Heredity (2014) 113, 156–166

www.nature.com/hdy

ORIGINAL ARTICLE

Using next-generation RNA sequencing to identify imprinted genes

X Wang[1,2] and AG Clark[1,2]

Genomic imprinting is manifested as differential allelic expression (DAE) depending on the parent-of-origin. The most direct way to identify imprinted genes is to directly score the DAE in a context where one can identify which parent transmitted each allele. Because many genes display DAE, simply scoring DAE in an individual is not sufficient to identify imprinted genes. In this paper, we outline many technical aspects of a scheme for identification of imprinted genes that makes use of RNA sequencing (RNA-seq) from tissues isolated from F1 offspring derived from the pair of reciprocal crosses. Ideally, the parental lines are from two inbred strains that are not closely related to each other. Aspects of tissue purity, RNA extraction, library preparation and bioinformatic inference of imprinting are all covered. These methods have already been applied in a number of organisms, and one of the most striking results is the evolutionary fluidity with which novel imprinted genes are gained and lost within genomes. The general methodology is also applicable to a wide range of other biological problems that require quantification of allele-specific expression using RNA-seq, such as *cis*-regulation of gene expression, X chromosome inactivation and random monoallelic expression.
Heredity (2014) **113,** 156–166; doi:10.1038/hdy.2014.18; published online 12 March 2014

INTRODUCTION

In diploid organisms, a subset of genes are expressed exclusively or preferentially from one of the two parental alleles, resulting in allelic imbalance (AI) in gene expression (Pastinen and Hudson, 2004). In some cases, the deviation from 50%:50% is dependent on the parent of origin, such as genomic imprinting (Reik and Walter, 2001; Barlow, 2011; Bartolomei and Ferguson-Smith, 2011) and imprinted X chromosome inactivation (Wake *et al.*, 1976; Huynh and Lee, 2001; Xue *et al.*, 2002; Dindot *et al.*, 2004; Wang *et al.*, 2013a). In other cases, the AI is random or sequence dependent, such as random X chromosome inactivation (Heard *et al.*, 1997), autosomal random monoallelic expression (RMAE; Gimelbrant *et al.*, 2007), allelic exclusion of immunoglobulin genes (Vettermann and Schlissel, 2010), *cis*-regulating expression quantitative trait loci (*cis*-eQTL), loss of heterozygosity in cancer (Thiagalingam *et al.*, 2001) and monoallelic expression of olfactory receptors (Chess *et al.*, 1994). Technical details of the methods for quantifying differential allelic expression (DAE) accurately are critical to study AI.

Genomic imprinting is a special form of AI in which the expression level of each allele depends on the parent-of-origin (Reik and Walter, 2001; Barlow, 2011; Bartolomei and Ferguson-Smith, 2011). To date, about 120 imprinted genes have been identified and successfully validated in humans and mice (Morison *et al.*, 2001, 2005; Prickett and Oakey, 2012), and we know this list is not complete. Genome-wide and transcriptome-wide approaches have been applied to detect genomic imprinting (Maeda and Hayashizaki, 2006; Henckel and Arnaud, 2010), including microarray expression profiling of parthenogenote and androgenote embryos (Mizuno *et al.*, 2002; Nikaido *et al.*, 2003; Kuzmin *et al.*, 2008; Sritanaudomchai *et al.*, 2010) and uniparental disomic mice (Choi *et al.*, 2001, 2005; Schulz *et al.*, 2006), expression profiling using allele-specific single-nucleotide polymorphism (SNP) arrays (Pollard *et al.*, 2008; Serre *et al.*, 2008; Brideau *et al.*, 2010; Morcos *et al.*, 2011) and computational prediction methods (Ke *et al.*, 2002; Yang *et al.*, 2003; Luedi *et al.*, 2005, 2007; Brideau *et al.*, 2010). Recently, RNA-seq has become the method of choice for in-depth analysis of the whole transcriptome of an organism (Ozsolak and Milos, 2011). Among many applications, quantification of DAE is crucial for understanding a number of fundamental biological questions, such as *cis*-acting factors regulating gene expression, identification of genomic imprinting and parent-of-origin effects, X chromosome inactivation and dosage compensation. To search for novel imprinted genes genome-wide in an unbiased way, we and others have applied next-generation RNA sequencing (RNA-seq) and quantitatively measured the allele-specific expression in RNA samples from reciprocal crosses of inbred mouse strains. This was accomplished by directly counting the number reference/alternative allele-containing reads at polymorphic SNP positions in the parental genome (Babak *et al.*, 2008; Wang *et al.*, 2008, 2011; Gregg *et al.*, 2010a, b; DeVeale *et al.*, 2012; Okae *et al.*, 2012). This approach has been successful in identifying a number of additional imprinted genes in mice.

Although the RNA-seq method is expected to allow discovery of the full complement of genes that undergo genomic imprinting, there are some discrepancies between different studies and some of the candidate genes could not be verified, suggesting that technical issues with RNA-seq for DAE quantification can produce false-positive calls,

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA and [2]Cornell Center for Comparative and Population Genomics, Cornell University, Ithaca, NY, USA
Correspondence: Dr AG Clark, Molecular Biology and Genetics, Cornell University, 107 Biotechnology Building, Ithaca, NY 14853, USA.
E-mail: ac347@cornell.edu
Received 22 August 2013; revised 2 December 2013; accepted 19 December 2013; published online 12 March 2014

necessitating a need for subsequent validation using an independent method (DeVeale *et al.*, 2012; Kelsey and Bartolomei, 2012; Okae *et al.*, 2012). Based on our experience with quantification of DAE from RNA-seq data and pyrosequencing validation (Wang and Elbein, 2007), the allele-specific expression ratios in reciprocal F1 hybrids could be accurately quantified using RNA-seq data with sufficient library complexity and coverage depth. In this report, we discuss the challenges for inference of allele-specific expression using RNA-seq, including correcting the reference genome mapping bias, assessing library complexity, filtering problematic SNPs and determining the true false discovery rate.

Genomic imprinting has been discovered in both animals and plants (Kohler and Weinhofer-Molisch, 2010). In animals, genomic imprinting is only known in therian mammals, and most imprinted genes are expressed and imprinted in the brain and placenta (Pask, 2012; Keverne, 2013; Renfree *et al.*, 2013). Having the full catalog of imprinted genes in several different mammalian species will greatly facilitate the understanding of many evolutionary questions about genomic imprinting, including but not limited to origin and fixation of genomic imprinting, conservation of degree of expression direction and imprinted genes, dynamic gain and loss of imprinting status on the mammalian phylogenetic tree and effect of loss of diploidy for imprinted genes in shaping the population genetic profile. However, to date, the majority of imprinted genes have been discovered in mouse or human, and studies in other mammals have mostly been surveys or confirmations of the imprinting status for known imprinted genes. This ascertainment bias has limited the ability of comparative evolutionary genomic analysis. And the presence of non-mammalian imprinting effects (de la Casa-Esperon, 2012) also motivates an unbiased way for searching novel imprinted genes without any previous knowledge. Quantification of DAE from RNA-seq data of reciprocal F1 individuals remains a powerful method to achieve this goal when the study design, experimental procedures and data analysis have been properly performed. In addition, the method for DAE quantification could be used to study other forms of AI as well.

QUANTIFICATION OF DAE IN RECIPROCAL CROSSES FROM INBRED OR SEMI-INBRED STRAINS

In order to identify novel imprinted genes by allele-specific expression analysis, informative SNP positions in exonic regions are needed to distinguish the relative transcript abundance from the two parental alleles. RNA samples from hybrid F1 individuals derived from reciprocal crosses of inbred strains/species are ideal for this purpose. The advantage of using inbred or semi-inbred strains/species is to have every single SNP in the F1 transcripts be informative with trackable transmission direction (Figure 1a). To identify novel imprinted genes, detecting significant AI (deviation from 50%:50% expression from the two parental alleles) is not sufficient, because there are three different possibilities: *cis*-eQTL effect, RMAE or parent-of-origin effect due to genomic imprinting or imprinted X chromosome inactivation. By *cis*-eQTL, we mean *cis*-regulating expression quantitative trait locus, which could be due to 5′ or upstream/downstream *cis*-acting polymorphisms near the promoter/enhancer regions that display variability in their interaction with transcription factors to produce differential expression, or cis-eQTLs could be 3′-untranslated region SNPs affecting mRNA stability or microRNA binding (Gilad *et al.*, 2008; Majewski and Pastinen, 2011). In *cis*-eQTL, the allelic expression is regulated in *cis* (Figure 1b). RMAE (Krueger and Morison, 2008; Chess, 2012) was observed in mouse olfactory receptors (Lomvardas *et al.*, 2006), mammalian X-linked genes that undergo random X chromosome inactivation (Clerc and Avner, 2006), allelic exclusion of immunoglobulin genes (Vettermann and Schlissel, 2010), loss of heterozygosity in cancer (Lasko *et al.*, 1991) and about 8% of the mouse autosomal genes are randomly expressed from one of the parental alleles (Gimelbrant *et al.*, 2007; Zwemer *et al.*, 2012). RMAE generally refers to the situation in which individual cells express one allele or the other, and across a population of cells, both alleles are generally expressed. To distinguish between RMAE and parent-of-origin monoallelic expression, adequate sample size is needed to capture representative individuals expressing an identified paternal/maternal allele in the two reciprocal F1 crosses (Figure 1c). Therefore, validation of novel candidate imprinted genes should be performed across a panel of informative individuals to exclude the possibility of RMAE (Wang *et al.*, 2011, 2013b). Consistent parent-of-origin allelic expression in both reciprocal crosses is needed to confirm a newly identified imprinted gene (Figure 1d). Besides the canonical imprinted genes with parent-of-origin expression in all F1 samples tested, a subset of imprinted genes show variable imprinting status, with biallelic expression in some individuals (Wang *et al.*, 2013b). This is different from RMAE because monoallelic expression of the other parental allele is never observed (Figure 1e).

The design of reciprocal crosses of two inbred strains/lines is efficient and has led to the discovery of most imprinted genes in mice. However, in crosses using semi-inbred lines and outbred organisms, including humans, it could be challenging to perform a genome-wide survey for novel imprinted genes. As the SNP positions are not always heterozygous in all genes of the assayed individuals, a larger sample size (at least >30) is needed to achieve genome-wide gene coverage with enough informative individuals. Due to the segregating variation in the genome, the heterozygous SNP positions are not always informative in terms of transmission direction (individual 3 in Figure 1f). To identify novel candidate imprinted genes in humans, the first step is to check informative individuals for transcribed genes and exclude the possibility of RMAE. The next step is to genotype the informative SNP positions in the parents and distinguish the imprinting candidates from a large number of *cis*-eQTL (Montgomery *et al.*, 2010; Majewski and Pastinen, 2011; Becker *et al.*, 2012). One useful method is to search for a 'flipped' allelic expression pattern at SNP positions (Pollard *et al.*, 2008). For example, shown in Figure 1f is the allelic expression profile of three individuals in a maternally expressed imprinted gene. A flipped pattern would be seen when some individuals have higher expression from the reference allele (A allele in Figure 1f), while others have higher expression from the alternative allele (G allele in Figure 1f), so observing a flipped pattern (individuals 1,2 or 1,3 but not 2,3) will exclude the possibility of a strong *cis*-eQTL effect (Figure 1f).

It is also critical to have adequate SNP density in the transcriptomes of the two strains for successful identification of novel imprinted genes in reciprocal F1 individuals. If the SNP density is too low, most genes will contain no exonic SNPs for allelic expression quantification, therefore genome-wide coverage will not be achieved. On the other hand, if the two strains/species are too distantly related, then the F1 hybrids may display dysregulation of the epigenetic profile, resulting in aberrant genomic imprinting (Shi *et al.*, 2005). In addition, high SNP density might affect the proper alignment of the alternative allele-containing reads, resulting in biased estimates of the allele-specific expression ratio, favoring the reference allele. The quality of the alignment will affect the accuracy of the DAE estimation. Insertions and deletions (INDELs), copy number variants

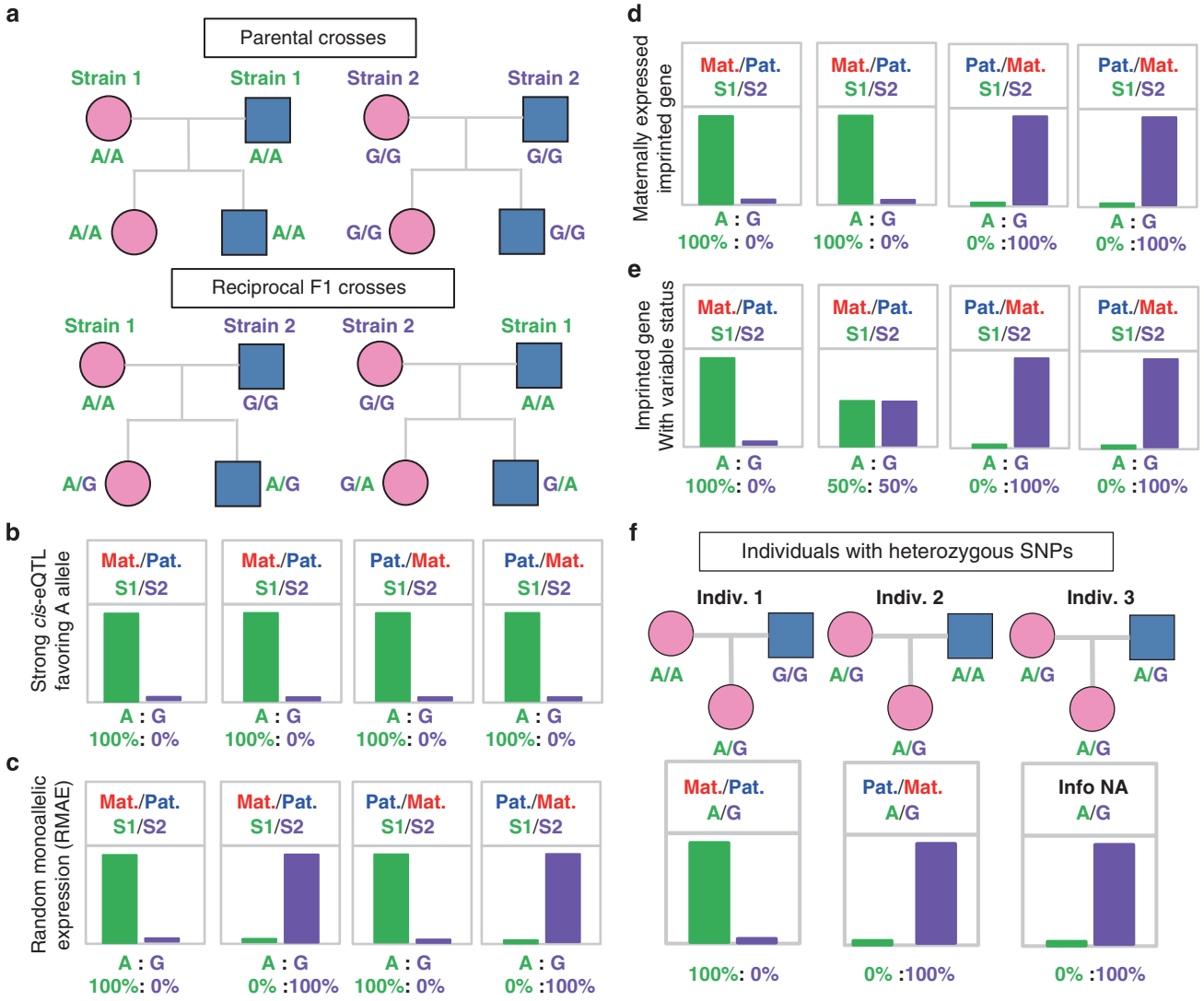

Figure 1 Reciprocal F1s for identification of novel imprinted genes. (**a**) Parental and F1 reciprocal crosses of inbred strains (S1 and S2). An A/G SNP position is shown. (**b**–**e**). Allelic expression profile under a *cis*-eQTL effect (**b**), RMAE (**c**), stable genomic imprinting (**d**) and variable genomic imprinting (**e**). (**f**) Informative SNP position (A/G) in individuals of polymorphic parents.

and unaligned regions between the two strains/species will also cause some problem in read alignment and the allelic expression ratio quantification. For crosses of semi-inbred strains, selecting a reference strain with full genome sequence will help the sequencing alignment step by minimizing the SNP differences between the selected and the reference genome. From our experience of several different species, 0.1–5% sequence divergence is the best for genome-wide survey of imprinted genes in reciprocal crosses.

ESTIMATION OF THE DAE RATIO AT INFORMATIVE SNP POSITIONS AND IDENTIFICATION AND VALIDATION OF NOVEL IMPRINTED GENES

After read alignment to a reference genome (Figure 2a), the DAE ratios are obtained by directly counting the number of reference allele and alternative allele-containing reads (Figure 2b). SNP agreement across the entire transcript should be checked before summarizing the total SNP count by gene (Figure 2c). Inconsistent SNP positions may suggest the presence of alternatively spliced transcript forms with different DAE ratios. Because positions with low read coverage will have large variance in DAE estimation, based on our validation experience, a minimum of $20\sim25\times$ depth is needed to estimate the DAE ratio from a single SNP. If there are multiple informative SNPs within a gene, a depth coverage cutoff of $8\sim10\times$ is used for any particular SNP.

To detect significant parent-of-origin expression in reciprocal F1 crosses of strain/species 1 (S1) and strain/species 2 (S2), we define p_1 as the S1 expression ratio in S1 mother × S2 father cross (maternal ratio) and p_2 as the S1 expression ratio in S1 father × S2 mother cross (paternal ratio). For the majority of the genes in the genome with 50% expression from the mother and 50% from the father, p_1 and p_2 are approximately equal to 0.5; under the case of a *cis*-eQTL effect favoring S1 allele, p_1 and p_2 will both be >0.5; for a paternally expressed imprinted gene, p_1 is <0.5 and p_2 is >0.5 (Figure 2d). Therefore, we could detect significant DAE between the two reciprocal F1 crosses (parent-of-origin effect) by rejecting the null hypothesis $p_1 = p_2$. Chi-squared test or Fisher's exact test could be used, but the Storer–Kim test has more power when the four counts are smaller (Storer and Kim, 1990). But the chi-squared or Storer–Kim test alone should not be relied on, because application of these tests shows a clear inflation of chi-squared statistics, indicating a consistent

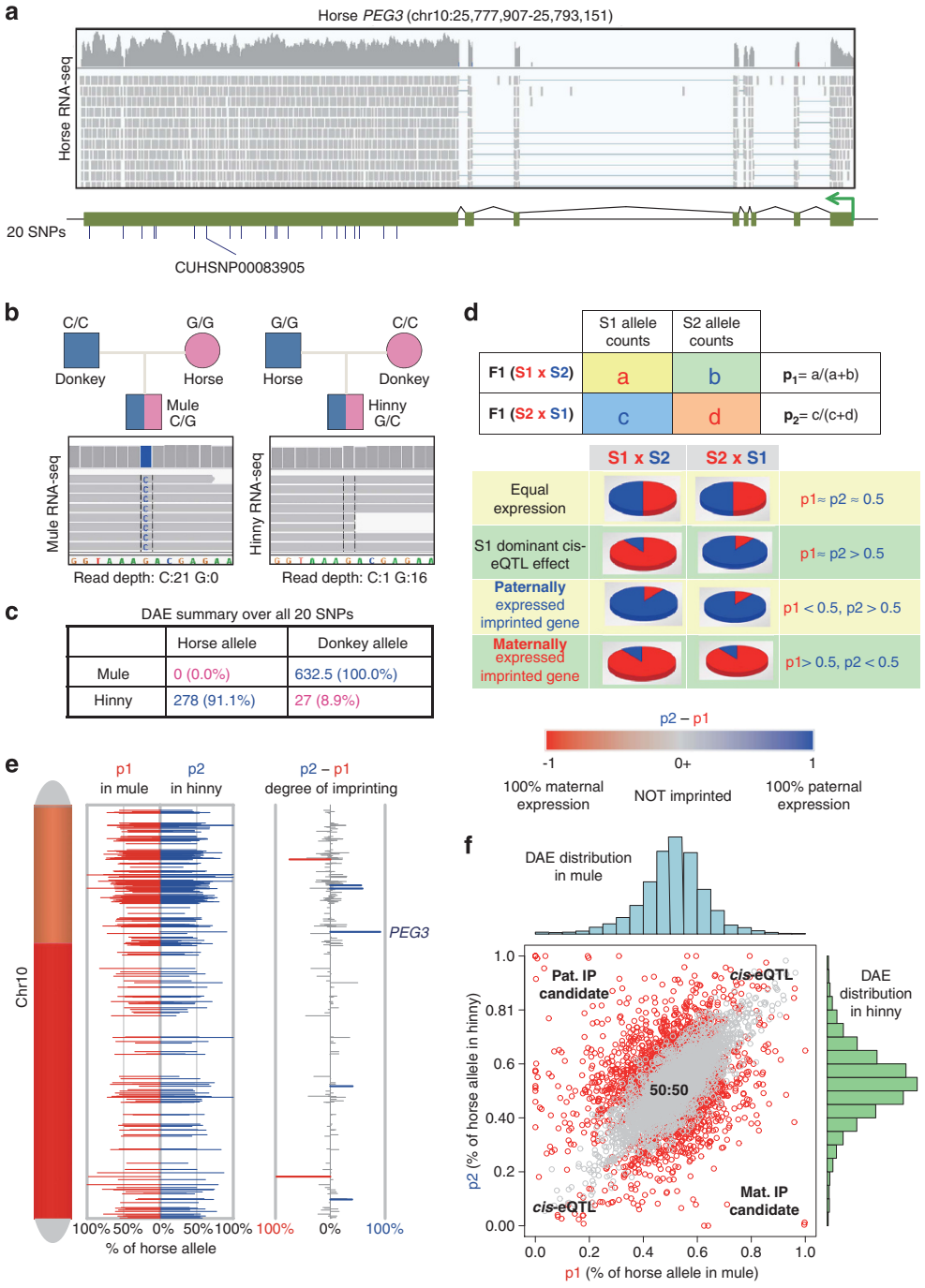

Figure 2 DAE for identification of novel imprinted genes. (**a**) Illumina RNA-seq alignment screenshot from IGV browser (Thorvaldsdottir *et al.*, 2013), showing the exon models for horse *PEG3* gene, a paternally expressed imprinted gene. Twenty informative SNP positions between horse and donkey were plotted as vertical bars under the gene model. (**b**) Estimation of DAE ratio in hybrid equids from RNA-seq data at SNP position CUHSNP00083905. The horse parents have G/G genotype and the donkey parents have C/C genotype. Reciprocal F1 individuals, mule and hinny, are both polymorphic at this position. Paternal expression could be inferred from the RNA-seq reads. (**c**) Summary of DAE for *PEG3* gene across 20 informative SNPs. (**d**) Detection of novel imprinted genes. Top: definition of p_1 and p_2. Bottom: value of p_1 and p_2 under nearly equal allelic expression, *cis*-eQTL effect and genomic imprinting. Red represents maternal expression ratio, and blue stands for paternal expression ratio in the pie chart. The degree of parent-of-origin effect is defined as (p_2-p_1). If a gene has 50%:50% DAE ratios in both crosses, (p_2-p_1) is zero, and this gene is not imprinted. If (p_2-p_1) equals -1, there is 100% maternal expression. If (p_2-p_1) equals $+1$, there is 100% paternal expression. (**e**) Scan of candidate imprinted genes in the genome, modified from (Wang *et al.*, 2013b). Left panel: Horse chromosome 10, color-coded with human chromosome synteny. Middle panel: DAE ratios for genes on chromosome 10 for both mule and hinny placenta samples. The *x* axis is the DAE ratio from the horse allele in the two reciprocal F1 hybrids. The red bar is drawn according to the horse allelic expression ratio in mule (p_1), and the blue bar is the horse allelic expression ratio in hinny (p_2). Right panel: Degree of parent-of-origin effect (p_2-p_1) along chromosome 10. Red: significant overexpression of the maternal allele; blue: significant overexpression of the paternal allele; gray: non-significant genes. (**f**) Plot of joint distribution of p_1 and p_2. The top panel is a histogram of p_1 (blue), and the right panel is a histogram of p_2 (green). In the middle is a scatterplot of the of p_1(on *x* axis) and p_2 (on *y* axis).

overdispersion of the data. The test assumes simple binomial sampling of alleles, and the RNA-seq experiment has several sampling steps and PCR amplification also inflates the variances. Consequently, we call the candidate imprinted genes based on a combination of statistical significance and the degree of allele-specific bias, using an arbitrary cutoff of $p_1 > 0.65$, $p_2 < 0.35$ for maternally expressed candidate imprinted genes and $p_1 < 0.35$, $p_2 > 0.65$ for paternally expressed candidate imprinted genes. This choice is empirical, based on the variance of DAE estimation and the results of many trials of validation methods. Most candidates with DAE ratio between 0.4 and 0.6 will not validate across multiple individuals, even if the initial test of $p_1 = p_2$ is rejected with high statistical confidence. The quantity $(p_2 - p_1)$ was used as an indicator of the degree of parent-of-origin effect to scan for novel candidate imprinted genes across the genome (Figures 2d and e). In the plot of the joint distribution of p_1 and p_2, most genes with a 50:50 expression ratio are in the middle; strong *cis*-eQTL effect genes are in the top-right and bottom-left corners; and paternally and maternally expressed imprinted genes are in the top-left and bottom-right corners, respectively (Figure 2f).

To confirm the newly identified imprinting candidates discovered from RNA-seq data, an independent validation method is needed. For crosses of polymorphic parents or semi-inbred lines (Figure 1f), Sanger sequencing, SNP genotyping arrays or gDNA-seq methods were used to genotype the informative SNPs in the F1 genomic DNA samples to make sure they are heterozygous and determine the transmission direction. Allele-specific quantitative reverse transcriptase–PCR or allele-specific pyrosequencing are common assays to verify the DAE ratios. Direct Sanger sequencing of the F1 cDNA could also be used for validation of all-or-none or nearly 100% imprinting candidates, but they cannot provide accurate allelic expression ratios for partially imprinted genes and sometimes will lead to wrong conclusions. Allele-specific pyrosequencing is an accurate method for DAE ratio quantification using the surrounding nucleotides as internal controls (Wang and Elbein, 2007). The error of DAE ratio estimation ranges from 2% to 5%, which has better resolution than quantitative reverse transcriptase–PCR assay (Singer-Sam and Gao, 2002). Parental samples can be included in verification as controls (Figure 3a). The validation assay is performed in a panel of independent individuals to exclude the possibility of RMAE and variable imprinting status (Figures 1c and e). To confirm the newly identified imprinted genes at the mechanism level, bisulfite sequencing can be used to check the differential allelic DNA methylation at the differentially methylated regions (DMRs) or imprinting control regions (Delaval and Feil, 2004; Figure 3b). Most known DMRs are located at promoter CpG islands or upstream or downstream enhancer regions (Bartolomei and Ferguson-Smith, 2011). As not all known imprinted genes have a characterized DMR, the DMR validation is not required for confirming the imprinted status, but it will provide some insight regarding the regulation of genomic imprinting at this locus.

CHALLENGE 1: DETECTION AND CORRECTION OF REFERENCE GENOME ALIGNMENT BIAS

When aligning the RNA-seq reads from the two reciprocal crosses, in most cases, only one reference genome is available for an organism, resulting in a bias toward the reference allele. In the F1 individuals from crosses of two species or two strains, using a cutoff of a fixed number of mismatches, sequencing reads coming from the parental strain that is more closely related to the reference genome will be over-represented, because some reads from the divergent strain will be rejected by the mismatch cutoff (Figure 4a). We could quantify the global alignment bias by the average percentage of reference allele across all informative SNP positions (Figure 4b). The degree of the bias depends on the SNP density between the two parental strains and the sequence alignment parameters. In some inter-specific hybrids between two *Nasonia* species, we have observed >70% alignment bias in divergent regions in the genome (Figure 4a).

The alignment bias caused by the distance to the reference genome will have a significant effect on the estimation of DAE ratios. The high fidelity of read mapping of the reference-matching allele, and greater chance of the non-reference allele to fail to map to the reference, produces this bias. The DAE ratios will shift toward the reference allele (Figure 4b), and the effect is more severe in regions with high SNP density, leading to spurious parent-of-origin effects. There are two methods for correcting the reference alignment bias. The first is to apply different mismatch cutoffs for the reference and alternative allele-containing reads (Wang *et al.*, 2008). This method requires the tracking of all SNP alleles in each read and is not applicable when a single read contains the reference allele for one SNP and the alternative allele for another SNP. The best way that we have found to remove the alignment bias is to align the reads to both the reference genome and a pseudogenome, constructed by substituting the reference allele with the alternative allele in all transcribed regions, and take the average counts from the reference and pseudogenome (Wang *et al.*, 2011, 2013b). With this approach, the alignment bias could be reduced to <1% (Figure 4b). It is not recommended to use species/strains with an exonic SNP density>10%, because one strain might not be mappable to a single reference genome.

CHALLENGE 2: RNA-SEQ LIBRARY COMPLEXITY

Studies in the literature using RNA-seq to identify novel imprinted genes in mouse have revealed dramatically different profiles (Gregg *et al.*, 2010b; DeVeale *et al.*, 2012), and in some studies the validation rate is extremely low even for the top candidate genes (Henckel and Arnaud, 2010). Although the counts of reads of the two alleles have several sources of variation, we think the main reason for the discrepancy is due to the problem of insufficient library complexity, as explained below.

Under ideal conditions, the read count distribution at any given SNP position is expected to be binomial. However, in practice the distribution across replicates of read counts has a variance much greater than binomial, and the sources of the inflated error enter at all stages, from the biological material collection and extraction to library preparation, to sequencing and read mapping. For lowly expressed genes with only a few mRNA copies in the transcriptome, only one of the two alleles could get randomly ligated to the adapter and be included in the final pool just by chance, due to a series of sampling steps during RNA-seq library construction (Figure 5a). After PCR amplification and sequencing, the data would suggest a monoallelically expressed gene, when in fact it is not (Figure 5a). The spurious monoallelic expression in both reciprocal F1 crosses may result in false positives of candidate imprinted genes. One of the two alleles is less likely to get lost by chance for moderately expressed genes, but the allele-specific expression ratio could shift. Biased DAE ratios in opposite directions in the reciprocal F1s could also cause a spurious parent-of-origin effect (Figure 5b). This library complexity problem will occur when the number of molecules in any step during the library preparation is small, and it is not easily noticed because the final concentration is not low after the PCR amplification step. The DAE ratios could not be accurately quantified for lowly and moderately expressed genes. However, the total expression level estimation will not be affected as much, and there may still be a

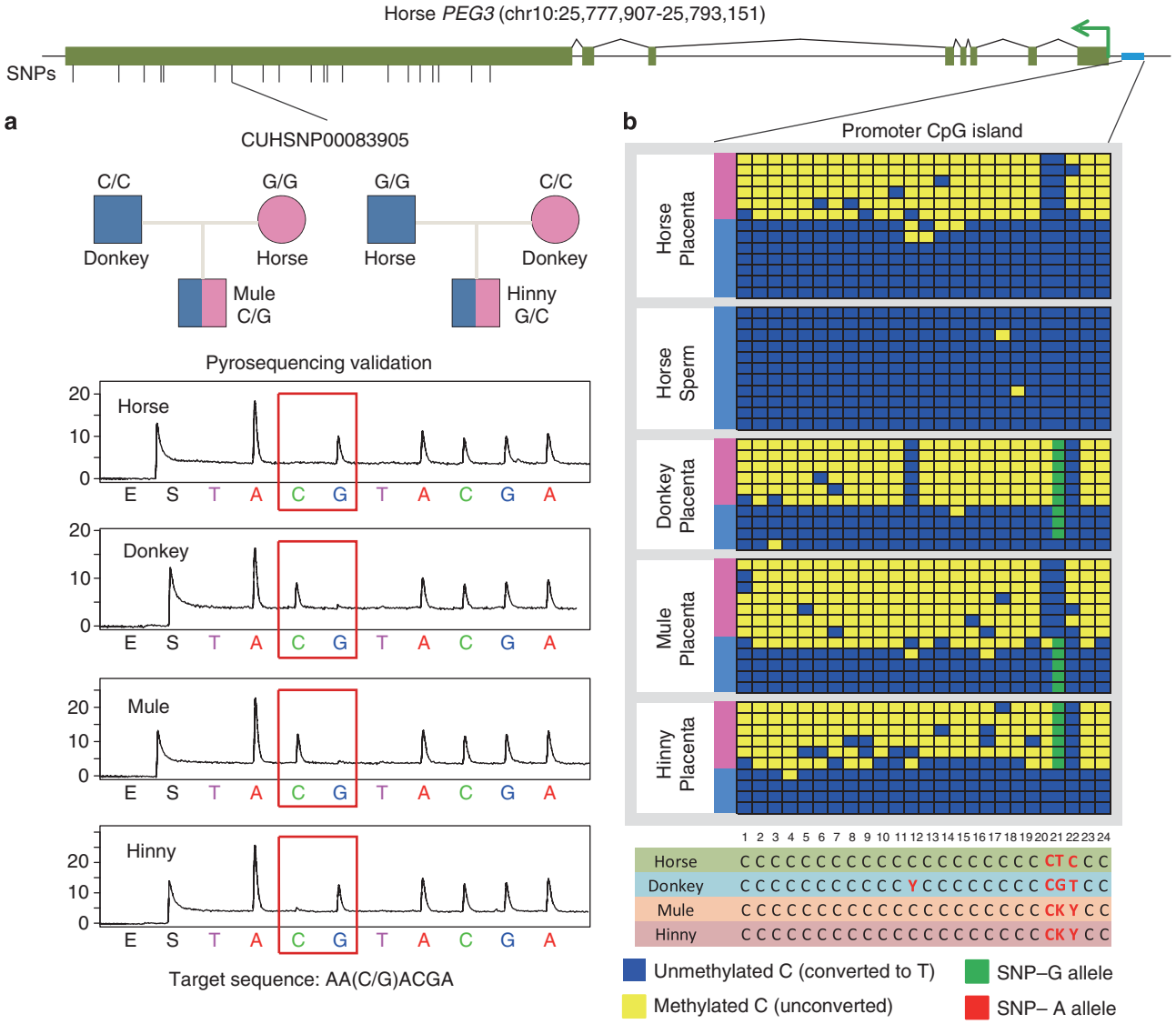

Figure 3 Pyrosequencing and bisulfite sequencing validation of imprinted gene *PEG3*. (**a**) Allele-specific pyrosequencing validation of the paternal expression of *PEG3* in mule and hinny placental samples, modified from Wang *et al.* (2013b). (**b**) Allele-specific differential methylation in the DMR at *PEG3* promoter CpG island from bisulfite sequencing is corresponding to the allelic expression bias. The *PEG3* DMR is differentially methylated with maternal methylation in horse, donkey, mule and hinny placental samples and is unmethylated in sperm, consistent with paternal expression. Parental allele-specific methylation in mule and hinny is inferred from horse–donkey single-nucleotide sequence differences (positions 21 and 22). Yellow boxes depict methylated CpGs, and blue boxes are unmethylated CpGs. Modified from Wang *et al.* (2013b).

good correlation of total expression level across replicates. This makes the library complexity problem easily overlooked.

The statistical distribution for the DAE ratios with read depth n can be modeled as beta-binomial distribution, Beta-binom (N, α, β), which is a compound distribution where p, the binomial parameter of the binomial distribution $X \sim \text{Binom}(N, p)$, is a random variable drawn from a beta distribution $p \sim \text{Beta}(\alpha, \beta)$. The mean of p can be calculated as $\mu = \alpha/(\alpha+\beta)$, and the variance is $\mu(1-\mu)(1+(N-1)\rho)$. In this case, $\rho = 1/(1+\alpha+\beta)$ is known as the overdispersion parameter. A value of $\rho = 1$ gives the usual binomial distribution. Higher values of ρ indicate overdispersion, and in this case the overdispersion can arise from insufficient library complexity. We simulated DAE ratio distributions for high, medium and low library complexity by using different overdispersion parameters in single or mixture beta-binomial distributions and found that the overdispersion parameter ρ could be used as a diagnostic check for library complexity (Figure 5c). When we compared RNA-seq data from reciprocal crosses of mouse inbred strains AKR and PWD generated in 2009 (Wang *et al.*, 2011) and the RNA-seq data from libraries made with the same total RNA samples in 2011 (unpublished), we found the library complexity is much better using the Illumina TruSeq (Illumina, Inc., San Diego, CA, USA) protocol compared with the older Illumina library preparation kit (Figure 5d).

As adequate library complexity is critical for quantitative RNA-seq analysis for allelic-specific expression, here we discuss the following aspects to improve the library complexity and ensure proper data analysis.

(1) Starting with more input RNA

Although the total RNA input for RNA-seq could be as low as 100 ng, based on our experience, if possible, having $\geqslant 2\,\mu g$ input total RNA will help solve potential library complexity problem by having more starting mRNA molecules.

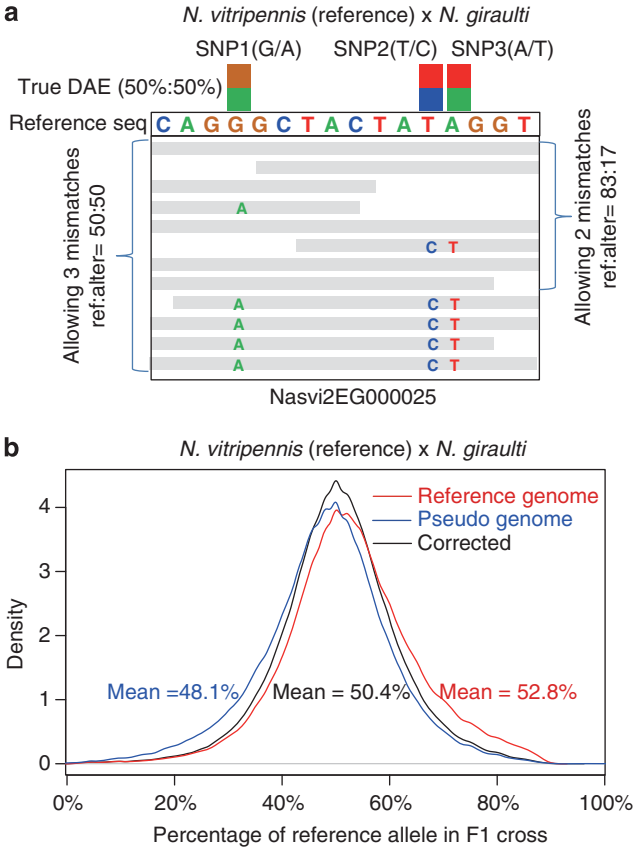

Figure 4 Detection and correction of the reference genome alignment bias. (**a**) A cartoon figure illustrating the reference genome alignment bias in crosses of two *Nasonia* species (*N. vitripennis* and *N. girault*). Shown in the aligned reads at three SNP positions (G/A, T/C and A/T) in the exonic region of gene Nasvi2EG000025. When allowing three mismatches during the sequence alignment in this region, all reads will be aligned, and the true DAE ratio (50:50) will be correctly estimated. If only two mismatches were allowed, the DAE ratio estimated will be biased toward the reference allele. (**b**) Density plot of the reference allelic expression percentage distribution for alignment to the reference genome (red), pseudogenome (blue) and corrected percentages by averaging the two (black).

(2) Biological and technical replicates

Independent biological replication in each of the two reciprocal crosses and/or different technique replicates for a single sample at library preparation level will help identify any library complexity problems. Deeper sequencing and pooling samples before library preparation cannot solve this problem. Gregg *et al.* (2010b) found >1300 novel imprinted genes in mouse brain, but other researchers only discovered a limited number of them, although the study by Gregg *et al.* (2010b) did use an unusual degree of identification and separation of many brain sub-regions. Given the very small size of these sub-regions, and the fact that pooled samples from four different F1 embryos were used for library preparation in Gregg *et al.* (2010b), it seems likely that these data could be subject to the library complexity problem.

(3) Detecting library complexity problems

To accurately estimate DAE in RNA-seq experiments, library complexity needs to be checked. The proportion of counts for one parental allele versus the other can be plotted as a frequency distribution and then fitted by beta-binomial distribution as described above. A higher value of the overdispersion parameter ρ (>0.05 based on our experience) can be used as an indicator of poor library complexity. Another method for characterizing the molecular complexity of sequencing library is an empirical Bayesian method implemented in preseq software (Daley and Smith, 2013).

(4) Independent single-gene validation in multiple individuals

Most of the earlier RNA-seq studies lack biological or technical replicates because of cost, making the validation across multiple individual samples absolutely essential to confirm the newly identified candidate imprinted genes from RNA-seq data. As we mentioned earlier, verification in multiple individuals also helps exclude the problem seen in the study by Gregg *et al.* (2010b), where the majority of the 1300 novel imprinted genes were not confirmed by an independent validation method.

(5) Candidate genes with multiple independent SNP support

For RNA-seq data with moderate library complexity, restricting the list of candidates to those with multiple independent SNP support will significantly reduce the false positive rate (Wang *et al.*, 2011). We define independent SNPs as informative SNPs separated by a distance greater than the read length so that each observation must be an independent sequence read. Because each SNP is an independent sampling trial, a significant candidate gene with multiple independent SNP support has a higher rate of successful verification.

(6) Calling novel imprinted genes based on validation rate

As the library complexity problem can result in numerous false positives, one solution is to determine an empirical statistical cutoff tuned for each study based on the validation rate using an independent single-gene method. In an acceptable study, when the candidate imprinted genes are rank ordered according to the *P*-values or *q*-values, genes on top of the list should be verifiable, and the significant candidate list should stop when the validation rate drops to an unacceptable level. This empirical validation approach is generally more conservative (and more reliable) than a strict statistical cutoff. In a recent RNA-seq data study from Okae *et al.* (2012), the researchers reported 133 paternally and 955 maternally expressed candidate imprinted gene in the mouse placenta. Only 1/6 paternal and 1/269 maternal candidates could be verified. Their conclusion is that there is high false positive rate for allele-specific expression inference from RNA-seq data (Okae *et al.*, 2012). We agree that the false positive rate is extremely high in this particular study, but this does not mean this is true in every study using RNA-seq. In fact, the validation rate of this study (2/275 = 0.7%) is even lower than the 1–2% of the genome that is estimated to be imprinted. The number of confirmed genes (2) is actually less than the expected number of imprinting genes by random sampling of 275 genes from the genome. It seems highly likely that this study suffers from a severe library complexity problem, which causes this high false positive rate. From our analysis pipeline, following the rules we describe above to assure good library complexity, the top 40 genes that we called were successfully validated using allele-specific pyrosequencing (Wang *et al.*, 2013b). In short, DAE ratios can be accurately estimated from RNA-seq data when the experiment is carefully designed and the data analysis is properly done.

CHALLENGE 3: FILTERING OF PROBLEMATIC SNP POSITIONS

Even after all the major sources of variation are considered and controlled, calls of strongly skewed DAE ratios from RNA-seq are not always confirmed by allele-specific pyrosequencing. Highly significant genes with moderate or high expression level and multiple SNP

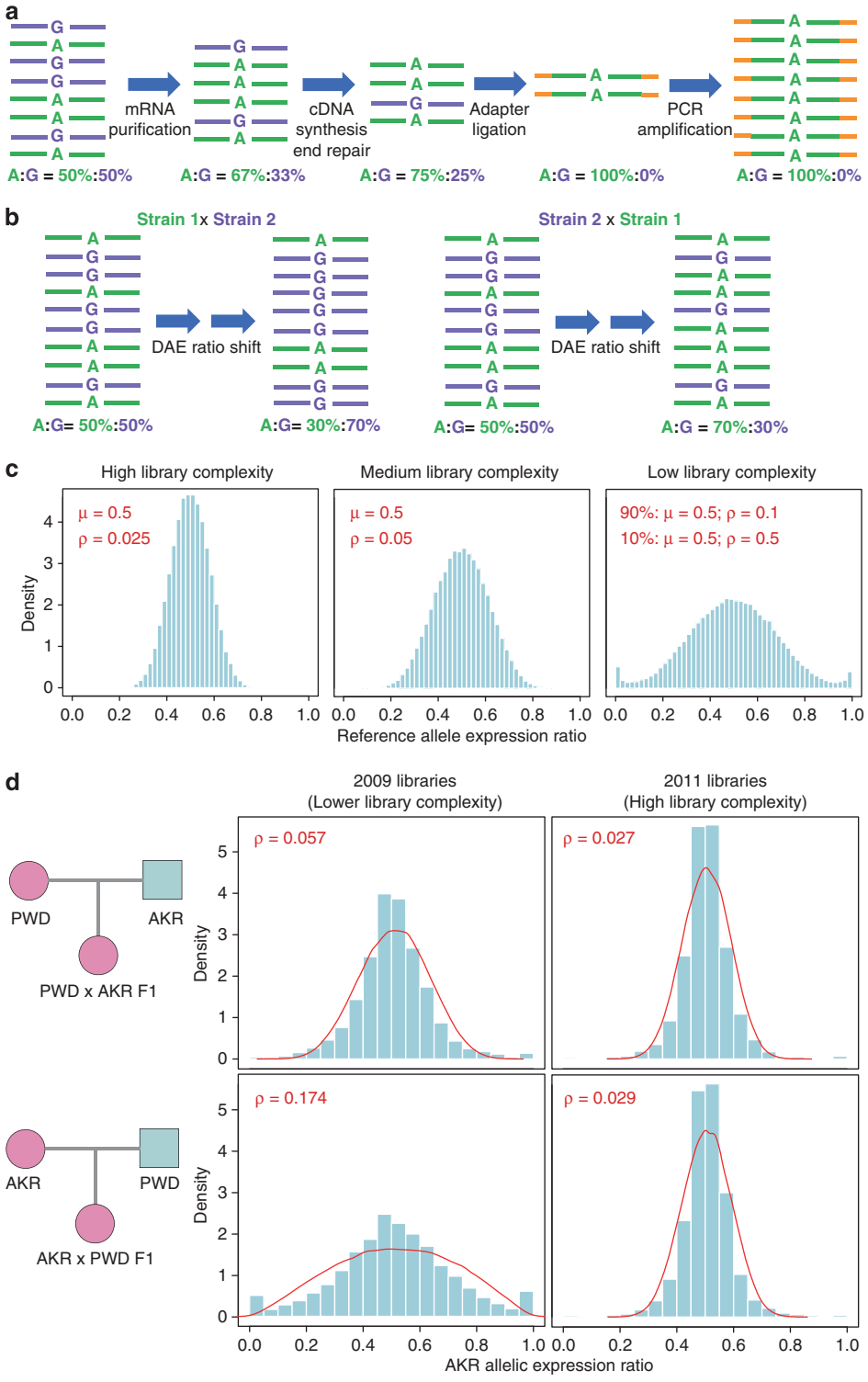

Figure 5 Low library complexity will result in false positives in the identification of novel imprinted genes. (**a**) A cartoon figure for the decrease of library complexity due to a series of sampling effect during library preparation for lowly expressed genes. (**b**) A cartoon figure for the shift of allelic expression ratio during library preparation for moderately expressed genes. (**c**) Simulated distribution of allelic expression ratio under high, medium and low library complexity using single or mixed beta-binomial distributions. (**d**) Histograms for distributions of AKR allelic expression ratio in E17.5 placental samples from reciprocal crosses of AKR and PWD inbred mouse strains. Top panels: AKR father × PWD mother crosses; bottom panels: PWD father × AKR mother crosses. The left panels are from RNA-seq data prepped and sequenced in 2009 (Wang *et al.*, 2011), and the right panels are from RNA-seq data of the same samples prepped and sequenced in 2011. The red curve is plotted with the fitted parameters of beta-binomial distribution.

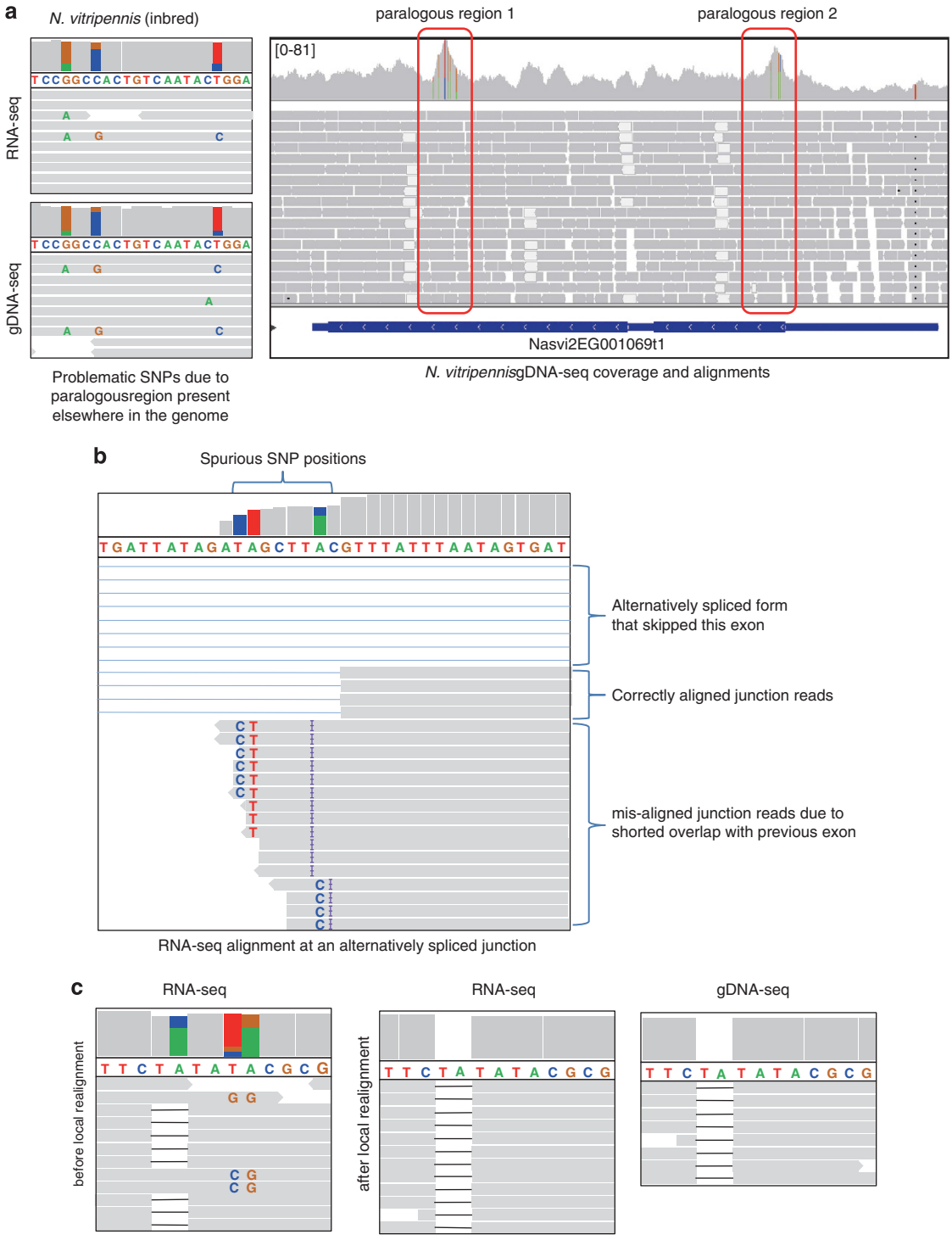

Figure 6 Problematic SNP positions in RNA-seq data. (**a**) Problematic SNP positions due to the presence of paralogous regions in the genome. The top-left panel is RNA-seq read alignments from an inbred jewel wasp species (*Nasonia vitripennis*), showing three potential SNP positions. The bottom-left panel is the gDNA-seq data from the same species, suggesting this region has paralogous sequences elsewhere in the genome. Shown in the right panel is the gDNA-seq coverage and alignments for a *Nasonia* gene. The two paralogous regions in the red boxes have elevated coverage and polymorphic positions within them. (**b**) Spurious SNP positions caused by mis-alignment of the exon–exon junction reads. (**c**) Local realignment over INDEL position could eliminate the problematic SNPs near INDEL positions.

support of the same direction have a higher rate of successful verification. When checking the SNP consistency within a gene, most of the inconsistent SNPs are due to the presence of the following categories of problematic SNPs.

(1) Some SNP positions are actually non-polymorphic between the two parental strains (reference genome sequenced strain 1 and another strain 2). They seem to be polymorphic due to sequencing errors in the reference genome if the SNP is

called by comparing gDNA-seq from strain 2 to the reference genome.
(2) RNA editing or partial RNA editing could also result in problematic SNPs.
(3) Some SNPs are located in repetitive/paralogous regions (including copy number variants) that are only present once in the reference genome, so they could not be filtered out using uniquely mapped reads.
(4) SNPs near insertion/deletion events might be problematic due to the mis-alignment over INDELs.
(5) Junction SNPs that are near the exon–intron boundaries, where some junction spanning reads were not properly aligned.
(6) At each SNP in double-stranded RNA-seq data, sequencing reads are equally likely to come from the sense and antisense direction. SNPs with strong strand bias have a low validation rate.

These problematic SNPs are generally rare, but they are highly enriched in the candidate imprinted genes identified by searching for SNPs with significantly skewed DAE ratios. They need to be filtered out to reduce the high false positive rate. Having RNA-seq and gDNA-seq data from inbred parental crosses (Figure 1a) will help filter out the above classes of problematic SNPs. We propose the following filters to eliminate the effect of problematic SNPs in the estimation of DAE ratios:

(1) Compare the gDNA-seq data from the reference strain to the reference genome, and correct all positions with sequencing errors in the reference genome.
(2) Search for different base positions between RNA-seq and gDNA-seq data in inbred parental strains to find the edited positions and exclude them from the analysis.
(3) Check for polymorphic positions in gDNA-seq data and scanning for regions with elevated gDNA-seq coverage in inbred parental strains (Figure 6a).
(4) Perform local realignment over INDEL positions using software such as GATK (McKenna *et al.*, 2010; Figure 6b).
(5) Remove SNPs near the exon–intron boundaries with significant parent-of-original effect and check for mis-alignments (Figure 6c).
(6) Discard SNPs with significant strand bias as quantified by the relative ratio of sense and antisense reads for reference and alternative alleles.

The above problematic SNPs should be removed before estimating DAE ratios from RNA-seq data. After finding SNPs that show statistical support for imprinting, it is important to double check each for the above issues in order to detect additional false positive calls.

CHALLENGE 4: MATERNAL CONTAMINATION IN PLACENTAL TISSUES

In placental tissues, the maternal contamination during post-implantation sample collection is inevitable for species with hemochorial placentation, such as human and mouse (Proudhon and Bourc'his, 2010). Such maternal contamination will make genes expressed in the uterus but not fetal placenta resemble maternally expressed imprinted genes, resulting in over-calling of maternally expressed imprinted genes. Computational prediction of imprinted genes remains notoriously challenging, and one effort called 10 novel imprinted genes in the mouse placenta, all with maternal expression (Brideau *et al.*, 2010). However, a subsequent study showed that 4 of the 10 genes are likely due to maternal contamination (Okae *et al.*, 2012). Therefore, it is critical to detect and minimize the possibility for maternal contamination for the identification of novel imprinted genes. One method to avoid this is to use placental tissues from pre-implantation embryos that are free of maternal contamination (Wang *et al.*, 2013b). If this is not possible, special dissection techniques need to be used to minimize potential maternal contamination. The degree of contamination could be quantified by measuring the expression level of uterus-specific genes, and samples with the least contamination could be selected for RNA-seq experiments. As maternal contamination is a genome-wide effect, after RNA-seq the degree of contamination could be quantified by the global maternal bias in DAE ratios for autosomal genes (Wang *et al.*, 2011). For the candidate imprinted genes identified from placenta, the expression level in uterus samples near the placenta need to be checked. Highly expressed genes in uterus are more likely to bias the gene toward preferential maternal expression.

CONCLUSION

The simple logic of scoring differential allele-specific expression in the offspring of reciprocal crosses to identify parent-of-origin effects has strong appeal. However, blind application of the method without considerable care at several key steps in the library construction and bioinformatics is bound to yield an inordinate number of false positive calls. This paper shares our insights from experience in applying this approach to samples from mouse, mule/hinny, opossum, honeybee and Nasonia. We emphasize the need to maintain library complexity, an issue that is very easy to miss until the final bioinformatics are done. With a low complexity library, the sampling variance grows to the point that many genes spuriously reject the null hypothesis of no parent-of-origin effects. Many problems may also occur at the bioinformatics steps, including biased mapping of reads to the reference allele. We stress the need for independent confirmation of imprinting status with replicate biological samples and an orthogonal technology. Expanded application of these methods, with careful validation, will eventually yield data that will provide a solid foundation on which to build well-supported models for the evolution of genomic imprinting, a better understanding of the diverse molecular mechanisms by which parent-of-origin expression is regulated and aspects of the fitness consequences of disruptions in genomic imprinting.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

AGC is supported by NIH Grants R01 GM64590, R01 AI064950 and R01 HD059060. We thank collaborators Douglas Antczak (Cornell), Susan Lamont (Iowa), Mark Roberson (Cornell), Gene Robinson (Illinois), Paul Samollow (Texas A&M), Paul Soloway (Cornell) and John Werren (University of Rochester) for their enthusiasm in helping us develop the methods reported here in various biological systems.

Babak T, Deveale B, Armour C, Raymond C, Cleary MA, van der Kooy D *et al.* (2008). Global survey of genomic imprinting by transcriptome sequencing. *Curr Biol* **18**: 1735–1741.
Barlow DP (2011). Genomic imprinting: a mammalian epigenetic discovery model. *Annu Rev Genet* **45**: 379–403.

Bartolomei MS, Ferguson-Smith AC (2011). Mammalian Genomic Imprinting. *Cold Spring Harb Perspect Biol* **3**: a002592.

Becker J, Wendland JR, Haenisch B, Nothen MM, Schumacher J (2012). A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals. *Eur J Hum Genet* **20**: 97–101.

Brideau CM, Eilertson KE, Hagarman JA, Bustamante CD, Soloway PD (2010). Successful computational prediction of novel imprinted genes from epigenomic features. *Mol Cell Biol* **30**: 3357–3370.

Chess A (2012). Mechanisms and consequences of widespread random monoallelic expression. *Nat Rev Genet* **13**: 421–428.

Chess A, Simon I, Cedar H, Axel R (1994). Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78**: 823–834.

Choi JD, Underkoffler LA, Collins JN, Marchegiani SM, Terry NA, Beechey CV *et al.* (2001). Microarray expression profiling of tissues from mice with uniparental duplications of chromosomes 7 and 11 to identify imprinted genes. *Mamm Genome* **12**: 758–764.

Choi JD, Underkoffler LA, Wood AJ, Collins JN, Williams PT, Golden JA *et al.* (2005). A novel variant of Inpp5f is imprinted in brain, and its expression is correlated with differential methylation of an internal CpG island. *Mol Cell Biol* **25**: 5514–5522.

Clerc P, Avner P (2006). Random X-chromosome inactivation: skewing lessons for mice and men. *Curr Opin Genet Dev* **16**: 246–253.

Daley T, Smith AD (2013). Predicting the molecular complexity of sequencing libraries. *Nat Methods* **10**: 325–327.

de la Casa-Esperon E (2012). Nonmammalian parent-of-origin effects. *Methods Mol Biol* **925**: 277–294.

Delaval K, Feil R (2004). Epigenetic regulation of mammalian genomic imprinting. *Curr Opin Genet Dev* **14**: 188–195.

DeVeale B, van der Kooy D, Babak T (2012). Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genet* **8**: e1002600.

Dindot SV, Kent KC, Evers B, Loskutoff N, Womack J, Piedrahita JA (2004). Conservation of genomic imprinting at the XIST, IGF2, and GTL2 loci in the bovine. *Mamm Genome* **15**: 966–974.

Gilad Y, Rifkin SA, Pritchard JK (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* **24**: 408–415.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A (2007). Widespread monoallelic expression on human autosomes. *Science* **318**: 1136–1140.

Gregg C, Zhang J, Butler JE, Haig D, Dulac C (2010a). Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329**: 682–685.

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

Heard E, Clerc P, Avner P (1997). X-chromosome inactivation in mammals. *Annu Rev Genet* **31**: 571–610.

Henckel A, Arnaud P (2010). Genome-wide identification of new imprinted genes. *Brief Funct Genomics* **9**: 304–314.

Huynh KD, Lee JT (2001). Imprinted X inactivation in eutherians: a model of gametic execution and zygotic relaxation. *Curr Opin Cell Biol* **13**: 690–697.

Ke X, Thomas NS, Robinson DO, Collins A (2002). A novel approach for identifying candidate imprinted genes through sequence analysis of imprinted and control genes. *Hum Genet* **111**: 511–520.

Kelsey G, Bartolomei MS (2012). Imprinted genes... and the number is? *PLoS Genet* **8**: e1002601.

Keverne EB (2013). Importance of the matriline for genomic imprinting, brain development and behaviour. *Philos Trans R Soc Lond B Biol Sci* **368**: 20110327.

Kohler C, Weinhofer-Molisch I (2010). Mechanisms and evolution of genomic imprinting in plants. *Heredity* **105**: 57–63.

Krueger C, Morison IM (2008). Random monoallelic expression: making a choice. *Trends Genet* **24**: 257–259.

Kuzmin A, Han Z, Golding MC, Mann MR, Latham KE, Varmuza S (2008). The PcG gene Sfmbt2 is paternally expressed in extraembryonic tissues. *Gene Expr Patterns* **8**: 107–116.

Lasko D, Cavenee W, Nordenskjold M (1991). Loss of constitutional heterozygosity in human cancer. *Annu Rev Genet* **25**: 281–314.

Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R (2006). Interchromosomal interactions and olfactory receptor choice. *Cell* **126**: 403–413.

Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ (2007). Computational and experimental identification of novel human imprinted genes. *Genome Res* **17**: 1723–1730.

Luedi PP, Hartemink AJ, Jirtle RL (2005). Genome-wide prediction of imprinted murine genes. *Genome Res* **15**: 875–884.

Maeda N, Hayashizaki Y (2006). Genome-wide survey of imprinted genes. *Cytogenet Genome Res* **113**: 144–152.

Majewski J, Pastinen T (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**: 72–79.

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

Mizuno Y, Sotomaru Y, Katsuzawa Y, Kono T, Meguro M, Oshimura M *et al.* (2002). Asb4, Ata3, and Dcn are novel imprinted genes identified by high-throughput screening using RIKEN cDNA microarray. *Biochem Biophys Res Commun* **290**: 1499–1505.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J *et al.* (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**: 773–777.

Morcos L, Ge B, Koka V, Lam KC, Pokholok DK, Gunderson KL *et al.* (2011). Genome-wide assessment of imprinted expression in human cells. *Genome Biol* **12**: R25.

Morison IM, Paton CJ, Cleverley SD (2001). The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res* **29**: 275–276.

Morison IM, Ramsay JP, Spencer HG (2005). A census of mammalian imprinting. *Trends Genet* **21**: 457–465.

Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M *et al.* (2003). Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res* **13**: 1402–1409.

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

Ozsolak F, Milos PM (2011). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**: 87–98.

Pask A (2012). Insights on imprinting from beyond mice and men. *Methods Mol Biol* **925**: 263–275.

Pastinen T, Hudson TJ (2004). Cis-acting regulatory variation in the human genome. *Science* **306**: 647–650.

Pollard KS, Serre D, Wang X, Tao H, Grundberg E, Hudson TJ *et al.* (2008). A genome-wide approach to identifying novel-imprinted genes. *Hum Genet* **122**: 625–634.

Prickett AR, Oakey RJ (2012). A survey of tissue-specific genomic imprinting in mammals. *Mol Genet Genomics* **287**: 621–630.

Proudhon C, Bourc'his D (2010). Identification and resolution of artifacts in the interpretation of imprinted gene expression. *Brief Funct Genomics* **9**: 374–384.

Reik W, Walter J (2001). Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**: 21–32.

Renfree MB, Suzuki S, Kaneko-Ishino T (2013). The origin and evolution of genomic imprinting and viviparity in mammals. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120151.

Schulz R, Menheniott TR, Woodfine K, Wood AJ, Choi JD, Oakey RJ (2006). Chromosome-wide identification of novel imprinted genes using microarrays and uniparental disomies. *Nucleic Acids Res* **34**: e88.

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E *et al.* (2008). Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet* **4**: e1000006.

Shi W, Krella A, Orth A, Yu Y, Fundele R (2005). Widespread disruption of genomic imprinting in adult interspecies mouse (Mus) hybrids. *Genesis* **43**: 100–108.

Singer-Sam J, Gao C (2002). Quantitative RT-PCR-based analysis of allele-specific gene expression. In: Ward A (ed). *Genomic Imprinting* vol. 181. Humana Press: New York, NY, USA, pp 145–152.

Sritanaudomchai H, Ma H, Clepper L, Gokhale S, Bogan R, Hennebold J *et al.* (2010). Discovery of a novel imprinted gene by transcriptional analysis of parthenogenetic embryonic stem cells. *Hum Reprod* **25**: 1927–1941.

Storer BE, Kim C (1990). Exact properties of some exact test statistics for comparing 2 binomial proportions. *J Am Stat Assoc* **85**: 146–155.

Thiagalingam S, Laken S, Willson JK, Markowitz SD, Kinzler KW, Vogelstein B *et al.* (2001). Mechanisms underlying losses of heterozygosity in human colorectal cancers. *Proc Natl Acad Sci USA* **98**: 2698–2702.

Thorvaldsdottir H, Robinson JT, Mesirov JP (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178–192.

Vettermann C, Schlissel MS (2010). Allelic exclusion of immunoglobulin genes: models and mechanisms. *Immunol Rev* **237**: 22–42.

Wake N, Takagi N, Sasaki M (1976). Non-random inactivation of X chromosome in the rat yolk sac. *Nature* **262**: 580–581.

Wang H, Elbein S (2007). Detection of allelic imbalance in gene expression using pyrosequencing. In: Marsh S (ed). *Pyrosequencing Protocols*. Humana Press: New York, NY USA, pp 157–175.

Wang X, Douglas KC, Vandeberg JL, Clark A, Samollow PB (2013a). Chromosome-wide profiling of X-chromosome inactivation and epigenetic states in fetal brain and placenta of the opossum, *Monodelphis domestica*. *Genome Res* **24**: 70–83.

Wang X, Miller DC, Harman R, Antczak DF, Clark AG (2013b). Paternally expressed genes predominate in the placenta. *Proc Natl Acad Sci USA* **110**: 10705–10710.

Wang X, Soloway PD, Clark AG (2011). A survey for novel imprinted genes in the mouse placenta by mRNA-seq. *Genetics* **189**: 109–122.

Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG (2008). Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One* **3**: e3839.

Xue F, Tian XC, Du F, Kubota C, Taneja M, Dinnyes A *et al.* (2002). Aberrant patterns of X chromosome inactivation in bovine clones. *Nat Genet* **31**: 216–220.

Yang HH, Hu Y, Edmonson M, Buetow K, Lee MP (2003). Computation method to identify differential allelic gene expression and novel imprinted genes. *Bioinformatics* **19**: 952–955.

Zwemer LM, Zak A, Thompson BR, Kirby A, Daly MJ, Chess A *et al.* (2012). Autosomal monoallelic expression in the mouse. *Genome Biol* **13**: R10.