Published in final edited form as:

Biometrics. 2019 September ; 75(3): 864–874. doi:10.1111/biom.13026.

A Statistical Method for Joint Estimation of *Cis*-eQTLs and Parent-of-Origin Effects under Family Trio Design

Vasyl Zhabotynsky,

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A

Kaoru Inoue,

National Institute for Environmental Health Sciences, Research Triangle Park, North Carolina, U.S.A.

Terry Magnuson,

Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

J. Mauro Calabrese,

Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

Wei Sun

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.

Summary:

RNA sequencing allows one to study allelic imbalance of gene expression, which may be due to genetic factors or genomic imprinting (i.e., higher expression of maternal or paternal allele). It is desirable to model both genetic and parent-of-origin effects simultaneously to avoid confounding and to improve the power to detect either effect. In studies of genetically tractable model organisms, separation of genetic and parent-of-origin effects can be achieved by studying reciprocal cross of two inbred strains. In contrast, this task is much more challenging in outbred populations such as humans. To address this challenge, we propose a new framework to combine experimental strategies and novel statistical methods. Specifically, we propose to study genetic and imprinting effects in family trios with RNA-seq data from the children and genotype data from both parents and children, and quantify genetic effects by *cis*-eQTLs. Towards this end, we have extended our method that studies the eQTLs of RNA-seq data (Sun, *Biometrics* 2012, 68(1): 1–11) to model both *cis*-eQTL and parent-of-origin effects, and evaluated its performance using extensive simulations. Since sample size may be limited in family trios, we have developed a data

wsun@fredhutch.org.

SUPPORTING INFORMATION

Supporting Information, referenced in Section 2–4, is available with this paper at the *Biometrics* website on Wiley Online Library. The raw RNA-seq data have been deposited in the NCBI SRA database at <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA385599/>. We have included our R package as well as our data analysis pipeline in a zip file as a web supplement. We will update this R package and pipeline and post the most recent version at <https://github.com/Sun-lab/rcppreqtl>, and <https://github.com/Sun-lab/eQTL-PoO-analysis-pipeline>, respectively.

analysis pipeline that borrows information from external data of unrelated individuals for *cis*-eQTL mapping. We have also collected RNA-seq data from the children of 30 family trios, applied our method to analyze this dataset, and identified some previously reported imprinted genes as well as some new candidates of imprinted genes.

Keywords

allele-specific expression; eQTL; imprinting; RNA-seq

1. Introduction

By combining RNA-seq data and phased genotype data from a diploid genome, one can estimate allele-specific expression (ASE). Allelic imbalance of gene expression (i.e., difference between the ASE of the two alleles) may be controlled by genetic factors or genomic imprinting. Genetic loci that affect gene expression variation are often referred to as gene expression quantitative trait loci (eQTLs). Sophisticated statistical methods have been developed to identify eQTLs using both ASE and total expression (Sun, 2012; Sun and Hu, 2013; Hu et al., 2015). Genomic imprinting is a phenomenon in which the ASE of a gene depends on whether the allele has maternal or paternal origin. Some genes are completely imprinted in the sense that they are exclusively expressed from one of the two parental alleles. However, recent studies have shown that many genes are partially imprinted, i.e., both alleles are expressed while the allele from one parent has higher expression than the other allele (Crowley et al., 2015). Although only a small percentage of human genes are (completely or partially) imprinted, studying imprinting is of great importance for at least two reasons. First, imprinted genes play important roles in many common diseases such as obesity or psychiatric disorders (Peters, 2014). Second, genomic imprinting is also a great model to study epigenetic regulation of gene expression (Barlow, 2011).

It is important to jointly estimate genetic and parent-of-origin (PoO) effects in order to avoid confounding of these two factors and to improve the power to detect either effect. This goal can be achieved by studying RNA-seq data from reciprocal crosses of two inbred strains, e.g., an F1 mouse population (Zou et al., 2014; Crowley et al., 2015). Specifically, when considering two inbred strains, A and B , a reciprocal cross refers to two crosses with ordered genotype $A^m \times B^p$ and $B^m \times A^p$, where the superscripts m and p indicate maternal and paternal origin, respectively. In the cross $A^m \times B^p$, if the expression of the A^m allele is larger than the B^p allele, it is not clear whether it is due to genetic effect (i.e., A allele has higher expression) or PoO effect (i.e., the maternal allele has higher expression). However, given a reciprocal cross, such confounding can be resolved.

In an outbred population such as the human population, separation of genetic and PoO effect becomes much more challenging. Previous studies have employed genotyping arrays to assess allelic imbalance using both RNA and genomic DNA samples and identify potential imprinted genes if the degrees of allelic imbalance are significantly different between RNA and DNA samples. Those potential imprinted genes are then screened using family pedigree data (Morcos et al., 2011) or experimental validation (Barboux et al., 2012). The accuracy of

such approaches is limited by the sensitivity of genotyping arrays and the ad hoc screening criterion. A more direct approach is to study ASE using RNA-seq data. Without deferring to family design, one can identify potentially imprinted genes from unrelated individuals by selecting genes with strong allelic imbalance that cannot be explained by other factors (Baran et al., 2015; Kumasaka et al., 2016). However, larger sample sizes are needed and only genes with very strong allelic imbalance can be identified from unrelated individuals.

Given the relatively low cost of RNA sequencing and genotyping arrays, we propose a new approach to identify imprinted genes using a new statistical method designed to analyze RNA-seq data from the children of family trios as well as genotype data from all family members. Our statistical method is built on our previous work to study the eQTLs of gene expression using both total expression (Total Read Count, or TReC) and ASE (TReCASE) (Sun, 2012). The main contributions of this paper are threefold: first, we extended the TReCASE model to study both eQTL and PoO effects in family trios; second, we developed a data analysis pipeline that borrows information from external population of unrelated individuals for eQTL mapping, and third, we generated RNA-seq data from 30 children of family trios, applied our method to analyze this dataset, and identified new candidate imprinted genes. Our method also provides a generalizable framework for analyzing the effects of multiple genetic or epigenetic factors on both total expression and allele-specific expression.

2. Method

2.1 An overview of our method

We first clarify the meaning of *cis*-eQTL because our method is designed to quantify genetic effects through this type of eQTL. *Cis*-acting regulation is due to DNA variation that directly influences the transcription process in an allele-specific manner. Alternatively, *trans*-acting regulation affects the gene expression by modifying the activity (or abundance) of the factors that regulate the gene, which leads to the same amount of expression changes for both alleles (Wittkopp et al., 2004). An eQTL is a *cis*-eQTL if it alters the expression of the two alleles of its target gene differently, otherwise it is a *trans*-eQTL (Sun and Hu, 2013). Most local eQTLs are *cis*-eQTLs and thus local eQTLs are often referred to as *cis*-eQTLs. In this paper, we use a more strict definition of *cis*-eQTL such that *cis*-eQTL must be associated with allelic imbalance of gene expression.

Our method models the expression of each gene separately, and thus when discussing the details of our method, we drop the index for each gene for notational convenience. We assume genetic effects can be captured by a small number of *cis*-acting eQTLs. This assumption can be justified by the following rationale. First, by definition, only *cis*-acting eQTLs are relevant for ASE study. Second, the degrees-of-freedom of such *cis*-acting eQTLs is limited because most *cis*-acting eQTLs are local eQTLs around a gene and the linkage disequilibrium (LD) limits their degree-of-freedom. In the following, we assume there is only one *cis*-eQTL to simplify the discussion. Our method can be extended to the cases with multiple *cis*-eQTLs, though we expect a larger sample size is needed in order to accurately detect multiple *cis*-eQTLs given that they are in strong LD. In the simulation Section, we also consider the situation when there is a secondary *cis*-eQTL but we miss it in our model.

The conclusion is that it may bias the estimate of an eQTL effect but has little impact on the estimation PoO effect (Supporting Information Section D.5).

To help explain the motivation of the model choice and parameterization of our method, we start by a toy example to illustrate how *cis*-eQTL and PoO factors affect both total expression and ASE. Consider the expression of one gene in four individuals, and a *cis*-eQTL of this gene with ordered genotype CC, TT, TC or CT with the first allele being maternal allele (Figure 1). In other words, “ordered genotype” is different from standard genotype in that two heterozygous genotypes TC and CT have different meanings. The genetic effect is that C allele has three times expression of T allele, and the PoO effect is that paternal allele has twice of the expression of maternal allele.

It is apparent that there is allelic imbalance in all four individuals. The degree of allelic imbalance depends on both *cis*-eQTL and PoO effects and it is very challenging to discern them without knowing the parent-of-origin of each allele. This toy example also demonstrates that in addition to ASE, both *cis*-eQTL and PoO factors can affect total expression. *Cis*-eQTL’s effect on total expression is more apparent since total expression decreases as genotype shift from CC, CT/TC, to TT. PoO modifies the total expression of the two individuals with ordered genotype CT or TC. Without PoO effect, these two individuals should have an equal amount of total expression.

We have previously developed a statistical method to map *cis*-eQTLs by jointly modeling TReC and ASE (Sun, 2012). In this paper, we extend this method to study PoO effect. We model TReC by a negative binomial distribution and ASE by a beta-binomial distribution. Our statistical framework is general. It can be used to study other types of *cis*-acting factors in other types of sequencing data, such as sequencing data for DNA methylation, open chromatin regions, or certain epigenetic marks.

2.2 Allele specific read counts (ASReCs)

We start with the statistical model for ASE, or more specifically, ASReCs for RNA-seq data. Then the model for total expression (or total read count, TReC) can be derived by adding up the expression of the two alleles. For a gene of interest, we denote the observed ASReCs in the k -th individual (for $k=1, \dots, K$) to be N_{k1} and N_{k2} for two arbitrarily defined haplotypes, denoted by h_{k1} and h_{k2} , respectively. Let $N_k = N_{k1} + N_{k2}$. To study ASReCs, we model the distribution of N_{k1} given N_k by a beta-binomial distribution, which extends the binomial distribution by allowing the variance to be larger than the binomial variance. This extra amount of variance is usually attributed to the biological variation across individuals. If one collects multiple technical replicates from the same individual, it is likely that a binomial distribution is sufficient to model the ASReC data.

Our model is an example to study the association between the ASE of one gene and multiple covariates using the framework of a generalized linear model. We need to first define the design matrix such that they reflect the allele-specific nature of the covariates. In our study, the two haplotypes of a gene are characterized by two covariates: parent-of-origin and a candidate *cis*-eQTL. Using genotype data from a family trio, we can impute the most likely haplotypes of each gene for both parents and the child, and then it is easy to assign the

parent-of-origin status of the two haplotypes of the child. To associate each haplotype with the eQTL genotype, we assume the haplotypes connecting a candidate *cis*-eQTL and the gene of interest have been inferred, e.g., using a phasing method on un-phased genotype data (Sun and Hu, 2013). Denote the two alleles of the candidate eQTL as A_1 and A_2 , and denote its ordered genotype in the k -th individual as g_k , which may take one of four values: A_1A_1 , A_1A_2 , A_2A_1 , or A_2A_2 . Here the ordered genotype is defined such that the first allele is from haplotype h_{k1} and the second allele is from haplotype h_{k2} . We model N_{k1} by a beta-binomial distribution (Griffiths, 1973), denoted by f_{BB} :

$$N_{k1} \sim f_{BB}(N_{k1}; N_k, \pi_k, \phi), \quad \log\{\pi_k/(1 - \pi_k)\} = b_0 z_k + b_1 x_k, \quad (1)$$

where ϕ is a gene-specific over-dispersion parameter, π_k is the expected proportion of reads coming from haplotype h_{k1} , and b_0 and b_1 model *cis*-eQTL and PoO effect, respectively. In other words, we model the contrast of the expression of haplotype h_{k1} versus the expression of haplotype h_{k2} by $\log\{\pi_k/(1 - \pi_k)\}$. Based on this model setup, we need to construct a design matrix for *cis*-eQTL effect (z_k) and PoO effect (x_k). If the eQTL has homozygous genotype, $\pi_k=0.5$, and thus $\log\{\pi_k/(1 - \pi_k)\} = 0$. This implies that $z_k=0$ for a homozygous genotype. For a heterozygous genotype, the *cis*-eQTL effect for phased genotype A_1A_2 should be the inverse of the effect for A_2A_1 , and thus we have

$$z_k = \begin{cases} 0 & \text{if } g_k = A_i A_i, \quad i = 1 \text{ or } 2, \\ 1 & \text{if } g_k = A_2 A_1, \\ -1 & \text{if } g_k = A_1 A_2. \end{cases}$$

For the PoO effect (x_k), it should be reversed when we switch the PoO of haplotype h_{k1} , and thus

$$x_k = \begin{cases} 1 & \text{if haplotype } h_{k1} \text{ is inherited from the paternal allele,} \\ -1 & \text{if haplotype } h_{k1} \text{ is inherited from the maternal allele.} \end{cases}$$

Therefore b_0 models the *cis*-eQTL effect of A_2 allele versus A_1 allele, and b_1 models the effect of paternal allele versus maternal allele. The hypothesis testing of either effect can be written as $H_0: b_0=0$ or $H_0: b_1=0$.

The above setting is general for modeling ASReC in the sense that the two haplotypes (h_{k1} and h_{k2}) are arbitrarily defined. In our study, the PoO information for each haplotype is known. Therefore in the following discussions, to simplify notations, we assume haplotypes h_{k1} and h_{k2} correspond to maternal and paternal haplotype, respectively. In this setting, we can elucidate the roles of b_0 and b_1 by considering the value of $\log\{\pi_k/(1 - \pi_k)\}$ for four ordered genotypes of the potential *cis*-eQTL:

$$\log\{\pi_k/(1-\pi_k)\} = \begin{cases} -b_1 & \text{if } g_k = A_i^m A_i^p, \ i = 1 \text{ or } 2, \\ b_0 - b_1 & \text{if } g_k = A_2^m A_1^p, \\ -b_0 - b_1 & \text{if } g_k = A_1^m A_2^p, \end{cases}$$

where the superscripts m and p denote maternal and paternal allele, respectively.

2.3 Total read counts (TReCs)

Denote the TReC of the gene of interest in the k -th individual by T_k . We model T_k by a negative binomial distribution with mean μ_k and a gene-specific over-dispersion parameter ϕ , denoted by $f_{NB}(T_k; \mu_k, \phi)$. Let $\mu_{A_1}^m$ be the expected value of TReC for A_1 allele if it is of maternal origin. $\mu_{A_1}^p$, $\mu_{A_2}^m$, and $\mu_{A_2}^p$ can be defined similarly. Based on the aforementioned parametrization for ASReC data, we have

$$\log(\mu_{A_2}^p / \mu_{A_1}^p) = \log(\mu_{A_2}^m / \mu_{A_1}^m) = b_0, \quad \log(\mu_{A_2}^p / \mu_{A_2}^m) = \log(\mu_{A_1}^p / \mu_{A_1}^m) = b_1. \quad (2)$$

Then assuming haplotypes h_{k1} and h_{k2} correspond to the maternal and paternal haplotypes, we can formulate the mean value of TReC data by adding up the expression of the two alleles:

$$\mu(g_k) = \begin{cases} \mu_{A_1}^m + \mu_{A_1}^p = \mu_{A_1}^p \left\{ \exp(-b_1) + 1 \right\} & \text{if } g_k = A_1^m A_1^p, \\ \mu_{A_1}^m + \mu_{A_2}^p = \mu_{A_1}^p \left\{ \exp(-b_1) + \exp(b_0) \right\} & \text{if } g_k = A_1^m A_2^p, \\ \mu_{A_2}^m + \mu_{A_1}^p = \mu_{A_1}^p \left\{ \exp(b_0 - b_1) + 1 \right\} & \text{if } g_k = A_2^m A_1^p, \\ \mu_{A_2}^m + \mu_{A_2}^p = \mu_{A_1}^p \left\{ \exp(b_0 - b_1) + \exp(b_0) \right\} & \text{if } g_k = A_2^m A_2^p. \end{cases}$$

If we use group $A_1^m A_1^p$ as a reference group, we can write the mean structure for negative binomial distribution as:

$$\log(\mu_k) = \gamma_0 + \beta_k \log(\kappa_k) + \sum_{u=1}^p \beta_u c_{ku} + \eta_k. \quad (3)$$

Here κ_k is the read-depth measurement for the k -th individual. In this paper, we set κ_k to be the 75-th percentile of TReC across all the genes in the k -th individual, which is more robust than using genome-wide total RNA-seq read counts (Bullard et al., 2010). c_{ku} denotes the value for the u -th covariate (e.g., age, gender, batch effects) for the k -th individual. There are

three sets of coefficients in the linear part of this model $\{\gamma_0 + \beta_k \log(\kappa_k) + \sum_{u=1}^p \beta_u c_{ku}\} : \gamma_0$ is an intercept, β_k is the regression coefficient associated with κ_k , and β_u 's ($u=1, \dots, p$) are the regression coefficients for p additional covariates. The nonlinear part of this model (η_k) is defined as $\eta_k = \mu(g_k) - \mu(g_k = A_1^m A_1^p)$. Note that the *cis*-eQTL and PoO effects are parametrized by b_0 and b_1 , which are the same as the b_0 and b_1 in equation (1) for ASReC.

2.4 Joint likelihood

For a specific gene, the joint likelihood of TReC and ASReC is

$$\mathbb{L}(\Theta) = \prod_{k=1}^K f_{BB}(N_{k1}, N_k; b_0, b_1, \phi) f_{NB}(T_k; b_0, b_1, \phi, \gamma_0, \beta_k, \beta_1, \dots, \beta_p), \quad (4)$$

where $\Theta = (b_0, b_1, \phi, \phi, \gamma_0, \beta_k, \beta_1, \dots, \beta_p)$. In this model, TReC and ASReC data share the same *cis*-eQTL and PoO effect. We test *cis*-eQTL and PoO effects by testing whether b_0 or b_1 equals 0, respectively, with a likelihood ratio test.

2.5 Optimization Algorithm

Let $\mathbf{X}$ be an $n \times (p+1)$ design matrix constructed using the covariates $\log(\kappa_k)$ and $c_{k1}, \dots, c_{kp}$ defined in equation (3). Let $\boldsymbol{\beta} = (\beta_k, \beta_1, \dots, \beta_p)^T$. For given initial values of all the parameters, our optimization algorithm includes the following steps.

Step 1: Optimize linear terms:

$$\beta_{r+1} = \beta_r + (X'W_r X)^{-1} (X'W_r \tilde{y}_r), \quad \text{diag}(W_r) = \frac{\mu_r}{1 + \phi_r^{-1} \mu_r} \text{ and } \tilde{y}_r = \frac{y_r - \mu_r}{\mu_r},$$

where r indexes iteration, W_r is a diagonal matrix.

Step 2: Estimate b_0 and b_1 . The following two sub-steps are redundant. We use both to improve the robustness of the algorithm when sample size is small.

1. Optimize b_0 and b_1 together using BFGS method.
2. Optimize b_0 and b_1 separately using Brent algorithm.
3. Proceed using the results of (1) or (2) depending on which one gives higher likelihood.

Step 3: Optimize over-dispersion parameters $\log(\phi)$ and $\log(\varphi)$ separately using Brent algorithm. For stability we limit the range of over-dispersion to be between 10^{-4} and 10^4 .

Step 4: If the likelihood change is larger than a small number ϵ , go to Step 1, otherwise finish the estimation process.

We have carefully checked the convergence and robustness of our algorithm by two approaches. First, we did grid searches on 10,000 simulation datasets with 1,000 grid points

around the initial MLE for each dataset. In all of the 10,000 simulation datasets, all 1,000 grid searches ended up with very similar MLEs. Next, we checked whether our algorithm is robust to the choice of initial values. For 10,000 simulation datasets, we randomly chose 100 initial values. MLE is wrong for 35 of the 10,000 100 simulation settings (0.0035%), and in all these 35 cases, the initial values are very far away from the true MLE. Therefore, we conclude that our algorithm is fairly robust to the choice of initial values.

3. Simulations

3.1 Simulation setup

We evaluated our method using two sets of simulations. In the first set of simulations, we simulated data for one gene with several effect sizes to evaluate power and highlight the benefit to jointly estimate *cis*-eQTL and PoO effects. This simple simulation setting helps us focus on a few issues that are generic for all genes. In the second set of simulations, we simulated data for 10,000 genes with 10 effect sizes for *cis*-eQTL effects, 10 effect sizes for PoO effects, as well as 10 categories for total read count per gene, and we simulated the over-dispersion parameter as well as the proportion of allele-specific reads according to empirical distributions observed in real data. Due to space limitation, we defer the results of the second set of simulations to Supporting Information Section D.5.

In the first set of simulations, we simulated TReC and ASReC from the model described by the joint likelihood in equation (4). We chose the smallest sample size to be 32, which is similar to the sample size we have for real data analysis. To demonstrate the asymptotic properties, we also simulated data of larger sample sizes of 64, 128, and 256. We set the over-dispersion parameters for the beta-binomial (ϕ) and negative binomial (ϕ) distributions to be $\frac{1}{4}$, and $\frac{4}{3}$, respectively, which are fairly typical in real data. We included several covariates to mimic the real data analysis setting, including intercept, two categorical variables and one continuous variable that can be interpreted as library size. We further scaled the expected counts so that the mean TReC was about 250. We set the proportion of reads that are allele-specific to be 10%, which is a reasonable value based on our real data analysis results.

3.2 Simulation results

Our simulation results show that the estimates of b_0 and b_1 from our method are unbiased. In contrast, if a model is fit with only *cis*-eQTL or PoO effect while in fact both effects are present, the parameter estimates have non-ignorable bias, even at a larger sample size of 256 (Figure 2 a-b, Table 1). This bias is unexpected since b_0 and b_1 should be orthogonal. We further explore the reason for such bias in Section E.3 of Supporting Information. It appears if we ignore one effect (e.g., b_0) and only estimate b_1 , it leads to larger estimates of over-dispersion of beta-binomial distribution, and that is likely the reason for the downward bias of b_1 .

We also compared the type I error and the power of our joint model versus a naive fit using the *R glm.nb* function to fit a negative binomial model for TReC data and the *R vglm* function to fit a beta-binomial model for ASReC data. As shown in Table 2, with a sample

size of 32, the simple models do not control type I error as well as the joint model and that the joint model has higher power to detect both *cis*-eQTL and PoO effects.

Even at a sample size of 32, our method has around 80% of power to detect either *cis*-eQTL or PoO effect at a two-fold change, which corresponds to the effect size of $\log_2 0.693$ (Figure 2 c-d, Table 2). The power to detect PoO effects is higher than the power to detect *cis*-eQTL effects. This is because the ASE of all samples can be used to quantify PoO effect (i.e., comparing gene expression of paternal allele vs. maternal allele). In contrast, ASE can be used to quantify *cis*-eQTL effect only if the genotype of the candidate eQTL is heterozygous, which will only happen in a subset of the samples.

The computational time of our method scales well with the sample size. The time per gene remains around one second or less for sample sizes under 200 (Figure S17 in Supporting Information).

3.3 Additional simulation results.

Our model assumes that imprinting effects apply to all samples, and both genotype calls and parent of origin assignments of haplotypes are accurate. We conducted additional simulations to evaluate the performance of our model when these assumptions are violated. Our method always controls type I errors when these assumptions are violated. As expected, when imprinting only affects a fraction of samples or when parent of origin assignments of haplotypes are wrong, we lose power to detect parent of origin effects and with no or minor effect on the power to detect genetic effects. In contrast, with genotyping errors, we lose power to detect genetic effects but the power to detect parent of origin effects remains similar. See Supporting Information Section D.2-D.4 for details.

4. Application

4.1 Data collection

We collected RNA-seq data from 30 HapMap CEU (Caucasian) samples (15 males + 15 females). All of these samples are children of family trios, where the genotypes of all three family members have been reported in the HapMap project. For most of these samples, the RNA reads were 150 bp paired-end reads, with an additional smaller run of 75 bp paired-end reads. The median of the total number of reads for each of these 30 samples is about 20 million. These reads were mapped to hg38 human reference genome using Tophat2. The Hapmap project genotyped about 3.9 million SNPs for these 30 trios. We phased and imputed the genotypes of these 30 trios using SHAPEIT2 (Delaneau et al., 2014) and IMPUTE2 (Howie et al., 2012), against the 1000 Genome reference panel containing 2,504 individuals and 82 million SNPs. Finally, based on phased and imputed genotype, we extracted allele-specific reads (i.e., those RNA-seq reads that overlap with at least one heterozygous SNP), and counted the number of allele-specific reads for each haplotype of a gene.

4.2 Identification of candidate *cis*-eQTLs

The parents of these 30 HapMap family trios are part of the samples included in the 1000 Genomes Project (1000 Genomes Project Consortium, 2012). To improve the power and precision for eQTL mapping, we first mapped *cis*-acting eQTLs using the Caucasian samples of the 1000 Genome Project. Then given the eQTLs, we jointly estimated *cis*-eQTL and PoO effects in the 30 children of HapMap trios. This strategy is useful for similar studies because eQTL data are often available in larger sample size for unrelated individuals than family trios. However, the same data processing and analysis pipeline should be applied to both populations and systematic checking is needed to ensure the eQTLs identified from the larger population of unrelated individuals are consistent with the ones from the children of family trios.

Following this guideline, we re-processed the RNA-seq data from the 1000 Genomes project from the raw fastq files. We downloaded fastq files for 465 samples from the Geuvadis consortium (Lappalainen et al., 2013). We mapped all the reads to the hg38 reference, and then performed the same process of phasing and imputation of the genotypes of these samples. Then we calculated TReC and ASReC per gene and per sample. After filtering out non-European samples as well as samples with low quality in terms of systematic bias of ASE (see Section B.1 in Supporting Information for details), we performed eQTL mapping for 227 samples using the TReCASE method (Hu et al., 2015). For each gene, we identified one candidate *cis*-acting eQTL by selecting the one with the smallest p-value. Note that some genes do not have any *cis*-acting eQTLs because all the nearby SNPs show certain evidence of *trans*-acting eQTLs, i.e., the SNP genotype is associated with total expression, but not allele-specific expression. The eQTL effect sizes estimated from these 227 samples are consistent with the estimates from the 30 children of HapMap trios (see Section B.2 in Supporting Information for details).

4.3 Identification of imprinted genes

We used our method to estimate *cis*-eQTL and imprinting effects for 12,386 genes in the 30 children of the family trios. These genes were selected because they had enough expression in the 30 samples and they had at least one candidate *cis*-eQTL based on eQTL mapping results from the 227 samples of 1000 Genome Project. For the negative binomial model of TReC, we fit our model with additional covariates to capture batch effect (the RNA-seq data were collected through 3 batches with 10 samples per batch) and read-depth. No additional covariate is needed for the analysis of ASReC because our model for ASE compared the expression of one allele versus the other allele, and thus the effects of a covariate is canceled. We found 16 genes with significant imprinting effects at q-value cutoff 0.05 (Table 3).

Out of these 16 genes, 10 were known to be imprinted from previous studies and 6 were novel candidates for imprinting. For all of those 10 known imprinted genes, our estimates of allelic imbalance agree with what were reported before. The majority of these imprinted genes (14 of the 16 genes) have higher expression from the paternal allele than the maternal allele. At a more liberal cutoff of q-value 0.25, we identified 15 additional candidate imprinted genes. The genes that don't make q-value cutoff of 0.05 tend to have lower counts

and smaller PoO effects (Figure S13). By manually examining the expression pattern of these 15 genes, we found that 12 of them have consistent direction of allelic imbalance across most samples, and thus they may miss the q-value cutoff 0.05 due to low power (e.g., low read-depth) or small effect size (see Table 3 and Section E of Supporting Information).

We illustrate the read count data of a clearly imprinted gene, ZNF497, which has higher expression on maternal allele (Figure 3).

The imprinting effect can be observed from both TReC and ASReC. For TReC, the two groups with ordered genotype AA and AB have similar expression because they share the same maternal allele and maternal allele has higher expression. Similarly, the two groups BA and BB have similar expression. The effect of imprinting is more apparent in allele-specific reads where we observe the proportion of reads from the paternal allele is far below 50%.

We selected known imprinted genes based on the list reported by Morison et al. (2005) or those genes recorded in the Geneimprint database (Jirtle, 2016). There are a total of 90 genes classified as imprinted. Out of those, 32 were expressed in our dataset and had a candidate eQTL. Of these 32 genes 10 were found to be significant (q-value 0.05) by our method. For several genes (such as RB1, KCNQ1, PEG3, and PLAGL1), we observed signals of imprinting, but the signals were too weak to produce significant q-values. In general, though, we observed that even for insignificant results, those with relatively smaller q-values tend to have their estimated imprinting direction matched with the reported imprinting direction.

The imprinted genes in mouse genome tend to form clusters. It is interesting to check whether this is true for human genome. Based on the findings from our study, partly due to relatively small number of imprinted genes, we do not observe any large cluster of imprinted genes, though the imprinted genes on chromosomes 15 and 16 appear to form two clusters (Figure S14).

5. Discussion

We developed a systematic approach to jointly estimate *cis*-eQTL and PoO effect in human. Our results have a good consistency with the imprinted genes reported by previous studies. However, none of the genes classified as “predicted imprinting” instead of “imprinting” in Geneimprint database were detected as imprinted genes with high confidence in our results. One possible reason is that most of these genes were selected based on a screening paper (Luedi et al., 2007) and were false positives. We also noted that for these “predicted imprinting” genes, the proportion of genes with predominantly paternal expression is roughly 50%, while this proportion is about 68% (61/90) for known imprinting genes. Several imprinted genes that we found are non-coding RNAs, such as PWAR6, SNHG14, RP11-701H24.4 and RP1-309F20.3. The former three non-coding RNAs are found in the Prader Willi/Angleman region of hg38, on chromosome 15, and Prader Willi/Angleman disease is linked with genomic imprinting. Further studies are warranted to elucidate the functional consequence of these imprinted non-coding RNAs.

In exonic regions with at least one heterozygous SNPs, one haplotype (e.g., the paternal one) may have more reference alleles than the other haplotype. This may create RNA-seq

mapping bias because we map RNA-seq reads to reference genome. As shown by a comprehensive study, mapping bias has minimum effect on eQTL mapping (Panousis et al., 2014). In addition, we have explored the potential impact of mapping bias on estimating parent-of-origin effect and found its impact is ignorable (Supporting Information, Section A.9).

We have assumed haplotypes around a gene are known or it can be inferred from un-phased genotypes. Phasing uncertainty is less of a concern for our family trio study because we performed phasing using family information and thus the phasing results are reasonably accurate within a short distance of a gene. For eQTL mapping in unrelated individuals, we employed the eQTL mapping method from Hu et al. (2015), which does handle phasing uncertainty.

Another potential issue is that due to genotyping error, some SNP genotypes may be incorrectly labeled as heterozygous. If such a genotyping error happens for a couple of individuals, it can be accommodated by increasing the over-dispersion parameter estimate for ASReC. If it happens for more individuals, the effect sizes of TReC and ASReC will be different and we would estimate genetic effects using only TReC data and thus avoid confounding due to genotyping error. It is possible that such false positives of heterozygous genotypes may also affect an estimation of imprinting effect if they happen in a way consistent with parent-of-origin. We expect that such a situation is extremely rare, especially for studies with relatively larger sample sizes. To safeguard our results from the aforementioned complications, we have used a permutation test to evaluate our results. We observed good consistency between our model-based p-values and permutation p-values (Supporting Information, Section E.2).

We estimate gene-specific over-dispersion parameters for both the beta-binomial distribution (for ASReC) and the negative binomial distribution (for TReC). The estimation of over-dispersion parameters for RNA-seq data has been very well studied. When sample size is small (e.g., a two-group comparison with 2 or 3 samples per group), it is important to borrow information across genes to accurately estimate over-dispersion parameters (Anders and Huber, 2010; Zhou et al., 2011; Love et al., 2014). When sample size is large, borrowing information across genes may not be helpful (Zhou et al., 2011; Love et al., 2014). In fact, Zhou et al. (2011) showed that even for modest sample size (e.g., 5 or more samples per group for a two-group comparison), the benefit of borrowing information across genes may not be apparent. In our studies, the sample size is larger and thus it is not necessary to borrow information across genes to estimate over-dispersion parameters. In addition, we observed that the over-dispersion parameters for either the beta-binomial distribution or the negative binomial distribution vary substantially from gene to gene even when conditioning on gene expression level (Figures S7-S8), and thus borrowing information across genes requires careful modeling to avoid estimation bias. Therefore, we chose to estimate gene-specific over-dispersion parameters.

In our model (equation (1)), we assume the observed expression ratio between two alleles is the summation of genetic effect and PoO effect. However, it is likely that there is an interaction effect. It is interesting to explore why the genetic effect may depend on parent of

origin, for example, as pointed by an anonymous reviewer, it may be related with allele-specific DNA methylation. Our model can be extended to include such interaction effect, though larger sample size may be needed to reliably estimate such interaction effect, and ideally in a population with additional epigenetic data to help explain the underlying causes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research is supported in part by NIH grant GM105785.

REFERENCES

- 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. [PubMed: 23128226]
- Anders S and Huber W (2010). Differential expression analysis for sequence count data. *Genome biology* 11, R106. [PubMed: 20979621]
- Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, et al. (2015). The landscape of genomic imprinting across diverse adult human tissues. *Genome research* 25, 927–936. [PubMed: 25953952]
- Barboux S, Gascoin-Lachambre G, Buffat C, Monnier P, Mondon F, Tonanny M-B, et al. (2012). A genome-wide approach reveals novel imprinted genes expressed in the human placenta. *Epigenetics* 7, 1079–1090. [PubMed: 22894909]
- Barlow DP (2011). Genomic imprinting: a mammalian epigenetic discovery model. *Annual review of genetics* 45, 379–403.
- Bullard JH, Purdom E, Hansen KD, and Dudoit S (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC bioinformatics* 11, 94. [PubMed: 20167110]
- Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, et al. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics* 47, 353–360. [PubMed: 25730764]
- Delaneau O, Marchini J, Consortium GP, et al. (2014). Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel. *Nature communications* 5,.
- Griffiths D (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* pages 637–648. [PubMed: 4785230]
- Howie B, Fuchsberger C, Stephens M, Marchini J, and Abecasis GR (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* 44, 955–959. [PubMed: 22820512]
- Hu Y-J, Sun W, Tzeng J-Y, and Perou CM (2015). Proper use of allele-specific expression improves statistical power for cis-eQTL mapping with RNA-seq data. *Journal of the American Statistical Association* 110, 962–974. [PubMed: 26568645]
- Jirtle RL (2016). Gene imprint, imprinted gene database (internet) Available at <http://www.geneimprint.com/site/genes-by-species>.
- Kumasaka N, Knights AJ, and Gaffney DJ (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature genetics* 48, 206–213. [PubMed: 26656845]
- Lappalainen T, Sammeth M, Friedländer MR, ACt Hoen P, Monlong J, Rivas, et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506. [PubMed: 24037378]
- Love MI, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15, 550. [PubMed: 25516281]

- Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, and Hartemink AJ (2007). Computational and experimental identification of novel human imprinted genes. *Genome research* 17, 1723–1730. [PubMed: 18055845]
- Morcos L, Ge B, Koka V, Lam KC, Pokholok DK, Gunderson KL, et al. (2011). Genome-wide assessment of imprinted expression in human cells. *Genome biology* 12, 1.
- Morison IM, Ramsay JP, and Spencer HG (2005). A census of mammalian imprinting. *TRENDS in Genetics* 21, 457–465. <http://www.otago.ac.nz/IGC>. [PubMed: 15990197]
- Panousis NI, Gutierrez-Arcelus M, Dermitzakis ET, and Lappalainen T (2014). Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome biology* 15, 467. [PubMed: 25239376]
- Peters J (2014). The role of genomic imprinting in biology and disease: an expanding view. *Nature Reviews Genetics* 15, 517–530.
- Sun W (2012). A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* 68, 1–11. [PubMed: 21838806]
- Sun W and Hu Y (2013). eQTL mapping using RNA-seq data. *Statistics in biosciences* 5, 198–219. [PubMed: 23667399]
- Wittkopp P, Haerum B, and Clark A (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88. [PubMed: 15229602]
- Zhou Y-H, Xia K, and Wright FA (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* 27, 2672–2678. [PubMed: 21810900]
- Zou F, Sun W, Crowley JJ, Zhabotynsky V, Sullivan PF, and de Villena FP-M (2014). A novel statistical approach for jointly analyzing RNA-seq data from F1 reciprocal crosses and inbred lines. *Genetics* 197, 389–399. [PubMed: 24561482]

Figure 1.

This diagram illustrates *cis*-eQTL and parent-of-origin effects on gene expression in four individuals (individuals 1–4). The middle/side boxes indicate exonic/intronic regions, respectively. Assume there is a heterozygous SNP in the exonic region of this hypothetical gene in all four individuals, and thus we can quantify ASE. There is a *cis*-eQTL with C or T allele, and its C allele has three times of expression of T allele. In addition, the paternal copy has twice expression of the maternal copy. This figure appears in color in the electronic version of this article.

Figure 2.

(a)-(b): Estimation of additive *cis*-eQTL effect (a) and PoO effect (b) for various sample sizes, using joint model or short model that ignore the other effect. Left legend describes number of simulated individuals, solid lines are density curves of parameter estimates of joint model, and dashed lines are density curves for short models. (c-d): The statistical power to detect additive *cis*-eQTL effect (c) or PoO effect (d) for various sample sizes using joint model for short model. The horizontal dash line indicates the p-value cutoff 0.05 to declare statistical significance. This figure appears in color in the electronic version of this article.

Figure 3. ZNF497 - a maternally expressed gene. (a) normalized total read counts ($\log_{10}$ scale), (b) percent of paternally expressed reads. Reads are classified into four categories by the ordered genotype with the first allele being maternal copy and the second allele being paternal copy. This figure appears in color in the electronic version of this article.

Table 1

Simulation results with sample size 256. All the bias, variance, and MSE have been multiplied by 100. The first part of this table shows estimation bias, variance and MSE for eQTL effect. The full model includes both eQTL and PoO effect, and the short model only includes eQTL effect. The second part of the table shows the results for PoO effect.

Estimation of eQTL effect							
True effect size		bias		variance		MSE	
eQTL	PoO	full	short	full	short	full	short
0	0	0.00	0.00	0.75	0.75	0.75	0.75
0	0.5	-0.05	-0.07	0.76	0.73	0.76	0.73
0	1	-0.02	-0.06	0.83	0.72	0.83	0.72
0.5	0	0.03	0.00	0.77	0.77	0.77	0.77
0.5	0.5	0.14	-1.62	0.81	0.77	0.81	0.79
0.5	1	0.11	-5.45	0.83	0.72	0.83	1.02
1	0	0.21	0.14	0.87	0.87	0.87	0.87
1	0.5	-0.01	-3.32	0.88	0.85	0.88	0.96
1	1	0.13	-10.64	0.91	0.79	0.91	1.93
Estimation of PoO effect							
True effect size		bias		variance		MSE	
eQTL	PoO	full	short	full	short	full	short
0	0	0.03	0.03	0.45	0.45	0.45	0.45
0	0.5	0.05	0.02	0.48	0.47	0.48	0.47
0	1	0.20	0.13	0.56	0.56	0.56	0.56
0.5	0	0.01	-0.02	0.45	0.44	0.45	0.44
0.5	0.5	0.06	-1.06	0.51	0.49	0.51	0.50
0.5	1	0.22	-1.89	0.59	0.57	0.59	0.61
1	0	-0.15	-0.17	0.48	0.42	0.48	0.42
1	0.5	-0.04	-4.02	0.50	0.44	0.50	0.61
1	1	0.15	-7.55	0.62	0.54	0.62	1.11

Table 2

Power Analysis

Parameters		eQTL			PoO		
b_0	b_1	Joint Model	Negative Binomial	Beta Binomial	Joint Model	Negative Binomial	Beta Binomial
0.00	0.00	0.06	0.11	0.07	0.06	0.12	0.07
0.13	0.13	0.09	0.12	0.08	0.11	0.11	0.10
0.25	0.25	0.19	0.14	0.11	0.27	0.11	0.18
0.50	0.50	0.53	0.21	0.21	0.72	0.12	0.46
0.75	0.75	0.83	0.32	0.33	0.96	0.14	0.76
1.50	1.50	1.00	0.76	0.49	1.00	0.39	0.93

Table 3

A list of genes with potential PoO effects. We classify genes with q-value between 0.05 and 0.25 into three groups.

ID	estimate (CI)	Name	q-value	Expression	Known
ENSG00000269821	95%(91%;97%)	KCNQ1OT1	1.2e-08	paternally	yes
ENSG00000204186	96%(93%;98%)	ZDBF2	9.6e-08	paternally	yes
ENSG00000167981	8.9%(5.9%;13%)	ZNF597	2.9e-07	maternally	yes
ENSG00000185513	87%(81%;91%)	L3MBTL1	5.1e-06	paternally	yes
ENSG00000177432	99%(98%; 100%)	NAPIL5	6.6e-06	paternally	yes
ENSG00000242265	97%(92%;99%)	PEG10	6.9e-06	paternally	yes
ENSG00000257151	100%(99%; 100%)	PWAR6	5.8e-05	paternally	no
ENSG00000224078	95%(86%;98%)	SNHG14	8.2e-05	paternally	no
ENSG00000130844	78%(71%;84%)	ZNF331	1.2e-04	paternally	no
ENSG00000261069	100%(100%; 100%)	RP11-701H24.4	3.8e-04	paternally	no
ENSG00000225806	96%(87%;99%)	RP1-309F20.3	3.8e-04	paternally	no
ENSG00000122390	36%(31%;40%)	NAA60	4.3e-04	paternally	yes
ENSG00000128739	100.0%(100%; 100%)	SNRPN	3.6e-03	paternally	yes
ENSG00000100138	56%(54%;58%)	SNU13	4.4e-03	paternally	no
ENSG00000145945	100%(100%; 100%)	FAM50B	2.5e-02	paternally	yes
ENSG00000101898	93%(87%;97%)	MCTS2P	3.7e-02	paternally	yes
ENSG00000279192 ¹	100% (97%; 100%)	PWAR5	8.6e-02	paternally	no
ENSG00000174851 ²	56%(53%;59%)	YIF1A	9.9e-02	paternally	no
ENSG00000182109 ¹	74%(64%;82%)	RP11-69E11.4	1.3e-01	paternally	no
ENSG00000171847 ¹	23%(13%;36%)	FAM90A1	1.5e-01	maternally	no
ENSG00000082781 ¹	75%(63%;84%)	ITGB5	2e-01	paternally	no
ENSG00000178057 ¹	1.1%(0.3%; 3.4%)	NDUFAF3	2e-01	maternally	no
ENSG00000254319 ¹	88%(68%;97%)	RP11-134O21.1	2e-01	paternally	no
ENSG00000253633 ²	69%(59%;77%)	KB-1980E6.3	2e-01	paternally	no
ENSG00000111678 ¹	26%(17%;38%)	C12orf57	2e-01	maternally	no
ENSG00000101160	47%(46%;49%)	CTSZ	2e-01	maternally	no
ENSG00000054967 ²	55%(53%;57%)	RELT	2.1e-01	paternally	no
ENSG00000126226 ²	44%(41%;48%)	PCID2	2.1e-01	maternally	no
ENSG00000135709	43%(39%;47%)	KIAA0513	2.1e-01	maternally	no
ENSG00000175643	45%(43%;48%)	RMI2	2.1e-01	maternally	no
ENSG00000262155 ¹	66%(58%;73%)	RP11-266L9.5	2.1e-01	paternally	no

¹ those that may miss q-value cutoff 0.05 due to low count

² those that may miss q-value cutoff due to smaller effect size, and the remaining ones that are likely false discoveries.