A statistical method for joint estimation of *cis*-eQTLs and parent-of-origin effects using an orthogonal framework with RNA-seq data

Shirong Deng¹, Feifei Xiao^{2*}

¹School of Mathematics and Statistics, Wuhan University, Wuhan, China

²Department of Epidemiology and Biostatistics, University of South Carolina, SC,
USA

**To whom correspondence should be addressed*

Corresponding to: Feifei Xiao, Ph.D.

Department of Epidemiology and Biostatistics, Arnold School of Public Health,
University of South Carolina

Room 449, 915 Greene St., Columbia, SC, 29201

Tel: (803) 777-8936

Email: xiaof@mailbox.sc.edu

Abstract

In the past few years extensive studies have been put on the analysis of genome function, especially on expression quantitative trait loci (eQTL) which offered promise for characterization of the functional sequencing variation and for the understanding of the basic processes of gene regulation. However, most studies of eQTL mapping have not implemented models that allow for the non-equivalence of parental alleles as so-called parent-of-origin effects (POEs); thus, the number and effects of imprinted genes remain important open questions. Imprinting is a type of POE that the expression of certain genes depends on their allelic parent-of-origin which are important contributors to phenotypic variations, such as diabetes and many cancer types. Besides, multi-collinearity is an important issue arising from modeling multiple genetic effects. To address these challenges, we proposed a statistical framework to test the main allelic effects of the candidate eQTLs along with the POE with an orthogonal model for RNA sequencing (RNA-seq) data. Using simulations, we demonstrated the desirable power and Type I error of the orthogonal model which also achieved accurate estimation of the genetic effects and over-dispersion of the RNA-seq data. These methods were applied to an existing HapMap project trio dataset to validate the reported imprinted genes and to discovery novel imprinted genes. Using the orthogonal method, we validated existing imprinting genes and discovered two novel imprinting genes with significant dominance effect.

Author Summary

In the past decades, an unprecedented wealth of knowledge has been accumulated for understanding variations in human DNA level. However, this DNA-level knowledge has not been sufficiently translated to understanding the mechanisms of human diseases. Gene expression quantitative trait locus (eQTL) mapping is one of the most promising approaches to fill this gap, which aims to explore the genetic basis of gene expression. Genomic imprinting is an important epigenetic phenomenon which is an important contributor to phenotypic variation in human complex diseases and may explain some of the “hidden” heritable variability. Many imprinting genes are known to play important roles in human complex diseases such as diabetes, breast cancer and obesity. However, traditional eQTL mapping approaches does not allow for the detection of imprinting which is usually involved in gene expression imbalance. In this study, we have for the first time demonstrated the orthogonal statistical model can be applied to eQTL mapping for RNA sequencing (RNA-seq) data. We showed by simulated and real data that the orthogonal model outperformed the usual functional model for detecting main effects in most cases, which addressed the issue of confounding between the dominance and additive effects. Application of the statistical model to the HapMap data resulted in discovery of some potential eQTLs with imprinting effects and dominance effects on expression of *RBI* and *IGF1R* genes.

In summary, we developed a comprehensive framework for modeling imprinting effect for eQTL mapping, by decomposing the effects to multiple genetic components. This study is providing new insights into statistical modeling of eQTL mapping with RNA-seq data which allows for uncorrelated parameter estimation of genetic effects, covariates and over-dispersion parameter.

Introduction

With the completion the 1000 Genomes Project [1], an unprecedented wealth of knowledge has been accumulated for understanding variations in human DNA level. However, much less of this DNA-level knowledge has been translated to understanding the mechanisms of human diseases. Gene expression quantitative trait locus (eQTL) mapping is one of the most promising approaches to fill this gap, which aims to explore the genetic basis of gene expression [2]. Among the eQTL techniques, *cis*-eQTL mapping is the most commonly used technique to map local eQTLs on the same chromosome of the gene. To evaluate gene expression levels, RNA sequencing (RNA-seq) technology has recently become a widely used high-throughput tool to assess the gene expression abundance, which has many advantages over microarray, especially for discovery of novel eQTLs.

Imprinting is a type of parent-of-origin effect that the expression of certain genes depends on their allelic parent-of-origin. The same alleles transmitted from the mother may have different expression levels on transcripts with those transmitted from the father. The influence on the phenotype between the two types of heterozygotes are different, as so-called parent-of-origin effect (POE). It is now known that there are at least 80 imprinted genes in humans, many of which are involved in embryonic and placental growth and development [3]. Studies suggested that POE is an important contributor to phenotypic variation in human complex diseases and may explain some of the “hidden” heritability. An earlier study showed that for type II diabetes, a variant of SNP rs2334499 in chromosome region 11p15 was protective when maternally transmitted, whereas confers risk when paternally transmitted [4]. Evidence also demonstrated the important roles of POEs in type I diabetes, breast cancer and other carcinomas [4, 5]. However, most studies of eQTL mapping have not implemented models that allow for the non-

equivalence of parental alleles; hence the number and effects of imprinted genes remain important open questions. In the past few years, there were a few approaches that modeled POEs while searching for eQTLs with RNA-seq data. For example, Zhabotynsky et al. proposed to jointly model genetic effect and POE with a RNA-seq statistical method which focused on modeling the allele specific expression [6].

In quantitative genetics, the partition of the variance in statistical components due to additivity, dominance does not reflect the biological (or functional) effect of the genes but it is most useful for prediction, selection, and evolution [7]. In 2013, we proposed a unified orthogonal framework for modeling genetic variants displaying imprinting effects [8]. It allowed for imprinting effect detection whereas maintained the power to detect the main allelic effect including the additive and dominance effects. From theory of quantitative genetics, statistical additive genetic effects are obtained from average allele substitution effects, whereas dominance effects reflect the deviation of the genotypic values of the heterozygotes and the expected midpoint of the two homozygotes. In recent years, the inclusion of dominance in genomic models have been proposed by several researchers. However, multi-collinearity is an important issue arising from modeling multiple genetic effects. To achieve straightforward model selection and variance component analysis, uncorrelated estimation of the additive and dominance effect was necessary. Therefore, in current study, we propose to implement an orthogonal model to jointly evaluate the effect from both additive and dominance effect along with POE in eQTL mapping.

Genetic imprinting affects complex diseases through regulating the gene expression and can reveal an important component of heritable variation that remains “hidden” in traditional complex trait studies. In this study, we hypothesized that POEs contribute to regulate gene

expression along with the main allelic effect from the gene. We proposed a statistical framework to test the main allelic effects (i.e., additive and dominance effects) of the candidate eQTLs along with the POE with an orthogonal model. Intensive simulations were conducted to evaluate the methods. We also applied the methods to an existing HapMap project trio dataset to validate the reported imprinting genes and identifying eQTLs for these genes.

Results

Simulations

The statistical power of the Stat-POE and Func-POE methods was illustrated when the POE was (a) $\tau = \log(1.1)$ and (b) $\tau = \log(1.2)$. The results are shown in Figures 1-2, respectively. In both scenarios, the additive and dominance effects were fixed at $\log(1.2)$. We simulated a small fold change in both allelic effect terms to demonstrate the desirable performance of the Stat-POE method when the effect size were relatively small to detect. Consequently, even at a sample size of 50 with moderate over-dispersion ($\phi = 0.2$), the Stat-POE method presented around 70.8% power to detect a genetic effect at 1.2-fold change in additive effect, corresponding to an effect size of $\log(1.2) = 0.18$ (Figure 1). To detect POE at the fold change of 1.2, the Stat-POE and Func-POE methods both reached a statistical power of 83% with a reasonable sample size of 100 (Figure 2). Even with a very small effect size from POE at fold change of 1.1, corresponding to an effect size of $\log(1.1) = 0.10$, the methods yielded 61% power when the sample size was 200, and 91% when the sample size was 500 (Figure 1). As expected, the Stat-POE method yielded same power in detecting POE but more desirable power in detecting main genetic effects compared to the Func-POE model (Figures 1-2). In conclusion, the Stat-POE method outperformed the Func-POE method in most simulation scenarios and our methods all achieved sufficient power for detection of POEs with a relatively sample size ($N=100$).

Figure 1. Statistical power to detect additive, dominance and POE effect when (a)-(c) overdispersion $\phi = 0.2$ and (d)-(e) $\phi = 0.2$ for various samples sizes, using Stat-POE model (stat) or Func-POE model (func). The covariate coefficient $\beta = 0.1$, the sample size (n) was set at 50, 100, 200 and 500. Addi: additive effect; domi: dominant effect; impr: imprinting effect; stat: statistical model; func: functional model. The additive effect $\alpha = \log(1.2)$, dominant effect $\delta = \log(1.2)$, imprinting effect $\iota = \log(1.1)$.

Figure 2. Statistical power to detect additive, dominance and POE effect when (a)-(c) overdispersion $\phi = 0.2$ and (d)-(e) $\phi = 0.5$ for various samples sizes, using Stat-POE model (stat) or Func-POE model (func). The covariate coefficient $\beta = 0.1$, the sample size (n) was set at 50, 100, 200 and 500. Addi: additive effect; domi: dominant effect; impr: imprinting effect; stat: statistical model; func: functional model. The additive effect $\alpha = \log(1.2)$, dominant effect $\delta = \log(1.2)$, imprinting effect $\iota = \log(1.2)$.

With the simulated data, we also evaluated the estimation bias for all the parameters (β, β_G, ϕ) from the Stat-POE model. Table 1 shows that the estimation of all genetic effects achieved higher accuracy when sample size increased. Interestingly, the estimation of the covariate β and over-dispersion parameter ϕ was not notably affected by the sample sizes. Also, the estimation of genetic effects was not obviously affected by the value of the over-dispersion parameter. These results revealed the accurate and stable estimation of the covariates and over-dispersion parameters by using the Stat-POE model. Moreover, large sample sizes and small over-dispersion ensured better overall performance of the proposed methods.

Table 1. Simulation results with different sample sizes. The estimation relative bias was defined as the difference between estimated value and true parameter value divided by the true parameter value. Relative bias and mean square of errors (MSE) have been multiplied by 100. for model parameters including additive effect (α), imprinting effect (ι), covariate (β), and dispersion parameter (ϕ).

Over-dispersion	True POE effect size	N=200						N=500					
		Bias			MSE			Bias			MSE		
		α	δ	ι	α	δ	ι	α	δ	ι	α	δ	ι
0.2	log(1.1)	-0.98	1.96	-0.47	0.21	0.44	0.20	1.15	-0.83	2.35	0.09	0.18	0.07
0.2	log(1.2)	-1.80	1.67	-1.12	0.21	0.39	0.22	-0.18	-0.03	-0.41	0.09	0.18	0.08
0.5	log(1.1)	-3.48	-1.57	-6.65	0.49	1.08	0.58	-2.47	0.46	-1.84	0.20	0.40	0.19
0.5	log(1.2)	-0.71	2.13	0.66	0.52	1.04	0.52	-0.34	0.85	0.07	0.22	0.42	0.17

To evaluate the Type I error, we also simulated scenarios where there no genetic effect or POE. Although there were slightly inflated false positives in detecting genetic effect and POEs when over-dispersion was large, the type I error rates were around the nominal level 0.05 for both methods in most scenarios especially when moderate over-dispersion existed ($\phi = 0.2$ (Table 2). The false positive rate for detecting the genetic effects was comparable between the Stat-POE and Func-POE models in most scenarios.

Table 2. Type I error of the methods in detecting genetic effects. Stat: Stat-POE model; Func: Func-POE model. α : additive effect; δ : dominant effect; ι : imprinting effect, and ϕ : dispersion parameter

N	Model	$\phi = 0.2$			$\phi = 0.5$		
		α	δ	ι	α	δ	ι
50	Stat	0.078	0.046	0.062	0.074	0.078	0.100
	Func	0.070	0.048	0.062	0.076	0.074	0.100
100	Stat	0.088	0.082	0.056	0.090	0.062	0.060
	Func	0.084	0.084	0.056	0.086	0.068	0.060
200	Stat	0.066	0.062	0.062	0.062	0.054	0.068
	Func	0.066	0.060	0.062	0.064	0.052	0.068
500	Stat	0.050	0.064	0.070	0.062	0.058	0.062
	Func	0.044	0.064	0.070	0.050	0.052	0.062

Real Data Application to HapMap Parent-Child Trio Data

Using 30 children of the family trios from the HapMap project, we applied the proposed methods

to estimate additive and imprinting effects for 22 genes with previous evidence of imprinting.

These selected genes were identified to be imprinted genes using 296 phased trios from the 1000 Genomes Project and the Genome of the Netherlands participants [9].

With the proposed Stat-POE method, we identified 33 significant *cis*-eQTLs (adjusted p-values in additive effects < 0.05) for seven genes including *LPAR6*, *RBI*, *PXDC1*, *IGF1R*, *AC069277.2*, *IGF2BP3*, *SNRPN* genes (S2 Table). Among them, most candidate *cis*-eQTLs presented maternal expression pattern in regulating the gene expression. Besides, we identified six genes with imprinting effects, including *LPAR6*, *PER3*, *RBI*, *PXDC*, *IGF1R*, and *IGF2BP3* with adjusted p-values < 0.05 (Table 3). Among the significantly imprinted genes, the gene expression of *LPAR6* and *IGF1R* had significant regulation from the candidate *cis*-eQTLs rs11633209, rs728075 and rs7329291 in additive effect (adjusted p-values = 1.96×10^{-66} , 3.57×10^{-64} and 3.57×10^{-64}). Interestingly, we also discovered two novel genes which presented significant dominance effect in gene expression (Table 4), including *RBI* gene from multiple candidate *cis*-eQTLs (adjusted p-value = 3.02×10^{-80}) and *IGF1R* gene from rs4965238 (adjusted p-value = 1.37×10^{-67}). Among the identified genes presenting dominance effect in eQTL mapping, the *RBI* gene locating on chromosome 13 is a tumor suppressor gene, the mutation inactivation of which has been found to be the cause of human cancer [10]. It was also found to be an imprinted gene earlier in 2009 [11]. Interesting, *IGF1R* is the only gene which presented both dominance effect and imprinting effect from the candidate *cis*-eQTL rs4965238 (Table 4).

In conclusion, our real data application validated several existing imprinted genes, mapped candidate eQTLs for these imprinted genes. More interestingly, we discovered a few genes presenting significant dominance effect which might be involved in tumorigenesis.

Table 3. A list of genes with potential imprinting effects (p-values < 0.05). ATL: alternative allele; a: additive effect; d: dominance effect; l: imprinting effect; expression: imprinting status. The p-values were adjusted by the BH multiple comparison method.

ID	Gene	Chr	SNP	Position	Allele	ATL	P value.a	P value.d	P value.l	Expression
ENSG00000049246	<i>PER3</i>	chr1	rs67537516	7864693	G	A	0.96	1	0	Maternal
ENSG00000139687	<i>RB1</i>	chr13	rs61949059	48923973	T	G	0.18	1	6.26E-82	Maternal
ENSG00000139687	<i>RB1</i>	chr13	rs150098624	49001371	T	C	0.9	1	3.06E-80	Paternal
ENSG00000139687	<i>RB1</i>	chr13	rs4258502	48989564	G	A	0.9	1	3.06E-80	Paternal
ENSG00000139687	<i>RB1</i>	chr13	rs9568143	48996494	T	A	0.9	1	3.06E-80	Paternal
ENSG00000140443	<i>IGF1R</i>	chr15	rs4965238	99456357	T	C	0.78	1.37E-67	1.12E-71	Paternal
ENSG00000140443	<i>IGF1R</i>	chr15	rs11633209	99221205	A	T	1.96E-66	1	5.01E-71	Maternal
ENSG00000168994	<i>PXDC1</i>	chr6	rs113644426	3738592	C	T	0.98	1	4.14E-69	Paternal
ENSG00000139679	<i>LPAR6</i>	chr13	rs4942796	49011882	T	C	0	1	5.81E-64	Maternal
ENSG00000139679	<i>LPAR6</i>	chr13	rs728075	48999087	G	A	3.57E-64	1	5.81E-64	Maternal
ENSG00000139679	<i>LPAR6</i>	chr13	rs7329291	48999148	G	A	3.57E-64	1	5.81E-64	Maternal
ENSG00000139679	<i>LPAR6</i>	chr13	rs4942795	49003191	A	G	0	1	7.86E-64	Maternal
ENSG00000139679	<i>LPAR6</i>	chr13	rs7998359	48983977	G	A	0	1	7.86E-64	Maternal
ENSG00000140443	<i>IGF1R</i>	chr15	rs62023801	99295429	T	C	0.99	1	0.001	Maternal
ENSG00000140443	<i>IGF1R</i>	chr15	rs6598236	99299001	G	A	1	1	0.001	Maternal
ENSG00000136231	<i>IGF2BP3</i>	chr7	rs56259943	23400797	C	T	0.15	1	0.0137	Maternal
ENSG00000136231	<i>IGF2BP3</i>	chr7	rs73073066	23447351	C	T	0.22	1	0.0207	Maternal

Table 4. A list of genes with potential dominance effects (p-values < 0.05). ATL: alternative allele; a: additive effect; d: dominance effect; l: imprinting effect; expression: imprinting status. The p-values were adjusted by BH multiple comparison.

ID	Gene	Chr	SNP	Position	Allele	ATL	P value.a	P value.d	P value.l	Expression
ENSG00000139687	<i>RB1</i>	chr13	rs9596041	48898297	C	T	0.9	0	1	
ENSG00000139687	<i>RB1</i>	chr13	rs2122276	48893306	T	C	0.9	0	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs1529370	48903892	G	T	0.9	0	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs1427167	48885561	A	G	0.9	0	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs1449574	48887861	A	T	0.9	0	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs7490216	48897733	A	T	0.9	0	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs4141954	48895766	A	G	0.9	0	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs7985609	48945285	T	G	0	0	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs9568118	48887412	T	C	0.99	3.02E-80	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs1449576	48887803	C	T	0.99	3.02E-80	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs1449577	48887205	G	T	0.99	3.02E-80	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs6561483	48885946	T	C	0.99	3.02E-80	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs9596038	48889438	T	C	0.99	3.02E-80	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs6561482	48884687	A	G	0.99	3.02E-80	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs1834735	48885404	C	T	0.99	3.02E-80	1	-
ENSG00000139687	<i>RB1</i>	chr13	rs9596040	48889980	T	C	0.99	3.02E-80	1	-
ENSG00000140443	<i>IGF1R</i>	chr15	rs4965238	99456357	T	C	0.78	1.37E-67	1.12E-71	Paternal

Discussion

This article stands on recent advances in genetic modeling for carrying out new methodological developments to the aid of the analysis of genetic imprinting. The orthogonal statistical model here developed improves previous methods by providing a solution to allowing for additive-by-dominance genetic effects for *cis*-eQTL mapping with RNA-seq data. We demonstrated the desirable power and preserved Type I error of the statistical POE model with un-biased estimation of the genetic effects and over-dispersion of the RNA-seq data. The application to the HapMap project validated previously reported imprinting genes and discovered significant *cis*-eQTLs for these imprinted genes. More interestingly, we also identified two novel imprinting

genes with significant dominance effect, although the interpretation of which is still unclear.

This is the first time the performance of the natural and orthogonal models in detecting POE with RNA-seq data are evaluated. This project is an extension of the natural and orthogonal interaction model (NOIA) proposed by Albarez-Castro et al., in 2007 [12], based on which we have extensively worked on the framework to the estimation of statistical epistasis, gene-environmental interactions and imprinting effect in genotype-phenotype mapping for quantitative traits and qualitative traits [8, 13, 14]. In previous findings, we proved that the one-locus Stat-POE model for quantitative trait was orthogonal in certain conditions including when HWE was satisfied [8]. The conclusion should be straightforwardly extended to negative binomial or Poisson regressions. Although it still needs more effort to evaluate the orthogonality of the Stat-POE model when HWE is violated for the current framework of eQTL mapping in RNA-seq, the proposed work is providing new insights into eQTL mapping which allows uncorrelated parameter estimation of genetic effects (i.e., additive effect, dominance effect and imprinting effect) and accurate estimation of covariates and over-dispersion parameters.

In this study, we compared the Stat-POE model and Func-POE model for analyzing POE effects for RNA-seq data. These two models can be transformed to each other, but they had different meanings for their parameters and their test statistics varied in formulas. The Func-POE model focus on the biological properties and describes allele substitution effects thus presented a natural and interpretable results. The Stat-POE model utilizes the population properties estimated from the sample therefore presented greater power in detecting additive effect compared to the Func-POE model, which was consistent with our previous methodology studies in genotype-phenotype mapping for quantitative traits [8]. However, using the estimates of genotype frequencies, parameters from application of the Stat-POE model describes the variance

components, rather than allele substitutions effects as in Func-POE model, so may be seen as having a less clear interpretation. The results generated from the functional model is not tied to population rendering easier interpretation however the model selection is not straightforward. In applications, we will suggest utilizing both Stat-POE and Func-POE models to further characterize the underlying genetic architecture of the eQTLs.

We also investigated the parameter estimation and hypothesis testing of the two models with NB regression and Poisson regression assumption of the read counts for different application scope (results not shown). NB regression will be suggested when the over-dispersion is relatively large and Poisson regression is suggested to be used for small over-dispersion. In conclusion, we provided a comprehensive and robust framework for joint estimation of *cis*-eQTLs and parent-of-origin effects.

The current study has several limitations. First, family data such as trios are needed to obtain the genotypes of heterozygotes in the offspring. Second, borrowing information from the whole samples will allow for better power for modeling the RNA-seq data. A third direction is to incorporate the allele specific gene expression(ASE) as conducted in Zhabotynsky's work so we can extend our study to model both ASE and POE for candidate *cis*-eQTLs [6]. The gradually decreased cost in RNA-seq technology and future studies in methodology are warranted to achieve more powerful estimation of decomposed variance from different genetic components.

Materials and Methods

Methods: the Stat-POE and Func-POE methods

Our methods are proposed upon a basic model of eQTL mapping of a single gene with RNA-seq data which are read counts. Therefore, we consider a single gene and study the association of its expression with the j -th candidate expression quantitative trait loci (eQTL). Let y_i be the total read

counts mapped to this gene in the i -th sample, where $i = 1, \dots, n$ and n is the sample size. We model y_i by a negative binomial distribution as they are sparse counts data, which is a generalization of Poisson distribution to allow for over-dispersion. Let $f_{NB}(y_i; \mu_i, \phi)$ be the probability mass function for a negative binomial distribution with mean μ_i and dispersion parameter ϕ :

$$f_{NB}(y_i; \mu_i, \phi) = \frac{\Gamma(y_i + 1/\phi)}{y_i! \Gamma(1/\phi)} \left(\frac{1}{1 + \phi \mu_i} \right)^{1/\phi} \left(\frac{\phi \mu_i}{1 + \phi \mu_i} \right)^{y_i}, y_i = 0, 1, 2, \dots$$

where $\Gamma(\cdot)$ is the gamma function. It's easy to find that the variance $\text{Var}(y_i) = \mu_i + \phi \mu_i^2$, in which $\phi \mu_i^2$ is the over-dispersion part. When the over-dispersion parameter ϕ goes to 0, $f_{NB}(y_i; \mu_i, \phi)$ reduces to $f_p(y_i; \mu_i) = \mu_i^{y_i} e^{-\mu_i} / y_i!$, which is the probability mass function for Poisson distribution with parameter μ_i . Let $\mathbf{X}_i$ be a set of p covariates and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ be the regression coefficients, and $\boldsymbol{\beta}_G = (R, a, d, l)'$ be the genetic effects from genotypes (G) of the eQTL on Y , where R is the baseline, a , d and l are the additive, dominance and imprinting effects from G , respectively. The covariate effect of $G = G_i$ and covariates $\mathbf{X} = \mathbf{x}_i$ on the gene expression, can be formulated through the following log-linear regression model

$$\log(\mu_i) = \boldsymbol{\beta}' \mathbf{x}_i + \omega(G_i, \boldsymbol{\beta}_G),$$

where $\omega(G_i, \boldsymbol{\beta}_G)$ is the function which reflects the genetic effects.

For a bi-allelic locus, let the major and minor alleles of the j -th candidate eQTL as A_1 and A_2 , respectively. The genotype G takes four possible values $\vec{A}_1 A_1, \vec{A}_1 A_2, \vec{A}_2 A_1$ and $\vec{A}_2 A_2$, the first allele of which with arrow denotes the paternal allele and the second denotes the one originated from maternal side. We use p_{11}, p_{12}, p_{21} and p_{22} to denote genotype frequencies in the population, and use M to denote the number of variant allele A_2 , which takes values of 0, 1, 1 and 2 for the four genotypes separately. $\bar{M} = 1 + p_{22} - p_{11}$ and $V = (p_{11} + p_{22}) - (p_{11} - p_{22})^2$ are the mean and variance of M . For estimation of the genetic effects, there are different methods we can express the

genetic effect function $\omega(G_i, \beta_G)$. Xiao (2013) proposed two related methods for identifying genetic variants influences quantitative traits with different characteristics, the statistical and functional POE estimation methods [8]. Accordingly, we can express $\omega(G_i, \beta_G)$ as a population-referenced formulation using an orthogonal model, such expression generated from orthogonality reparameterization:

$$\omega(G_i, \beta_G) = \begin{cases} R - \bar{M}a - 2p_{22}(p_{12} + p_{21})d/V, & \text{if } G_i = \vec{A}_1A_1 \\ R + (1 - \bar{M})a + 4p_{11}p_{22}d/V - 2p_{21}l/(p_{12} + p_{21}), & \text{if } G_i = \vec{A}_1A_2 \\ R + (1 - \bar{M})a + 4p_{11}p_{22}d/V + 2p_{12}l/(p_{12} + p_{21}), & \text{if } G_i = \vec{A}_2A_1 \\ R + (2 - \bar{M})a - 2p_{11}(p_{12} + p_{21})d/V, & \text{if } G_i = \vec{A}_2A_2 \end{cases} \quad (2.1)$$

With the orthogonality property, this model allows for un-correlated estimation of the genetic effects including additive, dominance and imprinting effect. Therefore it is referred to the statistical POE (Stat-POE) model. For a functional model with no orthogonalization property with a POE component, the genetic effect function $\omega(G_i, \beta_G)$ can be expressed as

$$\omega(G_i, \beta_G) = \begin{cases} R & \text{if } G_i = \vec{A}_1A_1 \\ R + a + d - l, & \text{if } G_i = \vec{A}_1A_2 \\ R + a + d + l, & \text{if } G_i = \vec{A}_2A_1 \\ R + 2a, & \text{if } G_i = \vec{A}_2A_2 \end{cases} \quad (2.2)$$

from which we obtain

$$\begin{aligned} a &= \frac{1}{2} \log(\mu_{i\vec{A}_2A_2} / \mu_{i\vec{A}_1A_1}), \\ d &= \frac{1}{2} \log \left(\frac{\mu_{i\vec{A}_1A_2} + \mu_{i\vec{A}_2A_1}}{\mu_{i\vec{A}_2A_2} + \mu_{i\vec{A}_1A_1}} \right), \\ l &= \frac{1}{2} \log(\mu_{i\vec{A}_2A_1} / \mu_{i\vec{A}_1A_2}), \end{aligned} \quad (2.3)$$

where $\mu_{iA_1A_1}, \mu_{iA_1A_2}, \mu_{iA_2A_1}$ and $\mu_{iA_2A_2}$ are the underlying means of the read counts for subjects with the four genotypes, respectively. Indeed, the additive effect a measures the fold change of gene expression between the two homozygotes; the dominance effect d measures the deviation of the heterozygotes from its additive expectation; and the imprinting effect l reflects the different

effect from the two types of heterozygotes. The model in Equations (2.3) is therefore defined as a functional POE (Func-POE) model. This genetic effect model can also be called as natural model since it uses natural effects of allele substitutions as parameters, mainly focusing on the biological properties [12].

Theoretically, the parameters from the Stat-POE and Func-POE models can be transformed to each other in linear regression setting, yet which still present different properties for various application scopes. As shown in the theoretical derivation from Xiao (2013) [8], the statistical model will present better power in detecting additive effect given the existence of this effect whereas the functional model generates more interpretable results.

Parameter Estimation and Hypothesis testing

To estimate the genetic effects and POE, we can write the likelihood based on the data (y_i, X_i, G_i) ($i = 1, 2, \dots, N$) as follows,

$$L(\boldsymbol{\beta}, \boldsymbol{\beta}_G, \phi; \{y_i, \mathbf{x}_i\}_{i=1}^N) = \prod_{i=1}^N f_{NB}(y_i; \mu_i(\boldsymbol{\beta}, \boldsymbol{\beta}_G), \phi)^{I_{NB}(y_1, \dots, y_N)} f_P(y_i; \mu_i(\boldsymbol{\beta}, \boldsymbol{\beta}_G))^{1 - I_{NB}(y_1, \dots, y_N)}, \quad (2.4)$$

where $I_{NB}(y_1, \dots, y_N)$ is an indicator function which equals to 1 if a negative binomial distribution is used and 0 if a Poisson distribution is used.

With equation 2.4, the maximum likelihood estimator (MLE) of the model parameters $(\boldsymbol{\beta}_\phi, \boldsymbol{\beta}_G)$ with $\boldsymbol{\beta}_\phi \triangleq (\boldsymbol{\beta}', \phi)'$ can be estimated by the following iterative procedure.

- (1) **Initialization:** We first fit a null model by a Poisson regression using the covariate $\mathbf{X}_i$, and then estimate $\boldsymbol{\beta}$, which is obtained by the Newton Rapsion iteration method based on the first and second derivatives for the log-likelihood function of the null model, whose formulas are given in Appendix A.1. Then a score test is conducted for the over-dispersion parameter, the whole hypothesis testing process of which is shown in Appendix A.2. If the p-value of the score test is smaller than a cutoff value, e.g., $\alpha = 0.05$, we will fit a

negative binomial regression model and estimate β_ϕ . Under the negative binomial model, the estimate of β_ϕ is obtained by an iterative procedure. The details of the iterative formulas for β and ϕ are given in (A.10) and (A.11) in Appendix A.3, which is based on the iteratively re-weighted least squares method [15] and the Newton-Raphson iterate method, respectively.

- (2) **Iteration:** (a). Given β or β_ϕ , estimate β_G by the Newton-Raphson method, the details of which are shown in **Appendix B**; (b). Given β_G , estimate β by a Poisson regression with offsets $\omega(G_i\beta_G)$, or estimate β_ϕ by a negative binomial regression with offsets $\omega(G_i\beta_G)$. The estimation for β under the Poisson regression is the same as that in the initialization step with the first and second derivatives given in Appendix B.6 for the statistical model and Appendix B.8 for the functional model, respectively. Under the negative binomial regression, the estimation method for β_ϕ described in the the initialization step is also used here with the detailed forlumas given in Appendix B.5 for the statistical model and in Appendix B.7 for the functional model respectively.

- (3) **Termination:** Iterate steps (1) and (2) until estimates of all the parameters converge.

In order to assess whether each regression covariate in the model is significant on the read counts of the gene or not, statistical hypothesis testing will be performed. We constructed three testing methods including the likelihood ratio test (LRT), score test and wald test as follows. For example, for the test of additive effect with hypotheses

$$H_0:a = 0 \text{ v.s. } H_1:a \neq 0. \quad (2.5)$$

Denote $\theta = (\beta_\phi', \beta_G')'$ for the Negative Binomial (NB) regression, or $\theta = (\beta', \beta_G')'$ for the Poisson regression, the unrestricted MLE and restricted MLE under (Appendix D.1) obtained by the algorithm given in above section are denoted as $\hat{\theta}$ and $\hat{\tilde{\theta}}$ respectively. Without loss of generality,

we put the parameter a in the first position of θ , denote the other parameters as ξ , i.e. $\theta = (a, \xi)'$.

Then the score function for θ is $U(\theta) = \begin{bmatrix} \frac{\partial l(\theta)}{\partial a} \\ \frac{\partial l(\theta)}{\partial \xi} \end{bmatrix}$, and the observed fisher information matrix is $I(\theta)$

$$) = - \begin{bmatrix} \frac{\partial^2 l(\theta)}{\partial a^2} & \frac{\partial^2 l(\theta)}{\partial a \partial \xi} \\ \frac{\partial^2 l(\theta)}{\partial \xi \partial a} & \frac{\partial^2 l(\theta)}{\partial \xi^2} \end{bmatrix} \triangleq \begin{bmatrix} I_{aa}(\theta) & I_{a\xi}(\theta) \\ I_{\xi a}(\theta) & I_{\xi\xi}(\theta) \end{bmatrix}, \text{ where } l(\theta) \text{ is the log-likelihood function given as in}$$

Appendix B.1 for NB regression and statistical model, in Appendix B.2 for Poisson regression and statistical model, in Appendix B.3 for NB regression and functional model, in Appendix B.4 for Poisson regression and functional model. The formulas of $U(\theta)$ and $I(\theta)$ are given in **Appendix**

C. The LRT statistic is

$$T_L = 2[l(\hat{\theta}) - l(\tilde{\theta})]. \quad (2.6)$$

According to the theory from Rao 2005 [16], in our statistical setting, the score test statistic is therefore defined as

$$T_S = \left(\frac{\partial l(\theta)}{\partial a} \right)^2 \big|_{\hat{\theta}} J_{aa}(\theta) \big|_{\hat{\theta}}. \quad (2.7)$$

where $J_{aa}(\theta) = (I_{aa}(\theta) - I_{a\xi}(\theta) I_{\xi\xi}^{-1}(\theta) I_{\xi a}(\theta))^{-1}$.

Moreover, the Wald test statistic is defined by:

$$T_W = \frac{\hat{a}^2}{J_{aa}(\theta) \big|_{\hat{\theta}}}. \quad (2.8)$$

Under H_0 , the statistics T_L , T_S , and T_W all converge in distribution to χ_1^2 . For a given significance level α , reject H_0 when the observed value of the statistics are greater than $\chi_{1,\alpha}^2$. The process of the hypothesis testing for the other parameters can be implemented similarly.

Simulations

To evaluate the performance of the proposed statistical methods in eQTL mapping with RNA-seq data, we carried out extensive simulation studies in realistic settings. First, we compared

power of the Stat-POE and Func-POE methods in detecting the main allelic effects (including the additive and dominance effects) and POE. We simulated y_i , the total number of read counts of a gene in the i -th sample by a negative binomial distribution with $\mu_{iG} = \exp(0.1x_i + \omega(G_i\beta_G))$ and over-dispersion parameter $\varphi = 0.2$, where $\mathbf{X}$ was a continuous covariate $\mathbf{X} \sim N(0,1)$. To evaluate the performance of the methods in estimating both genetic effects and over-dispersion, we generated data with different sample sizes $N = 50, 100, 200$ and 500 , respectively. In all simulation settings, Hardy-Weinberg Equilibrium (HWE) was assumed that the genotype frequencies in the sample were set at $[p_{11}, p_{12}, p_{21}, p_{22}] = [0.36, 0.24, 0.24, 0.16]$. The over-dispersion parameter was set at empirical values which was 0.2 or 0.5 . The additive effect α and dominance effect δ were both fixed at $\log(1.2)$ where the values of 1.2 reflected the fold change of the logarithm mean shift of the genotypic values, referring to Equation (2.3). The POE parameter τ was set at $\log(1.1)$ or $\log(1.2)$ respectively. Each simulation was replicated for 500 times to evaluate the performance of the Stat-POE and Func-POE methods. Relative bias and mean square of errors (MSE) were calculated for each parameter in the different scenarios to evaluate the estimation accuracy. The estimation relative bias was defined as the difference between estimated value and the true parameter value and then divided by the true parameter value. We also used simulated data to quantify the power and Type I error rates of the methods. To illustrate the performance of the proposed methods in detecting different genetic effect terms and POE, the statistical power for the hypothesis testing was calculated using a range of different critical values under different scenarios. Type I error was calculated under the null model where there was no genetic effect or POE.

Application to a HapMap RNA-seq dataset

Datasets

We used an RNA-seq dataset from 30 HapMap Caucasian samples obtained from the NCBI Bioproject (PRJNA385599). The samples were collected from lymphoblastoid cell lines from 15 males and 15 females. For most of these samples, the RNA reads were 150 bp paired-end reads, with an additional run with 75 bp paired-end reads. The median of the total number of reads for these 30 samples was approximately 20 million. All of these reads were mapped to hg38 human reference genome using Tophat2 [6].

Since all of these samples were children of family trios, the parents of which were also part of the samples included in the 1000 Genomes Project[17]. Genotype data of the 30 trios were used to obtain the phased genotype of the children. For these 30 trios, the HapMap project genotyped about 3.9 million SNPs. The phasing and imputation of these 30 trios were conducted by Zhabotynsky et al.'s study, where the phased and imputed genotypes in our study were directly obtained from [6]. Briefly, SHAPEIT2 [18] was used for phasing and IMPUTE2 [19] was used for imputation against the 1000 Genome reference panel containing 2,504 individuals and ~82 million SNPs. Based on the phased and imputed SNPs, we extracted 6,211,048 of high confidence imputed SNPs (the ones with at least one heterozygote in the sample) in total.

Identification of imprinted genes and genes with dominance effect

We selected 22 known imprinted genes based on the list reported by Jadhav et al. [9]. These genes were selected because they had abundant expression in the 30 samples (S1 Table). For each potential imprinted gene, all SNPs in the gene coding region were selected as candidate *cis*-eQTLs. For each *cis*-eQTL and gene expression pair, the Stat-POE method was applied to detect candidate eQTLs with additive, dominance and POE effects. Four covariates were adjusted in the model including the total read counts per individual and the first three principal components computed from the matrix of normalized expression. In all of the hypothesis testing, Benjamini-

Hochberg (BH) method was used for multiple comparison to adjust the p-values obtained from the LRT tests [20]. We tested the POE of the previously reported imprinted genes to evaluate the performance of our methods. For novel discovery of genetic effects of these imprinted genes, we tested the additive and dominance effects simultaneously.

Acknowledgments

This work was supported by the internal support from University of South Carolina for Dr. Feifei Xiao. We are extremely grateful to Dr. Wei Sun and Dr. Zhabotynsky's generous support who kindly provided the phased and imputed genotypes and pre-processed RNA-seq data.

Reference

1. Genomes Project, C., et al., *A global reference for human genetic variation*. Nature, 2015. **526**(7571): p. 68-74.
2. Cookson, W., et al., *Mapping complex disease traits with global gene expression*. Nat Rev Genet, 2009. **10**(3): p. 184-94.
3. Perry, J.R.B., et al., *Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche*. Nature, 2014. **514**(7520): p. 92-+.
4. Kong, A., et al., *Parental origin of sequence variants associated with complex diseases*. Nature, 2009. **462**(7275): p. 868-U59.
5. Wallace, C., et al., *The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes*. Nature Genetics, 2010. **42**(1): p. 68-U85.
6. Zhabotynsky, V., et al., *A statistical method for joint estimation of cis-eQTLs and parent-of-origin effects under family trio design*. Biometrics, 2019.
7. Huang, W. and T.F. Mackay, *The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis*. PLoS Genet, 2016. **12**(11): p. e1006421.
8. Xiao, F., J. Ma, and C.I. Amos, *A unified framework integrating parent-of-origin effects for association study*. PLoS One, 2013. **8**(8): p. e72208.
9. Bharati Jadhav, V.O.P.M., Kristina K. Galalova, Daniel Ho, Harmen H.M. Draisma, Mark A. van de Wiel, Lude Franke, View ORCID ProfileBastiaan T. Heijmans, Joyce van Meurs, Rick Jansen, GoNL Consortium, BIOS Consortium, Peter A.C. 't Hoen, Andrew J. Sharp, Szymon M. Kielbasa, *RNA-Seq in 296 phased trios provides a high resolution map of genomic imprinting*. bioRxiv, 2018.
10. Chinnam, M. and D.W. Goodrich, *Rb1, Development, and Cancer*. Cancer and Development, 2011. **94**: p. 129-169.
11. Kanber, D., et al., *The Human Retinoblastoma Gene Is Imprinted*. Plos Genetics, 2009. **5**(12).
12. Alvarez-Castro, J.M. and O. Carlborg, *A unified model for functional and statistical*

- epistasis and its application in quantitative trait Loci analysis*. Genetics, 2007. **176**(2): p. 1151-67.
13. Ma, J., et al., *Natural and orthogonal interaction framework for modeling gene-environment interactions with application to lung cancer*. Hum Hered, 2012. **73**(4): p. 185-94.
 14. Xiao, F., et al., *Natural and orthogonal model for estimating gene-gene interactions applied to cutaneous melanoma*. Hum Genet, 2014. **133**(5): p. 559-74.
 15. Green, P.J., *Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives*. . Journal of the Royal Statistical Society. Series B (Methodological). **46**(2): p. 149-192.
 16. CR., R., *Score Test: Historical Review and Recent Developments*. 2005.
 17. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes*. Nature, 2012. **491**(7422): p. 56-65.
 18. Delaneau, O., et al., *Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel*. Nat Commun, 2014. **5**: p. 3934.
 19. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing*. Nature Genetics, 2012. **44**(8): p. 955-+.
 20. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society Series B-Statistical Methodology, 1995. **57**(1): p. 289-300.

Supporting information

S1 Table. List of imprinted genes identified in previous studies. 22 genes were included to study the *cis*-eQTL and imprinting effects. These genes were included in Jadhav et al., 2018 which were compared to the identified imprinted genes reported as imprinted by the Genotype-Tissue Expression (GTEx) project (Baran et al. 2015). The genes were annotated by NCBI hg19 build.

S1 File.

S2 Table. A list of *cis*-eQTL (p-values < 0.05) for the 22 imprinting candidate genes. ATL: alternative allele; a: additive effect; d: dominance effect; l: imprinting effect; expression: imprinting status. The p-values were adjusted by the BH multiple comparison method.