

# MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues

Seth Blackshaw<sup>\*</sup>, Winston P Kuo<sup>†‡§</sup>, Peter J Park<sup>†</sup>, Motokazu Tsujikawa<sup>¶</sup>, Jenny M Gunnensen<sup>¥</sup>, Hamish S Scott<sup>¥</sup>, Wee-Ming Boon<sup>¥</sup>, Seong-Seng Tan<sup>¥</sup> and Constance L Cepko<sup>\*</sup>

Addresses: <sup>\*</sup>Department of Genetics, Harvard Medical School and Howard Hughes Medical Institute, <sup>†</sup>Children's Hospital Informatics Program, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA. <sup>‡</sup>Decision Systems Group, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. <sup>§</sup>Division of Health Sciences and Technology, Harvard University and Massachusetts Institute of Technology, Cambridge, MA 02139, USA. <sup>¶</sup>Department of Ophthalmology, Harvard Medical School/MEEI, 243 Charles Street, Boston, MA 02114, USA. <sup>¥</sup>Brain Development Laboratory, Howard Florey Institute, University of Melbourne, Parkville, Victoria 3010, Australia.

Correspondence: Constance L Cepko. E-mail: cepko@genetics.med.harvard.edu

Published: 18 February 2003

*Genome Biology* 2003, 4:R17

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/4/3/R17>

Received: 19 July 2002

Revised: 9 October 2002

Accepted: 2 January 2003

© 2003 Blackshaw et al.; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

## Abstract

**Background:** Serial analysis of gene expression using small amounts of starting material (microSAGE) has not yet been conclusively shown to be representative, reproducible or accurate.

**Results:** We show that microSAGE is highly representative, reproducible and accurate, but that pronounced differences in gene expression are seen between tissue samples taken from different individuals.

**Conclusions:** MicroSAGE is a reliable method of comprehensively profiling differences in gene expression among samples, but care should be taken in generalizing results obtained from libraries constructed from tissue obtained from different individuals and/or processed or stored differently.

## Background

Serial analysis of gene expression (SAGE) has been used to profile mRNA levels in a variety of organisms and systems [1-5]. Recently, several modifications have been made to the protocol, to create a protocol referred to as microSAGE [6-10], which allows the generation of SAGE libraries from small quantities of starting material, as little as 100,000 cells or 1 µg total RNA. These modifications include the direct binding of mRNA to magnetic oligo(dT) beads before the reverse transcription step, and a great reduction in the quantity of linkers used to generate ditags. While it has been

shown that conventional SAGE is representative and accurate [2], the same has not been clearly shown for microSAGE. Moreover, the extent to which data generated by either SAGE or microSAGE are reproducible, in the sense of giving identical tag representation from duplicate samples, has not been thoroughly addressed. Given the time and expense involved in constructing and sequencing SAGE libraries, concerns about how reproducible SAGE data are in general, as well as how representative and accurate microSAGE data are in particular, may deter potential users from pursuing this approach.

We have addressed these questions by constructing and analyzing a series of microSAGE libraries. We made libraries from the same mRNA samples to examine the variability due to library construction, and made sublibraries from identical pools of ditags to examine the variability due to the last steps in SAGE library construction, as well as due to sampling. We have also made and analyzed libraries from tissue obtained from both age-matched and non-age-matched individuals to examine the variability due to tissue preparation as well as individual differences in gene expression. We found that microSAGE data, as previously shown for conventional SAGE data [2], are highly accurate in the sense that they reflect known tissue and developmental gene-expression profiles. We found that tag distribution is virtually identical in samples constructed from either identical mRNA samples or ditag ligation reactions. However, we found large variations in tag distributions between both age-matched and non-age-matched human peripheral retinal libraries. We even saw a relatively large variation in tag distributions between libraries made from tissue pooled from small numbers of individual neonatal mice, a fact that may reflect developmental asynchrony as well as baseline individual variation in gene expression. This fact has implications for any expression-profiling approach, and suggests that in order to average out individual variations in gene expression, many different samples of a tissue of interest will have to be examined.

## Results

### MicroSAGE data are representative and accurate

To determine whether microSAGE data are representative, we conducted a virtual Rot analysis. A virtual Rot analysis, much like a Rot analysis generated by measurement of mRNA reassociation kinetics, examines the relative fraction of total mRNA made up of transcripts of high, medium or low abundance. It is constructed from SAGE data by cumulatively plotting the total fraction of all tags represented by tags of a given abundance level. Virtual Rot analyses of adult mouse whole retina and adult hypothalamus showed a distribution that closely reflects an actual Rot analysis of mouse brain RNA (Figure 1a-c) [11].

To determine whether microSAGE data can provide an accurate measure of mRNA abundance - that is, give tag levels that match levels of mRNAs for which abundance levels have previously been measured - we examined our adult mouse retinal library for tags corresponding to several such mRNAs. Although no absolute quantification of mRNAs in mouse retina has been made, measurement of rhodopsin, beta-tubulin and interphotoreceptor retinol-binding protein (IRBP) mRNAs in adult rat retina indicate that these represent 1.25-2.5%, 0.2% and 0.1-0.2% of total RNA, respectively [12]. These figures are based on the observation that mRNA consists of roughly 2% of total RNA in both rat and mouse retina (S.B. and C.L.C., unpublished work). In the mouse, we found that rhodopsin represented 2.1%, beta-tubulin 0.08%, and IRBP 0.13% of total

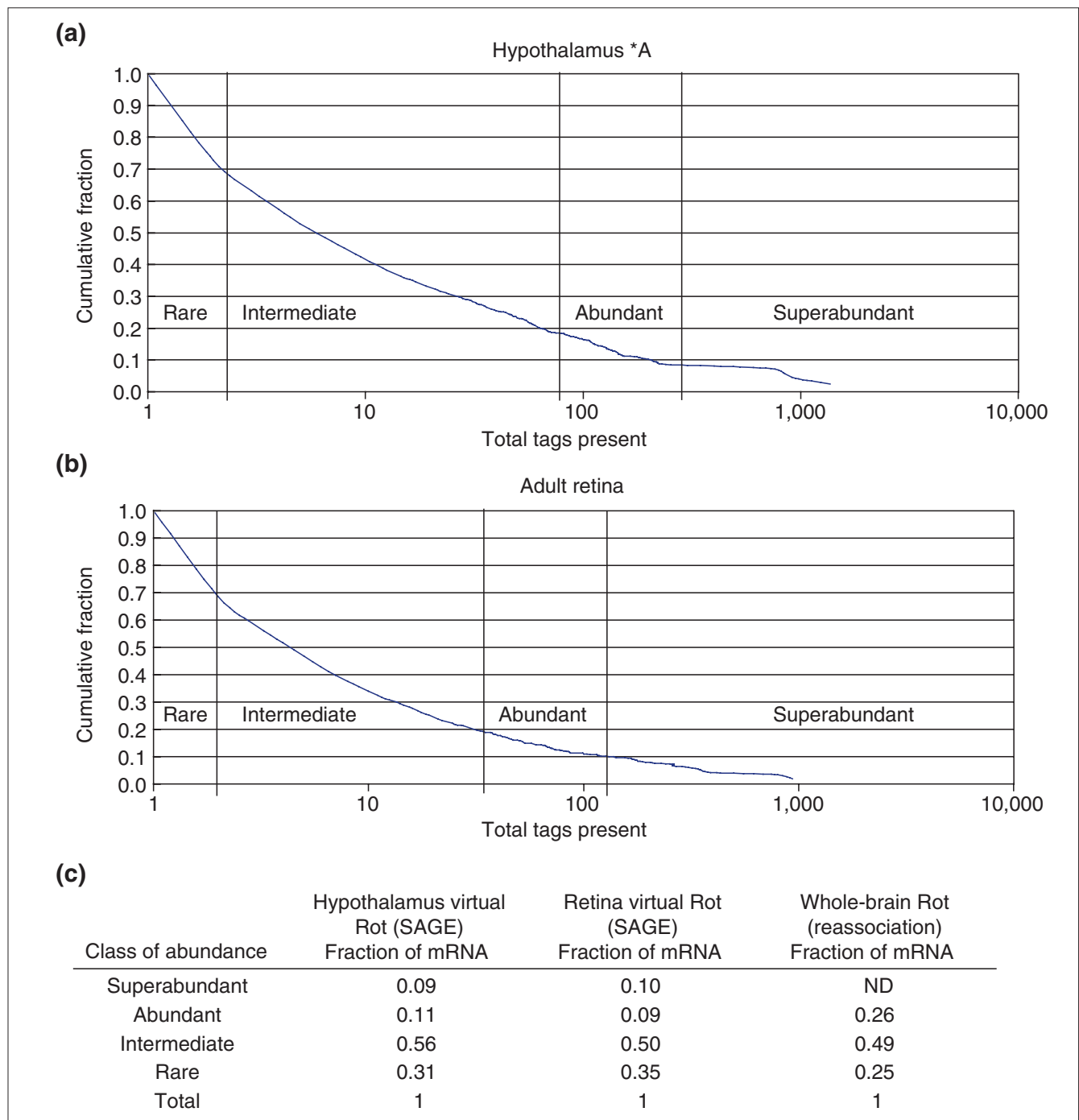
SAGE tags, values that on the whole approximate levels reported for rat. The slight discrepancy in beta-tubulin levels between rat and mouse may reflect species variability in gene-expression levels, or the failure to identify all alternative 3' ends of the beta-tubulin message by microSAGE.

Additional analyses to determine the accuracy of microSAGE data were carried out by examining a series of libraries from mouse retina and hypothalamus. We have constructed SAGE libraries from mouse hypothalamus; total wild-type mouse retina at 2-day intervals from embryonic day (E) 12.5 through to postnatal day (P) 6.5; P10.5 wild-type mice along with littermates mutant for the *crx* gene (which encodes a Paired type homeodomain transcription factor that directly regulates transcription of many photoreceptor-enriched genes); adult mouse retina; and microdissected outer nuclear layer, which consists entirely of rod and cone photoreceptors [11]. We examined this very large dataset for expression of genes known to be photoreceptor specific (Figure 2). We found that photoreceptor-specific genes typically showing high, retina-specific expression that began around one week after birth, coinciding exactly with the previously reported increase in expression of rhodopsin and other rod-specific genes [13-15]. We typically saw higher expression of photoreceptor-specific genes in microSAGE libraries made from microdissected outer nuclear layer (97% rods versus 70% rods for whole retina) and, in many cases, higher expression in P10.5 *crx*<sup>+/+</sup> samples relative to P10.5 *crx*<sup>-/-</sup> samples. These observations accurately reflect the known expression pattern of rod-specific genes [15]. Examining the temporal dataset for genes that are known to be selectively expressed in retinal progenitor cells (Figure 3), we found that high expression persisted until P2, followed by a rapid drop, with expression barely detectable in the adult. These data correlate well with the time course of mitosis within the developing retina [16,17].

### Cumulative distributions of *p*-values indicate the differences among microSAGE libraries

One way to measure the significance of the differences between libraries is to see the cumulative distribution function of the *p*-values. If the difference between two libraries is large, we would see a large number of SAGE tags with small *p*-values for that pair, compared to a pair in which the only source of variation is experimental variability in microSAGE library construction. The *p*-values we compute are the probability of observing the different tag counts for each gene when there is no differential expression. Although comparing the cumulative distributions does not match the libraries on a gene-by-gene basis, it gives an overall picture of the qualitative differences.

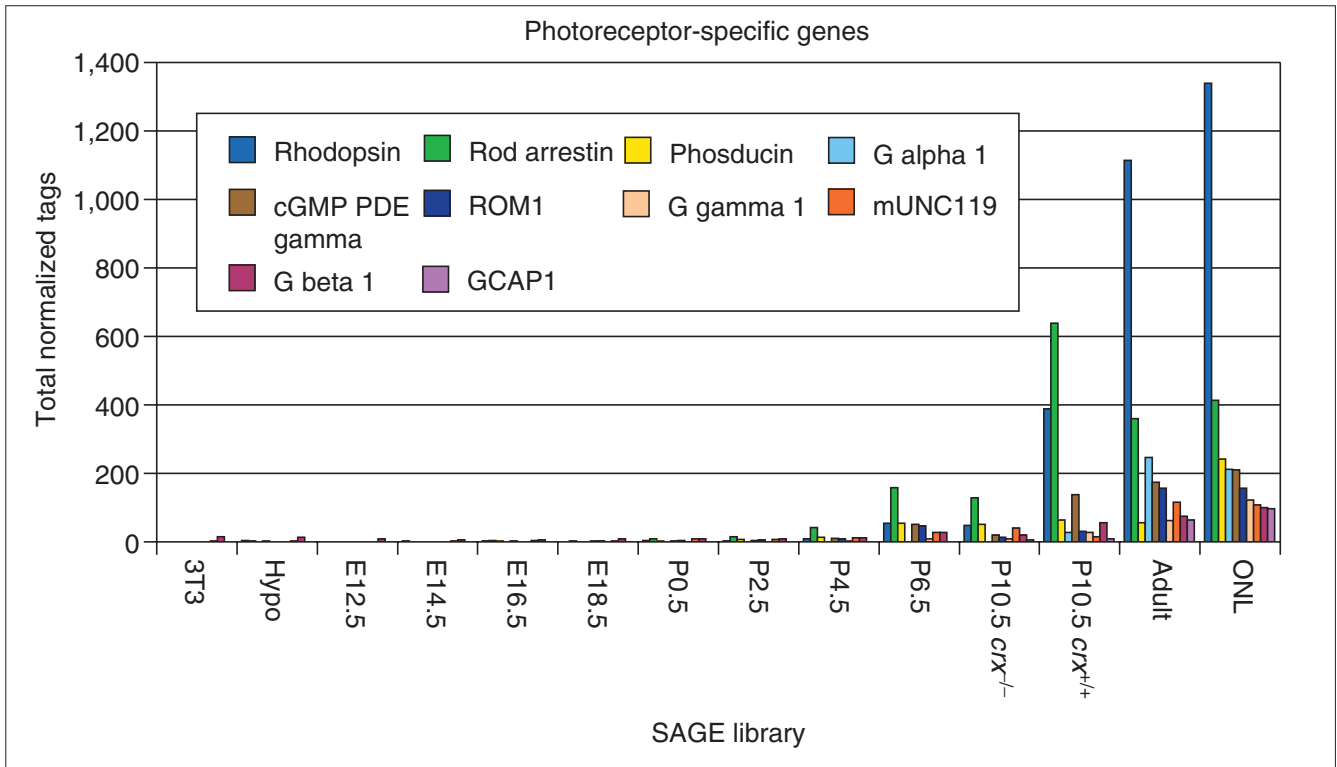
Cumulative *p*-value distributions for all libraries considered in this study are plotted in Figure 4. We investigated a variety of statistical methods of calculating *p*-values for differences in SAGE tag distribution in addition to the Monte Carlo analysis program included in the SAGE analysis software. These



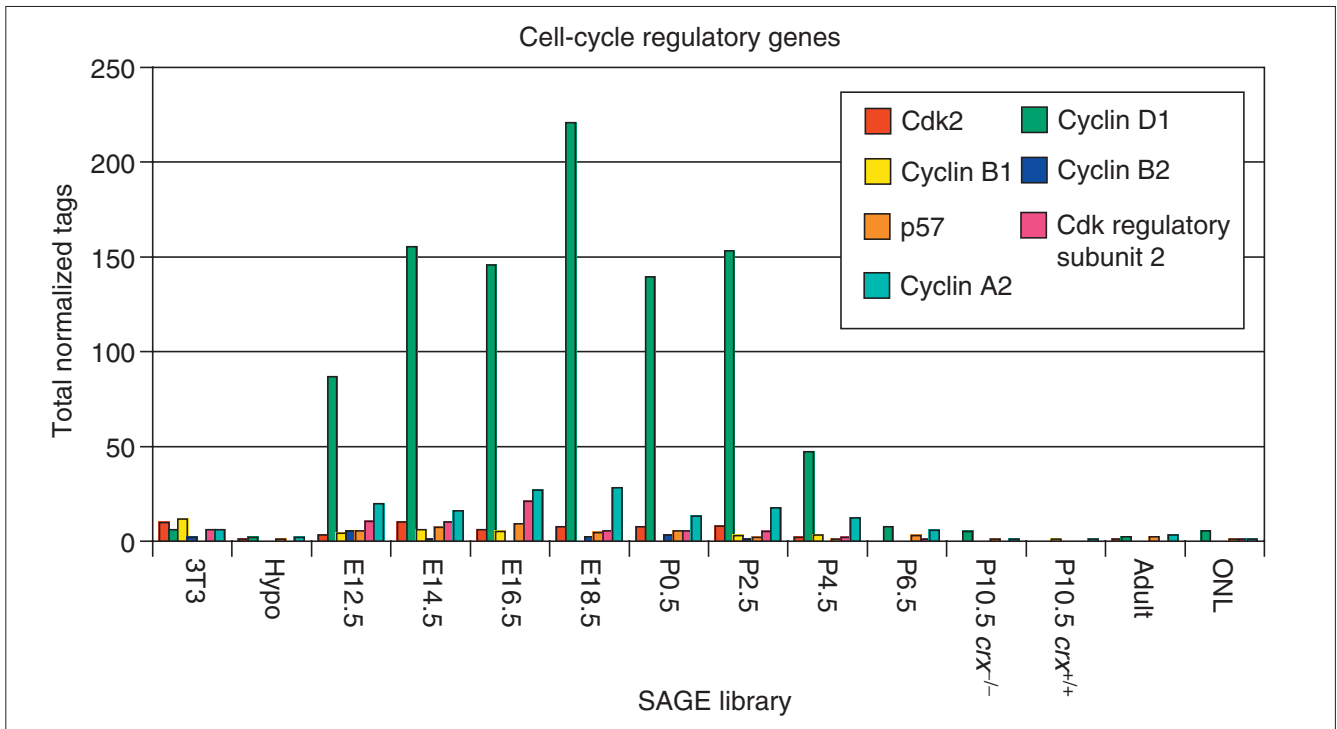
**Figure 1**  
 Virtual Rot analyses. **(a)** Virtual Rot analysis of hypothalamus library A. This plots the cumulative fraction of tags in the library against the total tags present for each tag-abundance level. The boundary between moderate- and low-abundance transcripts is set at the fraction of tags present only once in each library, as tag number (approximately 55,000) is not sufficient to comprehensively cover every cellular mRNA (around 300,000 molecules of mRNA/cell). **(b)** Virtual Rot analysis of adult retina library, carried out by the same method as (a). **(c)** mRNA abundance classes derived from the virtual Rot analysis in (a) and (b), compared to an actual Rot analysis of mouse whole-brain polysomal mRNA obtained from [11]. ND, not determined.

included a test for equality of proportions (equivalent to the chi-square test), modeling the sampling of tags using the Poisson distribution, Fisher's exact test, and a Bayesian

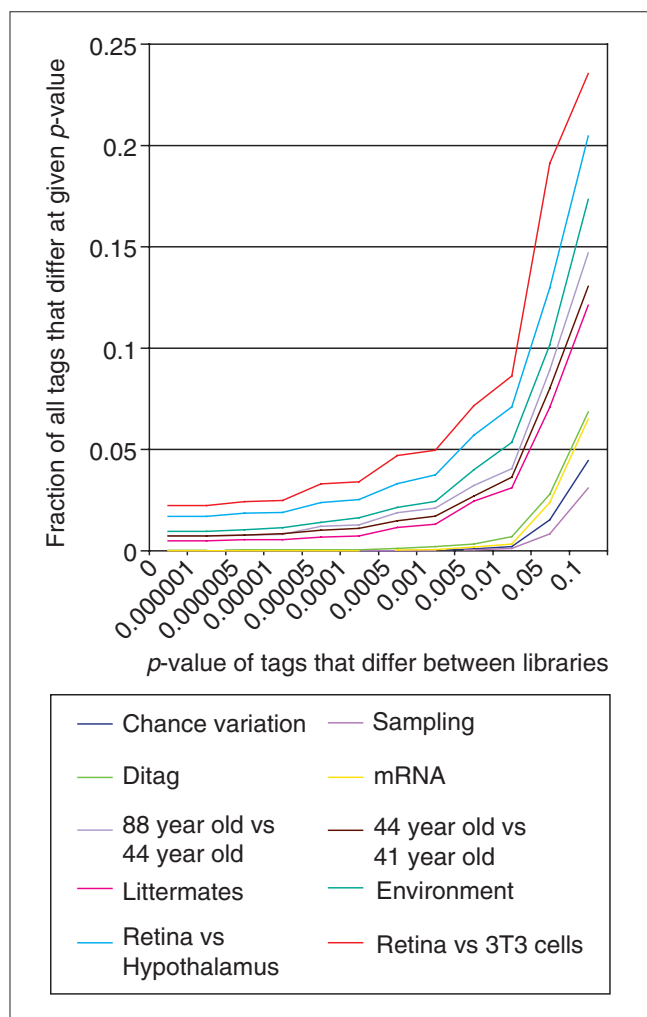
model. We found that all methods gave similar results for the range of *p*-values considered here (see Materials and methods for a more complete discussion of these comparisons).



**Figure 2**  
SAGE tag levels of photoreceptor-specific genes as measured by SAGE. Tag levels are normalized to 55,000.



**Figure 3**  
SAGE tag levels of cell-cycle regulatory genes as measured by SAGE. Tag levels are normalized to 55,000.



**Figure 4**  
 Cumulative *p*-value distributions for each tag comparison made between SAGE libraries in the study. Monte Carlo analysis using binomial distribution was used to generate *p*-values for both observed and chance tag *p*-values. The software for carrying out this analysis is included in the SAGE 3.0.1 software package [18]. Differences are expressed as a cumulative fraction of the total tags for which comparisons could be made, which in this case is all tags present more than once in both libraries combined. Chance variation, the cumulative *p*-value difference expected for chance variation in sampling (values computed for comparison of SAGE libraries ONL1 vs ONL2 - the plot of chance cumulative *p*-value differences is essentially the same for each library comparison considered); sampling, ONL1 vs ONL2 (same library, different tags); ditag, P2.5A vs P2.5B (same ditag ligation, different libraries); mRNA, *crx*<sup>+/+</sup>A vs *crx*<sup>+/+</sup>B (same starting mRNA, different libraries); 44-year-old vs 41-year-old, libraries from peripheral retina from different similarly aged males; 88-year-old vs 44-year-old, libraries from peripheral retina from individuals who differ in age and sex; littermates, P6.5A vs P6.5B (libraries made from retinas of three P6.5 C57B/6 mice vs libraries made from retinas of three littermates); environment, hypothalamus A vs hypothalamus B (libraries made from hypothalami of 3 8-week-old male C57B/6 mice housed in Boston, MA vs libraries made from hypothalami of 20 8-week-old male C57B/6 mice housed in Melbourne, Australia); retina vs hypothalamus, libraries from adult male C57B/6 mouse retina vs adult male C57B/6 mouse hypothalamus; retina vs 3T3 cells, libraries from adult male C57B/6 mouse retina vs mouse 3T3 fibroblast cell line.

**Visualization of the data in rotated coordinates reveals the relationship between average count and fold changes**

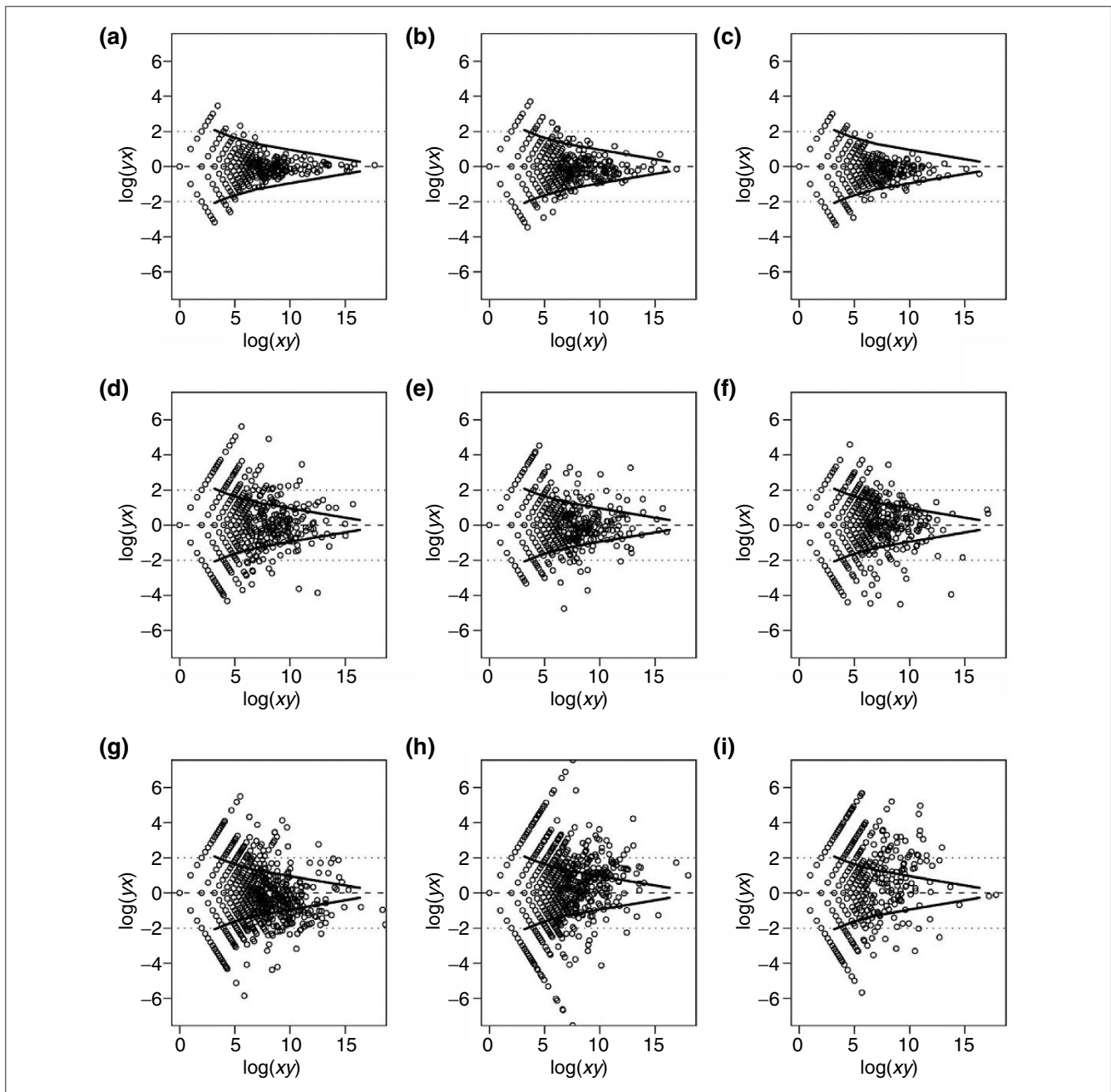
We find it convenient to view the data in a log scale in rotated coordinates, going from the original coordinates (*x*, *y*) to the new coordinate system ( $\log_2(xy)$ ,  $\log_2(y/x)$ ). This transformation corresponds to a 45-degree rotation and a linear stretching of the log-scaled data ( $\log_2(x)$ ,  $\log_2(y)$ ). This type of plot has been popular for microarray data, with the *x*- and *y*-axes corresponding to the mean intensity of the signals and their fold ratio, respectively, up to scaling constants [18,19]. For tag counts, the axes become the sum of the log counts and the fold change, respectively, with no rescaling necessary. This makes it easier to visualize the data, with horizontal lines rather than diagonal lines corresponding to fold changes. We can easily identify those genes with large fold changes and see the relationship between the average count and the fold change. We note that although it is advantageous to work in a log scale for the fold-change interpretation, it means that those genes with tag count of zero in either library are deleted from the picture. Another way of viewing the data is to plot differences in the two counts, rather than their fold ratios, as a function of the mean log count. Both ways contain the same information, and one can be derived from the other.

To gauge the baseline variability arising from library construction in the microSAGE process, we consider the two libraries made from the same mRNA sample (*crxA* versus *crxB*). In Figure 5c, we plot the data along with a ‘confidence band’ corresponding to two standard deviations away from the center at each average log count. (This band was computed using a locally weighted smoothing technique based on the standard deviations in the neighborhood of each point.) As we expect [20], the fold changes due to noise are higher at lower tag counts. This confidence band can serve as a rough measure of the variability in the system due to experimental procedures. The same confidence band from this comparison is superimposed on Figure 5a-i to see how the variability present in comparisons of different libraries compares to variation seen in libraries made from the same starting material. See the additional data files for a full list of tag counts in each of the SAGE library comparisons in the Results section.

**MicroSAGE data are highly reproducible**

To examine microSAGE data for reproducibility, several analyses were carried out. As a baseline gauge of variability in tag sampling, we compared the collection of tags comprising the first 26,000 tags obtained from the microdissected outer nuclear layer library to the collection of tags comprising the second 26,000 tags obtained from this library. A scatter plot of tag abundance showed a correlation coefficient of 0.976 when all tags were considered or 0.819 when only tags present less than 100 times in total in both libraries were examined (Table 1). A plot of cumulative observed



**Figure 5**

Examination of tag levels and fold-change in each of the SAGE library comparisons considered in Figure 4. The data is plotted in a log scale in rotated coordinates, where  $\log_2(xy)$  serves to indicate tag levels in the two libraries (where  $x$  = tag levels in SAGE library A, and  $y$  = tag levels in SAGE library B) and  $\log_2(y/x)$  serves to indicate fold changes in tag levels. Tags with a count of zero in either library were omitted from the analysis. The dark lines indicate the two standard deviation confidence limit calculated for the distribution of tags in Figure 5c (that is, same starting mRNA, different SAGE libraries) using a locally weighted smoothing technique based on the standard deviations in the neighborhood of each point. This is done to allow ready comparison of the observed tag differences between SAGE libraries to the baseline level of variability expected from SAGE library construction.

(a) Sampling, ONL1 vs ONL2 (same library, different tags); (b) ditag, P2.5A vs P2.5B (same ditag ligation, different libraries); (c) mRNA,  $crx^{+/+}A$  vs  $crx^{+/+}B$  (same starting mRNA, different libraries). (d) 44-year-old vs 41-year-old, libraries from peripheral retina from different similar-aged males; (e) 88-year-old vs 44-year-old, libraries from peripheral retina from individuals that differ in age and sex; (f) littermates, P6.5A vs P6.5B (libraries made from retinas of three P6.5 C57B/6 mice vs libraries made from retinas of three littermates); (g) environment, hypothalamus A vs hypothalamus B (libraries made from hypothalami of 3 8-week-old male C57B/6 mice housed in Boston, MA vs libraries made from hypothalami of 20 8-week-old male C57B/6 mice housed in Melbourne, Australia); (h) retina vs hypothalamus, libraries from adult male C57B/6 mouse retina vs adult male C57B/6 mouse hypothalamus; (i) retina vs 3T3 cells, libraries from adult male C57B/6 mouse retina vs mouse 3T3 fibroblast cell line. See Table 1 and Materials and methods for more information about these SAGE libraries.

*p*-values of tag differences in the samples normalized to total tag number showed no increase of observed variance over expected variance (Figure 4), and no high-abundance tags fell outside the confidence band of two standard deviations plotted for comparisons of libraries made from identical starting material (Figure 5a). P2.5 retinal libraries made from the same ditag ligation, but different large-scale PCR preparations, showed a correlation coefficient of 0.938 when all tags were considered, little difference between observed and expected *p*-values of variance (Figure 4), and showed relatively little difference between abundant tags (Figure 5b). Two libraries made from the same postnatal day 10.5 *crx*<sup>+/+</sup> mRNA samples (that is, the RNA sample was divided into two tubes at the onset of the reverse-transcription step) showed a correlation coefficient of 0.939, showed very little difference between observed and chance cumulative *p*-values of variance (Figure 4), and showed little difference between abundant tags (Figure 5c). These data show that microSAGE data are highly reproducible when identical starting material is used to generate libraries (see Table 2 for a full list of libraries and comparisons). Although the amount of starting material used to generate each of the libraries varied somewhat, we did not find any substantial changes in the fraction of duplicate ditags found in each library, which we have found to be more or less constant (approximately 5%) when more than 1 μg total RNA is used (S.B. and C.L.C., data not shown).

The correlation coefficients, particularly for the self-self comparison (ONL1 versus ONL2) when only tags present less than 100 times in total were considered, may seem somewhat low, compared to correlation coefficients computed for microarray data [21-23]. However, any digital gene-expression dataset shows a certain amount of noise due simply to sampling variability, particularly for lower-abundance transcripts [20]. Even if two tags are actually

found at identical levels in two libraries, the observed values of those tags will show a Poisson distribution about the actual values, thus resulting in a lower observed correlation coefficient. This is particularly true for lower-abundance transcripts, and is clearly seen in the substantially lower correlation coefficients that are observed when very high-abundant transcripts are excluded from analysis (see Table 1). Multiple replicate experiments, which are typically carried out in microarray studies, might filter out some of this noise, but such replicates are not feasible given the expense and time required to generate microSAGE data. In any case, the absolute *r*-values are not as important as the relative changes in the correlation coefficients observed as a result of experimental or sample variability when considering the data presented here.

**MicroSAGE data show pronounced differences in gene expression between age and sex-matched samples**

As the above data indicate that the microSAGE protocol does not introduce differences in libraries made from identical starting material, it is possible to interpret the gene-expression differences that might be found among individuals. To begin to probe such individual variations, we analyzed libraries from individual human retinal samples and from genotypically identical mice from the same litter, and genotypically identical mice raised in different animal facilities. MicroSAGE libraries generated from freshly dissected peripheral retina from similarly aged adult male humans gave a correlation coefficient of 0.673 when all tags were considered, a substantial difference between observed and chance *p*-values of variance between the samples (Figure 4), and many tags lie outside the confidence band plotted for comparisons of libraries made from identical starting material (Figure 5d). These differences were substantially similar to differences in expression seen between peripheral retinal

**Table 1**

Correlation coefficients calculated for all SAGE tags and for tags present at less than 100 times in total in both libraries examined

Comparisons made	Type of control	Correlation coefficient (all tags)	Correlation coefficient (tags present <100 times only)
ONL1 vs ONL2	Sampling	0.976	0.819
P2.5A vs P2.5B	Library construction from ditag stage	0.938	0.804
<i>Crx</i> <sup>+/+</sup> A vs <i>Crx</i> <sup>+/+</sup> B	Library construction from mRNA stage	0.939	0.824
41-year-old vs 44-year-old	Individual/environmental variation	0.673	0.580
44-year-old vs 88-year-old	Individual variation/age- and sex-dependent gene expression	0.760	0.628
P6.5A vs P6.5B	Library construction from tissue preparation stage/individual variation	0.661	0.578
Hypothalamus A vs hypothalamus B	Environmental/individual variation	0.782	0.500
Adult retina A vs hypothalamus A	Tissue-specific gene expression	0.763	0.349
Adult retina B vs 3T3 cells	Tissue-specific gene expression	0.479	0.182

**Table 2****Summary of all libraries used for statistical analysis and correlation coefficients for comparisons made between SAGE libraries**

Library name	Source	Tag number
ONL1	Adult mouse retina outer nuclear layer	26,016
ONL2	Same starting library as ONL1, but represents the second 26,000 tags sequenced in that library	25,956
P2.5A	P2.5 C57B/6 whole-mouse retina	23,525
P2.5B	Same ditag ligation as P2.5A was used to make the library, but ditag ligation stored at -20°C for 3 months before generation of library	23,566
Crx <sup>+/+</sup> A	P10.5 whole retina from Crx <sup>+/+</sup> mice (hybrid of C57B/6 and Sv129)	17,441
Crx <sup>+/+</sup> B	Same mRNA sample as in Crx <sup>+/+</sup> A was used to make the library	17,422
41-year-old peripheral retina	41-year-old Japanese male	23,662
44-year-old peripheral retina	44-year-old American caucasian male	23,758
88-year-old peripheral retina	88-year-old American caucasian female	23,773
P6.5A	P6.5 whole retina from three different C57B/6 littermate mice. Tissue stored at -80°C for 1 week before library preparation	25,466
P6.5B	P6.5 whole retina from three different C57B/6 littermate mice, from same litter as for P6.5A. Tissue stored at -80°C for 6 months before library preparation	25,473
Hypothalamus A	8-week-old male C57B/6 mouse hypothalamus from three different individuals. Animals housed at facilities in Boston, MA	48,978
Hypothalamus B	8-week-old male C57B/6 mouse hypothalamus from 20 different individuals. Animals housed at facilities in Melbourne, Australia	48,953
Adult whole retina A	8-week-old male total C57B/6 mouse retina	48,964
Adult whole retina B	8-week-old male total C57B/6 mouse retina (same library as adult total retina A)	28,593
3T3 cells	Mouse 3T3 fibroblast cell line	28,531

libraries obtained from a 44-year-old and an 88-year-old individual, which showed a correlation coefficient of 0.760 when all tags were considered, a large difference between observed and expected *p*-values of variance (Figure 4), and large number of differences in high-abundance tags (Figure 5e).

Retinas from one P6.5 mouse litter were used to generate two SAGE libraries, with retinas from three P6.5 littermates combined to generate one library (P6.5A) and three different P6.5 littermates combined to make another (P6.5B). Many differences in gene expression were observed, and an overall correlation coefficient of 0.661 was seen when all tags were considered (Figures 4,5f). Some of these differences appear to be in known rod-specific genes, and may reflect developmental asynchrony, as this is the time at which high-level expression of rod-specific genes is first observed, although most of the observed differences do not appear to be found in genes that show dramatic variation in expression through development. We also examined two hypothalamus libraries; each constructed from age-matched male mice of the same strain, obtained from animals raised in different facilities. These

samples showed a correlation coefficient of 0.782 when all tags were considered or 0.500 when highly abundant tags were excluded, a high value for cumulative *p*-value differences (Figure 4), and large numbers of tag differences (Figure 5g), suggesting that pronounced variation in gene expression is found even in genetically identical age- and sex-matched individuals housed in similar laboratory environments.

To understand the significance of these individual differences, we also compared disparate tissues. We compared two CNS tissues, the adult retina and hypothalamus, which showed a correlation coefficient of 0.763 when all tags were considered or 0.349 when highly abundant tags were excluded (Figure 5h). We also compared a CNS tissue and a non-CNS tissue, adult retina and 3T3 fibroblasts, which showed a correlation coefficient of 0.479 when all tags were considered or 0.182 when highly abundant tags were excluded (Figure 5i). Examination of the *p*-value plots (Figure 4) reveals many more tags that differ at low *p*-values in the CNS and non-CNS tissue comparison than in the comparison of the two CNS tissues. Likewise, plots of *p*-value differences versus tag abundance (Figure 5h-i) show



substantially more variability in the CNS versus peripheral tissue comparison. See additional data files for a full list of tag counts in each of the SAGE library comparisons discussed here.

## Discussion

### MicroSAGE is highly representative, reproducible and accurate

We found that tag abundance data measured by microSAGE reliably reflect mRNA abundance profiles, despite the fact that nearly 1,000-fold less starting mRNA was used here than is used to construct libraries in conventional SAGE. Tag distribution as determined by microSAGE closely mirrors profiles seen by Rot analysis of brain tissue. Likewise, tag abundance levels in libraries made from identical mRNA samples and ditag ligations showed very high similarity. The fact that the tag abundance levels differ slightly more in the libraries made from the same ditag ligation than in those made from the same mRNA sample may reflect a small amount of denaturation of AT-rich ditags during storage of the ligation reaction, as has been previously reported [24]. Overall, microSAGE passes all the tests for representativity, reproducibility and accuracy used previously for conventional SAGE [2].

### Differences in gene expression were observed among samples of the same tissues from different individuals

The high reliability of microSAGE data revealed in these analyses allows us to report the first comprehensive profiles of levels of difference in gene expression observed among tissues obtained from age- and sex-matched individuals. The variability among the samples observed does not appear to be due to experimental variability in library construction. Thus, it is due either to differences in the processing of the tissues used to generate the libraries - in variations in mRNA stability between tissue collection and mRNA purification or in heterogeneous tissue dissection - or to real variations in gene expression among individuals.

Although no studies to our knowledge have addressed the question directly, the limited body of data on expression profiling of age- and sex-matched humans generally reports significant variation in gene expression among individuals [25-27], although little of these data are publicly available. On the other hand, the literature on inter-individual variation in gene expression in inbred mice is both limited and somewhat inconsistent. While many studies have examined expression in inbred mice, the extent of variation in gene expression in age- and sex-matched individuals is rarely addressed. Published microarray experiments using inbred mice have been typically conducted using pooled samples from anywhere from 6 [28] to 17 [29] individuals. Likewise, for other methods of gene-expression profiling such as differential display and its variations [30] or SAGE [31], most previously published work has either been focused on samples from

large numbers of pooled individuals, or has considered individual mice that were not strain, age or sex matched.

Our results agree quite well with several recent reports in which inter-individual variation in gene expression is explicitly examined. A recent study reported a high degree of variation in RNA levels in rat hippocampus taken from different animals. A large fraction of candidate genes tested by quantitative reverse transcription PCR (RT-PCR) were found to vary significantly, even among individuals from inbred strains [32]. Another study reported analysis of inter-individual variations in gene expression in liver, testis and kidney of male C57B/6 mice using a 5000-spot microarray. This study found that 0.8% (liver), 1.9% (testes) and 3.3% (kidney) of all genes differed between individuals at  $p \leq 0.05$  [33]. The data from kidney compare fairly well with our data in Figure 4 for littermate-littermate variation (5.5% of SAGE tags different at  $p \leq 0.05$ , after subtracting the 1.5% of tags different at  $p \leq 0.05$  in the ONL1 versus ONL2 sampling control) or even human-human variation (7-8% of SAGE tags different at  $p \leq 0.05$  following subtraction of sampling variation). Because our libraries are all derived from neural tissue, our roughly twofold higher variability may reflect greater experience-dependent plasticity in the CNS. Several other microarray studies of inbred mice, while not specifically reporting inter-individual changes in expression, nonetheless reported considerable differences in both brain and peripheral tissues among different individuals [34,35].

There are, however, several published microarray experiments in which, although individual variation in gene expression was not explicitly addressed, variation in gene expression in age- and sex-matched individual mice or pooled samples of three or fewer mice was measured and found to be relatively small [36-39]. These include data from both peripheral tissue and brain that reported  $r$ -values for individual mice at around 0.98 [37,38], and one publicly available dataset of inter-individual comparisons of CNS tissues [38]. We analyzed this dataset and found it to have correlation coefficients in the 0.97-0.98 range (S.B., W.P.K. and C.L.C, data not shown). On the face of it, this seems at odds with our and other groups' data demonstrating considerable variation in gene expression - our  $r$ -values being in the 0.65-0.8 range for the developing retina and hypothalamus comparisons. None of these published data was from the hypothalamus or developing retina, so it is possible that the samples we analyzed show greater inter-individual variability. However, this sort of direct comparison between expression profiling methods may not be appropriate. Expression changes observed in microarray experiments can, in principle and in the hands of some investigators, accurately reflect expression-level changes observed by northern or quantitative RT-PCR data [40]. However, other investigators (including ourselves) have found that both oligonucleotide and cDNA microarray consistently underestimate true fold changes in mRNA levels as measured by

northern or RT-PCR [23,41]. This is most probably due to background hybridization to the array and signal normalization, and is not an issue for digital-expression profiling approaches such as SAGE. Underestimating differences in mRNA levels will give rise to higher observed  $r$ -values. While we cannot be sure that this was the case for the studies mentioned above, given the variation in microarray hybridization and scanning protocols among different investigators, it remains a distinct possibility. This concern, together with other recent studies that raise questions about the ability to accurately perform cross-platform comparisons of gene-expression datasets [42,43], prevents us from drawing firm conclusions about the true extent of inter-individual differences from the published microarray data. It will require SAGE and microarray analysis of identical starting material, coupled with northern or RT-PCR verification of genes that show inter-individual variation by either or both methods, to gauge exactly how accurately either method estimates inter-individual differences in gene expression. This, however, is beyond the scope of the current study.

#### Heterogeneous dissection cannot account for most of the inter-individual differences observed

The observed differences among microSAGE libraries made from identical tissue from different individuals could have resulted from heterogeneous tissue dissection. We addressed this question for our retinal samples. Because our P6.5 mouse retina libraries were made from total retina stripped of surrounding tissues (such as lens and retinal pigment epithelium), we determined whether tags corresponding to genes highly expressed in those tissues (such as crystallins, RPE65, RGR-like opsin) were found in these libraries, and determined that they were not expressed in either library, except occasionally at the single tag level (S.B. and C.L.C., data not shown). Moreover, inter-individual variation in cell-type composition among murine retinas appears to be minimal [44], and thus variable ratios of cell types in dissected retinas does not appear to account for the observed variation.

Rods represent a greater fraction of cells in the periphery of the human retina, while cones and ganglion cells represent a greater fraction of cells near the center. Observed differences in tag levels between the human peripheral retina libraries could be reflecting small differences in the radial distance from the fovea of the dissected region. We found, however, that although a number of the tags that differ by  $p \leq 0.005$  between the 44-year-old and 41-year-old individuals corresponded to known rod-enriched genes, the direction of change was not consistent between the two samples. Although more tags corresponding to known rod-enriched genes were at a higher level in the library from the 44-year-old individual, many other rod-enriched genes were more highly represented in the library from the 41-year-old individual. One of the tags that was higher in the 44-year-old individual corresponded to a known ganglion-cell-specific gene, whereas none of the tags that were more highly

expressed in the 41-year-old individual corresponded to ganglion-cell-specific genes (see Table 3 for a summary of this data). Moreover, when compared to data from our previously published work analyzing genes that show either peripheral or foveal enrichment in libraries taken from the same individual at  $p \leq 0.005$  [45], we likewise found no consistent pattern in the tags that differed between the 44-year-old and 41-year-old peripheral retinas (Table 3).

It is more difficult to rule out heterogeneous dissection as a variable in the differences seen in the hypothalamus libraries, particularly as tissue dissections were carried out by different investigators. Unlike in the retina, we lack good markers that are both highly expressed in adjacent tissues and excluded from the hypothalamus, and are thus unable to directly evaluate our hypothalamic libraries for contamination. The great majority of differences observed, however, seem to represent differences in very broadly expressed proteins, particularly ribosomal and mitochondrial proteins (S.B. and C.L.C., data not shown), suggesting that heterogeneous dissection is unlikely to account for most of the differences observed. However, we cannot rule out the possibility that some smaller fraction of the differences is due to heterogeneous dissection of the starting material for the two libraries.

#### RNA instability following tissue dissection does not account for most of the inter-individual differences observed

While major differences in gene expression were observed in libraries generated from different individuals, some of the samples used to generate these libraries differ in their storage conditions before processing. The interval between enucleation and freezing was 45 minutes for the 44-year-old and 88-year-old retinas, whereas the 41-year-old retina was kept at room temperature for 4 hours and then processed for RNA extraction. The tissue used to make the two P6.5 mouse

**Table 3**

#### Expression pattern of genes that differ in age- and sex-matched peripheral retinal libraries

Tags that differ at $p \leq 0.005$	Tags higher in 41-year-old	Tags higher in 44-year-old
Total	40	33
Known ganglion cell-enriched	0	1
Known rod-enriched	8	8
Higher in periphery	4	6
Higher in fovea	1	3

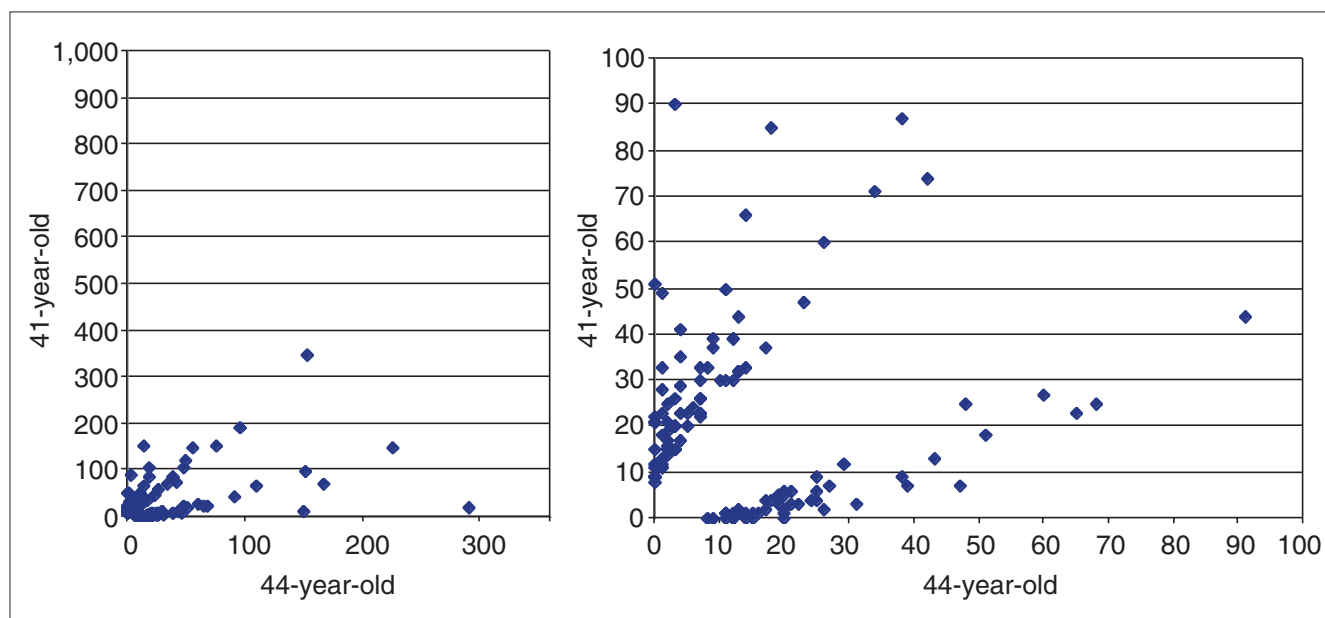
The table shows a breakdown of cell specificity (where known) of genes corresponding to all SAGE tags that differ at  $p \leq 0.005$  between the peripheral retina libraries from 44-year-old and 41-year-old individuals. In cases where cell specificity of the gene matching the tag is not known, the fact that the tag is enriched in peripheral retina or fovea (at  $p \leq 0.005$ ) is shown, based on previously published data [45].

libraries was harvested and frozen 5 minutes later, but the periods over which the two samples were stored at  $-80^{\circ}\text{C}$  were different - 1 week and 6 months, respectively. It is thus possible that differences in tag distribution among the libraries might represent differential mRNA stability following tissue collection. Several reports have indicated a high level of RNA stability for up to 48 hours postmortem in tissue kept at room temperature [46-49], and no reports to our knowledge have indicated instability of mRNA in tissue stored long term at  $-80^{\circ}\text{C}$ . However, as SAGE comprehensively profiles cellular mRNA expression, while previous studies examined only a handful of candidate transcripts [46-49], it could be that SAGE identifies unstable transcripts that are missed by a candidate-gene approach.

The correlation coefficient between tag abundance levels in the 44-year-old and 88-year-old retinal samples, which were both processed 45 minutes after enucleation, is substantially higher than the correlation coefficient between the retinal libraries from the 44-year-old and 41-year-old individuals. This is so despite the fact that the 44-year-old and 88-year-old individuals differ not only in age but also in sex. This could mean that a major portion of the observed variance might result from differential processing and/or storage of the tissue before mRNA preparation. However, comparison of expression of genes that differ by  $p \leq 0.005$  in the SAGE libraries from the peripheral retinas from the 44-year-old and the 41-year-old individuals indicates a roughly equal magnitude and direction of change of overall expression levels between the samples (Figure 6). Were the

differences between the samples largely due to differential mRNA stability, one would expect to see a subset of mRNAs that were much reduced in expression in the 41-year-old sample, as that one spent a longer period of time between enucleation and RNA extraction. One might also expect to see a fraction of more stable transcripts showing a small, compensatory increase in relative frequency, but this is not the case. The most likely explanation then for the differences between the 41-year-old and 44-year-old samples is variation in gene expression.

The observed variation between the two P6.5 mouse libraries was more surprising. As 3.8% of the tags that differ at  $p \leq 0.005$  represent known rod-specific genes (S.B. and C.L.C., data not shown), a small subset of these results probably partially reflect minor differences in developmental timing, and some of the other differences in tag abundance may also reflect variation in genes expressed in a dynamic manner during development. Other differences may reflect differences in mRNA stability during long-term frozen storage. However, as with the human samples, the direction of a roughly equal magnitude and direction of change of expression levels is seen in tags that differ at  $p \leq 0.005$  between the samples (Figure 7), suggesting that differential mRNA stability is not the cause of observed variation. This also suggests that for both the human-human and littermate-littermate comparisons, no major selective mRNA degradation occurs between tissue harvesting and binding of the mRNA to the magnetic oligo(dT)-coupled beads. The source of most of the variation between the samples thus



**Figure 6**  
SAGE tag abundance levels for tags in the 41-year-old peripheral retina SAGE library (retina kept at room temperature 5 h post-enucleation) and the 44-year-old peripheral retina library (retina kept at room temperature 45 min post-enucleation) for tags that differed at  $p \leq 0.005$ ; 139 such tags are shown.

remains unclear. One animal that was a developmental outlier is a possible explanation. For example, litters of inbred mice sometimes have a runt, suggesting a difference in rate of development and/or other factors. However, much of our data collected and analyzed for other purposes would suggest that such outliers are rare, at least with respect to eye development [13,23].

### Inter-individual differences in gene expression are most likely to account for the observed differences in tag distribution

A final source of the observed variability between individual samples probably reflects real individual differences in gene expression. It may partially reflect genetic polymorphisms that give rise to different *Nla*III sites and hence different SAGE tags, although this is unlikely to explain observed differences in gene expression between the inbred mice. The fact that the 41-year-old and 44-year-old human samples, as well as the murine hypothalamus samples, were both obtained from males implies that differential expression of X-inactivated polymorphic alleles cannot explain the observed differences. A fraction of the variation is likely to represent either differential environmentally induced patterns of gene expression or stochastic variation in gene expression that persists in the adult. In tissues such as the hypothalamus, where we see large differences in gene expression in age- and sex-matched mice of the same strain, this variability may reflect high levels of variability in behavior that are seen in inbred mice raised and tested under seemingly identical laboratory conditions [50].

Moreover, the high variability that is seen in both the development and gene-expression profiles of cloned mice [34,51], and the relatively low heritability of many quantitative traits such as lifespan or psychiatric illness that is observed in monozygotic twins [52,53], implies that there is substantial variability in gene expression even in genetically identical individuals, particularly in tissues of the CNS.

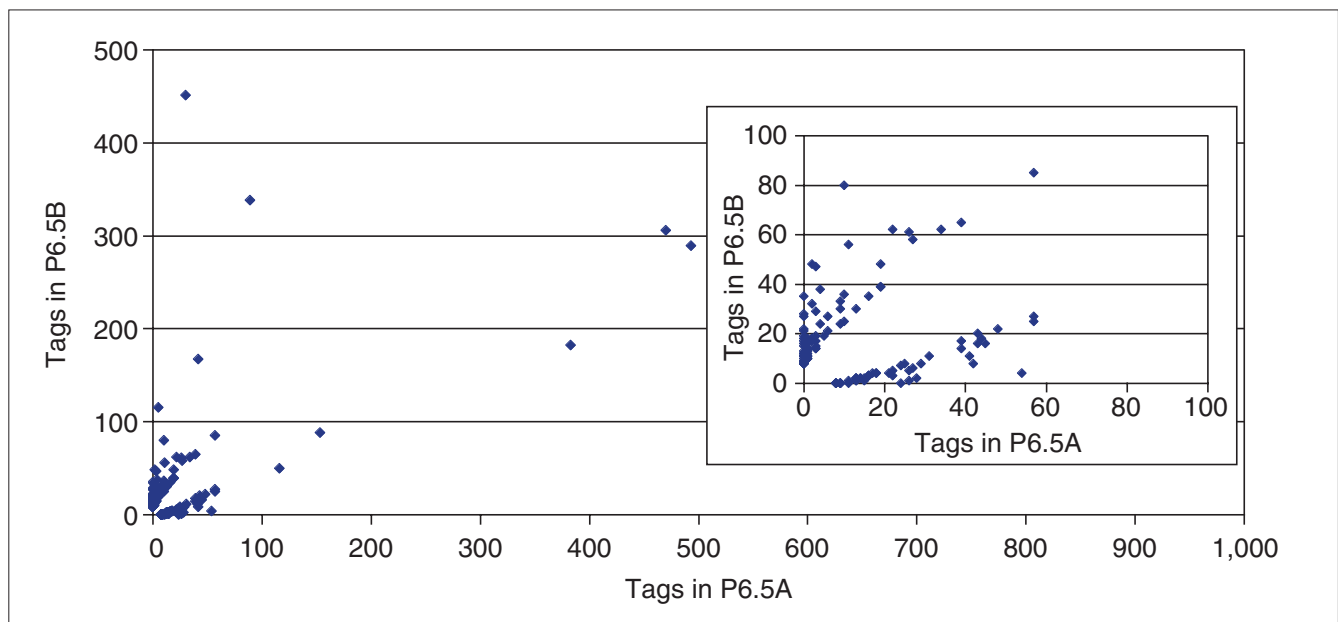
### Conclusions

We have shown that the microSAGE method, like conventional SAGE analysis, is highly accurate and representative. Furthermore, we have shown that it is highly reproducible when different libraries are made from the same starting mRNA samples. However, we find that considerable differences in gene expression are observed between samples of the same tissue taken from different individuals. These may reflect minor differences in the processing or storage of the samples, or may represent real differences in gene expression. Further experiments are required to resolve this issue. In any case, these data show that a measure of caution is required when comparing tag abundance levels in different SAGE libraries that were not processed identically or obtained from the same individual.

### Materials and methods

#### Isolation of mouse brain and retinal tissue

C57/B6 mice were used to obtain tissue from hypothalamus, adult retina and outer nuclear layer. The *crx*<sup>-/-</sup> and *crx*<sup>+/+</sup>



**Figure 7**

SAGE tag abundance levels for tags in the P6.5A SAGE library (generated from tissue stored for 1 week at  $-80^{\circ}\text{C}$ ) and the 44-year-old peripheral retina library (generated from tissue stored for 6 months at  $-80^{\circ}\text{C}$ ) for tags that differed at  $p \leq 0.005$ ; 135 such tags are considered.

littermates represent hybrids of C57B6 and Sv129 strains. Mice were killed between 2 p.m. and 4 p.m., and were housed on a 12:12 light:dark cycle. Microdissected outer nuclear layer tissue was obtained by flat-mounting freshly dissected retinas flattened by radial cuts onto nitrocellulose filters. The filters were then frozen at  $-80^{\circ}\text{C}$  in OCT mounting medium photoreceptor-side down. Orthogonal cryostat sections of  $20\ \mu\text{m}$  were cut onto Superfrost slides (Fisher), and the edge of the section containing the outer nuclear layer was hand dissected using a microscalpel and immediately frozen. Only sections containing retinas with excellent morphology were used for sample preparation. All libraries, with the exception of the P10.5  $crx^{-/-}$  and  $crx^{+/+}$  littermates were made immediately directly from tissue stored at  $-80^{\circ}\text{C}$ . The P10.5  $crx^{-/-}$  and  $crx^{+/+}$  littermate libraries were made directly from Trizol-purified total RNA and were the same starting material used in two previous studies from this lab [14,27]. Human tissue was obtained from patients with periocular malignancies, removal of which required enucleation of a healthy eye. The donors were an 88-year-old Caucasian female and 44-year-old Caucasian male (surgery conducted at Massachusetts Eye and Ear Hospital, Boston, MA) and a 41-year-old Japanese male (surgery conducted in Osaka, Japan). Human tissue was dissected and snap-frozen 45 min post-enucleation, and processed shortly afterwards. Deviations from this general schema are as follows: 41-year-old human peripheral retina: tissue spent 5 h at room temperature before RNA extraction; hypothalamus B, mice were housed in animal facilities at the Howard Florey Institute, University of Melbourne, Melbourne, Australia.

### Generation of SAGE libraries

A modification of the microSAGE method [10] was used to generate all SAGE libraries. In brief, 20–25 mg of tissue (containing 5–10 million cells) was used to generate each of the libraries with the exception of the P10.5  $crx^{-/-}$  and  $crx^{+/+}$  littermate libraries, for which 5  $\mu\text{g}$  total RNA was used, and the microdissected outer nuclear layer library, for which retinas from two different individuals were used. Six nanograms of each linker was used to generate the ditags and 28 cycles of amplification were used to obtain ditags from each of the samples, with the exception of the E12.5 library, which was amplified for 30 cycles, and the ONL library, which was amplified for 32 cycles. In the case of the P2.5B SAGE library, ditag amplification took place after the initial ditag ligation had been stored at  $-20^{\circ}\text{C}$  for 3 months. DNA sequencing was conducted on an ABI 3700. SAGE tag counts for each of the libraries: Figure 1, hypothalamus = 55,212, adult retina = 54,009; Figures 2,3, 3T3 = 28531 (obtained from [54]), hypothalamus = 55,212, E12.5 = 53,267, E14.5 = 54,884, E16.5 = 55,213, E18.5 = 56,779, P0.5 = 50,954, P2.5 = 58,140, P4.5 = 54,976, P6.5 = 59,979, P10.5  $crx^{-/-}$  = 53,827, P10.5  $crx^{+/+}$  = 52,016, adult retina = 54,009, adult outer nuclear layer = 52,207.

### Data analysis

We computed  $p$ -values for differences observed in SAGE tag levels using a binomial distribution-based Monte Carlo

analysis [55], which we found to be effectively equivalent to all other methods of calculating  $p$ -values we tested. All Monte Carlo analysis was conducted using SAGE 3.0.1, which was obtained from Ken Kinzler (Johns Hopkins University School of Medicine). 100,000 separate trials were run for each tag considered. Correlation coefficients were calculated for all SAGE tags in the libraries compared, and separately for tags present at  $< 100$  and  $> 1$  total in both libraries, using MS Excel.

A variety of methods have been proposed for calculating the  $p$ -values, and we examined the commonly used ones using the supposition that we have observed tag numbers  $x$  and  $y$  in libraries of size  $A$  and  $B$ .

#### Testing for equality of proportions

In this method, we want to see if the difference between the proportions  $p_1 = x/A$  and  $p_2 = y/B$  is significant [4]. Each tag here is viewed as a random variable following a binomial distribution, which can be approximated by a normal distribution in this case because the total count is large. The statistic of interest is

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{A} + \frac{1}{B}\right)}} \sim N(0,1), \text{ where } p = \frac{p_1A + p_2B}{A + B}.$$

We can compute the  $p$ -value or the confidence interval from this statistic. This test is conceptually simple and easy to implement. There is no need to perform Monte-Carlo simulations as done in the software by SAGE 3.0.1, as they seem to give similar results.

#### Poisson model

We can use Poisson distribution to model the sampling of tags, as the selection of a particular tag can be regarded as a random event with a small probability. Considering the probability of observing  $y$  occurrences of a clone already observed  $x$  times, we have the formula [20]

$$p(y|x) = \left(\frac{B}{A}\right)^y \frac{(x+y)!}{x!y! \left(1 + \frac{B}{A}\right)^{x+y+1}}$$

$p$ -values can be computed by summing up the appropriate tail of the distribution.

#### Fisher's exact test

When we think of the data as the  $2 \times 2$  table, the most common way to compute the  $p$ -value is the chi-square test, which can be shown to be equivalent to the equality of proportion test described above. For a small number of tag counts, Fisher's exact test may be used, which computes the exact probability of the observed  $2 \times 2$  table (containing  $x$ ,  $y$ ,



$A-x$ ,  $B-y$ ) based on a hypergeometric distribution assuming that the marginals are fixed.

#### Bayesian Model

We can also use a Bayesian model with a prior distribution for the likely fold changes [56]. The prior distribution is often specified using a beta function (for mathematical convenience) and the parameter is chosen to reflect the distribution of changes in expression levels between experimental conditions. This approach allows one to estimate the probability that the proportions differ by a given factor.

The first three of these methods are closely related. For example, the binomial distribution (related to the equality of proportions method, and which serves as the basis for the Monte Carlo analysis conducted by the SAGE 3.0.1 data analysis software) with small probability and large sample size converges to the Poisson distribution; Fisher's exact test should replace the chi-square test when the number of counts in the  $2 \times 2$  contingency table becomes very small ( $< 5$ , roughly speaking). For our data, we have found that all these methods give very similar results, and the difference in the  $p$ -values is noticeable only for the very small  $p$ -values (that is  $< 10^{-10}$ ).

#### Additional data files

All comparisons of SAGE libraries discussed in the text (see Table 2) are detailed in additional Excel files available with the online version of this paper as follows (the type of variation measured for each file is in brackets): ONL1 and ONL2 (sampling); P2.5A and P2.5B (library construction from ditag stage);  $crx^{+}/A$  vs  $crx^{+}/B$  (library construction from mRNA stage); 41-year-old peripheral retina and 44-year-old peripheral retina (individual/environmental variation); 44-year-old peripheral retina and 88-year-old peripheral retina (individual variation/age- and sex-dependent gene expression); P6.5A and P6.5B (library construction from tissue preparation stage/individual variation); adult retina A vs hypothalamus A (tissue-specific gene expression); Adult retina B vs 3t3 cells (tissue-specific gene expression). The type of variation measured (see Figure 4 and Table 1) is given in parentheses after the name of the library. For computing correlation coefficients and  $p$ -values, the SAGE linker-derived tags TCCC-CGTACA and TCCCTATTAA have been omitted from consideration. The data on the hypothalamus A and Hypothalamus B (environmental/individual variation) is freely available at [57].

#### Acknowledgements

We thank Dror Sharon and Ted Dryja for use of data from the 44-year-old and 88-year-old human retinal libraries, and George Church for helpful discussions. This work was supported through the Foundation for Retinal Research by the generosity of the Schwartz and Brint families and

by NIH grant EY0976. W.P.K. was funded through the Research Training in Health Informatics program by the National Library of Medicine, 5T15 LM07092-07. S.B. is a Life Sciences Research Foundation fellow, and C.L.C. is an investigator of the Howard Hughes Medical Institute.

#### References

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484-487.
2. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88**:243-251.
3. Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA: **Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*.** *Genome Res* 2001, **11**:1346-1352.
4. Kal AJ, van Zonneveld AJ, Benes V, van den Berg M, Koerkamp MG, Albermann K, Strack N, Ruijter JM, Richter A, Dujon B, et al.: **Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources.** *Mol Biol Cell* 1999, **10**:1859-1872.
5. Matsumura H, Nirasawa S, Terauchi R: **Technical advance: transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE).** *Plant J* 1999, **20**:719-726.
6. Datson NA, van der Perk-de Jong J, van den Berg MP, de Kloet ER, Vreugdenhil E: **MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue.** *Nucleic Acids Res* 1999, **27**:1300-1307.
7. Peters DG, Kassam AB, Yonas H, O'Hare EH, Ferrell RE, Brufsky AM: **Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite.** *Nucleic Acids Res* 1999, **27**:e39.
8. Ye SQ, Zhang LQ, Zheng F, Virgil D, Kwitrovich PO: **MiniSAGE: gene expression profiling using serial analysis of gene expression from 1 microgram total RNA.** *Anal Biochem* 2000, **287**:144-152.
9. Virlon B, Cheval L, Buhler JM, Billon E, Doucet A, Elalouf JM: **Serial microanalysis of renal transcriptomes.** *Proc Natl Acad Sci USA* 1999, **96**:15286-15291.
10. St Croix B, Rago C, Velculescu V, Traverso G, Romans KE, Montgomery E, Lal A, Riggins GJ, Lengauer C, Vogelstein B, Kinzler KW: **Genes expressed in human tumor endothelium.** *Science* 2000, **289**:1197-202.
11. Croizat B, Berthelot F, Felsani A, Gros F: **Complexity of polysomal polyadenylated RNA in mouse whole brain and cortex.** *FEBS Lett* 1979, **103**:138-143.
12. Treisman J, Morabito MA, Barnstable C: **Opsin expression in the rat retina is developmentally regulated by transcription activation.** *Mol Cell Biol* 1988, **8**:1570-1579.
13. Blackshaw S, Fraioli RE, Furukawa T, Cepko C: **Comprehensive analysis of photoreceptor gene expression and the identification of candidate retina disease genes.** *Cell* 2001, **107**:579-589.
14. Morrow EM, Belliveau MJ, Cepko CL: **Two phases of rod photoreceptor differentiation during rat retinal development.** *J Neurosci* 1998, **18**:3738-3748.
15. Morrow EM, Furukawa T, Cepko CL: **Vertebrate photoreceptor cell development and disease.** *Trends Cell Biol* 1998, **8**:353-358.
16. Alexiades MR, Cepko CL: **Quantitative analysis of proliferation and cell cycle length during development of the rat retina.** *Dev Dyn* 1996, **205**:293-307.
17. Young RW: **Cell proliferation during postnatal development of the retina in the mouse.** *Dev Brain Res* 1985, **21**:229-239.
18. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al.: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747-752.
19. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al.: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
20. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7**:986-995.
21. Lee ML, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci USA* 2000, **97**:9834-9839.

22. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Res* 2001, **29**:e41.
23. Livesey FJ, Furukawa T, Steffen MA, Church GM, Cepko CL: **Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene *Crx*.** *Curr Biol* 2000, **10**:301-310.
24. Margulies EH, Kardia SL, Innis JW: **Identification and prevention of a GC content bias in SAGE libraries.** *Nucleic Acids Res* 2001, **29**:e60.
25. Weindruch R, Kayo T, Lee CK, Prolla TA: **Gene expression profiling of aging using DNA microarrays.** *Mech Ageing Dev* 2002, **123**:177-193.
26. Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, Kunkel LM: **Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle.** *Proc Natl Acad Sci USA* 2002, **99**:15000-15005.
27. Yoshida S, Yashar BM, Hiriyanna S, Swaroop A: **Microarray analysis of gene expression in the aging human retina.** *Invest Ophthalmol Vis Sci* 2002, **43**:2554-2560.
28. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice.** *Genome Res* 2000, **10**:2022-2029.
29. Zirlinger M, Kreiman G, Anderson DJ: **Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei.** *Proc Natl Acad Sci USA* 2001, **98**:5270-5275.
30. Clinton M, Manson J, McBride D, Miele G: **Gene expression changes during murine postnatal brain development.** *Genome Biol* 2000, **1**:research0005.1-0005.11.
31. Chrast R, Scott HS, Papasavvas MP, Rossier C, Antonarakis ES, Barras C, Davisson MT, Schmidt C, Estivill X, Dierssen M, et al.: **The mouse brain transcriptome by SAGE: differences in gene expression between P30 brains of the partial trisomy 16 mouse model of Down syndrome (Ts65Dn) and normals.** *Genome Res* 2000, **10**:2006-2021.
32. Humpherys D, Eggan K, Akutsu H, Friedman A, Hochedlinger K, Yanagimachi R, Lander ES, Golub TR, Jaenisch R: **Abnormal gene expression in cloned mice derived from embryonic stem cell and cumulus cell nuclei.** *Proc Natl Acad Sci USA* 2002, **99**:12889-12894.
33. Tudor M, Akbarian S, Chen RZ, Jaenisch R: **Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain.** *Proc Natl Acad Sci USA* 2002, **99**:15536-15541.
34. Alfonso J, Pollevick GD, Castensson A, Jazin E, Frasch AC: **Analysis of gene expression in the rat hippocampus reveals high inter-individual variation in mRNA expression levels.** *J Neurosci Res* 2002, **67**:225-234.
35. Pritchard CC, Hsu L, Delrow J, Nelson PS: **Project normal: defining normal variance in mouse gene expression.** *Proc Natl Acad Sci USA* 2001, **98**:13266-13271.
36. Lee CK, Weindruch R, Prolla TA: **Gene-expression profile on the ageing brain in mice.** *Nat Genet* 2000, **25**:294-297.
37. Lee CK, Klopp RG, Weindruch R, Prolla TA: **Gene expression profile of aging and its retardation by caloric restriction.** *Science* 1999, **285**:1390-1393.
38. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C: **Regional and strain-specific gene expression mapping in the adult mouse brain.** *Proc Natl Acad Sci USA* 2000, **97**:11038-11043.
39. Rampon C, Jiang CH, Dong H, Tang YP, Lockhart DJ, Schultz PG, Tsien JZ, Hu Y: **Effects of environmental enrichment on gene expression in the brain.** *Proc Natl Acad Sci USA* 2000, **97**:12880-12884.
40. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Res* 2001, **29**:e41.
41. Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC: **Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays.** *Nucleic Acids Res* 2002, **30**:e48.
42. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18**:405-412.
43. Kothapalli R, Yoder SJ, Mane S, Loughran TP Jr: **Microarray results: how accurate are they?** *BMC Bioinformatics* 2002, **3**:22.
44. Jeon CJ, Strettoi E, Masland RH: **The major cell populations of the mouse retina.** *J Neurosci* 1998, **18**:8936-8946.
45. Sharon D, Blackshaw S, Cepko CL, Dryja TP: **Profile of the genes expressed in the human peripheral retina, macula, and retinal pigment epithelium determined through serial analysis of gene expression (SAGE).** *Proc Natl Acad Sci USA* 2002, **99**:315-320.
46. Johnson SA, Morgan DG, Finch CE: **Extensive postmortem stability of RNA from rat and human brain.** *J Neurosci Res* 1986, **16**:267-280.
47. Lukiw WJ, Wong L, McLachlan DR: **Cytoskeletal messenger RNA stability in human neocortex: studies in normal aging and in Alzheimer's disease.** *Int J Neurosci* 1990, **55**:81-88.
48. Marchuk L, Sciore P, Reno C, Frank CB, Hart DA: **Postmortem stability of total RNA isolated from rabbit ligament, tendon and cartilage.** *Biochim Biophys Acta* 1998, **1379**:171-177.
49. Castensson A, Emilsson L, Preece P, Jazin EE: **High-resolution quantification of specific mRNA levels in human brain autopsies and biopsies.** *Genome Res* 2000, **10**:1219-1229.
50. Crabbe JC, Wahlsten D, Dudek BC: **Genetics of mouse behavior: interactions with laboratory environment.** *Science* 1999, **284**:1670-1672.
51. Eggan K, Akutsu H, Loring J, Jackson-Grusby L, Klemm M, Rideout WM 3rd, Yanagimachi R, Jaenisch R: **Hybrid vigor, fetal overgrowth, and viability of mice derived by nuclear cloning and tetraploid embryo complementation.** *Proc Natl Acad Sci USA* 2001, **98**:6209-6214.
52. McGue M, Bouchard TJ Jr: **Genetic and environmental influences on human behavioral differences.** *Annu Rev Neurosci* 1998, **21**:1-24.
53. Cournil A, Kirkwood TB: **If you would live long, choose your parents well.** *Trends Genet* 2001, **17**:233-235.
54. **SAGE: serial analysis of gene expression** [<http://www.sagenet.org>]
55. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: **Gene expression profiles in normal and cancer cells.** *Science* 1997, **276**:1268-1272.
56. Chen H, Centola M, Altschul SF, Metzger H: **Characterization of gene expression in resting and activated mast cells.** *J Exp Med* 1998, **188**:1657-1668.
57. **Melbourne brain genome project** [[http://www.mbgproject.org/MBGP\\_Data\\_files.html](http://www.mbgproject.org/MBGP_Data_files.html)]