MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) **12**, 1079–1089 doi: 10.1111/1755-0998.12006

An improved algorithm for the detection of genomic variation using short oligonucleotide expression microarrays

MATTHEW L. SETTLES,* TRISTAN CORAM,† TERENCE SOULE*,‡ and BARRIE D. ROBISON*,§
*Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID83844-3051, USA, †Dow AgroSciences LLC, Indianapolis, IN46268-1053, USA, §Department of Computer Science, University of Idaho, Moscow, ID83844-1010, USA, ‡Department of Biological Sciences, University of Idaho, Moscow, ID83844-3051, USA

Abstract

High-throughput microarray experiments often generate far more biological information than is required to test the experimental hypotheses. Many microarray analyses are considered finished after differential expression and additional analyses are typically not performed, leaving untapped biological information left undiscovered. This is especially true if the microarray experiment is from an ecological study of multiple populations. Comparisons across populations may also contain important genomic polymorphisms, and a subset of these polymorphisms may be identified with microarrays using techniques for the detection of single feature polymorphisms (SFP). SFPs are differences in microarray probe level intensities caused by genetic polymorphisms such as single-nucleotide polymorphisms and small insertions/deletions and not expression differences. In this study, we provide a new algorithm for the detection of SFPs, evaluate the algorithm using existing data from two publicly available Affymetrix Barley (*Hordeum vulgare*) microarray data sets and compare them to two previously published SFP detection algorithms. Results show that our algorithm provides more consistent and sensitive calling of SFPs with a lower false discovery rate. Simultaneous analysis of SFPs and differential expression is a low-cost method for the enhanced analysis of microarray data, enabling additional biological inferences to be made.

Keywords: bioinfomatics/phyloinformatics, ecological genetics, genomics/proteomics, molecular evolution, transcriptomics

Received 24 May 2012; revision received 30 July 2012; accepted 1 August 2012

Introduction

Ecological population studies using transcriptome data are increasingly common and are leading to new biological insights (Oleksiak *et al.* 2002; Kammenga *et al.* 2007; Bay *et al.* 2009). A common experimental design is to compare gene expression differences across populations to gain insights into the transcriptional response to environmental conditions. These experiments are typically focused on the analysis of differential gene expression and do not consider potential genetic variation. Currently, high-throughput RNA sequencing methods (RNA-seq) can be used to assess both genetic variation and gene expression simultaneously; however, this technique is still costly, relative to microarrays, limiting the sample size of the experiment. Further, RNA-seq analysis methodology, while currently under active development, is not as established. DNA microarrays are still commonly in use today with a per sample cost that allows for the large sample sizes needed for broader population inferences. In addition, analysis methodology for microarrays is well established with over a decade's worth of development in peer reviewed publications.

Short oligonucleotide microarrays have been used to predict candidate locations of genomic and transcriptional polymorphisms between populations, using both genomic DNA (gDNA) (e.g. Borevitz *et al.* 2003; Winzeler *et al.* 2003; Kim *et al.* 2006; Kumar *et al.* 2007; and others) and messenger RNA (mRNA) (e.g. Cui *et al.* 2005; Ronald *et al.* 2005; Rostoks *et al.* 2005; West *et al.* 2006; Luo *et al.* 2007; and others). The advantage of using mRNA over gDNA is that no additional experiments need be performed, as both expression and genetic variability can be assessed simultaneously. However, mRNA-based polymorphism detection algorithms are generally more error prone, as they must also consider variation in gene expression. An algorithm to reliably predict candidate locations of genetic polymorphisms from microarray gene expression experiments will

Correspondence: Matthew L. Settles, Fax: (208) 885-5003;
E-mail: msettles@uidaho.edu

provide a high-throughput technique for studying both the genetic and transcriptional basis of phenotypic variation between populations from routine microarray gene expression experiments.

When a short oligonucleotide probe is designed at a position with a genomic or transcriptional polymorphism, the hybridization efficiency is reduced. Single feature polymorphisms (SFPs) are statistical differences in the probe level hybridization efficiency between two populations caused by an underlying genetic or transcriptional polymorphism. They are detected by comparing microarray probe level intensity signals, a proxy value for hybridization efficiency, between two populations. When hybridizing gDNA, SFPs are induced by single-nucleotide polymorphisms (SNPs) and small insertions/deletions (INDELs) (Borevitz *et al.* 2003). When hybridizing mRNA, SFPs can also be induced by splicing variation and polyadenylation differences (Rostoks *et al.* 2005). It is also important that the probe be a particular length, the shorter a probe is the greater the likelihood of nonspecific binding. Conversely, as a probe becomes longer, the impact that small localized polymorphisms (ie. SNPs) have on hybridization efficiency is reduced, making SFP detection increasingly difficult. The Affymetrix GeneChip microarray platform offers a whole genome solution with short oligonucleotide probes (25-mers) to detect both expression and SFPs (as performed in Kim *et al.* 2006; Coram *et al.* 2008; Bernardo *et al.* 2009; Childs *et al.* 2010).

Affymetrix GeneChip microarrays consist of oligonucleotide fragments of length 25 bp, termed probes. The mRNA molecule of interest is measured by multiple probe pairs, usually 11–20, assembled into a probe set. Each probe pair is composed of one perfect match (PM) probe and one mismatch (MM) probe, the PM probe matches the targeted mRNA sequence exactly, while the MM probe is generated by complementing the middle (13th bp) nucleotide of the PM probe. The MM probes were intended to provide an estimate of nonspecific binding and background. Labelled gDNA or cDNA is hybridized to a microarray, scanned, and image analysis is performed to provide intensity values for each of the PM and MM probes. When a SFP exists between populations, the probe's relative hybridization efficiency is expected to be different, and therefore, the probe's relative intensity values will also be different (see Fig. 1A). When hybridizing gDNA, prediction of SFPs is relatively straight forward; any significant probe intensity differences can be assumed to be attributed to hybridization efficiency differences and therefore an SFP. When predicting SFPs from mRNA, any prediction technique must also consider the expression level of the mRNA molecule and be able to distinguish probe level differences because of underlying genomic variation from differences because of variable gene expression (see Fig. 1B). A probe's hybridization efficiency can be calculated by computing the difference between the probe's observed intensity (raw intensity value) from the expected intensity value (intensity because of expression) for each probe (Ronald *et al.* 2005; Rostoks *et al.* 2005). The resulting measure has the signal because of expression removed and can be analysed in a similar manner to gDNA hybridization data.

In this study, we describe two variations on a new algorithm for the detection of SFPs in standard expression microarray experiments. The two variations on our SFP prediction algorithm use different probe set summarization methods, Robust Multichip Averaging (RMA) and MicroArray Suite (MAS), to compute the expected intensity values. The algorithms are evaluated using data from two publicly available Affymetrix Barley (*Hordeum vulgare*) microarray data sets and are compared to two previously published SFP detection algorithms (Ronald *et al.* 2005; Rostoks *et al.* 2005). The two barley data sets use the same two cultivars (Morex and Golden Promise), but differ in the number of microarrays and tissue types. A previously reported DNA sequence database of polymorphisms between Morex and Golden Promise is used to determine the sensitivity, specificity and false discovery rate of each method. We assess the overlap of called SFPs across the four detection methods as well as between the two barley data sets. We also explore the sensitivity of each algorithm to predict known polymorphisms by their known position within the probe. Finally, we compare the overlap between genes with called SFPs and their differential expression calls. Results show that our algorithm using the RMA summarization method better estimates and removes signal from expression than the other methods. The result is more consistent and sensitive calling of SFPs with a lower false discovery rate.

Materials and methods

Experiment data sets

The Affymetrix Barley Genome Array contains 22,840 probe sets, each with 11 probe pairs (PM and MM probes), developed by an international collaboration of barley researchers (Close *et al.* 2004). Raw data from two publicly available barley data sets were obtained from plexdb (http://www.plexdb.org, Experiment ID: BB3) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress, Experiment ID: E-TABM-110). The first barley microarray data set (BB3) was generated

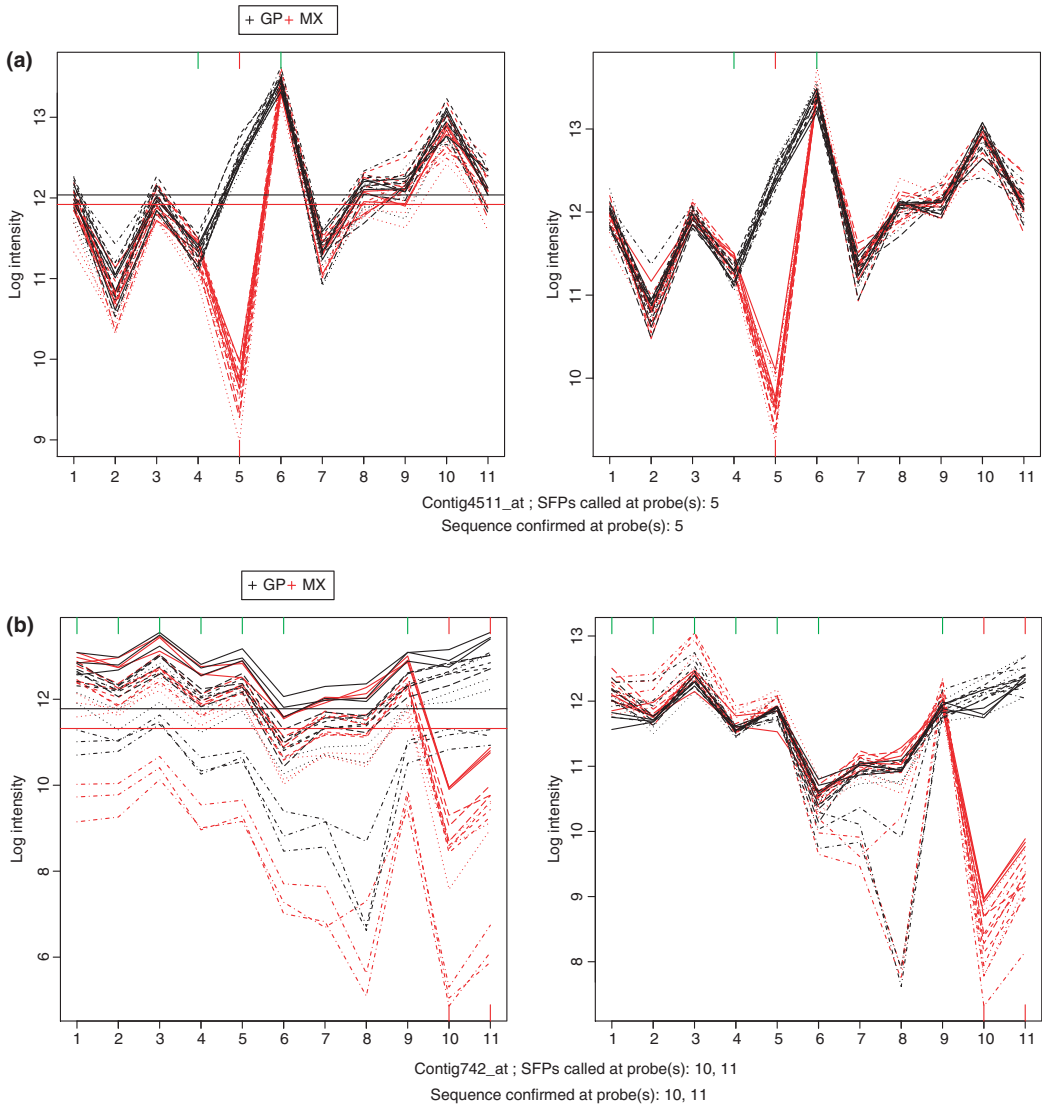

Fig. 1 Single feature polymorphism (SFP) detection using the Robust Multichip Averaging (RMA) preprocessing pipeline with a low expression variability gene (a) and one with high expression variability (b) in the BB3 data set. Each pane shows the $\log_2$ intensity values (*y* axis) of each array for each probe (*x* axis), with the Morex cultivar samples shown in red and Golden Promise samples in black. The left panes show unadjusted raw $\log_2$ intensity values with the average RMA computed expression values for each genotype (horizontal lines). The right pane shows $\log_2$ intensity values after RMA expression adjustment, each probe can then tested for a genotype effect. The bottom axis tick marks show the called SFPs (red for Morex, black for Golden Promise), hile the top axis tick marks show probes with known SFPs (red/black) and known SFP free (green) probes, according to the sequence confirmation dataset.

to provide a reference gene expression data set across 15 tissues, six of which (stem, seminal root, vegetative shoot, seedling leaf, coleoptile and hypocotyl) contained samples from both Morex and Golden Promise (GP) cultivars (Druka *et al.* 2006). The data set consists of three biological replicates per tissue and cultivar for a total of 36 microarrays and is also the same data set used to detect SFPs in (Rostoks *et al.* 2005). As determined in Rostoks *et al.* (2005), and verified here, one array of cultivar GP from the tissue type seminal root consistently clusters with the three replicates from the Morex cultivar (data not shown). The seminal root tissue type was therefore removed from this experiment, leaving five tissue types and 30 microarrays. The second data set (E-TABM-110) is also from seedling leaves and contains three replicates from the same two cultivars, Morex and GP, for a total of six microarrays. This data set is used here to determine the sensitivity to call SFPs using a number of microarrays more typical of a small gene expression

experiment and to assess consistency of SFP calls across experiments.

Sequence confirmation data set and algorithmic performance evaluation

The sequence confirmation data set used in this study is the same as that of (Rostoks *et al.* 2005) and can be found at their website (http://naturalsystems.uchicago.edu/naturalvariation/barley/SNPtable. csv). The sequences were collected from three barley sequence sources totalling 2699 sequences. Of those, 30 were duplicated in two of the three sources and one was duplicated in all three. Sixty-six probes contained polymorphisms in both Morex and GP genotypes as compared to the reference sequence on the microarray. After removing duplicates and sequences with polymorphisms in both genotypes, the sequence confirmation data set consisted of 223 sequences polymorphic for GP, 178 for Morex and 2200 sequences that did not contain a sequence polymorphism for either cultivar.

Results of each SFP prediction algorithm were evaluated by direct comparison to the sequence confirmation data set for calculations of sensitivity, specificity and false discovery rate. An algorithm's sensitivity is the proportion of known polymorphisms in the sequence confirmation data set correctly called as a SFP by the algorithm. The specificity is the proportion of known negatives correctly identified as such. False discovery rate is the proportion of called SFPs by the algorithm incorrectly identified as a polymorphism, when the sequence confirmation data set indicates that no SNP is present.

Data preprocessing

All analyses were conducted within the R statistical computing language using publicly available packages from CRAN and BIOCONDUCTOR (R version 2.11.0) (Gentleman *et al.* 2004; R Development Core Team 2010). Raw CEL files were read into R using the bioconductor package *affy* (Gautier *et al.* 2004) and checked for quality using pseudo-chip images and residual error visualizations. Quality assurance of microarray data was further checked using the *affyQAReport* function from the BIOCONDUCTOR package *affyQCReport* (Parman & Halling 2009). Hybridization and housekeeping controls, RNA degradation, sample clustering, Normalized Unscaled Standard Error plots, Local Pooled Error plots and Relative Log Expression plots all showed high-quality data (results not shown), and no additional microarrays were removed. MAS PMA (present/marginal/absent) calls were determined for each probe set within each mircroarray. In this study, a marginal call was considered to be absent. Microrrays were then grouped by common attributes (i.e. strain and tissue). Finally, present/absent calls were determined for each probe set within each group; a probe set was called as present, for the group, if at least five of six samples within the group were called as present. A probe set was retained for further analysis if at least one group was called as present. This procedure is expected to remove probe sets that are unexpressed across the entire experiment or expressed in only one cultivar.

Analysis of differential expression

Differential expression analysis was conducted on each data set in the following manner. First, each data set was normalized using the RMA preprocessing routine (Bolstad *et al.* 2003; Irizarry *et al.* 2003a,b). Differential expression was determined using the linear analysis of microarray technique from the *limma* package (Smyth 2005) with empirical Bayes adjustment to the variance, followed by Benjimini and Hochberg (BH) correction for multiple testing (Benjimini & Hochberg 1995; Smyth 2004). Differential expression was only determined for those probe sets that passed the PMA filter as described in preprocessing. A gene was considered to be differentially expressed if it had both an BH-adjusted P-value <0.05 and a log fold change >0.5. Using both P-value and fold change criteria to determine differential expression is recommended by the MicroArray Quality Control (MAQC) project (Shi *et al.* 2006; MAQC Consortium 2010).

Single feature polymorphism detection

Single feature polymorphism detection was conducted in the following manner. First, hybridization efficiencies for each probe were calculated using one of the four proposed models (models described below). Each probe was then fitted for a genotype effect, using the limma approach with empirical Bayes adjustment to the variance followed BH correction for multiple testing. SFP detection was only performed for those probes within probe sets that passed the PMA filter as described in preprocessing. A probe was considered to contain an SFP if it had a BH-adjusted P-value of <0.05 and a log fold change >0.5. The genotype with the reduced signal was determined to contain the SFP.

Model # 1: linear model (LM). This model of probe level hybridization efficiencies is the same as the model used in (Rostoks *et al.* 2005). In this model, the relative probe hybridization efficiencies are modelled as the residuals from fitting the following LM to each probe set on the array:

$$\log_2(I_{pgtr}) = \mu + \text{probe}_p + \text{genotype}_g + \text{tissue}_t + (\text{tissue}_t \times \text{genotype}_g) + \epsilon_{pgtr} \quad (1)$$

where, I_{pgtr} is the background corrected and normalized intensity of probe p, genotype g, tissue type t, replicate r in a probe set. The residuals from the model are extracted and fitted for a genotype effect using the procedure described earlier.

Model # 2: intensity to RMA expression ratio (RATIO). The Intensity to RMA Expression Ratio model (RATIO) is similar to the method used in (Ronald *et al.* 2005). A difference is that we used values from the RMA summarization procedure to compute the expected expression values instead of the probe-dependent nearest-neighbour (PDNN) model (Zhang *et al.* 2006). This was performed to make a more accurate comparison with the other models that use RMA, and because of the complicated nature of preparing the energy parameterization, files for the barley microarray needed for PDNN. Further, Ronald *et al.* (2005) reported that the RMA summary method exhibited similar and only a slightly weaker performance than the PDNN model. In this model, the relative probe hybridization efficiencies are modelled as:

$$\frac{I_{pa}}{\hat{I}_a} \quad (2)$$

where I_{pa} is the background adjusted and normalized intensity value of probe p in array a and $\hat{I}_a$ is the expected expression value of array a. The expected value of the ratio is 1 for probes that do not contain an SFP and significantly less, or greater, than 1 for probes that do contain a SFP. It should be noted that the RMA model (described below) is equivalent to the a $\log_2$ transformation of $I/\hat{I}$.

Model # 3: RMA subtraction (RMA). The RMA model is the first variant of our new SFP calling algorithm, where the RMA preprocessing procedure is used to estimate the expected expression value of the probe set. From the LM model above, we can interpret the sum of the genotype, tissue and any interaction terms as estimates of group level expression from the mean with its own corresponding error (replicate deviations) that are being nested inside the overall error term. This procedure adds unnecessary variance to the analysis of SFPs. We can remove this group specific error from the overall error and rewrite the LM as:

$$\log_2(I_{pa}) = \text{probe}_p + \log_2(\hat{I}_a) + \epsilon_{pa} \quad (3)$$

$$\mu + (\log_2(I_{pa}) - \log_2(\hat{I}_a)) = \mu + \text{probe}_p + \epsilon_{pa} \quad (4)$$

where, $\log_2(I_{pa})$ is the background adjusted and normalized $\log_2$ intensity value of array a and probe p in the probe set and $\log_2(\hat{I}_a)$ is the $\log_2$ expected expression value of array a for the probe set. The probe_p term in the model represents a scalar adjustment to each probe that accounts for the hybridization differences between the probes and can be ignored in this context. We further scale the relative probe hybridization efficiencies by the mean intensity for the probe set across all arrays which, when partnered with the empirical Bayes adjustment to the variance from the *limma* package, has the effect of giving decreased weight to those probes with low overall expression. The probe hybridization efficiencies are now modelled as the $\log_2$ differences in the probe intensity values from the expected expression values, adjusted by the mean. In the RMA model, the RMA summarization procedure (median polish) is used to compute the expected intensity values for each probe set of each array (Bolstad *et al.* 2003).

Model # 4: MAS subtraction (MAS). The MAS model is the second variation of our new SFP calling algorithm, where the MAS5 microarray preprocessing procedure is used to estimate the expected expression value of a probe set instead of the RMA procedure. The probe level hybridization efficiencies are calculated in the same manner as the RMA model, but the Affymetrix Microarray Suite (MAS5) summarization procedure (Tukey biweight procedure) is used to compute the expected intensity values for each probe set of each array (Affymetrix 2004). MAS5 is the default preprocessing procedure used in Affymetrix's MicroArray Suite (MAS) software for their 3′ IVT microarrays and, along with the RMA procedure, is a commonly used preprocessing procedure for Affymetrix microarray experiments.

Results

Sensitivity, specificity and false discovery rate

Microarray data from two publicly available barley microarray data sets (E-TABM-113 and BB3) were preprocessed according to the procedures described in Material and methods. Each microarray experiment was tested for both differentially expressed genes and SFPs, where probe sets had been both filtered by presence/absence calls (PMA calls) and without any filter applied. Filtering the BB3 data set by PMA calls reduced the number of probe sets from 22 840 to 17 457 (251 240–192 027 probes) and for the E-TABM-113 data set reduced the number of probe sets from 22 840 to 14 232

(251 240–156 552 probes). Table 1 lists the sensitivity, specificity and false discovery rate for each of the four models of hybridization efficiency when applied to filtered and unfiltered BB3 (Table 1A) and E-TABM-113 (Table 1B) data sets. Filtering the data by PMA calls resulted in an increase to the sensitivity to detect known SFPs (average increase of 8.5% in E-TABM-113; 2.0% BB3), a slight decrease in the specificity (average decrease of 0.6% in E-TABM-113; 0.3% BB3) and no consistent change in the expected false discovery rate. Comparison studies by Rostoks et. al and Ronald et. al did not perform any probe set filtering (Ronald *et al.* 2005; Rostoks *et al.* 2005). If a gene is unexpressed, any probe containing an SFP would not be detectable, nor would the gene be differentially expressed. It therefore makes little sense to include these probes in any detection analysis. The trends observed across the four models of hybridization efficiency for both filtered and unfiltered were the same; therefore, from this point on, we discuss results based only on the filtered data set.

Within the BB3 data set, the sensitivity to detect known SFPs varied widely from 62% using MAS to 76.7% using the RATIO model. The sensitivity reflects the ability of a model to detect known SFPs according to the barley sequence confirmation data set. Specificity was similar for each of the four models ranging from 94.0% using the LM model to 97.1% for RMA; specificity reflects the ability of a model to call a probe a non-SFP when it is known, no SFP exists. A more dramatic difference between the models was observed in the false discovery rate, ranging from 21.1% in the RMA to 36% for the LM model. Within the E-TABM-113 data set, the sensitivity to detect known SFPs also varied from 50.4% using the RATIO model to 56.9% using the RMA model. Specificity was again similar for each of the four models, ranging from 95.2% using the MAS model to 98.7% for RMA. Finally, the false discovery rate ranged from 11.5% using the RMA algorithm to 37.3% for the MAS algorithm.

Overall, RMA outperformed MAS and the LM models in every category and outperformed the RATIO algorithm in five of the six evaluated criteria for filtered data. The RATIO method outperformed the LM model in all cases except sensitivity in the E-TABM-113 data set, where RATIO performed the worst when compared to all other models. The MAS procedure performed comparable to the LM procedure for sensitivity and specificity but performed significantly worse than all other models for false discovery rate for the smaller E-TABM-113 data set; however, MAS performed better than both RATIO and LM in the larger BB3 data set. In general, the RMA method performed consistently well across both data sets.

We also tested the sensitivity of each model to call known SFPs by the SNP position within the probe (see Fig. S1, Supporting information). All models show an increased sensitivity to positively detecting known SFPs as the SNPs position moved towards the centre of the probe or if multiple SNPs occur within the same probe. Further, a sharp increase in sensitivity (≈40%) was observed when the variant did not occur in the outside three bases of the probe (occurred within bases 4 and 22 inclusive).

Comparison and overlap of called SFPs

Table 2 shows the frequency in the number of called SFPs per gene across all four models for the BB3 (Table 2A) and E-TABM-113 (Table 2B) data sets. Most genes contained a single called SFP and the number of SFPs per gene decreased steeply thereafter. Genes that contain many called SFPs are more likely to contain false

Table 1 Sensitivity, specificity and false discovery rate of each of the four single feature polymorphism (SFP) calling algorithms in the E-TABM-113 (A) dataset and BB3 (B) datasets as compared to the barley sequence confirmation dataset

	LM		RATIO		RMA		MAS	
	Filter	No filter	Filter	No filter	Filter	No filter	Filter	No filter
A. BB3								
Sensitivity (%)	68.8	67.3	**76.7**	**74.1**	72.5	70. 6	62%	59.9
Specificity (%)	94.0	94.4	95.0	95.5	**97.1**	**97.3**	96.8	97.0
FDR (%)	36.0	35.6	29.9	29.1	**21.1**	**20.5**	28.4	27.9
B. E-TABM-113								
Sensitivity (%)	51.6	44.6	50.4	38.2	**56.9**	**50.6**	53.3	43.9
Specificity (%)	98	98.2	98.6	**99.2**	**98.7**	98.9	95.2	96.5
FDR (%)	19.5	20.4	14	**11.6**	**11.5**	11.7	37.3	35.5

Category best values are shown in bold font. In 8 of 12 possible categories the RMA procedure outperforms the other 3 algorithms with the RATIO method performing the best in the remaining 4 categories. The Robust Multichip Averaging (RMA) procedure outperformed the MicroArray Suite (MAS) and linear model (LM) procedures in all categories.

Table 2 Frequency of single feature polymorphisms (SFPs) in the BB3 (A) and E-TABM-113 (B) datasets per gene, total SFPs found for the Morex (MX) and Golden Promise (GP) genotypes, total SFPs found across both genotypes and total number of genes containing an SFP

	1	2	3	4	5	6	7	8	9	10	11	MX SEP	GP SEP	Total SFPs	Genes
A. BB3															
LM	1826	570	250	190	120	127	133	123	99	116	68	5273	5279	10552	3622
RATIO	2804	1097	513	273	241	191	128	114	68	33	3	6565	6201	12766	5466
RMA	1921	738	330	172	131	103	98	91	56	53	7	4620	4253	8873	3700
MAS	1556	520	201	148	133	121	123	106	106	73	41	5419	3607	9026	3128
B. E-TABM-113															
LM	1516	447	200	116	100	73	48	41	26	10	2	2794	2638	5432	2579
RATIO	1949	636	233	85	40	18	3	1	1	0	0	2370	2249	4619	2969
RMA	1726	597	254	141	84	44	19	11	3	0	0	2609	2569	5178	2879
MAS	1976	648	301	174	160	143	166	130	112	92	50	4146	7063	11209	3952

LM, linear model; MAS, MicroArray Suite; RMA, Robust Multichip Averaging.

positives and are a warning sign that an accurate estimate of gene expression might not be obtained. As the number of true SFP containing probes in a probe set increases, the ability to accurately estimate gene expression decreases. The number of probes reflecting only expression becomes outnumbered by the number of probes reflecting both expression and genetic influences. Both the MAS and LM models contained a significant number of genes with greater than six called SFPs relative to the RMA and RATIO models. Further, the number of SFPs is evenly split between Morex and GP genotypes in RMA, RATIO and LM models across both data sets, a trend also observed in differential expression. However, the ratio was significantly skewed towards GP in the MAS method in the smaller E-TABM-113 data set and towards Morex in the larger BB3 data set. The RATIO model also has approximately 50% more called SFPs relative to the RMA model (12 766 to 8873) in the BB3 data set, and the MAS model has approximately twice as many called SFPs relative to all other models in the E-TABM-113 data set.

An important evaluation is the proportion of shared SFP calls across the four methods and conversely the proportion of unique SFPs (a SFP called in that model only). Figure 2 shows a Venn diagram of the overlap of all called SFPs for the two data sets across all four models. A large core of called SFPs exists across both data sets and across all models, with relatively few unique SFPs. The exceptions for unique SFPs are in the MAS and RATIO methods. The MAS method called a large number of unique SFPs in both data sets, 60% and 16% of all called SFPs were unique to the MAS method in the

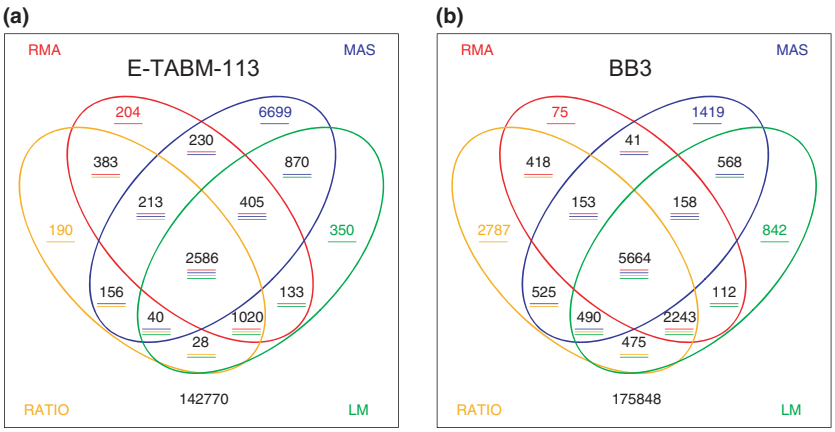

Fig. 2 Venn diagram of the overlap of single feature polymorphism (SFP) calls between the four algorithms and across the two datasets E-TABM-113 (a) and BB3 (b). The value inside a cell represents the number of called SFPs in common between the algorithms represented by the overlap between the ovals. The value outside all ovals is the number of probes without a called SFP in any of the four algorithms. R code available from http://webpages.uidaho.edu/msettles/Rcode/venn.R

E-TABM-113 and BB3 data sets, respectively. The RATIO method called a large number (22%) of unique SFPs only in the E-TABM-113 data set. In comparison, the RMA model had only 4% and 1% unique SFPs in the E-TABM-113 and BB3 experiments, respectively. The RMA model has the greatest overlap, while the LM method was intermediate to the RMA and the other two models, with regard to overlap. In the larger BB3 data set, there existed a significant core of SFPs representing a large portion of all called SFPs (64% RMA, 63% MAS, 44% RATIO, 54% LM. When not considering the MAS method, the remaining three methods showed an even more significant core of called SFPs (89% RMA, 62% RATIO, 75% LM). In the smaller E-TABM-113 data set, the overlap was less significant between the four methods (50% RMA, 23% MAS, 56% RATIO, 48% LM). Again not considering the MAS method, the overlap between the three remaining methods increases significantly again (70% RMA, 78% RATIO, 66% LM). Interestingly, the RATIO method has considerable overlap with the RMA and LM methods in the smaller E-TABM-113 data set, but has a lower overlap and a large number of unique SFP calls in the larger BB3 data set. Overall, the RMA method produced results that overlapped the most and had the fewest unique SFP calls as compared to the other three models across both data sets.

SFP call agreement between BB3 and E-TABM-113

Both BB3 and E-TABM-113 data sets use the same two cultivars (Morex and Golden Promise) and should therefore have similar called SFPs within the common set of genes tested. A total of 154 110 probes (14 010 genes) were tested for SFPs in both data sets. Of these, the percentage of probes with a called SFP in at least one of the two data sets was 8.9% MAS, 7.3% RATIO, 6.5% LM, and 5.6% for the RMA procedure. The agreement between the two data sets also varied greatly. Considering only probes that contained a called SFP in at least one of the two data sets, the agreement for the four models was 28.9% RATIO, 31.9% MAS, 27.5% LM, and 39.5% for the RMA methods. With the majority of disagreements being a called SFP in one data set and a no call in the other, the number of called disagreements (i.e. called as a Morex SFP in one data set and a GP SFP in the other) was relatively few across all four methods (15 RMA, 22 LM, 30 RATIO, and 89 MAS probes).

The relatively low concordance for all four data sets might be explained in the differences between the two data sets, which are in the number of samples, BB3 having five times more microarrays, and in the number of different tissues, BB3 contains five tissues, where E-TABM-113 contains only one. One would therefore expect the BB3 data set to have increased power and subsequently greater ability to detect SFPs. The BB3 data set, however, also has five different tissues and therefore will have five different expression profiles and potentially different gene splicing events. If for instance, a gene containing a SFP was not expressed in one (or more than one, but not all) tissues, it would be difficult for any SFP calling algorithm to account for both differences in hybridization efficiency in expressed tissue because of the SFP and the lack of a signal in the unexpressed tissues.

Comparison of SFP calls to differential expression

In addition to SFP calling, we also performed a differential expression analysis for both data sets. Differential expression analysis resulted in 549 genes up-regulated and 680 down-regulated in Morex relative to Golden Promise in the BB3 experiment (1229 total differentially expressed genes). For the BB3 data set, tissue was included as a blocking factor in the LM. The smaller E-TABM-113 experiment resulted in 760 genes up-regulated and 1043 genes down-regulated in Morex relative to Golden Promise (1803 total differentially expressed genes).

Of the 1229 genes differentially expressed in the BB3 data set, 74.3% (913 genes) contained at least one SFP when computing SFP calls using the RMA model. Conversely, of the 3700 genes containing at least one SFP, 24.7% were also differentially expressed. In the E-TABM-113 data set, of the 1803 total genes differentially expressed 53.6% (967 genes) contained at least one SFP, and of the 2879 genes containing at least one SFP, 33.5% were also differentially expressed. In general, as the number of probes within a gene called as an SFP increased, the likelihood that the gene was also differentially expressed also increased. Similar patterns were seen in both the MAS and LM models (Fig. 3 and Fig. S2, Supporting information). In the RATIO model ,however, differential expression and increased probes with SFPs did not appear to be associated with each other, and no relationship was observed.

Discussion

We have described two variations of a new algorithm for the prediction of SFPs from standard Affymetrix microarray gene expression experiments. We compared our two variants to two previously published methods for the prediction of SFPs between two barley cultivars (Morex and Golden Promise) across two existing barley microarray gene expression data sets. The differences between the two data sets are that one (BB3) contains more samples (30 microarrays) across five tissue types and the other (E-TABM-113) has a smaller number of

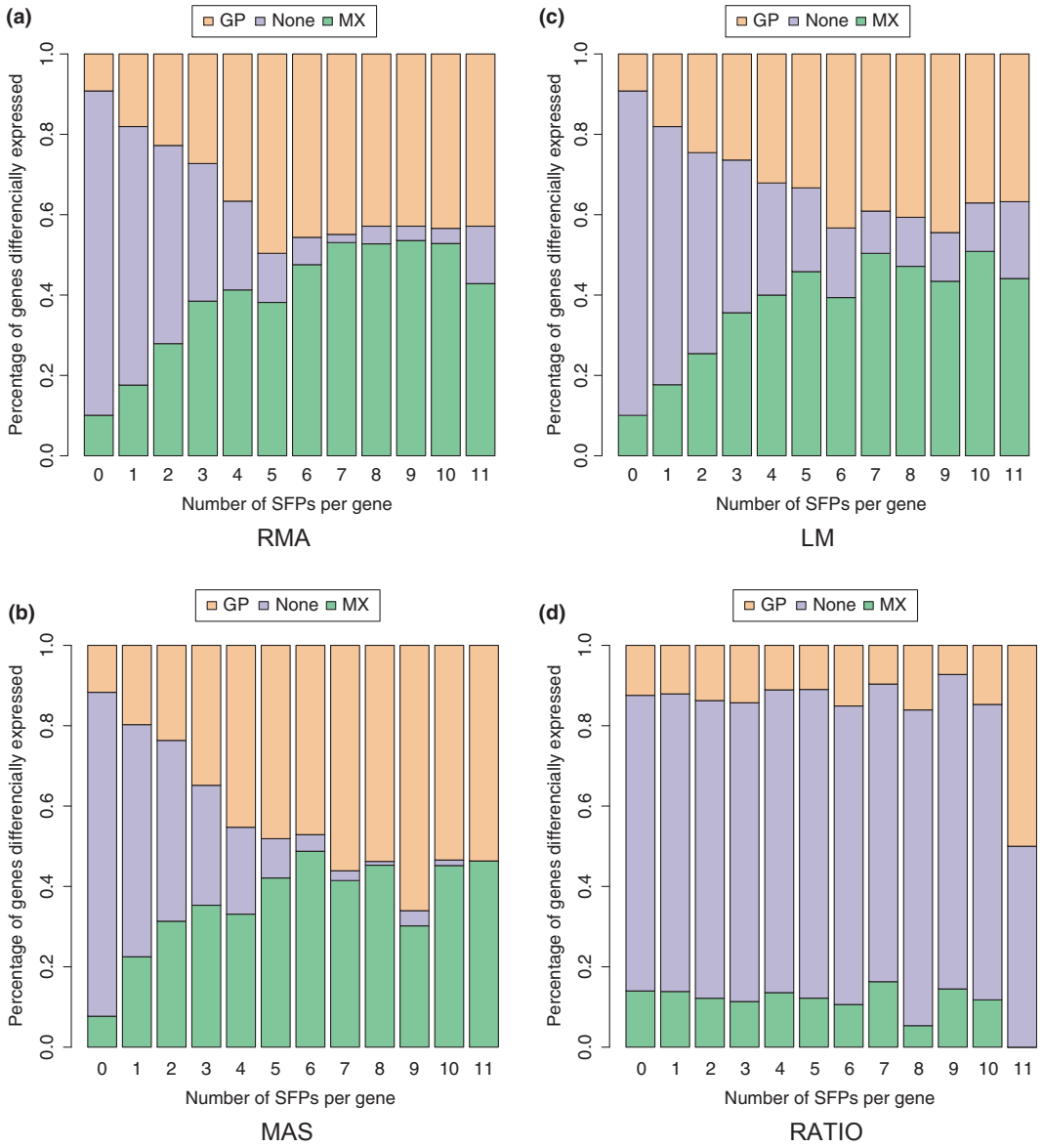

Fig. 3 Comparison of differential expression and single feature polymorphism (SFP) calling in the BB3 dataset. The *y*-axis gives the percentage of differentially expressed genes as the number of called SFPs, within the probe set, increases (*y*-axis)

samples (six microarrays) and only one tissue type. The E-TABM-113 data set, however, represents a common experimental design of a direct comparison of two genotypes across a single factor using a small number of microarrays. Any SFP detection technique should also be evaluated on this type of experimental design. Our algorithm, using the RMA summarization routine (RMA model), produced the overall best results across both data sets for sensitivity, specificity and false feature polymorphism rate against a database of known SNP differences between the two genotypes. Further, the RMA model for calling SFPs was the most conservative and consistent across all evaluated statistics and both data sets.

Relatively, few genes containing at least one SFP (24.7% BB3, 33.5% E-TABM-113), using the RMA method, were also called as being differentially expressed. This would imply that most SFPs are not cis-acting SFPs and are not associated with the gene's expression. However, when a gene is differentially expressed, a majority (74.3% BB3, 53.6% E-TABM-113) were also found to contain at least one called SFP. These SFP are potential candidates for cis-acting regulators that impact gene expression and may be important ecological markers.

Results for the LM model can also be compared to the original study as one of two data sets is the same (BB3). We were able to maintain the sensitivity, (67.3% in this study vs. 67% in the original studies) while decreasing the false positive rate from 40% to 35.5%. The slight sensitivity gain and lower false discovery rate are probably because of the use of the limma procedure for evaluating significance in differential hybridization rather than the significant analysis of microarrays (SAM) procedure (Tusher *et al.* 2001) used in the original study. The limma procedure further allows for a more standard choice in significance cut-off value (BH $\leq$ 0.05) rather than the SAM empirical p-value cut-off of $\leq$0.001 used in the original study, with approximately the same number of called SFPs (10 552 limma, 10 504 SAM). The RATIO method's original study did not use the same data set, so a direct comparison of the results is not possible.

As the results and the model algorithms indicate, the ability of an algorithm to successfully predict SFPs is largely dependent on two factors; the location of the SNP within the probe and the accuracy of the expression estimate. If the polymorphism occurred in the outside three bases (position 1–3 and 23–25) of the probe, the likelihood of detection was reduced approximately threefold. Factors that may impact expression estimates are the assumption that all probes within a probe set have a single target and the same target as the other probes within the probe set (i.e. cross-hybridization and nonspecific hybridization are rare). Therefore, before SFP prediction, care should be taken to update the probe to transcript mapping, ensuring that that each probe belongs within the probe set to which it is assigned and that it has a unique target. In addition, as the number of probes containing a polymorphism increases within a probe set, the likelihood that the corresponding estimate of gene expression will not represent the true level of gene expression also increases. Poor estimation of gene expression will lead to an increase in the false positive rates of both differential expression analysis and SFP prediction. Within the BB3 data set, as the number of SFPs within a gene increased, the relative frequency of the majority genotype to the minority genotype increases towards one (i.e. more even), this is likely to indicate an entire genotype's expression profile has been shifted (for example, see Fig. S3, Supporting information). However, these genes are candidates for splice variation, multi-probe INDELs and/or polyadenylation differences; these probes can then be mapped to exons for possible discovery of these types of polymorphisms. Finally, SFP prediction should be limited to only those probe sets where both genotypes are expressed (i.e. called present), and many of the observed errors can be attributed to a likely unexpressed transcript in one genotype (data not shown).

We provide a new algorithm (using RMA) for prediction of SFPs from standard expression microarray data sets. The algorithm is simple, fast to implement and produces results that are superior to the comparison algorithms. The algorithm can be applied to small data sets and be expected to perform well with a low FDR. Results indicate that the RMA algorithm is an effective technique for studying both gene expression differences and genetic polymorphisms in ecological microarray studies across populations.

Acknowledgements

The authors would like to acknowledge Ms. Maia Benner, Dr. Robert Heckendorn and Dr. Chris Williams for their comments and edits in preparation of this manuscript. This project was supported by grants from the National Center for Research Resources (5P20RR016448-10), the National Institute of General Medical Sciences (8 P20 GM103397-10) from the National Institutes of Health and the NSF Idaho EPSCoR Program and by the National Science Foundation under award number EPS-0447689.

References

Affymetrix (2004) GeneChip Data Analysis Fundamentals Manual [computer software manual].

Bay LK, Ulstrup KE, Nielsen HBR *et al.* (2009) Microarray analysis reveals transcriptional plasticity in the reef building coral acropora millepora. *Molecular Ecology*, **18**, 3062–3075.

Benjimini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, **57**, 289–300.

Bernardo AN, Bradbury PJ, Ma H *et al.* (2009) Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays. *BMC Genomics*, **10**, 251.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Borevitz JO, Liang D, Plouffe D *et al.* (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research*, **13**, 513–523.

Childs LH, Witucka-Wall H, Günther T *et al.* (2010) Single feature polymorphism (SFP)-based selective sweep identification and association mapping of growth-related metabolic traits in *Arabidopsis thaliana*. *BMC Genomics*, **11**, 188.

Close TJ, Wanamaker SI, Caldo RA *et al.* (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiology*, **134**, 960–968.

Coram TE, Settles ML, Wang M, Chen X (2008) Surveying expression level polymorphism and single-feature polymorphism in near-isogenic wheat lines differing for the Yr5 stripe rust resistance locus. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, **117**, 401–411.

Cui X, Xu J, Asghar R *et al.* (2005) Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. *Bioinformatics*, **21**, 3852–3858.

Druka A, Muehlbauer G, Druka I *et al.* (2006) An atlas of gene expression from seed to seed through barley development. *Functional & Integrative Genomics*, **6**, 202–211.

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Gentleman RC, Carey VJ, Bates DM *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**, e15.

Irizarry RA, Hobbs B, Collin F *et al*. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kammenga JE, Herman MA, Ouborg NJ, Johnson L, Breitling R (2007) Microarray challenges in ecology. *Trends in Ecology & Evolution (Personal Edition)*, **22**, 273–279.

Kim S, Zhao K, Jiang R *et al*. (2006) Association mapping with single-feature polymorphisms. *Genetics*, **173**, 1125–1133.

Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, Nguyen HT (2007) Single feature polymorphism discovery in rice. *PLoS ONE*, **2**, 9.

Luo ZW, Potokina E, Druka A, Wise R, Waugh R, Kearsey MJ (2007) SFP genotyping from affymetrix arrays is robust but largely detects cis-acting expression regulators. *Genetics*, **176**, 789–800.

MAQC Consortium (2010) MAQC-II: analyze that! *Nature Biotechnology*, **28**, 761.

Oleksiak MF, Churchill GA, Crawford DL (2002). Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261–266.

Parman C, Halling C (2009) affyQCReport : a package to generate QC reports for affymetrix array data.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing [Computer Software Manual]*. Vienna, Austria.

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

Rostoks N, Borevitz JO, Hedley PE *et al*. (2005) Single feature polymorphism discovery in the barley transcriptome. *Genome Biology*, **6**, R54.

Shi L, Reid LH, Jones WD *et al*. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**, 1151–1161.

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3.

Smyth GK (2005) Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W), pp. 397–420. Springer, New York.

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116–5121.

West MAL, Leeuwen H, van Kozik A *et al*. (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Research*, **16**, 787–795.

Winzeler EA, Castillo-Davis CI, Oshiro G *et al*. (2003) Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics*, **163**, 79–89.

Zhang Y, Ferreira A, Cheng C, Wu Y, Zhang J (2006) Modeling oligonucleotide microarray signals. *Applied Bioinformatics*, **5**, 151–160.

M.S. conceived the experiment, performed bioinformatic analysis and wrote the manuscript. T.C. contributed to experimental design. M.S., T.C., B.R., and T.S. contributed to writing, reviewing and editing of manuscript.

Data Accessibility

Barley data sets were obtained from plexdb (http://www.plexdb.org, Experiment ID: BB3) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress, Experiment ID: E-TABM-110).

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Fig. S1 Position dependent sensitivity of each algorithm to detection of known SFPs.

Fig. S2 Comparison of differential expression and SFP calling in the E-TAB-M data set.

Fig. S3 Single feature polymorphism detection using the RMA method when multiple probes are called as containing an SFP in the BB3 data set.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.