BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 21 2009, pages 2772–2779
doi:10.1093/bioinformatics/btp492

Gene expression

The effects of probe binding affinity differences on gene expression measurements and how to deal with them

Michael Dannemann[1,†], Anna Lorenc[1,‡], Ines Hellmann[2], Philipp Khaitovich[3] and Michael Lachmann[1,*]

[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Departments of Integrative Biology and Statistics, University of California, Berkeley, CA 94720, USA and [3]Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Yue Yang Road, Shanghai, 200031, P.R. China

Received on April 9, 2009; revised on July 17, 2009; accepted on August 9, 2009

Advance Access publication August 18, 2009

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: When comparing gene expression levels between species or strains using microarrays, sequence differences between the groups can cause false identification of expression differences. Our simulated dataset shows that a sequence divergence of only 1% between species can lead to falsely reported expression differences for >50% of the transcripts—similar levels of effect have been reported previously in comparisons of human and chimpanzee expression. We propose a method for identifying probes that cause such false readings, using only the microarray data, so that problematic probes can be excluded from analysis. We then test the power of the method to detect sequence differences and to correct for falsely reported expression differences. Our method can detect 70% of the probes with sequence differences using human and chimpanzee data, while removing only 18% of probes with no sequence differences. Although only 70% of the probes with sequence differences are detected, the effect of removing probes on falsely reported expression differences is more dramatic: the method can remove 98% of the falsely reported expression differences from a simulated dataset. We argue that the method should be used even when sequence data are available.

Contact: lachmann@eva.mpg.de

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

1 INTRODUCTION

To study the evolution of gene expression one can compare gene expression of species, strains, or populations (Brem *et al.*, 2002; Khaitovich *et al.*, 2004; Lai *et al.*, 2006; Nuzhdin *et al.*, 2004; Vuylsteke *et al.*, 2005). For this comparison to be valid, transcript detection and quantification should be equally efficient for all individuals compared. Otherwise, efficiency differences might be mistaken for differences in expression levels. Thus, when gene expression is compared using qPCR, primers are designed so that they do not cover sequence differences between individuals. Oligonucleotide arrays measure the expression of thousands of genes by binding mRNA molecules to probes. The density of molecules that bind to a probe, a patch of oligonucleotides on the array, indicates the original amount of mRNA present in the sample. Equal efficiency of detection requires that the mRNA targets for a probe are identical across all samples. When the samples to be compared have different transcriptomes, for example, belong to different species, subspecies or genetically different populations, some target sequences will differ between the groups, and thus their probe binding affinity might also differ. This would cause a difference in signal intensity even if no difference in expression level between the targets exists. Such sequence differences between targets are sometimes referred to as 'single-feature polymorphisms' (SFPs; Winzeler *et al.* 1998). Since we also address differences between species in this article, and not just polymorphisms, we will call this difference a 'binding affinity difference' (BAD) and a probe hybridizing with BAD targets, a 'BAD probe'. Since a part of a probe's signal comes from its hybridization with sequences others than the desired target (Binder and Preibisch, 2005), BAD probes can also arise from a difference in the secondary target—either a difference in sequence or in expression level. Finally, BAD probes can also be produced as a result of differences in the splicing of transcripts between the groups.

The impact of BAD probes on expression estimates, when comparing expression between species, was recognized by Hsieh *et al.* (2003), who identify some of the changes in gene expression as artifacts of species-specific probes: when chimpanzee expression is measured on human-specific arrays, a 1% nucleotide difference between species leads to >22% of probes with a sequence difference within them. It is difficult to estimate how serious the general bias introduced by those probes is—Hsieh *et al*. calculated that in the study by Enard *et al.* (2002), species–probe interaction is significant for more than half the genes.

BAD probes will have an especially strong impact on studies doing QTL analysis (Alberts *et al.*, 2007) of gene expression. When an expression difference stems from a sequence difference in a

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

‡Present address: Max Planck Institute for Evolutionary Biology, Ploen, Germany.

probe's target, it will give a very strong signal, mapping exactly to the region in the genome where the difference occurs. It will therefore look like *cis*-regulation when the sequence difference is in the target region of the gene, or as *trans*-regulation when the signal stems from a sequence difference in a secondary target. Alberts *et al.* (2007) use a method similar to the one used by Greenhall *et al.* (2007) to correct specifically for such problems.

How can one overcome this problem? Designing a special microarray for each of the species or strains involved in the study without taking into account the differences in target sequences, does not solve the problem, because in that case the same mRNA target will be measured using different probes and thus with different binding efficiency. One could design arrays with probes whose targets have no sequence difference between the species. Some authors have tried to design arrays for all studied species, and through a cross-hybridization design take the BADs into account (Gilad *et al.*, 2005). The approach that we describe in this article is finding BAD probes, and masking them, i.e. not using them in the expression analysis.

When several probes measure each mRNA target molecule—on Affymetrix gene expression arrays usually 11 or 16 probes measure a single mRNA target—it is possible to remove all BAD probes from the probesets used in a study (Khaitovich *et al.*, 2004) and estimate expression levels using only the remaining probes. Different methods for detecting BAD probes have been proposed. One approach uses genomic, mRNA or EST sequence data to identify and remove from analysis (i.e. mask) probes that have a sequence difference within the targets of the probe (Khaitovich *et al.*, 2004), thereby creating a sequence-derived mask. This requires all sequence differences within targets to be known for the compared groups—i.e. species, subspecies, strain or population sequence data. However, such a mask still does not account for any non-primary target sequence differences, as secondary targets remain mostly unknown (Rule *et al.*, 2009). On the other hand, the approach may also be too conservative, since not every probe that differs in sequence from the probe's target necessarily causes a difference in binding affinity (Naiser *et al.*, 2008). Another approach uses expression data to identify BAD probes. Since oligonucleotide arrays measure gene expression using several different probes to determine the expression level of a single mRNA molecule, comparisons between these probes can detect BAD probes. This approach was used by Càceres *et al.* (2003), Khaitovich *et al.* (2004) and Greenhall *et al.* (2007). Ronald *et al.* (2005) couple a similar approach with an analysis of the signal expected based on a probe's thermodynamic properties. Here, we expand and improve upon the method used by Khaitovich *et al.* (2004) to build a mask to remove BAD probes and present an analysis for evaluating the quality of the mask produced. We developed an R package to detect BAD probes in Affymetrix Gene Chip array data. The R package implementing our method and compatible with the Bioconductor package for analyzing microarray data can be downloaded from http://bioinf.eva.mpg.de/masking/. On the same web page, we also provide scripts for producing all results reported here.

One method to evaluate the efficiency of the masking algorithm is comparing it with a sequence-derived mask. However, since a sequence-derived mask does not remove any probes that are BAD due to secondary target differences, it only approximates an ideal mask. We also do not know the 'real' expression differences in the samples to which we could compare our results. To overcome this problem, we generated datasets in which the real expression differences are known. We use evaluation datasets in which we artificially create BAD probes, replacing the signal from perfect matching (PM) probes by the signal from their coupled mismatch (MM) probes. Since the expression differences are known in the original datasets, we can evaluate how well our mask recovers the original expression differences.

2 METHODS

Our method for detecting BAD probes uses an Affymetrix array design feature: the probeset. A probeset is a set of probes, all of which are designed to measure the same target transcripts, but each of which binds to a different region of the mRNA molecule. If the probes truly measure the same transcripts, the signal intensities of probes within one probeset should be correlated to the level of this transcript, and thus to each other. When there is no difference in binding affinity between groups of samples, this correlation will be the same for all groups. When binding affinities differ between groups, signal intensities remain correlated within groups, but the correlation differs between groups (Fig. 1). Notice that when several alternative transcripts for the target mRNA molecule exist, some of them might not contain targets to all the probes. In that case targets that contain two probes contribute to their correlation, and a target molecule that contains one but not the other will add noise on top of this correlation. In that sense the effect of alternative transcripts is similar to that of secondary targets, except that instead of adding expression, expression is subtracted.

In the following, we use Li and Wong's (2001) model for signal intensities. According to this model, the signal intensity for the i-th array and j-th probe

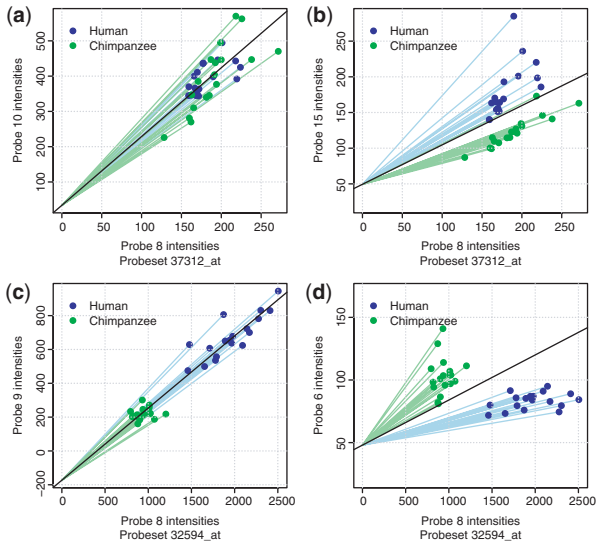

Fig. 1. Comparison of fluorescence level between two probes that measure the same mRNA target molecule—belonging to the same probeset. Each dot represents a sample—humans (blue) and chimpanzee (green). On the left, (**a**) and (**c**) relative fluorescence level when there is no sequence difference between humans and chimpanzees in either probe. In this case the relationship of fluorescence level between probes is expected to be linear. On the right, (**b**) and (**d**) probe comparison for the same probesets, but the probe on the y-axis has a sequence difference. On top, for probeset `37312_at`, there is no detectable expression difference between humans and chimpanzees, on the bottom, for probeset `32594_at` there is a difference. Our method performs a t-test of the slopes to each point (green and blue lines), assuming that the intercept is taken from all points (black line).

can be expressed as:

$$O_{ij} = \nu_j + \Phi_i \theta_j + \epsilon \quad (1)$$

Where ν_j is the baseline response of the j-th probe due to non-specific hybridization, Φ_i is the abundance of the target RNA sequence, θ_j is the rate of increase of probe j to the target sequence, and ϵ is an error term. We can rearrange this equation, where probes $j = 1$ and 2 produce:

$$O_{i2} = \frac{\nu_2\theta_1 - \nu_1\theta_2}{\theta_1} + \frac{\theta_2}{\theta_1} O_{i1} + \epsilon \quad (2)$$

Therefore, assuming the binding strength of probes 1 and 2 are equal across all samples and if the background binding level for the two probes is also identical, then the expression level measured by probe 2 will be a linear function of the expression level measured by probe 1. When samples come from different species, the binding affinities (θ_j) or background levels (ν_j) could be different for different arrays. In that case, there will be a different linear relationship between the signals O_{i1} and O_{i2}, or between probes 1 and 2 in the two species, but a linear relation will remain for within species comparisons.

The model used above is not an exact model for the fluorescence levels in microarrays, as these measurements are not linear at the target mRNA levels. Zhang *et al.* (2003) model this probe response function. The baseline response is also not constant it depends on cross-hybridization—additional transcripts that bind with a lower affinity to the probe. Finally, the error term is not known to have the same distribution across the whole range of expression. We therefore need to construct a method that is robust to these deviations. Since we wish to detect expression differences between species, the method must also be robust to real differences in expression levels between the species.

We tested several different methods (see Supplementary Material). Of these, the following test gave the best results. Our null hypothesis considers that in both species the two sequences compared are the same, as well as the background binding level. We estimate the intercept

$$\beta = \frac{\nu_2\theta_1 - \nu_1\theta_2}{\theta_1} \quad (3)$$

from the data using the Reduced Major Axis regression (RMAr—not to be confused with the Robust Multichip Average method used to analyze microarrays). If the RMAr slope is negative, the minimal value for probe 2 is taken as the intercept. We then have:

$$\frac{O_{i2} - \beta}{O_{i1}} = \frac{\theta_2}{\theta_1} + \epsilon \quad (4)$$

We can therefore test if $(O_{i2} - \beta)/O_{i1}$ has the same distribution for both species. This test will give us a P-value for the hypothesis that the two species have the same binding strength and background binding level for probes 1 and 2. When the hypothesis is rejected, we do not know which of the two probes has a difference in binding strength or background. So, we then perform the same test for all probe pairs. As the test result is not symmetric in the two probes used, we conduct the tests in both directions. We thus build a $J \times J$ matrix of all pairwise tests, less the diagonal.

A probe that has a BAD across the species cannot inform us about other probes, since the null hypothesis does not hold. Therefore, we remove probes from the probeset one by one and repeat the tests. To obtain a single value for each probe in a probeset, we use the following algorithm:

(1) For each probe, calculate the geometric mean of all P-values in the matrix where this probe is involved, and ignore comparisons of the probe with itself.

(2) Record the probe with the smallest geometric mean of P-values (we will call this mean the mP-value, to distinguish it from a real P-value). Exclude comparisons with this probe from the matrix.

(3) Repeat Step 1, until the matrix contains only two probes.

(4) Assign the last two probes the same mP-value, which is the geometric mean of their P-values.

Once each probe has been assigned an mP-value, we choose a cutoff and designate all probes with an mP-value below this cutoff as BAD probes.

2.1 Choice of the cutoff

The last step when constructing an expression-based mask involves masking all probes with an mP-value below a certain cutoff. The choice of this cutoff depends on the goal of the analysis. For detecting candidate sequence differences between species, where a strong type 2 error control is important, we might be more concerned that all differences reported are indeed sequence differences, than our concern that some sequence differences are missed. In this case a low cutoff is reasonable. In contrast, when masking BAD probes in between-species comparisons of gene expression, it is important to mask as many BAD probes as possible, eliminating any probes that cause false expression readings, as long as a sufficiently high number of probes for a reliable expression analysis is retained. Since mP-values depend on the particular dataset, an individual cutoff must be chosen for each dataset. In the next section, we outline our assessment of each cutoff by evaluating its effects on detecting differential gene expression. We will demonstrate that a good cutoff selection is the one that eliminates a fraction of probes close to the expected number of differences between the species. An alternative strategy for choosing the cutoff is to sequence some of the probes, and then use these data to calculate types 1 and 2 errors for different cutoffs, and select the desired cutoff.

3 RESULTS

3.1 Evaluating the expression-based mask

3.1.1 Comparison with a sequence-based mask First, we tested our mask's ability to detect probes with known sequence differences. We used a human–chimpanzee brain expression dataset from U95A arrays (Khaitovich *et al.*, 2004), comprising six different brain tissues from three individuals for each species. This array has 16 probes in every probeset. Notice that for constructing a mask, pooling different tissues should not pose a problem as long as differences in expression levels of secondary targets do not generate so much noise as to mask out the signal we are looking for. All probes on the U95A array were mapped to the human and chimpanzee genomes and cDNA (see Supplementary Material). For our comparison of sequence results with the expression-based mask, we used a conservative definition of matching and mismatching probes, so as to minimize mis-calling either (see Supplementary Material). We estimated type 1 errors—the fraction of probes undetected by our expression mask among all probes with a sequence difference, and type 2 errors—the fraction of probes identified as different by our mask among all probes with no sequence difference between the species. As illustrated in Figure 3, our expression-based mask detects >70% of probes with a sequence difference, while at the same time removing 18% of probes without a known sequence difference (cutoff = 0.05, masking 26% of the probes). At this cutoff, an average of 3.3 probes are removed from a probeset, leaving an average of 12.7 (See Supplementary Fig. 10 for the distribution of probes left per probeset). In comparison, our method is more powerful than Greenhall's method (Greenhall *et al.*, 2007), since we detect ~10% more probes with a sequence difference, while maintaining the same level of type 2 errors (Supplementary Fig. 7).

A sequence-based mask only masks probes where the primary target differs, but does not consider differences caused by the cross-hybridization of secondary targets between the two species. Thus, the intended target probe might have the same sequence in both species, but one of the cross-hybridizing targets might have a changed sequence or a changed expression level. In this case, an expression-based mask might detect a BAD probe that a sequence-based mask

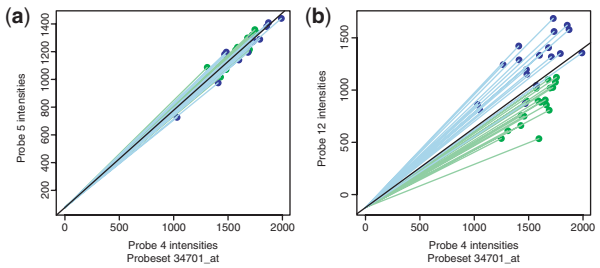

Fig. 2. An example of a BAD probe not detected by a sequence-derived mask based on human and chimpanzee data, in probeset `34701_at` measuring expression of gene DLG4 (post-synaptic density protein 95). (**a**) Probe 5 versus probe 4: no BADs between the species can be seen. (**b**) Probe 5 versus probe 12: even though none of the probes have a sequence difference between species, it is clear that probe 12 displays a difference between the species that is not the result of a difference in the expression level of DLG4. This probe could cause a spurious expression difference between the species.

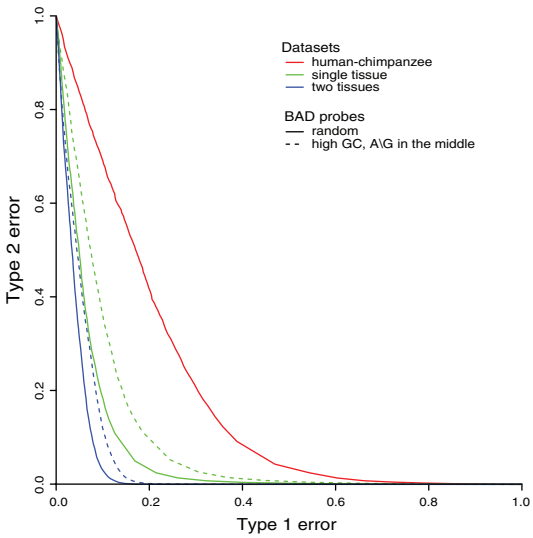

Fig. 3. Power of detecting sequence differences using an expression-based mask. The *x*-axis, type 1 error, refers to the fraction of probes without a sequence difference, that are still detected as BAD by the method. The *y*-axis, type 2 error, refers to the fraction of probes with a sequence difference that are not detected as BAD by the method. Shown are power curves for detecting BAD probes for the human–chimpanzee dataset, and for the two simulated datasets. Dashed lines are simulated datasets in which only the probes that were the most difficult to detect as BAD were used (highest GC content among probes with an A/G in the middle of the probe).

misses; therefore, not all type 2 errors of the expression-based mask are actual errors. An inspection of such probes often shows clear differences in binding affinity (Fig. 2). For such probesets, the assumptions underlying the calculation of expression levels using, for example, the RMA method, do not hold. Such probes should be excluded when calculating expression levels.

Similarly, not all type 1 errors have the same consequences. A sequence-based mask only points to a sequence difference, but not its effect on the probe's signal. A sequence change at the edge of a probe might have negligible effects on binding affinity and expression estimates. Indeed, we find that the position of the MM in the probe significantly impacts our ability to detect a sequence difference—changing from a detection rate of ∼30% at the edges to 80% in the middle of the probe (see Supplementary Fig. 2).

3.1.2 Datasets with artificially created BAD probes For a better estimate of our method's power to improve detection of expression differences, we constructed simulated datasets. In a subset of samples, we introduced probes with a 1 nt MM to the target. Affymetrix arrays contain probes matching the target perfectly (called PM probes), as well as probes containing a single MM to the target (called MM probes). MM probes are identical to the PM probes, except for a change in the middle nucleotide. In half of a dataset's samples, in a fraction of probes (e.g. 20%), we exchanged the PM probe intensity for the MM probe intensity, and thus created artificial BAD probes ('flipped probes'). We then used those artificial datasets to create masks, and compared expression values obtained before and after masking. Since PM and MM values are used for deciding whether a probeset is expressed, we used the 'expressed' calls from the original dataset.

The first dataset, the 'single-tissue dataset', consists of 30 samples of healthy human prefrontal cortex, arbitrarily divided into two groups (Ryan *et al.*, 2006). In one of the two groups, group 1, we randomly selected 20% of the probes, and replaced the PM values with the corresponding MM values. This is the percentage of BAD probes one would obtain from a sequence difference of ∼1% per nucleotide. The other half of this dataset, group 2, remained unchanged.

The second dataset, the 'two-tissue dataset', comprises expression data for two human brain regions (caudate nucleus and frontal cortex), from 12 individuals each (Hodges *et al.*, 2006). Here, we replaced the PM with MM values for the caudate nucleus samples. As two brain regions differ in their expression pattern, the second dataset mimics two genetically distinct groups differing in gene expression.

In the simulated datasets, types 1 and 2 errors of detecting the flipped probes for almost any cutoff are smaller than for the human–chimpanzee mask compared against sequence mask (Supplementary Fig. 3). This is mainly because in the simulated dataset, we replaced always the middle nucleotide, whereas the sequence differences in the human–chimpanzee data occur at any position, and also because MM probes always replace A↔T and C↔G, whereas for species differences all possible replacements occur. Supplementary Figure 3 shows that our ability to detect flipped probes in the simulated dataset is a function of the type of change in the MM probe and the GC content of the remainder of the probe. We can see that probes with a low GC content and central A/G in PM probes are detected best, whereas probes with a high GC content and central C/T are the most difficult to detect. Our greater ability to detect a difference in probes with a low GC content could imply that these changes have a larger effect on the binding affinity. This larger difference in affinity for A/G changes was observed by Binder and Preibisch (2005).

Because flipping probes with a high GC content and a central C or T produce error profiles closest to those produced in the human–chimpanzee data, we decided to only use such probes in the simulations described below, in order to get closer to the error profiles of real sequence differences. The results for randomly selected probes were similar (see Supplementary Material).

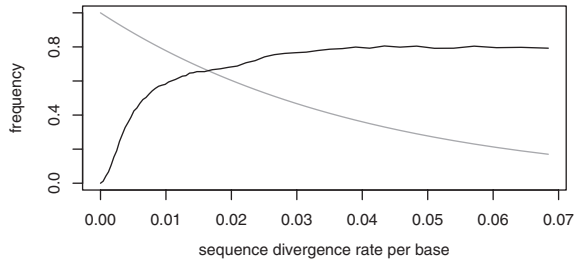

Fig. 4. Black: rate at which probesets gain spurious expression differences versus sequence divergence rate per base. In the single-tissue dataset, we flipped a number of probes equivalent to the sequence divergence rate shown on the *x*-axis (per base). Then we measured what fraction of probesets showing no difference in the original dataset show an expression difference after flipping (*t*-test *P*-value < 0.05). Notice that in this plot, we first applied flipping only to high GC, middle C/T probes, but as those ran out to all probes. Red (grey): fraction of probes without a sequence difference versus divergence rate per base.

3.1.3 Influence of masking on gene expression estimates We used the simulated datasets to study the influence of BAD probes and masking on expression differences. We can compare our results to both the expression differences observed before the probes were flipped, and to the results one would obtain with a 'perfect' mask, since we know exactly which probes should be removed (exactly those that were flipped). Notice that after applying this perfect mask, which eliminates all flipped probes, we do not obtain exactly the original expression values and expression differences. This is because even expression differences detected with the 'perfect' mask have some 'error' versus the original expression data, both due to the loss of power to detect differential expression when probes are discarded from probesets and to the noise in expression levels caused by using different subsets of probes.

In the single-tissue dataset, group averages should be the same, since all samples come from one population. Therefore, we examined only the newly introduced expression differences, but not the effect flipping may have on the few false positive expression differences that were present in the original data. First, we look at the effect of BAD probes on false detection of expression differences. Figure 4 shows the fraction of probesets that had no expression difference in the original dataset, which showed a difference after we flipped some of the probes. We can see that at a sequence divergence of 1% between the groups, which corresponds to ~20% of probes flipped, 55% of the probes that originally showed no difference between the groups show a difference (*t*-test, at $P < 0.05$ level). This number is similar to the result estimated by Hsieh *et al.* (2003).

Almost all of these introduced differences disappear after masking, and the expression-based mask is as effective as the 'perfect mask'—only 2% of the probesets without an original difference are marked as differentially expressed (Fig. 5a).

For the two-tissue dataset, we can consider both types of error. After flipping, >78% of probesets that showed no difference between the tissues, now show a significant difference. This is reduced to 27% after masking, compared with 8% with a perfect mask (Fig. 5b). Of the probesets that had a significant difference in the original dataset, 18% lose their significance after flipping. This is reduced to 10% of probesets with the expression-based mask, and 5% with the perfect mask.

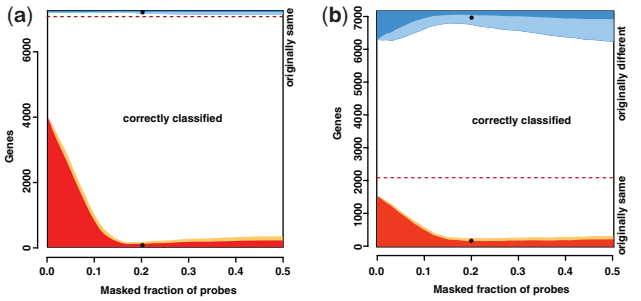

Fig. 5. Detecting differential expression with masked datasets. The *y*-axis represents all genes. Above the dashed line are differentially expressed genes, below are non-differentially expressed—in the original dataset. The *x*-axis represents the fraction of genes removed by masks with different cutoffs. The white area represents correctly classified genes. The shaded area, above the dashed line, represents misclassified genes—genes that were originally different, but after masking are classified as not differently expressed. The shaded area below the dashed line represents genes that were originally not different, and after masking are classified as different. In both the cases, we distinguish those genes where the misclassification occurred already after flipping (dark shaded), and those where the misclassification was only introduced after masking, and did not occur after just flipping (light shaded). Black dots: perfect mask—exactly those probes are removed to which sequence differences were introduced. (**a**) Single-tissue dataset with 20% of probes simulated as BAD. (**b**) Two-tissue dataset with 20% of probes simulated as BAD.

Why is there a difference between the two simulated datasets? In the single-tissue dataset, after masking, there are almost no new expression differences, whereas in the two-tissue dataset, after masking, 27% of the probesets without any original difference in expression now show a difference. One possibility is that when there are real differences in expression between the groups, our method has less power to detect a BAD probe. Another possibility is that our mask removes probes with a different cross-hybridization profile between the tissues (present only in the two-tissue dataset) and by doing that increases the power to detect differences between the groups.

To examine these possibilities, Figure 5 distinguishes between errors in detecting expression differences that were introduced by the mask and those errors that appeared when the BAD probes were introduced, when MM and PM were flipped. We can see that most of the errors that were introduced by flipping are removed by the mask—it performs almost as well as the perfect mask—4% versus 5%. The reason for the large difference between the two datasets, therefore, is not reduced power. Instead, it might be that masking makes more expression differences between the tissues apparent, an effect unrelated to flipping. In fact, most of the new expression differences we see after masking of flipped data, are also observed when we run masking algorithm for the raw (unflipped) two-tissue dataset and apply the resulting mask on it. We hypothesize that this is because masking removes probes that have a difference in cross-hybridization profiles between the tissues, present already in the raw data. After removing these probes, the power to see expression differences between the tissues increases.

3.1.4 Comparing gene expression between strains In populations with a large effective population size, the sequence difference between groups can be large enough to introduce a large number

of transcript differences, and thus a large number of spurious expression differences. One can estimate this effect by looking at Figure 4. At 1% sequence divergence between groups, as we would see, for example, between strains of *Saccharomyces cerevisiae* (Liti *et al.*, 2009) and strains of *Drosophila melanogaster* (Aquadro *et al.*, 2001), >50% of genes without a significant expression difference will be falsely identified as having such a difference. The effect decreases when the divergence between groups is down to 0.1%—only 5% of genes are falsely classified.

We applied our method to a dataset consisting of two strains of mice, C57BL/6J and A/J taken from Hovatta *et al.* (2005). This dataset consists of expression from seven brain regions: amygdala, bednucleus of the stria terminalis, cingulate cortex, hippocampus, hypothalamus, periaqueductal gray and pituitary gland, with two replicates each. The study used the U74Av2 array, in which the large majority of probesets contain 16 probes. We looked for BAD probes in the comparison between the strains. Our masking method is able to identify 60% of a set of 313 known SNPs (downloaded from the Mouse Genome Database, Bult *et al.* 2008, http://www.informatics.jax.org in June 2009) within probe target regions between these strains, while masking only 8.3% of all probes. In fact, of the 400 probes with the lowest m*P*-value, a third are among these known SNPs. As was the case in our other datasets, we can see obvious differences between the groups in correlation between some of the probes—both for probes with known SNPs in them, and for probes without known SNPs (Supplementary Fig. 1).

The divergence between these strains is ∼0.08% (Frazer *et al.*, 2007), and the divergence at the target regions for the probes is probably lower. At this divergence level, BAD probes have a low effect on expression differences. Of the 3430, 1163 probesets that were significant at $P < 0.05$ for the strain effect in a strain by tissue ANOVA, 249 become non-significant after masking, and 155 new probesets become significant. It is obvious that at such a low divergence between groups applying the mask is not crucial, but might improve the quality of the data.

As we mentioned in Section 1, when a study is mapping QTL for gene expression between strains, a spurious expression difference that stems from a BAD probe—in particular a sequence difference in the target molecule between the strains, or a sequence difference in a cross-hybridizing molecule, will give a strong signal, since it will map exactly to the sequence difference that causes it.

3.1.5 Effect of number of BAD probes We looked at the ability to detect BAD probes. We flipped 10%, 20% and 30% of the probes in the single- and two-tissue datasets. Note that by increasing the percentage of flipped probes, we are also increasing the average number of BAD probes per probeset. The overall effect on the detection rate of BAD probes was minimal (Supplementary Fig. 6). In all the cases, the number of removed probes required to eliminate errors in reported expression levels corresponds to the number of probes flipped 10%, 20% or 30%.

3.1.6 Influence of number of probes in a probeset Standard Affymetrix probesets contain usually 16 or 11 probes. We checked how our method performs for probesets of different size. This was done by artificially creating probests that contain fewer probes, and measuring the error rate in them. With three and five probes per probeset the error rate is significantly increased, but the effect for seven probes per probeset is already very small (see Supplementary Fig. 9). One can also infer that the additional power gained from going beyond 16 probes per probeset will be very small.

3.1.7 Data needed to build a reliable mask We tested how the size of the groups used to build a mask influences error rates for human–chimpanzee dataset (Supplementary Fig. 4). Using more individuals also increases the power to detect expression differences (Supplementary Fig. 5). We see that beyond six individuals, the effect of increasing the number of individuals per group is negligible.

4 DISCUSSION

Large-scale measurement of gene expression provides an invaluable tool for studying the phenotype of organisms. Comparing expression of orthologous genes or transcripts across species gives important insights into the evolution of their phenotypes. However, since expression differences are not the only difference between the species compared, we must ensure that the detected differences are indeed in the expression of the target transcripts, and not the result of some other difference between the species, such as the thawing rate of tissues, rate of RNA degradation or large-scale differences in the mRNA profile, which would invalidate our normalization assumptions. Methods that measure gene expression by binding molecules to an oligonucleotide probe will also misidentify expression level differences between transcripts because of sequence differences in the targets transcripts or cross-hybridizing transcripts or because of differences in the expression profile of such cross-hybridizing transcripts. In some cases custom arrays designed for each of the species compared are available. If we measure expression using these arrays, we are not measuring the expression levels using the same probe, and thus the relationship between fluorescence level and mRNA expression level will be different between the species. In this case, the differences detected will in most cases be probe differences.

There are two principal ways to overcome these obstacles and reliably measure expression differences between species: (1) experimentally decreasing or negating the binding differences to the target, and (2) only using targets without any binding differences. Methods based on (1) include application of longer targets sequences in the hope of washing out the effect of a few sequence differences (Walker *et al.*, 2006), or using probes with the sequences of targets in both (or all) species, and including binding affinity in the statistical analysis model (Gilad *et al.*, 2005). Our article, focuses on approach (2), only using probes with no binding differences to their targets—and masking out all the rest. Masking BAD probes has some shortcomings. One is that the number of usable probes decreases as sequence divergence between the species increases. At a divergence of 1% per nucleotide in the target region, ∼80% of 25mer probes have no sequence difference and can be used; at 5% divergence <30% of the probes have no sequence difference, and at 10% divergence, only 7% of the probes are usable (Fig. 4). The number of usable probes further decreases if we include more than two species in our comparison, as we then need to mask probes that have a binding difference for their targets in any of the species. We must also ensure that less conserved genes, which potentially have more sequence differences, do not preferentially drop out of our analysis, and do not suffer from a greater measurement error. Therefore, applying any kind of masking approach is limited

to closely related species and reasonably conserved genes. The approach is also applicable to expression comparisons between strains or populations, when the number of sequence difference is small.

In this article, we present a method to detect probes with BADs between species based on the expression data itself. Since the method is based on expression, it has no power to detect differences in unexpressed genes. As it does not rely on sequence data, it is especially useful when comparing expression in different subspecies, strains, populations and other genetically distinct groups when not all genetic differences are known.

We also present methods for estimating the efficiency of our method in removing falsely reported positive expression differences between the groups. When there are no or only a few expression differences between the compared groups, such as in our single-tissue dataset, our method is very effective, removing 98% of these false positives. When there are real expression differences between the samples, as was the case in our two-tissue dataset, after masking only 5% of the expression differences introduced by the sequence differences remain—this is as few as would remain with a perfect mask removing all introduced sequence differences. However, in addition, 20% of genes which did not show a gene expression difference—neither in the raw data, nor after using the 'perfect mask'—are detected as differentially expressed after our expression-based masking. Where do these additional differences come from? Since virtually no additional differences are introduced when there are no expression differences of a molecule, we interpret this to stem from an increase in the power to detect expression differences, by removing noise in the dataset. Therefore, expression-based mask is useful not only to avoid spurious expression differences, but also to improve detection of others, unidentified in noisy unmasked data.

We are aware that our method of simulating species-derived BAD probes suffers from a few shortcomings. First, in the MM probes, it is always the central base among the 25 bases in a probe that is changed. In flipping MM and PM readings, we therefore do not cover the full range of affinity differences that are produced by sequence changes on all 25 positions. Second, our simulation scheme flips the probe sequence for one of the groups, but the target sequence in that group stays just as it was before flipping. When comparing expression between species, one is using the same probe for both groups, but the target sequence in one of the groups has a difference. It is not clear what effect this difference has. Third, changes in cross-hybridization have different origins in *in vivo* experiments versus our *in silico* simulation with respect to secondary targets: when comparing between species, false expression differences can result from both sequence changes in secondary targets and from changes in the expression levels of secondary targets. In the simulated dataset, the whole population of secondary targets that bind to the MM versus the PM probes are potentially different (rather than just one of the secondary targets changing sequence). In addition, no expression level differences in the secondary targets are introduced by flipping, whereas in comparing between species, many secondary targets can change their expression. Notice, however, that in the two-tissue dataset, there could be differences in the expression levels of secondary targets that were present even before flipping. Fourth, sequence and splicing differences between the species beyond the probes themselves might affect the binding affinity of the target in real life. These are impossible to be introduced by flipping. Again, in two-tissue dataset such differences between the tissues might be present even before flipping.

We have shown that relatively few samples per group are sufficient to construct a mask—even three already provide some power, which increases until we reach about 10 samples per group. Note that all samples used in the analysis must be run on the same microarray scanner, because we use the fluorescence response curve of the probes, which differs not only from machine to machine, but even across calibrations of the same machine. When insufficient numbers of samples per species from a single tissue are available, several different tissues can be pooled for the purpose of building the mask. For example, our dataset for finding differences between humans and chimpanzees comprises data from five different brain regions, and the mouse dataset contains seven different brain regions. Notice that expression differences between the tissues do not necessarily hamper our ability to construct the mask. We are testing to what degree, when fluorescence level of one probe increases because more target molecules were available, the level of another probe targeting the same molecule will also increase. In such cases it should not hamper the test if the target molecules are present in different levels in the different samples—in fact that is exactly what powers the test. When differences between the tissues is too large, however, differences in expression levels of secondary targets will reduce the power of detection of BAD probes between the species. A similar source for noise within a group is sequence differences between individuals within it. The effect, again, will be that there are BAD within the group, and therefore probes will not lie on a single line. This will mainly be a problem if the same probe has differences between and within the groups, in which case it might not be detected as a BAD probe. It is not clear if such a probe will produce spurious expression differences between the groups, but it might be good to develop methods to detect these cases, when the rate of polymorphism within the groups is high enough.

The expression-based masking we propose, allows us to compare gene expression when insufficient sequence data is available to build a sequence-based mask. Even in datasets where a full sequence-based mask is available, additional masking based on the expression data will provide a benefit (Khaitovich *et al.*, 2004). This is because a simple sequence-based mask only removes probes due to differences in primary targets. Changes in sequence or expression levels of secondary targets can still falsify expression differences. Differences in splicing between the groups could also be identified as differences in expression levels of the mRNA target. Our mask detects cases where some of the assumptions of the methods used to calculate gene expression levels, such as the RMA method, break down. Thus, our method will point to additional problematic probes, and enhance the sequence-based mask.

ACKNOWLEDGEMENT

We thank S. E. Ptak, C. Green, and B. Hinnerk for helpful discussions. Three anonymous reviewers helped in making the article more interesting.

Funding: Max Planck Society.

Conflict of Interest: none declared.

REFERENCES

Alberts,R. *et al*. (2007) Sequence polymorphisms cause many false *cis* eqtls. *PLoS ONE*, **2**, e622.

Aquadro,C. *et al*. (2001) Genome-wide variation in the human and fruitfly: a comparison. *Curr. Opin. Genet. Dev.*, **11**, 627–634.

Binder,H. and Preibisch,S. (2005) Specific and non specific hybridization of oligonucleotide probes on microarrays. *Biophys. J.*, **89**, 337–352.

Brem,R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Bult,C. *et al*. (2008) The mouse genome database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–D728.

Càceres,M. *et al*. (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl Acad. Sci. USA*, **100**, 13030–13035.

Enard,W. *et al*. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.

Frazer,K.A. *et al*. (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, **448**, 1050–1053.

Gilad,Y. *et al*. (2005) Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.*, **15**, 674–680.

Greenhall,J.A. *et al*. (2007) Detecting genetic variation in microarray expression data. *Genome Res.*, **17**, 1228–1235.

Hodges,A. *et al*. (2006) Regional and cellular gene expression changes in human Huntington's disease brain. *Hum. Mol. Genet.*, **15**, 965–977.

Hovatta,I. *et al*. (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature*, **438**, 662–666.

Hsieh,W.-P. *et al*. (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, **165**, 747–757.

Khaitovich,P. *et al*. (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.*, **14**, 1462–1473.

Lai,Z. *et al*. (2006) Microarray analysis reveals differential gene expression in hybrid sunflower species. *Mol. Ecol.*, **15**, 1213–1227.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Liti,G. *et al*. (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.

Naiser,T. *et al*. (2008) Impact of point-mutations on the hybridization affinity of surface-bound DNA/DNA and RNA/DNA oligonucleotide-duplexes: comparison of single base mismatches and base bulges. *BMC Biotechnol.*, **8**, 48.

Nuzhdin,S.V. *et al*. (2004) Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol. Biol. Evol.*, **21**, 1308–1317.

Ronald,J. *et al*. (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.*, **15**, 284–291.

Rule,R.A. *et al*. (2009) Use of hidden correlations in short oligonucleotide array data are insufficient for accurate quantification of nucleic acid targets in complex target mixtures. *J. Microbiol. Methods*, **76**, 188–195.

Ryan,M.M. *et al*. (2006) Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Mol. Psychiatry*, **11**, 965–978.

Vuylsteke,M. *et al*. (2005) Genetic analysis of variation in gene expression in Arabidopsis thaliana. *Genetics*, **171**, 1267–1275.

Walker,S.J. *et al*. (2006) Long versus short oligonucleotide microarrays for the study of gene expression in nonhuman primates. *J. Neurosci. Methods*, **152**, 179–189.

Winzeler,E.A. *et al*. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.

Zhang,L. *et al*. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.