Genetics and population analysis

QuASAR: quantitative allele-specific analysis of reads

Chris T. Harvey¹, Gregory A. Moyerbrailean¹, Gordon O. Davis¹, Xiaoquan Wen², Francesca Luca¹,* and Roger Pique-Regi¹,*

¹Center for Molecular Medicine and Genetics, Department of Obstetrics and Gynecology, Wayne State University, 540 E Canfield, Scott Hall, Detroit, MI 48201, USA and ²Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 2, 2014; revised on November 15, 2014; accepted on November 26, 2014

Abstract

Motivation: Expression quantitative trait loci (eQTL) studies have discovered thousands of genetic variants that regulate gene expression, enabling a better understanding of the functional role of non-coding sequences. However, eQTL studies are costly, requiring large sample sizes and genome-wide genotyping of each sample. In contrast, analysis of allele-specific expression (ASE) is becoming a 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 testing the null hypothesis of a 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 existing methods that jointly infer genotypes and conduct ASE inference, while considering uncertainty in the genotype calls.

Results: 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 and implementation: http://github.com/piquelab/QuASAR.

Contact: fluca@wayne.edu or rpique@wayne.edu

Supplementary information: Supplementary Material is available at Bioinformatics online.

1 Introduction

Quantitative trait loci (QTLs) for molecular cellular phenotypes (as defined by Dermitzakis, 2012), such as gene expression [expression QTL (eQTL)] (e.g. Stranger et al., 2007), transcription factor 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 Gibbs et al., 2010; Gieger et al., 2008; Melzer et al., 2008; Stranger et al., 2007). eQTL studies in particular have been successful at identifying genomic regions associated with gene expression in various tissues and conditions (e.g. Barreiro et al., 2012; Dimas et al., 2009; Ding et al., 2010; Fairfax et al., 2014; Grundberg et al., 2011; Lee et al., 2014; Maranville et al., 2011; Nica et al., 2011; Smirnov et al., 2009). Although 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 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-Sallicani 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 are a small subset of the presumably larger number of regulatory variants that mediate specific GxE interactions. eQTL studies are 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 in vitro. Even if biospecimens are readily available at no cost, eQTL studies require large amounts of experimental work to measure 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/sample, whereas costs of sequencing are also very rapidly decreasing (e.g. 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 (Duitama et al., 2012; Piskol et al., 2013; Shah et al., 2009), detect and quantify isoforms (Katz et al., 2010; Trappnell et al., 2010) and to measure allele-specific expression (ASE), if enough sequencing depth is available (Degner et al., 2009; Pastinen, 2010).

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 all trans-acting modifiers of gene expression, such as the genotype at other loci (Cowper-Sallicani et al., 2012; Hasin-Brumshtein et al., 2014; Kasowski et al., 2010; Kukurba et al., 2014; McDanieli et al., 2010; McVicker et al., 2013; Pastinen, 2010; Reddy et al., 2012; Skelly et al., 2011). As such, ASE studies have greater statistical power to detect genetic effects in cis than a traditional eQTL mapping approach when using a small sample size. Additionally, ASE may also be useful to detect epigenetic imprinting of gene expression if ASE is present but no eQTL is detected (Degner et al., 2009; Seoighe et al., 2006).

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 of RNA-seq reads. To reject the null hypothesis and infer ASE, it is necessary to first identify heterozygous SNPs with high confidence and then conduct inference to detect a departure from a 50% allelic ratio. Although genotyping and ASE are usually considered two separate problems, missing a homozygous SNP as heterozygous is likely to induce an error in rejecting the ASE null hypothesis; thus, we argue that the two problems should be addressed together.

Although it is possible to obtain genotype information from RNA-seq (Duitama et al., 2012; Piskol et al., 2013; Shah et al., 2009), to the best of our knowledge, all existing methods for detecting ASE consider the genotypes known and error probabilities associated with genotyping are not taken into account for the ASE step. Although 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 detect allelic imbalances in high-throughput sequencing data. An approach that takes into account base-calling errors was previously proposed for detecting ASE in ESTs data (Seoighe et al., 2006), but for RNA-seq data, it is essential to also include overdispersion. 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 evaluated with two different datasets that demonstrate genotyping accuracy and the importance of incorporating the genotype uncertainty when performing ASE inference.

2 Materials and methods

2.1 QuASAR 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 preparations. Figure 1 illustrates the underlying problem: detecting SNPs covered by a number of transcripts with high allelic imbalance and for which homozgyosity (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 $s \in \{1,...,S\}$. All samples are 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 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 $l$, $N_{sl}$ represents the total number of reads and $r_{sl} = \{r_{slk}\}_{k=1}^{N_{sl}}$ 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 $D = \{r_{slk}\}_{s=1}^{S}_{l=1}^{L}$ as a mixture model

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

(1)

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

$$\Pr (r_{sl|g_l} = 0 ; \epsilon_l) = \prod_{k=1}^{N_{sl}} (1 - \epsilon_l)^{1-r_{slk}} \epsilon_l^{r_{slk}}$$

(2)

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

$$\Pr (r_{sl|g_l} = 2 ; \epsilon_l) = \prod_{k=1}^{N_{sl}} (1 - \epsilon_l)^{1-r_{slk}} \epsilon_l^{r_{slk}}$$

(3)

If the genotype is heterozygous $G_l = 1$, we observe reads from the reference allele with probability $\rho_l$ or the alternate allele with probability $(1 - \rho_l)$, resulting in the following model:

$$\Pr (r_{sl|g_l} = 1 ; \epsilon_l) = \prod_{k=1}^{N_{sl}} \left((1 - \rho_l)(1 - \epsilon_l) + \rho_l \epsilon_l\right)^{1-r_{slk}} \cdot \left(\rho_l(1 - \epsilon_l) + (1 - \rho_l)\epsilon_l\right)^{r_{slk}}$$

(4)

Considering that $R_{sl} = \sum_{k=1}^{N_{sl}} r_{slk}$ 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, the previous equations can be simplified as follows:

$$\Pr (r_{sl|g_l} = 0 ; \epsilon_l) = \psi (1, \epsilon_l)^{R_{sl}} (1 - \psi (1, \epsilon_l))^{A_{sl}}$$

(5)
ably problem by assuming that those cases with extreme ASE across the genome are coded as QuASAR.

\[ \Pr (\tau_j | g_i = 2; \epsilon_i) = \psi(0, \epsilon_i)^{R_{ij}} [1 - \psi(0, \epsilon_i)]^{\bar{R}_{ij}} \]  
\[ \Pr (\tau_j | g_i = 1; \epsilon_i, \rho_{ij}) = \psi(\rho_{ij}, \epsilon_i)^{R_{ij}} [1 - \psi(\rho_{ij}, \epsilon_i)]^{\bar{R}_{ij}} \]  
where \( \psi(\rho, \epsilon) = [\rho (1 - \epsilon) + (1 - \rho) \epsilon] \) and makes explicit that homozygotes, \( g_i = 2 \) (or \( g_i = 0 \)), are indistinguishable from \( \rho_{ij} = 0 \) (or \( \rho_{ij} = 1 \)) when \( g_i = 1 \). In QuASAR, we resolve this identifiability problem by assuming that those cases with extreme ASE across all replicates are more likely to be homozygous genotypes.

To fit the mixture model, we use an expectation maximization (EM) algorithm (see Section 2 for more details) which we estimate sample-specific base-calling error rates \( \epsilon_i \) (\( \rho \) is fixed to 0.5) and posterior probabilities for the genotypes. For the ASE inference step, we wish to test the null hypothesis \( \rho_{ij} = 0.5 \). We additionally consider that \( \psi \) in (5–7) is a random variable \( \Psi_{ij} \) sampled from an \( \sim \beta(\alpha_{ij}, \beta_{ij}) \) distribution with:

\[ \alpha_{ij} = \psi_{ij} M_i \quad \beta_{ij} = (1 - \psi_{ij}) M_i \quad \Psi_{ij} = \psi(\rho_{ij}, \epsilon_i) \]  
where the hyper parameter \( M_i \) controls for over-dispersion and results in a better calibrated test as shown in Section 3. Combining (7) and (8) results in a beta-binomial distribution (13), which we use to model the number of reads coming from the reference allele. Formalized as a likelihood ratio test (LRT), the inference step takes into account over-dispersion and genotype uncertainty:

\[ A_{ij} = \frac{\sup_{\alpha_{ij} \in [0,0.5]} \{ \Pr (\tau_j; \rho_{ij}, \epsilon_i, M_i) \}}{\sup_{\rho_{ij}} \{ \Pr (\tau_j; \rho_{ij}, \epsilon_i, M_i) \}} \]  
where the set of parameters \( \{ \epsilon_i, M_i \}_{i=1}^S \) are maximum likelihood estimates under the null hypothesis \( \rho_{ij} = 0.5 \) (see Section 2). To calculate a P value, we use the property that \(-2 \log (A_{ij})\) is asymptotically distributed as \( \chi^2_1 \).

2.2 Model fitting and parameter estimation

To use the EM procedure (McLachlan and Krishnan, 2007), we first convert (1) to a ‘complete’ likelihood, as if we knew the underlying genotypes \( G = \{ G_i \}_{i=1}^L \):

\[ L(\Theta) = \Pr (\mathcal{D} | \mathcal{G} \Theta) = \Pr (\mathcal{D} | g \Theta) \Pr (g \Theta) = \prod_{i=1}^L \left\{ \Pr (G_i = g) \sum_{\rho_{ij}} \Pr (\tau_i; g, \rho_{ij}, \epsilon_i) \right\} \]  
where \( G_i^j \equiv 1(G_i = g) \) are binary indicator variables and \( \Theta \) represents the set of all parameters of the model. In log likelihood form we have:

\[ l(\Theta) = \log L(\Theta) = \sum_{i=1}^L \sum_{g=0}^2 G_i^j \ln (\Pr (G_i = g)) + \sum_{i=1}^L \sum_{j=1}^S \left\{ G_i^j \ln \psi(1, \epsilon_i) + A_{ij} \ln \left( 1 - \psi(1, \epsilon_i) \right) \right\} + \frac{G_i^j \ln \psi(0, \epsilon_i) + A_{ij} \ln \left( 1 - \psi(0, \epsilon_i) \right)}{2} \]  

During the genotyping step, to maintain identifiability of the model, we fix \( \rho_{ij} = 0.5 \) for all loci. Although \( M_i \) 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 \( \epsilon_i \).

2.2.1 E step

From the complete likelihood function (11), we derive the expected values for the unknown genotype indicator variables \( E(G_i^j | \mathcal{D}, \mathcal{G}, \Theta) = \langle G_i^j \rangle \) given the observed data and the current estimates for the model parameters. These quantities are also of interest for genotyping because they represent the posterior probabilities of each genotype given the data, \( \langle G_i^j \rangle = \Pr (G_i = g | \mathcal{D}, \mathcal{G}, \Theta) \):

\[ C = \langle (G_i^0) \rangle + \langle (G_i^1) \rangle^{-1} \langle (G_i^0) \rangle \Pr (G_i = 0) \exp \left\{ \sum_{i=1}^S \left[ R_{ij} \ln (1 - \epsilon_i) + A_{ij} \ln (\epsilon_i) \right] \right\} \]  
\[ \langle G_i^0 \rangle = C \Pr (G_i = 0) \exp \left\{ \ln (0.5) \sum_{i=1}^S \left[ R_{ij} + A_{ij} \right] \right\} \]  
\[ \langle G_i^1 \rangle = C \Pr (G_i = 1) \exp \left\{ \ln \left( 1 - \epsilon_i \right) \sum_{i=1}^S \left[ R_{ij} + A_{ij} \right] \right\} \]  
\[ \langle G_i^2 \rangle = C \Pr (G_i = 2) \exp \left\{ \ln (0.5) \sum_{i=1}^S \left[ R_{ij} + A_{ij} \right] \right\} \]  

Fig. 1. Reference allele frequency from reads overlapping SNPs. (Left) Each dot represents an SNP covered by at least 15 RNA-seq reads. The x-axis represents the order of reads coming from the reference allele (observed \( \tilde{\epsilon}_i \)). The y-axis represents the order of the SNP position on a chromosome. (Right) Histogram showing the distribution of \( \tilde{\epsilon}_i \) 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).
The prior genotype probabilities \(Pr(G_i = g)\) are obtained from the 1KG allele frequencies assuming Hardy–Weinberg equilibrium, but the user can change this.

### 2.2.2 M step

Using the expected values from the E step, the complete likelihood is now a function of \(\hat{\epsilon}_i\) that is easily maximized

\[
\hat{\epsilon}_i = \log^{-1} \left[ \frac{\sum_{j=1}^{L} (G_i^j - A_{ij} + (G_i^j - A_{ij})R_{ij})}{\sum_{j=1}^{L} (G_i^j - A_{ij} + (G_i^j - A_{ij})A_{ij})} \right]
\]

After we run QuASAR to infer genotypes across samples from the same individual, for each site we have a posterior probability of each genotype \(G_i^j\), and a base-calling error \(\hat{\epsilon}_i\) 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.

### 2.2.3 ASE inference

To detect ASE, we only consider sites with an heterozygous MAP control overdispersion using a beta-binomial model (combining (7) and (8)):

\[
Pr(R_{ij}|N_{ij}, \psi_{ij}, M_i) = \frac{(N_{ij})^{(M_i)} \Gamma(M_i + \psi_{ij}M_i) \Gamma(A_{ij} + (1 - \psi_{ij})M_i)}{R_{ij}^{(M_i)} \Gamma(N_{ij} + M_i) \Gamma(\psi_{ij}M_i) \Gamma((1 - \psi_{ij})M_i)}
\]

where \(M_i\) controls the effective number of samples supporting the prior belief that \(\rho = 0.5\) and is estimated using grid search:

\[
\hat{M}_i = \arg \max_{M_i} \prod_{i=1}^{L} Pr(R_{ij}|N_{ij}, \hat{\epsilon}_i, \rho_{ij} = 0.5, M_i)\]

(14)

We estimate \(\hat{\rho}_{ij}\) using (13) with \(M_i = \hat{M}_i\) from (14) and a standard gradient method (L-BFGS-B) to maximize the log-likelihood function

\[
I(\rho_{ij}; M_i, \hat{\epsilon}_i) = \log Pr(R_{ij}|N_{ij}, \psi_{ij} = \psi(\rho_{ij}, \hat{\epsilon}_i), M_i)
\]

(15)

Finally, all parameters are used to calculate the LRT statistic in (9) and its \(P\) value.

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

\[
\hat{\sigma}_{\rho_{ij}} = \left[ \frac{\partial^2 I(\rho_{ij}; M_i, \hat{\epsilon}_i)}{\partial \rho_{ij}^2} \right]_{\rho_{ij} = \hat{\rho}_{ij}}^{1/2}
\]

(16)

Alternatively, we can also recover a standard error from (9) (as is asymptotically distributed as \(\chi^2_{L-1}\)) by using the \(P\) value to back solve for the standard error:

\[
\hat{\sigma}_{\rho_{ij}} = \frac{\hat{\rho}_{ij}}{Q(\hat{P})}
\]

(17)

where \(Q()\) is the quantile function for a standard normal distribution and \(\hat{P}\) is the \(P\) value from (9). We use the first form (16) when \(\hat{\rho} \sim 0.5\) and (17) otherwise, as each provides a better approximation at those ranges. Alternatively, if we do not need \(\hat{\sigma}_{\rho_{ij}}\), we can use (15) to obtain a profile likelihood confidence interval for \(\rho_{ij}\).

### 2.3 Experimental data

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 h 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 phosphate-buffered saline, lysed on the plate, using Lysis/Binding Buffer (Ambion) and frozen at −80°C. 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 bar-coded DNA adapters (NEXTflex-96 RNA-seq Barcodes, Bioneer Scientific). Double-sided SPRI size selection (SPRSelect Beads, Beckman Coulter) was performed to select 350–500 bp fragments. The libraries were then amplified by polymerase chain reaction and pooled for sequencing on the Illumina HiSeq 2500 at the University of Chicago Genomics Core.

For each LCL sample, libraries from nine replicates were pooled for a total of 42.3M and 34.9M 50-bp PE reads, respectively. We subsequently resequenced this libraries for a total of 81.0M and 78.6M 150-bp PE reads. For the HUVEC samples (267M reads total), we collected data for 18 replicates across six time points to capture a wide range of basal physiologic conditions.

### 2.4 Pre-processing

To create a core set of SNPs for ASE analysis, we first removed rare (MAF < 5%) variants from all 1KG SNPs (see Supplementary Material for an analysis without this step). We also removed SNPs within 25 bases upstream or downstream of another SNP or short InDels as well as SNPs in regions of annotated copy numbers or other blacklisted regions (Degner et al., 2012). 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 the core set of SNPs using samtools mpileup command. Reads with an SNP at the beginning or at the end of the read, or with indels were also removed to avoid any potential experimental bias. Finally, the pile-ups 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.

### 2.5 Comparison with other methods

To assess QuASAR genotyping quality, we focused on the LCL data and compared genotype calls to those of 1KG project. We also compared QuASAR genotyping accuracy to samtools + bcftools (Li and Durbin, 2009) and GATK (DePristo et al., 2011). To assess QuASAR ASE inference, we used QQ plots and eQTL derived from the GEUVADIS dataset (Lappalainen et al., 2013; Wen et al., 2014).
When we compared QuASAR ASE inference to other methods, we used our implementation of the binomial and beta-binomial test by forcing QuASAR to ignore the genotyping uncertainty in the ASE inference step; i.e. in (9) numerator we only consider $\rho_d = 0.5$. For the binomial test, we further force $M_s = \infty$. In the Supplementary Material, we also compared QuASAR performance to the method used in (Kukurba et al., 2014). To make the results comparable, we used QuASAR genotype information and the same pre-processing steps described above up to samtools mpileup.

3 Results

We implemented the QuASAR approach as detailed in Section 2 in an R package available at http://github.com/piquelab/QuASAR. First, to evaluate QuASAR genotyping accuracy, we collected RNA-seq data for two LCLs that already have high-quality genotype calls from the 1Kg project (GM18507 and GM18508). As illustrated in Table 1, we are able to accurately genotype thousands of loci with lower error rates than other methods commonly used for genotyping DNA-seq data. By design, QuASAR is more conservative in making heterozygous calls than homozygous calls, yet still retains a large number of heterozygous loci compared with GATK (DePristo et al., 2011) and samtools (Li and Durbin, 2009). In QuASAR, if there is contention between (i) an heterozygous genotype call with extreme allelic imbalance or (ii) an homozygous genotype call with base call errors, our model will favor the latter (ii). This is a crucial feature for accurate inference of ASE, which will be discussed later in more detail. As we increase the threshold on the QuASAR posterior probability of heterozygosity, we consistently reduce the genotype error rate, Supplementary Table S1. The error rates increase slightly if we include rare variants with MAF < 0.05 (Table S2) and are minimally affected when 1KG allele frequencies are not used as priors (Table S3). An important feature of QuASAR is that the genotyping information is also used for the next inference step.

We next sought to characterize QuASAR performance in genotyping and ASE inference from RNA-seq experiments sequenced at different depths. In total, we analyzed 18 samples (three replicates across six 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 (Fig. 2). As expected, we observed that our ability to detect heterozygotes (MAP > 0.99) increases with the sequencing depth. At a more modest sequencing depth of 16 million, we can still detect more than 1,000 heterozygous sites.

After obtaining the genotypes, we assessed whether there is evidence of ASE at any of the SNPs determined to be heterozygous. To conduct ASE inference, we used the LRT (9) statistic in QuASAR 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 increased with the number of SNPs we were able to genotype, which in turn is a function of coverage (Fig. 2). Using 100 million reads, we detected roughly 9,000 heterozygous SNPs of which 50 have ASE at 10%FDR.

To assess the calibration of the ASE test in QuASAR and to compare it to other ASE inference approaches, we used QQ plots, 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 1:1 line for the range of $P$ values with higher value. 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. Many existing approaches for ASE use either a binomial test (Degner et al., 2012; Kukurba et al., 2014; McDaniell et al., 2010; Reddy et al., 2012) or a beta-binomial test (Pickrell et al., 2010; Sun, 2012) that does not account for genotyping or base-calling error. To compare QuASAR ASE inference to these alternative methods, we calculated the empirical power to detect the difference of ASE, which will be discussed later in more detail. As we increased the number of SNPs, we used our implementation of the binomial and beta-binomial test by comparing it to other ASE inference approaches, we used QQ plots, which 50 have ASE at 10%FDR.

<table>
<thead>
<tr>
<th>Sample (no. of reads, read length)</th>
<th>Genotyping method</th>
<th>Number of hets</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA18507 (4.23 × 10^7, 50 bp)</td>
<td>Samtools</td>
<td>3158</td>
<td>1.20%</td>
</tr>
<tr>
<td></td>
<td>GATK</td>
<td>4568</td>
<td>0.91%</td>
</tr>
<tr>
<td></td>
<td>QuASAR</td>
<td>3310</td>
<td></td>
</tr>
<tr>
<td>NA18508 (3.49 × 10^7, 50 bp)</td>
<td>Samtools</td>
<td>2718</td>
<td>0.63%</td>
</tr>
<tr>
<td></td>
<td>GATK</td>
<td>3933</td>
<td>0.97%</td>
</tr>
<tr>
<td></td>
<td>QuASAR</td>
<td>2844</td>
<td>0.56%</td>
</tr>
<tr>
<td>NA18507 (8.10 × 10^7, 150 bp)</td>
<td>Samtools</td>
<td>19 515</td>
<td>0.63%</td>
</tr>
<tr>
<td></td>
<td>GATK</td>
<td>24 522</td>
<td>0.97%</td>
</tr>
<tr>
<td></td>
<td>QuASAR</td>
<td>17 434</td>
<td>0.56%</td>
</tr>
<tr>
<td>NA18508 (7.86 × 10^7, 150 bp)</td>
<td>Samtools</td>
<td>16 998</td>
<td>2.12%</td>
</tr>
<tr>
<td></td>
<td>GATK</td>
<td>21 587</td>
<td>2.09%</td>
</tr>
<tr>
<td></td>
<td>QuASAR</td>
<td>15 326</td>
<td>1.14%</td>
</tr>
</tbody>
</table>

**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, 2 – triangle). The x-axis represents 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.
approaches, we used the RNA-seq data collected on the LCL samples and the genotype calls from QuASAR. Figure 4 clearly shows that the binomial test is too optimistic and will likely lead to many false discoveries. Another binomial test (Supplementary Fig. S1) independently implemented (Kukurba et al., 2014) show a similar behavior to our own implementation. In contrast to the binomial test, the beta-binomial model seems better calibrated but uncertainty on the genotype being a true heterozygote can lead to very small \( P \) values and false positives. QuASAR combines the beta-binomial model with uncertainty on the genotype, resulting in the most conservative approach, likely avoiding a common cause of false positives in ASE analysis.

To further evaluate the differences between QuASAR and other methods, we focused on the high-coverage eQTL studies such as GEUVADIS (Lappalainen et al., 2013) including RNA-seq data for more than 600 individuals. From a recent reanalysis of the GEUVADIS dataset (Wen et al., 2014), we selected genes with eQTLs (5% FDR) and a leading SNP with a high posterior inclusion probability (PIP > 0.9), as those are the nucleotides more likely to be causal (i.e. the eQTN). We hypothesized that in our data, genes for which the eQTN was heterozygote would be more likely to show evidence for ASE than those for which the eQTN was homozygote. However, transcripts with homozygous eQTNs could still show ASE, if they have additional personal/rare associations or due to an imprinting mechanism. In any of the methods we tested (Supplementary Fig. S2), there was a significant difference between ASE \( P \) values in the expected direction (Mann–Whitney \( U \)-test \( P < 0.02 \)). The distribution of \( P \) values (Supplementary Fig. S2D) for ASE signals in transcripts with homozygous eQTNs are much closer to the uniform distribution in QuASAR when compared with other methods, especially those that do not account for overdispersion.

In general, the \( P \) values obtained across all methods tested are very much correlated to each other, as shown in Supplementary Figure S3, but there are key differences. The methods that use a binomial test tend to have lower \( P \) values than QuASAR, which uses a beta-binomial model. Supplementary Figure S4A and B show that we may observe more shrinkage for SNPs with deeper coverage. Compared with a test using the beta-binomial distribution but ignoring genotype uncertainty, our QuASAR test leads to similar \( P \) values, but SNPs with more uncertainty on being heterozygous are corrected in a higher degree toward a less significant \( P \) value (Supplementary Figs. S3C and S4C).

In terms of computational complexity, QuASAR is very fast. QuASAR runtime for genotyping with the EM procedure takes less than 20 s on any of our datasets (using a computer with a 6-core 2.93-GHz Intel processor and 8-GB RAM). Each EM iteration in the genotyping step is \( O(\text{LS}) \) linear with the number of SNPs and samples, and convergence is achieved in less than 20 iterations. GATK with the default options takes a longer amount of time (about 2 h) compared with samtools and QuASAR, as the latter methods focus directly on 1KG SNPs only. Comparatively, for both our pipeline and Kukurba et al. (2014), more time is spent in data pre-processing (roughly 10 min per sample depending on the sequencing depth and assuming reads are already aligned) than in genotyping or ASE inference.

### 4 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 available from the same individual across many experimental
samples (data types, conditions, cell types or technical replicates), the more certainty we can gather about the genotype.

A key aspect of the QuASAR ASE inference step is that it takes into account over-dispersion and genotype uncertainty resulting in a test that we have shown 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 what additional assumptions may be introduced and if such methods 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 article, 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.

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

Acknowledgements

We thank Wayne State University HPC 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

This work was supported by the National Institute of Health [R01GM109215-01 to R.P.-R. and F.L.] and the American Heart Association [14SDG20440188 to F.L.].

Conflict of Interest: none declared.

References


Nicola,D.L. et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet., 6, e1000888.


