BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 18 2014, pages 2603–2610
doi:10.1093/bioinformatics/btu342

Gene expression

Advance Access publication May 19, 2014

Estimates of allele-specific expression in *Drosophila* with a single genome sequence and RNA-seq data

Andrew Quinn, Punita Juneja and Francis M. Jiggins*

Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: Genetic variation in *cis*-regulatory elements is an important cause of variation in gene expression. *Cis*-regulatory variation can be detected by using high-throughput RNA sequencing (RNA-seq) to identify differences in the expression of the two alleles of a gene. This requires that reads from the two alleles are equally likely to map to a reference genome(s), and that single-nucleotide polymorphisms (SNPs) are accurately called, so that reads derived from the different alleles can be identified. Both of these prerequisites can be achieved by sequencing the genomes of the parents of the individual being studied, but this is often prohibitively costly.

Results: In *Drosophila*, we demonstrate that biases during read mapping can be avoided by mapping reads to two alternative genomes that incorporate SNPs called from the RNA-seq data. The SNPs can be reliably called from the RNA-seq data itself, provided any variants not found in high-quality SNP databases are filtered out. Finally, we suggest a way of measuring allele-specific expression (ASE) by crossing the line of interest to a reference line with a high-quality genome sequence. Combined with our bioinformatic methods, this approach minimizes mapping biases, allows poor-quality data to be identified and removed and aides in the biological interpretation of the data as the parent of origin of each allele is known. In conclusion, our results suggest that accurate estimates of ASE do not require the parental genomes of the individual being studied to be sequenced.

Availability and implementation: Scripts used to perform our analysis are available at https://github.com/d-quinn/bio_quinn2013.

Contact: fmj1001@cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 4, 2013; revised on May 6, 2014; accepted on May 12, 2014

1 INTRODUCTION

Recently, high-throughput RNA sequencing (RNA-seq), in which a library of millions of short cDNA fragments are sequenced in parallel, has emerged as the preferred method for genome-wide studies of gene expression (Stevenson *et al*., 2013; Wang *et al*., 2009). Gene expression polymorphism is the result of differences in either *cis*- or *trans*-regulatory elements. As *trans*-regulatory changes affect the expression of both alleles equally, the role of *cis*-regulatory variation can be determined by examining differences in expression between alleles, termed allele-specific expression (ASE) (Babak *et al*., 2010; Fraser *et al*., 2011; Wang *et al*., 2009).

Detection of ASE from RNA-seq data involves mapping sequence reads to their region of origin and assigning them to separate alleles. Both of these steps are non-trivial and in need of further development. Reads can be mapped to a single reference genome; however, this method is inherently biased (Satya *et al*., 2012). Reads representing reference alleles are more likely to map correctly than those representing non-reference alleles because they contain fewer mismatches, yielding estimates of ASE that favor the reference (Degner *et al*., 2009; Satya *et al*., 2012; Stevenson *et al*., 2013). Degner *et al*. (2009) illustrated this problem by generating a simulated human RNA-seq dataset that contained an equal number of reference and non-reference reads. The authors found that reads carrying the single-nucleotide polymorphism (SNP) allele found in the reference genome were significantly more likely to be mapped. Increasing the error rate in the sequence reads increases the bias by introducing additional mismatches to the reference genome (Degner *et al*., 2009). In both humans and *Drosophila*, the degree of bias is unequal across genes (Degner *et al*., 2009; Stevenson *et al*., 2013), with loci containing clusters of SNPs showing a strong bias toward the reference sequence. Therefore, even if the average bias across all genes toward mapping reads matching the reference genome is just a few percent, individual loci may have far greater biases (Stevenson *et al*., 2013). The importance of clusters of SNPs also means that the problem is expected to be greater in species like *Drosophila melanogaster* that have a far greater density of SNPs than humans (Li and Sadler, 1991).

One way in which researchers have attempted to overcome the bias is by aligning reads separately to maternal and paternal genomes (Coolon *et al*., 2012; Graze *et al*., 2012; McManus *et al*., 2010) or to transcriptomes (Pandey *et al*., 2013). These methods are effective, but are not useful in cases in which parental genotypes cannot be readily or cost-effectively obtained (Stevenson *et al*., 2013). Another strategy involves aligning reads to a reference genome supplemented with all possible haplotypes within one read-length (Satya *et al*., 2012). Again, this technique has been shown to reduce the reference bias; however, it is impractical for use in systems that contain many polymorphisms, as the number of haplotypes increases exponentially with the number of polymorphic sites (Stevenson *et al*., 2013). This problem can be reduced by phasing the data using population genetic data, provided that genotypes are available from multiple individuals (Turro *et al*., 2011). A third solution has been proposed by Stevenson *et al*. (2013), in which analysis of ASE is restricted to genomic regions with fewer differentiating

*To whom correspondence should be addressed.

sites than the number of mismatches allowed. This solution is not ideal, however, as it requires one to discard useful data, ultimately decreasing statistical power. Finally, altering the alignment parameters can reduce the bias but does not eliminate it, while a comparison of different alignment software packages revealed little difference among them (Degner *et al*., 2009; Stevenson *et al*., 2013).

Once reads have been mapped to the reference sequence, they are assigned to separate alleles and counted. This can be accomplished by identifying SNPs between alleles. Unfortunately, SNP calls from RNA-seq data are not reliable because of the fact that there is no a priori expectation regarding read frequencies for each SNP allele, as there is when sequencing genomic DNA from a diploid individual (Stevenson *et al*., 2013). Therefore, it is unclear whether unequal frequencies of reads from the two alleles are because of strong ASE or an incorrect SNP call. Furthermore, RNA editing can alter the sequence of RNA after transcription, and these changes can be mistaken for SNPs (Bahn *et al*., 2012). Accurate SNP calls are important for obtaining reliable estimates of ASE because errors in SNP calling, like mapping errors, will introduce bias toward one allele. Thus, SNPs are typically called from genomic data (Bullard *et al*., 2010; Fraser *et al*., 2011). Although this method is generally effective, there are situations in which one would want to determine ASE without the extra time and cost associated with acquiring genomic sequences.

Here we present a protocol for obtaining accurate ASE estimates in *D.melanogaster*, which involves creating an alternate reference sequence featuring SNPs called directly from an RNA-seq dataset after filtering out SNPs not observed in a high-quality SNP database. Our methodology is advantageous in that it only requires a single reference sequence, can be used in systems that contain many polymorphisms and does not involve discarding data based on the number of mismatches in a region.

2 METHODS

2.1 Datasets

We used three RNA-seq datasets. The first was produced by Massouras *et al*. (2012), and downloaded from the EMBL-EBI ArrayExpress Web site (accession number E-MTAB-1266). The dataset we used comprised 10 million 79-bp Illumina single-end reads generated from a cross between lines 362 and 765 of the Drosophila Genetic Reference Panel (DGRP), which is a set of highly inbred lines derived from a natural population (Mackay *et al*., 2012).

The second dataset was RNA-seq data simulated from the same cross as the first dataset (lines 362 and 765) and contained 10 million 79-bp single-end reads. The simulated data included SNPs reported in DGRP freeze 2 for each line but lacked indels or ASE. Reads were simulated separately for each line with an overall sequencing error rate of 1%, and the expression levels for each transcript were weighted by the expression levels observed in the first dataset. Reads were then merged together to mimic the cross.

The third dataset was our own RNA-seq data made up of 18 293 076 101-bp Illumina paired-end reads generated from a cross between a *D.melanogaster* genotype from Innisfail, Australia, and the isogenic reference stock used for the original genome sequence (Adams *et al*., 2000). The approach we took was to effectively isolate a single haploid genome from the Australian isofemale line (C12) and cross this to the reference stock. Specifically, we crossed virgin females from the Australian line to males of a *T(2;3)CyO-TM6/pr cn; mwh ry[506] e* balancer stock in which the second and third chromosomes co-segregate. A single male from the progeny exhibiting the balancer phenotype was then crossed with *y; cn bw sp* virgin females from the reference stock, and offspring not exhibiting the balancer phenotype were collected for sequencing. Six- to nine-day-old females were homogenized in Trizol and frozen on liquid nitrogen. RNA was extracted using Direct-zol RNA MiniPrep kits according to the manufacturer's protocol (Zymo Research, Irvine, CA). Libraries were constructed following poly-A selection using the standard non-strand-specific TruSeq RNA library preparation protocol and sequenced using Illumina HiSeq2000.

The *D.melanogaster* genomic sequence (Ensembl build BDGP5.25) and GTF transcript annotation files for coding and non-coding genes were downloaded from the TopHat Web site http://tophat.cbcb.umd.edu/igenomes.shtml. Variant data from DGRP freeze 2 were downloaded from the Baylor College of Medicine Web site http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze2_Feb_2013/.

2.2 Software

RNA-seq data were simulated using RNASeqReadSimulator (https://github.com/davidliwei/RNASeqReadSimulator). To ensure the expression level of genes in the real and simulated data was the same, reads mapping to each transcript in the default parameter alignment of the Massouras *et al*. dataset were enumerated using HTSeq (Anders *et al*., 2014) and used to weight simulations.

Sequences were quality trimmed using Trimmomatic version 0.30 (Bolger *et al*., 2014). Reads were trimmed from the 3′ end when average quality scores in sliding windows of 4 bp dropped below 20 or when the quality score at the end of the read dropped below 20. Sequences <50 bp in length were discarded.

We used TopHat (version 2.0.8 with Bowtie 2 version 2.1.0) for read alignment (Trapnell *et al*., 2012). For the alignment, we input the GTF file with known *D.melanogaster* transcripts (setting the –G parameter), instructed TopHat not to consider novel splice junctions (––no–novel–juncs), set the length of seed substrings to 20 (––b2–L 20) and set the number of mismatches allowed in an alignment during multiseed alignment to one (––b2–N 1). In addition, we varied the number of read-mismatches (–N) and the indel length (––read–gap–length, ––max–insertion–length, ––max–deletion–length, ––read–edit–dist) allowed in the final read alignment (see Section 3). SAMtools (version 0.1.19) was used to discard non-uniquely mapping reads and to produce an mpileup file (Li *et al*., 2009). To identify potential SNPs in the RNA-seq data, we used VarScan mpileup2snp (version 2.3.5) (Koboldt *et al*., 2009). For a potential SNP to be called by VarScan, the read-depth had to be greater than one (––min–coverage 2), and a read had to have an average base quality of at least 20 to be counted (––min–avg–qual 20). In addition, we set the *P*-value threshold for calling SNPs at 1.0 (––*P*-value 1), did not implement a strand filter (––strand–filter 0) and set the minimum variant allele frequency threshold at 1e-10 (––min–var–freq 1e-10). Bedtools intersect (version 2.17.0) was used to filter SNPs by known variants (by setting the –wa parameter) and to filter and annotate SNPs according to location within a gene (by setting –wa –wb) (Quinlan and Hall, 2010). We generated alternate FASTA sequences via GATK's FastaAlternateReferenceMaker (version 2.4.9) (McKenna *et al*., 2010). HapCUT (version 0.5) was used to phase SNPs (Bansal and Bafna, 2008).

In-house python scripts were used for all else not covered above. The scripts and data files required to recreate this analysis are available at https://github.com/d-quinn/bio_quinn2013. The sequences generated during this project have been submitted to the Sequence Read Archive and have the accession number SRP040244.

2.3 Aligning to a single reference

To examine the effects of mapping bias, we used a published RNA-seq dataset from a cross between DGRP lines 362 and 765 for which the

genomic sequences are available (Massouras *et al*., 2012). Reads were aligned to the *D.melanogaster* reference sequence using TopHat, and those that did not align uniquely to the genome were discarded. We identified candidate SNPs using VarScan mpileup2snp set with the extremely generous parameters listed earlier (which will have allowed many false-positive results), identifying any candidate SNP found in at least two supporting reads with average base quality >19. We also removed SNPs with more than two alleles, as SNPs were being called from a single individual so these represent errors. We then filtered SNPs according to a variable and fixed coverage cutoff. Where a gene has been sequenced to a high depth of coverage, it becomes increasingly likely that the same sequencing error will occur in multiple reads. To compensate for this, we required SNPs to have more supporting reads as the depth of coverage increased. We used a binomial distribution where the probability of an error in any one read was 1 in 100 (the expected rate with a phred score of 20). A threshold was set for each depth of coverage such that the probability of observing enough erroneous reads to exceed this threshold would be <0.0001, assuming every read had the lowest possible quality score (a phred score of 20). In addition to this variable coverage cutoff, we used a fixed cutoff, which eliminated SNPs at positions with fewer than 15 total reads. Furthermore, SNPs that were not found in known transcripts (those that did not intersect with genes in the GTF transcript annotation file) were discarded, as they are not relevant for measuring ASE. Finally, SNPs were filtered by a set of known variants, retaining only SNPs that have been previously reported (see Section 3).

2.4 Aligning to multiple references

To reduce mapping bias, we aligned reads to both the original reference genome as well as an alternate version that featured SNPs called from the RNA-seq data. The protocol was similar to that for aligning to a single reference, with the addition that the filtered SNPs identified from the initial alignment were then used to create an alternate reference sequence, using GATK's FastaAlternateReferenceMaker and an in-house python script to fix FASTA headers. FastaAlternateReferenceMaker replaces reference bases at variant positions with bases supplied by a file in variant call format (VCF). Raw reads were realigned to this alternate sequence, and SNPs were identified in the same manner as before. Finally, we combined unique reads from both alignments to get per-SNP ASE estimates. A flow chart summarizing these steps and some others explained later is shown in Supplementary Figure S1.

2.5 Aligning to parental genomes

We used SNPs from the DGRP freeze 2 dataset to generate parental genomes for the Massouras et al. (2012) data. Specifically, SNPs in DGRP lines 362 and 765 were separately used as inputs for FastaAlternateReferenceMaker, yielding two separate parental genome proxies. Only homozygous SNPs were included, as the DGRP lines are highly inbred. We aligned the RNA-seq data to each parental genome, called SNPs (again, implementing a variable and fixed coverage cutoff) and calculated ASE values for each position called as a SNP in either of the parental lines. Unique reads from the two alignments were combined to get per-SNP estimates of ASE. We refer to this as our benchmark alignment, as it should produce the most accurate ASE estimates for the ideal case where both parental genome sequences are known.

2.6 Phasing SNPs

To phase SNPs, we used HapCUT, which uses a max-cut-based algorithm for haplotype assembly (Bansal and Bafna, 2008). We input our set of filtered SNPs from the alignment, the corresponding alignment (BAM) file and the *D.melanogaster* reference sequence (in FASTA format) into HapCUT and used the output to create two VCF files with an in-house python script. The logic for the creation of the VCF files was as follows: (i) if a SNP in the original VCF (the set of filtered variants from mapping to a single genome) was homozygous (comprises all non-reference reads), the SNP was included in both VCF files; (ii) if a SNP was heterozygous (there were reference and non-reference reads) and phased, it was included in a single VCF along with the other SNPs on that haplotype; and (iii) heterozygous SNPs that were not phased were all placed into one of the VCFs. The resulting two VCFs were input into FastaAlternateReferenceMaker to create two alternate reference sequences. Reads were aligned to these separately, SNPs called and unique reads combined to get per-SNP estimates of ASE.

2.7 Combining SNPs by gene

We wrote a script to combine SNPs across a gene in such a way that reads spanning multiple SNPs are counted only once. As input, it takes two VCF files that have been intersected (using the -wa -wb parameters of Bedtools intersect) with a GTF as well as the corresponding BAM files used to call those variants. The script iterates through the VCF files and produces a list of SNPs for each gene. It then uses Pysam, a lightweight wrapper of the SAMtools C-API, to generate lists of read IDs for each state of each SNP. For instance, if a SNP on chromosome 2L has two states, A and G, representing the reference and non-reference bases, respectively, it will generate a list of read IDs that contain an A at that position and a list of reads IDs that contain G at that position, and it will repeat this for every SNP in the gene. Lists of reference and non-reference read IDs are concatenated, the duplicates are removed and the lengths are used as the new per-gene expression counts.

2.8 Removing conflicts

When a single read contains multiple SNPs, then all the SNPs should assign the read to the same parent. Errors can therefore be detected when the SNPs conflict and assign the same read to different parents. On the first pass through our pipeline, we recorded both which reads were assigned to both parents and which SNPs were causing these conflicts. These reads and SNPs can be ignored on a second pass through the pipeline, producing a new set of per-gene estimates.

3 RESULTS AND DISCUSSION

3.1 Mapping: alignment to single reference

To examine patterns of ASE, we aligned published RNA-seq data from the F1 progeny of a cross between two inbred *D.melanogaster* lines (Massouras *et al*., 2012) to the published reference genome sequence, allowing up to two mismatches between each read and the genome (the default) [Fig. 1 (i)]. The genomes of these two lines have been sequenced, so we called SNPs from the aligned reads but removed any SNPs missing from the genome sequences. In total, there were 29 999 SNPs in 5404 genes (of the 14 869 genes in the genome). With these SNPs, we were able to assess the contribution of mapping bias alone—independent of SNP calling errors—to estimates of ASE. The mean proportion of reads carrying the reference allele of the SNPs was 0.535, much higher than the expected value of 0.5 (Table 1). The strength of this bias was strongly correlated with the density of SNPs (Supplementary Fig. S2), as expected if the bias is caused by SNPs introducing mismatches to the reference genome and preventing reads from mapping. In contrast, indels seem less important, as there was no difference in the strength of the bias between genes with or without indels ($t = 0.10$, df = 3564, $P = 0.92$).

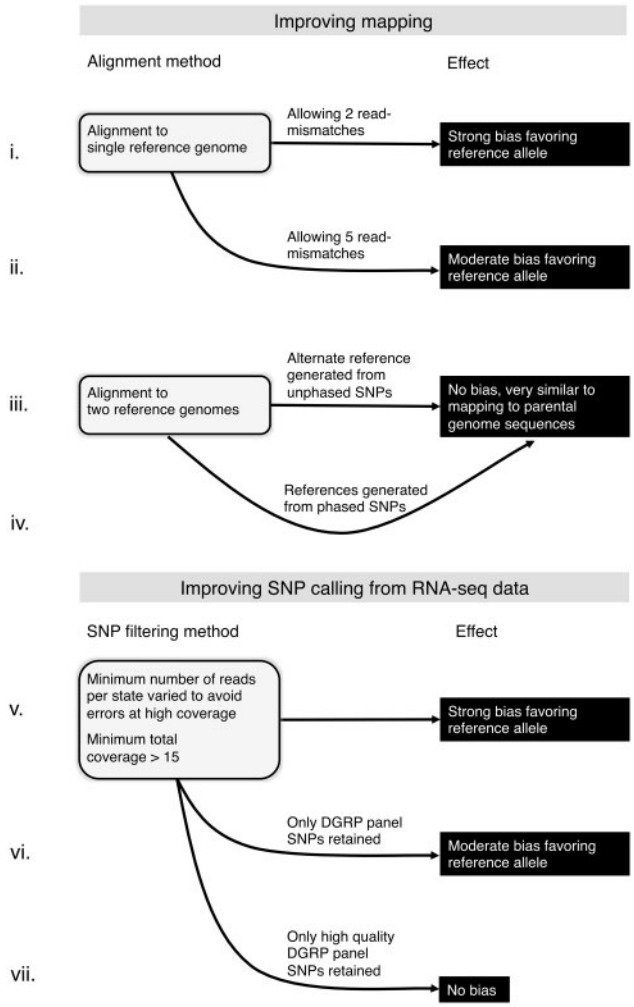

Fig. 1. Summary of the effect of different alignment and SNP filtering methods on measures of allele-specific expression. For **(i)**–**(iv)**, SNPs were filtered by those identified from the parental genomic sequences so that estimates of ASE would be independent of SNP calling errors. For **(ii)**–**(vii)**, we allowed five read-mismatches in the alignment, as this substantially reduced the reference allele bias. The DGRP was used to filter SNPs in (vi) and (vii)

Because mapping biases are being caused by SNPs preventing the read from mapping, they are affected by the number of mismatches allowed between the read and reference genome during mapping. To investigate this, we reran the previous protocol with several different sets of TopHat parameters. We found that increasing the number of mismatches allowed in an aligned read (–N) decreased the mapping bias substantially (Table 1 and Supplementary Fig. S3). The potential cost of allowing more mismatches is that the reads might map to multiple locations in the genome, decreasing the number of uniquely mapped reads. As multiple mapping might differentially affect the two alleles of a gene, the removal of these reads could give a false signature of ASE. However, although the number of uniquely mapped reads increased as we allowed more mismatches, the number of multiply mapped reads remained fairly constant (Table 1). We also varied the maximum indel length allowed in an aligned read; however, this had relatively little effect on the reference bias or the number of mapped reads (Table 1). Altering other parameters had little effect, so default settings were used for these [data not shown; number of read-mismatches allowed in a seed (––b2–N), the interval between seed substrings (––b2–i) and the length of the seed substrings (––b2–L)].

Based on our results in Table 1, we chose to allow up to five read-mismatches in future alignments and kept the allowed indel length at the default value of two. In the case of alignment to the published reference, the mean proportion of reads carrying the reference allele was 0.505 [Fig. 1 (ii) and Table 1], with this bias affecting 1.5% of SNPs (51.5% of SNPs have more than half the reads being the reference allele). Although allowing 10 mismatches gave a value closer to 0.5, we reasoned that the gain was not substantial enough to warrant the reduced stringency, which could compromise the accuracy of the alignment.

Others have demonstrated that accurate estimates of ASE can be obtained by aligning RNA-seq data to the genomes of the parents of the individual under study (Coolon *et al.*, 2012). Using this technique, we found that the mean proportion of reads carrying the reference allele after aligning to the parental genomes was 0.504, slightly closer to 0.5 than the mean from our single alignment (0.505). Although these numbers are similar, when we compared bias among individual sites, it is apparent that there are many SNPs in the single alignment that remain biased toward the reference allele (Fig. 2A). Thus, allowing multiple mismatches greatly reduces the bias toward the allele found in the reference genome, but does not eliminate it.

The analysis to this point has used real RNA-seq data, so true ASE might affect our results. To check that this was not the case, we simulated data with no ASE and repeated our analysis. The overall pattern was extremely similar to the true data, with mapping to a single genome generating a false signal of ASE at some SNPs (Fig. 2B, mean proportion reference mapping to reference genome is 0.503 compared with 0.501 when mapping to parental genomes).

It is unclear why a small bias toward the reference remains even when aligning to the parental genomes. One possibility is that there is still a mapping bias because our parental genomes are incomplete. In particular, these genomes still have the reference state for indels. However, this can only be a partial explanation, as there is still a small bias for the simulated RNA-seq reads that lack any indels (Fig. 2B), and the presence of indels is not correlated with the bias (see earlier text). Furthermore, any variants not included in the parental genomes are excluded from our analysis, so they will only create a systematic bias toward the reference allele at neighboring sites that were analyzed if they are in linkage disequilibrium. Alternatively, the remaining bias may result from errors in SNP calling rather than a failure to map reads.

3.2 Mapping: alignment to multiple references

In an effort to improve our estimates further, we aligned the Massouras *et al.* (2012) data to the published reference sequence and then to an alternate sequence featuring SNPs called from the RNA-seq data [Fig. 1 (iii)]. Thus, we generated an alternate reference sequence without having to sequence the genomes of the parents. Again, for both the alignment to the original reference

Table 1. Effect of varying alignment parameters on the bias toward mapping reads carrying the allele found in the reference genome

Mismatches	Indel length[a]	Proportion reference[b]	Uniquely mapped reads[c]	Multiply mapped reads[d]	Unmapped reads
2	2	0.535	8 859 264	98 028	1 042 708
3	2	0.517	9 210 621	101 515	697 864
5	2	0.505	9 405 429	104 288	490 283
10	2	0.502	9 506 855	105 891	387 254
3	3	0.517	9 219 892	101 576	678 532
5	5	0.505	9 434 418	104 457	461 125
10	10	0.502	9 598 279	106 599	295 122

[a]Indel length allowed in an aligned read. We specified this by setting the parameters --read–gap–length, --max–insertion–length, --max–deletion–length, --read–edit–dist, to the number shown in the table.
[b]Mean proportion of reads carrying the reference allele of SNPs.
[c]Number of uniquely mapped reads.
[d]Number of reads that mapped to more than one location and were therefore discarded.

sequence and to the alternate sequence, we only analyzed homozygous SNPs found in the parental genomic sequences, allowing us to separate the contribution of mapping bias from errors in SNP calling. Initially, we did not phase the SNPs, so the alternate sequence is a mixture of SNPs identified in both genomes. The mean proportion of reads carrying the reference allele was 0.504 (Fig. 2C), which is the same as when we aligned the reads to the parental genome sequences (our benchmark). Furthermore, the measures of ASE for individual SNPs were nearly identical to those obtained when we aligned to the parental genomes (Pearson's $R^2 = 0.999$, Fig. 2C). In addition, coverage for SNPs from aligning to multiple reference sequences is nearly identical to that for the benchmark (Pearson's $R^2 = 1.00$; Fig. 2E). This indicates that a virtually identical set of reads are mapped when aligning to multiple references generated from unphased SNPs as when the reads are mapped to the parental genomes. Overall, this represents a substantial improvement over the alignment to a single genome, with fewer SNPs exhibiting a reference bias (Fig. 2C versus A) because of this approach allowing us to map reads that would otherwise be missing from the single alignment (Fig. 2E versus F).

A potential weakness with our protocol for aligning to two genomes is that we have no information concerning the chromosome on which each SNP is found (i.e. its phase). This means that the alternate genome sequence we generated from the unphased SNPs includes variants from both chromosomes, so sequence reads may still not be perfect matches to either of the genomes we are using. It should often be possible to accurately phase clusters of SNPs because of the co-occurrence of SNPs within the same read pairs (Bansal and Bafna, 2008). We used HapCUT (Bansal and Bafna, 2008) to phase SNPs (restricted to SNPs found in the parental genomes as before) and generated two alternate genomes from these phased SNPs [Fig. 1 (iv)] After aligning to these, we found that the mean proportion reference was 0.504, nearly identical to that of the unphased alignment. Plotting this against estimates from our benchmark revealed that phasing has little effect on ASE estimates ($R^2 = 0.999$, compare Fig. 2C and D).

The finding that generating alternative reference genomes using unphased SNPs performs as well as using phased SNPs is unexpected. Presumably this is because the large number of mismatches allowed when the reads are mapped circumvents any need for phasing. For a read not to map to both the reference and an unphased alternate genome, it must contain six mismatches to both genomes (there must be a cluster of 12 SNPs within the space of one read). Linkage disequilibrium between nearby SNPs will make this scenario even more uncommon, as reference and alternate alleles will tend to be found on the same reads.

3.3 SNP calling: using the RNA-seq data

In the previous sections, we examined patterns of ASE using only SNPs found in the parental genomic sequences, which allowed us to isolate the effect of mapping bias from that of SNP calling errors. By altering our alignment parameters and aligning reads to multiple reference sequences, we were able to largely eliminate mapping bias. We next explored how errors in SNP calling can affect estimates of ASE and developed a strategy to reliably call SNPs from the RNA-seq data itself.

To investigate how errors in SNP calling from RNA-seq data can bias estimates of ASE, we aligned the Massouras *et al*. (2012) dataset to the published reference sequence as well as to an alternate reference generated using SNPs called from the RNA-seq data [Fig. 1 (v)]. Again, SNPs were filtered by a fixed and variable coverage cutoff (see Section 2). However, unlike before, we did not remove SNPs that had not been found in the genome sequences of these lines. The errors in SNP calling were substantial. Because sequencing errors tend to occur at low frequency, they inflate the frequency of the reference allele in the same way as mapping biases. The mean proportion of reads carrying the reference allele was 0.517 (Fig. 3A).

The strategy we took to improve the quality of our SNP calls was to remove any SNPs that had not been previously reported in *D.melanogaster*. Initially, we filtered our SNP calls from the RNA-seq data by those found in the DGRP lines, a panel of highly inbred lines whose genomes have been sequenced [Fig. 1 (vi)]. This resulted in a mean proportion reference of 0.510. This suggested that although this basic filtering is somewhat effective at removing errors, many still remain. The DGRP lines are highly inbred, so SNPs always called heterozygous are likely to be errors, and singletons may also be of lower quality. Therefore,

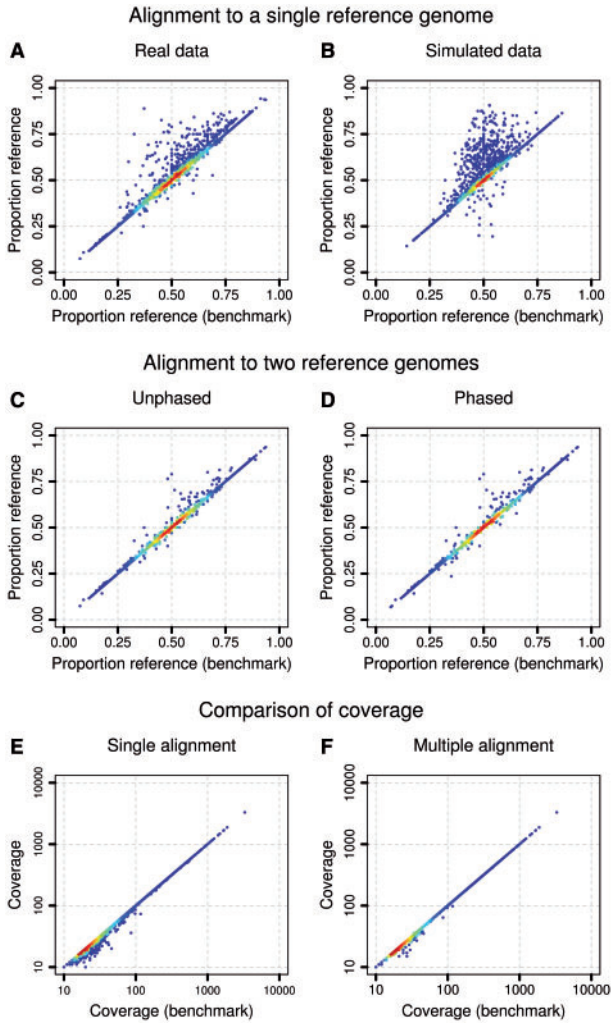

Fig. 2. The effect of aligning to different reference genomes on the bias toward the reference allele and the number of reads mapped. Panels **A** and **B**: There is a substantial bias toward the reference allele when mapping (A) real or (B) simulated data to the reference genome (Y axes) compared with mapping to the parental genomes (our benchmark, X axes) (Panel A: Pearson's $R^2 = 0.993$, Panel B: Pearson's $R^2 = 0.972$). Panels **C** and **D**: Aligning to genomes including SNPs called from RNA-seq data substantially improves estimates of allele-specific expression, regardless of whether SNPs are (C) unphased or (D) phased. In both cases, the proportion of the reference allele of SNPs is 0.504, and the correlation with the benchmark of aligning to the parental genomes is strong (Pearson's $R^2 = 0.999$). Panels E and F: Aligning to multiple reference sequences increases the number of mapped reads. The coverage of SNPs is shown from aligning reads to (**E**) the published reference genome and (**F**) both the published reference and an alternate reference generated including unphased SNPs called from the RNA-seq data. In both cases, the coverage is compared with our benchmark alignment to both parental genomic sequences. Shading indicates a greater density of superimposed points

we refined our criteria by keeping only SNPs that were called homozygous in two or more DGRP lines [Fig. 1 (vii)], giving a mean proportion reference of 0.505, an estimate that is nearly identical to that from our benchmark (Fig. 3B).

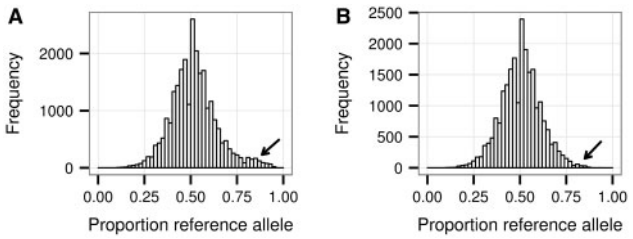

Fig. 3. Filtering out low-quality SNPs using a SNP database reduces bias. Reads were aligned to the published reference genome as well as to an alternate reference generated using SNPs called from the RNA-seq data. (**A**) Unfiltered SNPs have considerable bias (mean = 0.517), while (**B**) a substantial portion of the bias is removed by only retaining SNPs that were called homozygous in at least two DGRP lines (mean = 0.505). Arrows highlight portions of the distributions that change

We plotted estimates from variants filtered by SNPs called homozygous in two or more lines against those from the benchmark, which was based on aligning to and SNP calling from parental genome sequences. We found that the two datasets are similar (Pearson's $R^2 = 0.999$, Fig. 4A). Comparing this with the plot shown in Figure 2D illustrates that filtering by SNPs called homozygous in two or more lines gives much the same results to filtering by SNPs called from the parental genomes alone.

By relying on a high-quality database of known SNPs, we called SNPs using as little as 15X coverage with two reads supporting the variant. These fairly low coverage data appear to be sufficient, as there is no correlation between coverage and the reference bias, which can be used as a proxy for SNP calling errors (Supplementary Fig. S4).

These analyses have used real data, which may be affected by factors such as true ASE or our benchmark having SNP calling errors. We therefore repeated the analysis using simulated data where these factors are controlled, and again found that SNPs can be reliably called from RNA-seq data (Figs. 4B versus 2C).

3.4 Per-gene ASE estimates: removing pseudo-replication and conflicts

Thus far, we have focused on per-SNP estimates of ASE, but most research questions are concerned with gene-level estimates of ASE. Gene-level expression could be calculated by simply adding up ASE estimates from each SNP in a gene. However, this method is problematic because it leads to pseudo-replication when a single read overlaps two or more SNPs. To avoid this, we combined reference and non-reference counts across the gene without counting a read more than once.

Our technique relies on the data being accurately phased across the full length of the gene, so phasing using the sequences themselves may be unreliable. We suggest a simple alternative approach to phase the variants in *Drosophila* and other model organisms: if the genotype of interest is crossed to the homozygous strain used to generate the reference genome, then all the variants called are inherently phased. With this in mind, we generated a dataset composed of 100-bp Illumina paired-end reads from the F1 progeny of a cross between a haploid genome derived from a fly line collected in Innisfail, Australia,

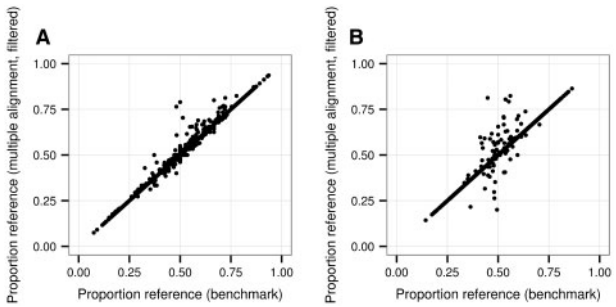

Fig. 4. Filtering by SNPs called homozygous in two or more lines from the DGRP produces accurate ASE estimates directly from RNA-seq data with (**A**) real and (**B**) simulated data. Proportion of reference allele in SNPs from alignment to multiple references and by SNPs called homozygous in two or more lines from the DGRP (y-axis) against estimates from alignment to parental genomic sequences (our benchmark) (x-axis) [Pearson's R^2 = (A) 0.999 and (B) 0.997]

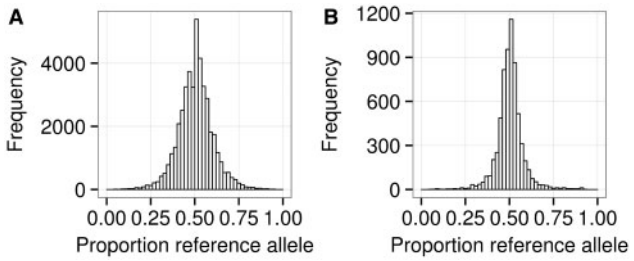

Fig. 5. No bias was observed when applying our pipeline to RNA-seq data generated from a cross between the *D.melanogaster* reference line and a line collected in Australia. The data were aligned to multiple references featuring SNPs called from the RNA-seq data and filtered by high-quality DGRP SNPs [scenario (vii) in Fig. 1]. Proportion reference allele distribution across (**A**) SNPs (mean = 0.504) and (**B**) across genes (mean = 0.503)

and the *D.melanogaster y; cn bw sp* stock sequenced for the reference genome (Adams *et al*., 2000). As in the previous section, we aligned reads to the published reference as well as to an alternate reference made up of SNPs called from the RNA-seq data. Again, we filtered variants by a fixed and variable coverage filter and by those found homozygous in at least two DGRP lines. The proportion of read-pairs carrying the reference allele was 0.504 when the SNPs were analyzed independently and 0.503 when they were combined within each gene (Fig. 5). As this bias is slightly less than what we observed earlier using two DGRP lines (lines whose genome sequences are included in our database of high-quality SNPs), this suggests that our approach at SNP calling can be applied to cosmopolitan populations of *Drosophila* that have not been included in the reference database.

With accurate phasing, read-pairs spanning multiple SNPs can be used to improve estimates further. Because one set of chromosomes in our dataset came from the Australian genome, whereas the other came from the reference genome, read-pairs should be assigned to only one genome. Of the 2 456 755 read-pairs spanning multiple SNPs, 99.7% were assigned to the same genome, whereas 0.3% were assigned to both genomes. Across SNPs, we found that 5% of 45 920 SNPs had a single conflicting read-pair, and 1.5% had greater than one conflicting read-pair.

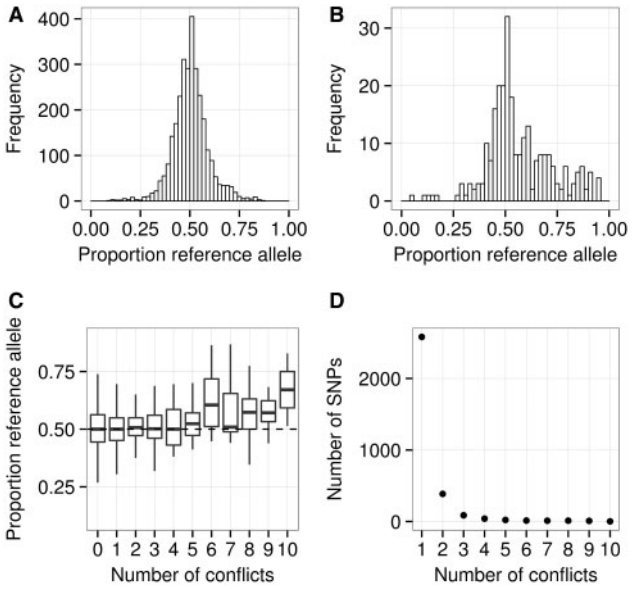

Fig. 6. The effect of conflicts on ASE estimates. A conflict occurs when a single read-pair is assigned to both parental genomes. (**A**) Distribution of the proportion of read-pairs carrying the reference allele for SNPs with one or two conflicts and (**B**) SNPs with greater than two conflicts. (**C**) Boxplots of the proportion of read-pairs carrying the reference alleles for SNPs with different numbers of conflicts (**D**) Number of SNPs with 1–10 conflicts

These conflicts indicate SNP calling or sequencing errors. If the conflict is caused by a sequencing error, the read should be excluded, but if it is caused by a SNP calling error, the SNP should be excluded. We found that SNPs with fewer than five conflicting read-pairs – which represented the grand majority of SNPs with conflicts (Fig. 6D)—had close to the expected distribution of proportion of reference read pairs (Fig. 6A and C). This suggests that ASE estimates for most of these SNPs are nearly correct. We reasoned that the majority of SNPs with a small number of conflicts were not called in error, and that the conflict was instead because of sequencing error. In contrast, those with five or more conflicts did not have the expected distribution (Fig. 6B and C), and therefore, many of these might be SNP calling errors. The most conservative course of action would be to remove all SNPs with conflicts, but this would involve discarding a large amount of data (Supplementary Table S1). With this in mind, we removed SNPs that contained three or more conflicts, and removed read-pairs that conflicted with one or two SNPs. The mean proportion of reference read-pairs did not change appreciably after removing conflicting SNPs and read-pairs. However, of the 2 384 931 reads that spanned multiple SNPs, 99.9% were now assigned to the same genome. Furthermore, the total number of conflicting SNPs decreased from 7 to 3.5%, and no SNPs exhibited more than two conflicting read-pairs (SNPs with two conflicts represented only 0.3% of the 44 643 total SNPs).

Our decision to remove only SNPs with three or more conflicting reads was driven by our desire to retain most of the useful

data, and this should be adjusted according to the experimental design and goals of a given project. For obtaining the most accurate estimates, we suggest removing all SNPs that contain a read-pair assigned to multiple genomes. For retaining more data, there is little bias if as many as five conflicts are allowed (Fig. 6C).

4 CONCLUSION

We have developed a protocol for obtaining accurate per-SNP and per-gene estimates of ASE from RNA-seq data. The main advantage of this protocol is that it does not require parental genomic sequences; SNPs used to get allele counts are called directly from the RNA-seq data. This can significantly reduce the time and cost associated with measuring ASE.

Unexpectedly, it does not appear that phased data are required for getting unbiased per-SNP measures of ASE using our datasets. We found that our estimates were nearly the same with and without phasing. However, phased data are required for combining SNPs to obtain per-gene measures of ASE, and this also allows us to remove unreliable reads and SNPs. We show that the data can be phased if genotype of interest is crossed to the homozygous strain used to generate the reference genome, although this will only be possible in *Drosophila* and other model organisms.

The main drawback to our method is that it requires knowledge of variation in the population of the organism of study, as this allows filtering to remove SNP calling errors. In *Drosophila* this data is readily available. As new sequencing technologies mean that comprehensive SNP databases are available from increasing numbers of species, this approach will become viable in even more species, especially those with short generations.

ACKNOWLEDGEMENTS

The authors thank Ary Hoffmann and Jennifer Shirriffs for providing the Australian fly line.

Funding: European Research Council grant DrosophilaInfection; the Royal Society University Research Fellowship to F.J.; and the Pomona College Downing Scholarship to A.Q.

Conflict of Interest: none declared.

REFERENCES

Adams,M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Anders,S. *et al.* (2014) HTSeq –A Python framework to work with high-throughput sequencing data. *bioRxiv*, doi: 10.1101/002824.

Babak,T. *et al.* (2010) Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. *BMC Genomics*, **11**, 473.

Bahn,J.H. *et al.* (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.*, **22**, 142–150.

Bansal,V. and Bafna,V. (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **24**, i153–i159.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bullard,J.H. *et al.* (2010) Polygenic and directional regulatory evolution across pathways in *Saccharomyces*. *Proc. Natl Acad. Sci. USA*, **107**, 5058–5063.

Coolon,J.D. *et al.* (2012) Genomic imprinting absent in *Drosophila melanogaster* adult females. *Cell Rep.*, **2**, 69–75.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Fraser,H.B. *et al.* (2011) Systematic detection of polygenic cis-regulatory evolution. *PLoS Genet.*, **7**, e1002023.

Graze,R.M. *et al.* (2012) Allelic imbalance in *Drosophila* hybrid heads: exons, isoforms, and evolution. *Mol. Biol. Evol.*, **29**, 1521–1532.

Koboldt,D.C. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,W.H. and Sadler,L.A. (1991) Low nucleotide diversity in man. *Genetics*, **129**, 513–523.

Mackay,T.F. *et al.* (2012) The *Drosophila melanogaster* genetic reference panel. *Nature*, **482**, 173–178.

Massouras,A. *et al.* (2012) Genomic variation and its impact on gene expression in *Drosophila melanogaster*. *PLoS Genet.*, **8**, e1003055.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

McManus,C.J. *et al.* (2010) Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.*, **20**, 816–825.

Pandey,R.V. *et al.* (2013) Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol. Ecol. Res.*, **13**, 740–745.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Satya,R.V. *et al.* (2012) A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res.*, **40**, e127.

Stevenson,K.R. *et al.* (2013) Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, **14**, 536.

Trapnell,C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.

Turro,E. *et al.* (2011) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.*, **12**, R13.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.