-types, conditions, cell-types or technical replicates), the more certainty we can gather about the genotype for any given site. The algorithm is computationally very fast, each EM iteration is $O(LS)$ linear with the number of SNPs and samples and convergence is achieved in about 10 or less iterations.

A key aspect of QuASAR ASE inference step is that it takes into account over-dispersion and genotype uncertainty resulting in a test that we have shown here to be well-calibrated. In many cases, the p-values obtained from biased statistics can be recalibrated to the true null distribution using a permutation procedure. Unfortunately, this is not possible for ASE inference, as randomly permuting the reads assigned to each allele would inadvertently assume that the reads are distributed according to a Binomial distribution. More complicated and computationally costly resampling procedures can be proposed, but it is not clear which additional assumptions may introduce and if they can correctly take into account genotyping uncertainty.

If prior genotype information is available, it can also be provided as input to the algorithm. The prior uncertainty of the genotypes should be reflected in the form of prior probabilities for each genotype. In this paper, we have shown that we can obtain reliable genotype information from RNA-seq reads, thus making additional genotyping unnecessary if the endpoint is to detect ASE. Instead, sequencing the RNA-seq libraries at a higher depth is probably a better strategy as it greatly improves the power to detect ASE signals.

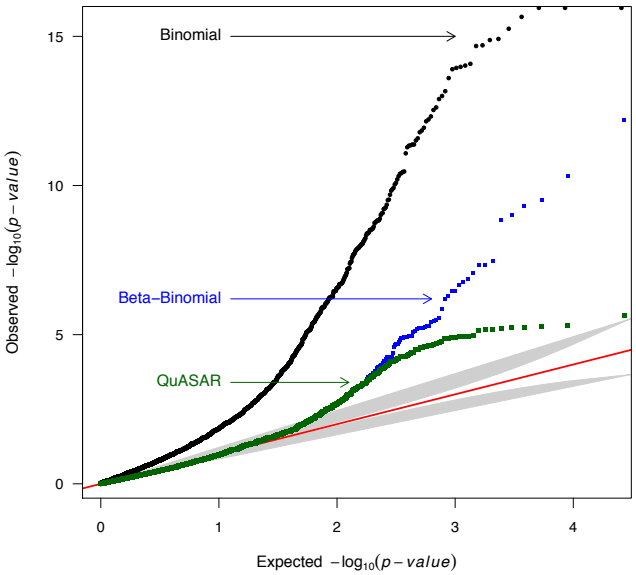

Fig. 4. QQplot comparing the p-value distribution of 3 alternative methods for determining ASE. The x-axis shows the $\log_{10}$ quantiles of the p-values expected from the null distribution. The y-axis shows the $\log_{10}$ quantiles of the p-values computed from the real data using 3 different methods: Binomial (black) assumes $M = \infty$ no overdispersion; Beta-binomial (blue) considers overdispersion but does not consider uncertainty in the genotype; iii) QuASAR uses the Beta-binomial distribution and uncertainty in the genotype calls. In all three cases the same set of SNPs are considered

Furthermore, as sequencing costs are decreasing very rapidly, ASE methods are becoming very attractive in applications where eQTL studies have been previously used. This is even more important in scenarios where collecting large number of samples is expensive or infeasible. Large scale eQTL studies are still very much necessary for fine-mapping, but allele specific analysis methods can provide unique insights into mechanisms that are uncovered only under specific experimental conditions, for example as a result of gene x environment interactions.

ACKNOWLEDGEMENT

We would like to thank Wayne State University High Performance Computing Grid for computational resources, the University of Chicago Genomics Core for sequencing services, and Jacob Degner and members of the Luca/Pique group for helpful comments and discussions.

Funding: NIH 1R01GM109215-01 (RPR, FL)
AHA 14SDG20450118 (FL)

REFERENCES

Barreiro, L. B., Tailleux, L., Pai, A. A., Gicquel, B., Marioni, J. C., and Gilad, Y. (2012). Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.*, **109**(4), 1204–9.

Cowper-Sal·lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., Moore, J. H., and Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.*, **44**(11), 1191–8.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

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. T. (2012). Cellular genomics for complex traits.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., Gagnebin, M., Nisbett, J., Deloukas, P., Dermitzakis, E. T., and Antonarakis, S. E. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.

Ding, J., Gudjonsson, J. E., Liang, L., Stuart, P. E., Li, Y., Chen, W., Weichenthal, M., Ellinghaus, E., Franke, A., Cookson, W., Nair, R. P., Elder, J. T., and Abecasis, G. R. (2010). Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet.*, **87**, 779–789.

Duitama, J., Srivastava, P. K., and Mndoiu, I. I. (2012). Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data. *BMC Genomics*, **13**(Suppl 2), S6.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., and Knight, J. C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, **343**(6175), 1246949.

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I. P., Troncoso, J., Johnson, R., Zielke, H. R., Ferrucci, L., Longo, D. L., Cookson, M. R., and Singleton, A. B. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, **6**, e1000952.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabé de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.-W., Wichmann, H.-E., Weinberger, K. M., Adamski, J., Illig, T., and Suhre, K. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.*, **4**, e1000282.

Grundberg, E., Adoue, V., Kwan, T., Ge, B., Duan, Q. L., Lam, K. C. L., Koka, V., Kindmark, A., Weiss, S. T., Tantisira, K., Mallmin, H., Raby, B. A., Nilsson, O., and Pastinen, T. (2011). Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. *PLoS Genet.*, **7**, e1001279.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M.-Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., and Snyder, M. (2010). Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**(12), 1009–15.

Lee, M. N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T. M., Imboywa, S. H., Chipendo, P. I., Ran, F. A., Slowikowski, K., Ward, L. D., Raddassi, K., McCabe, C., Lee, M. H., Frohlich, I. Y., Hafler, D. a., Kellis, M., Raychaudhuri, S., Zhang, F., Stranger, B. E., Benoist, C. O., De Jager, P. L., Regev, A., and Hacohen, N. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, **343**(6175), 1246980.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Maranville, J. C., Luca, F., Richards, A. L., Wen, X., Witonsky, D. B., Baxter, S., Stephens, M., and Di Rienzo, A. (2011). Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genet.*, **7**, e1002162.

McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., Scott, L. J., Morken, M. A., Kucera, K. S., Battenhouse, A., Keefe, D., Collins, F. S., Willard, H. F., Lieb, J. D., Furey, T. S., Crawford, G. E., Iyer, V. R., and Birney, E. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K. (2013). Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science*, **747**.

Melzer, D., Perry, J. R. B., Hernandez, D., Corsi, A.-M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., Paolisso, G., Rafiq, S., Simon-Sanchez, J.,

Lango, H., Scholz, S., Weedon, M. N., Arepalli, S., Rice, N., Washecka, N., Hurst, A., Britton, A., Henley, W., van de Leemput, J., Li, R., Newman, A. B., Tranah, G., Harris, T., Panicker, V., Dayan, C., Bennett, A., McCarthy, M. I., Ruokonen, A., Jarvelin, M.-R., Guralnik, J., Bandinelli, S., Frayling, T. M., Singleton, A., and Ferrucci, L. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.*, **4**, e1000072.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., and Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.*, **6**(4), e1000895.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., Hedman, A. K., Bataille, V., Tzenova Bell, J., Surdulescu, G., Dimas, A. S., Ingle, C., Nestle, F. O., di Meglio, P., Min, J. L., Wilk, A., Hammond, C. J., Hassanali, N., Yang, T.-P., Montgomery, S. B., O'Rahilly, S., Lindgren, C. M., Zondervan, K. T., Soranzo, N., Barroso, I., Durbin, R., Ahmadi, K., Deloukas, P., McCarthy, M. I., Dermitzakis, E. T., and Spector, T. D. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.*, **7**(2), e1002003.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **6**(4), e1000888.

Pastinen, T. (2010a). Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.*, **11**, 533–538.

Pastinen, T. (2010b). Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.*, **11**, 533–538.

Piskol, R., Ramaswami, G., and Li, J. B. (2013). Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.*, **93**(4), 641–51.

Reddy, T. E., Gertz, J., Pauli, F., Kucera, K. S., Varley, K. E., Newberry, K. M., Marinov, G. K., Mortazavi, A., Williams, B. A., Song, L., Crawford, G. E., Wold, B., Willard, H. F., and Myers, R. M. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression.

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., Steidl, C., Holt, R. a., Jones, S., Sun, M., Leung, G., Moore, R., Severson, T., Taylor, G. a., Teschendorff, A. E., Tse, K., Turashvili, G., Varhol, R., Warren, R. L., Watson, P., Zhao, Y., Caldas, C., Huntsman, D., Hirst, M., Marra, M. a., and Aparicio, S. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**(7265), 809–13.

Smirnov, D. A., Morley, M., Shin, E., Spielman, R. S., and Cheung, V. G. (2009). Genetic analysis of radiation-induced changes in human gene expression. *Nature*, **459**, 587–591.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. - Ser. B Stat. Methodol.*, **64**(3), 479–498.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.

Trapnell, C., Williams, B. a., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**(5), 511–5.