In Silico Biology 7, 0041 (2007); ©2007, Bioinformation Systems e.V.
B G R S 2 0 0 6

Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis

Yuriy L. Orlov[1], Jiangtao Zhou[1], Leonard L. Lipovich[1], Atif Shahab[2] and Vladimir A. Kuznetsov[1]*

[1] Genome Institute of Singapore, 60 Biopolis str., Genome, Singapore, 138672
[2] Bioinformatics Institute, Singapore, 30 Biopolis Street, Matrix, Singapore 138671

* Corresponding author

Email: kuznetsov@gis.a-star.edu.sg

Edited by S. Rodin (guest editor) and N. Kolchanov; received March 15, 2006; revised and accepted May 04, 2007; published August 19, 2007

Abstract

Careful analysis of microarray probe design should be an obligatory component of MicroArray Quality Control (MACQ) project [Patterson *et al.*, 2006; Shi *et al.*, 2006] initiated by the FDA (USA) in order to provide quality control tools to researchers of gene expression profiles and to translate the microarray technology from bench to bedside. The identification and filtering of unreliable probesets are important

preprocessing steps before analysis of microarray data. These steps may result in an essential improvement in the selection of differentially expressed genes, gene clustering and construction of co-regulatory expression networks. We revised genome localization of the Affymetrix U133A&B GeneChip initial (target) probe sequences, and evaluated the impact of erroneous and poorly annotated target sequences on the quality of gene expression data. We found about 25% of Affymetrix target sequences overlapping with interspersed repeats that could cause cross-hybridization effects. In total, discrepancies in target sequence annotation account for up to ~30% of 44692 Affymetrix probesets. We introduce a novel quality control algorithm based on target sequence mapping onto genome and GeneChip expression data analysis. To validate the quality of probesets we used expression data from large, clinically and genetically distinct groups of breast cancers (249 samples). For the first time, we quantitatively evaluated the effect of repeats and other sources of inadequate probe design on the specificity, reliability and discrimination ability of Affymetrix probesets. We propose that only functionally reliable Affymetrix probesets that passed our quality control algorithm (~86%) for gene expression analysis should be utilized. The target sequence annotation and filtering is available upon request.

Keywords: U133 microarray, target sequences, noise signals, cross-hybridization, human genome, gene expression, sense-antisense gene pairs, interspersed repeats, breast cancer

---

Introduction

Insufficient reliability of expression measurements is one of the key problems facing microarray experiments [Patterson *et al.*, 2006; Shi *et al.*, 2006]. In expression profiling studies, the quality of probesets is an essential piece of information in data analysis and data interpretation. The quality of probesets is therefore important for scientific research, clinical diagnostics and predictions. The problem could originate from poor gene identification by the probe sequences, whose design may not take into account the actual complexity of the transcriptome. Microarray sequence probes are designed to match particular mRNA transcripts often based on ESTs or incomplete cDNAs and also using imperfectly aligned sequences. As result, the relations between the probes and genes can change as sequence data are updated. Therefore, the problem of re-annotation and re-mapping of probes is common for all microarray platforms. In this work, we focused on Affymetrix Chips since this platform is widely used in gene expression profiling. Methods that can effectively extract reliable information from probe sets analysis in previous and future studies based on Affymetrix platforms are of considerable practical interest.

Affymetrix Corporation (http://www.affymetrix.com) provides one of the well established microarray technologies. *In situ* synthesized oligonucleotide microarray GeneChip uses a set (so called probeset) of 11-20 oligonucleotide probes, each 25 bases long, to represent a gene or a gene transcript. Affymetrix uses initial (target) sequences of

~150-450 nt of each gene to locate the probes. The perfect match probe comes together with a mismatch probe designed to measure non-specific cross-hybridization (CH) (Affymetrix, 2004, http://www.affymetrix.com/support/). The expression level for a gene is a summary of the data from the entire probesets.

Improper microarray probe design can influence expression analysis starting from hybridization signal measurements and ending in identification of differentially expressed genes and gene clustering. This has often been the norm and studies of the same biological samples have led, in many cases, to contradictory results [Gautier *et al.*, 2004; Harbig *et al.*, 2005; Zhang *et al.*, 2005]. In particular, high complexity of genome loci and diversity of transcriptome sequences (ESTs, mRNAs) could be the sources of incorrect annotation in databases. Probeset (and individual probes) design based on such incomplete or even erroneous annotation has a clear potential to generate downstream problems for correct interpretation of microarray experiments. In some cases, several probesets can specifically target a single genic sequence coding for protein (and thus allow to better define transcripts), in other cases a probeset is capable to hybridize more than one transcript (and provide uncertainty in transcript detection) [Stalteri and Harrison, 2007]. The "multiply targeting" probesets have a common component in expression signals, i. e. these probesets can lead to an increased number of spurious positive correlations between expressed genes [Okoniewski and Miller, 2006; Orlov *et al.*, 2006].

Recent papers [Gautier *et al.*, 2004; Mecham *et al.*, 2004; Dai *et al.*, 2005; Harbig *et al.*, 2005; Leong *et al.*, 2005] report re-evaluation of Affymetrix array probes using BLAST comparison of probe sequences to the "complete" human genome. The problem of accurate Affymetrix target sequence annotation is related to the complexity of multiple "gene models" based on (often old) ESTs without further validation. Reported re-identification using different genome database releases may affect 30-50% of probesets [Harbig *et al.*, 2005; Okoniewski and Miller, 2006].

The problem of multiple matching is not solved, although Affymetrix probeset names are supposed to identify probesets that are associated with multiple transcripts. In particular, those marked "_x_at" are identified as being "non-specific". Similarly, "_s_at" probesets are identified as potentially targeting different gene family members or splice variants. However, genome mapping shows that many of the probesets associated with multiple genomic loci are not identified correctly and the majority of them are not indicated by name convention [Okoniewski and Miller, 2006].

There are several basic quality control criteria for verification of the target sequence. These sequences (1) should detect a unique locus in the human genome; (2) should perfectly match a transcript; (3) should correspond to the sequence from the transcribed strand of the genome at the locus (correct strand orientation of original target sequence); (4) should not overlap with any other non-gene sequence that could cross-hybridize or even be independently transcribed (segmental duplications, interspersed repeats, microRNAs); (5) should correspond to mature RNA (not intronic sequences that are spliced). These basic criteria have not been well controlled, partly because transcript databases undergo continual growth and change.

Comprehensive reassessment of Affymetrix probesets is of great importance when considering the use of experimental data for analysis of differential expression of particular genes and especially when inferring expression networks. In the present study, we recomputed genomic mappings of the Affymetrix U133A and U133B GeneChip initial target sequences, reassessed mapping problems and annotations corresponding to those sequences, and evaluated the impact of erroneous and poorly-annotated initial probe target sequences. We developed novel quality control procedures based on sequence alignment, genome mapping, annotation and statistical analysis of GeneChip expression data from large and well-characterized clinical groups [Miller *et al.*, 2005; Ivshina *et al.*, 2006] as well as the treated and untreated human cells [Shames *et al.*, 2006]. We compared the good- and poor-quality probesets for their discriminating ability of biologically and clinically distinct cancer subtypes and the statistics of correlations between hybridization signals within these good- and poor-quality probesets for the same patient groups.

Method

Sequence data

Affymetrix sequence data for the U133A and U133B chips were downloaded from the NetAffx web site (http://www.affymetrix.com/analysis/index.affx). These sequences, intended to represent genes, are referred to as initial target sequences of the Affymetrix probesets. We used these target sequences to survey possible transcripts that each probeset might detect. To study target sequences assignment, we used BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat), UCSC Genome Browser tools, and our own programs developed at GIS (http://www.gis.a-star.edu.sg/internet/site/) and BII (http://www.bii.a-star.edu.sg/). We used BLAT search at 90% similarity level to match each Affymetrix target sequence to the genome. Then we annotated overlaps with exonic region(s) of RefSeq, mRNA and spliced EST variants on the NCBI Build 35 and 36.1 (hg17 and hg18) assemblies.

We mapped Affymetrix probesets to gene sequence blocks based on the initial target sequences, not based on the individual 25-mers in the probesets. The results of mapping (chromosome coordinates, orientation, details of overlapping with exons and repeats etc.) were stored in a local database associated with unique Affymetrix probesets ID.

Mapping onto human genome

An Affymetrix target sequence is defined as problematic if it (1) does not align by BLAT at 90% similarity criterion in the human genome; (2) shows more than one BLAT match at different loci of the human genome; and/or (3) shows an orientation opposite to the

intended gene sequence (perfect match of all the gene exons and the Affymetrix target sequence blocks, but on opposite strands). In complex cases of overlapping transcripts we consequently checked perfect matching of target sequence to antisense transcript for RefSeq gene annotation, then for mRNA, then for spliced EST corresponding to the intended target.

In addition, for each target sequence, we checked for exonic repetitive elements using RepeatMasker. We constructed a table of repeats found by family and repeat types (DNA, LTR, LINE, SINE, simple and low complexity repeats, etc.) indicating the length of the Affymetrix target sequence covered by the each type of repeats.

Some of the Affymetrix target sequences exactly or partially match a gene by mapped transcript blocks in opposite orientation (see Figure 9B, 232550_at probeset). Such a target sequence could be considered as wrongly selected sequence not representing the gene. Alternatively, such a target sequence could correspond to a mRNA or an EST located on the opposite strand of the given gene. We found that a large fraction of Affymetrix target sequences maps transcripts whose expression could be affected by transcription from a *cis*-antisense transcript of the opposite strand. A target sequence could completely or partially overlap a transcript mapped on the opposite strand and thus to be considered as wrongly designed. But such a *cis*-antisense transcript also could be an artifact of wrong EST mapping. However, substantial numbers of mRNAs and ESTs in *cis*-antisense loci represent natural antisense transcripts (NAST) derived from opposite strand of the given gene [Yelin *et al.*, 2003; Katayama *et al.*, 2005; Zhang *et al.*, 2006].

In order to distinguish the Affymetrix target sequences matching NAST from the Affymetrix target sequences having wrong orientation to transcript, we developed a pipeline and constructed the local United Sense-Antisense Pairs (USAP) database [Kuznetsov *et al.*, 2006b] collecting genomic information about sense-antisense (SA) gene pairs and Affymetrix target sequences matching such pairs. The database annotates and classifies SA pairs by three annotation tracks (RefSeq, mRNA and EST) using latest human genome release (hg18). USAP contains two times more SA transcript pairs than previously reported for the human genome by [Zhang *et al.*, 2006]

Data for expression validation

Cancer data and microarrays

To study functional usefulness of the problematic probesets, we analyzed the expression patterns of approximately 23,000 gene transcripts (represented by 44,928 probesets on Affymetrix U133A and U133B arrays) in 249 primary breast tumors (NCBI Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/; data sets GSE4922). The cancer samples were split into four groups (G1, G2a, G2b, G3): G1 and G3 groups correspond to histologic grades I and III tumors, respectively; G2a and G2b groups are the sub-types of histologic grade II tumors, which have been identified based on genetic

re-classification of the grade II breast cancer tissues resulting in computational pattern recognition of small and robust prognostic gene signatures [Ivshina *et al.*, 2006]. The order of G1, G2a, G2b and G3 corresponds to aggressiveness of breast cancer. The number of samples in G1, G1-like, G3-like and G3 was 68, 83, 43, and 55, respectively; for details, see [Miller *et al.*, 2005; Ivshina *et al.*, 2006].

In addition, we used U133A&B expression data from several normal and cancerous brain tissues (GEO data set: GDS1962) and lung cancer (GEO ID: GSE5816). 29 Affymetrix microarrays represent several human lung cancer cell lines before and after treatment with promoter hypermethylation agent 5-aza-2'-deoxycytidine (5-aza) [Shames *et al.*, 2006]. Expression pattern of 5-aza-treated lung cancer cells was associated with switch-on and switch-off for 132 tumor-specific promoter-hypermethylated genes and up- or down- expression of many other hundreds genes in the cancerous genome [Shames *et al.*, 2006].

All data passed the quality control of the expression signals on microarrays and MAS5 normalization was applied [MAS 5.0 algorithm. Affymetrix, 2002]. We then performed a global mean normalization to ln(500) to provide a standardization of signals of expressed genes across microarrays and to compare the frequency distributions of the signal within entire dynamical range of the signal in microarray transcriptome samples.

Analysis of the empirical frequency distribution function of hybridization signal for microarray transcriptome sub-sets

We constructed the empirical frequency distribution function of normalized hybridization signal (representing gene expression level) for individual microarrays. We also constructed the empirical distribution function for different classes of problematic target sequences (see below). To evaluate the quality of the problematic signals, we also constructed the distribution functions using non-problematic (well defined and high specific) target sequences collected by chance into a class of samples of the same size as we did for the problematic target sequences.

Discrimination ability of a set of Affymetrix probesets derived from a problematic class of target sequences

We evaluated the ability of a set of Affymetrix probesets derived from a given problematic class of target sequences to discriminate the biologically and clinically defined subtypes of tissue samples. In particular, we used tumor samples of histological grades I and III of breast cancer, which can be strongly discriminated by ~4000 differentially expressed genes [Chua *et al.*, 2006] and by patient disease free survival time (DFS) [Ivshina *et al.*, 2006].

To perform that evaluation quantitatively, we used Statistical Analysis of Microarrays software, SAM 3.1 [Tusher *et al.*, 2001]. This software provides the estimates of the

"significant differences" between two groups of samples by calculating a so-called "false discovery rate" value (SAM *q*-value). Taken a fixed *q*-value cut-off, the SAM program identified a set of Affymetrix IDs that have differentially expressed signals which discriminate two groups. Then, taken a fixed *q*-value cut-off, we obtained a set of Affymetrix IDs providing differentially expressed genes (represented by hybridization signals). We estimated the number of non-problematic target sequences expected by chance in a set of this size. The discrimination ability of Affymetrix probesets derived from a given problematic class of target sequences was estimated based on the statistical significance of the difference between expected and observed numbers of such probesets in the set of differentially expressed Affymetrix IDs.

Analysis of the correlations in the distinct groups of Affymetrix probesets

In a group of microarrays (representing tumor sub-type), we calculated the Kendall τ (tau) correlation coefficients matrix between all Affymetrix GeneChip signals. For every pair of Affymetrix probesets we calculated the Kendall τ correlation coefficient between their expression signals. We then counted the numbers of positive and negative correlation coefficients for Affymetrix probesets associated with the problematic classes of target sequences. We calculated the Kendall τ correlation coefficients for U133 Affymetrix probesets in microarrays representing tumors with G1, G2a, G2b, and G3 sub-types. Briefly, we calculated the Kendall τ rank order correlation coefficients between all pairs of probesets using data for 68, 43, 83 and 55 patients with G1, G2a, G2b, and G3 sub-types of breast cancer grades, respectively. Thus, for each of the four subtypes, we calculated symmetric matrices of size 44692 × 44692 (common probesets in array U133A and array U133B were used only once). To avoid noise correlations we calculated a number of positive and negative correlation coefficients at $P<0.05$ and $P<0.01$ (For example, for sample size 55 (G3 sample) only correlations higher than 0.31 are significant at 5% level, and higher than 0.409 are significant at 1% level, etc.). We did the same analysis for all the groups of probesets derived from problematic target sequences. For each problematic group we compared the number of observed correlation coefficients and the number of correlation coefficients expected by chance. The expected number we estimated by Monte Carlo sampling procedure using the same number of Affymetrix probesets that we have in the problematic group.

Statistical software

For group comparison Mann-Whitney *U*-test statistics were used for continuous variables and one-sided Fisher's exact test used for categorical variables (Statistica-6 and StatXact-6 software). We also used SAM 3.1 (Statistical Analysis of Microarrays) software [Tusher *et al.*, 2001] to estimate the number of differentially expressed Affymetrix probesets.

Results

Problematic classes of target sequences

We believe that target sequences of purportedly human microarray probes which, by BLAT, are completely absent in the human genome (sequences to which we hereafter refer to as Tag0 sequences) and target sequences which match multiple loci in the genome are sources of uncertainties in gene identification and cross-hybridization effects. They should be excluded from analysis of microarray experiments. We checked BLAT (hg18) mappings for all 44,692 sequences on U133A and U133B arrays, except the control sets.

We found: (i) 1212 (2.7%) initial target sequences which do not match any location in the human genome (Tag0 or mismatched sequences, see Tab. 1); (ii) 42708 (95.5%) target sequences with a single mapping (Tag1 / reliable target sequences); (iii) 772 (1.7%) target sequences with multiple locations in the human genome (Tag2+).

Tag2+ is defined as sum of Tag2, Tag3, Tag4 etc., based on the number of their BLAT-matched loci. Tag 0 and Tag2+ might cause noise and/or cross-hybridization signals. Tag0 probesets are related mostly to mRNA and EST, but not genomic DNA and were associated with poorly-designed target sequences, poorly-annotated transcripts, and non-human sequences which are mistakenly labeled "human" in GenBank (Affymetrix clearly designed probesets from those sequences without ever verifying by BLAST that the sequences are really human). For instance, about 45% of Tag0 were classified as "xeno-sequence/non-human" (mouse, cow, pathogens, rat etc; 224340_at is mouse c-*myc* with extra TGA insertion; 217283_at strongly maps mouse short stature homeobox; 217255_at 100% is cow SQSTM1); about 27% of Tag0 do not bring up any human sequences (207726_at falls to GeneCards estrogen-related receptor beta (ESRRB)); about 17% of Tag 0 were classified as low-accuracy sequences. Others belong to small groups of poorly-defined sequences (for instance, 222196_at falls to random (not assembled) chromosome parts).

Table 1: Statistics of Affymetrix target sequence matches in human genome.

# locations(Hg18)	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6+	Tag0	Total
#Affymetrix IDs	42708	450	129	67	42	84	1212	44692
%	95.56	1.0	0.28	0.14	0.09	0.18	2.71	100

Standard assignment of Affymetrix target sequences to genome provided by UCSC Genome Browser using default BLAT parameters partially missed problematic probesets

or has no hit indicated. Location of target sequences should correspond to mapping of the gene, but the latter may change from database releases and be not resolved. For example, 208303_s_at falls onto different chromosomes: X and Y following the mapping of *CRLF2* (cytokine receptor-like factor 2 isoform 1). The CDS of the gene is not complete. Another example is 207353_s_at probeset. Its target sequence mapped to the not assembled part of chromosome 4 (chr4_random) corresponding to the location of the *HMX1* gene (homeo box H6 family 1). The Affymetrix target sequence 221715_at is not mapped on neither human genome hg17 nor hg18 releases.

We identified multiple genome locations of some extraordinary redundant target sequences. For instance, 81737_at has 22 different locations in human genome; 213089_at also has more 11 hits to human genome. Some of these hits are presented both in hg17 and hg 18 maps.

Repeats in Tag1 target sequences

Surprisingly, we found up to 25% of target sequences covered by mobile elements (repeats) abundant in the human genome (Tab. 2).

Table 2: Statistics U133A&B Affymetrix target sequences overlapping genome.

Set of genome repeats	Repeat class	# in target sequences
Simple repeats	Simple repeat, Low complexity	3233
Short transposons (<300 bp)	DNA, SINE/Alu, SINE/MIR	4347
Long transposes (>300 bp)	LINE/CR1, LINE/L1, LTR/ERV1/ERVK/ERVL/MaLR	5420
Non-transposons and satellites	Other, RNA, rRNA, Satellite, scRNA, snRNA, srpRNA	80

The majority of repeats in target sequences are LINE, LTR and SINE. These target sequences might be a significant source of erroneous detection of expressed genes and cross-hybridization signals.

Inversely oriented target sequences

We consider an Affymetrix target sequence as inversely oriented if it matches the opposite strand to any RefSeq, mRNA, or EST-supported gene (Tab. 3). These target

sequences may refer to natural antisense transcripts (NAST), but not annotation errors [Harbig *et al.*, 2005]. The large fraction of Affymetrix target sequences (29.7% (13260/44692)) matches RefSeq or mRNA or EST on the opposite strand of a given gene, but the complete genomic coordinates (complete sets of exon and intron boundary coordinates) for these negative-strand transcripts are different from the mapped Affymetrix blocks. We used only verified coordinates of antisense transcripts from USAP database (column "Match to NAST", Tab. 3). These results suggest that a large fraction of Affymetrix probesets detect *cis*-antisense transcripts which are putatively non-coding transcripts located on the opposite strand of a given gene. The percentage is consistent with several published studies using an exon-to-exon *cis*-antisense overlap definition, which found that ~20% of mammalian transcriptional units have *cis*-antisense transcripts [Chen *et al.*, 2004].

Using our working definition of misoriented Affymetrix target sequences (see Appendix 1), we found further 1297 Affymetrix target sequences whose complete genomic coordinates perfectly matched the target gene on the opposite strand (gene mapping by UCSC browser onto hg18 genome assembly. See Appendix 1 for examples). These Affymetrix target sequences probably have been designed based on poorly defined mRNA sequences and ESTs in which orientation had not been defined accurately (e.g. EST clusters, pseudogene transcripts) and, perhaps, for which a gene name had been assigned later. Some inversely oriented target sequences might originate from reverse-oriented artifact singleton cDNA clones whose incorrect orientation is evident when their structures (complete genomic coordinates) are compared to those of newer and more accurate cDNA sequences mapping to the same locus. Importantly, a major fraction of probesets assigned to these target sequences showed low expression levels for our samples. This result supports our definition of these 1297 sequences as problematic Affymetrix targets. 370 of these 1297 problematic target sequences were found by manual curation and 927 by comparison of RefSeq, GenBank mRNA and EST annotation tracks.

In total, 810 (1.8%) Affymetrix target sequences were defined as misoriented sequences. This set was defined by manual curation and automatic comparison of blocks of Affymetrix target sequences with exons of RefSeq or mRNA sequences in opposite strand (Tab. 3).

Table 3: Classification of Affymetrix target sequences (Tag1) matching transcripts in opposite strand.

Sets	Correct orientation		Misoriented to intended transcript in as verified by				Total in Tag1
	Total	Match to NAST	Manual curation	RefSeq	mRNA	EST	
# Target	41898	13260	370	138	302	487	42708

sequences							
%	93.74	29.66	0.82	0.3	0.67	1.08	95.65

The number of Affymetrix target sequences misoriented relative to intended transcripts is about 2 times larger than reported by Harbig *et al.* [Harbig *et al.*, 2005]. Affymetrix target sequences matching transcripts in both strands may refer to the natural antisense transcripts (NAST) [Chen *et al.*, 2004; Harbig *et al.*, 2005] (see column "Match to NAST" in Tab. 3).

In addition, 487 (1.08%) Affymetrix target sequences perfectly match ESTs on opposite strand. However, we do not consider this set as a reliable set due to probable errors in EST annotation.

Classification of different categories of problematic Affymetrix target sequences

Tab. 4 shows the statistics of all categories of poorly-defined Affymetrix target sequences found using hg18 Assembly: Tag0, multiple genome matching Tag2+ (Tag2, Tag3 and others) targets sequences, misoriented target sequences and the target sequences covered by genome repeats. This table shows that only about 86% (38511/44692) U133A&B target sequences could be useful in expression analysis.

We suggest not use 1984 non-Tag1 Affymetrix target sequences, 810 misoriented sequences, and 3387 sequences covered by genome repeats by more than 40% of the target sequence length.

Tab. 4 also shows the numbers of Affymetrix target sequences covered by interspersed repeats grouped by percent interval of the sequence length. The number of target sequences that strongly overlapped with genome repeats was usually less than the number of partial overlaps (761 (or 1.7%) for overlap 80-100% and 1690 (3.78%) for 40-60% overlap).

Table 4: Joint classification of problematic Affymetrix U133 target sequences.

Target sequences groups	Non-redundant # of probesets	%
Total # of non-Tag1 sequences, including:	**1984**	**4.43**
Tag0	1212	2.71
Tag2+	772	1.72

Total # misoriented target sequences, including:	**810**	**1.81**
RefSeq IDs (by blocks)	138	0.3
mRNA GenBank (by blocks)	302	0.67
Manual curation protocol	370	0.82
Total # of target sequences overlapped with repeats, including:	**3387**	**7.57**
Overlap 80-100% of target sequence length	761	1.7
Overlap 60-80%	936	2.09
Overlap 40-60%	1690	3.78
Total # of useful Tag1 target sequences, including:	**38511**	**86.16**
Overlaps with observed transcripts in opposite strand	13260	29.66
Misoriented to ESTs in Tag1	487	1.08
Target sequences with 20-40% of repeats	2409	5.39
Target sequences with <20% of repeats	1210	2.7
TOTAL # of Affymetrix target sequences	**44692**	**100**

Comparison of mean gene expression levels detected by different classes of problematic target sequences

To study different groups of problematic Affymetrix target sequences, we used a large set of expression data of genetically and clinically well-separated breast cancer sub-types [Ivshina *et al.*, 2006]. Here we demonstrate that misoriented target sequences and strong overlapping of the target sequences with genome repeats provide lower mean expression signals and larger noise.

We compared average gene expression levels in these groups of problematic target sequences: Tag0, multiple loci matching, misoriented relative to given gene, and target sequences covered by repeats as 0-20%, 20-40%, 40-60%, 60-80%, 80-100% of target sequence length (in non-overlapping intervals of percents, i.e. [0;20), [20;40) ... and [80-100]). For each group of target sequences, we determined mean values (in natural log scale) of normalized hybridization signals for tumor samples averaged by all probesets in the group (Fig. 1).

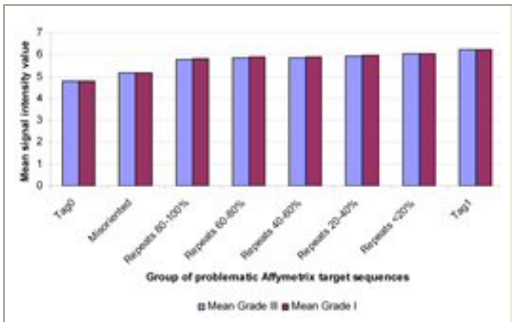

Figure 1: Averaged expression signal of probesets associated with problematic groups of Affymetrix target sequences. Designations for the groups: "Tag0": the sequences do not match to the human genome; "Misoriented": Misoriented by blocks to RefSeq or mRNA (see the text); "Repeats<20%": repeats cover less than 20% of target sequence size; "Repeats 20-40%": genome repeats cover from 20% to 40% of target sequence size; and so on for "Repeats 40-60%", "Repeats 60-80%", Repeats "80-100%"; "Tag1": do not contain problematic Affymetrix target sequences. Columns present averaged expression values in histologic Grade I and Grade III of breast cancer samples.

Affymetrix probesets derived from target sequences Tag1 which do not exhibit any complication and which are covered by genome repeats over less than 20% of target sequence length could be designated as "Normal". Fig. 1 shows a strong negative trend of the mean values of the hybridization signal from Normal group to misoriented target sequences group. In particular, problematic probesets from target sequences with larger fraction of genome repeats provide lower mean expression signals than target sequences with smaller repeat fraction.

Mean value of probesets from Normal group is close to 6.2 as should be in logarithmic scale of Fig. 1. All the problematic target sequence groups show lower average value of the hybridization signals. The differences between clinical sub-types (G1 or G3) exhibit relatively larger enrichment on differentially expressed genes represented by probesets of Normal group in comparison to the genes represented by problematic probesets.

Misoriented and multiple-matching target sequences provide most poor expressed probesets in comparison with other problematic sequence groups (Fig. 1 and Fig. 2). This trend is exhibited by the lower average expression signal (Fig. 1) and by a larger coefficient of variation (CV) (Fig. 3).

By our definition, misoriented Affymetrix target sequences perfectly match, but in the inverse orientation, the entire complete genomic coordinates of a protein-coding gene. Fig. 2 shows the mean expression values (in log-scale) probesets derived from wrongly orientated Affymetrix target sequences found in different annotation tracks. Tag1 data was used as a positive control.

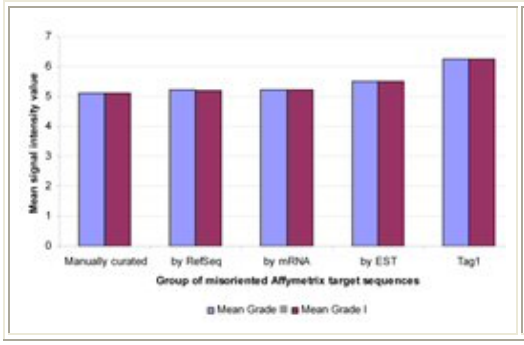

Figure 2: Average values of expression signals of probesets for misoriented Affymetrix target sequence groups.

Fig. 2 shows that manually curated, RefSeq and GenBank mRNA derived groups of misoriented target sequences provide lower averaged expression signal than the EST derived group. We joined probesets associated with these three groups as reliably-defined misoriented (corresponds to "Misoriented" in Fig. 1). We conditionally defined the Affymetrix target sequences which perfectly matching EST in opposite strand of the given gene as "putatively misoriented". All the groups of probesets derived from misoriented sequences have less average expression than Normal group.

In addition to the mean signal intensity value, we calculated the coefficient of variation (CV, ratio of standard deviation to mean) to find problematic Affymetrix probesets with higher noise in microarray experiments. Fig. 3 shows that among problematic groups, the trend of decreasing CV value reflects inversely the trend of increasing mean value presented in Fig. 1. Tag0 group provides the highest variability of the hybridization signals and lower mean value of the signal, perhaps due to low-specificity and cross-hybridization attributes of probesets derived from this group of target sequences. Interestingly, the CV value of G1 tumors is reproducibly smaller than CV value of G3 tumors among all compared groups. However, differences between clinical sub-types (G1 or G3) exhibit relatively minor impact on the CV value in comparing the hybridization signals between normal and the most problematic groups (Fig. 3: "Misoriented", "Tag0", and "Repeats 80-100%" groups).

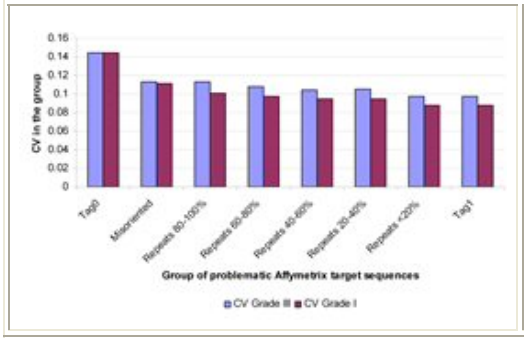

Figure 3: Comparison of average coefficient of variation (CV) values for Affymetrix probesets derived from different problematic target sequence groups.

Ability of probesets derived from repeat-overlapping target sequences to identify differentially expressed genes

We compared the discrimination ability of Normal and problematic probeset groups by identifying differentially expressed genes in human breast tumor sub-types. Histologic grades G1 and G3 of breast cancer are the subtypes that associated with low- and high-aggressive clinical behavior of the cancer. These tumor subtypes exhibit differential expression for at least 4000 U133A Affymetrix probesets [Chua *et al.*, 2006]. Assuming to use G1 and G3 microarray data sets in our quality control analysis of repeat-overlapping sequence targets, we suggested that if a given type of repeat elements covers a subset of target sequences, then probesets corresponding to these sequences should be relatively under-represented in a set of differentially expressed genes, because of the nonspecific matching of the exonic repeat fragment by multiple transcripts. We used a score estimated by a ratio of the numbers of differentially expressed probesets in repeats-overlapped and repeats-free groups of probesets.

First, we used U133A and U133B arrays and applied a statistical test (SAM 3.1 [Tusher *et al.*, 2001]), which calculated a "false discovery rate" (or SAM *q*-value). We selected 6144 differentially expressed probesets which can discriminate the low- and high-aggressive breast cancer samples at low *q*-value (less than 1.5%). Then, we counted the number of probesets derived from target sequences which were covered by a given type of repeats at 10%, 20%, ..., 100%.

Second, we compared observed ratio of the differentially expressed probesets for which the corresponding target sequences were overlapped with repeat sequences in the human genome and the ratio expected by chance using the formula:

$$f = (R_S / R) / (N_S / N) \quad (1)$$

where N is the total numbers of probesets, R is the total number of probesets with repeats-overlapped target sequences, N_S is the number of statistically significant probesets in SAM test; R_S is the number of statistically significant probesets derived from repeats-overlapped target sequences in SAM test. Note that even though theoretically this ratio could be larger than 1, in our analyses we found only smaller values indicating under-representation of problematic probesets in the G1 and G3 discriminating set.

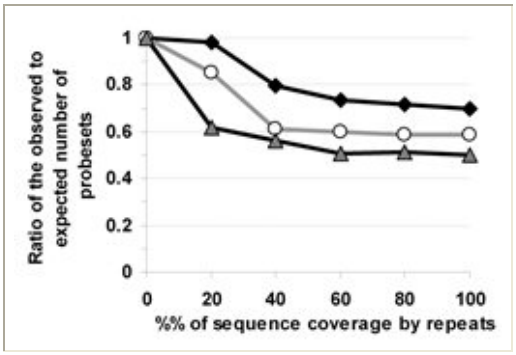

Figure 4: Index *f* as a function of percent of corresponding target sequence span covered by genome repeats. Shown are the results for all repeats (black diamonds), short repeats (SINE, DNA, Alu; white circles) and long repeats (LINE, LTR; gray triangles).

We found that simple repeats and low-complexity sequences do not affect the ability of corresponding probesets to discriminate tumor-type specific signals. As a general trend, probesets derived from target sequences with larger repeat overlap have progressively worsening proportions among differentially expressed genes, especially for longer repeats (LTR and LINE).

Cross-hybridization within groups of problematic probesets

We calculated Kendall τ rank correlation coefficients between probeset expression values on the breast cancer samples. We calculated the number of significant correlations (at the significance levels $P<0.01$ and $P<0.05$). We compared the number of such correlations within problematic groups and within randomly selected groups of the same size from Normal probesets (using Monte Carlo simulations). In general, whole array and random control groups have an approximately equal number of positive and negative correlation coefficients (around 50%). A fraction of significant positive correlations for Normal probesets group was slightly larger than a fraction of significant negative correlations (at $P<0.01$ and $P<0.05$). However, the fractions of significant positive correlation coefficients between probesets in the problematic Affymetrix target sequence groups were higher than in control groups. For example, the fraction of significant (at $P<0.05$) positive correlations between probesets derived from target sequences with repeat coverage of more than 80% is about 0.07 as compared with 0.05 for target sequences with repeat coverage of less than 20%. Fig. 5 shows that a fraction of significant correlation coefficients in G1 tumors has positive increment for the positive correlations and negative increment for the negative correlations when the fraction of repeats in a given target sequence becomes larger. Similar trends were observed for G2a, G2b, and G3 tumors (not presented).

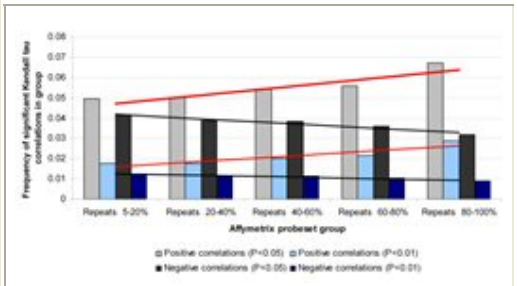

Figure 5: Fraction of significant positive and significant negative Kendall τ correlation coefficients (at levels $P<0.05$ and $P<0.01$) for Affymetrix probesets depends on the percent of the target sequence covered by genome repeats. The groups are from Tag1, target sequences with correct orientation. No simple and low complexity repeats were taken into account. Grade I breast cancer samples. Red and black lines show positive and negative trends, respectively.

These results support our assumption that problematic Affymetrix target sequences (in particular, the sequences which are mostly covered by genome repeats) can be an essential source of (positive) spurious correlations between many probesets (and respectively, genes) in microarray experiments.

Since the observed ratio of positive correlations versus negative correlations between probesets on Affymetrix U133A and U133B microarrays was about 1 (as expected by chance), the ratio between the number of positive and negative significant correlations in the samples also could be equal to 1. Indeed this ratio equals to approximately 1 in random samples taken from all probesets for each tumor sub-type (not shown). But for target sequences covered by genome repeats this ratio increases with repeat coverage up to 2 times (Fig. 6). The fraction of positive significant correlations within problematic groups monotonously increases when the covering of the target sequence by repeats becomes larger. However, there is no trend in the positive correlation proportion for the same size sub-groups of the probesets chosen randomly from Tag1 Normal group. The difference between the proportions of positive correlations expected by chance and observed becomes detectable when genome repeats cover on average more than 40% of original target sequence (Fig. 6). Fig. 6 shows the results for G1 tumors. Similar results were obtained for genetic grades G2a, G2b and histologic G3 breast cancer sub-types.

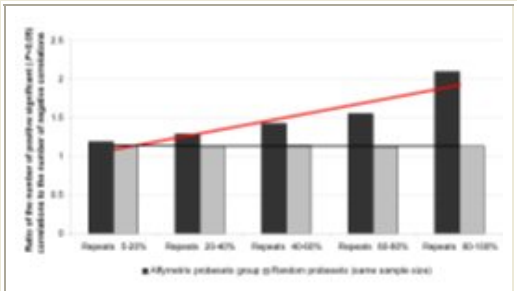

Figure 6: Ratio of significant positive Kendall τ correlation coefficients (at the significance level $P<0.05$) within problematic groups of probesets derived from target sequences covering by repeats and within samples of the same size chosen randomly from the Tag1 Normal group. Grade I breast cancer samples.

Comparison of the numbers and values of correlation coefficients of probesets derived from multiple matching target sequences with random samples from Normal group reveals similarly poor quality of these problematic groups. Our analysis of expression data for different sub-types of breast cancer samples reveals that a larger number of genome loci for the target sequence correlates with 1) higher expression noise (defined by CV-value), 2) lower average signal level, and 3) higher number of spurious positive correlations. This is what we expect for nonspecific hybridization of cDNA sequences.

Comparison of the signals on GeneChip U133A and U133B arrays

Arrays U133A and U133B show different statistical properties of sequence quality. Tab. 5 shows that the fraction of target sequences passed our QC (quality control, i.e. tag1, correct orientation on chromosome, and repeat coverage is less than 40% of target sequence length) on array U133A is a larger in comparison to array U133B. In general array U133A is better annotated and shows higher expression level than array U133B (about 89% of non-problematic probesets are in U133A vs ~83% for U133B). Fig. 7 shows that the density of signal intensity values averaging on 10 lung cancer cell line samples for arrays A and B are distinct. In particular, Fig. 7 shows that the fraction of noisy signals is significantly larger on array U133B (left side of the distributions), while array U133A exhibits much higher specific hybridization signal (right side). The effect of QC filtering is presented on both arrays.

Table 5: Comparison of genome annotation quality for U133 A and U133B arrays.

	# Probesets	# Correct probesets (passed QC)	% of correct probesets (passed QC)
A and B	100	98	98.0
Service probesets	68	N.A.	N.A.
Array U133A	22115	19753	89.3
Array U133B	22477	18660	83.0

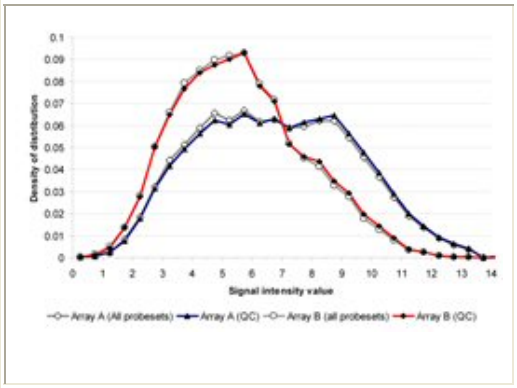

Figure 7: Signal intensity value distribution for Affymetrix U133A and U133B arrays (dots indicate density of the distribution in intervals [0;0.5), [0.5;1.0), [1.0;1.5), etc.) Probesets filtered by quality control (QC) have slightly higher average signals for both arrays U133A and U133B. MAS5 normalized and log-transformed data on lung cancer cell samples (GEO ID: GSE5816) [Shames *et al.* 2006].

There are many examples of problematic patterns distinguishing Tag1 U133A and U133B probesets corresponding to the same gene. For example, the c-*myc* gene is matched by three Affymetrix probesets: A.202431_s_at is from U133A array, B.239931_at and B.244089_at probesets are from U133B array. However, expression signals from these three probesets do not correlate to each other. 244089_at even is in opposite orientation to the gene. Probeset A.202431_s_at exhibits higher expression level 7.45-7.49 and appears in the right (most specific) side of the empirical frequency distribution of signal intensity value. Two other probesets exhibit much lower expression levels (4.387-4.388 and 3.404-3.569, respectively), they are located in left (non-specific, noisy) side of the frequency distribution of the gene expression value (Fig. 7, Fig. 8). The probesets B.239931_at and B.244089_at do not correspond to exons of c-*myc* gene, but correspond to ESTs located in the intron. These two target sequences perhaps were designed using incomplete mRNAs sequences and should be considered as noisy.

Note that the frequency distribution of signal intensity values in Fig. 7, Fig. 8 can be described by the mixture of additive and non-linear multiplicative noise-signal functions [Chua *et al.*, 2006]. Based on goodness-of-fit analysis of this model, signal intensity values of less than 6.2 (6.2=ln(500) is the mean value of the normalized signal on a microarray) are described by log-normal probability function and could be considered as mostly additive noise signals. By contrast, signals higher than 6.2 are distributed by the convolution of the (true) Generalized Pareto probability function with the (noise) gamma probability function. This convolution function can serve as a descriptor of the distribution of multiplicative noise-modulated true signals. In the case of data showed on Fig. 7, a fraction of additive noise on microarray U133A is significantly smaller than on microarray U133B: 41% and 60% for U133A and U133B microarrays, respectively. We calculated these numbers as the fractions of cumulative signals obtained between 0 and 6.2 for the frequency distributions in Fig. 7.

We found similar large differences between the quality of hybridization signals on U133A and U133B arrays for several different types of human cells (breast cancer cells, brain cells (not shown)).

To compare changes in signal intensity value distributions between U133A and U133B arrays and between different biological conditions on the same cell lines we constructed normalized distributions in untreated (10 control samples) and 5-aza treated (11 samples) human lung adenocarcinoma cells. It was shown that 5-aza treatment induces hypermethylation and higher expression of large number of genes [Shames *et al.*, 2006]. Fig. 8 shows that "technical" differences (between U133A and U133B arrays) are reproducibly larger than "biological" differences (control and treated samples). Moreover effect of QC filtering on signal intensity value distribution could be comparable with biological variation (Fig. 7 and Fig. 8). These results suggest a much better quality of the probesets on U133A versus U133B and also indicate that inadequate normalization of data on these arrays might be additional source of artifacts in analysis and interpretation of expression data.

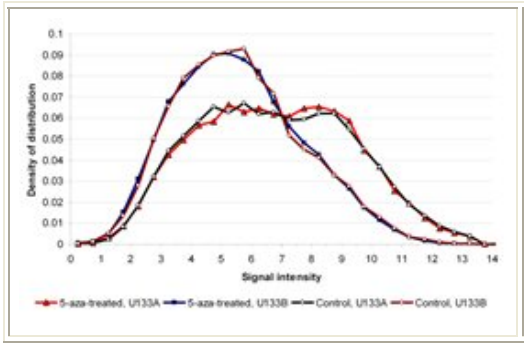

Figure 8: Signal intensity value distribution for two groups of lung cancer cell samples: control and treated by 5-aza groups measured on Affymetrix U133A and U133B arrays. (MAS5 normalization, log-transformed data). Curves present averages of 10 microarrays in control group and 11 microarrays in high-dose 5-aza treatment group.

Discussion and conclusion

Since the sources of noise in microarray experiments may be numerous [Harbig *et al.*, 2005; Wu *et al.*, 2005], the researchers try to minimize the influence of noise and/or estimate it through various quality control, normalization and outlier filtering procedures. One source of variation is cross-hybridization, which occurs when unintended sequences, along with the intended target, hybridize to the same probe, due to sequence homology and/or physicochemical reasons favoring such hybridization. In the case of Affymetrix microarrays, which use a set of short (typically 25-mer) oligonucleotide probes to target a transcript, hybridization conditions are carefully controlled with the aim of minimizing the effect of cross-hybridization due to non-specific binding [Wu *et al.*, 2005]. In addition, each Perfect Match (PM) probe is accompanied by a Mismatch probe (MM), in which the middle residue has been changed. The intention of the PM/MM system is to measure the level of CH associated with each PM probe. A more detailed discussion of cross-hybridization in short oligo microarrays may be found in [Gautier *et al.*, 2005; Harbig *et al.*, 2005]. Affymetrix also displays brief summaries of cross-hybridization

within their own NetAffx service [Liu *et al.*, 2003]. Rather than using homology comparison of every individual probe in our study, we have analyzed only whole target sequences presented by Affymetrix. Further analysis of individual probes can only increase the number of non-reliable probesets.

The evolution of gene definitions has altered Affymetrix target sequence annotation from one genome release to another since U133 GeneChip was designed in 2001 [Dai *et al.*, 2005]. It could increase exact number of unreliable probesets presented in our tables. Our approach for the probesets validation provides necessary background for further quality control filtering.

Decreased reliability of probesets containing or partially containing interspersed repetitive elements was suggested earlier, but statistical estimates from comprehensive datasets have been lacking. Here we presented quantitative estimations of the influence of the repeats on the mean signal intensity value, the CV, the structure of correlation matrix and on the definition of differentially expressed genes in distinct and relatively large groups of samples (from 10 to 83 microarrays per group).

The number of positive correlation coefficients increases as repeat coverage increases. There is a linear trend between repeat coverage and fraction of correlations, increasing for positive and decreasing for negative correlations. The existence of this trend implies that a large number of spurious positive correlations arises in Affymetrix probesets derived from target sequences that have repeats due to hybridization of transcript sequences to more than one probeset. These extra false correlations in the groups do not correspond to real gene co-regulation but solely to sub-optimal design of target sequences. Similarly, Tag2+ and Tag0 can be also a significant source of spurious correlations of signals of probesets (and representative genes) on microarrays. Methodologies such as hierarchical clustering, principal component analysis and relevance networks make direct use of the correlation coefficient of expression signal values between probesets, others methods (such as general linear models) are ultimately based on correlation-like principles. In all these cases, the spurious correlations can lead to serious erroneous interpretation of the microarray results.

Ivshina *et al.*, 2006, and Kuznetsov *et al.*, 2006a, imported Uppsala cohort expression data starting from the feature selection process for all original target sequences, and used the statistically weighted syndrome (SWS) method: a robust class prediction algorithm which discriminated patients with G1 and G3 breast cancers based on a statistically significant and biologically informative 264 gene signature. Interestingly enough, by the criteria presented here, almost all of these 264 probesets (with only two exceptions) were classified as Tag1. Hence, the automatic statistical selection procedure of the SWS method confirms results of Affymetrix probeset selection based on target sequence quality control.

Multiple-locus, non-human, misoriented, and nonspecific probe targets are a significant attribute of the U133A&B GeneChip probesets. The ability of probesets to hybridize to more than one gene product can lead to false positives when analyzing gene expression

data. The apparent artifacts in the data exist because the original target sequence annotations do not accurately reflect the transcripts bound by the targets' probes. For the first time, we quantitatively evaluated the influence of genome repeats and several other sources of inadequate probe sequence design on specificity, reliability and discrimination ability of individual Affymetrix probesets hybridization signals. We also evaluated the influence of probe design and annotation errors on generation of false-positive correlations, which may be an important source of errors in gene co-expression networks constructed using correlation/co-expression matrices.

In conclusion, we recapitulate our principal findings as follows:

2.7% of original Affymetrix target sequences don't match reliably any location in the human genome;

Another 1.7% of the target sequences have multiple locations (up to 10 times and more);

About 7.5% of the remaining Affymetrix target sequences are covered by repeat elements abundant in the human genome completely or over more than 40% of the target sequence length, yielding noisy expression signal;

1.8% of Affymetrix target sequences have wrong orientation relative to the transcript they are alleged to detect.

Identification and removal of the probesets derived from inaccurate target sequences can significantly improve the specificity, sensitivity and reliability of GeneChip technology.

Despite numerous wrongly designed and poorly annotated target sequences, we argue that Affymetrix U133A&B microarrays could show reproducible and quantitative hybridization signals, but about 14% of these signals need filtering based on robust criteria, genome re-annotation and statistical methods described in this paper. We recommend to restrict all analyses of Affymetrix U133A&B probesets to the 86% of artifact-free Tag1 probes with minimal repeat content. The Affymetrix probes annotation and mapping database is available by request to the authors.

Finally, we would like to conclude that careful re-analysis of microarray probe design for different microarray platforms (Affymetrix, Illimina, Agilent, etc) should be an essential component of MicroArray Quality Control (MACQ) project [Patterson *et al.*, 2006; Shi *et al.*, 2006]. This re-analysis would allow the evaluation of performance characteristics as well as of comparability between gene expression microarray techniques.

Acknowledgements

The authors are grateful to Joanne Chen, Li Yi, Yong How Choong and Caleb Khor for help in processing of Affymetrix probeset data.

Grant support: Agency for Science, Technology and Research (A*STAR), Singapore

References

- Chen, J., Sun, M., Kent, W. J., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R. Z. and Rowley, J. D. (2004). Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res. 32, 4812-4820.

- Chua, A. L.-S., Ivshina, A. V. and Kuznetsov, V. A. (2006). Pareto-Gamma Statistics reveals global rescaling in transcriptomes of low and high aggressive breast cancer phenotypes. *In*: Pattern Recognition in Bioinformatics (PRIB-2006), Ragapakese, J. C., Wong, L. and Acharya, R. (eds.), LNCS 4146, Springer-Verlag Berlin-Heidelberg, pp.49-59.

- Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J. and Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 33, e175.

- Gautier, L., Møller, M., Friis-Hansen, L. and Knudsen, S. (2004). Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics 5, 111.

- Harbig, J., Sprinkle R. and Enkemann S. A. (2005). A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. Nucleic Acids Res. 33, e31.

- Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Nordgren, H., Wong, J. E. L., Liu, E. T., Bergh, J., Kuznetsov, V. A. and Miller, L. D. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 66, 10292-10301.

- Katayama, S., *et al.*; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium (2005). Antisense transcription in the mammalian transcriptome. Science 309, 1564-1566.

- Kuznetsov, V. A., Senko, O. V., Miller, L. D. and Ivshina, A. V. (2006a). Statistically weighted voting analysis of microarrays for molecular pattern selection and discovery cancer genotypes. Intern. J. Computer Science and Network Security 6(12), 73-82.

- Kuznetsov, V. A., Zhou, J. T., George, J. and Orlov, Y. L. (2006b). Genome-wide co-expression patterns of human cis-antisense gene pairs. Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure. Novosibirsk, Inst. of Cytology & Genetics vol. 1, pp. 90-93.

- Leong, H. S., Yates, T., Wilson, C. and Miller, C. J. (2005). ADAPT: a database of affymetrix probesets and transcripts. Bioinformatics 21, 2552-2553.

- Liu, G., Loraine, A. E., Shigeta, R., Cline, M., Cheng, J., Valmeekam V., Sun S., Kulp, D. and Siani-Rose, M. A. (2003). NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res. 31, 82-86.

- MAS 5.0 algorithm. Affymetrix. (2002). Statistical Algorithms Description Document. Santa Clara, CA: Affymetrix, Inc. (http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf)

- Mecham, B. H., Wetmore, D. Z., Szallasi, Z., Sadovsky, Y., Kohane, I. and Mariani, T. J. (2004). Increased measurement accuracy for sequence-verified microarray probes. Physiol. Genomics 18, 308-315.

- Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T. and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. USA 102, 13550-13555.

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

- Orlov, Y. L., Zhou, J. T., Lipovich, L., Yong, H. C., Li, Y., Shahab, A. and Kuznetsov, V. A. (2006). A comprehensive quality assessment of the Affymetrix U133A&B probesets by an integrative genomic and clinical data analysis approach. *In*: Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure. Novosibirsk, Inst.of Cytology & Genetics, vol. 1, pp. 126-129.

- Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., Walker, S. J., Zhang, L., Hurban, P., de Longueville, F., Fuscoe, J. C., Tong, W., Shi, L. and Wolfinger, R. D. (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat. Biotechnol. 24, 1140-1150.

- Shames, D. S., Girard, L., Gao, B., Sato, M., Lewis, C. M., Shivapurkar, N., Jiang, A., Perou, C. M., Kim, Y. H., Pollack, J. R., Fong, K. M., Lam, C. L., Wong, M., Shyr, Y., Nanda, R., Olopade, O. I., Gerald, W., Euhus, D. M., Shay, J. W., Gazdar, A. F. and Minna, J. D. (2006). A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. PLoS Med. 3, e486.

- Shi, L., *et al.*; MAQC Consortium (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 24, 1151-1161.

- Stalteri, M. A. and Harrison, A. P. (2007). Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. BMC Bioinformatics 15, 8-13.

- Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl Acad. Sci. USA 98, 5116-5121.

- Wu, C., Carta, R. and Zhang, L. (2005). Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic Acids Res. 33, e84.

- Yelin, R., *et al.* (2003). Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. 21, 379-386.

- Zhang, J., Finney, R. P., Clifford, R. J., Derr, L. K. and Buetow, K. H. (2005). Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. Genomics 85, 297-308.

- Zhang, Y., Liu, X. S., Liu, Q. R. and Wei, L. (2006). Genome-wide i*n silico* identification and analysis of *cis* natural antisense transcripts (*cis*-NATs) in ten species. Nucleic Acids Res. 34, 3465-3475.

APPENDIX. Procedure for selection of misoriented sequences

We consider a target sequence as misoriented relative to the intended gene if the sequence:

1. is aligned perfectly in complete genomic coordinates, block by block (allowed shift is not more than 8 bp except of leftmost and rightmost block) to the transcript mapped to the opposite DNA strand;
2. has a number of blocks greater than one;
3. does not match any RefSeq gene in the same locus on the same strand;
4. perfectly matches, block by block, a majority of GenBank mRNAs in the locus which are located on the opposite strand, while there are none or only a single mRNA perfectly matching the Affymetrix target sequence blocks on the same strand.

We did not use the target gene names provided by Affymetrix. Instead, we independently attempted to determine the transcript whose expression is supposed to be measured by each Affymetrix probeset. Examples are given in Fig. A1.

Since several UCSC annotation tracks generally describe one gene, we checked orientation of Affymetrix target sequences hierarchically: first relative to RefSeq, then to GenBank mRNA, finally relative to EST annotations. We treated 3'ESTs separately from 5'ESTs to define correct transcript direction on chromosome, and assumed that 3'ESTs have a genomic orientation opposite to that of the transcript which they represent. To select correct representative strand we compared first orientation of RefSeq genes in the same locus (if any) orientation, then orientation of mRNA (if any), and only then EST.

Manual curation in UCSC Genome Browser revealed 370 probesets derived from target sequences misoriented relative to the genes at their loci.

Then, using an automated pipeline, we found 138 additional Affymetrix target sequences misoriented relative to RefSeq transcripts, 302 Affymetrix target sequences misoriented relative to transcripts with GenBank mRNA (but no RefSeq) support, and 487 target sequences misoriented relative to EST-supported transcripts with neither GenBank nor RefSeq support. If two different mRNAs were in the same locus and had the same mapped blocks in opposite strands, we selected the correct representative strand based on which of the two mRNAs had a matching RefSeq (if only one of two had it).

In total, we found 810 misoriented Affymetrix target sequences with complete genomic coordinates perfectly matching a transcript on the opposite strand.

After selecting Affymetrix target sequences wrongly oriented relative to intended genes, we also searched for target sequences located in regions of natural antisense transcription, i.e. real transcription from the opposite strand in the same locus. Natural antisense transcripts only partially overlap Affymetrix target sequences and have genomic coordinate sets (intron/exon boundaries) distinct from those of the Affymetrix targets. To avoid errors in mRNA/EST mapping, we used only verified transcript IDs stored in the sense-antisense transcript database developed at GIS.

A

B

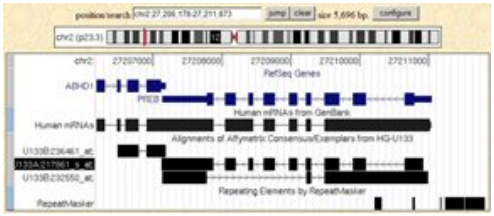

Figure A1: (A) Example of Affymetrix target sequence perfectly matching RefSeq in opposite strand. 200908_s_at matches RPLP2 gene by blocks. This target sequence is marked as "Misoriented". In contrast, probeset 200909_s_at matches the gene correctly. (B) Example of Affymetrix target sequence matching a gene in opposite strand. Probeset 217861_s_at corresponds to PREB gene, probeset 236461_at corresponds to ABHD1 gene. The genes overlap each other by 3'UTRs forming natural antisense transcript pair. Corresponding target sequences should be marked as matching antisense transcript. Probeset 232550_at has target sequence in wrong orientation to the PREB gene. It is marked as "misoriented as verified by manual curation".