Xie *et al. BMC Bioinformatics* (2019) 20:530
https://doi.org/10.1186/s12859-019-3141-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model

Jing Xie[1], Tieming Ji[1*], Marco A. R. Ferreira[2], Yahan Li[3], Bhaumik N. Patel[3] and Rocio M. Rivera[3]

Abstract

Background: High-throughput sequencing experiments, which can determine allele origins, have been used to assess genome-wide allele-specific expression. Despite the amount of data generated from high-throughput experiments, statistical methods are often too simplistic to understand the complexity of gene expression. Specifically, existing methods do not test allele-specific expression (ASE) of a gene as a whole and variation in ASE within a gene across exons separately and simultaneously.

Results: We propose a generalized linear mixed model to close these gaps, incorporating variations due to genes, single nucleotide polymorphisms (SNPs), and biological replicates. To improve reliability of statistical inferences, we assign priors on each effect in the model so that information is shared across genes in the entire genome. We utilize Bayesian model selection to test the hypothesis of ASE for each gene and variations across SNPs within a gene. We apply our method to four tissue types in a bovine study to de novo detect ASE genes in the bovine genome, and uncover intriguing predictions of regulatory ASEs across gene exons and across tissue types. We compared our method to competing approaches through simulation studies that mimicked the real datasets. The R package, BLMRM, that implements our proposed algorithm, is publicly available for download at https://github.com/JingXieMIZZOU/BLMRM.

Conclusions: We will show that the proposed method exhibits improved control of the false discovery rate and improved power over existing methods when SNP variation and biological variation are present. Besides, our method also maintains low computational requirements that allows for whole genome analysis.

Keywords: Allelic imbalance, Hierarchical generalized linear mixed model, High-throughput sequencing experiments, Single nucleotide polymorphism

Background

In a diploid cell, the two alleles of a gene inherited from maternal and paternal parents express roughly equally for most genes. However, research has uncovered a group of genes in the genome where two copies of a gene express substantially differently, a phenomenon known as allelic imbalance. One such example involves imprinted genes whose allele expression is based on the parent of origin [1, 2]; that is, imprinted genes are mainly or completely expressed from either the maternally or paternally inherited allele but not both, so the total expression from genomic copies is the appropriate amount for healthy and viable organisms [3]. Another prominent example is X-chromosome inactivation in mammals [4, 5], where one copy of the X chromosome is inactivated in female cells to maintain the same dosage of X-linked genes compared to male cells. The choice of which X chromosome is silenced is random initially, but once chosen, the same X chromosome remains inactive in subsequent cell divisions. In a third and rather random case, allelic imbalance occurs when there are mutations in *cis*-regulatory regions of one allele, leading to differential expression of two alleles [6, 7].

*Correspondence: jit@missouri.edu
[1] Department of Statistics, University of Missouri at Columbia, Columbia, MO 65211, USA
Full list of author information is available at the end of the article

Allelic imbalance affects approximately 5-10% of genes in the mammalian genome [5], but it is not biologically clear what series of mechanisms a cell employs to precisely initiate allele-specific expression (ASE) during fetal development and consistently maintain it through a lifetime. Several common congenital human disorders are caused by mutations or deletions within these ASE regions, such as Beckwith-Wiedemann syndrome (BWS) [8, 9], which characterizes an array of congenital overgrowth phenotypes; Angelman syndrome [10], which characterizes nervous system disorders; and Prader-Willi syndrome, in which infants suffer from hyperphagia and obesity.

To understand the molecular mechanisms underlying ASEs and human developmental defects due to misregulated ASE regions, a powerful and accurate computational algorithm to detect genome-wide ASEs is urgently needed. The binomial exact test, employed in AlleleSeq [11], is one of the most widely used methods to test ASEs due to its simplicity. [12] uses analysis of variance (ANOVA) in their proposed pipeline Allim. [13] fits a mixture of folded Skellam distributions to the absolute values of read differences between two alleles. However, these abovementioned statistical methods draw conclusions based on observations produced from one gene; due to the expensive cost of acquiring tissue samples and sequencing experiments, most laboratories can only afford three or four biological replicates. Depending on sequencing depth, genes may also have low read counts, limiting the power of the aforementioned methods.

In searching for more powerful and reliable ASE detection methods, several groups have proposed Bayesian approaches to share information across genes and thus improve gene-related inferences on average. For instance, the MBASED method [14] and the QuASAR method [15] all assume the read counts follow binomial distributions with a beta prior on the probability parameter. In their statistical models, they assume that ASE of a gene or a region is constant across SNPs. However, ASE is known to vary within a gene due to alternative splicing [16, 17], which is essentially universal in human multi-exon genes that comprise 94% of genes overall [17, 18]. Therefore, a highly desirable feature of ASE detection methods is identification of ASE genes and ASE variations within genes across multiple exons. [19] developed a flexible statistical framework that satisfied this requirement. It assumes a binomial distribution with a beta prior. Additionally, it places a two-component mixture prior on the parameters of the beta-binomial model. A Markov chain Monte Carlo (MCMC) method was adopted to compute posterior probabilities for inferences of genes and SNPs. However, due to the extensive computational power required in the MCMC calculation for one gene and the large number of genes in the entire genome, this method is not empirically appealing. Other relevant methods include the EAGLE method [20] that detects associations between environmental variables and ASEs, the WASP method [21] that addresses incorrect genotype calls, and the RASQUAL method [22] that detects gene regulatory effects.

In this paper, we propose a new statistical method that addresses the abovementioned challenges. Specifically, our proposed approach can detect ASE genes and ASE variations within genes simultaneously while maintaining a low computational requirement. Coupled with exon and RNA transcript information, our statistical predictions produce detailed, biologically relevant, intriguing results that enable researchers to examine the molecular mechanisms of ASE regulation in detail.

Particularly, we model the logistic transformation of the probability parameter in the binomial model as a linear combination of the gene effect, single nucleotide polymorphism (SNP) effect, and biological replicate effect. The random SNP effect permits ASE to vary within a gene; the random replicate effect accounts for extra dispersion among biological replicates beyond binomial variation. To overcome the low number of biological replicates and/or low number of read counts of a gene, we propose a hierarchical model with a Gaussian prior on the fixed gene effect and inverse gamma priors, respectively, on the variance components of the random SNP and replicate effects. We test hypotheses via Bayesian model selection method based on model posterior probabilities. To compute posterior probabilities, we propose combining the empirical Bayes method and Laplace approach to approximate integrations, leading to substantially reduced computational power requirements compared to MCMC. We illustrate the utility of our proposed method by applying it to the bovine genome in [23], which motivated our study; findings reveal for the first time highly detailed information regarding the testing results for whole-genome ASEs, unveiling inspiring ASE variations across exons and across tissue types. To compare our method with existing approaches, we simulate data that mimic real datasets to ensure that the comparison results can be reproduced in practice. The proposed method outperforms existing methods in false discovery rate (FDR) control of detecting ASEs and variations therein across SNPs. We call our method the Bayesian Logistic Mixed Regression Model (BLMRM) method. The R package, BLMRM, for the proposed method is publicly available for download at https://github.com/JingXieMIZZOU/BLMRM.

Results

Application for the de novo identification of ASE and imprinted genes in bovine

Most of the imprinted genes identified to date have been in the mouse [24]. Original work, identified the non-equivalency of the parental alleles by generating embryos which only had maternal chromosomes (gynogenotes and parthenogenotes) or paternal chromosomes (androgenotes) [25, 26]. By doing this, investigators identified which genes are expressed exclusively from each chromosome. Other studies used mice which had various types of genetic rearrangements including translocations, duplications and deletions and noticed that the direction in which the allele was inherited (either through the mother or the father) mattered for the successful development and wellbeing of the offspring [27]. Subsequent work turned to genetic manipulations to identify the function of imprinted genes in mice. More recent, with the advent of genome wide approaches, investigators have generated large datasets from F1 individuals generated from the breeding of two inbred (homozygous) strains of mice [28]. An advantage of using mice to do this type of work is that most strains have been sequenced and all animals within a strain will have the same maternal and paternal DNA sequence. While useful, the mouse model does not always faithfully represent other mammals [29]. In addition, most laboratory mice are inbred (homozygous) while other mammals are heterozygous which incorporates complexity to the analysis of identifying parental alleles. As imprinted gene expression is species-specific, tissue-specific, and developmental stage specific [24], investigators would have to do monetary and animal expensive studies to identify novel imprinted genes and their potential function in health and disease.

A current limitation for investigators working in the area of genomic imprinting in heterozygote animals such as bovine, is the difficulty to assess whether a gene or a region in a gene has ASE for the entire genome. For example, in the case in which 4 fetuses are obtained from the breeding of one cow and one bull, each of the fetuses may have a specific combination of alleles (penitentially 4 combinations), making the identification of imprinted gene expression a daunting task, not to mention extremely expensive. Therefore, new computational tools and analyses must be devised in order to provide investigators knowledge of allelic imbalances in the transcriptome which may then be used to do locus-specific wet bench work to determine the accuracy of the predictions.

Specifically, [23] measured gene expressions of four normal female F1 conceptuses (fetus and placenta) generated from the mating of Bos taurus (mother) and Bos taurus indicus (father). Tissues were retrieved from the brain, kidney, liver, skeletal muscle, and placenta of these four conceptuses. RNA-seq experiments were conducted on each tissue type for each replicate.

Aligning RNA-seq reads to a non-identical reference genome has been shown to introduce alignment bias [30, 31]. To address the mapping bias problem, [23] combined the reference genome (i.e., the *B. t. taurus* reference genome UMD3.1 build) and the pseudo *B. t. indicus* genome to create a custom diploid genome. Specifically, the sire's DNA was subjected to next generation sequencing (DNA-seq) to identify all SNPs between his genome and the *B. t. taurus* reference genome. Then Genome Analysis Toolkit (GATK) [32] and SAMtools [33] pipelines were applied for SNP calling and only SNPs identified by both pipelines were used to generate a pseudo *B. t. indicus* genome. At last, RNA-seq reads from the *B. t. indicus* × *B. t. taurus* F1 conceptuses were mapped to the diploid genome using both the HISAT2 [34] and BWA [35] pipelines and only variants identified by both methods were retained to minimize the potential effects of false positives. The resulting datasets are publicly available at the Gene Expression Omnibus database under accession number GSE63509.

We used the BLMRM method to separately analyze liver, kidney, muscle, and brain tissue data from [23]. Missing values are not uncommon in real datasets, especially when dealing with heterozygous species (for example, cattle and humans), as not all replicates share the same set of SNPs among parental alleles. We first filtered out genes containing only one SNP or for which all SNPs were not represented by at least two individuals. We also removed genes for which the observed maternal and paternal expression percentages were constant across all replicates and all SNPs as statistical inferences are straightforward in such a scenario. In total, 9,748 genes remained for analysis, among which many had low numbers of total RNA-seq read counts.

We then applied the proposed BLMRM method to these 9,748 genes. Hyperparameters were estimated using the method described in the "Method" section. For example, for liver tissue, we have $\widehat{\mu} = 0.43$, $\widehat{\sigma}^2$= 4.62, $\widehat{a}_s = 2.35$, $\widehat{b}_s = 1.37$, $\widehat{a}_r = 2.03$, and $\widehat{b}_r = 0.09$.

We identified several examples containing varied and informative patterns of tissue-specific and/or exon-specific ASEs. Here, we present four genes: *AOX1*, *HACL1*, *TMEM50B*, and *IGF2R*. Aldehyde oxidase 1 (*AOX1*; XLOC_003018) is a cytosolic enzyme expressed at high levels in the liver, lung, and spleen but at a much lower level in many other organs since this gene plays a key role in metabolizing drugs containing aromatic azaheterocyclic substituents [36, 37]. By controlling FDR at 0.05, the BLMRM method identified gene *AOX1* as exhibiting ASE at the gene level in the brain, kidney, and muscle, and biallelically expressed in the liver (top panel in Fig. 1). The vertical axis in Fig. 1 indicates the observed sample average percentage of gene expression from the maternal allele. The bar around each sample average denotes the 95% confidence interval at each SNP. SNPs are drawn with ascending genomic locations in a chromosome. The bottom of each panel in Fig. 1 shows the distribution of SNPs in exons from annotated RefSeq transcripts of this gene. Conclusions from our BLMRM method coincide with *AOX1* gene functional analysis. Using the binomial exact

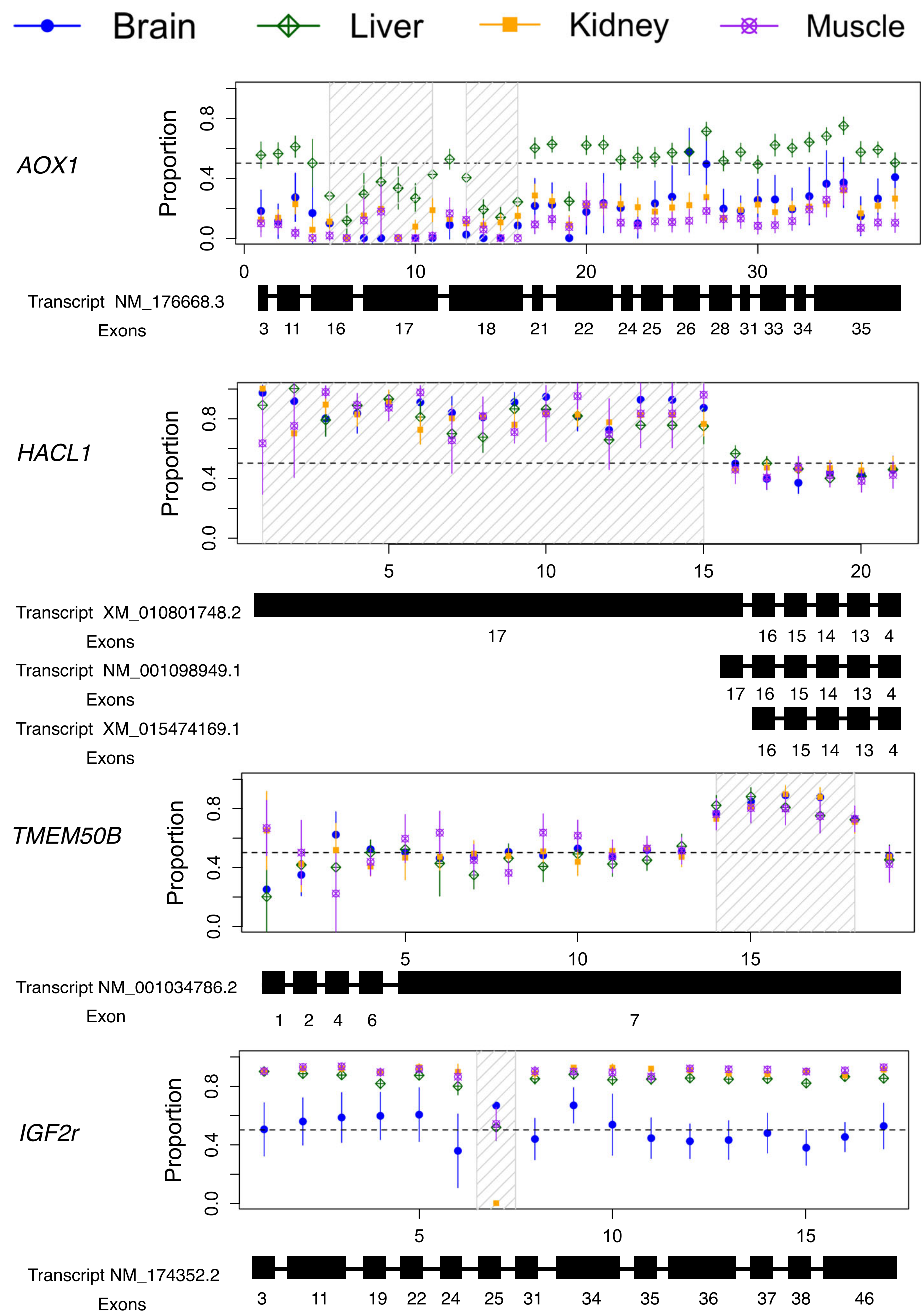

Fig. 1 Percentage of gene expression from maternal allele in brain, liver, kidney, and muscle, respectively. The top panel shows gene *AOX1*. The second panel shows gene *HACL1*. The third panel shows gene *TMEM50B*, and the bottom panel shows gene *IGF2r*. SNPs are drawn with ascending genomic locations. The bottom of each panel shows distribution of SNPs in exons from all RefSeq annotated transcripts of this gene. Rectangles represent exons (only those with SNPs are shown) with exon numbers indicated under each rectangle. Lengths of exons are not drawn to scale

test, [23] only found that *AOX1* had preferential paternal expression in bovine muscle and failed to detect ASE in the brain and kidney. Our proposed method also suggests significant ASE variations across SNPs in the liver, kidney, and muscle with FDR at the 0.05 level. Interestingly, regions in the liver showing ASE variations corresponded to the 16th, 17th, and 18th exons housing the 5-7th and 14-16th SNPs. Given this exon- and tissue-specific information, biologists can examine the ASE regulatory mechanism in detail.

2-hydroxyacyl-CoA lyase (*HACL1*; XLOC_001524) is involved in perixosomal branched fatty acids oxidation and primarily expressed in the liver [38]. Our proposed method identified *HACL1* as exhibiting significant ASE at the gene level and its variations across SNPs. Figure 1 Panel 2 visualizes our observations and shows a clear maternal preference of expression for the first 15 SNPs, whereas the remaining six suggest biallelic expression of this gene. This surprising finding spurred further investigation, upon which we identified that the first 15 SNPs belong to exon 17 of alternative splice variant XM_010801748.2 while the last SNPs are shared between two or three splice isoforms (i.e. NM_001098949.1, XM_015474169.1, and XM_010801748.2). No further information is available regarding the ASE mechanism of this gene, as this is the first time we have retrieved such detailed statistical results for each gene in an entire genome within a short computational window. Future work will identify whether this ASE gene is a novel imprinted gene and if, in fact, this gene shows variant-specific imprinted expression as has been documented for other genes [39].

Transmembrane protein 50B (*TMEM50B*; XLOC_000329) is a ubiquitously expressed housekeeping gene. Our method identified this gene to be biallelically expressed in all analyzed tissues (Fig. 1, Panel 3) as expected for a housekeeping gene. Interestingly, our proposed method also predicted significant variations across SNPs in each of these four tissue types. Upon investigating detailed activity of this gene, Fig. 1 indicates that a portion of the 3' UTR of this transcript appears to have maternal preference. The consistent pattern across tissues motivated us to understand the importance of this SNP variation. We hypothesize that this corresponds to a specific RNA variant required for maintaining cellular function.

Finally, insulin-like growth factor 2 receptor (*IGF2r*; XLOC_018398) is a well-known maternally expressed mannose receptor that targets IGF2 for degradation [40]. This gene is imprinted in the liver, kidney, and muscle (Fig. 1, Panel 4) but has biallelic expression in the brain of mice and cattle [41, 42]. In addition, *IGF2r* is lowly expressed in the cattle brain [42]. Prediction results from our proposed method coincide with the literature.

By controlling FDR at 0.05, Fig. 2 summarizes the numbers of detected ASE genes, numbers of genes with ASE variations across SNPs, and numbers of genes exhibiting ASE at the gene level and ASE variations across SNPs simultaneously, respectively, among the four tissues. We conducted some further analysis on these detected genes. For instance, in the top Venn diagram, among the 37 detected ASE genes shared by all four tissue types, 11 of them cannot be mapped to the set of annotated genes using the UMD 3.1 build. Among the rest of 26 annotated and detected ASE genes, we found that three of them had been documented as imprinted genes across all or most of these four tissue types. These three imprinted genes are (1) *GSTK1* that is maternally expressed in human placenta but unknown in other human tissues [43], paternally expressed in mouse kidney, liver, muscle, and maternally expressed in mouse brain [44], maternally expressed in bovine oocyte and unknown in other bovine tissues [45]; (2) *PLAGL1* that is paternally expressed in human kidney, muscle, and unknown in other human tissues [46], paternally expressed in mouse muscle, kidney, and brain [44], and paternally expressed in bovine brain, kidney, muscle, and liver [47]; (3) *BEGAIN*, which is unknown in human genome, preferentially expressed from the paternal allele in mouse neonatal brain [48], paternally expressed in bovine kidney and muscle with strong statistical evidence though no biological verification yet [42], and found to be paternally expressed in sheep kidney, liver, muscle, and brain (all four) tissue types [49]. Excluding these three documented imprinted genes, the other 23 annotated ASE genes detected by our BLMRM method are *de novo* detected ASE genes and their biological relevance await experimental verification.

Collecting all ASE genes from the first Venn diagram in Fig. 2, we summarized the number of detected ASE genes on each chromosome (see Additional file 1: Table S1). We found several interesting patterns. For instance, chromosomes 11 and 21 tend to have more ASE genes than other chromosomes for all tissue types. Besides, the X chromosome has more ASE genes in brain tissue than other tissue types. Additional file 1: Figure S1 plots distributions of these ASE genes in each chromosome, revealing several ASE clusters. Among all detected ASE genes, most ASE genes show preference of the maternal allele than the paternal allele. Specifically, 79%, 74%, 68%, and 71% ASE genes show maternal preference in the brain, liver, kidney, and muscle tissues, respectively.

At this stage, we are not able to statistically distinguish imprinted genes from other type of ASE genes as further experiment data are required to separate imprinting from other ASE molecular mechanisms. However, collecting all the detected ASE genes from all three Venn diagrams in Fig. 2, we found that seven *de novo* detected ASE genes are highly likely to be imprinted in the bovine genome but they have not been documented in any bovine study. They are: (1) *GATM*, *SNX14*, and *NT5E*, which are imprinted in mouse [50, 51]; (2) *IGF1R* and *RCL1*, which are imprinted in human [52, 53]; and (3) *KLHDC10* and *SLC22A18*, which are imprinted in both human and mouse [54, 55]. These genes are involved in varied physiological functions. For example, *GATM* encodes an arginine glycine amidinotransferase (AGAT) which is involved in creatine synthesis [56, 57]. *NT5E* encodes the protein CD73 (cluster of differentiation 73), a cell surface anchored molecule with ectoenzymatic activity that catalyzes the

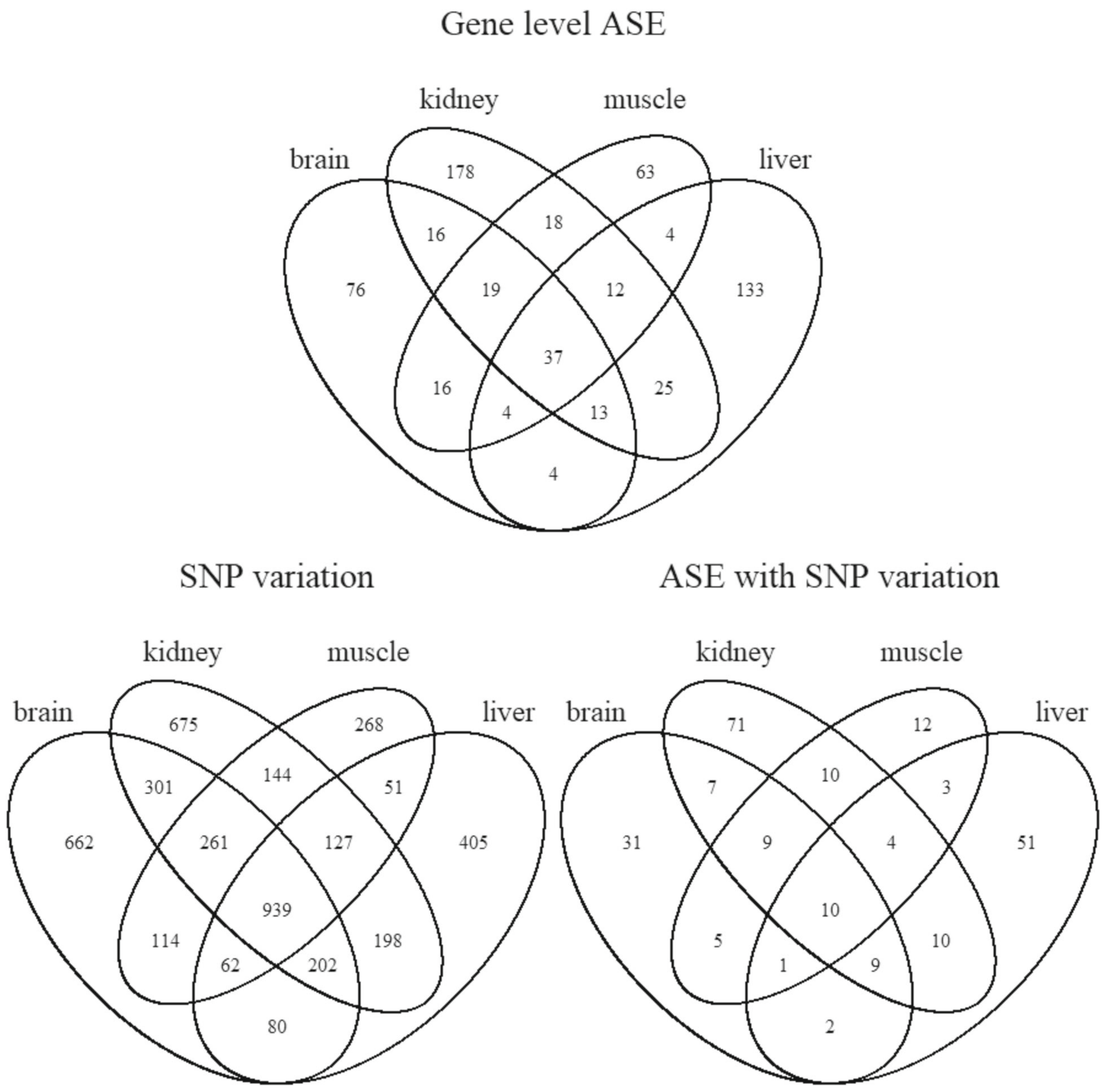

Fig. 2 Venn Diagram of detected ASEs across tissue types. Number of significant genes (estimated FDR=0.05) across four tissue types when testing ASE at the gene level, testing ASE variations across SNPs, and testing ASE gene and ASE variations within a gene simultaneously

hydrolysis of AMP into adenosine and phosphate and has been shown to mediate the invasive and metastatic properties of cancers [58, 59]. *SNX14* is a protein coding gene involved in maintaining normal neuronal excitability and synaptic transmission [51] and may be involved in intracellular trafficking [60]. *IGF1R* is a receptor tyrosine kinase that mediates the actions of insulin-like growth factor 1 (IGF1). *IGF1R* is involved in cell growth and survival and has a crucial role in tumor transformation and survival of malignant cells [61, 62]. *RCL1* is a protein-coding gene with roles in 18 S rRNA biogenesis and in the assembly of the 40 S ribosomal subunit [63, 64]. The Kelch repeat protein *KLHDC10* activates the apoptosis signal-regulating kinase 1 (ASK1) through the suppression of protein phophatase 5 [65] and activation of the ASK1 contributes in oxidative stress-mediated cell death through the activation of the JNK and p38 MAPK pathways [66]. *SLC22A18* plays a role in lipid metabolism [67] and also acts as a tumor suppressor [68]. Visualization of significant expression pattern of these seven genes are plotted in

Additional file 1: Figure S2 along with its significance level assessed by FDR.

Study on simulated data

Simulation design

Simulation studies based on real datasets can best evaluate empirical usage and performance. In this subsection, we introduce our approach to simulate data based on the real dataset in [23]. In the next subsection, we will compare the BLMRM method with the binomial test, ANOVA, MBASED, generalized linear mixed model (GLMM), and the BLMRM method with pure Laplace approximation.

In each simulation, we simulated 4000 genes in total with 1000 genes for each of the four models in $\mathcal{M}$. To base our simulation upon real datasets, we randomly selected 4000 genes from liver tissue in the real dataset and used the numbers of SNPs of these genes as the numbers of SNPs for the 4000 simulated genes. To ensure consistency with the real dataset, we set the number of biological replicates to be four.

Real data from liver tissue in [23] indicates a linear relationship between the logarithm of average total read counts and that of the sample standard deviation of total read counts within a gene across SNPs. Real data also indicates a roughly linear relationship between the logarithm of average total read counts and that of the sample standard deviation of total read counts within a SNP across four replicates. To simulate n_{gjk}, we utilized these two linear relationships. Specifically, let $\bar{n}_g$ denote the sample average of the total read count of gene g across SNPs; that is, $\bar{n}_g = \sum_{j=1}^{J_g}(\bar{n}_{gj})/J_g$ where $\bar{n}_{gj} = \sum_{k=1}^{K} n_{gjk}/\text{K}$. For the liver tissue in real data, by regressing $\log S(\bar{n}_g)$ on $\log(\bar{n}_g)$ with a simple linear model where $S(\cdot)$ denotes the sample standard deviation, we obtained fitted intercept $\widehat{\alpha}_1 = -0.36$ and slope $\widehat{\alpha}_2 = 0.97$. Hence, for each simulated gene, we independently sampled $\log \bar{n}_{g1}, \ldots, \log \bar{n}_{gJ_g} \sim \text{N}\,(\mu = \log \bar{n}_g, \text{ and } \sigma = \widehat{\alpha}_1 + \widehat{\alpha}_2 \log \bar{n}_g)$, where $\bar{n}_g$'s were computed from the 4,000 genes randomly selected from the real dataset. Next, we fit a linear regression model between $\log S(\bar{n}_{gj})$ and $\log(\bar{n}_{gj})$, which yielded an estimated intercept $\widehat{\alpha}_3 = -0.53$ and slope $\widehat{\alpha}_4 = 0.77$. Similarly, we simulated $n_{gj1}, \ldots, n_{gj4} \sim \text{N}\,(\mu = \log \bar{n}_{gj}, \sigma = \widehat{\alpha}_3 + \widehat{\alpha}_4 \log \bar{n}_{gj})$. We rounded the simulated values to ensure n_{gjk}'s were integers.

Given the simulated n_{gjk}'s, to simulate y_{gjk}'s, we needed to simulate p_{gjk}'s. We simulated gene effect β_g uniformly from $\{-4.39, -1.20, -0.41, 0.41, 1.20, 4.39\}$ for genes where $\beta_g \neq 0$. 0.41, 1.20, and 4.39 are the 10th, 50th, and 90th percentiles of absolute values of $\widehat{\beta}_g$'s, respectively, when significant gene ASEs are reported by the GLMM in (1). We simulated $\sigma^2_{sg} \overset{iid}{\sim} \text{IG}\,(\widehat{a}_s, \widehat{b}_s)$, $S_{gj} \overset{iid}{\sim} \text{N}\,(0, \sigma^2_{sg})$, and simulated $\sigma^2_{rg} \overset{iid}{\sim} \text{IG}\,(\widehat{a}_r, \widehat{b}_r)$, $R_{gk} \overset{iid}{\sim} \text{N}\,(0, \sigma^2_{rg})$, where $\widehat{a}_s$, $\widehat{b}_s$, $\widehat{a}_r$, and $\widehat{b}_r$ are hyperparameter estimates from the liver tissue whose values are given in real data analysis section. p_{gjk} was computed as $\exp(\beta_g + S_{gj} + R_{gk})/(1 + \exp(\beta_g + S_{gj} + R_{gk}))$. At last, we simulated $y_{gjk} \sim \text{Binomial}(n_{gjk}, p_{gjk})$. We repeated such simulation 10 times to assess variations in performance.

Simulation results

We compared our BLMRM method with the binomial test, ANOVA test in [12], MBASED method in [14], and GLMM in (1) without Bayesian priors. The binomial test and ANOVA test only detect the gene effect; the MBASED method can detect gene ASE and SNP variation separately but not simultaneously; and the GLMM and BLMRM methods can detect the gene effect, SNP variation, and gene ASE and SNP variation simultaneously. For the binomial, ANOVA, MBASED, and GLMM methods, we applied Storey's method [69] to estimate and control FDR. The FDR control for our BLMRM method was described in the "Method" section.

For the proposed BLMRM method, the hyperparameter estimation is accurate and stable across 10 simulations. The mean of absolute biases across 10 simulations are 0.61, 0.12, 0.08, and 0.06, respectively, for $\widehat{a}_s$, $\widehat{b}_s$, $\widehat{a}_r$, and $\widehat{b}_r$; and the standard deviations of these 10 absolute biases are 0.17, 0.08, 0.04, and 0.00.

Table 1 summarizes the average true FDR and average true positive rate (TPr) across 10 simulations when we control the estimated FDR at 0.05. Numbers in parentheses are sample standard deviations. Results suggested that among all methods under investigation, only our proposed method controlled FDR at the nominal level. The BLMRM method with pure Laplace approximation did not control FDR for simultaneous test on both gene effect

Table 1 Assess of FDR control and TPr when controlling estimated FDR at 0.05

Method	True FDR			TPr(%)		
	gene	SNP	gene-SNP	gene	SNP	gene-SNP
BLMRM	0.053	0.028	0.059	66.37	60.82	17.51
	(0.006)	(0.004)	(0.014)	(0.87)	(1.80)	(1.65)
BLMRM	0.060	0.030	0.094	68.87	56.82	17.50
(pure Laplace)	(0.006)	(0.002)	(0.008)	(0.29)	(1.19)	(0.91)
GLMM	0.073	0.006	0.625	68.66	57.20	86.72
	(0.010)	(0.002)	(0.004)	(1.52)	(1.49)	(0.86)
MBASED	0.358	0.032	-	91.34	64.32	-
	(0.006)	(0.005)	-	(0.54)	(1.51)	-
ANOVA	0.194	-	-	82.02	-	-
	(0.007)	-	-	(1.04)	-	-
Binomial	0.314	-	-	88.26	-	-
	(0.003)	-	-	(0.80)	-	-

and SNP variation. In addition, the proposed BLMRM method also had slightly higher TPr than the pure Laplace approximation approach in testing SNP variation. This suggested that the combined method of empirical Bayes and Laplace approximation provided more accurate results than three layers of Laplace approximation. The GLMM method was slightly liberal in testing gene ASE, overly conservative in testing the random SNP effect, and overly liberal in testing simultaneous gene ASE and SNP variation. The MBASED and binomial test methods did not control FDR when testing the gene effect. The MBASED method can not test gene ASE and ASE variation across SNPs simultaneously. Thus, under our simulation scenario, the MBASED method did not correctly separate observed variations among multiple sources of variations; i.e., gene ASE, SNP variation, biological variation, and error variation.

We plotted the box plots of true FDRs across 10 simulations in the left panel of Fig. 3, respectively, on testing the gene effect, SNP effect, and gene and SNP effects simultaneously when controlling the estimated FDR at 0.05, which represents same conclusions on FDR control in Table 1. The right panel in Fig. 3 displays the ROC curves when the false positive rate (FPr) was between 0 and 0.3. Compared to the other competing methods, the BLMRM method showed greater partial area under the ROC curves (AUCs) in testing gene ASE, SNP variation in ASE, and gene and SNP variation simultaneously. The GLMM and BLMRM methods were competitive for gene ranking when testing gene and SNP variation; however, the BLMRM method substantially outperformed the GLMM method in gene ranking when detecting simultaneous ASE gene effect and ASE variation within a gene.

Discussion

So far, no existing statistical methods can provide simultaneous inferences at both gene and exon (SNPs) levels for the entire genome in a short computational window, like the *de novo* detection for the bovine genome shown here. We are able to achieve this goal because we model

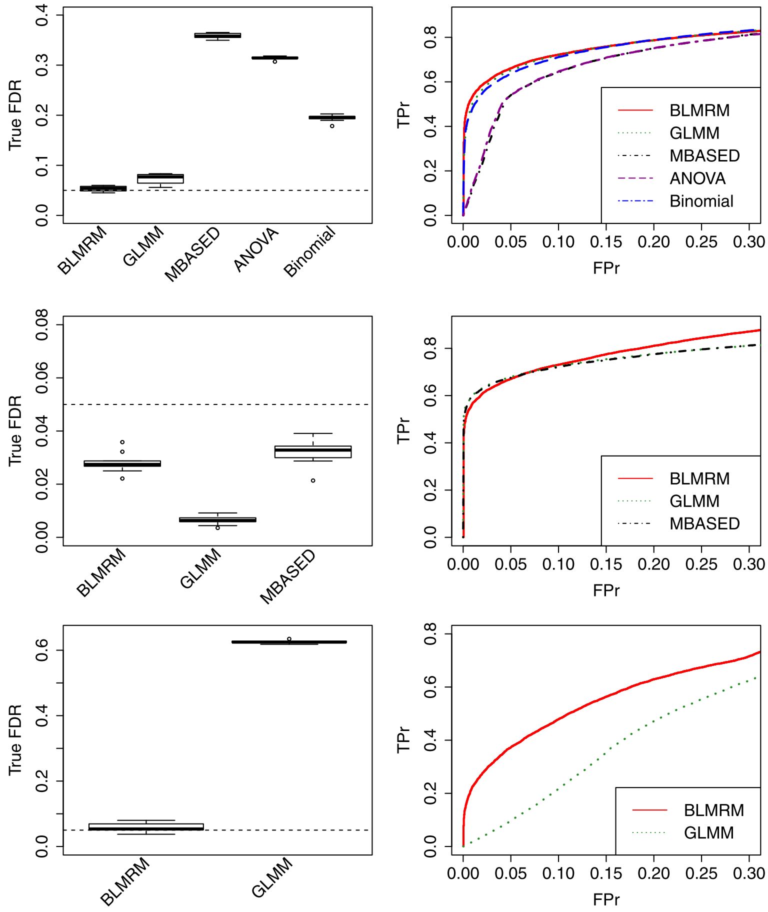

Fig. 3 FDR and ROC comparison. Top row shows results for testing the gene effect; middle row shows results for testing SNP variation within a gene; bottom row shows results for simultaneously testing gene ASE and SNP variation. Left panel shows box plots of true FDR across 10 simulations when controlling estimated FDR = 0.05; right panel presents ROC curves

multiple sources of variations (i.e., genes, SNPs, biological replicates, error variation) in one statistical model and adopt an efficient estimation method (i.e., a combination of empirical Bayes and Laplace approximation) for model selection, that is designed for whole genome analysis.

Conclusions

We have proposed a new method, BLMRM, to detect ASE for any RNA-seq experiment. Specifically, we propose a Bayesian logistic mixed regression model that accounts for variations from genes, SNPs, and biological replicates. To improve the reliability of inferences on ASE, we assign hyperpriors on genes, SNPs, and replicates, respectively. The hyperprior parameters are empirically estimated using observations from all genes in an entire genome. We then develop a Bayesian model selection method to test the ASE hypothesis on genes and variations of SNPs within a gene. To select a fitting model based on Bayes factors, we adopt a combination of the empirical Bayesian method and Laplace approximation method to substantially accelerate computation. To illustrate the utility of our method, we have applied the proposed approach to the bovine study that motivated our research; findings reveal the potential of our proposed method for application to real data analysis. We also conduct simulation studies that mimic the real data structure. Our data application and simulation study demonstrate the improved power, accuracy, and empirical utility of our proposed method compared to existing approaches. The R package, BLMRM, based on our method is available to download via Github at https://github.com/JingXieMIZZOU/BLMRM.

Method

Bayesian generalized linear mixed model

Let n_{gjk} denote the total number of read counts for the kth biological replicate of gene g at its jth SNP, where $g = 1, 2, \ldots, G, j = 1, 2, \ldots, J_g$, and$k = 1, 2, \ldots, K$. Let y_{gjk} denote the number of read counts from the maternal allele of replicate k. We model $y_{gjk} \sim \text{Binomial}(n_{gjk}, p_{gjk})$, where p_{gjk} denotes the proportion of gene expression from the maternal allele for gene g at SNP j of replicate k. It is known that using the RNA-seq approach to detect ASEs can produce bias during mapping because reads from the reference allele are more likely to be mapped due to fewer number of mismatches compared to reads from alternative alleles [30]. Potential solutions have been proposed in [23, 30, 70] to correct mapping bias. Here and throughout the paper, n_{gjk}'s and y_{gjk}'s denote the read counts after bias correction.

The objective of our study is to detect genes and regions within a gene whose expression is significantly different between the maternal and paternal alleles. Most existing methods assumed equal gene expression across all SNPs of a given gene; however, research discoveries have disproven this assumption for several reasons [71, 72], including alternative splicing and RNA variants. Thus, we model y_{gjk} as

$$y_{gjk} \sim \text{Binomial}(n_{gjk}, p_{gjk}), \text{ and}$$
$$\log \frac{p_{gjk}}{1 - p_{gjk}} = \beta_g + S_{gj} + R_{gk}, \tag{1}$$

where β_g is the fixed gene effect; S_{gj} is the random SNP effect and $S_{gj} \overset{iid}{\sim} \text{N}(0, \sigma^2_{sg})$; R_{gk} is the random replicate effect and $R_{gk} \overset{iid}{\sim} \text{N}(0, \sigma^2_{rg})$. We also assume S_{gj}'s and R_{gk}'s are mutually independent. Therefore, the null hypothesis $H_0 : \beta_g = 0$ is to test whether gene g exhibits imbalanced allelic expression. Furthermore, $H_0 : \sigma^2_{sg} = 0$ is to examine whether maternal (and/or paternal) gene expression percentage is the same across all SNPs of a gene.

Due to the expense of sample collection and sequencing experiments, most laboratories can only afford a few biological replicates, such as $K = 3$ or 4. In addition, the number of available SNPs in a gene also depends on the diversity between parental alleles. Often, only a small number of genes contain a large number of SNPs. Thus, for most genes, the estimates of β_g, σ^2_{sg}, and σ^2_{rg} are not robust, leading to unreliable statistical inferences. To improve estimation accuracy, we assume hierarchical priors on β_g, σ^2_{sg}, and σ^2_{rg} to share information across all genes in the genome. Specifically, we assume $\sigma^2_{sg} \overset{iid}{\sim} \text{IG}(a_s, b_s)$, $\sigma^2_{rg} \overset{iid}{\sim} \text{IG}(a_r, b_r)$, and a Gaussian prior on the gene effect $\beta_g \overset{iid}{\sim} \text{N}(\mu, \sigma^2)$. The hyperparameters a_s, b_s, a_r, b_r, μ, and σ^2 no longer have the subscript g because they are estimated by pooling observations from all genes. Given that there are tens of thousands of genes in the genome, the estimates of these prior hyperparameters are accurate.

Detection of imbalanced allelic gene expression through Bayesian model selection

Next, we describe our Bayesian model selection method to detect ASE at the gene level and corresponding variations across SNPs. Based on model (1), there are four models, indexed by $m \in \{1, 2, 3, 4\}$, in model space $\mathcal{M}$, where $\beta_g = 0$ and $\sigma^2_{sg} = 0$ in Model 1; $\beta_g \neq 0$ and $\sigma^2_{sg} = 0$ in Model 2; $\beta_g = 0$ and $\sigma^2_{sg} \neq 0$ in Model 3; and $\beta_g \neq 0$ and $\sigma^2_{sg} \neq 0$ in Model 4. For each gene g, we select model m in $\mathcal{M}$, which has the largest posterior probability defined as

$$P(m|\mathbf{y}^g, \mathbf{n}^g) = \frac{P(m)P(\mathbf{y}^g|m, \mathbf{n}^g)}{\sum_{m=1}^{4} P(m)P(\mathbf{y}^g|m, \mathbf{n}^g)}$$
$$\propto P(m)P(\mathbf{y}^g|m, \mathbf{n}^g), \tag{2}$$

where $\mathbf{y}^g = (y_{g11}, \ldots, y_{gJ_gK})'$ and $\mathbf{n}^g = (n_{g11}, \ldots, y_{gJ_gK})'$. $P(m)$ denotes the prior probability of model m. Without prior information, we assume a uniform prior on space

$\mathcal{M}$. Thus, our objective is to select a model m in $\mathcal{M}$ that maximizes the marginal likelihood $P(\mathbf{y}^g|m,\mathbf{n}^g)$, which, when comparing two models, is equivalent to choosing the model m using the Bayes factor. Let $\mathbf{b}_g$ denote all random effects; that is, $\mathbf{b}_g = (S_{g1},\ldots,S_{gJ_g},R_{g1},\ldots,R_{gK})'$. Hence,

$$P(\mathbf{y}^g|m,\mathbf{n}^g) = \iiiint P(\mathbf{y}^g|\beta_g,\mathbf{b}_g,\mathbf{n}^g,m)P(\beta_g)\times P(\mathbf{b}_g|\sigma^2_{sg},\sigma^2_{rg})P(\sigma^2_{sg},\sigma^2_{rg})\times d\beta_g\, d\mathbf{b}_g\, d\sigma^2_{sg}\, d\sigma^2_{rg}. \tag{3}$$

A direct integration of (3) is difficult because an analytical result of the density is not a closed form. An alternative approach is to use Laplace approximation to iteratively approximate each integral; however, in our experience, this leads to error accumulated through each layer of integration and thus affects the accuracy of results. To overcome this problem, we propose a combination of empirical Bayes estimation and Laplace approximation. Inspired by the approach in [73], we obtain the following empirical Bayes estimators.

$$\widetilde{\beta}_g = E(\beta_g|\widehat{\beta}_g) \approx \frac{\widehat{\mathrm{Var}(\beta_g)}\widehat{\mu} + \widehat{\sigma}^2\widehat{\beta}_g}{\widehat{\mathrm{Var}(\beta_g)} + \widehat{\sigma}^2}, \tag{4}$$

$$\widetilde{\sigma}^2_{sg} = E(\sigma^2_{sg}|\widehat{\sigma}^2_{sg}) \approx \frac{d_{sg}\widehat{\sigma}^2_{sg} + 2\widehat{b}_s}{d_{sg} + 2\widehat{a}_s}, \text{ and} \tag{5}$$

$$\widetilde{\sigma}^2_{rg} = E(\sigma^2_{rg}|\widehat{\sigma}^2_{rg}) \approx \frac{d_{rg}\widehat{\sigma}^2_{rg} + 2\widehat{b}_r}{d_{rg} + 2\widehat{a}_r}, \tag{6}$$

where $\widetilde{\beta}_g$, $\widetilde{\sigma}^2_{sg}$, and $\widetilde{\sigma}^2_{rg}$ denote the empirical Bayes estimates of β_g, σ^2_{sg}, and σ^2_{rg}, respectively. $\widehat{\beta}_g$, $\widehat{\mathrm{Var}(\beta_g)}$, $\widehat{\sigma}^2_{sg}$, and $\widehat{\sigma}^2_{rg}$ are maximum likelihood estimates from model (1). $\widehat{\mu}$, $\widehat{\sigma}^2$, $\widehat{a}_r$, $\widehat{b}_r$, $\widehat{a}_s$, and $\widehat{b}_s$ are estimated hyperparameters whose estimation method will be introduced in detail later in this section. d_{rg} and d_{sg} are degrees of freedom of the random SNP and random replicate effect, respectively, with $d_{sg} = J_g - 1$ and $d_{rg} = K - 1$. We enter these empirical Bayes estimates directly into (3), obtaining the approximation:

$$P(\mathbf{y}^g|m,\mathbf{n}^g) \approx \int P(\mathbf{y}^g|\widetilde{\beta}_g,\mathbf{b}_g,m,\mathbf{n}^g)\times P(\mathbf{b}_g|\widetilde{\sigma}^2_{sg},\widetilde{\sigma}^2_{rg})\, d\mathbf{b}_g. \tag{7}$$

Accordingly, (3) is reduced to (7), which requires only one step of Laplace approximation. Our objective in combining empirical Bayes estimates and Laplace approximation is to develop a method with improved power and accuracy while maintaining affordable computational power that allows for empirical application. In our simulation study, we compared our proposed approach with the method using pure Laplace approximation. We found that our proposed method is superior than purely using Laplace approximation with respect to FDR control and true positive rate (see "Simulation results" section). This approach also greatly decreases computational requirements compared to MCMC, considering there are tens of thousands of genes in an entire genome [74]. For instance, the method in [19] employs an MCMC algorithm for identifying ASE. With the default setting, their approach took approximately 1.5 hours to analyze 50 genes, whereas our method took approximately 3 minutes.

We still need to estimate hyperparameters μ, σ^2, a_s, b_s, a_r, and b_r. To avoid extreme values that produce unstable estimates, we first let $y^*_{gjk} = y_{gjk} + 1$ and $n^*_{gjk} = n_{gjk} + 2$. Then, based on y^*_{gjk}'s and n^*_{gjk}'s, μ and σ^2 are estimated by the method of moments using significant $\widehat{\beta}_g$ via likelihood ratio tests when controlling FDR at 0.05. a_s, b_s, a_r, and b_r are estimated based on y^*_{gjk}'s and n^*_{gjk}'s by the maximum likelihood method, where a_s and b_s are based on significant estimates of $\widehat{\sigma}^2_{sg}$'s via likelihood ratio tests and controlling FDR at 0.05, and a_s and b_s are based on $\widehat{\sigma}^2_{rg}$'s from all genes.

Finally, we test $H_0 : \beta_g = 0$ and $H_0 : \sigma^2_{sg} = 0$ for gene g by choosing Model m, where $m = \underset{\gamma\in\{1,2,3,4\}}{\arg\max} P(\gamma|\mathbf{y}^g,\mathbf{n}^g)$ for $g = 1,\ldots,G$. Let $P(g \in \{m\}|\mathbf{y}^g,\mathbf{n}^g)$ denote the posterior probability of gene g being sampled from Model m. The posterior probability of a gene exhibiting an ASE gene effect is $P(g \in \{2,4\}|\mathbf{y}^g,\mathbf{n}^g)$. Similarly, the posterior probability of a gene exhibiting ASE variations across SNPs is $P(g \in \{3,4\}|\mathbf{y}^g,\mathbf{n}^g)$. Finally, the posterior probability of a gene exhibiting an ASE gene effect and ASE variations across SNPs simultaneously is $P(g \in \{4\}|\mathbf{y}^g,\mathbf{n}^g)$. We adopt the following method to control FDR that have been used in [74, 75]. To control the FDR when testing the ASE gene effect, we order $P(g \in \{2,4\}|\mathbf{y}^g,\mathbf{n}^g)$, $g = 1,\ldots,G$, from largest to smallest. Let $g_{(1)},\ldots,g_{(G)}$ be the ordered genes; then, we find the largest l such that $\sum_{i=1}^{l}(1 - P(g_{(i)} \in \{2,4\}|\mathbf{y}^{g_{(i)}},\mathbf{n}^{g_{(i)}}))/l \leq \alpha$, where α is a pre-defined FDR threshold. We declare the first l genes are significant for testing $H_0 : \beta_g = 0$ when FDR is controlled at α level. The same strategy is used to control FDR for testing ASE variations among SNPs and gene and SNP variation effects simultaneously.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3141-6.

Additional file 1: Supplementary Materials for "Modeling Allele-Specific Expression at the Gene and SNP Levels Simultaneously by a Bayesian Logistic Mixed Regression Model".

Abbreviations

ANOVA: Analysis of variance; ASE: Allele-specific expression; AUC: Area under ROC curve; BLMRM: Bayesian logistic mixed regression model; BWS: Beckwith-Wiedemann syndrome; DNA-seq: next generation sequencing of

DNA; FDR: False discovery rate; FPr: False positive rate; GATK: Genome Analysis Toolkit; GLMM: Generalized linear mixed model; MCMC: Markov chain Monte Carlo; SNP: Single nucleotide polymorphism; TPr: True positive rate

Acknowledgements
We would like to thank the reviewers for their invaluable contributions to the peer review process.

Availability and requirements
Project name: BLMRM. **Project home page**: https://github.com/JingXieMIZZOU/BLMRM. **Operating system(s)**: Platform independent. **Programming language**: R.

Authors' contributions
JX, TJ, and MARF participated in the method design, simulation study, and result analysis. YL, BNP, and RMR annotated the biological results. All authors drafted the manuscript. TJ, MARF, and RMR revised the manuscript. JX implemented the method in R code. All authors have read and approved the final version of the manuscript.

Funding
This work was supported by the National Science Foundation (1615789 and 1853556 to TJ), National Science Foundation (1853549 to MARF), National Institutes of Health (5R21HD062920 to RMR), Agriculture and Food Research Initiative Competitive Grant No. 2018-67015-27598 from the USDA National Institute of Food (to RMR) and Agriculture, and the Food for the 21st Century Program at the University of Missouri - Animal Reproduction (to RMR). The funding bodies did not play any role in the design of the study, analysis and interpretation of data, or in writing the manuscript.

Availability of data and materials
The allele-specific expression data for the bovine study are publicly available at Gene Expression Omnibus with accession no. GSE63509. The R package, BLMRM, is publicly available at https://github.com/JingXieMIZZOU/BLMRM.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Statistics, University of Missouri at Columbia, Columbia, MO 65211, USA. [2]Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA. [3]Division of Animal Science, University of Missouri at Columbia, Columbia, MO 65211, USA.

Received: 3 May 2019 Accepted: 9 October 2019
Published online: 28 October 2019

References

1. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in x-linked gene expression in females. Nature. 2005;434:400–4.
2. Bartolomei MS. Genomic imprinting: employing and avoiding epigenetic processes. Gene Dev. 2009;23:2124–33.
3. Solter D. Differential imprinting and expression of maternal and paternal genomes. Annu Rev Genet. 1988;22:127–46.
4. Wutz A. Gene silencing in x-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat Rev Genet. 2011;12:542–53.
5. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding rnas in health and disease. Cell. 2013;152:1308–23.
6. Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KC, Gagné V, Dias J, Hoberman R, Montpetit A, Joly MM, Harvey EJ, Sinnett D, Beaulieu P, Hamon R, Graziani A, Dewar K, Harmsen E, Majewski J, Göring HH, Naumova AK, Blanchette M, Gunderson KL, Pastinen T. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat Genet. 2009;41:1216–22.
7. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet. 2010;11:533–8.
8. Cohen MJ. Beckwith-wiedemann syndrome: historical, clinicopathological, and etiopathogenetic perspectives. Pediatr Dev Pathol. 2005;8:287–304.
9. Weksberg R, Shuman C, Smith A. Beckwith-wiedemann syndrome. Am J Med Genet C. 2005;137:12–23.
10. Angelman H. "puppet" children: A report of three cases. J Dev Med Child Neurol. 1965;7:681–8.
11. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M. Alleleseq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011;7:522.
12. Pandey R, Franssen S, Futschik A, Schlötterer C. Allelic imbalance metre (allim), a new tool for measuring allele-specific gene expression with rna-seq data. Mol Ecol Resour. 2013;13:740–5.
13. Lu R, Smith RM, Seweryn M, Wang D, Hartmann K, Webb A, Sadee W, Rempala GA. Analyzing allele specific rna expression using mixture models. BMC Genomics. 2015;16:566.
14. Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, Watanabe C, Zhang Z. Mbased: allele-specific expression detection in cancer tissues and cell lines. Genome Biol. 2014;15:405.
15. Harvey C, Moyerbrailean G, Davis G, Wen X, Luca F, Pique-Regi R. Quasar: quantitative allele-specific analysis of reads. Bioinformatics. 2015;31:1235–42.
16. Nembaware V, Wolfe K, Bettoni F, Kelso J, Seoighe C. Allele-specific transcript isoforms in human. FEBS Lett. 2004;577:233–8.
17. Wang E, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore S, Schroth G, Burge C. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456:470–6.
18. Graveley BR. The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population. Trends Genet. 2008;24:5–7.
19. Skelly D, Johansson M, Madeoy J, Wakefield J, Akey J. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from rna-seq data. Genome Res. 2011;21:1728–37.
20. Knowles D, Davis J, Edgington H, Raj A, Favé M, Zhu X, Potash J, Weissman M, Shi J, Levinson D, Awadalla P, Mostafavi S, Montgomery S, Battle A. Allele-specific expression reveals interactions between genetic variation and environment. Nat Methods. 2017;14:699–702.
21. van de Geijn B, McVicker G, Gilad Y, Pritchard J. Wasp: allele-specific software for robust molecular quantitative trait locus disocvery. Nat Methods. 2015;12:1061–3.
22. Kumasaka N, Knights A, Gaffney D. Fine-mapping cellular qtls with rasqual and atac-seq. Nat Genet. 2016;48:206–13.
23. Chen Z, Hagen D, Wang J, Elsik C, Ji T, Siqueira L, Hansen P, Rivera R. Global assessment of imprinted gene expression in the bovine conceptus by next generation sequencing. Epigenetics. 2016;11:501–16.
24. Wilkins J, Úbeda F, Van Cleve J. The evolving landscape of imprinted genes in humans and mice: Conflict among alleles, genes, tissues, and kin. Bioessays. 2016;38:482–9.
25. McGrath J, Solter D. Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell. 1984;37:179–83.
26. Surani M, Barton S, Norris M. Nuclear transplantation in the mouse: heritable differences between parental genomes after activation of the embryonic genome. Cell. 1986;45:127–36.
27. Barlow D, Stöger R, Herrmann B, Saito K, Schweifer N. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the tme locus. Nature. 1991;349:84–7.
28. Hsu C, Chou C, Huang S, Lin C, Lin M, Tung C, Lin C, Lai I, Zou Y, Youngson N, Lin S, Yang C, Chen S, Gau S, Huang H. Analysis of experience-regulated transcriptome and imprintome during critical periods of mouse visual system development reveals spatiotemporal dynamics. Hum Mol Genet. 2018;27:1039–54.
29. Okamoto I, Patrat C, Thépot D, Peynot N, Fauque P, Daniel N, Diabangouaya P, Wolf J, Renard J, Duranthon V, Heard E. Eutherian mammals use diverse strategies to initiate x-chromosome inactivation during development. Nature. 2011;472:370–4.
30. Degner J, Marioni J, Pai A, Pickrell J, Nkadori E, Gilad Y, Pritchard J. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. Bioinformatics. 2009;25:3207–12.

31. Stevenson K, Coolon J, Wittkopp P. Sources of bias in measures of allele-specific expression derived from rna-seq data aligned to a single reference genome. BMC Genomics. 2013;14:536.
32. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo M. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res. 2010;20:1297–303.
33. Li H. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
34. Kim D, Langmead B, Salzberg S. Hisat: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
35. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.
36. Li-Calzi M, Raviolo C, Ghibaudi E, De Gioia L, Salmona M, Cazzaniga G, Kurosaki M, Terao M, Garattini E. Purification, cdna closing, and tissue distribution of bovine liver aldehyde oxidase. J Biol Chem. 1995;270: 31037–45.
37. Fu C, Di L, Han X, Soderstrom C, Snyder M, Troutman MD, Obach RS, Zhang H. Aldehyde oxidase 1 (aox1) in human liver cytosols: quantitative characterization of aox1 expression level and activity relationship. Drug Metab Dispos. 2013;41:1797–804.
38. Foulon V, Sniekers M, Huysmans E, Asselberghs S, Mahieu V, Mannaerts GP, Van Veldhoven PP, Casteels M. Breakdown of 2-hydroxylated straight chain fatty acids via peroxisomal 2-hydroxyphytanoyl-coa lyase: a revised pathway for the alpha-oxidation of straight chain fatty acids. J Biol Chem. 2005;280:9802–12.
39. Stelzer Y, Bar S, Bartok O, Afik S, Ronen D, Kadener S, Benvenisty N. Differentiation of human parthenogenetic pluripotent stem cells reveals multiple tissue- and isoform-specific imprinted transcripts. Cell Rep. 2015;11:308–20.
40. Denley A, Cosgrove LJ, Booker GW, Wallace JC, Forbes BE. Molecular interactions of the igf system. Cytokine Growth Factor Rev. 2005;16: 421–39.
41. Ludwig T, Eggenschwiler J, Fisher P, D'Ercole AJ, Davenport ML, Efstratiadis A. Mouse mutants lacking the type 2 igf receptor (igf2r) are rescued from perinatal lethality in Igf2 and Igf1r null backgrounds. Dev Biol. 1996;177:517–35.
42. Chen Z, Hagen D, Elsik C, Ji T, CJ M, Moon L, Rivera R. Characterization of global loss of imprinting in fetal overgrowth syndrome induced by assisted reproduction. Proc Natl Acad Sci. 2015;112:4618–23.
43. A S, Papageorghiou A, Nicolaides K, Alley M, Jim A, Nargund G, Ojha K, Campbell S, Banerjee S. Temporal regulation of the expression of syncytin (herv-w), maternally imprinted peg10, and sgce in human placenta. Biol Reprod. 2003;69:286–93.
44. Piras G, El Kharroubi A, Kozlov S, Escalante-Alcalde D, Hernandez L, Copeland N, Gilbert D, Jenkins N, Stewart C. Zac1 (lot1), a potential tumor suppressor gene, and the gene for epsilon-sarcoglycan are maternally imprinted genes: identification by a subtractive screen of novel uniparental fibroblast lines. Mol Cell Biol. 2000;20:3308–15.
45. Ruddock N, Wilson K, Cooney M, Korfiatis N, Tecirlioglu R, French A. Analysis of imprinted messenger rna expression during bovine preimplantation development. Biol Reprod. 2004;70:1131–5.
46. Kamiya M, Judson H, Okazaki Y, Kusakabe M, Muramatsu M, Takada S, Takagi N, Arima T, Wake N, Kamimura K, Satomura K, Hermann R, Bonthron D, Hayashizaki Y. The cell cycle control gene zac/plagl1 is imprinted – a strong candidate gene for transient neonatal diabetes. Hum Mol Genet. 2000;9:453–60.
47. Robbins K, Chen Z, Wells K, Rivera R. Expression of kcnq1ot1, cdkn1c, h19, and plagl1 and the methylation patterns at the kvdmr1 and h19/igf2 imprinting control regions is conserved between human and bovine. J Biomed Sci. 2012;19:95.
48. Tierling S, Gasparoni G, Youngson N, Paulsen M. The begain gene marks the centromeric boundary of the imprinted region on mouse chromosome 12. Mamm Genome. 2009;20:699–710.
49. Smit M, Tordoir X, Gyapay G, Cockett N, Georges M, Charlier C. Begain: a novel imprinted gene that generates paternally expressed transcripts in a tissue- and promotor-specific manner in sheep. Mamm Genome. 2005;16: 801–14.
50. Sandell L, Guan X, Ingram R, Tilghman S. Gatm, a creatine synthesis enzyme, is imprinted in mouse placenta. Proc Natl Acad Sci. 2003;100: 4622–7.
51. Huang HS, Yoon BJ, Brooks S, Bakal R, Berrios J, Larsen RS, Wallace ML, Han JE, Chung EH, Zylka MJ, Philpot BD. Snx14 regulates neuronal excitability, promotes synaptic transmission, and is imprinted in the brain of mice. PLoS ONE. 2014;9:98383.
52. Sun J, Li W, Sun Y, Yu D, Wen X, Wang H, Cui J, Wang G, Hoffman A, Hu J. A novel antisense long noncoding rna within the igf1r gene locus is imprinted in hematopoietic malignancies. Nucleic Acids Res. 2014;42: 9588–601.
53. Stelzer Y, Ronen D, Bock C, Boyle P, Meissner A, Benvenisty N. Identification of novel imprinted differentially methylated regions by global analysis of human-parthenogenetic-induced pluripotent stem cells. Stem Cell Rep. 2013;1:79–89.
54. Hamada H, Okae H, Toh H, Chiba H, Hiura H, Shirane K, Sato T, Suyama M, Yaegashi N, Sasaki H, Arima T. Allele-specific methylome and transcriptome analysis reveals widespread imprinting in the human placenta. Am J Hum Genet. 2016;99:1045–58.
55. Babak T, DeVeale B, Tsang E, Zhou Y, Li X, Smith K, Kukurba K, Zhang R, Li J, van der Kooy D, Montgomery S, Fraser H. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. Nat Genet. 2015;47:544–9.
56. Joncquel-Chevalier Curt M, Voice PM, Fontaine M, Dessein AF, Porchet N, Mention-Mulliez K, Dobbelaere D, Soto-Ares G, Cheillan D, Vamecq J. Creatine biosynthesis and transport in health and disease. Biochimie. 2015;119:146–65.
57. Stockler-Ipsiroglu A, Apatean D, Battini R, DeBrosse S, Dessoffy K, Edvardson S, Eichler F, Johnston K, Keller DM, Nouioua S, Tapir M, Verma A, Dowling MD, Wierenga KJ, Wierenga AM, Zhang V, Wong LJ. Arginine: glycine amidinotransferase (agat) deficiency: Clinical features and long term outcomes in 16 patients diagnosed worldwide. Mol Genet Metab. 2015;116:252–9.
58. Zhi X, Cheng S, Zhou P, Chao Z, Wang L, Ou Z, Yin L. Rna interference of echo-5'-nucleotidase (cd73) inhibits human breast cancer cell growth and invasion. Clin Exp Metastasis. 2007;24:439–48.
59. Ghiringhelli F, Bruchard M, Chalmin F, Rébé C. Production of adenosine by ectonucleotidases: A key factor in tumor immunoescape. J Biomed Biotechnol. 2012;2012:473712.
60. Teasdale RD, Collins BM. Insights into the px (pho-homology) domain and snx (sorting nexin) protein families: structures, functions and roles in disease. Biochem J. 2012;441:39–59.
61. Wang Y, Cheng Z, Elalieh HZ, Nakamura E, Nguyen MT, Mackem S, Clemens TL, Bikle DD, Chang W. Igf-1r signaling in chondrocytes modulates growth plate development by interacting with the pthrp/ihh pathway. J Bone Miner Res. 2011;26:1437–46.
62. Kasprzak A, Kwasniewski W, Adamek A, Gozdzicka-Jozefiak A. Insulin-like growth factor (igf) axis in cancerogenesis. Mutat Res / Rev Mutat Res. 2017;772:78–104.
63. Billy E, Wegierski T, Nasr F, Filipowicz W. Rcl1p, the yeast protein similar to the rna 3'-phosphate cyclase, associates with u3 snornp and is required for 18s rrna biogenesis. EMBO J. 2000;19:2115–26.
64. Karbstein K, Jonas S, Doudna JA. An essential gtpase promotes assembly of preribosomal rna processing complexes. Mol Cell. 2005;20:633–43.
65. Sekine Y, Hatanaka R, Watanabe T, Sono N, Immure S, Natsume T, Kuranaga E, Miura M, Takeda K, Ichijo H. The kelch repeat protein klhdc10 regulates oxidative stress-induced ask1 activation by suppressing pp5. Mol Cell. 2012;48:692–704.
66. Soga M, Matsuzawa A, Ichjio H. Oxidative stress-induced diseases via the ask1 signaling pathway. Int J Cell Biol. 2012;2012:439587.
67. Ito S, Honda G, Fujino Y, Ogata S, Hirayama-Kurogi M, Ohtsuki S. Knockdown of orphan transporter slc22a18 impairs lipid metabolism and increases invasiveness of hepg2 cells. Pharm Res. 2019;36:39.
68. Schwienbacher C, Gramantieri L, Scelfo R, Veronese A, Calin GA, Bolondi L, Croce CM, Barbanti-Brodano G, Negrini M. Gain of imprinting at chromosome 11p15: A pathogenetic mechanism identified in human hepatocarcinomas. Proc Natl Acad Sci. 2000;97:5445–9.
69. Storey J. A direct approach to false discovery rates. J R Stat Soc Ser B. 2002;64:479–98.
70. Satya RV, Zavaljevski N, Reifman J. A new strategy to reduce allelic bias in rna-seq readmapping. Nucleic Acids Res. 2012;40:127.
71. Blagitko N, Mergenthaler S, Schulz U, Wollmann HA, Craigen W, Eggermann T, Ropers H-H, Kalscheuer VM. Human grb10 is imprinted an expressed from the paternal and maternal allele in a highly tissue- and isoform-specific fashion. Hum Mol Genet. 2000;9:1587–95.

72. Croteau S, Charron M-C, Latham KE, Naumova AK. Alternative splicing and imprinting control of the meg3/gtl2-dlk1 locus in mouse embryos. Mamm Genome. 2003;14:231–41.
73. Ritchie M, Phipson B, Wu D, Hu Y, Law C, Shi W, Smyth G. *limma* powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res. 2015;43:47.
74. Ji T, Liu P, Nettleton D. Estimation and testing of gene expression heterosis. J Agric Biol Environ Stat. 2014;19:319–37.
75. Cui S, Guha S, Ferreira MAR, Tegge AN. hmmseq: a hidden markov model for detecting differentially expressed genes from rna-seq data. Ann Appl Stat. 2015;9:901–25.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.