Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays

C. Hummert[1], F. Mech[1], F. Horn[1], M. Weber[1], S. Drynda[2], U. Gausmann[3], R. Guthke[1]
[1]Research Group: Systems Biology / Bioinformatics, Hans Knöll Institute Jena
[2]Clinic of Rheumatology, Otto von Guericke University Magdeburg, Medical Faculty
[3]Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)

Abstract—*Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. Nevertheless, the quality of the technology is still capable for further improvements. One of the main problems is cross-hybridization of the transcripts to non-corresponding probes on the array by unspecific binding.*

Four different Affymetrix GeneChip arrays are analyzed, namely the Human Genome arrays HG-U133A, HG-U133B, HG-U133 Plus 2.0 and the Mouse Genome 430 2.0 array. It is shown that putative cross-hybridizations are common for the examined arrays (e.g., 45 % of all probes for the U133A). Furthermore, a considerable amount of probes does not match the annotated transcript correctly. A new set of CDFs is created avoiding putative cross-hybridization completely. It is compared with three other CDFs (Affymetrix, Dai et al.*, Ferrari* et al.*) with the help of correlation between microarray and qRT-PCR results for two datasets. The newly created and the Ferrari CDFs perform significantly better than the original Affymetrix CDFs. The new CDFs are available as R-packages at http://www.sysbio.hki-jena.de/software and have been submitted to BioConductor.*

Keywords: microarrays, unspecific binding, cross-hybridization, Chip Definition Files

1. Background

Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. They are applied to measure changes in expression levels for thousands of genes simultaneously. Until 2011, more than 20,000 measurement series based on microarray technology have been published in public repositories like NCBI's Gene Expression Omnibus.

Nevertheless, the quality of the technology is still capable for further improvements [1], [2]. Several studies tried to compare data derived from different types of arrays and showed a rather poor consistency [3], [4]. Although microarrays are commonly used, this is a daunting problem. In addition, although there has been extended work on this field [5], there is still a lack of standardized experimental protocols among different laboratories [6].

The main problem of microarray analysis is unspecific binding of transcripts by cross-hybridization. This means that RNA fragments hybridize to a probe which is not designed for this gene. It was shown that fragments longer than 8 nucleotides are able to hybridize and that cross-hybridization can emerge from alignments ranging from 10 to 16 nucleotides. Further, the 5'-ends were found to cross-hybridize more likely than the 3'-ends [7].

Unspecific binding may lead to false-positive and false-negative results following in incorrect hypotheses about gene expression [8], [9]. Affymetrix, a technology widely used [10], accounts for the influence of cross-hybridization by introducing internal controls: each probepair comprises a Perfect Match (PM) and a Mismatch (MM) probe which are statistically evaluated [11]. Unfortunately, this procedure cannot solve the problem of cross-hybridization completely [12] and further refinements are suggested [13]. For example, Wu *et al.* [7] stated that the MM probes can also cross-hybridize, even though by another mechanism as the PM probes. Therefore, they recommended ignoring the MM probes.

Generally, expressed transcripts are represented on the array by a series of probepairs called probesets. The signal intensities are summarized to a single value per probeset. A large number of single transcripts are represented by multiple probesets. Multiple probesets representing the same gene are expected to show similar fold changes calculated from the signal intensities of the hybridized samples. However, this is in fact not the case [14], [15], [16]. This problem arises from single probes in the probeset which are capable of cross-hybridization. Ways to deal with this problem is either a probe-based analysis, leaving out the probe-to-probeset summarization step [17], [18], or the composition of the probesets could be improved by setting up alternative Chip Definition Files (CDFs) based on information contained in different sequence databases. For example, the group of Ferrari *et al.* [19] created a set of custom CDFs based on the GeneAnnot database [20]. In these CDFs the probesets that match the same gene were merged into one probeset. Hence, the existence of more than one probeset per gene was eliminated, avoiding discordant expression signals for the same transcript.

Another set of custom CDFs relying on a broad repertoire of databases like RefSeq or Unigene has been created by the group of Dai *et al.* [21]. Probesets matching the same gene were merged, but remained divided if they were able to discriminate different isoforms of a gene. Probes causing cross-hybridizations were removed from the new probesets, but the filter had been not very strict.

Several groups dealt with the question of the minimum probeset size [19], [21]. For example, the group of Lu *et al.* [22] sets the minimum probeset size to 4 probes because smaller probesets result in high error rates. In this study the minimum probeset size was set to 4 [19], [21]. From these new probesets custom CDFs and the corresponding Bioconductor libraries for Affymetrix GeneChips were created.

In the work presented here, a new set of CDFs is introduced avoiding putative cross-hybridization completely. These CDFs are compared with those from Affymetrix, Ferrari, and Dai by validation of the respective microarray results using qRT-PCR for two different datasets.

2. Results

Four different Affymetrix GeneChip arrays are analyzed, namely the HG-U133A, HG-U133B, HG-U133 Plus 2.0 designed for human samples, and the Mouse Genome 430 2.0 array. For the detection of putative cross-hybridizations, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using blastn [23] as described in the methods section.

The GeneChip HG-U133A consists of 22,283 probesets, each of 11–20 probepairs and 247,937 probepairs in total. Additional 1,155 probepairs are controls and are furthermore ignored. About 44 % of the PM probes (109,245) match exactly one single gene. 11 % of the probes (26,159) do not match any annotated gene. 45 % of the probes (112,533) match more than one gene and thus have cross-hybridization potential.

Furthermore, the direction of the probes was analyzed. Normally, sense strand RNA fragments are expected, although there are some loci in the human genome [24], as well as in the mouse genome [25], where both sense and antisense strands are transcribed. However, mixing up probes detecting sense or antisense strands in one single probeset could cause wrong expression results. Here, only probes matching the sense strand are considered as correct. For the U133A microarray all probes match the sense strand.

The GeneChip HG-U133B consists of 22,645 probesets, each of 11–20 probepairs and 249,491 probepairs in total. Again, there are additional probesets containing more than 11 probes as controls and are ignored (1,100). About 35 % of the probes (87,067) are found to match exactly one gene. 2 % of the probes (5,453) match more than one gene, so they possibly cross-hybridize, 5 % of the probes (12,805) match at least one gene but in the wrong direction (antisense direction) and no gene in the sense direction, and 58 % of the probes (144,166) do not match any annotated gene.

The GeneChip HG-U133 Plus 2.0 consists of 54,675 probesets and 604,247 probepairs. Like in the other arrays, additional probesets containing more than 11 probes are controls and are discarded. Here, 37 % of the remaining probes (221,821) match exactly one gene, 23 % of the probes (141,146) match more than one gene, 11 % of the probes (65,327) match at least one gene but in the wrong direction (antisense direction) and no gene in the sense direction, and 29 % of the probes (175,953) do not match any annotated gene.

The Mouse Genome 430 2.0 array consists of 45,036 probesets and 496,457 probepairs. About 52 % of the counted probes (257,331) match exactly one gene and 5 % of the probes (27,112) match more than one gene. About 1 % of the probes (4,661) match genes only in the wrong direction and 42 % of the probes (207,353) do not match any annotated gene.

Nearly all Affymetrix probesets contain at least one probe which has cross-hybridization potential. In fact, for the HG-U133 Plus 2.0 Chip about 65 % of all probesets include more cross-hybridizing probes than non-ambiguous ones.

All probes matching exactly one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes, that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. Only the good probes are used to create the new CDFs as described in the methods chapter. Accordingly, for the HG-U133A microarray originally measuring 14,500 genes by 22,283 probesets the newly created CDF contains 12,400 probesets representing 12,400 genes. For the HG-U133 Plus 2.0 the number of probesets is reduced from 54,675 (representing 38,500 genes) to 18,800 (representing 18,800 genes). The HG-U133B comprises 22,645 probesets measuring the expression of 18,400 genes. Here, the number of probesets is reduced to 6,500 matching 6,500 transcripts. The Mouse 430 2.0 microarray consists of 45,036 probesets for 39,000 genes. With the new CDF there are 16,400 probesets matching 16,400 genes. Hence, the number of identifiable genes is reduced in order to achieve a higher specificity of the probesets. The result for the HG-U133 Plus 2.0 is in good agreement to the results of Barnes *et al.* [26], who used BLAT and the Golden Path database and achieved a number of 17,143 genes that can be measured.

Small probesets lead to higher error rates and result in lower statistical significance. In the Affymetrix CDFs the size is 11 for nearly all probesets, but in the newly created probesets the size is not fixed. Some probesets are smaller than those from Affymetrix due to the removal of the problematic probes. However, many probesets increase in size due to useful probes on the array that have not been used for the matching gene before and probesets measuring the same gene beeing merged. For example, for the HG-U133 Plus 2.0 the mean probeset size increases from 11 to 17.

For the validation of all CDFs two test datasets are chosen: (i) the Etanercept (ETC) and (ii) the MAQC dataset. The first of the two datasets is derived from a study analyzing the effect of the TNF-α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points [27]. It is a typical dataset that arises in medical studies and is rather representative. One Affymetrix HG-U133A array experiment was performed for each time point. The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. The subset of the 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here.

qRT-PCR results are considered to reflect the real transcript concentrations with higher reliability than those determined by microarrays. Therefore, qRT-PCR experiments are regarded as a 'gold standard' for chip analyses [29], [30]. The Pearson correlation coefficient (PCC) of the microarray and the qRT-PCR data is computed for each gene using the different CDFs.

For the Etanercept dataset we performed qRT-PCR experiments for 16 genes. In total, this dataset now contains results from 51 microarrays and 816 qRT-PCR experiments. In addition, the genes with qRT-PCR data in both records are analyzed in more detail.

The perfomance of these CDFs were compared: the original Affymetrix CDFs (A), the two alternative CDFs of Ferrari *et al.* (F) [19] and Dai *et al.* (D) [21], and the new CDFs (H) presented here. The CDFs from Ferrari, using the GeneAnnot database, contain merged probesests (see background chapter), and cross-hybridization was not considered. The group of Dai offers a broad spectrum of different CDFs based on different databases. The one using RefSeq is chosen for comparison because it corresponds best to the new CDFs, using RefSeq as well. In the Dai CDFs different probesets matching a single gene are combined, although there are exceptions for genes comprising different isoforms. A check for cross-hybridization is also included. However, it applies a different algorithm than the new CDFs and the filter is much less strict.

For the probe to probeset summarization step two algorithms are used as described in the methods section: (i) the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and (ii) the Affymetrix Microarray Suite MAS5 [32]. These were compared repeatedly, but it is difficult or even impossible to decide which of the both algorithms performs better in any case [33], [34], [35].

For the Etanercept dataset, the mean correlation coefficient of all 16 genes for the Affymetrix CDF is 0.61 using the robust multi-array analysis algorithm (RMA) and 0.60 using the Affymetrix Microarray Suite MAS5. These values include 31 probesets in total matching these 16 genes according to the Affymetrix annotation file. If only the best correlating probeset for each gene is considered, the average correlation coefficient increases to 0.73 for RMA and 0.71 for MAS5. However, this value is more of theoretical interest because the knowledge which probeset will perform best is gained not until the qRT-PCR experiments and correlation analysis is finished. On average, the incorporated probesets contain 5.58 putative cross-hybridizations calculated by BLAST (4.47 including only the best performing probesets).

The Dai CDF contains 23 probesets for the 16 genes of the Etanercept dataset. Their mean correlation coefficient increases to 0.67 for both RMA and MAS5 compared to the 0.60 using the Affymetrix CDF. Considering the best correlating Dai probesets only, the values further increase to 0.73 for RMA and 0.69 using MAS5. The mean size of the Dai probesets increases to 20.59 probes containing 8.82 putative cross-hybridizations. This number changes to 4.71 if normalized to a probeset size of 11. Here, normalization means the number of putative cross-hybridizations calculated for a hypothetical Dai probeset size of 11. Considering only the best Dai probesets, the number of putative cross-hybridizations decreases to 7.88 on average.

For the Ferrari CDF, the mean correlation coefficient equals 0.73 for RMA and 0.69 using MAS5 on average. The mean probeset size increases to 19.56, harboring 10.81 possible cross-hybridizations (6.07 if normalized).

Using the new CDF the mean correlation coefficient amounts to 0.72 for RMA and 0.68 for MAS5. The mean probeset size decreases to 10.25 with no cross-hybridizations at all. The detailed results are shown in the table below:

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset- size
TNF	A: 207113_s_at	0.88	0.85	N/A	8	11
	D: NM_000594_at	0.88	0.85	N/A	8	11
	F: GC06P031652_at	0.88	0.85	N/A	8	11
	H: gi_25952110	0.86	0.81	N/A	0	3
IL1B	A: 205067_at	0.95	0.90	0.37	6	11
	A: 39402_at	0.95	0.87	0.82	6	16
	D: NM_000576_at	0.96	0.89	0.74	12	27
	F: GC02M113303_at	0.96	0.89	0.74	12	27
	H: gi_27894305	0.95	0.88	0.86	0	15
IL6	A: 205207_at	0.69	0.71	0.81	3	11
	D: NM_000600_at	0.69	0.71	0.81	3	11
	F: GC07P022732_at	0.69	0.71	0.81	3	11
	H: gi_10834983	0.65	0.72	0.71	0	8
IL8	A: 202859_x_at	0.88	0.81	0.90	6	11
	A: 211506_s_at	0.86	0.73	0.98	6	11
	D: NM_000584_at	0.88	0.73	0.96	12	22
	F: GC04P074845_at	0.88	0.73	0.96	12	22
	H: gi_28610153	0.89	0.73	0.95	0	10
IL1RN	A: 212657_s_at	0.75	0.87	N/A	2	11
	A: 212659_s_at	0.77	0.84	N/A	4	11
	A: 216243_s_at	0.75	0.86	N/A	6	11
	A: 216244_s_at	0.13	0.07	N/A	4	11
	A: 216245_at	0.21	0.11	N/A	10	11
	D: NM_173841_at	0.80	0.88	N/A	12	33
	D: NM_000577_at	0.80	0.88	N/A	12	33
	D: NM_173842_at	0.80	0.88	N/A	12	33
	D: NM_173843_at	0.84	0.86	N/A	15	42
	F: GC02P113591_at	0.83	0.86	N/A	16	44
	H: gi_27894315	0.78	0.88	N/A	0	23
ICAM1	A: 202637_s_at	0.63	0.73	0.97	7	11
	A: 202638_s_at	0.62	0.72	0.98	4	11
	A: 215485_s_at	0.71	0.73	0.94	3	11
	D: NM_000201_at	0.70	0.76	0.99	14	33
	F: GC19P010247_at	0.70	0.77	0.99	14	33
	H: gi_4557877	0.72	0.74	0.97	0	20
SOD2	A: 215078_at	0.25	0.35	N/A	10	11

Continued on next page

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset- size
	A: 215223_s_at	0.15	0.28	N/A	7	11
	A: 216841_s_at	0.18	0.39	N/A	3	11
	A: 221477_s_at	0.32	0.44	N/A	10	11
	D: NM_001024466_at	0.16	0.33	N/A	6	12
	D: NM_000636_at	0.19	0.37	N/A	10	22
	D: NM_001024465_at	0.16	0.33	N/A	6	13
	F: GC06M160020_at	0.20	0.36	N/A	20	33
	H: gi_67782304	0.20	0.39	N/A	0	12
TRAF1	A: 205599_at	0.61	0.50	0.88	6	11
	D: NM_005658_at	0.61	0.50	0.88	6	11
	F: GC09M122704_at	0.61	0.50	0.88	6	11
	H: gi_53759116	0.59	0.47	0.89	0	5
ZFP36	A: 201531_at	0.84	0.86	N/A	5	11
	A: 213890_x_at	-0.01	-0.46	N/A	8	11
	D: NM_003407_at	0.84	0.86	N/A	5	11
	F: GC19P044589_at	0.84	0.86	N/A	5	11
	H: gi_141802261	0.85	0.82	N/A	0	6
PTGS2	A: 204748_at	0.91	0.71	0.97	4	11
	D: NM_000963_at	0.91	0.71	0.97	4	11
	F: GC01M184907_at	0.91	0.71	0.97	4	11
	H: gi_4506264	0.89	0.72	0.95	0	9
TNFAIP3	A: 202643_s_at	0.78	0.82	0.97	4	11
	A: 202644_s_at	0.87	0.85	0.93	6	11
	D: NM_006290_at	0.82	0.83	0.96	10	22
	F: GC06P138230_at	0.82	0.83	0.96	10	22
	H: gi_26051241	0.80	0.82	0.98	0	13
DUSP2	A: 204794_at	0.75	0.66	N/A	5	11
	D: NM_004418_at	0.75	0.66	N/A	5	11
	F: GC02M096230_at	0.75	0.66	N/A	5	11
	H: gi_12707563	0.74	0.60	N/A	0	6
ADM	A: 202912_at	0.80	0.67	0.92	5	11
	D: NM_001124_at	0.80	0.67	0.92	5	11
	F: GC11P010283_at	0.80	0.67	0.92	5	11
	H: gi_4501944	0.82	0.67	0.94	0	6
CROP	A: 203804_s_at	0.44	0.56	N/A	5	11
	A: 208835_s_at	0.43	0.36	N/A	5	11
	A: 220044_x_at	0.43	0.44	N/A	4	11
	D: NM_016424_at	0.49	0.50	N/A	13	32
	D: NM_006107_at	0.49	0.45	N/A	13	30
	F: GC17P046151_at	0.48	0.48	N/A	14	33
	H: gi_52426741	0.46	0.47	N/A	0	17
NFκBIA	A: 201502_s_at	0.81	0.73	N/A	4	11
	D: NM_020529_at	0.81	0.73	N/A	4	11
	F: GC14M034940_at	0.81	0.73	N/A	4	11
	H: gi_10092618	0.82	0.77	N/A	0	7
JUNB	A: 201473_at	0.44	0.44	0.94	7	11
	D: NM_002229_at	0.44	0.44	0.94	7	11
	F: GC19P012763_at	0.44	0.44	0.94	7	11
	H: gi_44921611	0.54	0.44	0.73	0	4
Ø	all Affymetrix	0.61	0.59	0.88	5.58	11.16
	best Affymetrix	0.73	0.71	0.92	4.47	11.00
	Dai	0.67	0.67	0.91	8.82	20.59
	best Dai	0.73	0.69	0.91	7.88	18.69
	Ferrari	0.73	0.69	0.91	10.81	19.56
	Hummert	0.72	0.68	0.89	0.00	10.25

Evaluating the PM and MM probes statistically, the MAS5 software assigns ’present’, ’absent’ or ’marginal’ to each expression value, and Affymetrix recommends to use only the ’present’ detection call for further analysis. Following this recommendation and using only those results for the correlation analysis that are marked as ’present’ the mean correlation coefficient increases from 0.59 to 0.66 (0.74 including only the best performing probesets). Hence, incorporating the Affymetrix detection call indeed improves the correlation, but using alternative CDFs is still better than using the Affymetrix probesets and the detection call.

Analyzing the MAQC reference dataset using the RMA suite, the results are almost in accordance with those of the Etanercept data described above. The mean correlation coefficient for all 1,000 genes is 0.47 for the Affymetrix CDF (0.71 incorporating only the best probeset for each gene). Using the Dai CDF, the mean correlation increases to 0.63 (0.64 for the best probesets). With the Ferrari and the new CDF the mean correlations are 0.63 and 0.58, respectively. The detailed results for all MAQC genes can be downloaded.

Discussion

Results from microarray experiments contain considerably high error rates [36]. Due to error propagation, it is of particular importance to minimize errors in the beginning of the analysis chain [37]. Therefore, especially the preprocessing of the chip data has to be done as accurate as possible. Many efforts were spent on these problems before [38], such as the notable results of the ’Golden Spike Project’ [6]. The question which statistical method should be adequately chosen is even more complicated if experimental data from different laboratories are incorporated in one single analysis [39].

For microarray analyses algorithms are essential which combine the 11-20 probepair intensities for a given gene and define a measure of expression that represents the amount of the corresponding mRNA species. In this study, two of these algorithms are compared, the robust multi-array analysis algorithm (RMA) and the Affymetrix Microarray Suite MAS5. Applying both algorithms to the Etanercept dataset RMA outperforms MAS5 on average. Other studies revealed similar results. However, their performance is assumed to be dependent on the actual dataset [40]. In fact, normalisation steps are applied after the probe to probeset summarization. Some of these steps depend on global parameters (e.g. mean of total gene expression) which depend on the total set of probesets. Therefore, identical probesets within different CDFs vary slightly in the final gene expression values.

Analyzing the probes of the Affymetrix microarrays discloses many inaccuracies. A large number of problematic probes are based on the fact that Affymetrix had to rely on genome annotation available at the time the chips were designed (U133A and U133B: 2001; U133 Plus 2.0 and Mouse 430 2.0: 2003). Because genome annotation improves permanently, the chip design does not properly match the present annotations anymore. Due to compatibility reasons, Affymetrix is not able to keep the design of their microarrays up to date.

The problem of cross-hybridization is well known. The first work on custom CDFs examining this error source was published by the group of Dai in 2005 [21]. They created a large amount of high quality custom CDFs related to different reference databases. Some probes, causing cross-hybridizations, are deleted from the probesets, but the filter is quite loose, so the number of problematic probes decreased but did not vanish. The use of the new CDFs can avoid full length, i.e., 25 mer long, cross-hybridizations completely. Cross-hybridization of shorter fragments are very difficult to handle due to the fact that the number of putative bindings grows exponentially the shorter the considered fragments are. Hence, if all putatively cross-hybridizing probes are excluded the amount of measurable genes will be reduced extremely.

The underlying gene annotation which is used for sequence alignment has a big impact on the number of cross-hybridizations. Manually curated mRNA sequences have a high chance of missing transcripts. Therefore, the inclusion of computational proposed gene annotations decreases the number of false negative predicted cross-hybridizations. The drawback is that a number of false positive hybridizations increases. A more strict approach should be preferred, because it does not significantly decrease the number of covered transcripts as there is a high amount of availabe probes. In this study, the exclusion of XM-RefSeq-accessions results in smaller differences between the different CDFs in the number of putative cross-hybridzing transcripts. Interestingly, the correlation coeefficents of the newly created probesets do not change significantly.

Evaluating the four different CDFs, we figured out that the usage of the original Affymetrix CDFs leads to poorer results than the usage of the custom CDFs, although the best Affymetrix probesets give equally good or even better results than the other CDFs. However, as already mentioned, this cannot be taken into account, because it is not known which probeset will perform best before the correlation analysis is completed. The Dai probesets perform better, but the problem of several probesets representing a single gene had not been solved. Although multiple probesets representing the same gene are expected to show similar signal intensities, this is in fact not the case [14], [15]. Thus, it is difficult to decide which of the probesets matching the same gene is the most reliable. The Ferrari and the new CDFs comprise only one probeset per gene, which is of great advantage. The Ferrari CDFs perform slightly better on the Etanercept dataset and both CDFs perform equally well on the MAQC data.

The analysis of the genes for which qRT-PCR results are available in the Etanercept dataset as well as in the MAQC dataset clearly shows higher correlation coefficients in the MAQC dataset. This is most likely due to the fact that the U133 Plus 2.0 arrays which were used in the MAQC dataset outperform the older U133A microarrays.

The results show that probesets consisting of more probes, i.e., larger probesets, lead to better correlation results in general, whereas smaller probesets perform poorer. This finding correlates to the results of the study of Cui *et al.* [14] that merges probesets matching the same transcript. Interestingly, probesets containing many putative cross-hybridizations do not considerably perform poorer than probesets containing only a few. This result is very surprising, because it is obvious that cross-hybridization is one of the main error sources in microarray experiments [8], [9]. The normalization step in the two summarizing algorithms RMA and MAS5 may explain for that because they possibly eliminate some cross-hybridization effects. Another explanation is that leaving out the problematic probes does not compensate the influence of cross-hybridization. Unspecific binding leads to two types of error: (i) false-positives because RNA fragments bind to problematic probes of the probeset, and (ii) gene expression events are missed or underestimated, leading to a false-negative error if the RNA fragments are already bound to problematic probes of other probesets (competitive binding). Custom CDFs can only account for the first type of error by leaving out the problematic probes, the second effect could only be overcome by better array design.

The newly created CDFs perform slightly poorer than the Ferrari probesets (0.72 vs. 0.73) on the Etanercept dataset and equally well on the much larger MAQC dataset. On the one hand, the Ferrari CDFs can obviously countervail the negative effect by their much larger probesets in comparison to the new CDFs. On the other hand, using the new CDFs, putative cross-hybridizations are systematically excluded whereas using the Ferrari CDFs, the negative effect vanishes for statistical reasons due to the larger probesets. For exact studies, it is better to avoid a putative error source instead of averaging the cross-hybridization effects out as the Ferrari CDFs do. In addition, it has to be mentioned that the new CDFs provide as good or better results as the other CDFs using only about half the amount of probes (HG-U133A: 44 %, HG-U133B: 35 %, HG-U133 Plus 2.0: 37 %, Mouse Genome 430 2.0 Array: 52 %). Hence, designing new microarrays without the problematic probes, the dimension can be reduced by half without loosing any information and minimize the costs of the technology tremendously. Future microarray design using only the good probes and incorporating probesets of large sizes like in the Ferrari CDFs will certainly provide optimal solutions.

Methods

Probe Analysis

For the detection of putative cross-hybridizations by sequence alignment, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using blastn [23]. For the U133A and the U133 Plus 2.0 the RefSeq release from 05/14/07 was used (download from ftp://ftp.ncbi.nih.gov/-refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz), for the U133B the realease from 01/10/08, and for the Mouse 430 2.0 microarray the release from 05/09/08 (∼M_musculus/mRNA_Prot/mouse.rna.fna.gz) was used. These parameters were applied: ValW = 7, ValE = 1000, ValHspmax = 1.

In this work all those RefSeq accession numbers beginning with XM or NM are used. The XM-identifiers indicate mRNA-RefSeq-accessions which are produced by computationally annotated genome submissions. The NM-identifier show that the RefSeq records are subsequently curated. Using both accessions in our model leads to more predicted cross-hybridizations which increases the reliability of the specificity of the probes.

The strand direction of the probes is analyzed. For each probe it is counted how many genes match and checked whether the match has the correct direction, i.e., the sense direction.

All BLAST hits for different transcript isoforms are merged, i.e., if the probe hybridizes to alternative splice variants of one gene but not to another gene, it is considered as unambiguous. Different gene isoforms of one gene are identified by screening the gene descriptions of the RefSeq database.

All probes matching only one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. For the creation of the new CDFs only the good probes are used. The probe sequences are annotated with GeneIDs derived from RefSeq. The GeneID is a database cross-reference qualifier, which supports access to the Entrez Gene database and provides a distinct tracking identifier for a gene or locus. Probes sharing the same GeneID are grouped together into a new probeset. The intersection between two different probesets is therefore always empty for all probesets. The size of the newly created probesets is variable and not fixed to 11 like in the Affymetrix CDFs.

Datasets

Two datasets were chosen for the validation of the different CDFs. The first of the two datasets chosen is derived from a study published by Koczan *et al.* [27] analyzing the effect of the TNF-α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points. One Affymetrix HG-U133A array was performed for each time point. The data are available at the Array Express archive [41] with the accession number E-MTAB-11.

Expression levels of 16 genes were measured by quantitative real-time RT-PCR (qRT-PCR) performed with TaqMan assay reagents according to the manufacturer's instructions on a 7900 High Throughput Sequence Detection System (Applied Biosystems, Foster City, CA, USA) using predesigned primers and probes (GAPDH Hs99999905_m1, ICAM1 Hs00164932_m1, TNFAIP3 Hs00234713_m1, IL1B Hs00174097_m1, NFκBIA Hs00153283_m1, IL8 Hs00174103_m1, ADM Hs00181605_m1, TNF Hs00174128_m1, IL6 Hs00174131_m1, IL1RN Hs00277299_m1, SOD2 Hs00167309_m1, TRAF1 Hs00194638_m1, ZFP36 Hs00185658_m1, PTGS2 Hs00153133_m1, DUSP2 Hs00358879_m1, CROP Hs00538879_s1, JUNB HS00357891_s1).

The threshold cycle values (C_T) for specific mRNA expression in each sample were normalized to the C_T values of GAPDH mRNA in the same sample. This provides ΔC_T values that were used for the correlation analysis. In total, 816 qRT-PCR experiments were performed and complement the 51 microarray experiments (17 patients, 3 time points) described in [27]. The results of the qRT-PCR experiments can be downloaded.

The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. All available 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here. The MAQC data discussed in this publication are available in NCBI's Gene Expression Omnibus with accession number GSE5350. In addition, the nine genes for which qRT-PCR results are available in both datasets, are analyzed in more detail.

Comparison of the CDFs

For the comparison of different CDFs, the correlation between the microarray and the qRT-PCR experiments is used [29], [30]. As a performance index the Pearson correlation coefficient of the microarray results and the qRT-PCR experiments is calculated. Calculation of the Spearman correlation coefficient showed very similar results (data available at http://sysbio.hki-jena.de/software).

The raw chip data (CEL Files) are analyzed using the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and the Affymetrix Microarray Suite MAS5 [32] in combination with the different CDFs.

The MAS5 software assigns 'present', 'absent' or 'marginal' to each expression value, and Affymetrix recommends to use only the 'present' detection call for further analysis [32]. For an additional correlation analysis only the 'present' probesets are used to check if the calculated detection call from MAS5 gives a good prediction for the probeset quality.

Availability

The newly created CDFs as R-packages and additional files are available for download at http://www.sysbio.hki-jena.de/software. Using the CDFs does not interfere with all further steps of microarray analysis.

Acknowledgements

This work was supported by the ILRS - International Leibniz Research School for Microbial and Molecular Interactions (CH, FH) and by the ERASysBio+ project Linconet (MW).

References

[1] S. Heber and B. Sick, "Quality assessment of Affymetrix GeneChip data," *OMICS A Journal of Integrative Biology*, vol. 10, no. 3, pp. 358–368, Fall 2006.

[2] O. Modlich and M. Munnes, "Statistical framework for gene expression data analysis," *Methods in Molecular Biology*, vol. 377, pp. 111–130, May 2007.

[3] P. K. Tan, T. J. Downey *et al.*, "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5676–5684, October 2003.

[4] A.-K. Järvinen, S. Hautaniemi *et al.*, "Are data from different gene expression microarray platforms comparable?" *Genomics*, vol. 83, no. 6, pp. 1164–1168, June 2004.

[5] A. Brazma, P. Hingamp *et al.*, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nature Genetics*, vol. 29, no. 4, pp. 365–371, December 2001.

[6] S. E. Choe, M. Boutros *et al.*, "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset," *Genome Biology*, vol. 6, no. 2, p. R16, January 2005.

[7] C. Wu, R. Carta, and L. Zhang, "Sequence dependence of cross-hybridization on short oligo microarrays," *Nucleic Acids Research*, vol. 33, no. 9, p. e84, May 2005.

[8] Z. Chen, M. McGee *et al.*, "A distribution free summarization method for Affymetrix GeneChip® arrays," *Bioinformatics*, vol. 23, no. 3, pp. 321–327, February 2007.

[9] A. C. Cambon, A. Khalyfa *et al.*, "Analysis of probe level patterns in Affymetrix microarray data," *BMC Bioinformatics*, vol. 8, no. 146, May 2007.

[10] H. R. Ueda, S. Hayashi *et al.*, "Universality and flexibility in gene expression from bacteria to human," *The Proceedings of the National Academy of Sciences (US)*, vol. 101, no. 11, pp. 3765–3769, March 2004.

[11] Affymetrix Inc, "GeneChip custom express array design guide. part no. 700506 rev. 4," Tech. Rep., 2003.

[12] L. Zhang, M. F. Miles, and K. D. Aldape, "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnology*, vol. 21, no. 7, pp. 818–821, July 2003.

[13] B. M. Bolstad, R. A. Irizarry *et al.*, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 3, pp. 185–193, January 2003.

[14] X. Cui and A. E. Loraine, "Consistency analysis of redundant probe sets on Affymetrix three-prime expression arrays and applications to differential mRNA processing," *PLoS One*, vol. 4, no. 1, p. 4229, January 2009.

[15] T. R. Hughes, M. Mao *et al.*, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer," *Nature Biotechnology*, vol. 19, no. 4, pp. 342–347, April 2001.

[16] M. A. Stalteri and A. P. Harrison, "Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips," *BMC Bioinformatics*, vol. 8, no. 13, January 2007.

[17] X. Liu, M. Milo *et al.*, "Probe-level measurement error improves accuracy in detecting differential gene expression," *Bioinformatics*, vol. 22, no. 17, pp. 2107–2113, September 2006.

[18] G. Sanguinetti, M. Milo *et al.*, "Accounting for probe-level noise in principal component analysis of microarray data," *Bioinformatics*, vol. 21, no. 19, pp. 3748–3754, October 2005.

[19] F. Ferrari, S. Bortoluzzi *et al.*, "Novel definition files for human GeneChips based on GeneAnnot," *BMC Bioinformatics*, vol. 8, no. 446, November 2007.

[20] V. Chalifa-Caspi, I. Yanai *et al.*, "GeneAnnot: Comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes," *Bioinformatics*, vol. 20, no. 9, pp. 1457–1458, June 2004.

[21] M. Dai, P. Wang *et al.*, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data," *Nucleic Acids Research*, vol. 33, no. 20, p. e175, November 2005.

[22] J. Lu, J. C. Lee *et al.*, "Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays," *BMC Bioinformatics*, vol. 8, no. 108, March 2007.

[23] S. McGinnis and T. L. Madden, "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Research*, vol. 32, pp. W20–W25, July 2004.

[24] R. Yelin, D. Dahary *et al.*, "Widespread occurrence of antisense transcription in the human genome," *Nature Biotechnology*, vol. 21, no. 4, pp. 379–386, April 2003.

[25] H. Kiyosawa, N. Mise *et al.*, "Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized," *Genome Research*, vol. 15, no. 4, pp. 463–474, April 2005.

[26] M. Barnes, J. Freudenberg *et al.*, "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5914–5923, October 2005.

[27] D. Koczan, S. Drynda *et al.*, "Molecular discrimination of responders and nonresponders to anti-TNFalpha in rheumatoid arthritis therapy by Etanercept," *Arthritis Research & Therapy*, vol. 10, p. R50, May 2008.

[28] L. Shi, L. H. Reid *et al.*, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, September 2006.

[29] J. S. Moray, J. C. Ryan, and F. M. Van Dolah, "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR," *Biological Procedures Online*, vol. 8, no. 1, pp. 175–193, December 2006.

[30] R. D. Canales, Y. Luo *et al.*, "Evaluation of DNA microarray results with quantitative gene expression platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1115–1122, September 2006.

[31] R. A. Irizarry, B. Hobbs *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, April 2003.

[32] Affymetrix Inc, "Statistical algorithms description document. whitepaper. part no. 701137 rev. 3," Tech. Rep., 2002.

[33] R. A. Irizarry, Z. Wu, and H. A. Jaffee, "Comparison of Affymetrix GeneChip expression measures," *Bioinformatics*, vol. 22, no. 7, pp. 789–794, July 2006.

[34] J. Seo and E. P. Hoffman, "Probe set algorithms: is there a rational best bet?" *BMC Bioinformatics*, vol. 7, no. 395, August 2006.

[35] S. D. Pepper, E. K. Saunders *et al.*, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, no. 273, July 2007.

[36] M. Eisenstein, "Microarrays: Quality control," *Nature*, vol. 442, pp. 1067–1070, August 2006.

[37] M. Grabe, *Measurement Uncertainties in Science and Technology*. New York: Springer Press, 2005.

[38] P. Boutros, "Systematic evaluation of the microarray analysis pipeline," in *Proceedings of the First 11th MGED Meeting: 1-4 September 2008; Riva del Garda*, G. Sherlock, Ed. MGED, 2008, pp. 16–27.

[39] H.-C. Liu, C.-Y. Chen *et al.*, "Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 570–579, August 2008.

[40] K. Shedden, W. Chen *et al.*, "Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data," *BMC Bioinformatics*, vol. 6, no. 26, 2005.

[41] H. Parkinson, M. Kapushesky *et al.*, "ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D868–D872, January 2009.