G3 Genes | Genomes | Genetics

A Generalized Linear Model for Decomposing *Cis*-regulatory, Parent-of-Origin, and Maternal Effects on Allele-Specific Gene Expression

Yasuaki Takada,* Ryutaro Miyagi,† Aya Takahashi,†,‡ Toshinori Endo,* and Naoki Osada*,1
*Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060-0814, Japan, and †Department of Biological Sciences and ‡Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Hachioji, 192-0397, Japan
ORCID IDs: 0000-0002-8391-7417 (A.T.); 0000-0002-5711-2527 (T.E.); 0000-0003-0180-5372 (N.O.)

ABSTRACT Joint quantification of genetic and epigenetic effects on gene expression is important for understanding the establishment of complex gene regulation systems in living organisms. In particular, genomic imprinting and maternal effects play important roles in the developmental process of mammals and flowering plants. However, the influence of these effects on gene expression are difficult to quantify because they act simultaneously with *cis*-regulatory mutations. Here we propose a simple method to decompose *cis*-regulatory (*i.e.*, allelic genotype), genomic imprinting [*i.e.*, parent-of-origin (PO)], and maternal [*i.e.*, maternal genotype (MG)] effects on allele-specific gene expression using RNA-seq data obtained from reciprocal crosses. We evaluated the efficiency of method using a simulated dataset and applied the method to whole-body *Drosophila* and mouse trophoblast stem cell (TSC) and liver RNA-seq data. Consistent with previous studies, we found little evidence of PO and MG effects in adult *Drosophila* samples. In contrast, we identified dozens and hundreds of mouse genes with significant PO and MG effects, respectively. Interestingly, a similar number of genes with significant PO effect were detect in mouse TSCs and livers, whereas more genes with significant MG effect were observed in livers. Further application of this method will clarify how these three effects influence gene expression levels in different tissues and developmental stages, and provide novel insight into the evolution of gene expression regulation.

KEYWORDS
gene expression
cis-regulatory mutation
genomic imprinting
maternal effect

Epigenetics, which refers to phenotypic modifications in the absence of changes to information encoded in DNA molecules, has become a central topic in biological research in order to understand the development of multicellular organisms and maintenance of highly differentiated cells and tissues (Waddington 1942). Although epigenetic effects can contribute to a wide array of phenotypes, most studies of epigenetic effects in the era of molecular biology have concerned gene expression, which is much more easily quantified than other phenotypes on a genome-wide scale. Epigenetic effects on gene expression can be classified as either *cis*- or *trans*-epigenetic effects (Bonasio *et al.* 2010). In diploid organisms, *cis*-epigenetics refers to chromosome-specific modification of gene expression. For example, histone protein modification and cytosine methylation could affect the expression of genes located on the same chromosome. In contrast, *trans*-epigenetics refers to epigenetic modifications of gene expression that have equal effects on both chromosomes of diploid organisms. In a broad sense, *trans*-epigenetics would therefore include all gene expression changes caused by intrinsic and extrinsic environmental changes, such as those observed in cell differentiation and reaction to environmental change.

Genomic imprinting is a well-known phenomenon in mammals and flowering plants and refers to the process by which genes inherited from a particular sex are downregulated or completely silenced (Köhler *et al.* 2012; Barlow and Bartolomei 2014). By the above definition, genomic imprinting is caused by *cis*-epigenetic mechanisms. Among mammals, genomic imprinting has been most extensively studied in laboratory mice (*Mus musculus*), and ∼150 loci, including both protein-coding

doi: https://doi.org/10.1534/g3.117.042895
Manuscript received March 22, 2017; accepted for publication May 6, 2017; published Early Online May 17, 2017.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.117.042895/-/DC1.
[1] Corresponding author: Graduate School of Information Science and Technology, Hokkaido University, Kita14 Nishi9, Kita ku, Sapporo, Hokkaido 060-0814, Japan. E-mail: nosada@ist.hokudai.ac.jp

genes and noncoding RNAs, have been experimentally identified as imprinted (Blake *et al.* 2010). In contrast, it remains unclear whether genomic imprinting can be detected in nonmammalian animals. In particular, there have been conflicting results whether fruit flies (*Drosophila melanogaster*), which lack DNA methyltransferases [except for the *Dnmt2* (*MT2*) product], are subject to genome-wide imprinting effect (Menon and Meller 2010; Coolon *et al.* 2012; McEachern *et al.* 2014; Takayama *et al.* 2014). Although the underlying mechanisms and causes of imprinting are not entirely clear, genomic imprinting is necessary to our understanding of the complex relationships between genotypes and phenotypes (Ferguson-Smith 2011). Therefore, the effects of genomic imprinting in different organisms should be determined using standardized methods.

Recent advances in sequencing technology have enabled the evaluation of a genome-wide imprinting pattern. RNA-seq transcriptome sequencing has allowed the measurement of chromosome-specific (or allele-specific) gene expression levels for paternally and maternally inherited genes that harbor genetic markers, such as single nucleotide variations (SNVs) (Wittkopp 2005). The comparison of patterns in allele-specific gene expression between reciprocal crosses is informative because of potential differences in gene expression levels as a consequence of *cis*-regulatory mutations (Wittkopp 2005); *i.e.*, observation of parent-of-origin (PO)–dependent allelic imbalance in both reciprocally crossed individuals suggests genomic imprinting rather than a *cis*-regulatory effect. Accordingly, such comparisons are widely used to discern *cis*-genetic and *cis*-epigenetic effects; *i.e.*, if allelic imbalance depending on PO is observed in both reciprocally crossed individuals, the imbalance is likely due to genomic imprinting rather than the *cis*-regulatory effect. Several studies have implemented these strategies to identify genes subject to genomic imprinting on a genome-wide scale (Babak *et al.* 2008; Gregg *et al.* 2010; Coolon *et al.* 2012; Calabrese *et al.* 2015). However, this method tends to be conservative if the *cis*-regulatory effect is prevalent, because it may reduce the power of statistical tests to detect imprinting effects.

In addition, comparisons of reciprocal crosses are complicated by additional confounding factors because reciprocally crossed individuals have different maternal environments. This finding was first described as the maternal effect in a classical experiment by Walton and Hammond (1938). Although classical family studies and embryo transplantation studies have shown that the environmental effect on offspring phenotype is generally larger than the genetic effect (Gluckman and Hanson 2004), genetic effects may contribute to the maternal effect to some extent. One example would be the genotype effect in oocyte cytoplasm, as these cells inherit mRNA and mitochondrial DNA from the mother in a process that meets the definition of a *trans*-regulatory effect, which is a genetic effect equally affecting both chromosomes in a diffusible way (Emerson and Li 2010). In addition, prenatal and postnatal environments determined by maternal genotype (MG) will contribute to offspring phenotypes. Here we use the term MG effect, assuming appropriate control of nongenetic environmental factors. Although the MG effect may be subtle, it might contribute to gene expression pattern in a *trans* manner. The PO and MG effects are hardly distinguished in a conventional genetic analysis, because the phenotypes of offspring are defined by the sum of the maternally and paternally inherited alleles, and the maternal and paternal contributions are indistinguishable at the phenotype level (Hager *et al.* 2008). However, by directly measuring gene expression level of maternally and paternally inherited alleles, there is an opportunity to separately evaluate the PO and MG effects. A previous study proposed a method to jointly estimate the genetic *cis*-regulatory, or allelic genotype (AG) and PO effects, but the MG effect was not examined (Zou *et al.* 2014).

Here, we have proposed a simple statistical framework for simultaneously and separately estimating the AG, PO, and MG effects on gene expression in reciprocally crossed individuals when the allele specific gene expression level is provided, and have demonstrated the effectiveness of this method using a simulated dataset. We used a generalized linear model (GLM) to quantify each effect, assuming a lack of interaction. The previous genome-wide study of the PO effect, which was designed without replication, suggested the importance of biological replicates (Coolon *et al.* 2012). GLMs efficiently deal with the contribution of each factor and fluctuations among biological replicates. We applied this method to two different organisms, *Drosophila* and mice. For the former, we obtained a new adult female whole-body gene expression dataset using two pairs of reciprocal crosses: F1 hybrids of the *Drosophila* Genetic Reference Panel (DGRP) strains for which genomic sequences were made publicly available (Mackay *et al.* 2012). For mice, we reanalyzed recently published datasets of trophoblast stem cells (TSCs) and livers from reciprocal crosses between CAST/EiJ and C57BL/6NJ (Cast/B6) animals (Goncalves *et al.* 2012; Calabrese *et al.* 2015). Although we identified statistically significant AG effect for a considerable number of genes in both organisms, we identified a very small number of genes with significant PO and MG effects in *Drosophila*, consistent with an earlier report by Coolon *et al.* (2012). In contrast, we found that dozens of genes in mouse TSCs and livers were subject to significant PO effect. In addition, considerably higher number of genes in mouse livers exhibited a significant MG effect compared to genes in mouse TSCs, indicating that the MG effect tends to be specific to tissues or developmental stages.

MATERIALS AND METHODS

GLM design

Suppose that there are two different isogenic strains, A and B. Following the general rule, A × B would denote F1 hybrids generated by a cross between females of strain A and males of strain B. When strains A and B exhibit sufficient genetic differences, we could measure allele-specific gene expression levels using RNA-seq for each reciprocal cross, A × B and B × A, with biological replications. The allele-specific expression value E would then be defined using the following linear regression model expression:

$$E = \mu + \mathrm{AG} + \mathrm{PO} + \mathrm{MG} + \varepsilon, \qquad (1)$$

where μ and ε represent average expression level and biological/statistical noise, respectively. Here, we assumed each effect was a fixed effect and assigned binary codes to the effects. For AG, we assigned values of 0 and 1 to A and B, respectively. For PO, we assigned a value of 0 if the chromosome was inherited from the mother, and 1 if the chromosome was inherited from the father. For MG, we assigned a value of 0 to sample A × B (MG A) and 1 to sample B × A (MG B). The error term was estimated using biological replicates of samples. A schematic representation of this design is shown in Figure 1.

We propose two GLM models to utilize allele-specific gene expression level to estimate the AG, PO, and MG effects. The first model is a log-normal GLM. In a typical RNA-seq data analytical pipeline, gene expression levels are normalized by gene length and total read count and represented as FPKM values, which can be assumed to exhibit a log-normal distribution (Bengtsson *et al.* 2005). Therefore, by log-transforming allele-specific FPKM values, we could apply a Gaussian distribution to the distribution of response variable in the GLM. The second model is a negative binomial GLM. Since an actual RNA-seq dataset is count data represented by the number of reads mapped on

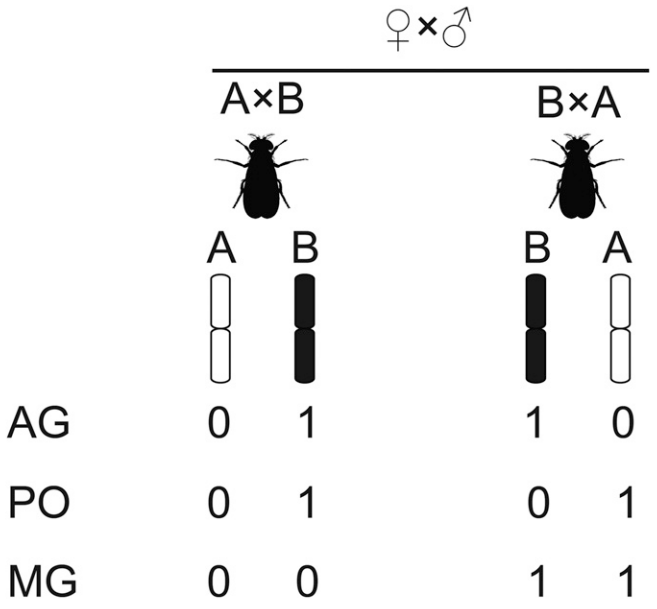

Figure 1 Schematic representation of the generalized linear model (GLM) design. A hypothetical reciprocal cross between fly strains A and B is assumed. Black and white chromosomes represent the A and B genotype, respectively, and binary code specifies the allelic genotype (AG) effect (A: 0, B: 1). The parent-of-origin (PO) effect is set to 0 when the chromosome is inherited from the mother (left side of diploid chromosomes) and 1 when the chromosome is inherited from the father (right side of diploid chromosomes). The maternal genotype (MG) effect is specified by the maternal genotype (A: 0, B: 1).

the transcript sequences, a negative binomial model has been widely adopted in many statistical packages for analyzing RNA-seq data, such as EdgeR (McCarthy *et al.* 2012) and DESeq (Anders and Huber 2010). The log-normal and negative binomial GLM analyses were performed using the glm function and EdgeR libraries in the R statistical package (R core team 2016). The R script and expression data files used for the GLMs are provided as Supplemental Material, File S1.

Computer simulations

In computer simulations, we only considered the log-normal GLM. We assumed normally distributed allele-specific gene expression levels with the fixed additive effects of AG, PO, and MG. Following the design shown in Figure 1, we considered eight different cases for the presence and absence of fixed effects: no effect, AG, PO, MG, AG + PO, AG + MG, PO + MG, and AG + PO + MG. As the statistical detection power for each fixed effect was determined by the magnitude of the fixed effect size relative to biological, environmental, and/or statistical fluctuations, we evaluated the power using the ratio of the fixed effect to the SD of experimental noise, which was equivalent to Cohen's d statistic. A larger d indicated more power for effect detection.

For each simulated gene, we arbitrarily assigned a basal gene expression level and added random noise drawn from a standard normal distribution $N(0, 1)$. After adding errors, a fixed effect was added to the expression value. For example, when $d = 5$, we added 5 to the expression value when the binary code of samples (Figure 1) was 1 for each fixed effect. For each condition, 1250 genes were simulated (in total, 10,000 genes for one replicate) with two or five replications, and the simulated dataset was analyzed using the log-normal GLM method. The significance of each gene test was evaluated using the criterion of false discovery rate (FDR) = 0.05 (Benjamini and Hochberg 1995).

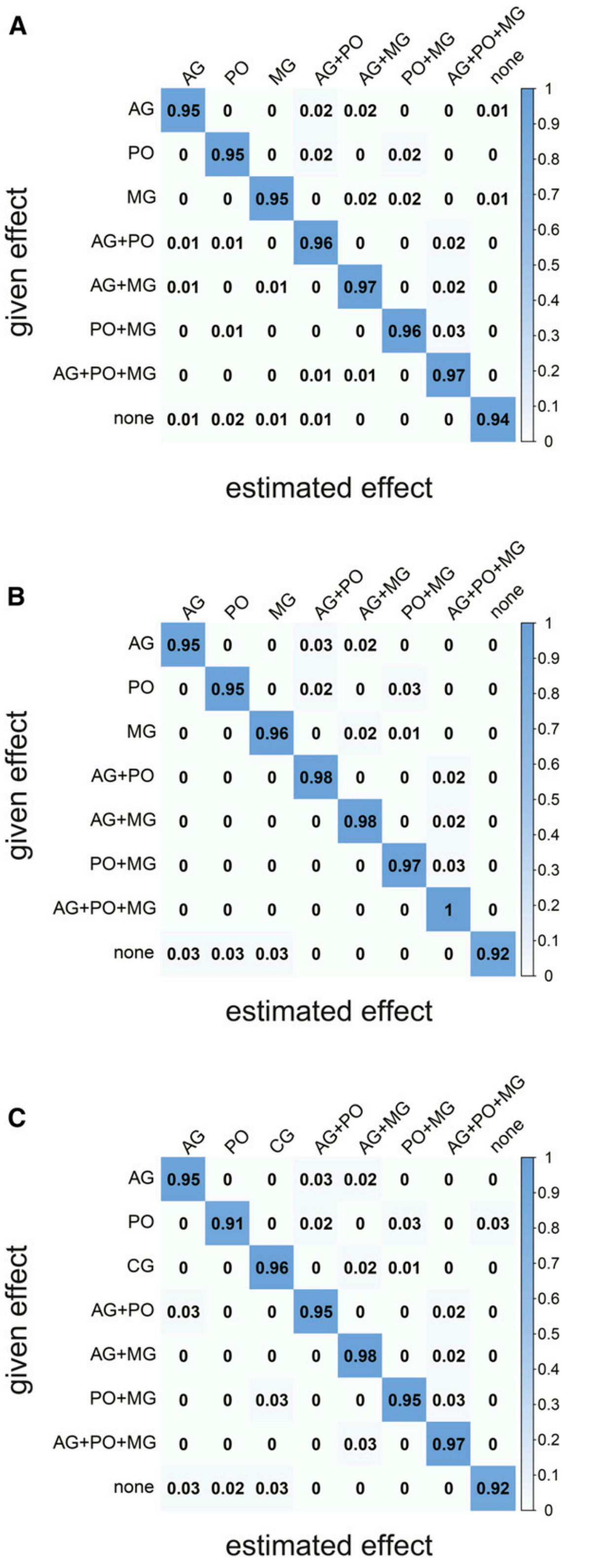

Figure 2 Evaluation of method using a simulated dataset. For each panel, the rows represent effects given in the simulations and the columns represent effects estimated using the generalized linear model (GLM). Numbers in cells denote the fractions of correctly estimated effect among 1250 simulated genes. (A) d = 5 with two replicates; (B) d = 3 with five replicates; and (C) d = 5 for allelic genotype (AG) and maternal genotype (MG), and d = 2 for parent-of-origin (PO) effects.

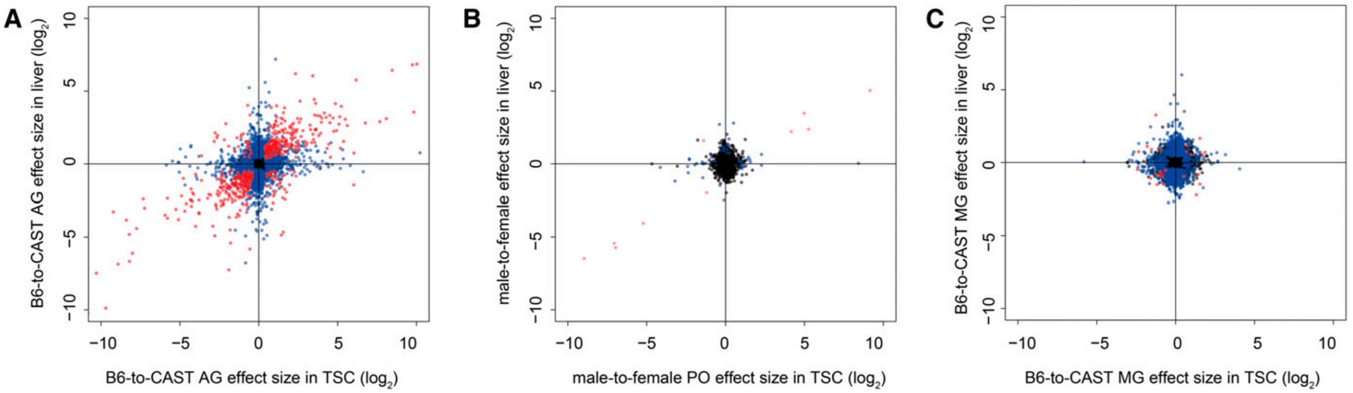

Figure 3 Comparison of effect sizes of mouse trophoblast stem cells (TSCs) and livers in the negative binomial generalized linear model (GLM). The estimated effect size of each gene is indicated by a colored circle. The effect sizes of the TSCs and livers are shown on the *x*- and *y*-axes, respectively. Red circles represent genes with significant effects in both tissues and blue circles represent genes with significant effects in either tissue. Genes indicated by black circles did not exert significant effects. The allelic genotype (AG), parent-of-origin (PO), and maternal genotype (MG) effects are shown in (A–C), respectively.

RNA-seq dataset

In this study, we obtained new gene expression data of two pairs of reciprocal crosses of *D. melanogaster* from the DGRP (Mackay *et al.* 2012; Massouras *et al.* 2012), representing crosses between RAL324 and RAL852 and between RAL799 and RAL820. These strains were arbitrarily chosen from a list of DGRP strains. The flies were grown at 25° with a 12-hr light/dark cycle and were fed standard cornmeal fly medium. F1 virgin females were collected within 8 hr of eclosion and maintained separately on the regular food media. After 4–7 d of isolation, 20 flies per sample were flash frozen in liquid nitrogen and stored at −80°. The whole-body total RNA was extracted using the TRIzol Plus RNA Purification Kit (Thermo Fisher Scientific, Waltham, MA). The concentration of extracted total RNA was measured using a Nanodrop 2000c (Thermo Fisher Scientific) and quality was evaluated using a TapeStation (Agilent Technologies, Foster City, CA). For RNA-seq, 250 ng total RNA was used for library construction with the TruSeq Stranded mRNA Library Prep Kit (Illumina, San Diego, CA). Samples were barcode-indexed and pooled for each sequencing lane. Raw read data were deposited into the DDBJ SRA database under the Bioproject ID PRJDB5381. The accession number and index type of each library are provided in Table S1. Mouse TSC expression data were retrieved from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE63968, and mouse liver data were downloaded from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) using the accession number E-MTAB-1091.

Estimation of allele-specific expression data

We obtained genomic sequences of focal strains to estimate allele-specific gene expression levels. For *Drosophila*, we used the version dm3 reference genome sequence, and obtained a VCF file (freeze 2.0 call) containing the information about the SNVs in DGRP strains from the DGRP website (http://dgrp2.gnets.ncsu.edu/). Genome sequences of RAL324, RAL799, RAL820, and RAL852 were reconstructed using the FastaAlternateReferenceMaker command in GATK software (McKenna *et al.* 2010). A mouse reference genome sequence (GRCm38) and VCF files of CAST/EiJ and C57BL/6NJ strains were retrieved from the ENSEMBL database (http://ensembl.org/) and the Sanger Mouse Genomes Project website (http://www.sanger.ac.uk/science/data/mouse-genomes-project), respectively. The genome sequences of CAST/EiJ and C57BL/6NJ were reconstructed using the same procedure described for *Drosophila* data.

We used ASE-TIGER software, which is based on Bayesian inference, to estimate the allele-specific FPKM and number of allele-specific mapped reads (Nariai *et al.* 2016). Briefly, RNA-seq reads were mapped on transcriptome sequences reconstructed from two parental genomes. Strain-specific *Drosophila* and mice transcriptome sequences were generated from the reconstructed genome sequences using the annotation file for the build 5 *D. melanogaster* genome (downloaded from NCBI: https://www.ncbi.nlm.nih.gov/) and Mus_musculus.GRCm38.84.gtf for mice (downloaded from ENSEMBL), respectively. We used bowtie2 software to map RNA-seq reads, using the option of "–very sensitive" (Langmead and Salzberg 2012). Because we could not accurately estimate the allele-specific expression levels of genes with small numbers of SNVs within genes, we filtered out transcripts with less than three SNVs in the exons. Because ASE-TIGER reported FPKM and the number of mapped reads for each transcript, those values were summed across isoforms to estimate the value at the gene level. For the log-normal model, weakly expressed genes (average FPKM $<$ 0.1) were filtered out. Before log-transformation, we replaced FPKM values $<$0.01 with 0.01 to avoid legalism associated with very small or 0 values. For the negative binomial GLM, genes with less than one count per million mapped reads in less than half of the chromosomes were filtered out.

Gene ontology enrichment analysis

We utilized the DAVID 6.7 webserver to identify significantly enriched gene ontology terms from a list of genes with significant effects (Jiao *et al.* 2012). Lists of background genes were extracted from all analyzed genes in each dataset. For each gene ontology term, terms with $p < 0.05$, determined using a modified Fisher's exact test after correcting for multiple testing, were selected as significantly overrepresented functional categories (Hosack *et al.* 2003).

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

RESULTS

Design of the GLM

We conducted a GLM analysis in order to jointly estimate the effects of AG, PO, and MG. Two different GLMs, the log-normal and negative binomial GLM, were applied. A full description of the GLMs is presented

Table 1 Summary of the log-normal generalized linear model (GLM) analysis of *Drosophila* and mice

	Drosophila melanogaster (Female Whole Body)		*Mus musculus* (Cast/B6)	
Samples	RAL799/RAL820	RAL324/RAL852	TSC	Liver
No. of analyzed genes	6176	6971	12,963	11,995
AG (FDR = 0.05)	776	1570	1456	1584
PO (FDR = 0.05)	0	0	22	16
MG (FDR = 0.05)	0	0	4	304

TSC, trophoblast stem cells; AG, allelic genotype effect; FDR, false discovery rate; PO, parent-of-origin effect; MG, maternal genotype effect.

in the *Materials and Methods* section. Briefly, in the log-normal GLM, we estimated the allele-specific gene expression level as FPKM for each gene and transformed these values to a $\log_2$ scale. The $\log_2$-transformed expression values were used as response variables in the GLM, assuming a Gaussian distribution. In the negative binomial GLM, the estimated number of reads mapped on the transcriptome sequences from each chromosome were used as count data. Three fixed effects (AG, PO, and MG) were set as the explanatory variables in the model and binary codes were assigned to the values. A schematic representation of the model is shown in Figure 1.

Computer simulations

Before analyzing real data, we performed computer simulations to confirm whether the GLM could successfully decompose three different effects (AG, PO, and MG). We only considered the log-normal model for the simulations because both log-normal and negative binomial models assume additive effects of the three factors and their underlying assumptions are essentially the same. We evaluated a range of Cohen's d ($1 \leq d \leq 5$), a ratio of the fixed effect to the SD of statistical noise, for datasets replicated two and five times. In the GLM with Gaussian distribution, p values monotonically decrease with $|d|$ and we expected that statistical power would increase with higher d values and more replicates.

Our simulation using a duplicated dataset showed that we could accurately estimate each effect at $d = 5$ (Figure 2A), where the true positive rate of the effect was ~0.95 with an FDR of 0.05. As expected, the statistical power of the test increased remarkably with more replicates (Figure S1and Figure S2). We attained very high statistical power (true positive rate ~0.95) with five replicates when $d = 3$ (Figure 2B). We also tested whether any unbalanced effects could result in a biased estimation of each effect. Figure 2C shows results from five replicates wherein the d values were 5 for AG and MG and 2 for PO. Despite the somewhat biased effect, we could accurately detect each significant effect.

Analysis of Drosophila whole bodies

We first analyzed two adult female *D. melanogaster* datasets, using a duplicated experimental design. In the log-normal GLM, after the initial filtering (see *Materials and Methods*), 6716 genes in the RAL799/RAL820 cross and 6971 genes in the RAL852/RAL324 cross were analyzed. We identified 776 and 1570 genes exhibiting signatures of the AG effect (FDR = 0.05) in the RAL799/RAL820 and RAL852/RAL324 crosses, respectively (Table 1). In the negative binomial GLM, 6536 genes in the RAL799/RAL820 cross and 6797 genes in the RAL852/RAL324 cross were analyzed. We identified 922 and 1732 genes exhibiting signatures of the AG effect (FDR = 0.05) in the RAL799/RAL820 and RAL852/RAL324 crosses, respectively (Table 2). Although none of the genes showed significant PO and MG effects with the log-normal GLM, 4 to 11 genes showed significant PO and MG effects with the negative binomial GLM.

Both methods agreed that 10–25% of genes in the DGRP strains have a significant *cis*-regulatory effect. Among them, 221 and 400 genes showed significant AG effect in both reciprocal crosses in the log-normal and negative binomial GLM, respectively. On the other hand, none of the genes with significant PO and MG effects overlapped between the reciprocal crosses. Detailed results are provided in File S2, File S3, File S4, and File S5.

Analysis of mouse TSCs

The second dataset was obtained in TSCs from mouse reciprocal cross Cast/B6 as reported by Calabrese *et al.* (2012), and composed of three biological replicates. However, because one of the replicates had been obtained in a previous study, we only used the dataset with duplicates in our analysis. Using the log-normal GLM, we identified 1493, 273, and four genes with significant AG, PO, and MG effects, respectively (FDR $<$ 0.05), among 13,343 genes in this dataset.

Although the sexes of analyzed TSC samples are unknown, we expect that genes on the X chromosome should show significant PO effect when the samples are males because a male inherits the X chromosome only from the mother. Indeed, most genes with significant PO effect (251 out of 273) were located on the mouse X chromosome, which implies that the samples included male TSCs. When we examined the pattern of gene expression on the Y chromosome and the expression level of the *Xist* gene on the X chromosome, one of the TSC samples (GSM1561520) showed similar gene expression pattern to the male liver samples, further demonstrating that the TSC sample was from a male (data not shown). Therefore, we excluded the genes on the X chromosome from further analysis. After the filtering, the number of genes with significant AG, PO, and MG effects became 1456, 22, and four, respectively, in the log-normal GLM (Table 1).

Similar to the results of *Drosophila*, we observed slightly more genes with significant AG and PO effects using the negative binomial GLM; after filtering out X chromosomal genes, in total, 2102 genes showed significant AG effect and 64 genes showed significant PO effect (Table 2). However, the negative binomial GLM identified 393 genes with significant MG effect, considerably higher than those identified by the log-normal GLM. Detailed results are provided in File S6 and File S7.

Analysis of mouse livers

The third dataset comprised mouse liver expression data with six replicates, as performed by Goncalves *et al.* (2012) using the same Cast/B6 reciprocal cross combination. Using the log-normal GLM, we identified 1608, 249, and 312 genes with significant AG, PO, and MG effects, respectively, among the 12,293 genes in the liver dataset. Because the samples were derived from male livers, most of the PO genes were on the X chromosome. After filtering out the genes on the X chromosome, the numbers of genes with significant AG, PO, and MG effects were 1584, 16, and 304, respectively (File S4 and Table 1). Likewise, among 11,169 autosomal genes, the negative binomial GLM identified 2014, 35, and 1355 genes with significant AG, PO, and MG effects,

Table 2 Summary of the negative binomial generalized linear model (GLM) analysis of *Drosophila* and mice

	Drosophila melanogaster (Female Whole Body)		*Mus musculus* (Cast/B6)	
Samples	RAL799/RAL820	RAL324/RAL852	TSC	Liver
No. of analyzed genes	6536	6797	12,219	11,169
AG (FDR = 0.05)	922	1732	2102	2104
PO (FDR = 0.05)	5	15	64	35
MG (FDR = 0.05)	6	12	393	1355

TSC, trophoblast stem cells; AG, allelic genotype effect; FDR, false discovery rate; PO, parent-of-origin effect; MG, maternal genotype effect.

respectively (Table 2). Detailed results are provided in File S8 and File S9.

Evaluation of the log-normal and negative binomial GLMs

In both *Drosophila* and mice, the negative binomial model identified more genes with significant effects, which indicates that the negative binomial GLM has a lower rate of type II error and/or a higher rate of type I error. Although the difference is small for the AG and PO effects, the number of genes with significant MG effect considerably differs between the log-normal and negative binomial GLMs. Despite of some discrepancy, FDR-corrected p values were highly correlated between the two models and the log-normal GLM gives more conservative estimate of p values. Although the two methods have both advantages and disadvantages, we primarily show the results of negative binomial GLM in the following analyses.

Comparison between mouse TSCs and livers

Because the mouse TSC and liver data were obtained from the same reciprocal crosses, we contrasted the difference between the two tissues. In Figure 3, we present plots of the estimated effect sizes (fixed effect to the expression level in $\log_2$ scale) using the negative binomial GLM, for each gene in the TSCs and livers. We observed relatively small overlap (24%) of genes with significant AG effect between TSCs and livers, and many genes showed opposite AG effect in the two tissues (Figure 3A). In contrast, although the significance level for the PO effect was different between the TSCs and livers, probably attributable to different sample size and noise level, the sign and size of PO effect were highly consistent between the two tissues (Figure 3B). As for the MG effect, a very small number of significant genes (29 genes) were overlapped between the TSCs and livers, suggesting that the MG effect is highly tissue specific (Figure 3C).

Functional analysis of genes with significant effects

We investigated whether there was significant enrichment of gene ontology terms among the genes with significant AG, PO, and MG effects. In *Drosophila*, none of the gene ontology terms were overrepresented after controlling for FDR = 0.05. In the mouse TSCs, only the genes with significant MG effect showed the enrichment of annotated gene functions; gene ontology terms neuron differentiation (GO:0030182) and neuron development (GO:404866) were slightly overrepresented in the genes with significant MG effect. In the liver, 30 gene ontology terms were significantly enriched among genes with the AG effect (FDR = 0.05); the most highly overrepresented gene category was oxidation reduction process (GO: 0055114). Enriched terms for the AG effect in the liver were mostly related to oxygen metabolism process, protein binding activity, and membrane components (Table S2). Genes with significant PO effect in the liver did not exhibit any statistically significant enrichment. In contrast, genes with significant MG effect exhibited statistically significant enrichment gene annotation for 21 gene ontology terms, mostly related to ribosomal and mitochondrial components (Table S3).

DISCUSSION

Here, we proposed a novel approach to decompose the three confounding effects affecting gene expression levels in reciprocally crossed F1 hybrids. Although we applied the two different GLMs, log-normal and negative binomial GLMs, we first focused on the log-normal GLM and performed computer simulations because the relationship between the effect size and error distribution in the log-normal model is much more intuitively understandable. Our simulation study showed the efficiency of this method when in the presence of sufficiently strong effects relative to statistical noises. In our duplicated *Drosophila* dataset, the average SDs of $\log_2$-transformed error were 0.255 for the RAL799/RAL820 reciprocal cross and 0.185 for the RAL852/RAL324 reciprocal cross. The higher error variance observed in the RAL799/RAL820 cross was likely due to the higher number of genes with significant AG effect in that line (Table 1 and Table 2). In mouse samples, the average SDs of $\log_2$-transformed error were 0.310 and 0.656 for TSCs and livers, respectively. As described above, we only focused on the log-normal model in the simulations, but we should note that the assumptions for the distribution of biological and technical noises are different between the models, leading to the difference in statistical power.

Our analysis of two different reciprocal crosses of *Drosophila* is largely corroborated by the Coolon *et al.* (2012) study that demonstrated an absence of genomic imprinting in *Drosophila*. In addition, we did not find strong evidence of the MG effect in the adult female flies. However, the negative binomial GLM identified a small number of genes with significant PO and MG effects. There was no overlap of those candidate genes between two different reciprocal crosses, and we were not able to conclude whether those candidate genes were true or false positive genes. We also should note that studies to date have used only adult files. Therefore, further experiments based on samples from early developmental stages with more replicates are required to conclude the status of genomic imprinting and maternal effects in *Drosophila*.

In contrast to *Drosophila*, mouse datasets yielded dozens to hundreds of genes with significant PO and MG effects. Although the two datasets were conducted by different research groups, our comparison between TSCs and livers provided a good opportunity to investigate differences in each effect in the tissues and at developmental stages. Although many genes are imprinted in a tissue-specific manner [*e.g.*, DeChiara *et al.* (1991)], our results showed a generally consistent genome-wide pattern of the PO effect across tissues and developmental stages (Figure 3). In contrast, small (24%) overlap between tissues was observed among genes with significant AG effect, although a similar number of genes were identified in both tissues. These results imply that a majority of *cis*-regulatory mutations are tissue specific. This pattern corroborates the modularity of gene regulation, wherein many mutations in *cis*-regulatory regions, such as enhancers, exhibit tissue-specific effects (Wray 2007). Moreover, the number of genes with significant MG effect differed strikingly between TSCs and livers both in the log-normal and negative binomial GLMs. As we identified similar numbers of genes with significant AG and PO effects in both tissues, this difference might reflect important tissue-specific biological features. Although we cannot convincingly explain weaker MG effect in the TSC dataset, we suspect that TSCs, which are derived from embryos before implantation, spend less time in maternal–fetal crosstalk compared with other fetal and adult tissues.

We examined whether our PO candidate genes in mice agree with the 150 known imprinted genes (Blake *et al.* 2010). In TSCs, 26 out of 64 candidate genes were known as imprinted genes. *Ano1* and *Gab1* genes, which are not included in the list of 150 known imprinted genes but are actually imprinted specifically in the placenta (Okae *et al.* 2012), were identified as PO-biased genes in the negative binomial GLM. Notably, the PO bias in *Ano1* was not detected by Calabrese *et al.* (2015) who shared a part of their dataset with our study. Likewise, in livers, 10 out of 35 candidate genes were known as imprinted genes. Among the 10 known imprinted genes, paternally biased *Peg13* and maternally biased *Rian* were not identified as imprinted genes in the study using the same dataset (Goncalves *et al.* 2012). In addition to the known imprinted genes, we identified 60 candidate genes with the PO effect in TSCs and livers. Among them, *Gm11407* and *Snhg14* showed a signature of PO bias both in TSCs and livers. Although *Gm11407* is a pseudogene, *Snhg14* is a long-noncoding RNA located within an imprinted locus. Because human *SNHG14* is known to be imprinted (Babak *et al.* 2008), mice *Shng14* is also likely imprinted. The list of imprinted genes identified in this study is shown in Table S4. Some known mouse imprinted genes did not achieve statistical significance, probably because of our statistical method. For example, paternally imprinted *H19* genes did not meet our criteria for a significant PO effect. We examined expression data for this gene in TSCs and found that the imprinting status was not highly consistent among replicates. In addition, this gene was not expressed in livers. Therefore, our method requires sufficient replicates with good experimental conditions. In general, our method identified relatively fewer genes with PO effect than previous studies using RNA-seq data (Gregg *et al.* 2010; Wang *et al.* 2011; DeVeale *et al.* 2012; Goncalves *et al.* 2012).

One of our most important methodologic achievements was the ability to evaluate the maternal effect (MG effect) without nuclear or embryo transplantation. Interestingly, the functional categories of genes with significant MG effect were largely different between the TSCs and livers. In particular, genes with significant MG effect in livers contained many ribosomal and mitochondrial genes. Although it is reasonable that the mitochondrial genotype plays an important role in the maternal effect, it is unlikely that the maternal cytosolic effect is still active in adult liver tissues. Therefore, the enrichment of ribosomal components in the genes with significant MG effect in livers should be a consequence of maternal effect during development. We note, however, that the maternal effect sizes were generally much smaller than those of the other two effects; even though statistical significance was detected, effect sizes of MG hardly exceeded 2, suggesting that the maternal effect is prevalent but has relatively minor effects on gene expression pattern (Figure 3C) compared with the AG and PO effects.

Conclusions

We have reported a novel method to decompose the three confounding effects on allele-specific gene expression level in reciprocal crosses, and have demonstrated the effectiveness of this method using simulated data. Although available data are currently limited, this method yielded many biologically important observations in fruit flies and mice. This method will contribute greatly to our understanding of how genetic and epigenetic signals regulate patterns of gene expression and induce phenotypic diversity among tissues and individuals.

ACKNOWLEDGMENTS

We thank Kenta Sumiyama and Hidemi Watanabe for valuable comments. We also thank the Bloomington *Drosophila* Stock Center for providing *Drosophila* stocks. This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI) on Innovative Areas and Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (grant nos. 23113008 and 26440202).

LITERATURE CITED

Anders, S., and W. Huber, 2010 Differential expression analysis for sequence count data. Genome Biol. 11: R106.

Babak, T., B. DeVeale, C. Armour, C. Raymond, M. A. Cleary *et al.*, 2008 Global survey of genomic imprinting by transcriptome sequencing. Curr. Biol. 18: 1735–1741.

Barlow, D. P., and M. S. Bartolomei, 2014 Genomic imprinting in mammals. Cold Spring Harb. Perspect. Biol. 6: a018382.

Bengtsson, M., A. Ståhlberg, P. Rorsman, and M. Kubista, 2005 Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Res. 15: 1388–1392.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289–300.

Blake, A., K. Pickford, S. Greenaway, S. Thomas, A. Pickard *et al.*, 2010 Mousebook: an integrated portal of mouse resources. Nucleic Acids Res. 38: D593–D599.

Bonasio, R., S. Tu, and D. Reinberg, 2010 Molecular signals of epigenetic states. Science 330: 612–616.

Calabrese, J. M., W. Sun, L. Song, J. W. Mugford, L. Williams *et al.*, 2012 Site-specific silencing of regulatory elements as a mechanism of X inactivation. Cell 151: 951–963.

Calabrese, J. M., J. Starmer, M. D. Schertzer, D. Yee, and T. Magnuson 2015 A survey of imprinted gene expression in mouse trophoblast stem cells. G3 5: 751–759.

Coolon, J. D., K. R. Stevenson, C. J. McManus, B. R. Graveley, and P. J. Wittkopp, 2012 Genomic imprinting absent in *Drosophila melanogaster* adult females. Cell Rep. 2: 69–75.

DeChiara, T. M., E. J. Robertson, and A. Efstratiadis, 1991 Parental imprinting of the mouse insulin-like growth factor II gene. Cell 64: 849–859.

DeVeale, B., D. van der Kooy, and T. Babak, 2012 Critical evaluation of imprinted gene expression by RNA–seq: a new perspective. PLoS Genet. 8: e1002600.

Emerson, J. J., and W.-H. Li, 2010 The genetic basis of evolutionary change in gene expression levels. Philos. Trans. R. Soc. B 365: 2581–2590.

Ferguson-Smith, A. C., 2011 Genomic imprinting: the emergence of an epigenetic paradigm. Nat. Rev. Genet. 12: 565–575.

Gluckman, P. D., and M. A. Hanson, 2004 Maternal constraint of fetal growth and its consequences. Semin. Fetal Neonatal Med. 9: 419–425.

Goncalves, A., S. Leigh-Brown, D. Thybert, K. Stefflova, E. Turro *et al.*, 2012 Extensive compensatory *cis-trans* regulation in the evolution of mouse gene expression. Genome Res. 22: 2376–2384.

Gregg, C., J. Zhang, B. Weissbourd, S. Luo, G. P. Schroth *et al.*, 2010 High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science 329: 643–648.

Hager, R., J. M. Cheverud, and J. B. Wolf, 2008 Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. Genetics 178: 1755–1762.

Hosack, D. A., G. Dennis, Jr., B. T. Sherman, H. C. Lane, and R. A. Lempicki, 2003 Identifying biological themes within lists of genes with ease. Genome Biol. 4: R70.

Jiao, X., B. T. Sherman, W. Huang da, R. Stephens, M. W. Baseler *et al.*, 2012 DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics 28: 1805–1806.

Köhler, C., P. Wolff, and C. Spillane, 2012 Epigenetic mechanisms underlying genomic imprinting in plants. Annu. Rev. Plant Biol. 63: 331–352.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The *Drosophila melanogaster* genetic reference panel. Nature 482: 173–178.

Massouras, A., S. M. Waszak, M. Albarca-Aguilera, K. Hens, W. Holcombe *et al.*, 2012 Genomic variation and its impact on gene expression in *Drosophila melanogaster*. PLoS Genet. 8: e1003055.

McCarthy, D. J., Y. Chen, and G. K. Smyth, 2012 Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res. 40: 4288–4297.

McEachern, L. A., N. J. Bartlett, and V. K. Lloyd, 2014 Endogenously imprinted genes in *Drosophila melanogaster*. Mol. Genet. Genomics 289: 653–673.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Menon, D. U., and V. Meller, 2010 Germ line imprinting in Drosophila: epigenetics in search of function. Fly (Austin) 4: 48–52.

Nariai, N., K. Kojima, T. Mimori, Y. Kawai, and M. Nagasaki, 2016 A Bayesian approach for estimating allele-specific expression from RNA-seq data with diploid genomes. BMC Genomics 17: 2.

Okae, H., H. Hiura, Y. Nishida, R. Funayama, S. Tanaka *et al.*, 2012 Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. Hum. Mol. Genet. 21: 548–558.

R core team, 2016 R: A Language and Environment for Statistical Computing.

Takayama, S., J. Dhahbi, A. Roberts, G. Mao, S. J. Heo *et al.*, 2014 Genome methylation in *D. melanogaster* is found at specific short motifs and is independent of *DNMT2* activity. Genome Res. 24: 821–830.

Waddington, C. H., 1942 The epigenotype. Endeavour 1: 18–20.

Walton, A., and J. Hammond, 1938 The maternal effects on growth and conformation in Shire horse-Shetland pony crosses. Proc. R. Soc. Lond. B Biol. Sci. 125: 311–335.

Wang, X., P. D. Soloway, and A. G. Clark, 2011 A survey for novel imprinted genes in the mouse placenta by mRNA-seq. Genetics 189: 109–122.

Wittkopp, P. J., 2005 Genomic sources of regulatory variation in *cis* and in *trans*. Cell. Mol. Life Sci. 62: 1779–1783.

Wray, G. A., 2007 The evolutionary significance of *cis*-regulatory mutations. Nat. Rev. Genet. 8: 206–216.

Zou, F., W. Sun, J. J. Crowley, V. Zhabotynsky, P. F. Sullivan *et al.*, 2014 A novel statistical approach for jointly analyzing RNA-seq data from F1 reciprocal crosses and inbred lines. Genetics 197: 389–399.

Communicating editor: D. J. de Koning