BMC
Genomics

METHODOLOGY ARTICLE

Open Access

A flexible Bayesian method for detecting allelic imbalance in RNA-seq data

Luis G León-Novelo[1], Lauren M McIntyre[2], Justin M Fear[2] and Rita M Graze[3*]

Abstract

Background: One method of identifying *cis* regulatory differences is to analyze allele-specific expression (ASE) and identify cases of allelic imbalance (AI). RNA-seq is the most common way to measure ASE and a binomial test is often applied to determine statistical significance of AI. This implicitly assumes that there is no bias in estimation of AI. However, bias has been found to result from multiple factors including: genome ambiguity, reference quality, the mapping algorithm, and biases in the sequencing process. Two alternative approaches have been developed to handle bias: adjusting for bias using a statistical model and filtering regions of the genome suspected of harboring bias. Existing statistical models which account for bias rely on information from DNA controls, which can be cost prohibitive for large intraspecific studies. In contrast, data filtering is inexpensive and straightforward, but necessarily involves sacrificing a portion of the data.

Results: Here we propose a flexible Bayesian model for analysis of AI, which accounts for bias and can be implemented without DNA controls. In lieu of DNA controls, this Poisson-Gamma (PG) model uses an estimate of bias from simulations. The proposed model always has a lower type I error rate compared to the binomial test. Consistent with prior studies, bias dramatically affects the type I error rate. All of the tested models are sensitive to misspecification of bias. The closer the estimate of bias is to the true underlying bias, the lower the type I error rate. Correct estimates of bias result in a level alpha test.

Conclusions: To improve the assessment of AI, some forms of systematic error (e.g., map bias) can be identified using simulation. The resulting estimates of bias can be used to correct for bias in the PG model, without data filtering. Other sources of bias (e.g., unidentified variant calls) can be easily captured by DNA controls, but are missed by common filtering approaches. Consequently, as variant identification improves, the need for DNA controls will be reduced. Filtering does not significantly improve performance and is not recommended, as information is sacrificed without a measurable gain. The PG model developed here performs well when bias is known, or slightly misspecified. The model is flexible and can accommodate differences in experimental design and bias estimation.

Keywords: Allelic imbalance, Allele-specific expression, RNA-seq, Systematic error, Bayesian model

Background

Sequence polymorphisms which impact gene expression have been identified as an important factor in human disease (Reviewed in [1-6]); explaining phenotypic differences between individuals (e.g., drug response [7]; biometric traits [8]) and species (e.g., ecological and reproductive traits [9-18]). A variety of experimental designs and analytical methods have been employed to identify the genetic basis of regulatory variation, finding abundant variation in both *cis* and in *trans* regulatory mechanisms [19-32].

In this study we focus on analytical approaches for the analysis of allelic imbalance (AI), a common method used to identify genetic differences in gene regulation. Allelic imbalance occurs when regulatory processes result in different steady-state transcript levels for the two alleles (within a single individual). Genetic differences in the regulation of transcript abundance for a focal gene can arise from regulatory sequence variation occurring within regulatory regions of that gene (*cis* effects) or in regulatory

*Correspondence: rmgraze@auburn.edu
[3]Department of Biological Sciences, Auburn University, 101 Rouse Life Science Building, 36849 Auburn, AL, USA
Full list of author information is available at the end of the article

or coding regions of *trans* acting factors (*trans* effects) or through indirect or epistatic effects. The two alleles in a diploid individual are expressed in a common cellular environment. Alleles expressed in a common cellular environment can differ in regulatory sequence, but share a common pool of *trans* acting factors. Therefore, allelic imbalance between alleles in a common cellular environment reveals functional differences between alleles in *cis* regulatory regions [20,22]. While comparing the same allele in different cellular environments (e.g., between genotypes) reveals differences in *trans* regulation because *cis* regulatory elements are identical while *trans* environments differ.

Early studies of AI focused on a limited number of genes and a few genotypes (e.g., parental genotypes and their F1 progeny, [22,29]). Different technologies have also been employed, including custom platforms [33], SNP detection [20,22] and arrays [29,34]. Currently, the technology most frequently used to asses allele-specific expression is RNA-seq (e.g., [32,35-38]).

The null expectation, when there is no AI, is that the two alleles are expressed equally (termed allelic balance or AB throughout). That is, the proportion of the total expression level contributed by the maternal/paternal allele is equal to a half. A common approach to analysis of allelic imbalance in RNA-seq data is use of a binomial or chi-square test to determine if allele-specific read counts depart from this expected proportion [32,35,39,40]. However, these tests do not necessarily have the correct error variance [41-45]. Bayesian models have been proposed to improve estimates of allele-specific expression and/or to identify allelic imbalance [46-48].

These Bayesian methods have primarily focused on proper handling of error variance in the statistical model. However, bias in estimation of AI is an important issue for both intraspecific [35,49] and interspecific [38,48,50] studies. Biases are present when aligning to a single reference, a single reference with SNPs masked, and multiple references; which can result in false positives for AI [35,49-51]. Bias in estimation of allele-specific expression or allelic imbalance has multiple sources, including sequence differences between reads and reference (missed SNPs/false SNPs), properties of alignment algorithms, genome features that result in ambiguity of read alignments and other technical sources of error (Figure 1) [35,52,53].

There are several approaches to dealing with bias in studies of AI, which are not necessarily mutually exclusive. Some analytical methods reduce bias resulting from differences between allele-specific reads and a single reference by the use of multiple strain specific references (e.g., [46,48]) or by allele augmented references [47,51]. To account for bias due to properties of the alignment, simulated reads have also been used to filter SNPs or

other units of analysis that show bias in alignments [35,46,51,54]. For simple F1 experimental designs, the use of DNA controls works quite well [21,22,29] to either filter biased regions [50] or to estimate bias and directly account for bias, technical error variance and biological error variance in the statistical model [48].
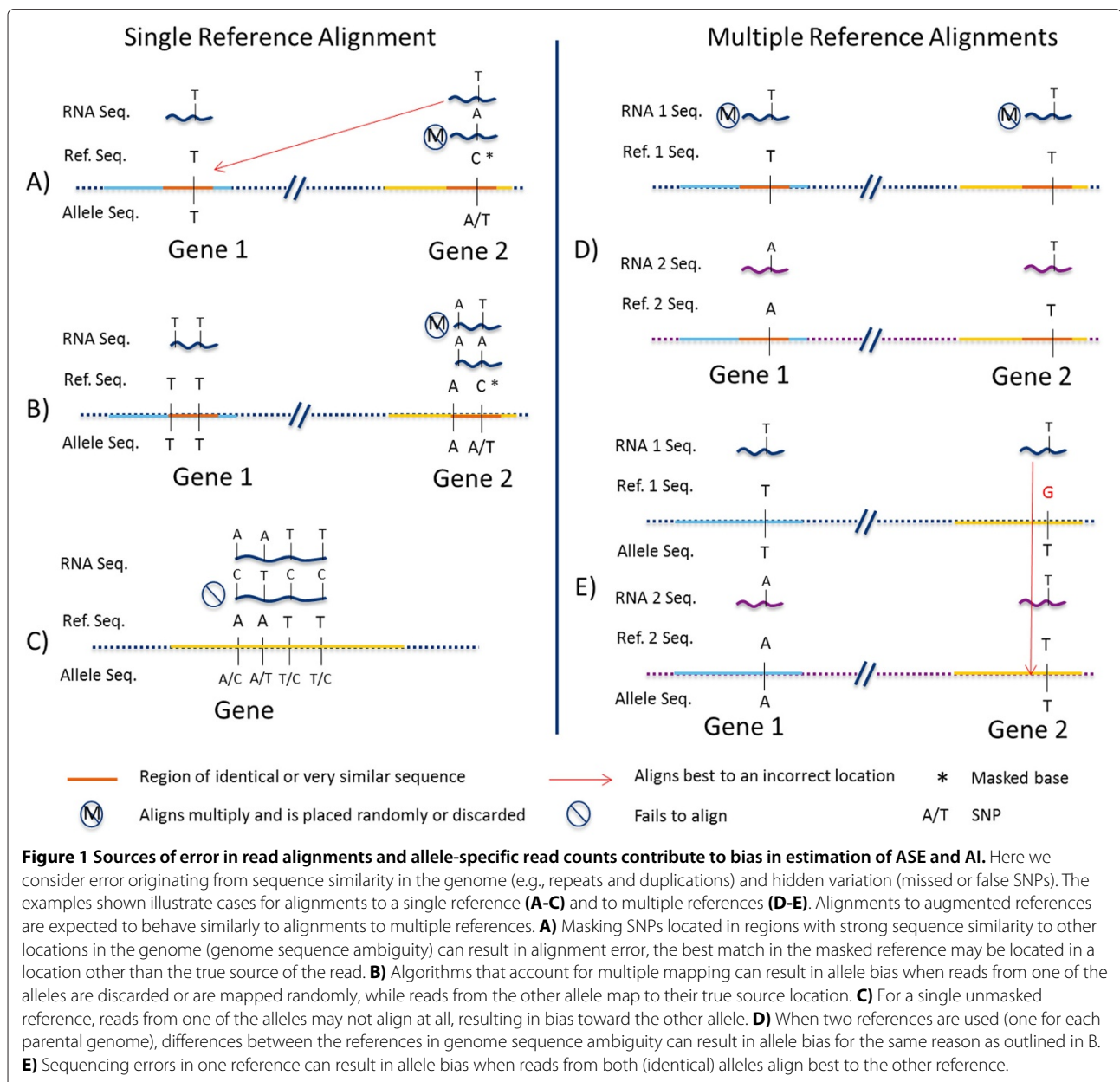
DNA sequencing of F1 heterozygotes (DNA controls) is used to determine allele-specific read counts in the case where equal amounts of each allele are present in the sample. If there is bias in the DNA read counts the paternal/maternal proportion of reads will deviate from 0.5. Because the measurements from DNA controls reflect the complete process of allele-specific read assessment, the error can originate from properties of the genome (ambiguity, missed variation), mapping, or sequencing related technical bias. This is in contrast to existing filtering approaches, which only capture sources of bias related to ambiguity and mapping. However, use of DNA controls can be cost prohibitive in intraspecific experiments, where the number of genotypes evaluated is expected to be quite high.

In this manuscript we introduce a Bayesian Poisson-Gamma (PG) model for analysis of allelic imbalance. The PG model is a Bayesian version of Poisson regression. This model can be used when DNA controls are not available through use of a parameter representing bias which is incorporated into the structure of the model. The parameter can be fixed (q) or random (ϕ) and can be used in conjunction with simulation to account for genome ambiguity and map bias.

Results and discussion

To compare model performance under different scenarios we generated allele-specific read counts for both RNA and DNA controls from a Poisson distribution (see Additional file 1). While total allele-specific reads are distributed similar to real data, bias and the ratio of the two allele mean counts are specified by the parameters B and R respectively. We investigated model performance for a previously developed Bayesian negative binomial (NB) model [48], the newly developed Bayesian Poisson-Gamma (PG) model, and a binomial model under three different scenarios: a null expectation data set with no bias and no AI, $B = 0.5$ and $R = 1$; a null expectation data set with bias and no AI, $B \neq 0.5$ and $R = 1$; and a model with bias and AI, $B \neq 0.5$ and $R \neq 1$.

To assess the performance of the PG model relative to the NB model when the bias parameter is random, we incorporate simulated DNA control counts ($\phi = DNA$) into the PG model and consider ϕ, assuming the same model that we assume for p in the NB model (as in (1) below). To determine the impact of using a fixed versus a random bias parameter, we examined both the PG model with q and the PG model with ϕ. For the NB and PG

Figure 1 Sources of error in read alignments and allele-specific read counts contribute to bias in estimation of ASE and AI. Here we consider error originating from sequence similarity in the genome (e.g., repeats and duplications) and hidden variation (missed or false SNPs). The examples shown illustrate cases for alignments to a single reference **(A-C)** and to multiple references **(D-E)**. Alignments to augmented references are expected to behave similarly to alignments to multiple references. **A)** Masking SNPs located in regions with strong sequence similarity to other locations in the genome (genome sequence ambiguity) can result in alignment error, the best match in the masked reference may be located in a location other than the true source of the read. **B)** Algorithms that account for multiple mapping can result in allele bias when reads from one of the alleles are discarded or are mapped randomly, while reads from the other allele map to their true source location. **C)** For a single unmasked reference, reads from one of the alleles may not align at all, resulting in bias toward the other allele. **D)** When two references are used (one for each parental genome), differences between the references in genome sequence ambiguity can result in allele bias for the same reason as outlined in B. **E)** Sequencing errors in one reference can result in allele bias when reads from both (identical) alleles align best to the other reference.

models with a random bias parameter, the value is taken from replicate simulated DNA control allele-specific read counts. For comparison, we examine the performance of the PG model with $q = 1/2$. In practice, any single value estimate of bias can be used in the PG model with q, under a null expectation of no bias $q = 1/2$ is appropriate.

The type I error rate was examined for the null case where allele-specific read counts are generated with no bias ($B = 0.5$) and no AI ($R = 1$). Type I error is less than 5% in all cases, with the NB and PG models showing similar levels of type I error that are lower than that of the binomial, but all tests are valid in this case (Table 1).
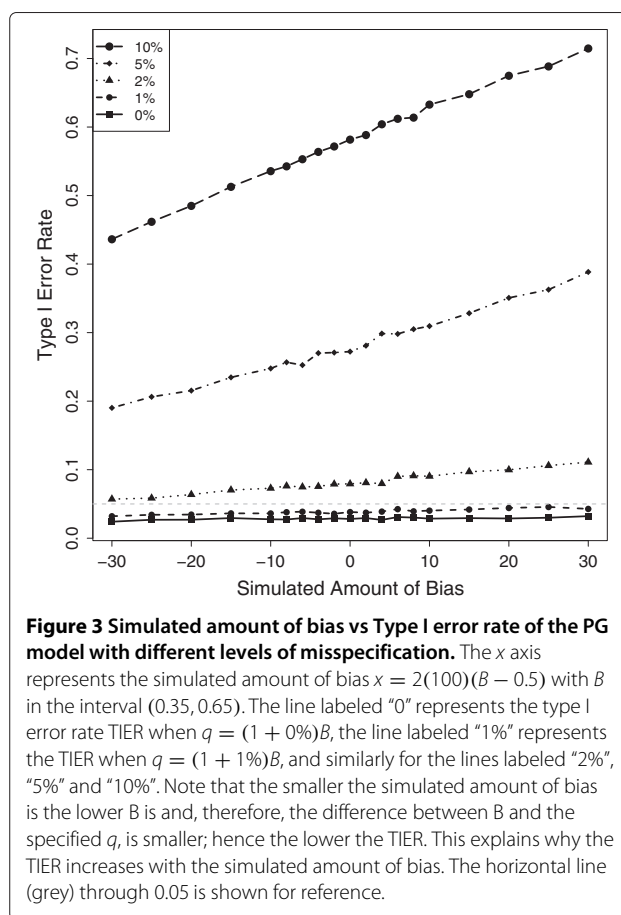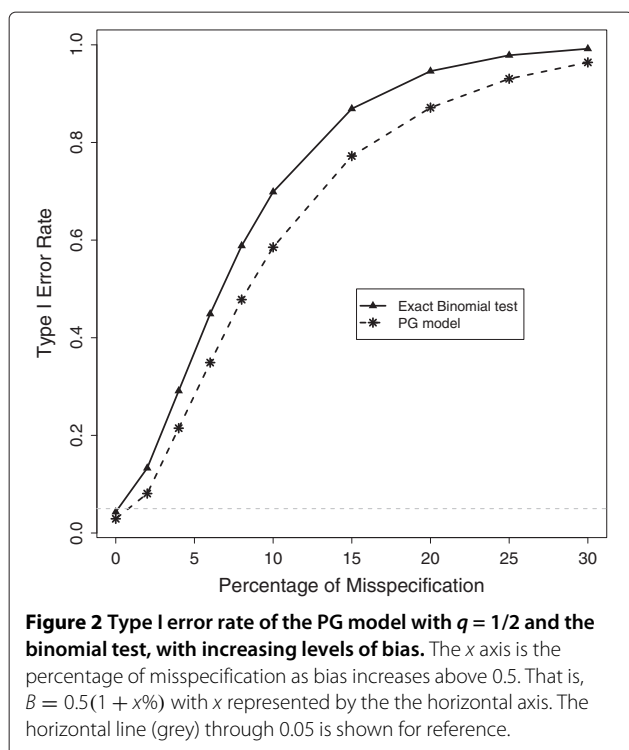
Table 1 Estimate of the type I error rate

Model	Type I error rate
Binomial	4.9%
NB: $p = DNA$	3.5%
PG: $\phi = DNA$	3.8%
PG: $q = 1/2$	3.2%

Allele-specific counts from DNA and RNA-seq data were simulated with no bias and no allelic imbalance for three replicates each of 10,000 exonic regions and analyzed using the binomial exact test, the random bias parameter NB and PG models that use DNA controls and the PG model with fixed bias parameter, q = 1/2. Even when there is no bias, the Bayesian models have better performance than a binomial exact test.

Both the PG model with a fixed effect of $q = 1/2$ and the binomial exact, assume that there is no bias. That is, the null expectation is equal amounts of reads from the paternal and maternal alleles. However, error variance is handled differently by the two approaches. Is the PG model with $q = 1/2$ different from the binomial test? We compare these models using simulated data sets of allele-specific read counts under a null scenario in which there is bias ($B \neq 0.5$) and no allelic imbalance ($R = 1$) and simulated data sets with both bias and allelic imbalance ($R \neq 1$). Comparing the type I error rates for data generated with increasing levels of bias shows that while in both cases the model assumptions are violated and type I error increases with increasing bias, **the PG model always has a lower type I error rate** (Figure 2).

To understand how bias affects model performance, we further investigated the behavior of the PG model with $q = 1/2$ and PG model with $q = B$, for $B = 0.5 \pm 10\%$ *error*. The model performs well when there is bias, while the binomial and $q = 1/2$ perform poorly when there is bias (Figure 2; Additional file 1: Figure S1). The model with $q = B$ controls the type I error rate (2.6%) even when there is bias in the allele-specific read counts. When bias is accounted for but misspecified, the type I error rate depends on the amount of misspecification (Figure 2; Figure 3). Interestingly, when the amount of bias is large misspecification of small amounts (5%) can result in large type I errors. As expected, this appears as slightly

Figure 3 Simulated amount of bias vs Type I error rate of the PG model with different levels of misspecification. The *x* axis represents the simulated amount of bias $x = 2(100)(B - 0.5)$ with B in the interval $(0.35, 0.65)$. The line labeled "0" represents the type I error rate TIER when $q = (1 + 0\%)B$, the line labeled "1%" represents the TIER when $q = (1 + 1\%)B$, and similarly for the lines labeled "2%", "5%" and "10%". Note that the smaller the simulated amount of bias is the lower B is and, therefore, the difference between B and the specified *q*, is smaller; hence the lower the TIER. This explains why the TIER increases with the simulated amount of bias. The horizontal line (grey) through 0.05 is shown for reference.

asymmetric with respect to the binomial. This is simply due to 1% of 0.65 being a larger absolute amount of bias than 1% of 0.35 (Figure 3). When bias is large and unaccounted for, the PG model with $q = 1/2$ and the binomial can have dramatic type I error rates (Figure 2).

Using DNA controls or simulated reads to measure bias

What causes bias in estimates of AI? Genome sequence ambiguity and mapping ambiguity can lead to bias, often collectively referred to as map bias. Graze et al. 2012 [48] and Satya et al. 2012 [51] found that the use of separate reference sequences for each allele or augmented single references that contain both alleles reduces map bias. Satya et al. 2012 [51] also found that ambiguity in the reference genome is associated with bias and showed that simple masking of biased SNPs is not sufficient to reduce systematic error in studies of allelic imbalance. Stevenson et al. 2013 [50] observed that changing mapping parameters and filtering can reduce the impact of map bias on estimates of AI when mapping to a single reference.

Simulation studies in which equal numbers of simulated reads from each allele are created and counted, after

Figure 2 Type I error rate of the PG model with $q = 1/2$ and the binomial test, with increasing levels of bias. The *x* axis is the percentage of misspecification as bias increases above 0.5. That is, $B = 0.5(1 + x\%)$ with *x* represented by the the horizontal axis. The horizontal line (grey) through 0.05 is shown for reference.

mapping to each reference, capture genome sequence ambiguity and bias in the mapping algorithm. To incorporate this measure into the PG model, we simulated reads from both a maternal (*D. simulans*) and paternal (*D. melanogaster*) reference. This creates a simulated set of reads that are analogous to sequencing the F1 heterozygote. Mapping these back to both references we counted allele-specific reads corresponding to each allele, using the proportion of allele-specific reads corresponding to the paternal reference as a measure of map bias. Bias was detected approximately 32% of the time in the interspecific read simulation study.

Intraspecific studies are expected to have smaller differences, but are still expected to have regions of genome ambiguity and map bias. Using the DGRP as inspiration [55]. We simulated 94 different F1 genotypes formed by crossing each line to a common tester. Approximately 18% of the time, gene regions were always biased, implicating shared regions of ambiguity among genotypes. In 3% of cases bias was specific to the genotype constructed, implicating a combination of ambiguity and SNP variation among lines. This supports the conclusions of previous studies that bias is likely present in intraspecific studies of AI.

To compare statistical modeling of bias and filtering strategies we examined the behavior of the models using real RNA-seq allele-specific read counts from an interspecific F1 genotype (see Methods for details). We compare a modeling approach with bias measured as the frequency of the paternal allele using allele-specific read counts from DNA sequencing of the same F1 interspecific genotype (PG model with $q = DNAcontrols$) with a model that uses an interspecific simulation study to measure map bias (PG model with $q = simulation$). Additionally, allele

assignment error and genome ambiguity based filtering strategies are investigated.

Read simulation and alignment generally produce smaller estimates of bias than the DNA controls. Often the simulation results in estimates of a half even when the DNA indicates bias. However, using q from the simulation study ($q = simulation$) does identify a portion of the bias and only rarely does this measure estimate a larger amount of bias then the DNA. This indicates that the PG model with $q = simulation$ should perform better than the PG model with $q = 1/2$ (Figure 4). Using the DNA controls as "truth", the proportion of false positives is notably smaller and the number of false positives and false negatives are more balanced in the PG model with $q = simulation$, relative to the PG model with $q = 1/2$ (Table 2). The specificity using $q = simulation$ is larger (0.74) than when using $q = 1/2$ (0.41), but the sensitivity using $q = simulation$ is smaller (0.69) than when $q = 1/2$ (0.81). Among biased exonic regions there is an exorbitant false positive rate. The false positive rate (FP) is equal to 0.59 ($q = 1/2$), similar to what we observed in analysis of simulated RNA-seq data sets. This is substantially better than binomial (FP = 0.69), but still indicates considerable unaccounted for bias. In comparison, the false positive rate is 0.26 when $q = simulation$ indicating that using simulated reads to estimate map bias and incorporating this measure into the statistical model dramatically reduces the false positive rate.

For those exons where the interspecific F1 simulated read counts do not capture the bias indicated by the DNA controls, we examined other possible sources of the bias. Using a new mapping tool BWA-MEM [56] and the variant caller FreeBayes [57], we identified variants not identified in the initial study [48]. Of the exons where $|q - 1/2| \leq$
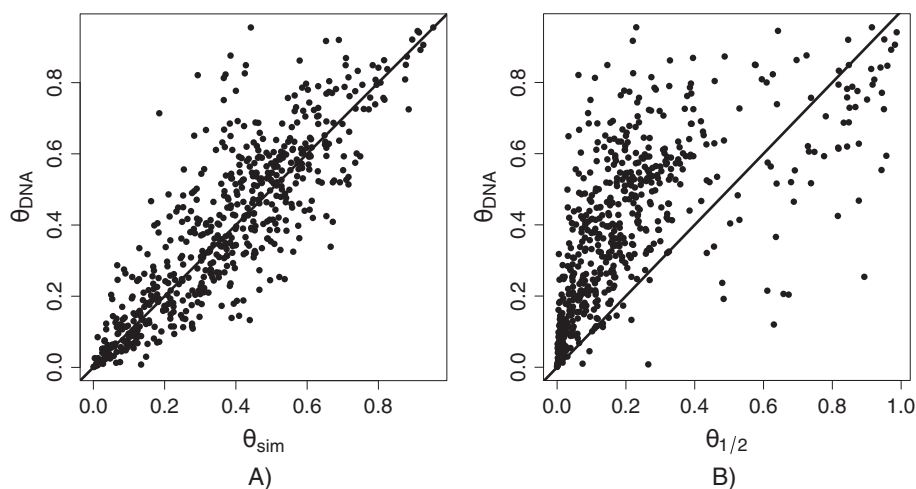

Figure 4 Comparison of the estimated θ. The proportion of reads coming from the paternal allele when $q = DNAcontrols$, as compared to $q = simulation$ **(A)** and $q = 1/2$ **(B)**. The $n = 617$ exons with simulated $q \neq 0, 1$ and $|q - 1/2| > 0.2$ are shown.

Table 2 Model comparisons

		PG $q = DNAcontrols$	
		AB	AI
PG	AB	0.31	0.18
$q = simulation$	AI	0.11	0.40
	(a)		
PG	AB	0.17	0.11
$q = 1/2$	AI	0.25	0.47
	(b)		
Binomial	AB	0.13	0.10
	AI	0.29	0.48
	(c)		

Comparison of the PG model with q = DNA controls (columns) to (a) the PG model with q = simulation, (b) the PG model with q = 1/2, and (c) the binomial test. The n = 2,230 exons with simulated $q \neq 0, 1$ and |q−1/2| >0.05 are considered.

0.05 for $q = simulation$, there are $n = 3,923$ where $q = DNAcontrols$ indicates there is no AI but the model based on $q = simulation$ finds AI. Approximately 38% of these have evidence for previously undetected polymorphism while only 27% of the $n = 14,338$ where both models indicate no AI is present show evidence for previously undetected polymorphism. This shows that unidentified variants are a source of bias captured in DNA controls, but not in read simulations.

A related source of bias that has been identified as contributing to systematic error in estimation of AI is allele assignment error. False positives and false negatives in SNP calls can result in reads either not mapping at all or to the wrong location, regardless of whether a single or multiple reference approach is used. To identify allele assignment error, RNA derived reads from each parent were aligned onto both the matching and non-matching reference. An exon was considered to show allele assignment error when reads originating from one parent were assigned (based on higher quality alignments to the wrong reference) to the other parent more than 5% of the time. There were 26,896 exons for which both bias as measured by DNA controls and allele assignment could be assessed. There is a positive association between allele assignment error and bias as measured by read simulations and these regions account for a portion of the bias (22%) in DNA controls. However, a large number of exons (2,129 and 5,376 respectively) show allele assignment error, but do not show bias in DNA controls or in simulated interspecific F1 read alignments.

Regions for which parental reads show error in allele assignment may be filtered from the results, as this is expected to contribute to bias. However filtering exons which show allele assignment error reduces the amount of data considered in the analysis, but does not appreciably reduce the false positive rate. Similarly, Satya et al. 2012 [51] noted that regions where alignment simulations found strong bias generally showed genome ambiguity. We assessed genome ambiguity by sequence identity and read mapping. Comparing genome ambiguity with simulated bias we found that in nearly all cases where bias is detected, genome ambiguity is also detected. However, the reverse is not true. There are many regions of the genome which show sequence similarity that do not show strong bias. For approaches which due not account for bias in the statistical model, a filtering strategy based on both ambiguity and allele assignment error has a small effect on the resulting percentage of false positives. However, this eliminates almost a quarter of all the data available. While simulations will not necessarily uncover these biases, filtering based upon them does not improve the overall inferences. Thus, filtering does not seem to be an effective strategy to control type I error rate.

Conclusions

Even for cases where no control is available, the PG model with $q = 1/2$ is preferable to a binomial test. The PG model had a consistently lower false positive rate than the binomial test. Considering extreme values of AI (greater than 95% of reads from the maternal/paternal allele) the PG model with $q = DNAcontrol$ is the least likely to reject, followed by the PG model with $q = 1/2$ or $q = simulation$. The binomial model always rejects in these cases. The PG model with $q = 1/2$ is more conservative even though bias is not corrected or filtered. This is expected if there is extra variance that is not accounted for when using the binomial. This extra variance has been discussed [43,58] and may well be due to reads being random draws from a distribution rather than the fixed number of trials the binomial assumes.

Accounting for bias by using simulated alignments is a better alternative than using a model in which no bias is assumed. The PG model with q estimated from simulated read alignments performs better than using $q = 1/2$, reducing the false positive rate by more than 50%. Simulation captures genome ambiguity and map bias, as well as allele assignment error. When filtering strategies are coincident with bias identified in simulated alignments, they can lower the false positive rate by removing those regions likely to be affected. However, these strategies also remove regions from consideration that do not show allele bias in either simulations or in DNA. Filtering does not provide an advantage over incorporating bias directly into the model and instead removes regions form consideration that can be evaluated using an appropriate model.

Unfortunately, there are additional sources of bias not captured by simulations or filtering strategies. The result

is large type I errors. A large source of this bias is likely to be variants that were not initially detected. While this is likely to decrease as variant callers improve, it is worth cautioning that even small amounts of unaccounted for bias result in steep increases in type I error rates.

The flexible Bayesian model proposed here allows for use of DNA controls when they are present. It also has the ability to use a fixed or random parameter for the estimate of bias. If desired, the confidence intervals around θ could be widened by allowing for variation in $q = simulation$ using an external estimate of variability. While we have explored the use of an estimate of bias from read simulations, the model is flexible with respect to other approaches. For example, in the absence of DNA controls or simulations an empirical Bayes approach with a sliding value of bias could be used and the robustness of AI estimates explored across a range of likely values of bias. Alternatively, in cases where there is known bias toward one reference or the other, a single best guess value of the bias could be used, similar in spirit to the skewed binomial test [59]. The model is general enough to accommodate many subtle differences depending on the particular experimental design and approach to estimating bias.

As sequencing costs continue to plummet, population studies of *cis* regulation are on the horizon. Large population studies mean rethinking approaches to evaluating AI, as DNA controls are no longer a viable prospect. Simulations can be effective, but hidden variation can cause significant bias in estimation of AI. Along with improving modeling capabilities, it will be necessary to improve variant callers and to spend more time and effort on large scale population genomic assemblies.

Methods

Simulated reads and alignments

Intraspecific read simulations: We simulated 95 *D. melanogaster* genotypes by randomly incorporating 160,000 SNPs into the exonic regions in a single reference sequence (FlyBase 5.51). All possible one hundred base pair reads were simulated from each genotype using a sliding window approach. To create an intraspecific cross using a reference design, one genotype was selected as a reference (Tester), simulated reads from the Tester strain were mixed with each of the remaining 94 genotypes (Lines). The mixed sets of reads were independently aligned to the exon regions in the Tester or Line references using bowtie [-k1 -m1] [60] and LAST [-l 20] [61]. Alignments were compared to determine which reads aligned better to the Tester or Line references and which reads aligned equally well to both the Tester and Line references. Bias towards the line was calculated by taking the number of Line specific reads divided by the total number of allele-specific reads (Tester + Line).

Interspecific read simulations: Starting from the set of strain specific reference exonic sequences [48], all possible 36 bp reads were created from each exon region for each parental strain of the F1 interspecific cross. The number of reads created for each exon region is (L - 36) + 1, where L is the length of the exon region. Exon regions shorter than 36 bp (in either reference) were excluded. Reads were aligned to all exonic sequences in the reference genomes using bowtie [-k1 -m1] and LAST [-l 20]. Reads were separated into three categories based on the highest quality alignment. Reads mapping ambiguously to both references were excluded. If a read mapped with equal quality (and uniquely) to both the maternal (*D. melanogaster*) and paternal (*D. simulans*) references they were assigned to the 'both' category. Reads were assigned to the 'maternal' category if they aligned best (and uniquely) to the maternal reference. Reads corresponding to the paternal allele were similarly assigned to the 'paternal' category. The value used for $q = simulation$ in the PG model is the proportion of allele-specific reads corresponding to the paternal allele.

RNA-seq data set

To measure allelic imbalance, reads from two alleles were quantified to estimate allele-specific expression from RNA-seq data for 3 independent replicate samples of RNA from an interspecific F1 hybrid ([48], GEO accession number GSE34591). Briefly, reads were aligned to species references (denoted as maternal and paternal) that were specific to the genotypes used in the experiment. Each reference contains the exonic sequences for one of the parents of the F1 genotype. For each exon, reads contributed to the allele-specific count for each allele (maternal or paternal) when they aligned better to the corresponding reference.

DNA-seq data set

To control for bias introduced by alignment error or by other technical sources, genomic DNA from the same F1 genotype was collected ([48], SRA accession number SRA048616). Allele-specific read counts for each exon were quantified for DNA as for the RNA data. Inferences from the DNA controls and RNA-seq data are used as the basis by which to compare other models and approaches.

Identifying ambiguity in the genome

Genome ambiguity was assessed using both sequence identity and read mapping. For FlyBase 5.26, there were 726 exonic regions with an identical sequence to at least one other exonic region. These 726 regions could be grouped into 224 sets of identical sequences. A unique reference was created with all unique exonic regions and only a single representative from each of the 224 identical sets.

Next we identified regions where alignment algorithms would have difficulty uniquely placing reads. All possible 36-mers were simulated from the unique reference using a sliding window approach. Some exonic regions (1,261) could not be simulated because they were less than 36 bp. Simulated reads were aligned back to the reference uniquely using bowtie [-k1 -m1]. Ambiguous reads were then re-aligned using bowtie [-k1 -a] placing an ambiguous read at all locations that it mapped. In comparisons of filtering strategies, an exonic region was considered ambiguous (6,229) if there was at least 1 ambiguous read aligning. Ambiguous regions were also identified using BLAST as the alignment algorithm or with the genome mappability analyzer (GMA) [62]. Bowtie, BLAST, and GMA all gave similar results. This process was applied to both species specific references.

Identifying allele assignment error

Parental RNA from both maternal (*D. melanogaster*, n = 6) and paternal (*D. simulans*, n = 3) lines ([63], GEO accession number GSE54069) were aligned to updated genotype-specific references [48]. Reads were aligned using a multiple step alignment process. First reads were aligned uniquely using bowtie [-k1 −m1]. Unaligned and ambiguous reads were then quality trimmed removing low quality ends. Trimmed reads were then aligned uniquely a second time using bowtie [-k1 −m1]. Finally unaligned and ambiguous reads were aligned uniquely using LAST [-l 20]. For each sample, reads were assigned to a genotype based upon the highest quality alignment. Reads mapping equally well to both genotype-specific references were assigned to a "both" category. In comparisons of filtering strategies, an exonic region was considered to show evidence of allele assignment error when greater than 5% of the reads aligned better to the wrong reference.

Model 1- binomial test

Let θ be the unknown proportion of reads from the paternal allele and let n be the total number of reads aligning to the exon. This is the standard binomial test of the null hypothesis of no allelic balance $H_0 : \theta = 1/2$ vs the alternative of allelic imbalance $H_1 : \theta \neq 1/2$. Here we reject the null hypothesis if $|z| > 1.965$ where $z = (\hat{\theta} - 1/2)/\sqrt{(1/2)(1 - 1/2)/n}$, and $\hat{\theta}$ is the observed proportion of paternal reads. The advantages of this test are that it is easy to implement and that there are statistical techniques that control for the fact that we are testing multiple tests at the same time. For example, the false discovery rate criterion of Benjamini and Hochberg [64]. Nevertheless, the binomial test does not control for systematic error and bio/tec variation. The standard practice is to compute a measure of bias based on simulated reads and alignments, similarly to $q = simulation$, and remove biased regions from the analysis.

Model 2- negative binomial- with DNA controls, p = bias in DNA

Systematic error in inference of allelic imbalance can arise from asymmetry in genetic differences between reads and references used in alignments or differences between references in ambiguity (e.g., CNVs), in combination with the specific alignment algorithm used [35,51]. Technical sources of systematic error arising from library construction and sequencing may also contribute [52]. Graze et al. 2012 [48] integrated information from a DNA control into the prior for the model used to estimate allelic imbalance in RNA in a Bayesian approach to inference of allelic imbalance. This approach adjusts the estimates of allelic imbalance based on the null expectation for the relative abundance of the two alleles estimated from the DNA controls, p. The model that estimates bias using DNA controls estimates the p hyper-parameter used in the model that estimates allelic imbalance in RNA-seq data. In the Negative Binomial model the number of reads is random, rather than fixed. For the RNA model: θ is the parameter for the proportion of reads coming from the paternal (*D. simulans*) allele, y_i and x_i is the number of RNA reads assigned to the paternal and maternal (*D. melanogaster*) references, respectively, for the replicate i. Similarly, for the DNA model: $y_{i^\star}^\star$ is the number of paternal assigned reads and $x_{i^\star}^\star$ is the number of maternal assigned reads. I ($i = 1, 2, \ldots, I$) and $I^\star$ ($i^\star = 1, 2, \ldots, I^\star$) are the number of replicate RNA and DNA sample respectively. The RNA model is,

$$x_i \mid y_i, \theta \sim \text{Negative Binomial}(y_i, \theta) \text{ for } i = 1, \ldots, I;$$
$$\theta \mid p \sim \text{Beta}((1 - p)t, pt);$$

and the DNA model is,

$$x_{i^\star}^\star \mid y_{i^\star}^\star, p \sim \text{Negative Binomial}(y_{i^\star}^\star, p) \text{ for } i^\star = 1, \ldots, I^\star;$$
$$p \sim \text{Beta}(v, v).$$

$$(1)$$

Here the parameterization of the Negative Binomial distribution is such that, if $\eta \sim \text{NegativeBinomial}(k, \epsilon)$, then $\eta \in \{0, 1, \ldots\}$ denotes the number of failures before the first k successes with probability of success equal to ϵ.

Model 3- poisson gamma

The PG model can be used when DNA control is not available by using q (fixed). Unlike the NB model which incorporates p as a prior, the PG model incorporates the parameter q (fixed) into the structure of the model. The model can also be specified with ϕ (random) if replicate measures of bias are available as for DNA controls. Let x_i and y_i be the maternal and paternal, respectively, RNA reads in the biological replicate i, $i = 1, \ldots, I$. We assume,

$$y_i \mid \mu, \alpha, \beta_i, q \sim \text{Poisson}(\mu \alpha \beta_i q) \quad \text{and}$$
$$x_i \mid \mu, \beta_i, q \sim \text{Poisson}(\mu \beta_i (1 - q)).$$

$$(2)$$

Here μ is the overall mean, a nuisance parameter. The parameter β_i, $i = 1, \ldots, I$ models the biological replicate variation. q is a constant that incorporates the information about the bias into the model. q in the PG model plays a role similar to that of p in the NB model. When we perform the simulation we make q equal to the proportion of simulated reads aligning to the paternal reference.

When we have DNA information we make q equal to the proportion of DNA sequencing reads from an F1 genotype aligned to the paternal reference. If DNA information is present, then a random bias parameter ϕ can be sampled from the posterior of the DNA model and used as a true value in the RNA model, and θ can then be sampled from the posterior, under the RNA model. For example, we sample $\phi^m = p$ from the posterior of model (1) and obtain a posterior sample of size 1, $(\alpha^m, \mu^m, \beta_1^m, \ldots, \beta_I^m) \mid q = \phi^m$, under model in (2) for every $m = 1, \ldots, M$. We flag the exon as in AI if the CI for α does not contain 1 (or, equivalently, the CI for θ does not contain 1/2). When comparing the PG model and the NB model we follow this approach, using p (random) and ϕ (random). For fair comparison, we contrast the PG model with $q = simulation$ (fixed) and with $q = DNAcontrols$ (fixed). The parameter of interest is the treatment effect, α. If θ is the "real proportion of reads from the paternal allele",

$$\theta = \frac{\mu\alpha\beta_i}{\mu\beta_i + \mu\alpha\beta_i} = \frac{\alpha}{1 + \alpha}.$$

So when there is no AI, $\alpha = 1$ and

$$E\left(\frac{y_i}{x_i + y_i}\right) = E\left(\frac{E(y_i \mid x_i + y_i)}{x_i + y_i}\right) = q,$$

the parameters q, θ, x and y in the Poisson Gamma model play the role of p, θ, x and y in the negative binomial model. We give standard priors to the parameters: $\mu \sim Gamma(a_\mu = 1/2, b_\mu = 1/2)$, $\beta_1, \ldots, \beta_I \sim Gamma(1/2, 1/2)$ and $\alpha \sim Gamma(1/2, 1/2)$. Here $\eta \sim Gamma(a, b)$ is parameterized such that $E(\eta) = a/b$.

Note that the model is parameterized to estimate the abundance of one of the alleles (the paternal), rather than the relationship between two alleles. If the bias is in the opposite direction relative to the the paternal allele, then the Type I error rate will be lower than if the bias is in the direction of the paternal allele. If it is likely that a global bias in one direction exists — perhaps due to a difference in reference quality — and the type I error is a greater concern than power, the model should be parameterized such that the allele estimated is the allele that is not favored by bias.

Availability of supporting data

The data sets supporting the results of this article are available in the following repositories: The F1 interspecific hybrid RNA sequencing data and corresponding parental RNA sequencing data are available from the Gene Expression Omnibus database (GEO) repository with accession numbers GSE34591 and GSE54069, respectively. The F1 interspecific hybrid DNA sequencing data are available from the NCBI Short Read Archive (SRA), accession number SRA048616.

Additional file

Additional file 1: Allele-specific read count simulation study.

Author details
[1]Department of Mathematics, University of Louisiana at *Lafayette*, 70503 Lafayette, LA, USA. [2]Department of Molecular Genetics and Microbiology, University of Florida, 32611 Gainesville, FL, USA. [3]Department of Biological Sciences, Auburn University, 101 Rouse Life Science Building, 36849 Auburn, AL, USA.

References
1. Conne B, Stutz A, Vassalli JD: **The 3′ untranslated region of messenger RNA: A molecular 'hotspot' for pathology?** *Nat Med* 2000, **6**(6):637–641.
2. Mendell JT, Dietz HC: **When the message goes awry: disease-producing mutations that influence mRNA content and performance.** *Cell* 2001, **107**(4):411–414.
3. Hollams EM, Giles KM, Thomson AM, Leedman PJ: **MRNA stability and the control of gene expression: implications for human disease.** *Neurochem Res* 2002, **27**(10):957–980.
4. Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease.** *Genes Dev* 2003, **17**(4):419–437.
5. Buckland PR: **The importance and identification of regulatory polymorphisms and their mechanisms of action.** *Biochim Biophys Acta* 2006, **1762**(1):17–28.
6. Chen J-M, Férec C, Cooper DN: **A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes I: general principles and overview.** *Hum Genet* 2006, **120**(1):1–21.
7. Johnson AD, Wang D, Sadee W: **Polymorphisms affecting gene regulation and mRNA processing: broad implications for pharmacogenetics.** *Pharmacol Ther* 2005, **106**(1):19–38.
8. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I,

Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, et al.: **Genetics of gene expression and its effect on disease.** *Nature* 2008, **452**(7186):423–428.

9. Lai Z, Gross BL, Zou YI, Andrews J, Rieseberg LH: **Microarray analysis reveals differential gene expression in hybrid sunflower species.** *Mol Ecol* 2006, **15**(5):1213–1227.

10. Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB: **The evolution of gene regulation underlies a morphological difference between two Drosophila sister species.** *Cell* 2008, **132**(5):783–793.

11. Martin-Coello J, Dopazo H, Arbiza L, Ausió J, Roldan ER, Gomendio M: **Sexual selection drives weak positive selection in protamine genes and high promoter divergence, enhancing sperm competitiveness.** *Proc R Soc Biol Sci* 2009, **276**(1666):2427–2436.

12. Wittkopp PJ, Stewart EE, Arnold LL, Neidert AH, Haerum BK, Thompson EM, Akhras S, Smith-Winberry G, Shefner L: **Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila.** *Science* 2009, **326**(5952):540–544.

13. Barbash DA, Siino DF, Tarone AM, Roote J: **A rapidly evolving MYB-related protein causes species isolation in Drosophila.** *Proc Nat Acad Sci USA* 2003, **100**(9):5302–5307.

14. Michalak P, Noor MAF: **Association of misexpression with sterility in hybrids of Drosophila simulansand D. mauritiana.** *J Mol Evol* 2004, **59**(2):277–282.

15. Sun S, Ting CT, Wu CI: **The normal function of a speciation gene, Odysseus, and its hybrid sterility effect.** *Science* 2004, **305**(5680):81–83.

16. Haerty W, Singh RS: **Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of Drosophila.** *Mol Biol Evol* 2006, **23**(9):1707–1714.

17. Michalak P, Malone JH, Lee IT, Hoshino D, Ma D: **Gene expression polymorphism in Drosophila populations.** *Mol Ecol* 2007, **16**(6):1179–1189.

18. Shirangi TR, Dufour HD, Williams TM, Carroll SB: **Rapid evolution of sex pheromone-producing enzyme expression in Drosophila.** *PLoS Biol* 2009, **7**(8):e1000168.

19. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast.** *Science* 2002, **296**(5568):752–755.

20. Yan H, Yuan W, Velculescu VE, Vogelstein B: **Kinzler KW: Allelic variation in human gene expression.** *Science* 2002, **297**(5584):1143.

21. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP: **Allelic variation in gene expression is common in the human genome.** *Genome Res* 2003, **13**(8):1855–1862.

22. Wittkopp PJ, Haerum BK, Clark AG: **Evolutionary changes in cis and trans gene regulation.** *Nature* 2004, **430**(6995):85–88.

23. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR: **Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid.** *Genetics* 2005, **169**(4):2295–2303.

24. Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L: **Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays.** *Genome Res* 2005, **15**(2):284–291.

25. Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, Cáceres CE, Paige KN: **Segregating variation in the transcriptome: cis regulation and additivity of effects.** *Genetics* 2006, **173**(3):1347–1355.12.

26. Genissel A, McIntyre LM, Wayne ML, Nuzhdin SV: **Cis and trans regulatory effects contribute to natural variation in transcriptome of Drosophila melanogaster.** *Mol Biol Evol* 2008, **25**(1):101–110.

27. Guo M, Yang S, Rupe M, Hu B, Bickel DR, Arthur L, Smith O: **Genome-wide allele-specific expression analysis using massively parallel signature sequencing (MPSSâĎć) reveals cis-and trans-effects on gene expression in maize.** *Plant Mol Ecol* 2008, **66**(5):551–563.

28. Lemos B, Araripe LO, Fontanillas P, Hartl DL: **Dominance and the evolutionary accumulation of cis-and trans-effects on gene expression.** *Proc Nat Acad Sci* 2008, **105**(38):14471–14476.

29. Graze RM, McIntyre LM, Main BJ, Wayne ML, Nuzhdin SV: **Regulatory divergence in Drosophila melanogaster and D. simulans, a genomewide analysis of allele-specific expression.** *Genetics* 2009, **183**(2):547–61121.

30. Tirosh I, Reikhav S, Levy AA, Barkai N: **A yeast hybrid provides insight into the evolution of gene expression regulation.** *Science* 2009, **324**(5927):659–662.

31. Zhang X, Borevitz JO: **Global analysis of allele-specific expression in Arabidopsis thaliana.** *Genetics* 2009, **182**(4):943–954.

32. McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ: **Regulatory divergence in Drosophila revealed by mRNA-seq.** *Genome Res* 2010, **20**(6):816–825.

33. Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee J-H, Aach J, Leproust EM, Eggan K, Church GM: **Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human.** *Nat Methods* 2009, **6**(8):613–618.

34. Yang Y, Graze RM, Walts BM, Lopez CM, Baker HV, Wayne ML, Nuzhdin SV, McIntyre LM: **Partitioning transcript variation in Drosophila: abundance, isoforms, and alleles.** *G3 (Bethesda)* 2011, **1**(6):427–436.

35. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nikadori E, Gilad Y, Pritchard JK: **Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.** *J Bioinformatics* 2009, **25**(24):3207–3212.

36. Main BJ, Bickel RD, McIntyre LM, Graze RM, Calabrese PP, Nuzhdin SV: **Allele-specific expression assays using Solexa.** *BMC Genomics* 2009, **10**(1):422.

37. Emerson JJ, Hsieh LH, Sung HM, Wang TY, Huang CJ, Lu HH-S, Lu M-YJ, Wu S-H, Li W-H: **Natural selection on cis and trans regulation in yeasts.** *Genome Res* 2010, **20**(6):826–836.

38. Fontanillas P, Landry CR, Wittkopp PJ, Russ C, Gruber JD, Nusbaum C, Hartl DL: **Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing.** *Mol Ecol* 2010, **19**(212–227.

39. Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C: **High-resolution analysis of parent-of-origin allelic expression in the mouse brain.** *Science* 2010, **329**(5992):643–648.

40. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjorson 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**(1):522.

41. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics* 2008, **9**(2):321–332.

42. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**(10):106.

43. Auer PL, Doerge RW: **A two-stage Poisson model for testing RNA-seq data.** *Stat Appl Genet Mol Biol* 2011, **10**(1):1–26.

44. Langmead B, Hansen KD, Leek JT: **Cloud-scale RNA-sequencing differential expression analysis with Myrna.** *Genome Biol* 2010, **11**(8):R83.

45. Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results.** *Genome Biol* 2010, **11**(12):220.

46. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM: **A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data.** *Genome Res* 2011, **21**(10):1728–1737.

47. Turro E, Su SY, Gonçalves Â, Coin LJ, Richardson S, Lewin A: **Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads.** *Genome Res* 2011, **12**(2):13.

48. Graze RM, Novelo LL, Amin V, Fear JM, Casella G, Nuzhdin SV, McIntyre LM: **Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution.** *Mol Biol Evol* 2012, **29**(6):1521–1532.

49. DeVeale B, Kooy DVD, Babak T: **Critical evaluation of imprinted gene expression by RNAâĂŞSeq: a new perspective.** *PLoS Genet* 2012, **8**(3):e1002600.

50. Stevenson KR, Coolon JD, Wittkopp PJ: **Sources of bias in measures of allele-specific expression derived from rna-seq data aligned to a single reference genome.** *BMC Genomics* 2013, **14**(1):536.

51. Satya RV, Zavaljevski N, Reifman J: **A new strategy to reduce allelic bias in RNA-Seq readmapping.** *Nucleic Acids Res* 2012, **40**:e127.

52. Heap GA, Yang JH, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, Albert TJ, Rodesch MJ, Clayton DG, Todd JA, van Heel DA, Plagnol V: **Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing.** *Hum Mol Genet* 2010, **19**(1):122–134.

53. Nothnagel M, Wolf A, Herrmann A, Szafranski K, Vater I, Brosch M, Huse K, Siebert R, Platzer M, Hampe J, Krawczak M: **Statistical inference of allelic imbalance from transcriptome data.** *Hum Mutat* 2011, **32**(1):98–106.
54. Pandey RV, Franssen SU, 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**(4):740–745.
55. Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, Magwire MM, Blankenburg K, Carbone MA, Chang K, Ellis LL, Fernandez S, Han Y, Highnam G, Hjelmen CE, Jack JR, Javaid M, Jayaseelan J, Kalra D, Lee S, Lewis L, Munidasa M, Ongeri F, Patel S, Perales L, Perez A, et al.: **Natural variation in genome architecture among 205 drosophila melanogaster genetic reference panel lines.** *Genome Res* 2014, **24**:1193–1208.
56. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** 2013, **1303.3997** *arXiv Prepr. arXiv1303.3997.*
57. Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** 2012, **1207.3907** *arXiv Prepr. arXiv1207.3907.*
58. Law CW, Chen Y, Shi W: **Smyth GK: Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol* 2014, **15**:29.
59. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464**(7289):773–777.
60. Langmead B, Trapnell C, Pop M: **Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):25.
61. Frith MC, Wan R, Horton P: **Incorporating sequence quality data into alignment improves DNA read mapping.** *Nucleic Acids Res* 2010, **38**(7):e100–e100.
62. Lee H, Schatz MC: **Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score.** *Bioinformatics* 2012, **28**(16):2097–2105.
63. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ: **Nuzhdin SV: RNA-seq : technical variability and sampling.** *BMC Genomics* 2011, **12**(1):293.
64. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57**(1):289–300.