nature biotechnology

A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies

Winston Patrick Kuo[1,2,3,18], Fang Liu[4,18], Jeff Trimarchi[2], Claudio Punzo[2], Michael Lombardi[5], Jasjit Sarang[5], Mark E Whipple[6], Malini Maysuria[7], Kyle Serikawa[7], Sun Young Lee[8], Donald McCrann[9], Jason Kang[10], Jeffrey R Shearstone[11], Jocelyn Burke[2,12], Daniel J Park[2,12], Xiaowei Wang[1,12], Trent L Rector[2], Paola Ricciardi-Castagnoli[13], Steven Perrin[11], Sangdun Choi[14], Roger Bumgarner[7], Ju Han Kim[15], Glenn F Short III[2,12], Mason W Freeman[2,12], Brian Seed[2,12], Roderick Jensen[5], George M Church[2], Eivind Hovig[4], Connie L Cepko[2], Peter Park[16], Lucila Ohno-Machado[3] & Tor-Kristian Jenssen[17]

Over the last decade, gene expression microarrays have had a profound impact on biomedical research. The diversity of platforms and analytical methods available to researchers have made the comparison of data from multiple platforms challenging. In this study, we describe a framework for comparisons across platforms and laboratories. We have attempted to include nearly all the available commercial and 'in-house' platforms. Using probe sequences matched at the exon level improved consistency of measurements across the different microarray platforms compared to annotation-based matches. Generally, consistency was good for highly expressed genes, and variable for genes with lower expression values as confirmed by quantitative real-time (QRT)-PCR. Concordance of measurements was higher between laboratories on the same platform than across platforms. We demonstrate that, after stringent preprocessing, commercial arrays were more consistent than in-house arrays, and by most measures, one-dye platforms were more consistent than two-dye platforms.

Gene expression microarray technology has greatly matured over the past decade, and it is expected that the technology will extend its current role as an experimental tool for basic science research and become increasingly applied in clinical practice. Several large efforts to create standardized protocols for microarray experiments (from probe annotation to data analysis) have been initiated: the Minimum Information About a Microarray Experiment (MIAME) standards (http://www.mged.org/Workgroups/MIAME/miame.html), The External RNA Controls Consortium (ERCC) (http://www.cstl.nist.gov/biotech/Cell&TissueMeasurements/GeneExpression/ERCC.htm) and The MicroArray Quality Control (MAQC) project (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/). All these initiatives aim at improving the quality of microarray data through standardization.

Major portals for deposition and retrieval of microarray data, such as the Gene Expression Omnibus (GEO)[1] and ArrayExpress[2], will be truly useful only if experiments are sufficiently reliable and annotated so that meaningful results can be extracted across platforms. The diversity of platforms and microarray data raise the questions of whether and how data from different platforms can be compared and combined. The results from previous cross-platform comparisons have been mixed and continue to be debated[3–30]. Although a body of information continues to develop, at least one of the following factors may have biased the results of previous comparative studies: (i) nonidentical samples on different platforms; (ii) samples not sufficiently distinct; (iii) samples processed using different protocols; (iv) lack of technical replicates; (v) data preprocessing steps not standardized; (vi) only a few types of platforms directly compared; (vii) measurements matched using probe annotations; (viii) 'agreement' not unambiguously quantified or (ix) insufficient biological validation. Although some of the above conditions may be reflective of the actual limitations of these platforms, in practice they complicate assessing the magnitude of disagreement attributable to the platforms.

[1]Department of Developmental Biology, Harvard School of Dental Medicine, 188 Longwood Ave., Boston, Massachusetts 02115, USA. [2]Department of Genetics, Harvard Medical School, Howard Hughes Medical Institute, Boston, Massachusetts, USA. [3]Decision Systems Group, Brigham and Women's Hospital, Boston, Massachusetts, USA. [4]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Montebello, Oslo, Norway. [5]Department of Physics, University of Massachusetts Boston, Boston, Massachusetts, USA. [6]Department of Otolaryngology-Head and Neck Surgery, [7]Department of Microbiology, University of Washington, Seattle, Washington, USA. [8]Division of Biology, California Institute of Technology, Pasadena, California, USA. [9]Department of Biochemistry, Boston University School of Medicine, Boston, Massachusetts, USA. [10]Macrogen Inc., Seoul, Korea. [11]Research Molecular Discovery, Biogen Idec, Cambridge, Massachusetts, USA. [12]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA. [13]Department of Biotechnology and Bioscience, University of Milano-Bicocca, Milano, Italy. [14]Department of Molecular Science and Technology, Ajou University, Suwon, Korea. [15]Seoul National University Biomedical Informatics, Seoul National University College of Medicine, Seoul, Korea. [16]Childrens' Hospital Informatics Program, Harvard Medical School, Boston, Massachusetts, USA. [17]PubGene AS, Vinderen, Oslo, Norway. [18]These authors contributed equally to this work. Correspondence should be addressed to W.P.K. (wkuo@genetics.med.harvard.edu).

Received 4 January; accepted 25 April; published online 2 July 2006; doi:10.1038/nbt1217

Table 1 Intra-platform performance

	Unfiltered					Filtered				
	Correlation coefficient			Accuracy		Correlation coefficient			Accuracy	
	Pearson	Spearman	S.d.	Score	No. of genes	Pearson	Spearman	S.d.	Score	No. of genes
Affymetrix	0.78	0.71	0.73	0.39	149	0.95	0.91	0.23	0.47	117
Amersham	0.89	0.84	0.52	0.41	149	0.95	0.93	0.25	0.44	135
Mergen	0.68	0.71	1.27	0.54	136	0.91	0.86	0.37	0.37	111
ABI	0.81	0.70	0.74	0.60	137	0.97	0.95	0.26	0.66	130
cDNA	0.71	0.72	1.08	0.17	37	0.66	0.65	1.03	0.18	36
MGH	0.85	0.84	0.74	0.55	99	0.93	0.91	0.45	0.61	92
MWG	0.87	0.80	0.25	0.17	87	0.92	0.84	0.20	0.17	87
Agilent	0.95	0.88	0.19	0.27	148	0.95	0.88	0.18	0.27	148
Compugen	0.83	0.87	0.91	0.28	43	0.97	0.96	0.24	0.35	34
Operon	0.87	0.89	0.68	0.40	136	0.97	0.96	0.25	0.47	118

Each platform was evaluated on internal performance through a number of statistics. Results are shown both for filtered and unfiltered data to illustrate the positive effects of filtering. Pearson and Spearman correlation coefficients calculated from normalized $\log_2$ ratios suggest good overall agreement between technical replicates. The Pearson and Spearman columns show the corresponding average correlation coefficients. S.d. over paired data from technical replicates are given for each platform to indicate the magnitudes of deviation between replicated $\log_2$ ratio measurements. Based on the biological validations, using the QRT-PCR $\log_2$ ratios as nominal values, we calculated an accuracy score for each platform as the slope of the regression line for measurements of common genes. The 'no. of genes' columns show the number of measurement pairs included in the regression for each platform. For the single-dye platforms, five technical replicates of $\log_2$ ratios were created by randomly pairing (without replacement) technical replicates of the single-sample experiments.

Because of the diversity of technical and analytical sources that can affect the results of an experiment and hence a comparison among experiments, standardization within a single platform may be insufficient. Nonetheless, several recent comparison studies involving microarrays have justified guarded optimism for the reproducibility of measurements across platforms but have also indicated the need for further large-scale comparison studies[4,7,10,31].

We present a comprehensive framework for cross-platform comparison of DNA microarrays based on data from ten different mouse microarray platforms. The study includes single- and dual-dye platforms, cDNA and oligonucleotide microarrays, and both commercial and in-house fabricated microarrays (see **Supplementary Data** online). Hybridizations were conducted in five replicates to enhance statistical reliability[32], and for three platforms, experiments were replicated at two different facilities. Each laboratory (see **Supplementary Data** online) received aliquots from two different RNA samples, mouse retina and mouse cortex prepared in the Cepko Laboratory at Harvard Medical School (http://genetics.med.harvard.edu/~cepko/). Pooling tissue from many animals before extracting the RNA minimized the biological variations within tissue RNA preparations. Large pools of each sample were collected to allow future inclusion of emerging technologies into the study.

Following methods from recent studies[13,33,34], we used probe sequence information to map probes at the level of both genes and exons to improve the stringency with which measurements are compared across platforms. To the best of our knowledge, this is the first study of this scale using probe sequences in this way. For the data analyses, we combined well-described, commonly used and publicly available analytical approaches in a framework that can be used every time the reliability of a new platform needs to be assessed.

RESULTS

Experimental study design

We evaluated intra-platform, inter-platform and inter-laboratory comparisons on ten different microarray platforms. The platforms included the following; Affymetrix, Agilent, Applied Biosystems (ABI), Amersham (now GE Healthcare), cDNA arrays provided by the Cepko laboratory (academic cDNA), Compugen (now Sigma-Genosys), Mergen, long oligonucleotide arrays from the Microarray Core facility at Massachusetts General Hospital (MGH long oligo), MWG BioTech (now Ocimum Biosolutions) and Operon platforms. Five replicate assays for each sample were processed for each platform. Biological validations were conducted by QRT-PCR. For further details about the experimental design and protocols, and data analyses, see **Supplementary Data** online and Methods, respectively.

Intra-platform comparisons

The consistency of measurements from technical replicates was generally good, both for intensities (absolute) and $\log_2$ ratios (relative). For intensities, the average Pearson correlation over all pairs of technical replicates was as high as 0.96 (after filtering; see **Supplementary Table 1a** online). Per platform, ABI had the highest correlation (>0.995) and academic cDNA had the lowest (0.88). As expected, the one-dye platforms had the highest correlations of intensities. When relative measurements were evaluated, the two-dye platforms, except for academic cDNA, had correlations similar to the single-dye platforms (**Table 1**). Spot quality filtering increased the intra-platform correlations for all platforms, except for the academic cDNA platform. Based on the s.d. of differences between paired data points, internal consistency was comparable across most of the platforms (**Table 1**).

Very similar patterns of performance were found when calculating coefficients of variation (CVs) per gene (see **Supplementary Table 1b** online). Based on CVs, platforms from ABI, Affymetrix, Amersham and Agilent performed best, whereas academic cDNA performed poorest.

The dynamic range of relative measurements provides information about how well a particular platform can reliably identify fold changes. Overall, the dynamic ranges were comparable for most of the platforms, except for those of Agilent and MWG, which had less than half the dynamic range of the others.

Inter-platform comparisons

To compare platforms, probe measurements were mapped to the following gene identifiers: UniGene (UG), LocusLink (LL), RefSeq (RS) and RefSeq exon (RSEXON). Mapping and matching

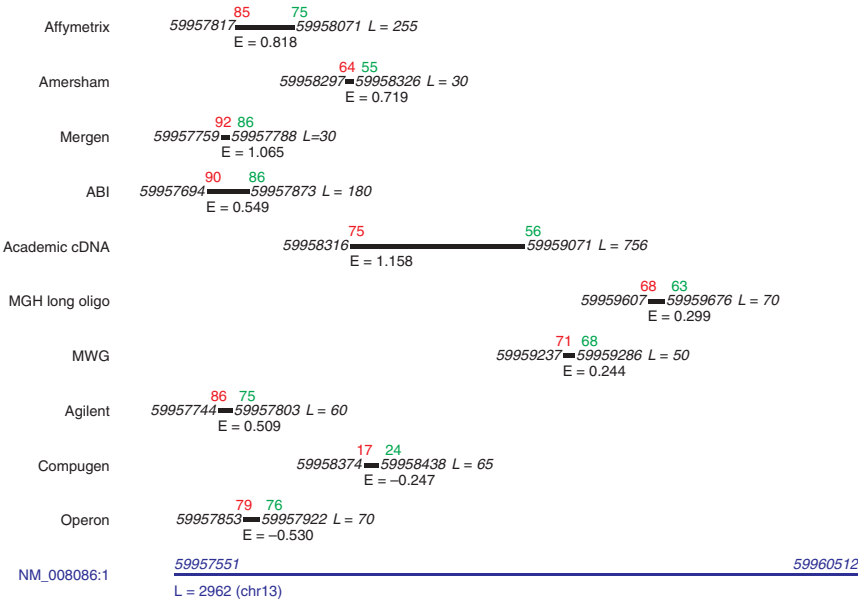

Figure 1 Cross-platform agreement of probes matched within one exon. For *Gas1* (RefSeq NM_008086, LocusLink 14451, UniGene Mm.22701), all ten platforms had probes that could be mapped completely within the boundaries of the first exon. The diagram shows the location of the probes from the different platforms. The complete exon is indicated at the bottom, with the 3′ end on the left hand side. The start and end positions on chromosome 13 are given just above the left-hand end and the right-hand end, respectively, of the bar representing the exon. The probes are indicated with black bars, flanked by the start and end coordinates, as given by the sequence alignments of the probes to the genome. The length of the alignment between the probe and the exon (L) is shown to the right of the probe bars. The relative gene expression (E) shown below each probe bar is the $\log_2$ ratio of mouse retina versus mouse cortex. The percentile transformed intensities from each platform are shown above the respective probe bars. For each platform, the number in red on the left-hand side is the intensity from mouse retina and the number in green on the right-hand side is the intensity from mouse cortex.

procedures are described in Methods. Amersham, Mergen and Compugen had the highest percentages (>72%) of probes that could be mapped completely within a single RSEXON. For the other platforms, <63% of the probes could be exon mapped. As expected, the longer cDNA probes had the lowest percentage of exon-mapped probes (8.7%).

Across platforms, the number of common genes decreased from UG, through LL and RS to RSEXON (see **Supplementary Table 2a** online). Only four RSEXONs were common across all ten platforms: NM_008086:1 (*Gas1*), NM_008686:1 (*Nfe211*), NM_018798:1 (*Ubqln2*) and NM_018871:1 (*Ywhag*). In general, when probe sequences were mapped within the same exon for a given gene, the expression measurements (both $\log_2$ ratios and intensities) were found to be very similar across the platforms, even though the probe sequences did not overlap (**Fig. 1**). Overall, data mapped through probe sequences (RS and RSEXON) showed higher inter-platform correlations than data mapped to UG and LL.

The Pearson correlation, as calculated from $\log_2$ ratios paired from two platforms, was reasonably good for all platforms (0.63–0.92), except for academic cDNA and partly for Compugen, whose correlations with other platforms were mixed. Spot quality filtering had a profound positive effect for all the inter-platform correlations, except for academic cDNA arrays (**Supplementary Table 2b** online). This table also shows that the correlations improved with increasing match stringency. For example, the Pearson correlations for the pair-wise comparison of Affymetrix and Amersham were 0.76, 0.76, 0.81 and

Table 2 Assessment of measurement deviation from pseudo-nominal values

	Outliers				Deviations				
	Unfiltered (1,690 genes)		Filtered (881 genes)			Distance from median		Distance from QRT-PCR	
	Counts	%	Counts	%	No. of genes	Mean	S.d.	Mean	S.d.
Affymetrix	284	19.6	127	18.7	149	0.27	0.32	1.65	1.28
Amersham	298	20.5	150	19.5	149	0.33	0.47	1.55	1.20
Mergen	331	21.3	164	21.7	136	0.42	0.68	1.61	1.30
ABI	413	27.9	219	28.8	137	0.39	0.63	1.53	1.16
cDNA	164	63.5	135	68.9	37	0.92	0.84	1.73	1.33
MGH	228	45.4	129	50.2	99	0.64	0.78	1.68	1.52
MWG	353	32.5	209	38.5	87	0.82	1.07	1.93	1.94
Agilent	304	27.0	223	31.1	148	0.50	0.77	1.66	1.11
Compugen	263	48.2	84	38.5	43	0.73	0.78	1.92	1.35
Operon	636	38.7	279	33.6	136	0.41	0.61	1.58	1.42

For each gene (RS) that had been represented on at least five platforms, an outlier $\log_2$ ratio was defined as a measurement that was outside the range defined by mean ± 1 s.d., as calculated from all $\log_2$ ratios for that gene. For each platform, we counted the number of times any of its measurements had been identified as an outlier $\log_2$ ratio. To indicate how often a given platform would be far away from the consensus measurement (as given by the cross-platform mean), the '%' columns show the outlier counts relative to the total number of genes on that platform included in the set of genes represented on at least five platforms. For a set of 153 genes (RS) validated by QRT-PCR, we used the median $\log_2$ ratios and the QRT-PCR $\log_2$ ratios as nominal measurements to assess magnitudes of deviations. The 'no. of genes' column shows, for each platform, the number of genes from this set represented on the given platform. For each choice of nominal value, the mean and s.d. are computed over all gene deviations for each platform.

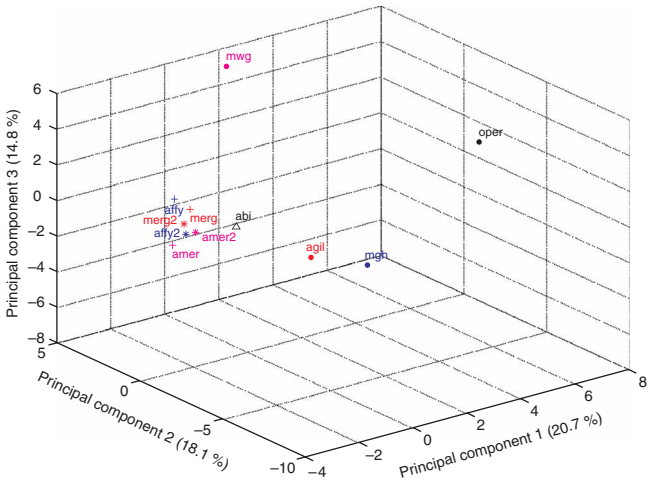

Figure 2 Cross-platform PCA plot. The plot illustrates PCA performed on $\log_2$ ratios corresponding to 130 RS identifiers common to eight of the ten platforms. As academic cDNA and Compugen had few RS identifiers in common with the other platforms, we chose to exclude these from this analysis to increase the number of genes applicable to PCA without missing values. For Affymetrix (affy2), Amersham (amer2) and Mergen (merg2) expression profiles obtained from a second laboratory were included in the analysis. The other expression profiles are labeled with abbreviations of the platform names used elsewhere. Each expression profile is plotted according to the first, second and third principal components. For each axis, the number in parenthesis gives the amount of variation (in percent of total) accounted for by the corresponding principal component.

0.85 for UG, LL, RS and RSEXON, respectively. It can also be seen that, for all matching criteria, the correlations were higher for comparisons within single-dye platforms than within two-dye platforms, although Agilent, MGH long oligo and Operon platforms were comparable to single-dye platforms. Moreover, for all platforms, permutation tests indicated that the correlations were highly significant for the RSEXON matched measurements (data not shown). Also s.d. of relative measurements improved with match stringency (see **Supplementary Table 2c** online).

We created CAT (Correspondence At the Top) plots based on all mapping options to assess and illustrate cross-platform agreement (see **Supplementary Fig. 1** online). One-dye platforms usually agreed with each other, whereas the two-dye platforms were more variable. In terms of outliers, the one-dye platforms usually performed better than the two-dye platforms in that they had fewer outliers when considering matched measurements (**Table 2**).

Principal component analysis (PCA) was used to illustrate the overall similarity of expression profiles. **Figure 2** shows the PCA plot based on the three first principal components calculated from 130 genes common to eight of the platforms. One-dye platforms were clustered together whereas two-dye platforms were more spread apart.

Inter-laboratory comparison

Data from three platforms, Affymetrix, Amersham and Mergen, were analyzed for cross-laboratory consistency. The intra-platform Pearson and Spearman correlations for intensities between laboratories were high for both samples ($r > 0.95$). For $\log_2$ ratios, Amersham had the highest cross-laboratory correlation (0.93), followed by Affymetrix (0.89) and Mergen (0.79). In contrast, the highest cross-platform Pearson correlation involving Amersham was 0.81 (Affymetrix versus Amersham, for RSEXON matched data), indicating that cross-laboratory variations are considerably smaller than cross-platform variations (see **Supplementary Table 2d** online). PCA focused on these three platforms showed that results from experiments on identical platforms conducted at different sites clustered much closer than measurements obtained from experiments on different platforms (see **Supplementary Fig. 2** online).

Quantitative biological validations

As an independent validation strategy, two methods using QRT-PCR were used to obtain RNA levels for a total of 160 unique genes. $\log_2$ ratios for 91 genes were obtained using TaqMan (see **Supplementary Table 3b** online) and for 74 genes using Universal ProbeLibrary (see **Supplementary Table 3c** online). As a replacement for a true gold-standard, we considered $\log_2$ ratios from QRT-PCR as nominal values and used the slope of the regression line of the $\log_2$ ratios from each microarray platform against QRT-PCR results as an accuracy measure to evaluate the platforms. By this statistic, ABI would be ranked the highest, followed by Affymetrix and Operon, whereas academic cDNA, MWG and Agilent were the lowest (see **Table 1** for accuracies using the TaqMan subset). These findings were confirmed by correlation coefficients on QRT-PCR data paired with data from the microarray platforms (data not shown). We observed slightly lower correlations for ProbeLibrary results than for TaqMan results when investigating the two subsets of QRT-PCR separately.

Overall, the measurements from most microarray platforms agreed well with QRT-PCR. However, the dynamic range for QRT-PCR was noticeably larger than that for the microarrays (**Fig. 3**). The median of all the microarray measurements had a Pearson correlation with QRT-PCR of 0.76, indicating reasonably good agreement. In terms of $\log_2$ ratio difference, one-dye platforms

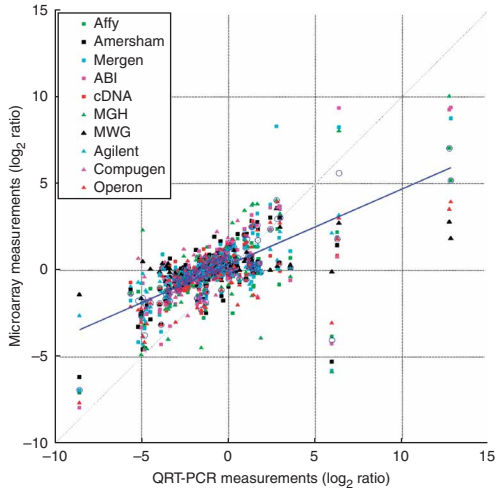

Figure 3 Scatter plot of QRT-PCR versus all microarrays. Log_2 ratios from the microarray platforms are plotted (y axis) versus the corresponding $\log_2$ ratios from QRT-PCR (x axis). From each platform, all $\log_2$ ratios based on probes that could be mapped to any of the 153 RS identifiers used in QRT-PCR were used. The regression line between the median microarray $\log_2$ ratios (over all platforms including a given gene) and the QRT-PCR $\log_2$ ratios is shown in blue. The slope of the line is 0.437, indicating a smaller dynamic range for the microarrays as compared to QRT-PCR. The Pearson correlation coefficient between the median measurements and QRT-PCR was 0.76 (P value 1.36 E-30), indicating relatively good, and highly significant correlation.

Table 3 Microarray measurements and QRT-PCR validation results for seven retina-related genes

	Gene						
	Rho, Mm.2965	*Neurod1*, Mm.4636	*Fgf3*, Mm.4947	*Crx*, Mm.8008	*Nr1*, Mm.20422	*Opn1sw*, Mm.56987	*Arr3*, Mm.95518
QRT-PCR	12.7	2.6	0	11.3	12.8	6.4	2.8
Affymetrix	8.5	3.7	0.6	5.6	5.2	5.6	NA
Amersham	–0.3	3.0	0.5	NA	NA	NA	NA
Mergen	9.2	3.5	0.6	8.8	8.8	8.2	NA
ABI	9.3	3.5	–0.1	6.6	9.4	NA	7.4
cDNA	5.6	–0.1	0.3	NA	1.3	NA	0.5
MGH	NA	3.3	1.2	9.5	NA	8.0	NA
MWG	NA	2.5	0	–0.1	1.8	2.7	NA
Agilent	0.3	2.2	0.2	NA	NA	3.2	NA
Compugen	NA	NA	1.3	NA	NA	NA	NA
Operon	3.3	2.7	2.5	3.8	4.0	3.0	NA

Seven genes previously verified to be highly expressed in the mouse retina were chosen for a detailed comparison of microarray versus QRT-PCR measurements. The table shows $\log_2$ ratios measured with QRT-PCR (ProbeLibrary) and the microarray platforms. Genes are named by official gene symbol and UniGene cluster number. Genes that were not represented on a given platform are listed with 'NA' in the corresponding cell. FGF3 is highly expressed in embryonic retina and not in adult retina[35,36].

had shorter distances compared to those of two-dye platforms. Within the two-dye platforms, Agilent and Operon had the shortest average distances (**Table 2**).

Most of the platforms had results consistent with QRT-PCR for genes of high and medium expression. For highly expressed genes, single-dye platforms had markedly better agreement with QRT-PCR than two-dye platforms did. In the case of genes with low expression, the agreement with QRT-PCR was much poorer for all platforms. For a subset of seven retina-related genes (confirmed to be highly expressed in retina from multiple studies using classical techniques), high expression in mouse retina versus mouse cortex was found in almost all platforms with probes for these genes and also confirmed by QRT-PCR (see **Table 3** and **Supplementary Table 3d** online).

DISCUSSION

In this study we compared gene expression data from ten different microarray platforms using identical samples. From the outset, we aimed to develop a sound and consistent framework for cross-platform comparisons. Our goals were to (i) provide unbiased results with clear metrics for performance evaluations using well-established analytical techniques, (ii) conduct the experiments for different platforms as systematically and as similarly as possible and (iii) allow inclusion of novel academic and commercial platforms as they develop. This is an ongoing comparative effort and we plan to include future platforms as they become available.

For platform comparisons, one would like to have two unlimited sources of RNA to ensure that identical samples are used on different platforms. For the general usefulness of the comparison, the RNA samples should be selected from a commonly used organism and should have a diverse set of transcripts covering a wide expression range. When this study was started, such sources were not available. We therefore chose two samples that had some of the above features. We extracted and pooled RNA from tissues of cortex and retina from the well-studied *Mus musculus*, selecting inbred mice to eliminate genetic variability. Cortex was chosen because brain tissues are generally considered to have broad expression profiles, whereas retina has some well-known tissue-specific transcripts[35,36]. Both tissue samples can be considered as replenishable sources of RNA with little variability between different pools, as demonstrated by laser-based capillary electrophoresis of labeled samples (data not shown). Mouse universal reference RNA sources have recently become commercially available, such as those from Ambion (http://www.ambion.com/catalog/CatNum.php?6050) and Stratagene (http://www.stratagene.com/manuals/740000.pdf) and will be considered for future studies.

We aimed to include all platforms claiming to perform whole-genome scale profiling of mouse mRNA that could provide probe sequence information. This set of hybridization-based platforms is considered homogenous relative to other gene expression technologies, for example, SAGE[37] and MPSS[38]. Each platform, however, may have a distinct set of laboratory or quality control features affecting the ease of inclusion for comparison purposes, including external spikes, alien probes and positive and negative controls. Such features were not present in all platforms and we chose to use internal controls where available. This may introduce biases in the comparisons that are not easily compensated for, but reflects the current usage of these platforms in laboratory environments.

Differences in technical and instrument choices, such as image analysis algorithms, make direct comparisons based on raw (intensity) signals impossible. We applied two transformations bringing the signal ranges to a uniform scale to compensate for differences in signal intensity ranges between platforms. This was found useful in comparing intra-platform variations. In spot quality filtering procedures, we chose to prioritize the quality flags generated by image analysis software according to recommendations from the platform vendors. Our results demonstrated that stringent spot quality filtering can improve data consistency, confirming reports of previous studies[22,39].

Despite efforts to optimize conditions for all experiments and analyses, we identified possible confounding factors biasing the results for three of the platforms: academic cDNA, Compugen and Agilent. The academic cDNA platform from our lab consistently performed poorest on all evaluations. We do not believe this reflects the quality of spotted cDNA arrays in general, as we found technical problems specific to our cDNA platform: (i) sequence-verification revealed that 15–20% of checked probes had false probe identity (about 50% of the probes were checked), and (ii) the amount of DNA varied from spot to spot. For Compugen, which also had low performance scores, one possible explanation is the relatively small number of probes, potentially causing a bias in the gene selection available for comparisons. For Agilent, a compression of the dynamic range was observed, possibly

due to suboptimal hybridization conditions (corrections have been made to Agilent's current protocols). Scanner saturation was observed for some experiments in some platforms. It is difficult to assess to what extent the limitations of scanner intensity ranges influenced the comparisons reported. However, a dual-scan procedure was tested for one platform having saturated spots but did not result in better agreement (data not shown). The above observations emphasize the need for careful design of cross-platform protocols and performance tuning throughout the execution of the experimental procedures.

Overall, the results based on different mapping strategies showed good agreement. The agreement between platforms on matched data tended to increase with increasing mapping specificity: UG, LL, RS, RSEXON. A possible interpretation is that the RefSeq mapping eliminates biases due to splice variants, being on the transcript level, and that the RSEXON mapping possibly forces the probes of different platforms to be more similar, as they are confined to a limited region of each gene. These hypotheses require further investigation. A systematic analysis of the effect of different probe designs was not performed, but could give more insights into the relative role of probe sequence versus other technological properties of the platforms.

In summary, the commercial platforms performed better than in-house platforms, both on internal consistency and agreement with other platforms. The performance of the one-dye platforms, Affymetrix, ABI and Amersham, was consistently among the best. The high internal consistency of Affymetrix and Amersham was also confirmed in the experiments conducted at a second laboratory. The observation that cross-laboratory variability using the same technology was lower than that of cross-platform variability confirmed results of other studies[7,40].

QRT-PCR, although commonly accepted as a gold standard for relative gene expression measurements[41], also has technical limitations and potential biases. Overall, the microarray results were in agreement with QRT-PCR for genes with medium and high expression, whereas there was little agreement for genes with lower or variable expression. We interpret this as stochastic variation appearing at low transcript numbers in both microarrays and validation procedures. We also found evidence for the importance of careful primer design when using QRT-PCR as the results from TaqMan were more consistent than those from Universal ProbeLibrary. For the former, primers had been designed to be on the same exon as the microarray probes. This was not enforced for the latter, where the primers were designed to be optimal for their kit using proprietary software. The differences in measurements of the two QRT-PCR methods, suggest that the use of QRT-PCR for biological validations must be carried out carefully.

In future studies, we will use other samples specifically selected to address biological and technical issues. A second pool of mouse retina RNA was collected to examine biological variability of the same sample across the arrays. Rat retina and yeast samples have been created to address issues related to cross-species specificity and cross-hybridization. The experimental design for two-dye platforms will also be extended to investigate dye swaps, self-self hybridizations and single-sample hybridizations. In addition to the platforms evaluated in this manuscript, data sets have been generated but not analyzed for these platforms: Agilix[42], Illumina[43], MPSS[38] and SAGE[37]. We will include these platforms in future comparisons.

The goal of this study was to illustrate a comparison framework that matched the transcripts at the sequence level. This is a first report of this relatively large-scale initiative in which the sequences of all probes were known. The results presented here indicate that there are many platforms available that provide good quality data, especially on highly expressed genes, and that between these platforms, there is generally good agreement. However, the results from different platforms vary substantially, both overall and for subsets of genes. Therefore, despite considerable developments toward standardization of gene expression profiling, many issues remain open for investigation.

METHODS

Sample collection and isolation. RNA samples used for all platforms were divided into aliquots from two pools of samples: C57/B6 adult mouse retina and Swiss-Webster postnatal day one (P1) mouse cortex. Mouse retina and mouse cortex were chosen because of their availability and biological interest. Mouse retina samples were obtained from a pool of C57/B6 mice (n=350) and mouse cortex were obtained from P1 Swiss-Webster mice (n=19), which were both purchased from Charles River Laboratories. The animal experiments were approved by the Institutional Animal Care Facility at Harvard University. The mouse cortex was used as a reference sample for the dual-dye platforms. The remaining total RNA from both samples was stored at –80 °C.

Labeling, hybridization and image processing. Sample preparation and hybridization steps were conducted following the protocols provided for each platform. Eight of the ten microarray platforms evaluated are currently commercially available: Affymetrix, Agilent, Applied Biosystems (ABI), Amersham (now GE Healthcare), Compugen (now Sigma-Genosys), Mergen, MWG BioTech (now Ocimum Biosolutions) and Operon. The remaining two platforms— academic cDNA and MGH long oligo arrays—are from academic laboratories. The cDNA arrays were provided by the Cepko laboratory, and comprised retinal and brain cDNAs from the Soares laboratory (University of Iowa, BMAP project (http://trans.nih.gov/bmap/), which were amplified and printed by J.R.S. and S.P. Oligonucleotides from both Compugen and Operon were purchased by the Division of Biology at California Institute of Technology and were printed together onto the same slide. A total of eight research laboratories were involved in this collaboration. Descriptive details of the different platforms and sites where the hybridizations were performed are shown in **Supplementary Data** online.

To evaluate cross-laboratory consistency, a subset of the platforms was conducted independently at a second laboratory using identical samples. This portion of the study is still ongoing, but results from Affymetrix, Amersham and Mergen platforms have already been completed and are reported here. The laboratories in which the hybridizations were conducted are shown in **Supplementary Data** online. Each laboratory provided the raw data sets and scanned images for analysis.

Six of the ten microarray platforms (Agilent, academic cDNA, Compugen, MGH long oligo, MWG and Operon) are considered to be two-dye platforms, as they require the hybridization of two samples, whereas the others (ABI, Affymetrix, Amersham and Mergen) are one-dye platforms. Because data based on a single array are often considered insufficient to obtain conclusive results[32], five replicates of each sample were used to assess the degree of variation in the expression data within each platform. The number five was chosen as a reasonable compromise between the wish to reduce the effect of array-to-array variability and resource limitations. The experimental design is shown in **Supplementary Data** online. A total of 91 hybridizations were completed and are reported in this manuscript.

All labeling and hybridization methods were completed as specified by each manufacturer's hybridization protocol. Image processing of the scanned images were conducted using the manufacturer's recommended scanners and settings. Detailed description of the protocols used for each platform is provided in **Supplementary Data** online.

Preprocessing of microarray data. Preprocessing methods included normalization, transformation and filtering. Specific normalization methods were chosen based on past microarray studies that have indicated their potential advantages over other methods in single and dual-dye platforms[44–46]. In the case of microarray data from single-dye platforms, normalization was performed using quantile normalization[46], where ten arrays (five for retina samples and five for cortex) were considered as one group. Data from two-dye platforms were normalized using Locally Weighted Scatterplot Smoothing (LOWESS) normalization[44,45]. Because probes from Compugen and Operon platforms were printed onto the same slide, LOWESS normalization was performed on the whole chip before they were separated and analyzed in the

study. We also examined and confirmed that when this normalization was performed for each platform independently, the results were similar (data not shown).

Data transformation included both linear and percentile scaling of the raw intensities, as well as $\log_2$ ratios between the two samples. The scaling transformations were needed to allow comparison of raw intensities quantified by different software packages. Linear scaling mapped the intensities of each slide/channel into a scale of 1 to 100, linearly and analogously. This method was used in measuring intra-platform coefficient of variations of the intensities. Percentile transformation projected the data to a hundred discrete levels (that is, 1 to 100) according to percentiles of the intensity values. Beyond making the measurements among various platforms comparable, percentile scaling may be useful to correct the artifacts introduced by different intensity distribution characteristics among various platforms, as well as to purposefully neglect some minor fluctuations in expression levels. Percentile transformation was mainly used in the inter-platform comparisons.

$\log_2$ ratios were computed to allow the comparison of single-dye and two-dye platforms. When we evaluated intra-platform variations, five $\log_2$ ratios were obtained from five technical replicates of each two-dye platform. For single-dye platforms, $\log_2$ ratios were obtained from five randomly paired arrays across samples without replacement. The averaged $\log_2$ ratios of technical replicates for each platform were used to assess inter-platform variation.

Stringent filtering for spot quality has been reported to improve consistency across different platforms[22,39]. The filtering criteria chosen in the study were either recommended by the vendors or have been broadly adopted by the research community. Filtering was conducted at the spot (image) level, taking into account both quality flags and signal-to-noise ratio thresholds. Ideally, all the platforms should have been scanned and quantified using the same scanner, with similar scanner settings. Because of the diversity of the technical approaches of the various platforms, different scanners were used, and this limited our ability to apply the same filtering criteria to all the platforms. In the case of the Affymetrix and Amersham platforms, probe set and spot quality flags were referenced, respectively. These meant only 'present' and 'good' calls were adopted, for Affymetrix and Amersham, respectively. The signal-to-noise ratio threshold of 3 was used for ABI, in addition to removal of flagged spots as recommended by the vendor. A signal-to-noise ratio threshold was set to 2 for Agilent, Compugen, Mergen and Operon platforms. For academic cDNA, MGH long oligo and MWG arrays, the images were scanned using GenePix software 3.0 (Molecular Devices). The software automatically generated flags at default settings for poor and missing spots, which were removed. The effects of filtering for each platform are detailed in **Supplementary Table 1c** online.

Mapping of genes across platforms. Gene mapping was conducted using annotation-based and sequence-based approaches. For the annotation-based approach, MatchMiner[47] was used to map UniGene (UG) clusters (UniGene Build 136) and LocusLink (LL) identifiers by using the GenBank accession numbers provided by each platform.

For the sequence-based approach, the February 2003 version of the mouse reference sequences (UCSC version mm3) was downloaded from the UCSC Genome Site (http://genome-archive.cse.ucsc.edu/goldenPath/mmFeb2003/bigZips/) and used for mapping the probe sequences. The probe sequences from each microarray platform were mapped to the mouse genome using the BLAT stand-alone program[48]. The sequence alignment results were also parsed so that only probe-to-exon matched pairs were extracted. Probe-to-exon meant only aligned sequences positioned completely within an exon were considered as a match. In the instances where multiple within-exon matches for a probe sequence occurred, the best match in terms of the length of 'hit' was selected. If no match was found, that probe was excluded. In this way, the probes from different platforms were matched both at the gene level by RS and on the exon level by RSEXON identifiers. The probe sequences used for mapping ABI and Affymetrix had lengths of 180 and 255 base pairs, respectively. Affymetrix uses 11 probe pairs to measure the expression level for each gene and the 255 base pairs correspond to the length of the sequences spanned by the 11 probe pairs (complete probe). The context sequences for Affymetrix were obtained from their NetAffx analysis center[49]. For ABI, the probe sequence for each gene on the array lies within the 180 base pairs used in the mapping.

In cases where there was more than one probe that matched to a particular identifier, the values were averaged. In most instances, however, each gene was represented by only one probe on all platforms.

Evaluation of intra-platform and inter-platform data consistency. We chose to measure data consistency by calculating CVs, correlation coefficients and standard deviations of the difference between measurements. PCA was performed to allow the display of axes corresponding to the largest variance in multiple platforms. Additionally, the degree of deviation of each platform from other platforms was quantified by defining outliers across various platforms' measurements for each gene.

The CV is defined as the variation among multiple measurements in proportion to their mean. We used CV to measure the reproducibility among multiple replicate experiments within each platform. Besides the conventional use of CV on channel-specific intensities, we also defined a segmental function for the CV of $\log_2$ ratios. When the mean of $\log_2$ ratios was between -1 and $+1$, the CV was equal to the s.d., otherwise, the conventional definition of CV was applied. This was to avoid including small denominators to distort the CVs considerably when a large proportion of probes having a mean of $\log_2$ ratio close to zero are expected in microarray experiments.

Pearson and Spearman correlation coefficients were calculated for both intra- and inter-platform comparisons. Intra-platform correlations consisted of computing the correlations for both linearly transformed intensities within each sample and their $\log_2$ ratios. For inter-platform comparisons, the correlations were calculated based on the averaged $\log_2$ ratios. As the expression data were not normally distributed (data not shown), we conducted two permutation tests on inter-platform correlations, aiming to estimate the significance of the correlation coefficients for the cross-platform probe matching. In both tests, for any pair of platforms, averaged log-ratios and paired measurements were randomly selected. The first test involved measurements chosen from the whole data set for the given platforms. In the second test, the measurements where chosen from a subset of data included in the list of matched probes between the platforms. In both cases, 10,000 randomized sets of matched measurements were created for each pair of platforms. Thus, empirical distributions of correlation coefficients were calculated and empirical confidence intervals of correlation coefficients were obtained to assess statistical significance.

The s.d. of the differences between matched measurements were computed as another measure of data consistency. In the case of intra-platform agreement, technical replicates were referred as 'matched measurements,' whereas each pair of platforms was considered in the case of inter-platform agreement. Annotation-based identifiers had a higher number of matched probes. Since there were very few overlaps across ten platforms, the relative measurements that were extracted for each matching option had to be observed in at least six platforms. Furthermore, among these, the four platforms ABI, Affymetrix, Agilent and Amersham were required to be represented, plus any additional two platforms.

PCA was performed on data from eight of the platforms, excluding the academic cDNA and Compugen platforms. In addition, PCA included second lab data from three one-dye platforms. PCA was conducted after standardization so that each gene has a zero mean and unit s.d.

To examine which platforms were more prone to have measurements that were markedly different from the others, we computed the frequency of outliers for each platform. For a given gene that has been measured in at least five platforms, if a platform's measurement lies outside of the range of the mean expression ratios $\pm$ 1 s.d., it was identified as an outlier.

The analyses were conducted using the R software environment (http://www.R-project.org/), BioConductor packages[50] and MATLAB (The MathWorks).

Biological validations. Molecular confirmation of microarray results is important when checking for consistencies of expression measurements across different platforms. We used the following criteria for selecting genes for validation: (i) genes should be present in at least six platforms, (ii) they should span the dynamic range and (iii) they should also include pairs with measurements that were in disagreement. Since the gene coverage varied across the platforms, we decided to select genes that were common across a minimum of six platforms based on RSEXONs. The six platforms had to include four

common microarray platforms: ABI, Affymetrix, Agilent and Amersham, and any additional two others. Based on this criterion, 399 genes were identified where two groups of genes were created based on their intensity. The groups were derived from the percentile-transformed data, where three categories of expression measurements were created: high (67–100 percentiles), medium (34–66 percentiles), and low (1–33 percentiles). The first group included genes that had combinations of high-high, high-medium and medium-medium expression measurements for both samples. In the second group, the expression measurements for both samples included genes that had combinations of high-low, medium-low and low-low expressions. A total of 158 genes were validated by QRT-PCR from these groups.

Biological validations for this study were conducted using QRT-PCR. Samples identical to the ones used for the microarray experiments were used for the biological validation step. The validation methods were conducted using ProbeLibrary, now Roche Universal ProbeLibrary, on two different Roche LightCyclers and TaqMan Gene Expression Assays on ABI PRISM 7900 HT Sequence Detection System (Applied Biosystems). The present version of the software allows exon-based primer designs, whereas we used a prior version having lower sensitivity. We verified 74 and 91 genes using Universal ProbeLibrary and ABI TaqMan Gene Expression Assays, respectively. Methods for both approaches are described in **Supplementary Data** online. The primer sequences for both Universal ProbeLibrary and TaqMan assay identifiers are provided as **Supplementary Table 3a** online.

A total of 165 genes were validated by QRT-PCR. Pearson correlation coefficients were computed for the $\log_2$ ratios for the set of genes validated by QRT-PCR and the corresponding platform. Expression ratios measured by QRT-PCR were calculated as follows:

$$\log_2 ratio(MR/MC) = -(\overline{Ct}_{MR} - \overline{Ct}_{MC})$$

where $\overline{Ct}_{MR}$ and $\overline{Ct}_{MC}$ correspond to the mean cycle thresholds for mouse retina and mouse cortex, respectively.

Accession numbers. The microarray data for the manuscript has been submitted to GEO OmniBus. The series record number is GSE4854.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We would like to thank vendors, Applied Biosystems, GE Healthcare, and Mergen for providing and running microarrays as part of this large-scale evaluation. In addition, we would like to thank Applied Biosystems for running TaqMan assays and Exiqon for supplying us with the ProbeLibrary kit as well as Roche Diagnostics for allowing us to use their 480 LightCycler. We thank Robert A. Greenes for reviewing the manuscript. W.P.K. was supported by the National Institutes of Health (NIH) EY014466 grant and by the Bioinformatics Division of the Harvard Center for Neurodegeneration and Repair. C.L.C. was supported by the Howard Hughes Medical Institute. F.L. and E.H. were supported by the functional genomics program (FUGE) in the Research council of Norway. G.M.C. was supported by NIH-NHGRI-CEGS. M.W.F., B.S. and G.F.S. were supported by Programs for Genomic Applications grants HL66678 and HL72358. R.B. was supported by NIH grants HL072370 and ES011387.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
2. Brazma, A. *et al.* ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
3. Ali-Seyed, M. *et al.* Cross-platform expression profiling demonstrates that SV40 small tumor antigen activates Notch, Hedgehog, and Wnt signaling in human cells. *BMC Cancer* **6**, 54–68 (2006).
4. Bammler, T. *et al.* Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods* **2**, 351–356 (2005).
5. Barczak, A. *et al.* Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.* **13**, 1775–1785 (2003).
6. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B. & Pavlidis, P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* **33**, 5914–5923 (2005).
7. Irizarry, R.A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).
8. Kothapalli, R., Yoder, S.J., Mane, S. & Loughran, T.P., Jr. Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22–32 (2002).
9. Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. & Kohane, I.S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
10. Larkin, J.E., Frank, B.C., Gavras, H., Sultana, R. & Quackenbush, J. Independence and reproducibility across microarray platforms. *Nat. Methods* **2**, 337–344 (2005).
11. Lee, J.K. *et al.* Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. *Genome Biol.* **4**, R82–94 (2003).
12. Li, J., Pankratz, M. & Johnson, J.A. Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol. Sci.* **69**, 383–390 (2002).
13. Mecham, B.H. *et al.* Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.* **32**, e74–82 (2004).
14. Park, P.J. *et al.* Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J. Biotechnol.* **112**, 225–245 (2004).
15. Parrish, M.L. *et al.* A microarray platform comparison for neuroscience applications. *J. Neurosci. Methods* **132**, 57–68 (2004).
16. Petersen, D. *et al.* Three microarray platforms: an analysis of their concordance in profiling gene expression. *BMC Genomics* **6**, 63–77 (2005).
17. Pylatuik, J.D. & Fobert, P.R. Comparison of transcript profiling on *Arabidopsis* microarray platform technologies. *Plant Mol. Biol.* **58**, 609–624 (2005).
18. Rogojina, A.T., Orr, W.E., Song, B.K. & Geisert, E.E., Jr. Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines. *Mol. Vis.* **9**, 482–496 (2003).
19. Schlingemann, J. *et al.* Patient-based cross-platform comparison of oligonucleotide microarray expression profiles. *Lab. Invest.* **85**, 1024–1039 (2005).
20. Shi, L. *et al.* Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6** suppl. Suppl. 2, S12–S26 (2005).
21. Tan, P.K. *et al.* Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676–5684 (2003).
22. Shippy, R. *et al.* Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* **5**, 61–76 (2004).
23. Walker, S.J., Wang, Y., Grant, K.A., Chan, F. & Hellmann, G.M. Long versus short oligonucleotide microarrays for the study of gene expression in nonhuman primates. *J. Neurosci. Methods* **152**, 179–189 (2005).
24. Wang, H., He, X., Band, M., Wilson, C. & Liu, L. A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics* **6**, 71–80 (2005).
25. Wang, H.Y. *et al.* Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biol.* **4**, R5–R18 (2003).
26. Warnat, P., Eils, R. & Brors, B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* **6**, 265–280 (2005).
27. Woo, Y. *et al.* A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J. Biomol. Tech.* **15**, 276–284 (2004).
28. Yauk, C.L., Berndt, M.L., Williams, A. & Douglas, G.R. Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* **32**, e124–e131 (2004).
29. Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J. & Sealfon, S.C. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30**, e48–e57 (2002).
30. Zhu, B., Ping, G., Shinohara, Y., Zhang, Y. & Baba, Y. Comparison of gene expression measurements from cDNA and 60-mer oligonucleotide microarrays. *Genomics* **85**, 657–665 (2005).
31. Sherlock, G. Of fish and chips. *Nat. Methods* **2**, 329–330 (2005).
32. Lee, M.L., Kuo, F.C., Whitmore, G.A. & Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* **97**, 9834–9839 (2000).
33. Mecham, B.H. *et al.* Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics* **18**, 308–315 (2004).
34. Carter, S.L., Eklund, A.C., Mecham, B.H., Kohane, I.S. & Szallasi, Z. 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**, 107–122 (2005).
35. Blackshaw, S., Fraioli, R.E., Furukawa, T. & Cepko, C.L. Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell* **107**, 579–589 (2001).
36. Blackshaw, S. *et al.* Genomic analysis of mouse retinal development. *PLoS Biol.* **2**, E247–E268 (2004).
37. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
38. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).

39. Pounds, S. & Cheng, C. Statistical development and evaluation of microarray gene expression data filters. *J. Comput. Biol.* **12**, 482–495 (2005).
40. Chu, T.M., Deng, S., Wolfinger, R., Paules, R.S. & Hamadeh, H.K. Cross-site comparison of gene expression data reveals high similarity. *Environ. Health Perspect.* **112**, 449–455 (2004).
41. Qin, L.X. *et al.* Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics* **7**, 23 (2006).
42. Roth, M.E. *et al.* Expression profiling using a hexamer-based universal microarray. *Nat. Biotechnol.* **22**, 418–426 (2004).
43. Gunderson, K.L. *et al.* Decoding randomly ordered DNA arrays. *Genome Res.* **14**, 870–877 (2004).
44. Workman, C. *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3**, research0048 (2002).
45. Berger, J.A. *et al.* Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics* **5**, 194–207 (2004).
46. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
47. Bussey, K.J. *et al.* MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.* **4**, R27–34 (2003).
48. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
49. Liu, G. *et al.* NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.* **31**, 82–86 (2003).
50. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80–R96 (2004).