BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 8 2012, pages 1102–1108
doi:10.1093/bioinformatics/bts089

Gene expression

Advance Access publication February 21, 2012

Allele-specific expression analysis methods for high-density SNP microarray data

Ruijie Liu[1], Ana-Teresa Maia[2], Roslin Russell[2], Carlos Caldas[2], Bruce A. Ponder[2] and Matthew E. Ritchie[1,3,*]

[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [2]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK and [3]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia

Associate Editor: Trey Ideker

ABSTRACT

Motivation: In the past decade, a number of technologies to quantify allele-specific expression (ASE) in a genome-wide manner have become available to researchers. We investigate the application of single-nucleotide polymorphism (SNP) microarrays to this task, exploring data obtained from both cell lines and primary tissue for which both RNA and DNA profiles are available.

Results: We analyze data from two experiments that make use of high-density Illumina Infinium II genotyping arrays to measure ASE. We first preprocess each data set, which involves removal of outlier samples, careful normalization and a two-step filtering procedure to remove SNPs that show no evidence of expression in the samples being analyzed and calls that are clear genotyping errors. We then compare three different tests for detecting ASE, one of which has been previously published and two novel approaches. These tests vary at the level at which they operate (per SNP per individual or per SNP) and in the input data they require. Using SNPs from imprinted genes as true positives for ASE, we observe varying sensitivity for the different testing procedures that improves with increasing sample size. Methods that rely on RNA signal alone were found to perform best across a range of metrics. The top ranked SNPs recovered by all methods appear to be reasonable candidates for ASE.

Availability and implementation: Analysis was carried out in R (http://www.R-project.org/) using existing functions.

Contact: mritchie@wehi.edu.au.

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

Received on November 3, 2011; revised on January 23, 2012; accepted on February 15, 2012

1 INTRODUCTION

Understanding the genetic basis of variation in gene expression is of considerable interest in biomedical research. The preferential expression of one of the two alleles of a gene, known as allele-specific expression (ASE) is one important class of such variation. Allelic expression imbalances have been widely studied in the context of development, where key epigenetic mechanisms such as X-inactivation and genomic imprinting lead to the silencing of one allele (Knight, 2004). Recent studies have also linked ASE to the susceptibility of a number of human diseases (Feng *et al*., 2006; Maia *et al*., 2009; Meyer *et al*., 2008; van Bilsen *et al*., 2008).

Genome-wide assessment of ASE has recently become feasible with the availability of high-density single-nucleotide polymorphism (SNP) chips and second-generation sequencing. A number of microarray-based studies which quantify ASE in a high-throughput way have been published in the past decade (Bjornsson *et al*., 2008; Daelemans *et al*., 2010; Ge *et al*., 2009; Gimelbrant *et al*., 2007; Lee, 2005; Lo *et al*., 2003; Pant *et al*., 2006; Serre *et al*., 2008; Tan *et al*., 2008). With microarrays, oligonucleotide probes that distinguish between the signal from allele A and allele B for SNPs in genomic DNA can measure the relative expression from each allele when mRNA (converted to cDNA) is hybridized to the chip. For individuals who are heterozygous (AB) at a particular SNP, distortions in the expected 1:1 ratio of allele A to allele B in the RNA signal can be an indication of ASE. Both DNA and RNA hybridizations are necessary for each sample, in order that genotype calls (AA, AB, BB) and allelic expression ratios are available for each SNP on the array.

Illumina's two-color Infinium technology has been used to measure ASE in lymphoblastoid cell lines (Ge *et al*., 2009). Infinium chemistry differentially labels allele A and allele B with red and green dye, respectively (Peiffer *et al*., 2006; Steemers *et al*., 2006). The fluorescence of each probe is quantified by Illumina's scanning software and summarized values for each SNP are output by the GenomeStudio software. To process the raw signal from the DNA hybridizations into genotype calls, Illumina's proprietary GenCall algorithm, implemented in the GenomeStudio software is typically used.

A number of ASE detection methods have been proposed. These include locus-specific approaches, that search for ASE on a SNP-by-SNP basis (Ritchie *et al*., 2010; Serre *et al*., 2008; Tan *et al*., 2008) and tests that combine data from neighboring SNPs in the genome (Ge *et al*., 2009; Wagner *et al*., 2010). Irrespective of the testing method, careful preprocessing to remove non-informative SNPs and dye-bias has been shown to be important in array-based ASE data sets that make use of the Illumina genotyping platform (Ge *et al*., 2009; Ritchie *et al*., 2010).

*To whom correspondence should be addressed.

2 APPROACH

In this article, we analyze data from two Illumina Infinium ASE studies, one involving cell lines and a second involving tissue samples. We first focus on the importance of careful preprocessing of the raw data. We then compare three ASE testing procedures (one existing approach and two novel methods) on the preprocessed data sets. SNPs from known imprinted genes are used as true positives to compare the performance of the different methods. All methods trialled are by their very nature, or the way in which we apply them, expected to identify SNPs for which ASE is relatively common. In both experiments, biological replication in the form of samples of the same tissue type from many individuals is available. All analysis was carried out in the R software environment [(R Development Core Team, 2011); http://www.R-project.org/, version 2.13.1].

3 METHODS

3.1 Data sets

Data from Ge *et al*. (2009) which made use of Illumina's Human 1M-Duo SNP genotyping BeadChip platform were provided by Bing Ge (personal communication). Each 1M-Duo array contains over 1 million SNPs, and two samples are processed in parallel per BeadChip. Both RNA and DNA samples from 53 HapMap individuals (HapMap, 2007) derived from lymphoblastoid cell lines were hybridized to these arrays. Raw X (allele A) and Y (allele B) signal for each SNP, along with genotype calls from GenomeStudio were used in the analysis. Data from 841 816 SNPs were available.

A second data set was generated using Illumina's HumanExon510s-Duo genotyping BeadChip platform. This platform contains 511 354 markers, of which ~330 000 target the coding region of the genome. Normal breast tissue was obtained from 76 individuals, with RNA and DNA samples from a given individual hybridized on the same BeadChip. Raw X and Y signal, along with genotype calls from GenCall (version 1.0, in BeadStudio version 3.1.3) were used in the analysis. This data set is available from GEO (www.ncbi.nlm.nih.gov/geo/) under accession number GSE35023.

3.2 Preprocessing

3.2.1 Sample filtering Samples for which there was a poor dynamic range of intensities in either channel (IQR of non-normalized $\log_2 X$ or $\log_2 Y$ < 1), or clear contamination of RNA with DNA sample [observed by visual inspection of the $M = \log_2 X - \log_2 Y$ versus $S = 0.5(\log_2 X + \log_2 Y)$ plots] were removed from further analysis (see Supplementary Figure S1). This left 53 and 64 samples in the 1M and 510s data sets, respectively.

3.2.2 Normalization Normalization is a key step in any microarray analysis, and we use within-array strip-level quantile normalization to correct for dye-biases between the two channels (in general $X =$ Cy5 = allele A and $Y =$ Cy3 = allele B) from each array, as previously recommended for Illumina Infinium BeadChip data (Ritchie *et al*., 2009) to obtain normalized intensities (X^* and Y^*) for each SNP. The function `crlmm:::stripNormalize` in the *crlmm* R package (version 1.10.0, available from http://www.bioconductor.org) was used to normalize the data. Log-ratios ($M = \log_2 X^* - \log_2 Y^*$) and average log-intensities [$S = 0.5(\log_2 X^* + \log_2 Y^*)$] were then calculated for each SNP on each array. We work on the log-ratio scale rather than the β-scale [$\beta = X^*/(X^* + Y^*)$, as used in (Ge *et al*., 2009)], as it has been shown to have more desirable properties for statistical testing in other applications of Infinium technology (Du *et al*., 2010).

3.2.3 SNP and call filtering Figure 1 shows the RNA and DNA signal from three typical SNPs. In general, we observe a linear relationship between the homozygous RNA and DNA M-values, which motivates us to fit a SNP-wise linear model (fitted using the `lm` function from the *stats* R package, available from http://www.R-project.org/) with an intercept and slope term to regress the RNA values on the DNA values. The strength of this linear relationship (magnitude of the slope) dictates how well the two alleles can be distinguished from one and other. Figure 2 shows that the majority of SNPs with low slope are also weakly expressed, with low average intensity. For SNPs with a slope <0.2 (such as the example given in Fig. 1C), the two alleles were deemed not to be differentially expressed in the tissue being analyzed. This cut-off was chosen by careful inspection of the distribution of slopes. After applying this filter, we are left with 288 396 and 172 123 SNPs in the full analysis of all samples in the 1M and 510s data sets, respectively.

All ASE detection methods rely on accurate genotype calls. To avoid false positives or false negatives driven by genotyping errors, such as the examples illustrated in Figure 3, it is necessary to remove erroneous calls. To do this,

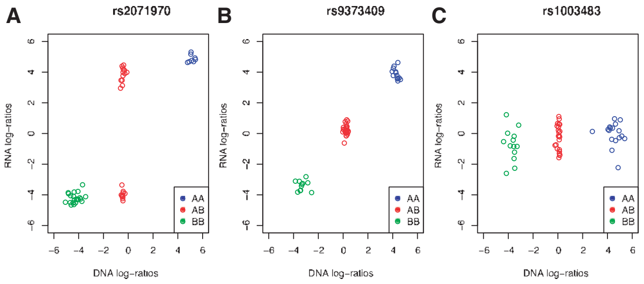

Fig. 1. Plot of signal from three SNPs from the 1M lymphoblastoid data set, illustrating typical signal characteristics for RNA (*y*-axis) versus DNA (*x*-axis) log-ratios. All three SNPs are from imprinted genes. (**A**) SNP rs2071970 from the paternally imprinted gene L3MBTL. This example illustrates bi-allelic expression, with the RNA log-ratios from the heterozygotes (AB, red) showing signal more like that from a homozygote (AA, blue; BB, green). (**B**) SNP rs9373409 from the gene PLAGL which is known to be imprinted in foetal tissues but not in lymphocytes. Accordingly, there is no evidence for ASE for this SNP, with the AB RNA log-ratios around zero. For the SNPs in panels A and B, there is a strong linear trend between the homozygous RNA and DNA log-ratios. (**C**) SNP rs1003483 (from the gene IGF2AS) provides an example of non-specific signal, where the RNA log-ratios lie around zero irrespective of the alleles present. In this final example, there is a very weak linear trend between the homozygous log-ratios. These observations motivate the removal of SNPs with low slope for the regression between the RNA and DNA homozygous log-ratios.

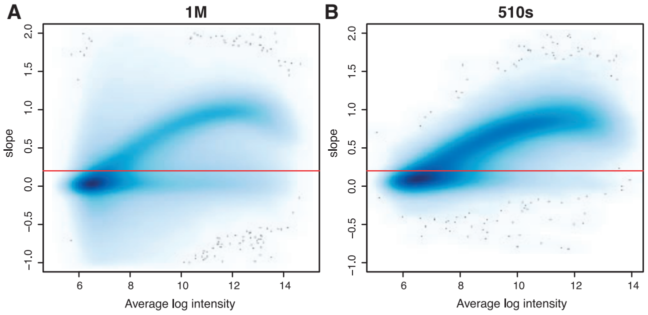

Fig. 2. Smoothed scatter plot of slope versus average log-intensity ($\overline{S}$) for SNPs from the 1M (**A**) and 510s (**B**) data sets. In these plots, a higher density of points is represented by a darker shade of blue. A clear trend is for slope to increase as expression level increases. To remove non-informative SNPs which show no evidence of differential expression between the two alleles in homozygous individuals, we filter based on slope (*y*-axis). SNPs with slope <0.2 (red line) were removed from further analysis, as they were deemed non-informative for ASE in these data sets.

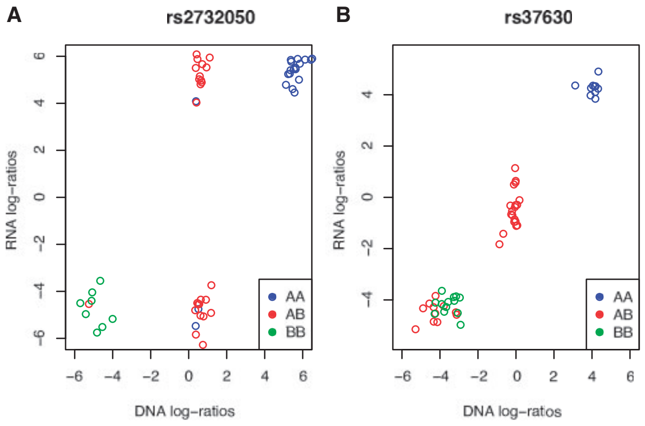

Fig. 3. Plot of signal from two SNPs from the 1M data set with possible genotyping errors. (**A**) SNP rs2732050 has bi-allelic expression and multiple genotyping errors, particularly for the AA calls. Such errors will effect Tan's method, and lead to false negative ASE calls. (**B**) SNP rs37630 has errors with the heterozygous SNP calls, which will lead to a false positive declaration of ASE by all methods compared. Individual calls such as these which are clear genotyping errors were removed from further analysis. Errors were detected as described in the Methods, using the DNA signal alone.

Table 1. Summary of the ASE detection methods compared

Method	Approach	Data used (log-ratios)
Tan	Heuristic (confidence intervals)	RNA and DNA
Variance	Statistical (variance test)	RNA
Cluster	Heuristic (clustering)	RNA (heterozygotes only)

we apply *kmeans* clustering with $k=3$ clusters to the DNA signal from each SNP using the `kmeans` function from the *stats* R package (available from http://www.R-project.org/). We then compare the *kmeans* cluster assignment to the genotype calls from GenCall, and remove individual calls which are discrepant. Both the DNA and matched RNA signal from these discordant calls are excluded from ASE testing.

3.3 Testing methods

A summary of the three methods that will be applied to these data sets is given in Table 1.

3.3.1 Tan The first is from Tan *et al.* (2008) and Serre *et al.* (2008), which we refer to as the method of *Tan* in the remainder of this article. This method interpolates the upper and lower limits (mean +/− 2SD) of the RNA heterozygous log-ratios from the homozygous RNA and DNA log-ratios to give a 'normal' range of allelic imbalance amongst the heterozygotes. For each heterozygous SNP, the ratio of the absolute distance between a given observation and the heterozygous cluster mean ($|d|$) and the absolute distance from the cluster mean and nearest limit ($|r|$) is calculated.

We use the proportion of heterozygous SNPs for which this value ($\theta=|d|/|r|$) is greater than a given threshold (typically $\theta>2$) to rank SNPs. An average ASE score ($\overline{\theta}$) is also calculated for each SNP with at least one heterozygous observation that satisfies this ASE criteria, which can help resolve ties that occur when ranking by proportion. Tan's method can be applied to SNPs with a minimum of three heterozygote and five homozygote observations, and at least one observation for the homozygote of the minor allele.

3.3.2 Variance The second method is motivated by the observation that the RNA log-ratios of SNPs from genes which show ASE are more variable compared with the RNA log-ratios from either homozygote (AA or BB) cluster (Fig. 1A). To formally test this, we use the `var.test` function from the *stats* package to perform an F-test for each SNP to look for excess variation between the heterozygotes versus the pooled homozygotes (adjusted to have mean zero). SNPs are then ranked by their P-value (and F-statistic in the event of ties) to assess evidence for ASE. We require at least three homozygote and three heterozygote observations to perform this testing procedure. This method is referred to as the *Variance* method in the remainder of this article.

3.3.3 Cluster The final method makes use of RNA signal from the heterozygotes only. Given the observation that two distinct clusters (Fig. 1A) are often observed for imprinted genes, this approach applies *kmeans* clustering with $k=2$ clusters to the RNA log-ratios for each SNP separately. SNPs are then ranked based on the distance between the two clusters (largest to smallest), with larger distances indicative of stronger ASE. At least three heterozygote observations are required to perform this testing procedure, which makes use of the `kmeans` function in the *stats* package. This method is referred to as the *Cluster* method in the remainder of this article.

3.3.4 Independent truth A list of 65 imprinted genes in human were downloaded from the Imprinted Gene Database (http://www.geneimprint.com/). SNPs which were assigned to these genes according to Illumina's chip annotation files (downloaded from Illumina's iCom customer website, https://icom.illumina.com/) were used as true positives for ASE in our subsequent analysis to allow comparison of the sensitivity of the three methods described above. In total, there were 1669 SNPs on the 1M and 575 SNPs on the 510s which could be assigned to these genes, and had sufficient slope or number of heterozygous or homozygous calls to be used in the comparison. The specificity of the different methods could not be assessed due to the lack of a suitable true negative set for ASE. SNPs from genes in the true positive set that are either not expressed in the tissue being analyzed (Fig. 1C) or expressed but not imprinted (Fig. 1B) can be considered true negatives. However, there is not a convenient way to independently identify such cases in advance. Also, SNPs in the first class are likely to have been removed prior to testing by our pre-filtering step (refer to Section 3.2.3), and those that are expressed but not imprinted are expected to receive a low ranking by all three methods. As such, we might expect all methods to have comparable, low true negative rates.

3.3.5 Gene set testing We applied a Wilcoxon rank-based gene set testing procedure to the positional (C1: gene sets corresponding to chromosome and cytogenetic bands with at least one gene) and curated (C2: gene sets collected from online pathway databases and publications in PubMed) gene sets available from MsigDB (Subramanian *et al*., 2005) to assess consistency between the two platforms and analysis methods. SNPs were assigned a gene symbol based on the information provided with Illumina's chip annotation. Genes were ranked according to proportion for Tan's method (with ties broken using $\overline{\theta}$), P-value for the Variance method (with ties broken by the F-statistic) and the inter-cluster distance measure for the Cluster method. The `geneSetTest` function in the *limma* package [version 3.8.3, (Smyth, 2005)] as used in (Michaud *et al*., 2008) was used to calculate a P-value testing whether a given gene set was highly ranked relative to other genes on the array.

4 RESULTS

Figure 4 shows the data from the RNA hybridizations from the 1M data set before and after strip-level quantile normalization. Before normalization, some curvature of the log-ratios at around $M=0$ is evident (Fig. 4A). This is no longer present after normalization (Fig. 4B), which is desirable, as the zero line is the level of the

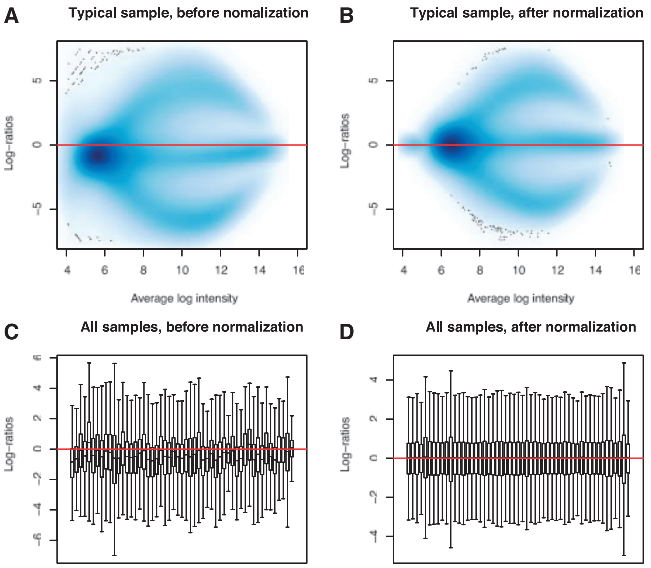

Fig. 4. (**A**) Smoothed scatter plot showing non-normalized M versus S-values from a typical RNA sample from the 1M data set. Curvature of the log-ratios at $M=0$ is evident in this figure. (**B**) Smoothed scatter plot of the M- versus S-values for the same sample after strip-level between-channel quantile normalization. (**C**) Boxplots of the log-ratios from all RNA samples from the 1M data set before normalization. The median values tend to be shifted below the equality line ($M=0$) in most samples. (**D**) Boxplots of the log-ratios from all RNA samples from the 1M data set after strip-level quantile normalization. The distributions are much more consistent and symmetric around $M=0$ after normalization.

heterozygous M-values expected a priori (ignoring dye-effects and other biases). Looking across all samples, we see that the distribution of RNA log-ratios is more variable before normalization (Fig. 4C). After normalization, these distributions are more comparable and centered around zero (Fig. 4D). This normalization procedure also improved the comparability of the DNA log-ratios from the 1M data set (data not shown) and both the RNA and DNA data from the 510s data set (data not shown).

We next examined the ability of the different ASE testing methods to detect SNPs in known imprinted genes. Figure 5 shows the effect sample size has on the sensitivity of the different methods using the top-ranked 1000 SNPs from each. To carry out this analysis, a subset of 10 and 20 samples chosen at random as well as the full 1M data set ($n=53$) were analyzed using each of the 3 methods (Tan, Variance and Cluster). Across this range of sample sizes, we see that the Cluster method is consistently the best, followed by the Variance method and Tan's method. As sample size increases, all methods recover a greater number of true positive SNPs, although the improvement in performance is most pronounced for the Variance method, with a true positive rate approaching that of the Cluster method for the full data set (Fig. 5C). The performance of the different methods was similar for the 510s data set (data not shown).

For genes represented by 5 or more SNPs on the array, we were able to examine how well each method was able to detect multiple occurrences of ASE within a gene. Supplementary Figure S2 shows the number of genes detected independently by 2 or more SNPs amongst the top 1000 candidates from each method. In both data sets, the Variance and Cluster methods detected multiple ASE hits in more genes than Tan's approach.

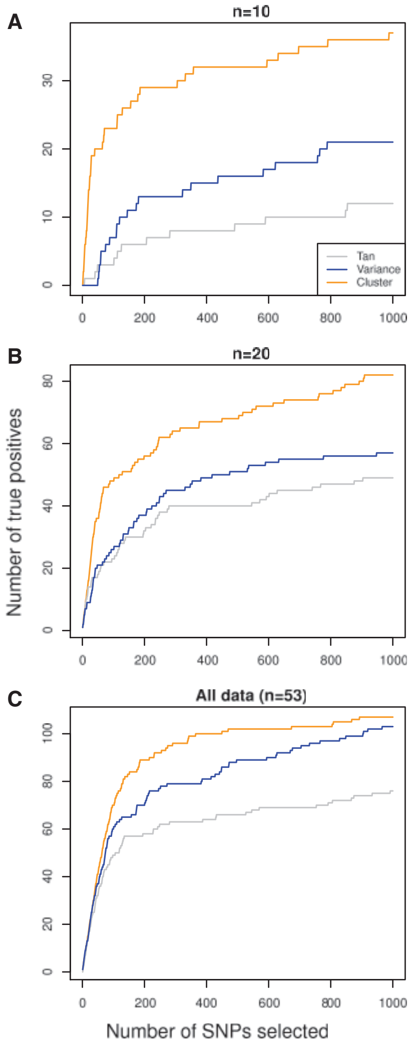

Fig. 5. Plot showing the number of 'true positives' recovered in the 1M data set as the sample size increases from $n=10$ samples (**A**), to $n=20$ samples (**B**), to the full data set (**C**, $n=53$). The Cluster method has the highest sensitivity across the range of sample sizes tested. The performance of all methods improves with increasing sample size, with the Variance method improving the most.

We next assessed whether the three methods had any preference for detecting SNPs at different expression levels. Boxplots of the average intensities for the top-ranked 1000 SNPs from each method for the full 1M and 510s data sets were generated (Supplementary Figure S3). In both cases, the Variance method has a tendency to detect SNPs with higher average expression levels than Tan's method and the Cluster method. The SNPs selected by the Cluster method have a distribution most similar to that of the reference distribution of all SNPs on the array (post-filtering).

Typical examples of SNPs recovered by the respective methods are shown in Figure 6. Most of these appear to be good candidates for ASE. Besides examples of classic bi-allelic expression (first column), many SNPs with log-ratios on the continuum from the homozygous AA to the homozygous BB level (second column) are also recovered. These two kinds of ASE profile are the most commonly observed in the top ranked SNPs from each method.

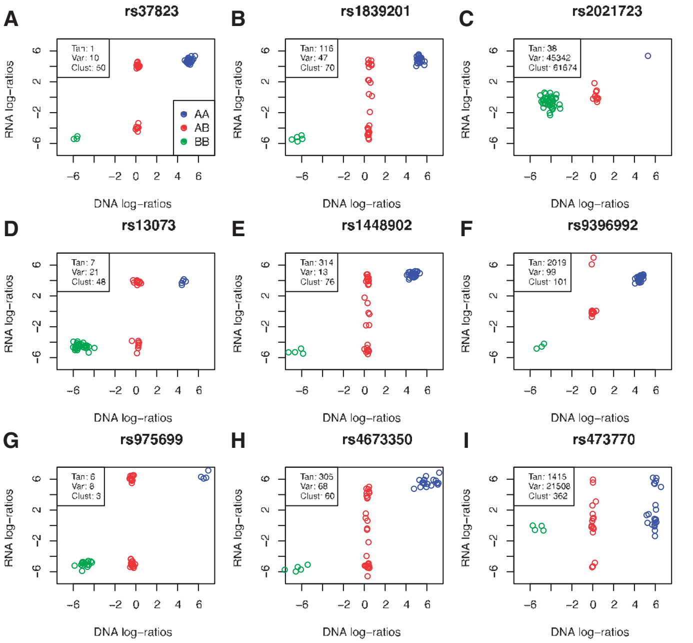

Fig. 6. Plot showing typical examples of SNPs for which ASE was detected in the 1M data set by Tan's method (**A–C**), the Variance method (**D–F**) and the Cluster method (**G–I**). The ranks for each method are given in the top left of each plot. The majority of these examples appear to be good candidates for ASE.

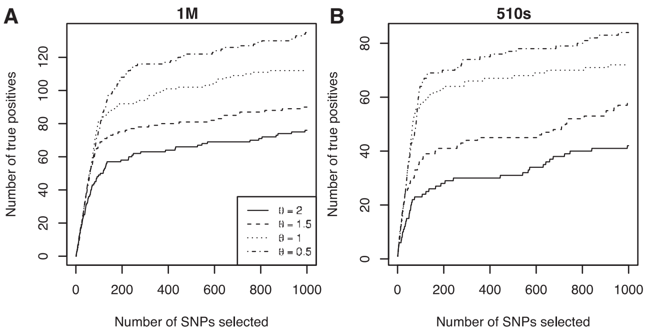

Fig. 7. Plot showing the sensitivity of Tan's method for the 1M (**A**) and 510s (**B**) data sets as the threshold for ASE (θ) is varied. In both data sets, as θ decreases from high (2) to low (0.5) stringency, the recovery of 'true positive' SNPs from known imprinted genes improves dramatically.

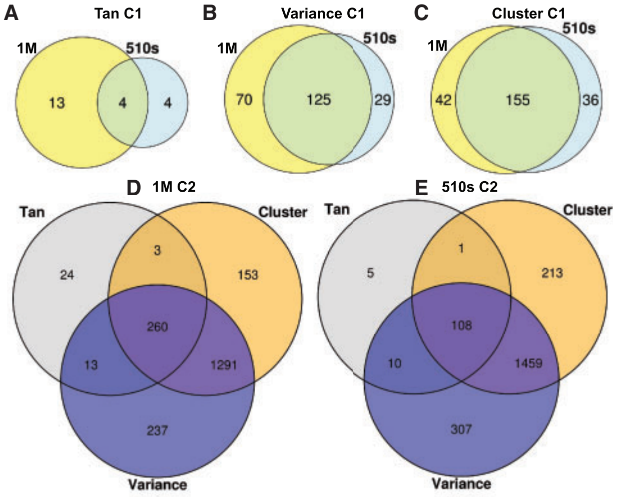

Fig. 8. Consistency of gene set testing results between and within data sets. Panels **A–C** show the degree of overlap between the 1M and 510s data sets for C1 positional gene sets which are over-represented by SNPs with ASE according to a Wilcox rank test for each method. A cut-off of P-value <0.0001 was chosen to select these sets. Not only do the Variance (B) and Cluster (C) methods have greater power, detecting more significant positions than Tan's method (A), but the regions selected are more consistent between platforms and data sets, which would be expected due to the clustering of imprinting in the genome. Similar results were obtained when other gene sets were tested (such as Gene Ontology and C2 curated gene sets from MsigDB, data not shown). Panels **D** and **E** show the overlap between methods within each data set for the curated C2 gene sets. The Variance and Cluster methods have a greater degree of overlap with each other than Tan's method. This figure was generated using the VennDiagram R package (Chen and Boutros, 2011).

A third class particular to Tan's method is shown in Figure 6C. Such examples are characterized by low minor allele frequency, which means the homozygous cluster center for the minor allele will be less reliably estimated than the cluster center for the major allele. Around a third of SNPs from the top 200 identified by Tan's method fall into this category, and many of these are likely to be false positives. A fourth class, that was relatively common for the Variance method, were cases for which ASE is relatively rare. Figure 6F shows an example in which two outlying RNA log-ratios drive the ASE signal for this SNP. Around a quarter of the top 200 SNPs ranked by the Variance method fall into this category. This SNP was also highly ranked by the Cluster method. Profiles of this kind were less common amongst the top ranked SNPs from the Cluster method (∼3 in every 100 from the top 200 SNPs inspected).

Unlike the Variance or Cluster method, Tan's method contains the tuning parameter θ which can be varied depending upon the desired stringency for ASE. To assess whether the performance of Tan's method can be improved, we re-analyzed both the full 1M ($n=53$) and 510s ($n=64$) data sets relaxing the threshold used for θ. The results of this analysis are shown in Figure 7, which plots the cumulative number of true positives recovered from known imprinted genes from the top 1000 SNPs. For both data sets, it is clear that using a lower, less conservative threshold for θ produces much better results than higher values ($\theta \geq 2$) as used in a number of previous studies (Morcos *et al.*, 2011; Serre *et al.*, 2008; Tan *et al.*, 2008).

We next assess the results from the gene set analysis to look for concordance between the methods. Figure 8 shows the overlap in significant positional C1 gene sets (those with P-value <0.0001) between platforms for each method. The Variance (Fig. 8B) and Cluster (Fig. 8C) methods have greater power, giving more significant gene sets overall than Tan's method (Fig. 8A). The gene sets reported are also highly consistent between platforms for the Variance and Cluster method, which is desirable given that imprinted genes are clustered in the genome, and many of the strongest ASE signals will be originating from these imprinted loci. If we look within a given data set and perform an analysis of the curated C2 gene sets, we similarly find that the Variance and Cluster methods recover more significant gene sets than Tan's method (Fig. 8D–E). There is also much greater consistency between the gene sets identified by the Variance and Cluster methods than between Tan's method and the others. The reason for the reduced power of Tan's method in each analysis is that beyond the first 20 000 or so SNPs from each data set, the results for the remaining SNPs are tied, which results in a loss of information and power in the rank-based gene set test used. The Variance and Cluster methods do not suffer from this problem, allowing all SNPs to be ordered by their respective test statistics.

5 DISCUSSION

We have shown that careful preprocessing of data from Illumina array-based ASE studies is important for improving signal consistency between different samples, and for avoiding potential false positives. The normalization and filtering approaches we have described ensure that the input data for a given ASE testing method is appropriately standardized and error free (as far as practicable). A similar strategy is likely to be useful when other microarray platforms, such as Affymetrix's two-color Axiom technology, are used to measure ASE.

A summary of the findings of our methods comparison is given in Table 2. The new heuristic approach we describe (Cluster) outperforms the other methods at recovering SNPs from known imprinted genes across a range of sample sizes, making it ideal in small studies. The Variance method was demonstrated to perform well on larger sample sizes, and was particularly good at detecting rare cases of ASE, driven by one or two individuals with extreme allelic expression imbalances. Both methods are appealing on account of their simplicity and the absence of any tuning parameters that the user must pre-set to obtain a result. Tan's method was found to perform worst initially, however, we observed that the performance could be improved considerably by relaxing the threshold for ASE (θ). Less conservative values of θ recovered more true positive imprinted SNPs than the default value used in previous studies ($\theta \geq 2$), which indicates that true positive candidates for ASE will have been missed in previous analyses. Tan's method is also less powerful and less consistent than both the Cluster and Variance methods in the gene set analyses carried out, and was less likely to detect multiple independent ASE hits in genes represented by several SNPs on the array. Together, these results indicate that the newer, simpler methods proposed in this article (Cluster and Variance) offer a good alternative to Tan's methods for detecting ASE in SNP microarray data sets.

Future work will be to compare the results obtained from SNP microarrays with other high-throughput technologies such as RNA-sequencing to assess the relative advantages and disadvantages of each platform. Whether similar detection methods can be applied to the ratios from RNA-seq allele counts remains to be seen.

Table 2. Summary of comparison findings

Method	Tuning parameter?	Study size for optimal performance	Power for gene set testing
Tan	Yes	Large	Low
Variance	No	Large	High
Cluster	No	Any size	High

ACKNOWLEDGEMENT

We thank Terry Speed for helpful discussions and feedback on this manuscript.

Funding: This work was supported by NHMRC Program grant 406657 (R.L. and M.E.R.), the University of Cambridge, Cancer Research UK, and Hutchison Whampoa Limited (A.T.M., R.R., C.C. and B.A.P.).

Conflict of Interest: none declared.

REFERENCES

Bjornsson,H.T. *et al.* (2008) SNP-specific array-based allele-specific expression analysis. *Genome. Res.*, **18**, 771–779.

Chen,H. and Boutros,P.C. (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, **12**, 35.

Daelemans,C. *et al.* (2010) High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. *BMC Genet.*, **11**, 25.

Du,P. *et al.* (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Feng,X. *et al.* (2006) Allele-specific silencing of Alzheimer's disease genes: the amyloid precursor protein genes with Swedish or London mutations. *Gene*, **371**, 68–74.

Ge,B. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.*, **41**, 1216–1222.

Gimelbrant,A. *et al.* (2007) Widespread monoallelic expression on human autosomes. *Science*, **318**, 1136–1140.

International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Knight,J.C. (2004) Allele-specific gene expression uncovered. *Trends Genet.*, **20**, 113–116.

Lee,M.P. (2005) Genome-wide analysis of allele-specific gene expression using oligo microarrays. *Methods Mol. Biol.*, **311**, 39–47.

Lo,H.S. *et al.* (2003) Allelic variation in gene expression is common in the human genome. *Genome Res.*, **13**, 1855–1862.

Maia,A.-T. *et al.* (2009) Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. *Breast Cancer Res.*, **11**, R88.

Meyer,K.B. *et al.* (2008) Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biol.*, **6**, e108.

Michaud,J. *et al.* (2008) Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics*, **9**, 363.

Morcos,L. *et al.* (2011) Genome-wide assessment of imprinted expression in human cells. *Genome Biol.*, **12**, R25.

Pant,P.V. *et al.* (2006) Analysis of allelic differential expression in human white blood cells. *Genome Res.*, **16**, 331–339.

Peiffer,D.A. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.

R Development Core Team. (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ritchie,M.E. *et al.* (2009) R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. *Bioinformatics*, **25**, 2621–2623.

Ritchie,M.E. *et al.* (2010) Data analysis issues for allele-specific expression using Illumina's GoldenGate assay. *BMC Bioinformatics*, **11**, 280.

Serre,D. *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.

Smyth,G.K. (2005). Limma: linear models for microarray data. In Gentleman,R. *et al.* (eds) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 397–420.

Steemers,F.J. *et al.* (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods*, **3**, 31–33.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tan,A.C. *et al*. (2008) Allele-specific expression in the germline of patients with familial pancreatic cancer: an unbiased approach to cancer gene discovery. *Cancer Biol. Ther.*, **7**, 135–144.

van Bilsen,P.H. *et al*. (2008) Identification and allele-specific silencing of the mutant huntingtin allele in Huntington's disease patient-derived fibroblasts. *Hum. Gene. Ther.*, **7**, 710–719.

Wagner,J.R. *et al*. (2010) Computational analysis of whole-genome differential allelic expression data in human. *PLoS Comput. Biol.*, **6**, e1000849.