American Journal of Bioinformatics 1 (1): 1-9, 2012
ISSN 1948-9862

A Sequence Based Validation of Gene Expression Microarray Data

[1]Gerhard G. Thallinger,
[2]Eva Obermayr, [1]Pornpimol Charoentong,
[2]Dan Tong, [1]Zlatko Trajanoski and [2,3]Robert Zeillinger
[1]Institute for Genomics and Bioinformatics,
Faculty of Electrical and Information Engineering,
Graz University of Technology, Graz, Austria
[2]Department of Obstetrics and Gynecology,
Comprehensive Cancer Center,
Medical University of Vienna, Vienna, Austria
[3]Ludwig Boltzmann Gesellschaft-Cluster Translational Oncology, Vienna, Austria

Abstract: Problem statement: Quantitative Reverse Transcription PCR (RT-qPCR) is often used to validate microarray data. Previous studies show different levels of correlation, without further investigation of influencing factors. **Approach:** We compared expression levels of 381 genes obtained from microarray hybridizations and from TaqMan based RT-qPCR assays. Correlation of expression levels was determined by comparing: (i) single genes across samples, (ii) all genes within a sample and (iii) the expression ratios of all genes in a sample using another sample as the reference. The influence of several parameters on the correlation was analyzed: (i) variation in transcript set targeted by the microarray probe and the PCR assay, (ii) variation in amplicon probe position relative to 3' end of transcript, (iii) variation in efficiency of the PCR reaction and (iv) normalization of the PCR data. **Results:** The 381 genes covered by RT-qPCR had 494 matching probes on the microarray. 397 probes with a matching transcript set were identified via a rigid sequence-based validation. Correlation was significantly higher among matching transcript sets and probes closer to the 3' end. Adjustments for different amplification efficiencies had either no influence or decreased correlation. Normalization of qPCR data consistently reduced correlation for all analysis approaches. **Conclusion:** Current clinical research uses microarrays to select genes of interest and evaluates these genes using qPCR. Therefore, it is important that expression levels measured by both techniques be highly correlated. High correlation can be achieved if the targeted transcript sets match, whereas normalization and efficiency correction can have a negative influence.

Key words: DNA microarrays, TaqMan quantitative reverse-transcription PCR, validation

INTRODUCTION

Quantitative Reverse Transcription PCR (RT-qPCR) is often used to validate gene expression measurements from DNA microarray experiments. Comparison with PCR is done primarily to underpin findings derived from microarray data, even as a requirement for publication (Rockett and Hellmann, 2004). An additional application is molecular fingerprinting (Veer *et al*., 2002), where an unknown sample (in general a tumor specimen) is classified by RT-qPCR using a small subset of discriminating genes, which typically has been determined by microarray hybridizations of known samples. In both applications, it is very important to know what level of agreement can be expected between RT-qPCR and microarray data. For microarray data confirmation, agreement on the direction of fold-change could be enough (Rajeevan *et al*., 2001), whereas sample classification has much more stringent agreement requirements (Perreard *et al*., 2006).

Several studies comparing microarray and RT-qPCR data have been published, which show differing levels of agreement between the platforms, ranging from negative correlation for some genes to almost perfect agreement for others (Zhang *et al*., 2000; Etienne *et al*., 2004; Abruzzo *et al*., 2005; Beckman *et*

Corresponding Author: Eva Obermayr, Department of Obstetrics and Gynecology, Medical University of Vienna, Währinger Gürtel 18-20 5Q, 1090 Wien, Austria Tel: +43-1-40400-7827 Fax: +43-1-40400-7832

al., 2004; Dallas *et al*., 2005; Walker *et al*., 2006; Wang *et al*., 2006; Canales *et al*., 2006). Only a few of these studies address potential factors influencing the correlation (Canales *et al*., 2006; Morey *et al*., 2006; Barbacioru *et al*., 2006).

Microarray and RT-qPCR techniques differ in many respects, including in the method for reverse transcription, in their reaction dynamics and in their dynamic range. Additionally, normalization methods applied vary considerably between the two methods, due to the up to 100-fold higher number of genes measured in a microarray experiment than in RT-qPCR. Systematic errors in RT-qPCR data are compensated for by the use of reference genes, which are assumed to show almost constant expression across samples and experimental conditionsc (Schmittgen and Livak, 2008; Vandesompele *et al*., 2002), In contrast, microarray normalization methods (Do and Choi, 2006) take into account all or a large subset of genes on the array to calculate correction factors. Additionally, the region targeted by the microarray probe and by the RT-qPCR primers, respectively, influences the correlation. Microarray probes are in general designed to hybridize to all splice variants of a gene and lay therefore completely within an exon. RT-qPCR primers, in contrast, are designed to span exons to avoid amplification of genomic DNA or unspliced mRNA and these primers target only a subset of the splice variants of a gene, depending on which exon boundary is spanned. Sequence-based validation of targeted transcripts has been applied for microarray platform comparisons (Ji *et al*., 2006; Carter *et al*., 2005; Mecham *et al*., 2004), but not yet applied for comparisons of microarray probes and RT-qPCR primers.

Validation of the microarray results can be done using three distinct approaches:

- Calculating the correlation for each gene individually across all samples (Dallas *et al*., 2005)
- calculating the correlation for all genes within a sample
- calculating the correlation of expression ratios of all genes within a sample using an arbitrary reference sample (Wang *et al*., 2006)

For this reason, we compared for the first time the agreement of both measurements using the three approaches mentioned above and investigated the influence of the mRNA region targeted by probes and primers, normalization and efficiency correction.

The presented comparison of mRNA expression measurements from DNA microarray and RT-qPCR experiments was part of a project which aimed to detect circulating tumor cells in the peripheral blood of patients suffering from gynecological cancers (Obermayr *et al*., 2010). To identify genes differentially expressed in tumor cells compared to peripheral blood cells, microarray analysis of 38 tumor cell lines and of PBMC from 12 healthy female volunteers was performed. The resulting gene expression levels obtained by the microarray analysis were validated with RT-qPCR for a subset of 381 genes.

MATERIALS AND METHODS

RNA samples: Total RNA was extracted from 38 cancer cell lines and from 12 PBMC samples taken from healthy female donors using the Total RNA Isolation Mini Kit (Agilent Technologies, Waldbronn, Germany). The quality and integrity of the total RNA was assessed on the Agilent 2100 Bioanalyzer and the same samples were divided into individual aliquots for the gene expression analysis on the microarray platform and for the TaqManbased RT-qPCR analysis. All RNAs used in the present study were of high quality and un-degraded (Supplemental Data Table 1). All peri-pheral blood was collected with the patients' written consent. The study was approved by the Ethics Committee of the Medical University of Vienna, Austria (Obermayr *et al*., 2010).

Gene expression array analysis: Gene expression profiles from the tumor cell lines and from the PBMC samples were generated using the AB Human Genome Survey Microarray Hs v1. Kits and reagents were used according to the manufacturer's protocols. Image acquisition and analysis were performed using the AB 1700 Chemiluminescent Analyzer Software (version 1.0.0.3). Signals from the autogridded images were background corrected and normalized first by feature, then by spatial effects for each slide and finally by global normalization across slides. The Assay Normalized Signal (ANS) and the Signal to Noise ratio of the measurements (S/N) were used during further analysis. No additional normalization was applied. Filtering data with a flag of greater than 5000 indicating a low quality spot and with a $S/N \leq 3$ (Wang *et al*., 2006) excluded 1290 measurements leaving 6075 for the comparisons to the RT-qPCR data. Finally, we identified genes with differential expression levels in each group of tumor cell lines and in part of the tumour cell lines, respectively, compared to the healthy control group using the maxT test on log transformed expression values from the R (RDCT, 2010) "multtest" package (Ge *et al*., 2003) and the 50% one-sided trimmed maxT-test (Gleiss *et al*., 2011).

Thus, 377 genes were selected for RT-qPCR-based validation and further investigation.

TaqMan gene expression assay based RT-qPCR: Microarray data was validated in 15 cancer cell lines using the AB TaqMan Low Density Array (TLDA) format 384, which allows the analysis of 380 gene targets in single reactions and of one mandatory endogenous control gene (GAPDH) in a quadruplicate reaction. Matching TaqMan Gene Expression Assays were selected according to a mapping of microarray probe IDs to assay IDs provided by AB. Additionally, three TaqMan Endogenous Controls (B2M, TBP and PGK1) were analyzed. Template cDNA was generated using M-MLV Reverse Transcriptase, RNase H Minus (Promega, Madison WI, USA) and random hexamers as primers. The Low Density Arrays were loaded with the sample specific mix containing the cDNA and TaqMan Universal PCR Master Mix, No AmpErase UNG. The RT-qPCR was run on the AB 7900HT Fast Real-time PCR System using default conditions (1 cycle of 2 min., 50°C; 1 cycle of 10 min. 95°C; 50 cycles of 15 s, 95°C; 1 min., 60°C). Raw data were analyzed with the AB7900 Sequence Detection Software version 2.2.2 using automatic baseline correction and manual cycle threshold setting.

Calculation of RT-qPCR efficiencies: The amplification efficiencies of 95 differentially expressed genes were assessed using the TLDA format 96A, which allows the amplification of 95 gene targets and of one mandatory endogenous control gene (GAPDH) in duplicate reactions. Equal cDNA amounts from eight cancer cell lines were pooled and fourfold serially diluted. Each template dilution was amplified in two TLDAs to compensate for experimental variations. Amplification and data analysis were performed as described above. The efficiencies were estimated both from the slope of log input template amount versus C_q-value ($E_g = 10^{(-1/k=slop)}$-1) and directly from the raw fluorescence intensities as proposed by (Zhao and Fernald, 2005). The resulting efficiencies were averaged across samples, assuming inhibition and amplification as being very similar in the cell lines.

Filtering and normalization of RT-qPCR data: C_q-values ≥ 35 were considered unreliable and filtered as described in Wang *et al*. (2006). Of the 5760 RT-qPCR measurements, 414 were below the detection limit and an additional 87 were removed according to the C_q-value quality criteria. The remaining 5214 Cq-values were converted to relative quantities (RQs) on a linear scale as follows: $RQ_g = (1+E_g)^{(Cq_{max}-Cqg)}$ where g denotes the gene and Cq_{max} is the maximum Cq-value over all 15 TLDAs. E_g is the efficiency of the PCR reaction for gene g ranging from 0 (no amplification) to 1 (perfect amplification).

Genes suitable for normalization of the RT-qPCR data were selected using NormFinder (Andersen *et al*., 2004) based on the ANS. The 10 most stable genes across the 15 cell lines were verified with geNorm (Vandesompele *et al*., 2002). This list was further reduced to three genes (CENPA, CDCA5 and CRYZL1) which were detected in all 15 cell lines by RT-qPCR and had a validated probe-assay pair (Supplemental Data Table 2).

Normalization to the geometric mean of these reference genes was performed as suggested by Vandesompele *et al*. (2002).

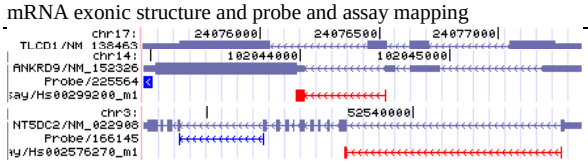
First, Cq-values of the reference gene h and the assay a were individually normalized across assays using the equation $NQ_{ha} = (1+E_h)^{(Cq_{min}-Cq_{ha})}$, where Cq_{hmin} denotes the minimum Cq-value of the reference gene h across all assays a. The normalization factor for an assay a (NF_a) is the geometric mean of all n NQ_{ha} of assay a: $NF_a = \sqrt[n]{\prod_{h=1}^{n} NQ_{ha}}$ The normalized relative quantity (Norm_RQ) of a gene g in assay a is finally calculated by $Norm\_RQ_{ga} = RQga/NF_a$ In the following, the abbreviation RQ is used for relative quantities calculated with a constant efficiency of 1 and ERQ for relative quantities derived with a gene specific efficiency. Additionally, normalization was performed with every gene available on the TLDA, to check whether there is any gene, which improves the correlation of microarray and RT-qPCR data.

Sequence-based mappings of microarray probes to RT-qPCR assays: Sequences of the 60-mer oligonucleotide microarray probes were retrieved from the Panther homepage SRI International 2011. RT-qPCR amplicon sequences were assembled by retrieving assay information consisting of accession number of targeted transcript, amplicon start position and amplicon length from the AB product homepage Applied Biosystems 2007. The corresponding mRNA sequence was retrieved either from GenBank (Benson *et al*., 2007) or the Panther homepage and the amplicon sequence was extracted based on the amplicon start position and length. Both sequence lists were subjected to a nucleotide BLAST (Benson *et al*., 2007) with high similarity against the Homo sapiens RefSeq database (Pruitt *et al*., 2005) Release 14 using the Comparative Transcriptomics Framework (Sturn, 2005). Only complete matches on the sense strand were accepted; hits to experimental sequences (XM_* and XR_*) were removed. An extended and annotated mapping between microarray probe and corresponding TaqMan assay was created by comparing the set of targeted transcripts. For identical transcript sets the respective probe-assay pair was added to the extended mapping list.

Table 1: Correlation coefficients (R) for single genes across samples for the different mapping sets. The P-value has been assessed by drawing random samples from the correlation coefficients of the initial mapping (n = 20000) and comparing the resulting distribution difference to the one observed for the set investigated

Probe-assay pair mapping set	N Probe-assay pairs	Min R	Median R	Mean R	Max R	Average cumulative distribution difference	P-value for difference
Initial mapping	494	0.6109	0.8657	0.7625	0.9999	NA	NA
Invalid pairs	93	0.5621	0.7809	0.6108	0.9903	-9.16	< 5.0e-05
Validated pairs	397	0.6109	0.8759	0.7845	0.9999	1.33	7.5e-04
Only probes closest to 3' end	345	0.3576	0.8792	0.7936	0.9999	1.88	3.5e-04
Pairs targeting only a single transcript	264	0.1486	0.8833	0.8149	0.9999	3.16	< 5.0e-05

Table 2: Exon/intron structure and probe/assay location of three mRNAs showing no correlation. RefSeq mRNA (slate blue), microarray probe (dark blue) and assay amplicon (red) are shown. Solid bars indicate exons; thin lines intronic sequences. mRNA 3' end is on the left side, exons are numbered starting from the right. (Images generated by the UCSD Genome Browser (Kuhn *et al*., 2007)

mRNA exonic structure and probe and assay mapping	Distance to 3' end (bp)	R
	618	0.05
	2	0.12
	591	0.15

Comparison of microarray and RT-qPCR data: Data were compared by calculating the Pearson's correlation coefficient R (Pearson, 1896) of the probe-assay pairs for (i) single genes across samples, (ii) all genes within a sample and (iii) the expression ratios of all genes in a sample using another sample as the reference. In some instances the comparison was done based on the Spearman's rank correlation coefficient (Spearman, 1904) or Kendall's Tau-b (Kendall, 1938). Significance of correlation differences observed between the probe-assay mappings were determined by a one-sided Wilcoxon's rank test. For the comparison of single genes across samples, the cumulative distribution of the correlation coefficients was used. The average difference between the distribution of the initial mapping and the derived mappings was calculated. The significance of the difference was assessed by a permutation test as follows: 20000 random samples from the correlation coefficients of the initial mapping were drawn (without replacement, with the number correlation coefficients matching the size of the mapping set investigated) and the average distribution difference of this sample was calculated. The p-value is the proportion of samples with a higher difference than the original set.

RESULTS

Sequence-based mappings of microarray probe to RT-qPCR assay: The 381 TaqMan assays corresponded to 491 unique microarray probes based on the annotation of the Human Genome Survey Array v1 supplied by AB. Three assays mapped to multiple probes yielding an initial mapping of 494 probe-assay pairs. Ninety-three of these were excluded by the sequence based mapping validation due to inconsistent RefSeq transcript sets targeted by the probe and assay. For seven pairs the BLAST search did not yield any results for both the probe and the amplicon sequence. Three new pairs were added during this process resulting in 397 validated probe-assay pairs (Supplemental Data Table 3). Two additional probe-assay mapping sets were defined:

- set (iii), where, for transcripts targeted by multiple probes, only the probe closest to the 3' end of the transcript was retained and
- set (iv), where all probe-assay pairs from set (iii) were removed which target multiple transcripts

Set (iii) was defined to compensate for the bias introduced by the oligo (dT) primed reverse transcription of mRNA for the microarray hybridizations. Sequences closer to the 3' end of the mRNA are more likely transcribed into cDNA, because the probability that the mRNA transcription terminates prematurely increases with the distance to the 3' end Applied Biosystems 2004. This bias is not present if random hexamer primers are used during reverse transcription for the RT-qPCR as transcription starts at random positions on the mRNA. With set (iv) differences in the detection of multiple transcripts by the two technologies were avoided (e.g., due to mRNA secondary structure). Therefore, four probe-assay sets were used in the subsequent comparison:

- The initial mapping as supplied by the manufacturer (n = 494)
- The mapping containing validated probe-assay pairs and new ones not present in the initial mapping (n = 397)

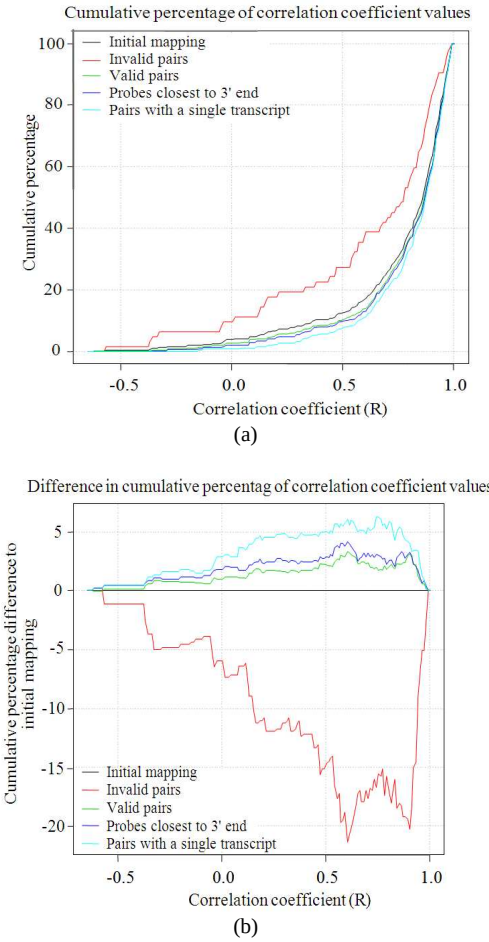


Fig. 1: Cumulative distribution of Pearson's correlation coefficients for single gene comparisons. (A) Absolute distribution for five probe-assay sets. (B) Distribution deviation relative to the initial mapping from the manufacturer (black). The largest positive difference can be observed for the set targeting a single transcript only (light blue). The only negative difference results from the pairs which were excluded through the validation process (red)

- The previous set, reduced to a single probe per transcript, where the probe with the least distance to the 3' end of a transcript was retained (n = 365)
- The previous set containing only probe-assay pairs targeting a single transcript (n = 264)

Correlation between the filtered microarray and RT-qPCR expression data was assessed using the three approaches mentioned previously: (i) across samples, (ii) within a sample and (iii) expression ratios within samples. Each approach was applied to the four probe-assay mappings described above.

Comparison of single genes across samples: The correlation coefficients (R) of the genes range from a minimum -0.61 and -0.14 (in the initial probe-assay set and in the set with probe-assay pairs targeting a single transcript only, respectively) to a maximum of almost 1.0 in all sets (Table 1). To assess the quality of the agreement of a certain set, the cumulative distributions of the calculated correlation coefficients in a probe-assay set were compared. The higher the number of correlation values close to 1, the better the agreement between the two technologies. All three subsets derived from the initial mapping showed an increased fraction of higher correlation coefficients. For the correlation value distribution of the excluded probe-assay pairs, a shift toward lower correlation coefficients could be observed (Fig. 1). The average distribution difference relative to the initial mapping was between -9.16 (for the excluded pairs) to 3.16 (for the pairs targeting a single transcript only). All differences were statistically significant ($p<0.001$).Using the normalized relative RT-qPCR quantities with the three selected reference genes (Norm_RQ) in the comparison shifted the correlation coefficient distribution considerably toward smaller values (data not shown).

The same could be observed for all datasets normalized to a single reference gene. Although the correlation was already quite high, there were still 18 probe-assay pairs with a correlation below 0.5. Three of these pairs targeting transcripts of the genes TLCD1, ANKRD9 and NT5DC2 were further investigated.

The distance to the 3' end of the transcript, the exon-intron structure and the location of the probes and assays relative to the mRNA exonic structure are shown in Table 2. A scatter plot of the expression for the transcript of gene TLCD1 (Fig. 2A) reveals an outlying measurement on the microarray for the cell line BT-549, indicating the existence of a TLDC1 splice variant without exon 1 in this cell line. Excluding this outlying measurement increased the correlation to 0.90. The measurements for the two other genes showed low expression in general, without obvious outliers (Fig. 2B for NT5DC2). The differing measurements are most likely explained by splice variants, which are not detected equally by primers and probes, as the exons targeted differ significantly.

Comparison of all genes within a sample: For this comparison the Pearson's correlation of $\log_2$ (ANS) vs. $\log_2$ (RQ) of all genes within a sample was calculated. The correlation was low ranging between 0.44 and 0.59 (for the initial mapping) and between 0.47 and 0.63 (for the mapping including pairs targeting a single transcript only).

Table 3: Summary of the pearson's correlation coefficients for the comparison of all genes within a sample for all mapping sets. For each mapping 15 correlations were calculated corresponding to the 15 different samples. P-values are based on a wilcoxon's signed rank test of the correlations of a mapping compared to these of the initial mapping

Probe-assay pair mapping set	Probe-assay pairs	Min R	Median R	Mean R	Max R	Average correlation difference	P-value of difference
Initial mapping	494	0.44	0.50	0.51	0.59	0.000	1.0000
Validated pairs	397	0.43	0.50	0.51	0.59	-0.007	0.0603
Only probes closest to 3' end	345	0.43	0.50	0.50	0.59	0.003	0.6807
Pairs targeting a single transcript only	264	0.47	0.56	0.55	0.63	-0.047	0.0006

Table 4: Summary of the correlation coefficients using the averaged values of all samples as a reference for all mapping subset. The P-value is calculated using a one-sided Wilcoxon's signed rank test between the results of the initial mapping and the results of the other subset (n = 15)

Probe-assay pair mapping set	Probe-assay pairs	Min R	Median R	Mean R	Max R	P-value
Initial mapping	494	0.63	0.79	0.78	0.88	1.0000
Validated pairs	397	0.70	0.83	0.81	0.89	0.00623
Only probes closest to 3' end	345	0.68	0.83	0.81	0.89	0.00269
Pairs targeting a single transcript only	264	0.77	0.84	0.83	0.90	0.00058

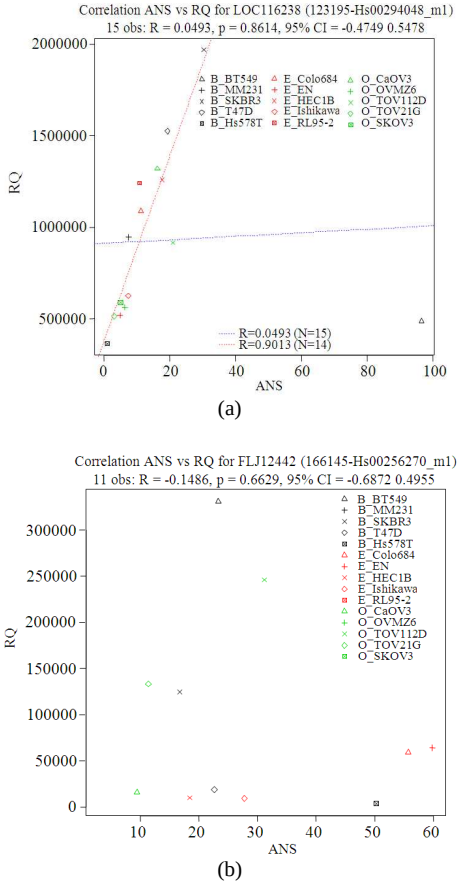


Fig. 2: Scatterplot of microarray (ANS) vs. RT-qPCR (RQ) of transcripts with low correlation. (top) LOC116238 (TLDC1) shows an outlying microarray measurement for the breast cancer cell line BT-549. Correlation without this measurement increases to 0.90. (bottom) FLJ1244 (NT5DC2) shows low expression (especially for RT-qPCR), but no striking outliers

A significant difference in the correlation coefficients was observed only for the single transcript mapping (Table 3 and Supplemental Data Fig. 1). In a second analysis, the $\log_2$ of the Efficiency corrected Relative Quantities (ERQ) were compared to $\log_2$(ANS). Correlation coefficients were consistently lower compared to the uncorrected values (data not shown). Using the RQs normalized to the reference genes did not change the results, because the individual assays are corrected by a constant factor only.

Comparison of the expression ratios of all genes of a sample: For all samples, the Ratios for the ANS (RANS) and the Ratio for the RQ (RRQ) were calculated using the averaged values of all samples as the reference and the correlation of $\log_2$ (RANS) and $\log_2$ (RRQ) was determined. For the initial mapping, values for R ranged between 0.63 and 0.88, for all other mapping subsets the minimum and maximum R increased up to 0.77 and 0.90, respectively (Table 4). The shift towards higher correlation coefficients compared to the initial mapping was statistically significant for all subsets (one-sided Wilcoxon's signed rank test, $p<0.01$).Using the RQs normalized with the internal reference genes did not change the results, because the individual assays are corrected only by a constant factor. Efficiency corrected Relative Quantities (ERQ) produced the same results as the RQ values, because the different PCR efficiencies cancel each other out when the ratios are calculated (data not shown).

DISCUSSION

Results from gene expression profiling with DNA microarrays are often validated by RT-qPCR. The level of agreement reported varies significantly between as well as within studies (Etienne *et al*., 2004; Abruzzo *et al*., 2005; Beckman *et al*., 2004; Dallas *et al*., 2005;

Walker *et al*., 2006; Wang *et al*., 2006; Canales *et al*., 2006). Here we have studied several parameters influencing the agreement and have shown that the correlation of the two technologies is significantly increased when microarray probes and RT-qPCR primers target the same set of transcripts. To this end, we have used a rigorous validation approach to exclude probes-assay pairs with a discordant set of targeted transcripts. The identification of valid probe-assay pair is based (i) on alignment of the probe and amplicon sequence against the human transcripts in RefSeq and (ii) on probe distance to 3' end and (iii) on the number of transcripts targeted. With the resulting four probe-assay pair sets, measurements of microarray and RT-qPCR experiments were compared for individual genes across samples, all genes within a sample and for expression ratios within a sample. In all three comparisons, sets with validated and bias avoiding probe-assay pairs showed a significantly higher correlation than the initial set derived from the microarray annotation supplied by the manufacturer.

Specifically, the correlation of the technologies for single genes across samples was greater than 0.70 for 80% of the genes for probe assay pairs targeting single transcripts. The correlation of the measurements for all genes of a sample was low, with a maximum of 0.63. Nevertheless, it was possible to observe a significant positive effect of the rigorous validation of the probe-assay pairs. Platform differences (especially different RT-qPCR efficiencies) have a pronounced influence on the results in these types of comparisons.

When assessing the correlation of the expression ratios of the genes in a sample with the average values of all samples as the reference, the median R was between 0.79 and 0.84. By calculating ratios to assess the correlation, differences in technologies (like RT-qPCR efficiencies or microarray hybridization dynamics) cancel out largely. The results achieved here are at the same level as those reported from Wang *et al*. (2006), where the same microarray and RT-qPCR platforms have been utilized. In Wang's study, the comparison was performed with robust linear regression fitting using bisquare weights, which resulted in slightly higher correlation coefficients due to down-weighing of outliers.

Both microarray and PCR technologies are subject to handling inaccuracies (pipetting, reaction conditions), which have to be compensated for by normalization methods. However, correlation decreased significantly when normalized data was used in the calculations instead of the raw data. This applied both to normalized data in the genes across samples correlation as well as efficiency corrected data in the within sample correlation.

Although the use of the correlation coefficient as a measure of agreement between the two technologies may not be optimal for low expressing genes and genes with a low variance in expression across samples (Abruzzo *et al*., 2005), it has been applied successfully in this study. Other means to assess the agreement like the concordance correlation coefficient (Lin, 2000), which was used by Miron *et al*. (2006) or measurement of agreement by direct comparison of expression values as suggested by Bland and Altman (2010) are not applicable as they yield poor correlation or agreement in presence of offsets and scaling factors between the measurements. The scale especially differs considerably, because the dynamic range is 4 orders of magnitude for the AB 1700 platform (Stefano *et al*., 2005) and up to 8 orders of magnitude for the TLDAs (Canales *et al*., 2006; Yang *et al*., 2004).

CONCLUSION

For a reliable validation of microarray measurements by RT-qPCR, it is of utmost importance that microarray probes and RT-qPCR primers target both the same exon of the mRNA. To avoid possible bias introduced by the secondary structure of the cDNA, the same region of the exon should be targeted. Special care has also to be directed to the selection of the internal references and normalization methods, because they can influence the results significantly.

ACKNOWLEDGEMENT

This study was supported by the GEN-AU projects "Cancer Transcriptomics" and "Bioinformatics Integration Network" (BIN) of the Austrian Federal Ministry of Science and Research. We are particularly grateful to Fatima Sanchez-Cabo (Genomics Unit, Centro Nacional de Investigaciones Cardiovasculares CNIC, Madrid, E) and Georg Heinze (Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, A) for valuable discussions. The authors thank Ravi Tharakan (Bayview Proteomics Center, Johns Hopkins University, Baltimore, MD, USA) for critically reading the manuscript.

REFERENCES

Abruzzo, L.V., K.Y. Lee, A. Fuller, A. Silverman and M.J. Keating *et al*., 2005. Validation of oligonucleotide microarray data using microfluidic low-density arrays: A new statistical method to normalize real-time RT-PCR data. BioTechniques, 38: 785-792.

Andersen, C.L., J.L. Jensen and T.F. Orntoft, 2004. Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res., 64: 5245-5250. DOI: 10.1158/0008-5472.CAN-04-0496

Barbacioru, C.C., Y. Wang, R.D. Canales, Y.A. Sun and D.N. Keys *et al*., 2006. Effect of various normalization methods on Applied Biosystems expression array system data. BMC Bioinformatics, 7: 533. DOI: 10.1186/1471-2105-7-533

Beckman, K.B., K.Y. Lee, T. Golden and S. Melov. 2004. Gene expression profiling in mitochondrial disease: Assessment of microarray accuracy by high-throughput Q-PCR. Mitochondrion, 4: 453-470. DOI: 10.1016/j.mito.2004.07.029

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell and D.L. Wheeler, 2007. GenBank. Nucleic Acids Res., 35: D21-D25. DOI: 10.1093/nar/gkl986

Bland, J.M. and D.G. Altman, 2010. Statistical methods for assessing agreement between two methods of clinical measurement. Int. J. Nurs. Stud., 47: 931-936. DOI: 10.1016/j.ijnurstu.2009.10.001

Canales, R.D., Y. Luo, J.C. Willey, B. Austermiller and C.C. Barbacioru *et al*., 2006. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat. Biotechnol., 24: 1115-1122. DOI: 10.1038/nbt1236

Carter, S.L., A.C. Eklund, B.H. Mecham, I.S. Kohane and Z. Szallasi, 2005. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. BMC Bioinformatics, 6: 107. DOI: 10.1186/1471-2105-6-107

Dallas, P.B., N.G. Gottardo, M.J. Firth, A.H. Beesley and K. Hoffmann *et al*., 2005. GGene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR – how well do they correlate? BMC Genomics, 6: 59. DOI: 10.1186/1471-2164-6-59

Do, J.H. and D.K. Choi, 2006. Normalization of microarray data: Single-labeled and dual-labeled arrays. Mol. Cells, 22: 254-261. PMID: 17202852

Etienne, W., M.H. Meyer, J. Peppers and R.A.M. Jr, 2004. Comparison of mRNA gene expression by RT-PCR and DNA microarray. Biotechniques, 36: 618-626. DOI: 10.2144/3604A0618

Ge, Y., S. Dudoit and T.P. Speed, 2003. Resampling-based multiple testing for microarray data analysis. Test, 12: 1-77. DOI: 10.1007/BF02595811

Gleiss, A., F. Sanchez-Cabo, P. Perco, D. Tong and G. Heinze, 2011. Adaptive trimmed t-statistics for identifying predominantly high expression in a microarray experiment. Stat. Med., 30: 52-61. DOI: 10.1002/sim.4093

Ji, Y., K. Coombes, J. Zhang, S. Wen and J. Mitchell *et al*., 2006. RefSeq refinements of UniGene-based gene matching improve the correlation of expression measurements between two microarray platforms. Appl. Bioinformatics, 5: 89-98. PMID: 16722773

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika, 30: 81-93. DOI: 10.1093/biomet/30.1-2.81

Kuhn, R.M., D. Karolchik, A.S. Zweig, H. Trumbower and D.J. Thomas *et al*., 2007. The UCSC genome browser database: Update 2007. Nucleic Acids Res., 35: D668-D673. DOI: 10.1093/nar/gkl928

Lin, L.I.K., 2000. Correction: A note on the concordance correlation coefficient. Biometrics, 56: 324-325. DOI: 10.1111/j.0006-341X.2000.-00324.x

Mecham, B.H., G.T. Klus, J. Strovel, M. Augustus and D. Byrne *et al*., 2004. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucl. Acids Res., 32: e74-e74. DOI: 10.1093/nar/gnh071

Miron, M., O.Z. Woody, A. Marcil, C. Murie and R. Sladek *et al*., 2006. A methodology for global validation of microarray experiments. BMC Bioinformatics, 7: 333. DOI: 10.1186/1471-2105-7-333

Morey, J.S., J.C. Ryan and F.M.V. Dolah, 2006. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. Biol. Proced. Online, 8: 175-193. DOI: 10.1251/bpo126

Obermayr, E., F. Sanchez-Cabo, M.K.M. Tea, C.F. Singer and M. Krainer *et al*., 2010. Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients. BMC Cancer, 10: 666. DOI: 10.1186/1471-2407-10-666

Pearson, K., 1896. Mathematical contributions to the theory of evolution. III. Regression, Heredity and Panmixia. Phil. Trans. R. Soc. Lond., 187: 253-318. DOI: 10.1098/rsta.1896.0007

Perreard, L., C. Fan, J.F. Quackenbush, M. Mullins and N.P. Gauthier *et al*., 2006. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res, 8: R23. DOI: 10.1186/bcr1399

Pruitt, K.D., T. Tatusova and D.R. Maglott. 2005. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res., 33: D501-D504. DOI: 10.1093/nar/gki025

Rajeevan, M.S., D.G. Ranamukhaarachchi, S.D. Vernon and E.R. Unger. 2001. Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. Methods, 25: 443-451. DOI: 10.1006/meth.2001.1266

RDCT, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Rockett, J.C. and G.M. Hellmann, 2004. Confirming microarray data--is it really necessary? Genomics, 83: 541-549. DOI: 10.1016/j.ygeno.2003.09.017

Schmittgen, T.D. and K.J. Livak, 2008. Analyzing real-time PCR data by the comparative C(T) method. Nat. Protoc., 3: 1101-1108. DOI: 10.1038/nprot.2008.73

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol., 15: 72-101. DOI: 10.2307/1412159

Stefano, G.B., J.D. Burrill, S. Labur, J. Blake and P. Cadet, 2005. Regulation of various genes in human leukocytes acutely exposed to morphine: Expression microarray analysis. Med. Sci. Monit., 11: MS35-MS42. PMID: 15874898

Sturn, A., 2005. Comparative analysis of human and mouse transcriptomes [PhD Thesis]. Graz, Austria: Institute for Genomics and Bioinformatics, Graz University of Technology.

Vandesompele, J., K.D. Preter, F. Pattyn, B. Poppe and N.V. Roy *et al*., 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol., 3. DOI: 10.1186/gb-2002-3-7-research0034

Veer, L.J.V., H. Dai, M.J.V.D. Vijver, Y.D. He and A.A.M. Hart *et al*., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415: 530-536. DOI: 10.1038/415530a

Walker, S.J., Y. Wang, K.A. Grant, F. Chan and G.M. Hellmann, 2006. Long versus short oligonucleotide microarrays for the study of gene expression in nonhuman primates. J. Neurosci. Methods, 152: 179-189. DOI: 10.1016/j.jneumeth.2005.09.007

Wang, Y., C. Barbacioru, F. Hyland, W. Xiao and K.L. Hunkapiller *et al*., 2006. Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. BMC Genomics, 7: 59. DOI: 10.1186/1471-2164-7-59

Yang, D.K., C.H. Kweon, B.H. Kim, S.I. Lim and S.H. Kim *et al*., 2004. TaqMan reverse transcription polymerase chain reaction for the detection of Japanese encephalitis virus. J. Vet. Sci, 5: 345-351. PMID: 15613819

Zhang, Z., S. Schwartz, L. Wagner and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol., 7: 203-214. DOI: 10.1089/10665270050081478

Zhao, S. and R.D. Fernald, 2005. Comprehensive algorithm for quantitative real-time polymerase chain reaction. J. Comput. Biol., 12: 1047-1064. DOI: 10.1089/cmb.2005.12.1047