BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 24 2008, pages 2887–2893
doi:10.1093/bioinformatics/btn571

Gene expression

Cross-hybridization modeling on Affymetrix exon arrays

Karen Kapur[1], Hui Jiang[2], Yi Xing[3] and Wing Hung Wong[1,4,*]

[1]Department of Statistics, [2]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, [3]Department of Internal Medicine and Department of Biomedical Engineering, University of Iowa, Iowa City, IA and [4]Department of Health Research and Policy, Stanford University, Stanford, CA, USA

Received on July 15, 2008; revised on October 3, 2008; accepted on October 30, 2008

Advance Access publication November 4, 2008

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Microarray designs have become increasingly probe-rich, enabling targeting of specific features, such as individual exons or single nucleotide polymorphisms. These arrays have the potential to achieve quantitative high-throughput estimates of transcript abundances, but currently these estimates are affected by biases due to cross-hybridization, in which probes hybridize to off-target transcripts.

Results: To study cross-hybridization, we map Affymetrix exon array probes to a set of annotated mRNA transcripts, allowing a small number of mismatches or insertion/deletions between the two sequences. Based on a systematic study of the degree to which probes with a given match type to a transcript are affected by cross-hybridization, we developed a strategy to correct for cross-hybridization biases of gene-level expression estimates. Comparison with Solexa ultra high-throughput sequencing data demonstrates that correction for cross-hybridization leads to a significant improvement of gene expression estimates.

Availability: We provide mappings between human and mouse exon array probes and off-target transcripts and provide software extending the GeneBASE program for generating gene-level expression estimates including the cross-hybridization correction http://biogibbs.stanford.edu/~kkapur/GeneBase/.

Contact: whwong@stanford.edu

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

1 INTRODUCTION

Microarrays are a widely used tool for conducting high-throughput analyses in many areas of biological research. They have been used in many different applications including to query mRNA transcript abundance, determine transcription factor binding sites or characterize genomic sequences (Gresham *et al*., 2008; Stoughton, 2005). The technology is based on attaching DNA fragments, or probes, to microarray slides, with each probe having exact nucleotide sequence complementarity to a specific transcript. Once labeled and amplified transcripts are hybridized to the array, transcript abundances are estimated using the fluorescent intensities of matching probes.

An important source of noise in microarray probe signals is due to artifacts of cross-hybridization. Although gene expression microarray probes are designed with perfect complementarity to target mRNA transcripts, they may also share sequence similarity with additional transcripts. As a result, probes may hybridize to specific off-target transcripts (Eklund *et al*., 2006). Gene-level summarization algorithms, which combine multiple probe intensities to generate an overall expression estimate for the gene, can suffer from biases of probe cross-hybridization. Such biases have been found to adversely affect downstream analysis based on correlation of gene expression profiles (Casneuf *et al*., 2007; Okoniewski and Miller, 2006).

Several methods for estimation of gene-level expression indexes have been proposed to reduce the effects of cross-hybridizing probes. In dChip (Li and Wong, 2001), an outlier removal procedure removes probes with intensities which differ substantially from the remaining set of probes. Methods which use robust estimation procedures to protect against outlier probes, such as Robust Multichip Average (RMA) (Irizarry *et al*., 2003) may also mitigate the effects of a small number of cross-hybridizing probes. Additionally, probe selection strategies (Xing *et al*., 2006), in which a subset of probes which show highly correlated intensities across multiple samples are selected for summarization of overall expression, guards against cross-hybridization biases originating from a minority of probes.

While the above approaches work well when only a minority of probes are affected by cross-hybridization, these methods are unlikely to be able to generate reliable estimates in the case where a large percentage of probes bind to off-target transcripts. Gene sequences which are closely related to each other may have a substantial number of potentially cross-hybridizing probes. Additionally, with the increasing oligonucleotide density on microarray chips, specific features are targeted by a small number of probes. For example, the Affymetrix exon array has, on average four probes to interrogate each exon. The analysis of alternative splicing on exon arrays, which is based on exon-level expression of individual probe intensities (Clark *et al*., 2007; Xing *et al*., 2008; Yeo *et al*., 2007), will be improved through consideration of cross-hybridization. A recent study (Xing *et al*., 2008) showed that cross-hybridization is a major cause of false predictions of differential alternative splicing. Therefore, to improve estimates of gene-level expression and to take advantage of emerging probe-rich microarray designs, it is important to understand cross-hybridization behavior and to develop methods to correct for cross-hybridization biases.

In this article, we describe how matches between short oligonucleotide probes and off-target transcripts can affect

*To whom correspondence should be addressed.

probe intensities. Using a detailed matching between probes and off-targets, we propose a correlation-based filtering method to detect and remove probes showing sequence-specific cross-hybridization to off-target transcripts. This method takes advantage of the tiling of probes across all transcribed regions to compare the observed probe intensity with the expression pattern of the putative cross-hybridizing transcript. Probes which follow the off-target expression pattern are removed while the remaining probes are retained. This strategy allows us to include as many informative probes as possible for summarization of gene-level expression. We validate our predictions of gene-level expression resulting from the cross-hybridization correction using ultra high-throughput sequencing data. Our results show that cross-hybridization modeling improves estimates for gene-level expression and can be used for exon-level analysis.

2 METHODS

2.1 Description of Affymetrix exon arrays

We use Affymetrix exon array data to illustrate our cross-hybridization modeling methods. Exon arrays are a high-density microarray platform with ∼6.5 million probes designed to target all annotated and predicted exons in the genome. A probeset, consisting of four probes, is designed to target a single putative exonic region (Affymetrix, 2005b). Furthermore, each exonic region is classified based on the supporting type of annotation. Probes targeting exons with full-length mRNA evidence, such as RefSeq mRNAs are regarded as the most confident and are referred to as 'core probes', probes targeting exons with partial mRNA evidence such as ESTs have intermediate evidence and are referred to as 'extended probes', and probes targeting exons supported solely by computational predictions have the least annotational confidence and are referred to as 'full probes'. For further details, see the Affymetrix documentation (Affymetrix, 2005a).

Due to the placement of four probes targeting each exonic region, genes have variable numbers of overall probes. However, each RefSeq sequence corresponds to a median of 30–40 core probes. In general, the core probes, which correspond to the well-annotated exonic regions, are used for summarization of gene-level expression.

2.2 Assessing probe-level cross-hybridization

We analyzed exon array data for a panel of mouse tissues, (brain, embryo, heart, kidney, liver, lung, muscle, ovary, spleen, testes and thymus), each with three replicates http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx. We applied a probe sequence-specific background correction (Johnson *et al*., 2006; Kapur *et al*., 2007) and normalized the data using a multiplicative normalization factor, chosen to set the median of core probe intensities on each array to a predefined value (100).

Genes on exon arrays, called transcript clusters, are mapped to RefSeq mRNA transcripts based on the overlap of the transcript cluster start and stop coordinates and RefSeq transcription start and transcription stop sites (see Supplementary Material). A total of 16 767 mouse transcript clusters are annotated to a RefSeq transcript.

To search for transcripts with a large degree of sequence similarity to each of the more than 6 million probes on Affymetrix exon arrays, we use the SeqMap software (Jiang and Wong, 2008). SeqMap is used to match multiple short sequences to a query transcript, allowing up to a small number of mismatches or insertion/deletions (Fig. 1). Next, a local alignment algorithm is used for the candidate probes to determine the final set of probes with the required type of match to the query transcript.

We matched probes to a set of transcribed regions, which consists of core probe selection regions (psrs), RefSeq mRNA transcripts, and ribosomal RNA sequences. Core probe selection regions (psrs), described in Affymetrix

Fig. 1. Shown here is an illustration of a probe (top sequence) matching a transcript (bottom sequence) with a 1-bp mismatch and a 1-bp insertion/deletion.

annotation files, consist of transcribed exonic regions supported by full-length mRNA annotations. We use the subset of RefSeq mRNA transcripts (release 26, December 3, 2007) which are mapped to transcript clusters on exon arrays. Additionally, we match probes to ribosomal RNAs, with sequences listed in the Supplementary Material, none of which map directly to transcript clusters. We consider only those matches between probes and off-target transcripts.

Probes are classified according to the type of match between the probe and transcript sequence. We define the match edit distance as the sum of mismatches and indels between the two sequences, which is also referred to as the Levenshtein distance.

To estimate the extent to which the expression of an off-target transcript explains the matching probe intensity, we model the background-corrected, normalized probe intensity, y_{ij} , of probe j in sample i, as

$$y_{ij} = \phi_j \hat{\theta}_i + \epsilon_{ij} \quad (1)$$

motivated by (Li and Wong, 2001). Here, $\hat{\theta}_i$ is the expression of the off-target transcript in sample i, ϕ_j represents the affinity of probe j to the off-target transcript and ϵ_{ij} is a random error term. The R^2 statistic of this model represents the proportion of the probe intensity variance which can be explained by cross-hybridization to the off-target transcript. We report R^2 statistics for full probes which match a single off-target transcript allowing up to 3-bp mismatches or indels. We used the set of full probes to study cross-hybridization instead of the set of background probes due to the larger number of full probes on the array and because these probes were not specifically chosen by Affymetrix to avoid potential cross-hybridization as is the case with the set of background probes (Affymetrix, 2005b). Gene expression was estimated using GeneBASE (Kapur *et al*., 2007) on the set of core probes which uniquely map to the transcript, allowing an edit distance of up to 3 bp.

2.3 Detecting genes enriched for non-unique probes

We calculate the proportion p_3 of core probes matching one or more off-target transcripts, allowing an edit distance of 3. For each gene k, we define a standardized residual statistic,

$$r_k = \frac{o_k - n_k p_3}{\sqrt{\frac{(n_k p_3 (1-p_3)(N-n_k)}{N-1}}} \quad (2)$$

where o_k is the observed number of non-unique probes, n_k is the number of core probes associated with gene k and N is the total number of core probes on the array. This statistic gives the observed minus expected number of non-unique probes divided by the SD of a hypergeometric distribution.

Paralog information was derived using the Ensembl compara homology database version 49 (Hubbard *et al*., 2007). We mapped transcript clusters to Ensembl gene ids by first mapping transcript clusters to RefSeq transcripts and subsequently mapping RefSeq transcripts to Ensembl genes. For mouse, a total of 16 636 transcript clusters were mapped to Ensembl genes. We classified each mapped gene as belonging to a paralog family provided it has at least one predicted paralog Ensembl gene.

2.4 Correcting cross-hybridization biases: strategy and validation

We developed the following strategy to correct for cross-hybridization. Start with GeneBASE expression estimates using all core probes (Kapur *et al*., 2007). Exclude each core probe matching an off-target transcript with

up to 2-bp mismatches or indels and with correlation between the probe intensity and off-target expression estimate above 0.7. Set probe and off-target SD to be the maximum of 100 and the observed value to guard against high correlation due to low SD. Re-calculate expression estimates using only the included probes. Iterate until gene-level expression estimates stabilize. We label this strategy GeneBASE-xhyb. To compare GeneBASE with GeneBASE-xhyb, we use our own implementation of GeneBASE to reduce implementation effects.

To study whether the cross-hybridization correction improves expression estimates, we compare GeneBASE and GeneBASE-xhyb to estimates obtained by Solexa sequencing of RNA sequences for mouse liver, skeletal muscle and brain samples. The publicly available samples, described in Mortazavi *et al*. (2008), consist of independent samples pooled from adult mouse tissues. Each tissue library resulted in 10–30 million 25-bp reads mapping to unique sites in the mouse genome. In our analysis, we combine the two independent libraries generated for each source tissue. We generated estimates of gene-level expression from sequencing reads by counting the reads per kilobase gene exon per million mapped reads (RPKM) (Mortazavi *et al*., 2008). Here, we use only those reads which have an exact match to genic regions or exon–exon junctions. Reads which map to multiple locations, multireads, were re-assigned proportional to the number of unique reads per kilobase gene exon. In the case where neither gene contains any unique reads, the reads are divided equally. For each RefSeq transcript, we generate an expression estimate by counting the number of normalized reads which fall in exonic regions. Gene expression estimates are included provided the gene has at least one assigned read. To map between gene-level estimates from the sequencing data to exon array data, we use the mapping between RefSeq transcripts and exon array transcript clusters. Using quantile normalization, we transform microarray estimates, taking the median across sample replicates, to have the same distribution as the sequencing estimates. To compare the agreement with sequencing for each strategy we use the statistic,

$$T = \frac{|s_{gb} - s| - |s_{gb_x} - s|}{(1/2)(|s_{gb} - s| + |s_{gb_x} - s|)} \quad (3)$$

where s_{gb} is the transformed expression estimate from GeneBASE, s_{gb_x} is the transformed expression estimate from GeneBASE-xhyb and s is the expression estimate from sequencing. If the GeneBASE-xhyb estimates are no more concordant with sequencing estimates than the GeneBASE estimates, then we would expect T to be centered at 0. If GeneBASE-xhyb tends to show smaller deviations from the sequencing estimates than GeneBASE, then we expect T to be shifted to the right. Using the set of genes with expression estimates changed from the cross-hybridization correction, we carry out a statistical test of the null that T is centered at 0 versus the alternative that T is centered at some positive value. We use a one-sided Wilcoxon signed rank test.

In addition to testing whether the entire set of genes with altered expression between GeneBASE and GeneBASE-xhyb are more concordant with sequencing estimates, we also perform the Wilcoxon signed rank test using a subset of genes with large changes in expression between the two sets of expression estimates, defined by the requirement

$$|\log_2(gb + c_{15,gb}) - \log_2(gb_x + c_{15,gb_x})| > 1 \quad (4)$$

where $c_{15,gb}$ and c_{15,gb_x} are the 15th quantile of gene expression from the GeneBASE estimates gb and GeneBASE-xhyb estimates gb_x, respectively. The purpose of the addition of the moderation constants, c, is to de-emphasize genes expressed at low levels which may have large fold-change statistics due to minor fluctuations in expression.

2.5 Comparison of different estimates of gene expression from Affymetrix exon arrays

We compare several methods of gene-level summarization. In addition to the GeneBASE-xhyb strategy, described above, we also computed GeneBASE estimates (Kapur *et al*., 2007). RMA, Plier and IterPlier estimates were downloaded from the set of mouse APT results from the Affymetrix website, http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx.

Genes with available expression estimates from all methods and which were mapped to sequencing estimates (requiring 1+ read) were included to compare exon array expression estimates to ultra high-throughput sequencing estimates.

2.6 Results for mouse and human

We present the results for mouse exon arrays and compare them to ultra high-throughput sequencing estimates of liver, muscle and brain tissues. Results of applying SeqMap to human exon arrays can be found in the Supplementary Material.

3 RESULTS

3.1 Matching probes to off-target transcripts

Although microarray probes are designed to have exact nucleotide complementarity to a corresponding target transcript, they may also share a large degree of sequence similarity to off-target transcripts. To determine the extent to which off-target transcripts affect the intensities of matching probes, we carry out a detailed mapping between each 25-bp probe on Affymetrix exon arrays and off-target transcripts using the SeqMap algorithm (see Section 2). We search for matches between probes and off-target transcripts which differ by any combination of mismatches, in which the sequence has a mismatched base-pair in a given position, or insertion/deletions (indels), in which one sequence contains one or more nucleotides which do not align to the complementary sequence, provided the match edit distance between the two sequences is not more than 3 bp. An example of a match between a probe and transcript is shown in Figure 1.

The results of the mapping in Table 1 show that most core probes uniquely map to their target transcripts when allowing a matching distance of up to 3 bp. A smaller percentage of probes match one or more off-target transcripts, and for an edit distance of 3, 5.09% match one transcript, 3.00% match two transcripts, 0.59% match three transcripts, and 0.89% match four or more transcripts. The results for extended and full probes are provided in the Supplementary Tables S2–S3.

Although overall only a small number of probes match to off-target transcripts, individual genes may have a large number of such

Table 1. The number of matches between core probes and off-target transcripts, allowing variable matching edit distances

Distance	Number matching transcripts				
	0	1	2	3	4+
0	839580 (98.03%)	11312 (1.32%)	3937 (0.46%)	573 (0.07%)	1069 (0.12%)
1	834693 (97.46%)	13174 (1.54%)	5501 (0.64%)	1042 (0.12%)	2061 (0.24%)
2	831059 (97.03%)	14534 (1.70%)	6395 (0.75%)	1438 (0.17%)	3045 (0.36%)
3	774502 (90.43%)	43623 (5.09%)	25673 (3.00%)	5083 (0.59%)	7590 (0.89%)

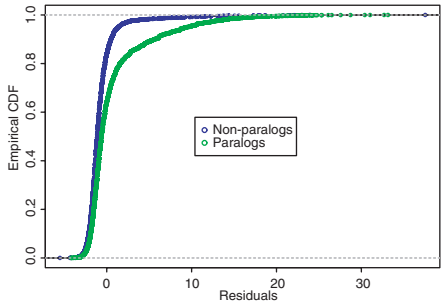

Fig. 2. The empirical cdf of gene standardized residuals, separated by paralog classification.

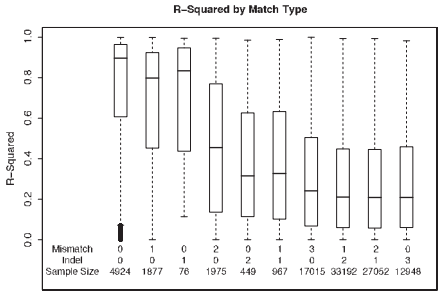

Fig. 3. R^2 statistics modeling full probe intensities by corresponding matching off-target gene expression levels, separated by the number of mismatches and insertion/deletions.

probes. We characterize the number of genes with large proportions of non-unique probes using a standardized residual statistic, given in Equation (2). While the majority of transcript clusters tend to have residuals near zero, there are still a large number of transcript clusters that are enriched for potentially cross-hybridizing probes, with 1136 transcript clusters having residuals greater than 7.0 (Supplementary Fig. S1).

To further characterize the set of genes enriched for non-uniquely matching probes, we investigated whether the set of genes with the largest residuals are enriched for genes belonging to paralog families. Using the Ensembl compara homology database of paralog predictions, we classified each gene as belonging to a paralog family if it has one or more predicted paralogs in the compara database. Overall, 65.29% of genes were classified as belonging to a paralog family. Among the top 1% and 10% of transcript clusters ranked according to standardized residuals ($n = 166, 1664$), the percent of genes belonging to paralog families increased to 92.13% and 89.76%, respectively. Plots of the empirical cumulative distribution function (cdf) separately for non-paralogs and paralogs are shown in Figure 2 and a one-sided Kolmogorov–Smirnov test that the distribution function of the non-paralog residuals lies above the distribution function of the paralog residuals is significant ($P < 2.2e-16$). This analysis indicates that genes belonging to paralog gene families are enriched for probes with sequence similarity to off-target transcripts.

3.2 Effects of match type on probe cross-hybridization behavior

The matching between probes and off-target transcripts can be used to study how probe behavior is affected by the corresponding off-target transcript expression pattern. We compute estimates of gene-level expression using GeneBASE from the set of core probes which uniquely match the target transcript when allowing an edit distance of up to 3 bp. These estimates were generated to be robust to artifacts of cross-hybridization (see Section 2).

The set of full probes are ideal for studying cross-hybridization because they target exonic regions supported purely by computational predictions, and are therefore less likely to target truly transcribed regions. Furthermore, full probes do not influence gene-level expression estimates. Through the Affymetrix annotation pipeline, a subset of full probes are assigned to nearby genes. However, the majority of such full probes have weak correlations with the assigned gene expression, shown by the distribution of R^2 statistics modeling full probe intensities using the assigned gene expression estimates (Supplementary Fig. S2).

Due to the lack of correlation between full probes and their target transcripts, we can study how full probes behave when they have a given type of match to an off-target transcript. We find that the type of match between full probes and off-target transcripts has a large effect on the extent of cross-hybridization. For each full probe we compute an R^2 statistic representing the proportion of the full probe intensity which can be explained by cross-hybridization to an off-target transcript [see Equation (1)]. To avoid complications from probes matching multiple off-target transcripts, we only consider probes which match exactly one off-target transcript. Probes are classified according to the match type in terms of number of mismatches and indels. We show in Figure 3 that for full probes with fewer mismatches/indels to a given transcript, their intensities can be better explained by cross-hybridization than for probes with a larger number of mismatches/indels. As the edit distance between probe and transcript increases a smaller proportion of probes show strong correlations with the off-target patterns of expression.

3.3 Implications for gene-level analysis

Cross-hybridization mapping has implications for gene-level analysis. Although probes are designed to hybridize to a target transcript, off-target transcripts which share sequence similarity to the probe may also bind to the probes. The response of probes which share sequence similarity to off-target transcripts may subsequently bias the resulting gene-level estimates.

For example, Figure 4 shows the probe intensities across the 33 tissue panel experiments for the pair of genes Scd3 and Scd1. In each plot, probes are partitioned into two groups—those probes which match uniquely to the target transcript and those probes which match to the corresponding off-target. Probes matching uniquely to the gene Scd3 show a different expression pattern than the probes matching Scd1, whereas for the gene Scd1, the probes either matching uniquely to Scd1 or those which also match Scd3 show the same expression patterns. This plot suggests that some of the Scd3 probes are cross-hybridizing to the Scd1 transcript. The resulting gene-level expression pattern changes depends on the set of probes used for summarization. By identifying potentially cross-hybridizing probes, we are able to remove the cross-hybridization bias and use the uniquely matching probes to estimate the expression level of Scd3. This example shows that cross-hybridizing probes can

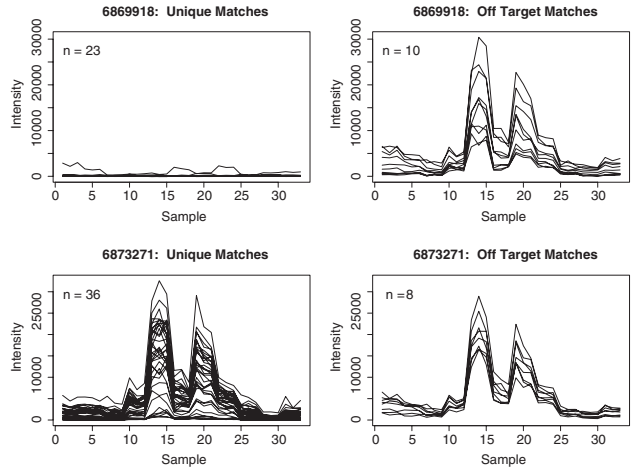

Fig. 4. Plots of probe intensities from genes Scd3 (transcript cluster 6869918) and Scd1 (transcript cluster 6873271), separated by those which uniquely match the target transcripts and those which match the corresponding off-target transcript. Unique probes from Scd3 show a different intensity patterns from those which match Scd1, suggesting a cross-hybridization bias.

result in large biases of gene-level expression estimates with only a relatively small number of cross-hybridizing probes.

We design a strategy for generating gene-level expression estimates which removes probes showing strong evidence of cross-hybridization, while retaining probes which have matches to off-target transcripts but show little evidence of cross-hybridization. Probes with up to 2-bp mismatches or indels and correlation above 0.7 with the off-target transcript expression level are excluded from gene-level summarization (see Section 2 for details). We label this strategy for generating gene-level expression estimates as GeneBASE-xhyb.

For the majority of transcript clusters, the set of GeneBASE and GeneBASE-xhyb estimates agree well. However, after excluding probes which have high correlation to off-target transcripts, some genes have fewer than five probes. Due to the small number of probes targeting the transcript, gene expression estimates are likely to suffer from a large amount of variation, and as a result we do not output gene-level expression estimates. We refer to the genes with insufficient numbers of probes for estimating gene-level expression as a result of the cross-hybridization correction as masked. The GeneBASE-xhyb method results in 129 transcript clusters with masked expression estimates. In general, a transcript cluster will have altered expression between GeneBASE and GeneBASE-xhyb if at least one selected probe is excluded from gene-level summarization. As a result, the transcript cluster will have altered expression between GeneBASE and GeneBASE-xhyb in all samples. Comparing the set of transcript clusters with available expression estimates from the two methods, the number with altered expression is 612.

To compare GeneBASE and GeneBASE-xhyb strategies of estimating gene expression, we use an independent set of expression estimates derived from ultra high-throughput sequencing data. Using Solexa sequencing reads of mouse liver, muscle and brain samples, we generated estimates of gene expression for each RefSeq transcript provided the transcript had at least one mappable read (see Section 2 for details). We then map RefSeq transcripts to exon array transcript clusters.

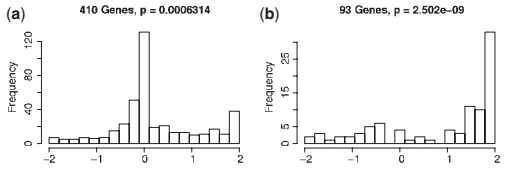

Fig. 5. Histogram of T-values to compare the concordance of the two sets of exon array expression estimates, GeneBASE and GeneBASE-xhyb, with sequencing in liver. Histogram of (**a**) all genes with altered estimates between GeneBASE and GeneBASE-xhyb and (**b**) genes with absolute log moderated fold-change between GeneBASE and GeneBASE-xhyb estimates above 1.

To compare the concordance between GeneBASE or GeneBASE-xhyb and sequencing estimates, we transform exon array expression estimates so that they have the same distribution as sequencing estimates (quantile transformation). We use the statistic, T defined in Equation (3) to compare the agreement with sequencing. In the liver, muscle and brain samples, a total of 410, 413 and 422 genes with altered expression between GeneBASE and GeneBASE-xhyb estimates are mapped to sequencing estimates, respectively, with each gene giving rise to a separate value of T. Under the assumption that GeneBASE and GeneBASE-xhyb are equally concordant with sequencing estimates, the distribution of T values will be centered at zero. On the other hand, if GeneBASE-xhyb is more concordant with sequencing estimates, we would expect positive values of T. Therefore, we carry out a Wilcoxon signed rank non-parametric test of the null hypothesis $T=0$ against the alternative, $T>0$. From the distribution of T values, we see enrichment of T values above zero. The corresponding P-values in liver, muscle and brain are 0.0006314, 0.002495 and 0.01994, respectively. Furthermore, we select a subset of genes with large changes in expression comparing GeneBASE and GeneBASE-xhyb estimates, selected for large absolute log moderated fold-change values, described in Equation (4). Using the T-values corresponding to this subset of genes ($n=93,89,72$), the distribution is dramatically skewed right and the statistical test of $T=0$ against $T>0$ yields P-values of $2.502e-09$, $4.107e-06$ and $7.865e-06$ in liver, muscle and brain, respectively. The two histograms for the liver sample are given in Figure 5 and histograms for the remaining samples can be found in Supplementary Figures S3–S4. We find that, among the genes showing large changes in expression as a result of the cross-hybridization correction, the expression estimates are more concordant with sequencing expression.

Although we chose specific parameters for the GeneBASE-xhyb strategy to strike a balance between keeping the total number of genes with altered expression estimates low and removing many true cross-hybridization biases, other choices of parameters are possible. A histogram of the liver T values from a simple filter which excludes probes matching off-targets with up to 2-bp mismatches or indels, with no requirement of correlation between probe intensities and off-target expression is shown in Figure 6. The test of $T=0$ against $T>0$ in liver is not significant when considering the entire set of genes with altered expression between GeneBASE and the estimates from the simple filter ($P=0.5198$). However, the test still gives a very low P-value of $9.137e-14$ when considering the genes with the absolute log moderated fold-changes above 1. The simple filtering procedure

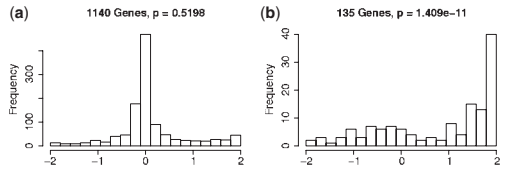

Fig. 6. Histogram of T-values to compare the concordance of the two sets of exon array expression estimates, GeneBASE and GeneBASE-xhyb-Filter, with sequencing. Histogram of (**a**) all genes with altered estimates between GeneBASE and GeneBASE-xhyb-Filter and (**b**) genes with absolute log moderated fold-change between GeneBASE and GeneBASE-xhyb-Filter estimates above 1.

Table 2. Spearman correlation between exon array estimates of gene expression and ultra high-throughput sequencing estimates for different summarization methods

	GeneBASE-xhyb	GeneBASE	RMA	Plier	IterPlier
Liver (N = 12 339)	0.8539	0.8521	0.8064	0.8198	0.8125
Muscle (N = 13 136)	0.8500	0.8481	0.8072	0.8109	0.8080
Brain (N = 13 783)	0.7542	0.7535	0.7443	0.7275	0.7132

may be able to detect more genes affected by cross-hybridization biases, but has the drawback of introducing additional variation to gene expression estimates, especially among those genes with small numbers of uniquely matching probes. Additionally, more transcript clusters will have masked expression resulting from the simple filter, with 599 masked from the simple filter compared with the 129 masked from GeneBASE-xhyb.

3.4 Comparison of competing methods to estimate gene expression

Although a comprehensive comparison of competing methods of gene expression on exon arrays is beyond the scope of this article, we believe it is important to point out that GeneBASE-xhyb and GeneBASE are competitive with other methods. We compare several gene expression indexes with the sequencing data using the Spearman rank correlation. The results, presented in Table 2, show that GeneBASE-xhyb and GeneBASE perform slightly better than competing methods RMA, Plier and IterPlier.

3.5 Implications for exon-level analysis

The cross-hybridization biases which we observed to affect gene-level expression are even more likely to affect exon-level expression, due to the smaller number of probes targeting exonic regions. For example, on Affymetrix exon arrays there are only four probes for each putative exonic region. Incorporating cross-hybridization information will be useful for exon-level analysis to reduce false positive predictions of differential splicing and for novel exon discovery (Xing *et al*., 2008).

3.6 Available software/mappings

We provide mappings between human and mouse exon array probes and off-target transcripts and provide software to run GeneBASE-xhyb http://biogibbs.stanford.edu/∼kkapur/GeneBase/. Additionally, we include an option to output several variables relating to the correlation between probes and off-target transcripts which can be used to remove effects of cross-hybridization from exon-level analysis http://biogibbs.stanford.edu/∼yxing/MADS/.

4 DISCUSSION

In this work, we demonstrate that probe cross-hybridization signals can be mapped to specific off-target transcripts. Incorporating exon array probe mapping information, we exclude probes showing strong correlations with corresponding off-target transcripts to remove cross-hybridization biases from resulting gene-level expression estimates. We evaluated our strategy for gene-level expression using independent estimates of transcript abundances from Solexa ultra high-throughput sequencing. We find that expression estimates for a number of genes can be improved by removing cross-hybridization artifacts.

Our work gives further understanding to factors affecting microarray probe cross-hybridization. The set of exon array full probes, designed to target computationally predicted exonic regions, tends to have probe intensities near background levels and can be used to study how probes respond to transcripts to which they share sequence similarity. We found decreasing correlation between probes and the expression patterns of matching off-target transcripts as the match edit distance between the probe and transcript is increased. Allowing an edit distance of 3 bp between probes and off-target transcripts, probes may show strong signals of cross-hybridization, compared with signals expected by chance (see also Supplementary Fig. S5).

Matches including insertion/deletions between probe and transcript sequences can also give rise to strong cross-hybridization signals. Therefore, it will be important to apply sequence mapping programs which have the ability to detect these types of alignments. In previous work, it has been reported that probes with much shorter alignments of 10–16 nt may be sufficient for cross-hybridization (Wu *et al*., 2005). However, as the number of matches between probes and transcripts rapidly increases with larger edit distance, it will be important to develop more sophisticated models to predict individual probe cross-hybridization. Future models may incorporate factors, such as the type of probe-transcript alignment, probe sequence (Wu *et al*., 2005) or transcript secondary structure. For example, we found that the probe sequence GC-content affects the extent of cross-hybridization and may affect cross-hybridization in different ways, depending on the type of alignment between probes and transcripts. For perfect matches between probes and transcripts, probes with intermediate GC-content tend to have the highest correlation with the transcript expression level. However, for larger match edit distances between probes and transcripts, probes with larger GC-content show higher correlation with the transcript expression levels (Supplementary Fig. S6). With more detailed knowledge of how probe sequence affects cross-hybridization, we will be able to design probes to be more specific to target transcripts.

In the absence of sequence-based predictive models of cross-hybridization, we found that the empirical data can be used to detect cross-hybridization. For a matching probe and off-target transcript, we use the transcript's expression pattern to determine whether the probe intensity reflects cross-hybridization. This approach takes advantage of the large amount of annotation of the transcriptome and can be used on other arrays with genome-wide coverage. The number of samples for which the cross-hybridization correction can improve gene expression estimates will depend on the expression pattern of the off-target transcripts. For example, from Figure 4, removing the cross-hybridizing probes will dramatically change expression estimates in many different tissue types. In a few other tissue types the estimates will not be affected because the off-target is not highly expressed in those samples. As a result, the set of genes which are affected by the cross-hybridization correction will tend to overlap among the different samples. We found that our method based on the empirical data is limited by the array design. Genes with small numbers of probes uniquely matching the target transcript can yield less reliable estimates of gene expression. For example, we found that many genes where GeneBASE-xhyb estimates are less concordant with sequencing than the GeneBASE estimates have fewer than five uniquely matching probes.

In many microarray applications careful selection of probes to uniquely match target transcripts can be used to eliminate cross-hybridization biases. In future microarray designs, SeqMap (Jiang and Wong, 2008) or similar algorithms (Li *et al*., 2008; Smith *et al*., 2008) can be used to select probes which do not share sequence similarity to off-target transcripts, allowing up to a certain number of mismatches or insertion/deletions. However, there are many microarray applications where it is unavoidable for probes to share some sequence similarity to off-target transcripts. For querying exon–level expression or for certain paralog gene families there may be only a small number of probes which uniquely match the target transcript. For probes designed to target individual sequences which differ at a particular locus by a single nucleotide polymorphism (SNP), each probe will have a single nucleotide difference between the competing genomic transcripts. Additionally, probes which target exon-exon junctions may be subject to cross-talk from hybridization to mRNA transcript isoforms which include only one of the exons (Boutz *et al*., 2007; Srinivasan *et al*., 2005). In these situations detailed knowledge of cross-hybridization will be useful to design probes with high specificity to their target transcripts.

ACKNOWLEDGMENTS

We thank Barbara Wold's lab for making the Solexa sequencing data available and Zhengqing Ouyang, Xi Chen and Xiao Tong for discussions.

Funding: National Institutes of Health (R01HG003903 and U54GM62119 to W.H.W; R01HG004634 to W.H.W. and Y.X.); Hereditary Disease Foundation (research grant to Y.X.); California Institute for Regenerative Medicine (to W.H.W., partial).

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2005a) Affymetrix technical documentation for exon array probe annotations. Available at http://www.affymetrix.com/support/technical/whitepapers/exon_probeset_trans_clust_whitepaper.pdf

Affymetrix (2005b) Exon array design datasheet. Available at www.affymetrix.com/support/technical/datasheets/exon_arraydesign_datasheet.pdf

Boutz,P.L. *et al*. (2007) A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev.*, **21**, 1636–1652.

Casneuf,T. *et al*. (2007) In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics*, **8**, 461.

Clark,T.A. *et al*. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, **8**, R64.

Eklund,A.C. *et al*. (2006) Replacing cRNA targets with cdna reduces microarray cross-hybridization. *Nat. Biotechnol.*, **24**, 1071–1073.

Gresham,D. *et al*. (2008) Comparing whole genomes using dna microarrays. *Nat. Rev. Genet.*, **9**, 291–302.

Hubbard,T.J.P. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D6107.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jiang,H. and Wong,W.H. (2008) SeqMap : mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, doi:10.1093/bioinformatics/btn429.

Johnson,W.E. *et al*. (2006) Model-based analysis of tiling-arrays for chip-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.

Kapur,K. *et al*. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biol.*, **8**, R82.

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.

Li,R. *et al*. (2008) Soap: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.

Okoniewski,M.J. and Miller,C.J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, **7**, 276.

Smith,A.D. *et al*. (2008) Using quality scores and longer reads improves accuracy of solexa read mapping. *BMC Bioinformatics*, **9**, 128.

Srinivasan,K. *et al*. (2005) Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, **37**, 345–359.

Stoughton,R.B. (2005) Applications of dna microarrays in biology. *Ann. Rev. Biochem.*, **74**, 53–82.

Wu,C. *et al*. (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res.*, **33**, e84.

Xing,Y. *et al*. (2006) Probe selection and expression index computation of affymetrix exon arrays. *PLoS ONE*, **1**, e88.

Xing,Y. *et al*. (2008) Mads: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *RNA*, **14**, 1470–1479.

Yeo,G.W. *et al*. (2007) Alternative splicing events identified in human embryonic stem cells and neural progenitors. *PLoS Comput. Biol.*, **3**, 1951–1967.