BMC Genomics

Methodology article

Open Access

Universal Reference RNA as a standard for microarray experiments

Natalia Novoradovskaya*[1], Michael L Whitfield[2,5], Lee S Basehore[1], Alexey Novoradovsky[1], Robert Pesich[2], Jerry Usary[3], Mehmet Karaca[3], Winston K Wong[1], Olga Aprelikova[4], Michael Fero[2], Charles M Perou[3], David Botstein[2,6] and Jeff Braman[1]

Address: [1]Stratagene, 11011 N. Torrey Pines Road, La Jolla, CA 92037, USA, [2]Department of Genetics, Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA, [3]Departments of Genetics and Department of Pathology and Laboratory Science, Lineberger Comprehensive Cancer Center, CB# 7295, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, [4]National Cancer Institute, NIH, Bethesda, MD 20892, USA, [5]Department of Genetics, Dartmouth Medical School, 7400 Remsen, Hanover, NH 03755, USA and [6]Lewis-Sigler Institute for Integrative Genomics 140 Carl Icahn Laboratory, Princeton University Washington Road, Princeton, NJ 08544, USA

Email: Natalia Novoradovskaya* - nnovo@stratagene.com; Michael L Whitfield - Michael.Whitfield@Dartmouth.edu; Lee S Basehore - Scott.Basehore@stratagene.com; Alexey Novoradovsky - anovo@stratagene.com; Robert Pesich - rspesich@genome.stanford.edu; Jerry Usary - usary@unc.edu; Mehmet Karaca - mehmet_karaca@med.unc.edu; Winston K Wong - Winston.Wong@stratagene.com; Olga Aprelikova - apreliko@mail.nih.gov; Michael Fero - mfero@genome.stanford.edu; Charles M Perou - cperou@med.unc.edu; David Botstein - botstein@princeton.edu; Jeff Braman - Jeff.Braman@stratagene.com

* Corresponding author

Abstract

Background: Obtaining reliable and reproducible two-color microarray gene expression data is critically important for understanding the biological significance of perturbations made on a cellular system. Microarray design, RNA preparation and labeling, hybridization conditions and data acquisition and analysis are variables difficult to simultaneously control. A useful tool for monitoring and controlling intra- and inter-experimental variation is Universal Reference RNA (URR), developed with the goal of providing hybridization signal at each microarray probe location (spot). Measuring signal at each spot as the ratio of experimental RNA to reference RNA targets, rather than relying on absolute signal intensity, decreases variability by normalizing signal output in any two-color hybridization experiment.

Results: Human, mouse and rat URR (UHRR, UMRR and URRR, respectively) were prepared from pools of RNA derived from individual cell lines representing different tissues. A variety of microarrays were used to determine percentage of spots hybridizing with URR and producing signal above a user defined threshold (microarray coverage). Microarray coverage was consistently greater than 80% for all arrays tested. We confirmed that individual cell lines contribute their own unique set of genes to URR, arguing for a pool of RNA from several cell lines as a better configuration for URR as opposed to a single cell line source for URR. Microarray coverage comparing two separately prepared batches each of UHRR, UMRR and URRR were highly correlated (Pearson's correlation coefficients of 0.97).

Conclusion: Results of this study demonstrate that large quantities of pooled RNA from individual cell lines are reproducibly prepared and possess diverse gene representation. This type of reference provides a standard for reducing variation in microarray experiments and allows more reliable comparison of gene expression data within and between experiments and laboratories.

Background

Techniques used for microarray experiments are similar in principle and practice to methods developed for Southern and Northern blots [1,2]. However, in contrast to those methods, cDNA microarray experiments employ nucleic acid probes of known nucleotide sequence attached to a solid support such as a glass microscope slide [3]. Hybridization of fluorescently labeled cDNA, reverse transcribed from an RNA sample, to a microarray measures the relative level of mRNA in the sample. Due to variability in microarray spot geometry, quantity of DNA deposited at each spot and hybridization efficiency, absolute fluorescence intensity cannot be used as a reliable measure of RNA level. However, if two RNA samples are differentially labeled and co-hybridized to spots on the same microarray, the ratio of their signal intensities accurately reports the relative quantity of RNA targets in both samples. Two basic types of 2-color microarray experimental designs exist, which are the "loop" and "reference" designs [4]. Advantages and disadvantages of each design have been discussed in several publications [5,6]. In the "loop" design, samples are compared to one to another in circular or multiple-pairwise fashion. This design might be useful when small numbers of samples are compared, but it becomes inefficient for more than 10 samples. In the "reference" design each sample is compared to a common RNA reference sample, serving as a common denominator between different microarray hybridizations [7]. This experimental design has been widely used to study diversity in cell lines [8,9] and patterns of gene expression allowing classification of breast and lung carcinoma samples [10,11]. Reference RNA can also be used for time-course experiments in which the response of cells to drugs or other perturbations to the biological system is monitored. In addition, comparing microarray data sets produced in different laboratories will be more reliable by employing the use of reproducible common reference RNA and can be also used to normalize data from one set of Affymetrix experiments to another [12].

An ideal universal reference RNA for normalizing gene expression data should provide positive hybridization signal at each probe element on the microarray and should be achievable by pooling RNA from a mixture of cell lines [7]. The properties required of a reference RNA sample and the number of pooled RNAs were examined by Yang et al. [13], who showed that pools of cell lines can be an efficient reference sample. The ideal reference should also be available in large quantities, sufficient to satisfy long-term requirements of many researchers, and reproducible such that different batches are indistinguishable from one another. Many researchers prepare their own reference RNA from a single cell line or by blending RNA derived from several cell lines or tissues [8-11,14,15]. Alternatively, amplified RNA from multiple cell lines has been employed as reference RNA [16]. Several other materials have been used as a reference, such as a mixture of cDNA products spotted onto arrays [17], a mix of labeled oligos complementary to every microarray probe [18] or genomic DNA [19].

Although these approaches serve the immediate needs of each research group, when the reference is exhausted, the problem of producing an equivalent large batch is significant. In addition, it is difficult to compare data sets between different experiments, and between different laboratories. Large batches of reproducible reference RNA solve these problems.

We describe here the development and performance of large, reproducible quantities of human, mouse and rat Universal Reference RNA (URR). These materials provide high microarray coverage and allow normalization of data in two-color microarray experiments.

Results

Development of large quantities of URR

Sufficient total RNA was isolated from cell lines, representing the tissues described in the Materials and Methods section, to produce multi-gram quantities of UHRR, UMRR and URRR. Individual cell line and pooled total RNA was devoid of RNAse activity, was intact as evidenced by discreet 28S and 18S ribosomal RNA bands revealed by agarose gel electrophoresis and possessed A260:A280 ratios of 1.8 or greater (data not shown). UHRR, UMRR and URRR were consistently tested on a variety of microarrays to validate both high percent spot coverage and batch-to-batch reproducibility (data presented in following sections).

Use of Universal Reference RNA

We illustrate one intended use of URR for two-color microarray experiments in the following example. Gene expression in ten human cell lines was compared using UHRR as a common reference. Each individual cell line RNA was compared to UHRR directly on the same microarray. For simplicity the KIAA0923 gene expression is shown only for two samples, brain and testis (Figure 1), to illustrate an inefficiency of the intensity values and advantages of using Cy5/Cy3 ratios. When a common reference sample is used, the multiple microarray hybridizations can be compared. Total RNA from brain and testis human cell lines were reverse-transcribed into cDNA and labeled with Cy5. UHRR was reverse-transcribed into cDNA and labeled with Cy3. Each experimental Cy5-labeled cDNA was co-hybridized with Cy3-labeled cDNA from UHRR onto 43,000-spot cDNA microarrays (Stanford University). The microarrays were processed and data collected and analyzed as described in the Material and Methods. Ratios of Cy5/Cy3 intensities were compared to
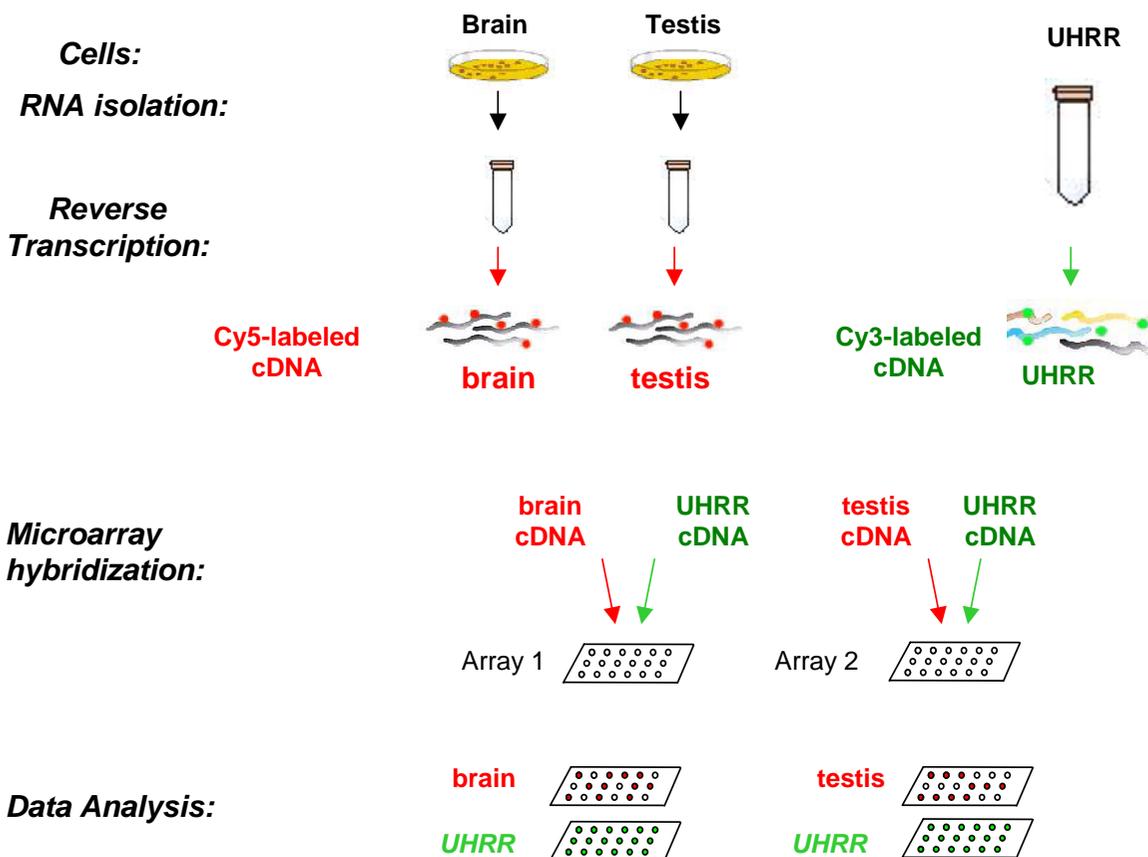
Figure 1
Microarray experimental design: Comparison of gene expression in two human cell lines using UHRR as a common reference sample. Total RNA isolated from human brain and testis cell lines were reverse-transcribed to cDNA and labeled with Cy5. UHRR was reverse-transcribed to reference cDNA and labeled with Cy3. Each Cy5-labeled cDNA was co-hybridized with Cy3-labeled reference cDNA on microarrays and data analyzed as described in Material and Methods.

Table 1: Row data of KIAA0923 protein gene.

| Spot	KIAA0923 protein Hs.22587 Sample ID	Brain Cy5 Intensity	Testis Cy5 Intensity	Brain Cy3 Intensity	Testis Cy3 Intensity	Brain R/G Ratio	Testis R/G Ratio	Brain log2 Ratio	Testis log2 Ratio
1	1291986	382	125	187	189	2.04	0.66	0.713	-0.416
2	IMAGE:46091	2663	668	971	769	2.74	0.87	1.008	-0.139
3	IMAGE:1505784	321	163	140	274	2.28	0.59	0.824	-0.528
4	IMAGE:51460	912	368	575	477	1.58	0.77	0.457	-0.261
5	IMAGE:132954	319	99	203	182	1.57	0.54	0.451	-0.616
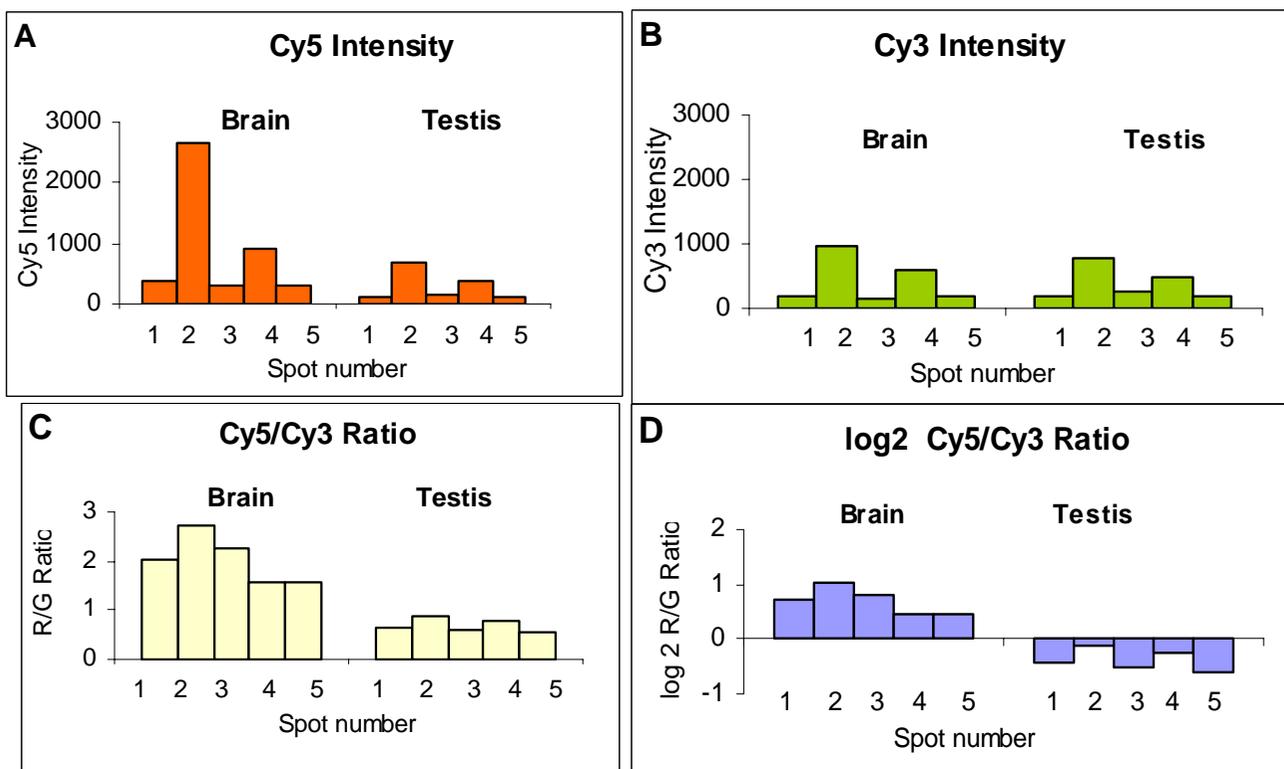
Figure 2
Gene expression of human brain and testis cell line KIAA0923 protein. Total RNA isolated from brain and testis cell lines were reverse-transcribed to cDNA, labeled with Cy5 and co-hybridized with Cy3-labeled UHRR onto two 43,000-spot cDNA microarrays (Stanford University). Each microarray had five separate spots containing the probes for KIAA0923 protein gene: (A) Comparison of Cy5 absolute fluorescence intensities registered on five KIAA0923 spots on arrays co-hybridized with Cy5-labeled brain and testis cDNA, and Cy3-labeled UHRR; (B) Comparison of Cy3 absolute fluorescence intensities registered on five KIAA0923 spots on arrays co-hybridized with Cy5-labeled brain and testis cDNA, and Cy3-labeled UHRR; (C) Ratios of intensity values in red and green channels (Cy5/Cy3 ratio); (D) Log2 of the Cy5/Cy3 ratios.

Table 2: Microarray Coverage of UHRR on 43,000-spot microarrays with yeast control spots.

Microarray print	Dyes	1% False-positive Yeast Intensity	UHRR microarray coverage	5% False-positive Yeast Intensity	UHRR microarray coverage
shat 041	Cy5	36	77%	26	82%
shaz 136	Cy5	25	53%	18	59%
shau 100	Cy5	79	57%	32	73%
shau 085	Cy5	29	57%	21	65%
shav 141	Cy3	33	63%	19	75%
shau 099	Cy3	96	59%	65	67%
shaz 137	Cy3	94	65%	48	75%
			62 ± 3%		**71 ± 3%**

each other and differentially expressed genes tabulated. Table 1 and Figure 2 illustrate how UHRR allows interpretation of microarray data that would be misinterpreted by ignoring its use. Five different probes for KIAA0923 protein (UniGene cluster Hs.22587), printed on Stanford 43K arrays, are characterized by a seven to eight fold difference in absolute fluorescence intensities in both Cy5 (red) and Cy3 (green) channels on the same microarray (Fig. 2A and 2B). This difference in fluorescence intensity is due to variable spot geometry and probe concentration (data not shown). Using UHRR in combination with brain and testis cell line RNA and computing the Cy5/Cy3 ratio minimizes spot-to-spot variability and allows characterization of high and low gene expression relative to the reference sample. Using Cy5/Cy3 ratios, the data reveals that KIAA0923 gene expression in brain cells is approximately 2-fold higher compared to UHRR (Cy5/Cy3 = 1.57 – 2.74) and 1.5-fold lower in testis cells compared to UHRR (Cy5/Cy3 = 0.54 – 0.87) (Fig. 2C). Logarithmic transformation of the Cy5/Cy3 ratio results in symmetric distribution about zero (Fig. 2D). The transformed data clearly demonstrate expression of the KIAA0923 gene in brain cells and suppression in testis cells.

Microarray coverage
Ideally, URR should hybridize to a large fraction of microarray spots, allowing accurate ratios to be determined for as many spots as possible. We define "microarray coverage" as the percentage of spots with hybridization signal above a selected threshold value. The statistical identification of "expressed" genes is a controversial topic, and therefore, we chose two different and simple methods to define threshold values for the identification of "expressed genes". One method is based on signal intensities of microarray control spots, such as 384 yeast ORFs on the 43,000-spot human microarrays from Stanford University. The second method uses background intensity adjacent to spots as the threshold. Microarray coverage of UHRR was determined by counting the number of spots with intensities exceeding the 1% and 5% false positive rates defined using the yeast negative control spots (Table 2). Seven microarrays from different print runs were used. In four experiments, the UHRR was labeled with Cy5 and in three experiments with Cy3. Labeled UHRR was co-hybridized with Common Reference RNA (CRF) developed at Stanford University (8) labeled with the alternate dye. When a 1% false positive cut off was used, fluorescence intensity from the yeast control spots varied between 25 and 96 fluorescence units. The analogous values for the 5% false positive cut off were 18 to 65. Microarray coverage of UHRR was 62 ± 3% and 71 ± 3% using 1% or 5% false positive cut off, respectively.

To evaluate microarray coverage of UHRR, UMRR and URRR on other microarrays, average background intensity values were used as thresholds (either 1X or 2X the background intensity for each channel; Table 3). We calculated the microarray coverage by counting spots with background-subtracted intensity exceeding 1X or 2X background, expressed as a percentage. For example, the Cy3 intensity value of one spot was 500, the local background intensity was 50 and the average background intensity of all spots in the Cy3 channel was 100. First, we subtracted local background from the intensity value of the spot (500-50 = 450). Spots were called "present" if background-subtracted intensity exceeded 100 (1X) and 200 (2X) background intensity thresholds in the Cy3 channel. When 1X background was used as the threshold, UHRR demonstrated 98–99% coverage of 7,600-spot and 10,000-spot human microarrays obtained from NCI, 92% of 12,000-spot Agilent Human I cDNA microarrays, and 85–86% of 41,000 and 43,000-spot human microarrays obtained from Stanford University. UMRR had 97% coverage of 8,700-spot and 7,500-spot mouse microarrays (NCI and UNC, respectively) and 87% coverage of 8,500-spot mouse microarrays (Agilent). URRR demonstrated 81% coverage of 6,500-spot rat microarrays (NCI) and 86% coverage of 14,500-spot rat microarrays (Agilent). When 2X background was used as the threshold, UHRR microarray coverage ranged from 97% of 7,600-spot microarrays (NCI) to 60% coverage of 43,000-spot microarrays (Stanford). UMRR showed 93% coverage of 8,700-spot microarrays (NCI), 85% of 7,500-spot microarrays (UNC) and 70% of 8,500-spot microarrays (Agilent). URRR demonstrated 72% and 62% coverage of 6,500-spot (NCI) and 14,500-spot (Agilent) microarrays, respectively. It should be noted that results obtained from cDNA microarrays might actually underestimate the "microarray coverage" because as was shown by Weil M.R. (19) that 12.8% of cDNA clones fail to show signal intensity under any condition.

Contribution of unique genes from individual cell lines to URR
We identified the number of unique genes contributed by each cell line to URR by performing the following experiment. Each cell line RNA comprising UHRR, UMRR, and URRR was reverse transcribed to cDNA, labeled with Cy5, and co-hybridized with the corresponding Cy3-labeled UHRR, UMRR or URRR. Stanford 43,000-spot human microarrays, UNC 7,500-spot mouse microarrays, and Agilent 14,500-spot rat microarrays were used for this experiment. Data from 10 human, 11 mouse and 14 rat microarrays were analyzed using GeneTraffic and are presented in Figures 3 through 5. Criteria for gene uniqueness were spot fluorescence intensity greater than 1000 in the Cy5 channel and Cy5/Cy3 ratio greater than 2 in a single cell line. If a spot possessed these same fluorescent inten-

Table 3: Microarray Coverage of UHRR, UMRR and URRR.

Array platform	Array used for analysis	Spots number	Spots with intensity value above zero %	Spots with intensity value above 1x background %	Spots with intensity value above 2x background %
UHRR					
NCI, NIH	3	7,600	99.9	99	97
NCI, NIH	6	10,000	99.9	98	82
Agilent	4	12,000	99.9	92	68
Stanford University	6	41,000	99.9	86	63
Stanford University	11	43,000	99.9	85	60
UMRR					
NCI, NIH	3	8,700	99.9	97	93
University of North Carolina	4	7,500	99.9	97	85
Agilent	4	8,500	98	87	70
URRR					
NCI, NIH	2	6.500	98	81	62
Agilent	8	14,500	99.7	86	72

sity characteristics in more than one cell line, it was not considered unique and eliminated from the list of unique genes. The result of this experiment is that each cell line contributes a different number of unique genes to URR. For example, human brain, breast and liver cell lines contribute, respectively, 394, 343 and 335 unique, tissue-specific genes to UHRR (Fig. 3). The total number of highly expressed genes in UHRR contributed by individual cell lines is 2393. Eliminating one or more cell lines from the reference pool would result in reduction of microarray coverage by several hundred unique tissue-specific genes. UMRR includes 1673 uniquely expressed genes out of 7,500 sequences represented on a mouse microarray (Fig. 4) and URRR includes 2205 unique genes out of 14,500 genes on a rat microarray (Fig. 5). These results point out the rationale for pooling RNA from several cell lines to make URR, rather than relying on single cell line RNA.

Batch-to-batch reproducibility
Two batches of UHRR were compared by co-hybridization to 12,000-spot human microarrays. In one experiment (two arrays), Cy3-labeled cDNA reverse transcribed from UHRR (batch 1) was co-hybridized with Cy5-labeled cDNA derived from UHRR (batch 2). In a second experiment (two arrays), UHRR (batch 1) was labeled with Cy5 and UHRR (batch 2) with Cy3. Data was normalized using LOWESS sub-grid normalization method (GeneTraffic) and the signal intensities for the two fluorescent images were compared. The scatter-plot of background-subtracted mean values (normalized data) obtained from Cy3 and Cy5 channels are shown in Figure 6A. Each data point can be interpreted as the relative content of a given transcript in two batches of UHRR. The data suggest that there is nearly 1:1 correspondence between all expressed genes in both batches of UHRR. The Pearson's correlation coefficient between two channels was 0.9736 ± 0.004 (n = 4 microarrays). To evaluate the significance of the correlation values, UHRR samples from the same batches were

characterized as follows. Two aliquots of UHRR (batch 1) were reverse-transcribed with Cy5 and Cy3 and co-hybridized to the same microarray (n = 4). The scatter-plot of background-subtracted mean values (normalized data) is shown in Figure 6B. The Pearson's correlation coefficient of 0.9930 ± 0.0009 (n = 4) was obtained between the two channels. Gene expression comparison of two batches of UMRR and URRR also resulted in Pearson's correlation coefficients ≥ 0.97 (data not shown).

To identify statistically significant changes in gene expression between two batches of UHRR, the following statistical analysis was performed. First, all log2 intensity ratios resulting from 4 hybridizations from each of two URR batches were compared using a 2-tail non-parametric Mann-Whitney test. The critical z-value was calculated to be 4.448 using the Bonferroni correction for the multiple of 5,926 tests. None of the 5,926 individual z-parameters exceeded the critical z-value, which does not allow rejection of the null-hypothesis about the equality of log2-ratios between two batches of UHRR.

Second, Significance Analysis of Microarrays (SAM) was applied to the set of 5,926 genes (UHRR batch 1, n = 4; UHRR batch 2, n = 4) using batch number as the "supervising" parameter. SAM identified 78 significant genes with a 2-fold change and an estimated false discovery rate of (FDR) of 10% when $\Delta = 0.2$ was used as a threshold, and 61 significant genes (FDR = 1.1%) when $\Delta = 1$ was used as a threshold. Therefore, only 0.65% of the genes presented on these microarrays are differentially expressed in two batches of UHRR as assed by SAM, and no differentially expressed genes were identified when using the non-parametric Mann-Whitney test. These results demonstrate that URR was independently produced with high batch-to-batch consistency.

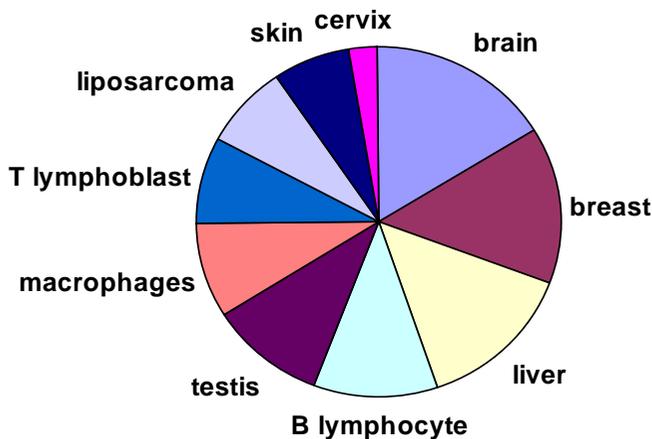| Human cell lines	Genes number (LEX.E >1000, R>2)
brain	394
breast	343
liver	335
B lymphocyte	256
testis	251
macrophages	212
T lymphoblast	193
liposarcoma	184
skin	163
cervix	62
Total	*2393*

Figure 3
Unique genes contributed by individual cell lines to UHRR. Total RNA isolated from 10 individual human cell lines were reverse-transcribed to cDNA, labeled with Cy5 and co-hybridized with Cy3-labeled UHRR onto 43,000-spot cDNA microarrays (Stanford University). The data was analyzed using GeneTraffic software. Approximately 6000–8000 spots out of 43,000 (14–18%) were flagged on each microarray and excluded from further analysis. Spots with hybridization signals in Cy5 channel higher than 1000 and with Cy5/Cy3 ratio greater than 2 were collected and the number of spots with these characteristics on only one microarray was determined.

Discussion

A universal human reference sample could be obtained from "normal" tissue samples. Unfortunately, human samples are difficult and expensive to procure and "normal" is difficult to define due to genotypic, pathological and environmental variability. Large-scale production of URR from tissues is also problematic as the following example illustrates. If lymphocytes were isolated from whole human blood, instead of the lymphocyte cell line, then the desired amount of RNA to make multi-gram batch of UHRR would require 10,000 liters of blood. Genotype, pathological state and environmental variables are more easily controlled in laboratory mice and rats. However, isolating large quantities of high quality RNA from tissue is problematic. URR from cell lines eliminates these problems. The properties of a cell line derived reference RNA sample was examined by Yang et al. [13], however, they failed to address the long term needs of the research community for a very large scale and highly reproducible reference sample. We build upon these studies and show

that a highly reproducible reference sample can be created from a pool of RNAs derived from cell lines, and demonstrate that this reference samples gives good "coverage" across many diverse microarray platforms. Immortal cell lines are obviously not "normal", but most are easy to cultivate, their growth is controllable and large quantities of high quality material can be obtained at a reasonable cost providing multi-gram batches of human, mouse and rat URR.

Whole laboratory animals or embryos are candidate sources for UMRR and URRR. However, this approach does not take into account organ mass differences and would result in overrepresentation of transcripts from certain tissues in the final RNA pool. Whole mouse URR would contain 30% liver RNA, 27% muscle RNA, and 27% intestine RNA. The balance of RNA would be derived from all remaining small-mass organs.
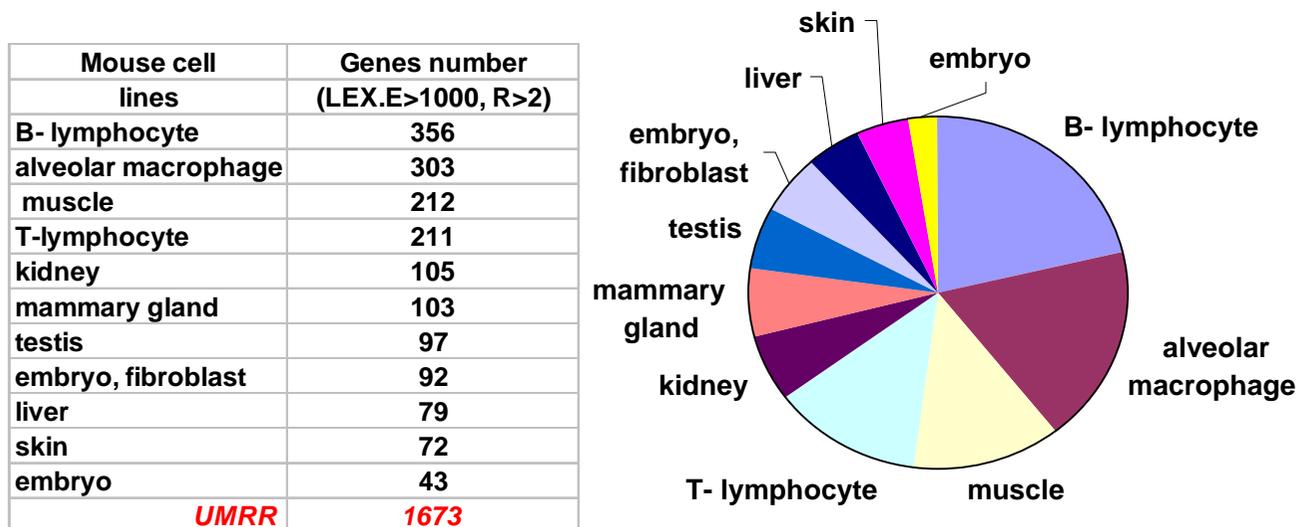
| Mouse cell lines	Genes number (LEX.E>1000, R>2)
B- lymphocyte	356
alveolar macrophage	303
muscle	212
T-lymphocyte	211
kidney	105
mammary gland	103
testis	97
embryo, fibroblast	92
liver	79
skin	72
embryo	43
UMRR	*1673*

Figure 4
Unique genes contributed by individual cell lines to UMRR. RNA from 11 individual mouse cell lines were reverse-transcribed to cDNA, labeled with Cy5 and co-hybridized with Cy3-labeled UMRR onto 7,500-spot mouse oligo microarrays (UNC). The data was analyzed using GeneTraffic software. 300–1000 spots out of 8,000 were flagged on each microarray and excluded from further analysis. Spots with hybridization signals in Cy5 channel higher than 1000 and with Cy5/Cy3 ratio greater than 2 were collected and the number of spots with these characteristics on only one microarray was determined.

Data presented in this paper demonstrate highly correlated gene expression for different batches of cell line URR. We recognize that even though the Pearson's correlation coefficient for two batches of URR is 0.97, gene expression levels undoubtedly vary considerably among the thousands of genes measured in a variety of microarrays. Indeed, approximately 1% of the genes analyzed in this study were significantly differentially represented in batch to batch comparisons. Since these genes are likely to be related physiologically or developmentally, this variation has the potential to result in errors in gene expression measurements in which different URR batches are used. A solution to this problem will be to perform comparative hybridizations of two or more URR batches and generate a batch-correction factor. The factor will be validated by applying it to analyses of several different experimental samples and measuring how close the reproducibility of these analyses are, using different batches of URR, relative to the reproducibility of the same analysis using a single batch of URR.

We also recognize that achieving the ultimate goal of increasing positive hybridization signal above background at all microarray spots might be realized by adding synthetic RNA or genomic DNA to URR. The increased use of oligonucleotide microarrays with known sequences, may allow the synthesis of pre-labeled complementary oligonucleotides that would provide signal for every gene. The drawback of this strategy is that it would not control for variation in the RNA labeling step. Data presented in this paper demonstrates that pools of RNA derived from a limited yet diverse set of cell lines results in URR nearly accomplishing that goal.

Conclusions
Results of this study demonstrate that large quantities of pooled RNA from individual cell lines are reproducibly prepared and possess diverse gene representation. The Universal Reference RNA provides a long-sought solution for the standardization and cross-referencing of microarray experiments by offering a high-quality standard for accurate and consistent data comparison.

Rat cell lines	Ggenes number (LEX.E>1000, R>2)
lung	356
alveolar macrophage	340
T lymphocyte	253
B lymphoblast	122
basophil	199
brain	167
liver	66
kidney	92
mammary gland	124
muscle	72
skin	158
embryo, fibroblast	47
embryo, yolk	209
URRR	*2205*
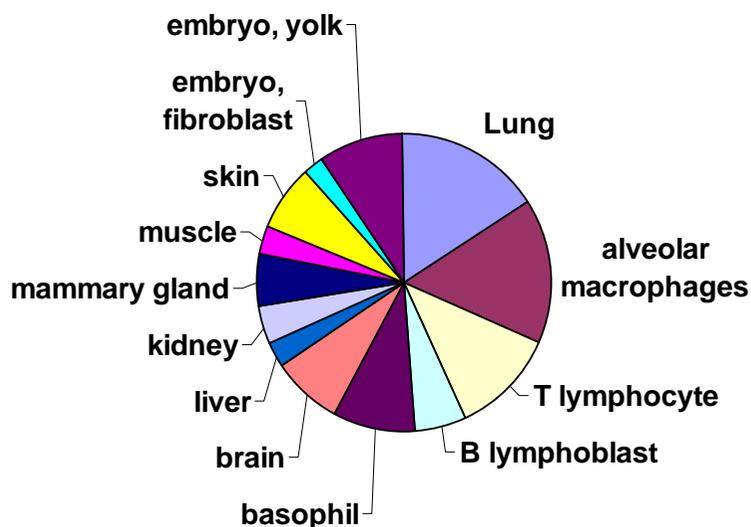
Figure 5
Unique genes contributed by individual cell lines to URRR. RNA from 14 individual rat cell lines were reverse-transcribed to cDNA, labeled with Cy5 and co-hybridized with Cy3-labeled URRR onto 14,000-spot rat cDNA microarrays (Agilent). The data was analyzed using GeneTraffic software. 400–1200 spots out of 14,000 were flagged on each microarray and excluded from further analysis. Spots with hybridization signals in Cy5 channel higher than 1000 and with Cy5/Cy3 ratio greater than 2 were collected and the number of spots with these characteristics on only one microarray was determined.

Methods

Cell lines

Ten human cell lines derived from the following human tissues were selected for UHRR including liver, testis, mammary gland, cervix, brain, skin, liposarcoma, macrophage, T-lymphoblast and B-lymphocyte [20]. Eleven mouse cell lines, representing liver, kidney, testis, mammary gland, embryo, alveolar macrophages, skin, muscle, macrophage, T-lymphocyte and B-lymphocyte, were chosen for UMRR [21]. Fourteen rat cell lines derived from liver, kidney, brain, testis, mammary gland, embryo, lung, skin, fibroblast, muscle, macrophage, basophil, T-lymphocyte and B-lymphoblast, were used for URRR [22]. The human and mouse cells were grown to 60–80% confluence in RPMI-1640 media while rat cell lines were grown to the same extent in DMEM media. Both media's were supplemented with 2 mM L-glutamine, 10% fetal bovine serum and penicillin/streptomycin. At this point, old media was replaced with fresh media and the cells harvested after 24 hours by trypsinization, washed with 1X PBS, the cell pellets frozen in liquid nitrogen, and stored at -80°C.

RNA isolation

Total RNA was isolated using modified StrataPrep™ Total RNA isolation kit [23] (Stratagene, La Jolla, CA). One to three × 10^8 cells were lysed in 15 ml of lysis buffer containing guanidine isothiocyanate and filtered in a spin cup to remove particles and reduce contaminating DNA. An equal volume of 70% ethanol was added to the cell lysate followed by vortexing. The mixture was transferred to a second spin cup containing an RNA binding filter and centrifuged for 5–10 min at 5000 × g followed by washing with 15 ml of low-salt buffer. DNase I (500 U of DNase I in 500 μl of DNase digestion buffer) was added directly to the spin cup filter and incubated at 37°C for 15 min. The filter was washed with 10 ml of high-salt wash buffer, 15 ml of low-salt wash buffer and finally with 10 ml of low-salt wash buffer. Total RNA was eluted from the filter using 1 ml of elution buffer added directly to the spin cup filter. The cup was incubated for 2 min at room temperature, and centrifuged for 3 min at 3000 × g. The latter step was repeated twice more. The total volume of elution buffer added was 3 ml. The quantity and quality of isolated RNA was determined by spectrophotometry. RNA integrity was determined by two methods; formaldehyde-
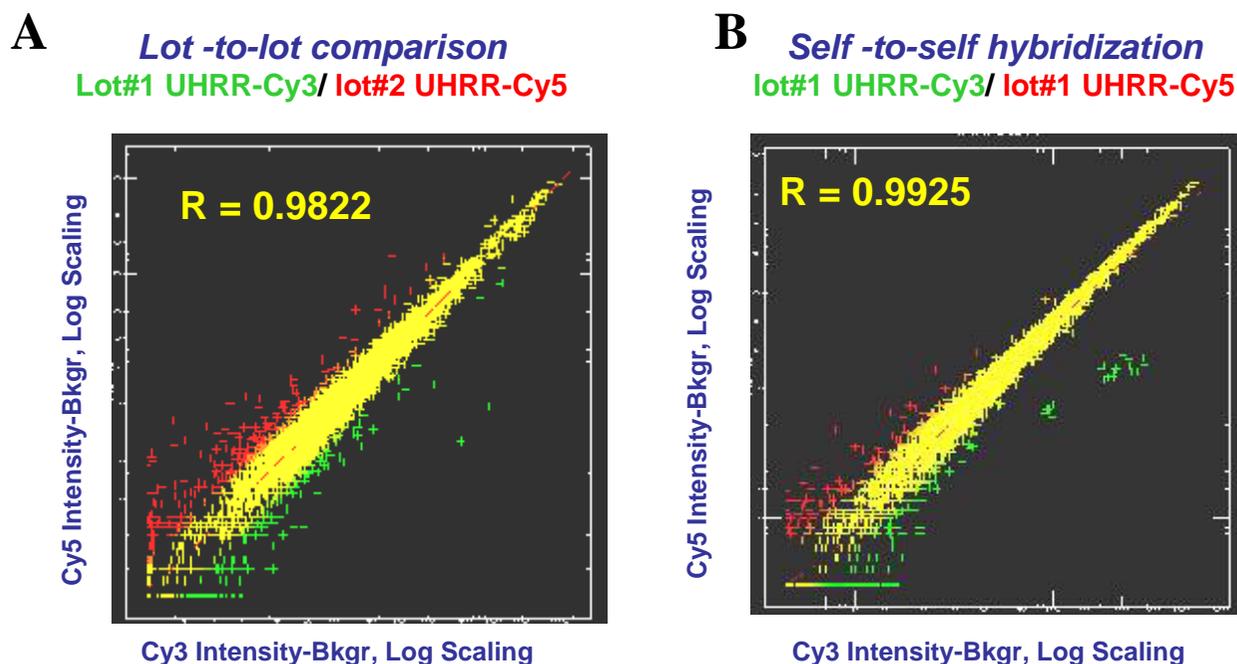
Figure 6
UHRR batch-to batch comparison. (A) Scatter plot of signal intensities from 12,000 spot human microarray (normalized data obtained using GeneTraffic), using Cy3-labeled UHRR batch 1 co-hybridized with Cy5-labeled UHRR batch 2. (B) Scatter plot of Cy3-labeled UHRR batch 1 co-hybridized with Cy5-labeled UHRR batch 1 to 12,000-spot human microarray.

agarose gel electrophoresis and Agilent Bioanalyzer analysis (Agilent Technologies, Palo Alto, CA). URR was prepared by pooling equal mass quantities of total RNA from each cell line, dividing the pool into 200 µg aliquots followed by ethanol precipitation and storage at -80°C.

cDNA synthesis and labeling
Labeled cDNAs were synthesized with the FairPlay cDNA labeling kit (Stratagene, La Jolla, CA). 20 µg of total RNA (individual cell line RNA or URR) in 12 µl of DEPC-treated water was combined with 1 µl of 500 ng/µl oligo-d(T)$_{12-18}$. The mixture was incubated at 70°C for 10 min and cooled on ice. For each reaction, 2 µl of 10X StrataScript reaction buffer, 1 µl of unlabeled 20X dNTP mix containing amino-allyl dUTP, 1.5 µl of 0.1 M dithiothreitol and 0.5 µl of RNase Block (40 U/µl) were prepared and mixed with the RNA sample and 1 µl of 50 U/µl StrataScript RT. After incubation at 48°C for 30 min an additional 1 µl of StrataScript RT was added and incubation was continued for 30 additional minutes. RNA was

degraded by adding 10 µl of 1 M NaOH, followed by a 10-min incubation at 70°C and the mixture neutralized with 10 µl of 1 M HCl. Unincorporated nucleotides were removed by precipitation of the cDNA with 4 µl of 3 M sodium acetate, 1 µl of 20 mg/ml glycogen and 100 µl of 95% ethanol at -20°C for 1 hr. After centrifugation and washing the pellet with 70% ethanol, it was resuspended in 5 µl of 2x coupling buffer provided in the kit. Cy3 or Cy5 dye (Amersham), resuspended in 45 ul of DMSO, was added and the reaction incubated for 30 min at room temperature in the dark. Dye-coupled cDNA was purified with a DNA-binding spin cup, as described in the FairPlay cDNA labeling kit protocol, and the final volume adjusted to 5 µl.

Microarrays
Human 7,600-spot and 10,000-spot, mouse 8,700-spot and rat 6,500-spot cDNA microarrays were printed at the National Cancer Institute (NCI; NIH, Gaithersburg, MD). Human 12,000-spot (human 1), mouse 8,500-spot and

14,500-spot rat cDNA microarrays were purchased from Agilent Technologies (Palo Alto, CA). Human 41,000 and 43,000-spot cDNA microarrays were printed at the Stanford Functional Genomics Core Facility (Stanford University; http://www.microarray.org). Mouse 7,500-spot oligonucleotide microarrays were printed at the University of North Carolina (UNC; Chapel Hill, NC) using the Compugen – Sigma murine oligo set.

Microarray pre-hybridization (blocking)

Microarrays were pre-hybridized at 42°C for at least 1 hr in 20–30 μl of pre-hybridization buffer (5X SSC, 0.1% SDS and 1% BSA) covered with coverslips. The slides were then washed by rapidly dipping them in distilled water for 2 min, followed by dipping in isopropanol for 2 min followed by air drying.

Microarray hybridization and data processing

5 μl each of Cy3-labeled and Cy5-labeled cDNA targets were combined with 2 μl of 10 μg/μl human Cot 1 DNA (mouse Cot 1 DNA was used for mouse and rat microarrays; Gibco-BRL), 2 μl of 8 μg/μl poly d(A)$_{40-60}$ and 2 μl of 4 μg/μl yeast tRNA (Gibco-BRL). Labeled cDNA target (16 μl) was denatured at 100°C for 1 min and cooled on ice. 16 μl of 2X hybridization buffer (50% formamide, 10X SSC and 0.2% SDS) was added and 30 μl of the mixture was applied to a single microarray under a glass coverslip. Microarrays were incubated at 42°C for 16 hr in sealed chambers with humidity maintained by a small reservoir of 3X SSC. Arrays were washed in 2X SSC, 0.1% SDS for 4 min, 1X SSC, 0.1% SDS for 4 min, 0.2X SSC for 4 min, 0.05X SSC for 1 min and air dried. Hybridization signal was visualized and collected using an Axon microarray scanner.

Data analysis

Data from each array was collected with GenePix 3.0 (Axon Instruments). Each spot was defined by manual positioning of a grid over the array image. Aberrant and empty spots were manually flagged and excluded from further analysis. The average pixel intensity within each circle was determined and local background was computed for each spot. Net signal was determined by subtracting local background from the average intensity.

Microarray coverage of UHRR on 43,000-spot microarrays using control spots

Data files generated by GenePix 3.0 were exported into the Stanford Microarray Database (SMD). After background subtraction, normalization and filtering the raw intensity values were used for analysis. Fluorescence intensities of spots representing human genes and negative control spots were compared to estimate the number of genes represented in the UHRR. Human 43,000-spot microarrays (Stanford University) have 384 yeast gene spots used as negative controls. Signals produced at these control spots when hybridized to human cDNA were considered non-specific. (BLAST analysis of 384 yeast ORF nucleotide sequences against UniGene human database did not show cross-reactivity between the yeast spots and human cDNA with the expected values lower than 1E-14). A signal intensity threshold from 25 to 96 fluorescence units, depending on microarray print, was defined such that 1% of the control spots showed greater signal intensity (four of the most intense yeast control spots). 62 ± 3% of the human gene spots had signal intensity greater than this threshold. A second threshold of 18–65 fluorescence units was defined such that 5% of the control spots showed greater signal intensity (nineteen of the most intense yeast control spots). 71 ± 3% of the human gene spots had signal intensity greater than this threshold.

Microarray coverage of URR on different microarrays using the average background intensity

For microarrays lacking the control spots described above, average background intensity values in each channel were used as the threshold. Data files generated by GenePix 3.0 were exported into GeneTraffic (Iobion Bioinformatics, Toronto, Canada). After background subtraction, normalization, and filtering of spot intensities, those spots with intensity above 1X and 2X (1 and 2-fold higher than average of background intensity, respectively) were considered positive.

Evaluation of gene number contributed by individual cell line RNA to the reference pool

Highly expressed genes in individual cell lines were identified by comparing microarrays hybridized with total RNA prepared from each cell line and URR pools. This evaluation was performed for 10 human, 11 mouse and 14 rat cell lines. Data was analyzed using GeneTraffic. Signal intensities between two fluorescent images were normalized using Locally Weighted Scatter Plot Smoother (LOWESS) sub-grid normalization [24]. All the spots with signal intensity less than the local background for Cy5 and Cy3 channels were flagged. In the second step, spots with signal intensity greater than 1000 (highly expressed genes) in the Cy5 channel, and with Cy5/Cy3 ratio's greater than 2, were selected and the number of spots with these parameters average of background intensity in one cell line was evaluated.

Batch-to-batch comparison

Two batches of UHRR were compared by co-hybridizing Cy3-labeled cDNA reverse-transcribed from batch 1 and Cy5-labeled cDNA from batch 2 to the same microarray. The experiment was repeated by switching the dyes. The Pearson correlation coefficient for background-subtracted mean intensities for Cy5 and Cy3 channels was calculated. Identification of genes whose expression level sig-

nificantly differed between two batches of URR was accomplished using the following statistical analysis (only data with intensity value exceeding 2X background were used). First, individual values of the log2 intensity ratios in 4 hybridizations from each of two URR industrial batches were compared using a 2-tail non-parametric Mann-Whitney test. The critical z-value was calculated as 4.448 using the Bonferroni correction for the multiple of 5,926 tests (e. g. see http://home.clara.net/sisa/bonhlp.htm). Second, *Significance Analysis of Microarrays (SAM)* software was used [25]. False discovery rate (FDR; percentage of genes selected by chance) was determined by recursive permutations. Data files generated by Gene-Pix 3.0 were entered into GeneTraffic and signal intensities between two fluorescent images were normalized using the LOWESS sub-grid normalization method. All spots with intensity less than local background in Cy5 or Cy3 channels were flagged and excluded from further analysis. A table including all non-flagged spots was generated and exported to SAM. We used the following parameters for data analysis: two class, unpaired data (log, base2), number of permutations 100 and average of background intensity number of neighbors 10.

Authors' contributions

NN participated in development of URR and experimental design, including cell lines selection and culturing, RNA isolation, microarray hybridization experiments, microarray data analysis and drafting of manuscript. CP and MW participated in experimental design, microarray hybridization experiments, microarray data analysis and contributed to writing and revising the manuscript. AN was involved in microarray data analysis and drafting of the manuscript. SB participated in developing URR, including cell lines culturing, RNA isolation and microarray hybridization experiments. WW was involved in experimental design and implementation of the study. RP, JU, MK and OA performed microarray hybridization experiments. MF participated in UHRR development, including cell lines selection and microarray experiments performance. DB generated the idea of using RNA from multiple cell lines as reference RNA for two-color microarray hybridization experiments, developed the first human reference RNA, in collaboration with Dr. Patrick O. Brown, and participated in URR design and development. JB participated in study design, coordination of the study and drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

References

1. Southern EM: **Detection of specific sequences among DNA fragments separated by gel electrophoresis.** *J Mol Biol* 1975, **98(3):**503-17.
2. Kafatos FC, Jones CW, Efstratiatis : **Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure.** *Nucl Acids Res* 1979, **7(6):**1541-52.
3. Brown PO, Botsein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21:**33-37.
4. Kerr MK, Churchill GA: **Experimental design for gene expression microarrays.** *Biostatistics* 2001, **2(2):**183-201.
5. Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Gen Supp* 2002, **32:**490-496.
6. Townsend JP: **Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays.** *BMC Genomics* 2003, **4:**41.
7. Eisen MB, Brown PO: **DNA arrays for analysis of gene expression.** *Methods Enzymol* 1999, **303:**179-205.
8. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
9. Chang HY, Chi JT, Dudoit S, Bondre C, van De Rijn M, Botstein D, Brown PO: **Diversity, topographic differentiation, and positional memory in human fibroblasts.** *Proc Natl Acad Sci USA* 2002, **99:**12877-12882.
10. Perou CM, Sorlie T, Eisen MB, Van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
11. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I: **Diversity of gene expression in adenocarcinoma of the lung.** *Proc Natl Acad Sci U S A* 2001, **98(24):**13784-13789.
12. **Maximizing gene expression microarray data quality with Stratagene's Universal Reference RNA.** *Strategies* 2003, **16(1):**10-11.
13. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biology* 2002, **3(11):**RESEARCH0062.1-0062.12.
14. Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M, Botstein D, Brown PO: **Gene expression patterns in human liver cancers.** *Mol Biol Cell* 2002, **13(6):**1929-39.
15. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D: **Identification of genes periodically expressed in the human cell cycle and their expression in tumors.** *Mol Biol Cell* 2002, **13(6):**1977-2000.
16. Puskas LG, Zvara A, Hackler LJr, Micsik T, van Hummelen P: **Production of bulk amounts of universal RNA for microarrays.** *Biotechniques* 2002, **33(6):**898-900.
17. Sterrenburg E, Turk R, Boerg JM, van Ommen GB, den Dunnen JT: **A common reference for cDNA microarray hybridizations.** *Nucleic Acids Res* 2002, **30(21):**e116.
18. Dudley AM, Aach J, Steffen A, Church GM: **Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range.** *Proc Nat Acad Sci USA* 2002, **99(11):**7554-7559.
19. Weil MR, Macatee T, Garner HR: **Toward a universal standard: comparing two methods for standardizing spotted microarray data.** *Biotechniques* 2002, **32(6):**1310-4.
20. Novoradovskaya N, Payette T, Novoradovsky A, Braman J, Chin N, Pergamenschikov A, Fero M, Botstein D: **Pooled, high-quality Reference RNA for human microarrays.** *Strategies* 2000, **13(4):**121-122.
21. **Pooled total RNA for mouse microarray experiment.** *Strategies* 2001, **14(4):**1-2.
22. **Common Standard for rat microarray experiment.** *Strategies* 2002, **15(4):**3-4.

23. Dolter K, Braman J: **Fast isolation of total RNA from small numbers of cells.** *Strategies* 2000, **13(2):**12-14.
24. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **Suppl 32:**496-501.
25. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Nat Acad Sci USA* 2001, **98:**5116-5121.