Microarrays as a Tool to Investigate the Biology of Aging: A Retrospective and a Look to the Future

Simon Melov and Alan Hubbard

(Published 20 October 2004)

The use of microarrays as a tool to investigate fundamental biological questions has become ubiquitous over the past several years. Microarrays are becoming as common as the polymerase chain reaction or any of the other tools in the molecular biologist's armory. Unlike experiments involving other tools, however, the design and analysis of microarray experiments present some unique problems to molecular biologists, problems with which statisticians have long been familiar. In this overview of microarrays and aging-related research, we will review selected highlights of microarray studies that have been carried out to study aging to date, as well as discuss some of the potential problems that routinely arise during these types of experiments, especially in the context of aging.

Microarray Platforms

Gene expression profiling using microarrays measures the expression level of thousands of genes simultaneously from RNA extracted from a tissue, a cell line, or in some cases an entire organism. The microarray platform broadly consists of two-color and one-color approaches (Fig. 1). Two-color approaches are usually based on differential fluorescent labeling of RNA derived from two samples, the experimental sample and the reference sample. The labeling is carried out with dyes of different spectral properties, for example Cy3 and Cy5 (1). The labeled RNAs from the two samples to be compared are cohybridized on a single microarray. This procedure allows the experimental sample to be compared with a control in the same experiment. The eventual quantitative output is calculated as a ratio value derived from the value of the reference (labeled, for example, with Cy3) over the value from the experimental sample (labeled with Cy5). Other dyes can also be employed for differential labeling, but Cy3 and Cy5 are currently the most often used. The microarray chip itself most frequently consists of a glass slide (substrate) spotted with DNA corresponding to many thousands of genes, either synthesized as synthetic oligomers (oligos) or cDNA-based (1). The DNA can represent the entire genome of the species in question, or it can represent a subset of genes of interest. The latter is termed a “sub-array” or “focused array” (2).

Alternatively, one-color approaches, in which all the RNA samples are labeled with the same dye, can encompass a wide variety of methods and platforms, all of which treat each sample as a discrete entity to be compared with the other samples in the experiment via bioinformatic approaches. Examples of one-color approaches include “macro-arrays,” which are nylon membranes spotted with genes and hybridized via traditional molecular approaches (3), or high-density oligo- or cDNA-based arrays on a glass matrix (1). This type of microarray is perhaps most commonly seen in chips produced for the Affymetrix (www.affymetrix.com) platform. “Affy” arrays are made by laying down oligos on the glass slide via photolithographic principles, a process analogous to that used in the computer industry (1). Other types of high-density arrays are made by ink-jet deposition of phosphoramidites on the substrate (as done by Agilent at www.agilent.com) or by tethering oligos to glass beads with unique identifiers, after which the tethered oligos are randomly placed on a glass matrix. This latter approach, pioneered by Illumina (www.illumina.com) facilitates the delivery of an extremely high density of genes to a single substrate and allows 30-fold redundancy of gene coverage per array, even for the entire human genome. This redundancy allows very accurate estimates of gene expression to be made, as a result of the high number of replicates per array, and this approach will do much to address issues of chip-to-chip variability in the future. All these approaches have various merits with regard to use for biological research, but in the absence of appropriate experimental design and statistical analysis, substantial risks for both false positive and false negative statistical inferences lie in wait for the unwary.

Study Design and Standards

Since the introduction of microarrays several years ago, there has been a steady improvement in the methodology as well as in what constitute acceptable criteria for publication (4). Arguably, the technology and the expertise to evaluate microarray data and experimental design are only now reaching critical mass. This idea is perhaps indicated by the recent publication of a set of standards for describing and reporting microarray experiments, although the guidelines do not explicitly describe minimum criteria for success or suggest valid approaches for analysis (5). Numerous journals now require authors who submit manuscripts that contain microarray data to adhere to these standards for publication (although many other journals still do not stipulate this requirement, it is likely that it will eventually become commonplace). In addition to the substantial improvements to the microarray methodology itself, there has recently been more emphasis on appropriate analysis of the data from the microarray experiment. Microarrays epitomize more than any other technique in molecular biology the saying originally developed for computing: “GIGO,” or “garbage in, garbage out.” Microarray experimental design requires attention to statistical issues in combination with the biological question, as well as more than a passing familiarity with the statistical issues of multiple testing (more on this point later). In reviewing the

Simon Melov is at the Buck Institute for Age Research, Novato, CA 94945, USA and Alan Hubbard is in the Division of Biostatistics at the University of California, Berkeley, Berkeley, CA 94720, USA. E-mail: smelov@buckinstitute.org (S.M.)
aging-related microarray literature for this article, we found very few studies published to date that analyze aging over the course of an organism’s life span with any real temporal resolution, which could be obtained, for example, by analyzing expression patterns for more than three or four time points over the life span being studied (Fig. 2). We were also struck by how well the published microarray studies on aging exemplify the development and maturation of microarray methodology and analytical techniques over the past several years. We have tried to select various highlights from the literature, rather than make this Review a comprehensive summary of every study claiming to study aging via microarray. Further, it is beyond the scope of this article to discuss the biological inferences made in each study, other than to describe a few key bullet points. One of the common goals of all the reports we review in this article was to identify a transcriptional profile unique to old animals, a molecular “fingerprint” of aging, which at the very least may be an excellent biomarker of aging (see Miller Perspective at http://sageke.sciencemag.org/cgi/content/full/2001/1/pe2).

Biomarkers, Aging, and the Microarray Paradigm

More than 20 years ago, one of the earliest searches for biomarkers of aging was initiated by the National Institute on Aging [NIA (http://www.nia.nih.gov/)] during a workshop entitled “Nonlethal Biological Markers of Physiological Aging” (6). These workshops continued over the next 20 years or so, in conjunction with requests from the NIA for proposals with the aim of discovering such biomarkers. A great deal of effort was put into identifying biomarkers that define physiological age, as opposed to chronological age, as well as how best to define aging itself (6). Unfortunately, despite substantial investment by the NIA, no biomarkers were discovered that could robustly identify physiological age or that were predictive of life span itself (6). Because calorie restriction prolongs life span in a variety of organisms, including mice (http://sageke.sciencemag.org/cgi/genedata/sagekeGdbIntrvn;9) (see Masoro Subfield History at http://sageke.sciencemag.org/cgi/content/full/2003/8/re2), much of the work concentrated on comparing calorically restricted animals to ad libitum-fed animals, in an attempt to define markers that scaled with the effects of the caloric restriction paradigm.

In the late 1990s, toward the end of this particular search for biomarkers of aging, microarrays were starting to become widely used in the biological sciences, principally in the field of cancer research. In this field, the uniformity of gene expression patterns within specific tumors was being exploited to give unique “molecular fingerprints” that allowed tumors to be differentiated from one another (7, 8). Microarray studies on tumors have now revealed biomarkers of cancer and are effective at differentiating tumor types, as well as whether or not tumors are responding to specific interventions (chemotherapies) (9–13). In principle, such studies can be seen as a research paradigm for the NIA-driven search for biomarkers of aging, especially because the cancer experiments led to the types of results that were desired by the NIA program. In cancer studies, tumors are now being classified, often via biopsy, and predictions are being made as to the probability for survival of the patient (either with or without chemotherapy) by means of the transcriptional profile generated by microarray analysis of RNA from the tumor (13). The hope is that this procedure will be
<table>
<thead>
<tr>
<th>Year published (reference)</th>
<th>Array platform</th>
<th>Number of genes; % of genome on chip</th>
<th>Normalization method</th>
<th>Technical replicates</th>
<th>Biological replicates</th>
<th>Number of arrays in study</th>
<th>Statistical analysis</th>
<th>Study design</th>
<th>Tissues, Pools, or Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999 (14)</td>
<td>Affymetrix</td>
<td>6347 ~ 21%</td>
<td>Affymetrix</td>
<td>-</td>
<td>3</td>
<td>9</td>
<td>Used ranking of DE between treatments</td>
<td>3 young vs 3 old vs 3 old CR</td>
<td>Gastrocnemius</td>
</tr>
<tr>
<td>2001 (3)</td>
<td>Nylon membranes</td>
<td>588 ~ 2%</td>
<td>Novel</td>
<td>4</td>
<td>3-5</td>
<td>24</td>
<td>t test, Bonferroni adjusted</td>
<td>4 young vs 5 middle vs 4 old; Dwarf: 4 young vs 3 middle vs 4 old</td>
<td>Liver</td>
</tr>
<tr>
<td>2002 (23)</td>
<td>Nylon membranes</td>
<td>2352 ~ 8%</td>
<td>Novel</td>
<td>-</td>
<td>8 per treatment</td>
<td>32</td>
<td>SAM</td>
<td>8 9-month old vs 8 9-month old GH receptor KO or CR treated mice</td>
<td>Liver</td>
</tr>
<tr>
<td>2002 (26)</td>
<td>Affymetrix</td>
<td>Full Genome*</td>
<td>dChip</td>
<td>2 (not for each biological replicate)</td>
<td>3</td>
<td>71</td>
<td>ANOVA, linear regression, Bonferroni adjusted</td>
<td>6-8 timepoints throughout life span of ad lib or CR flies</td>
<td>Pools of 10 flies per timepoint</td>
</tr>
<tr>
<td>2002 (30)</td>
<td>Custom cDNA array</td>
<td>Full Genome*</td>
<td>Novel</td>
<td>-</td>
<td>3-6</td>
<td>26</td>
<td>ANOVA, corrected for multiple testing at 0.001</td>
<td>6 timepoints throughout life span</td>
<td>Pools of thousands of mass cultured nematodes per timepoint</td>
</tr>
<tr>
<td>2004 (34)</td>
<td>Custom cDNA array</td>
<td>921 ~ 5%</td>
<td>Lowess</td>
<td>4</td>
<td>4-5</td>
<td>37</td>
<td>LIMMA, Benjamini &amp; Hochberg adjusted</td>
<td>4 timepoints throughout N2 life span, compared to equivalent timepoints in dat-2</td>
<td>Individual nematodes</td>
</tr>
<tr>
<td>2004 (37)</td>
<td>Affymetrix</td>
<td>Full Genome*</td>
<td>Novel (37)</td>
<td>-</td>
<td>5-6</td>
<td>17</td>
<td>SAM</td>
<td>Young vs old (2 timepoints)</td>
<td>Pools of 30 male hybrid Drosophila</td>
</tr>
<tr>
<td>2004 (38)</td>
<td>Affymetrix</td>
<td>12000 ~ 40%</td>
<td>dChip</td>
<td>-</td>
<td>30</td>
<td>30</td>
<td>SAM on a subset of the 30 cases (2-way comparison)</td>
<td>Partition of age-range into young and old, most analysis carried out on subsets</td>
<td>Brodmann area 10 from frontal cortex of different aged individuals</td>
</tr>
</tbody>
</table>

Fig. 2. Summary of results from selected microarray papers focusing on aging published between 1999 and 2004. *Indicates full genome at time of publication; currently may not be full genome as a result of improvements in gene identification as well as annotation updates. CR, calorie restriction.
translated into a common practice at the bedside. Of course, the comparison of cancer-related microarray studies with aging-related microarray studies is not perfect. Tumors, unlike aging organisms or tissues, are often homogeneous in their make-up, and this characteristic plays to the strength of the microarray approach. Nevertheless, it is a useful comparison to make, as much of the microarray methodology that was pioneered in cancer research has now been applied to a wide variety of biological problems, including aging.

Microarrays and Aging: The Beginning of the Beginning (1999-2001)

Studies by Prolla, Weindruch, and colleagues

The first studies to be published using the microarray platform to address the biology of aging were those pioneered by Prolla, Weindruch, and colleagues (14). Essentially the authors used Affymetrix array to profile differences in gene expression in a skeletal muscle in the leg (the gastrocnemius) between young (5-month-old) versus old (30-month-old) C57BL/6 mice (an inbred strain commonly used in research) (Fig. 2). In this particular experiment, caloric restriction was also used, in an attempt to determine whether this intervention would shift the transcriptional profile of the old mice to that of the younger 5-month-old animals. A surprisingly small number of animals were used in the study (N = 3 at each time point/treatment). The number of genes being assessed on the microarray was 6347, the highest available at that time for mice. No formal statistical inference was used to determine whether there was a transcriptional profile associated with an aged state in mouse muscle. Instead, the researchers chose to rank genes based on the absolute value of the mean of the log_{2} ratio of expression [abs(log_{2}(relative expression))] for young versus old mice. Log transformation is typically used because microarray data are often skewed, and the transformation results in a more symmetric (normal) distribution of expression values. This is important because the statistical inference often applied to microarray data relies on the assumption of normality. Normality is particularly important if there are a small number of arrays in an experiment.

This type of approach identifies one of the first crucial decisions required for microarray studies: What statistic should one use to rank the genes? That is, how should it be determined which genes show the most dramatic differences in expression between the two states being studied? Typically, the first attempt to find patterns in microarray studies results in a simple ordered list of the genes based on their distribution of expression. Lee et al. (14) chose to use mean (log_{2}) differential expression, but other choices are possible [for example, a t statistic, which normalizes the parameter estimate of interest (for instance, the mean differential expression) by the variability of the estimate]. The next step is to determine how to reduce the number of genes on the original list (usually a very large number) to a smaller, more tractable set of genes targeted for further study. Lee et al. selected genes that exhibited expression levels that were more than 2-fold induced or repressed [which is equivalent to the abs(log_{2}(relative expression))] > 1] in the young versus old mice and regarded these expression changes as “significant.” This was a common technique at the time (1999), and many other groups also made arbitrary cut-offs for significance at the 2-fold level. Why 2-fold was determined to be significant in these early studies (not just in this report, but in many others as well) was never adequately explained. This approach is still seen in some publications, but is indefensible given the current number of acceptable analytical techniques available.

One possible reason for the use of “2-fold as significant” might derive from the results of expression studies in yeast or bacteria, which generally show large changes in transcriptional abundance as a requirement for the activation of various pathways. Changes in abundance of various transcripts in yeast or bacteria that are less than 2-fold generally don’t have meaningful biological effects. However, the situation in mammals is far less clear, because our knowledge of the mammalian genome and the transcriptional changes that occur over a wide range of physiological states and tissues is far from complete (one could even say that the nature of such changes is almost completely unknown).

Lee et al. used no statistical methods to determine whether or not the distribution of expression among the genes identified as “interesting” was consistent with chance variation. It could be argued that the principal weakness of this initial report is the number of animals used; an N of 3 is generally regarded as not being sufficient for drawing robust conclusions in aging-related studies. However, this issue is clearly related to the variance of gene expression in the population. If gene expression is extraordinarily tightly controlled with age, then perhaps no more than three or four animals, tissues, or cells need be sampled per time point. In the absence of data demonstrating negligible gene expression variance with age, caution must be used in interpreting any results with a biological N of 3 or 4, given that we know there is a great deal of variability with age in many species, for many traits. The report by Lee et al. generated a great deal of excitement at the time; it was widely perceived as being at the forefront of the molecular biology of aging and a portent of great things to come. It also elicited some controversy and sharp criticism, the nature of which was primarily statistical in nature (15, 16). The types of concerns raised in the controversies surrounding the studies by Prolla and Weindruch remain a valid issue in all microarray studies to date (15). Prolla and Weindruch have continued to publish microarray studies on aging, with their main focus being the effects of caloric restriction and regard these expression changes as “important.”

Studies by Miller and colleagues

One of the first studies seriously to address aging with microarray methods using established statistical principles did not use high-density arrays at all, but used “macroarrays.” Spotted nylon membranes containing a few hundred (588) genes were hybridized to radioactively labeled RNA extracted from aged and young mice (3). Miller and colleagues also compared the expression patterns of RNA isolated from wild-type mouse livers versus the Ames dwarf mouse (http://sageke.sciencemag.org/cgi/content/full/2001/1/tg11) (a long-lived mutant mouse) in an attempt to identify biomarkers of aging in this tissue. They used the level of significance (the P value) from a standard statistical test (where the null hypothesis is no differential expression) to rank the genes. In addition, they recognized (for a fixed raw P value cut-off) that the potential for making type I errors (that is, the possibility of identifying false positives) goes up with the number of genes compared. Thus, one will expect by chance a certain number of “small” P values (implying significant differential expression) if enough tests are computed (in this case, enough genes are examined), and so a further step beyond simply reporting the raw P value is necessary to conclude that rela-
tive expression is (statistically) different in the groups being compared. The problem of reporting “interesting genes” when thousands of potentially dependent tests are performed is a classic problem in statistics known as multiple testing. This has been addressed in various ways (20–22); in this case, the authors tried to control the false discovery rate (FDR), which is the expected proportion of false positives in a ranked (by P value) subset of genes. They chose a list where the attempt to control the FDR was set at 5%. Intuitively, the further one goes down the list (as the P values increase) the higher the FDR. Other ways of controlling false positives are to control the family-wise error rate (FWER), the probability of making any type I errors by declaring all genes with a specified P value or smaller as being significantly differentially expressed. A typically conservative approach (that is, one that overestimates the false positive error rate) to controlling FWER is the Bonferroni method, although more precise (resulting in true error rates closer to the nominal rate) and typically less conservative methods are available (20–22). Although the authors were familiar with the requirement to correct for multiple testing, the study design was somewhat limited; the number of animals used per time point was between 3 and 5, and little rationale was given as to why liver was chosen as being a critical tissue to examine for age-related changes (Fig. 2).

These caveats aside, the authors make many good points that are still relevant to microarray studies carried out today. Specifically, it is important to ask (i) What is the variance of gene expression in the target tissue or animal between individuals at a given age? (ii) How does the statistical threshold for significance affect the lists of genes identified as being differentially expressed with age? (iii) How can statistical noise be reduced in these studies? These researchers constructed various filtering paradigms to eliminate noise and winnowed gene lists down to identify genes that are likely differentially expressed with age.

Early in 2002, these same authors reported the results of similar experiments using calorically restricted and long-lived growth hormone deficient mice (http://sageke.sciencemag.org/cgi/content/full/2002/8/5g1), in addition to increasing the numbers of animals used to examine gene expression and age in wild-type, ad libitum-fed animals (23). Arguably, this report is still the best mammalian study carried out to date to test a specific hypothesis on aging in a rigorous and well-designed fashion (Fig. 2). The reason we make this statement is that this particular study asked a very narrow question and then applied microarray methodology to it. Often in microarray studies, many questions are posed, and this sometimes results in compromise of an optimal study design. This particular experiment used 32 mice in total (the exact makeup of sex was not reported), half of which were calorically restricted and half of which were mutant for the growth hormone receptor. A single time point of 9 months of age was analyzed in the life span of these mice. The argument to support this experimental design was that at 9 months of age, the mice were actively experiencing the transcriptionally mediated beneficial effects of caloric restriction, whereas at later time points such benefits might no longer be present. For analysis of the microarray data, the authors used a “plug-and-play” Excel program called SAM (http://www-stat.stanford.edu/~tibs/SAM/) (significance analysis of microarrays) (24, 25). Essentially, this program facilitates multiple testing correction and assignation of statistical significance of differential expression for a number of experimental designs (25).

It is very easy to use and is easily applicable to a variety of different experimental design scenarios. Biologically speaking, the conclusions of the Miller 2002 paper were modest, and probably somewhat limited by the number of genes (2352) on the membrane array, as well as by the single time point and tissue chosen (Fig. 2). Essentially, the authors did not find evidence that the genes that display altered expression under caloric restriction fell into specific classes (whereas some previous studies did); however, the researchers did find some evidence of coordinated changes within a class of genes in animals undergoing caloric restriction. They made some logical inferences about what direction the technology could go in future studies, such as toward higher resolution of sampling across time.


Longitudinal studies of aging, in which individual animals are followed through time, are difficult to do and usually require substantial expenditure of resources as well as careful consideration of the initial experimental design to maximize the value of the resultant data. From a logistical point of view with regard to longevity studies, working with the invertebrate model systems as compared with mice is much easier. This is due to a combination of simple economic and temporal factors, that is, one can carry out “high-resolution” studies of aging in Caenorhabditis elegans because this organism only lives about 3 weeks, but comparable studies in mice take 3 years, which adds a substantial number of confounding economic, logistical, and methodological factors. “High resolution” with regard to aging-related studies is of course an entirely arbitrary term. In microarray studies of aging in the worm, high resolution might mean that samples are analyzed every 2 days or so over the course of the life span of the worm (a difficult but definitely feasible experiment), whereas high resolution in a mouse study might mean that samples are analyzed every 6 months over the course of the life span.

Why is it desirable to have high-resolution sampling over the course of the life span in microarray studies? This is not just a case of more is better, but rather speaks to our lack of knowledge as to when critical events occur at the transcriptional level in different tissues or cells of animals as they age. We really have no idea whether there are critical transcriptional events that occur during the course of life that mediate or modulate the functional decline of a tissue or an organism’s ability to maintain homeostasis. The earlier studies discussed above hinted at some transcriptional changes occurring somewhere during life that might be related to the aging process itself, but by only measuring gene expression levels at one or two time points between early adulthood and near the end of life, the dynamics of the “aging transcriptome” remain opaque.

Studies by Pletcher, Partridge, and coworkers

With any given gene’s expression over the course of life, is there a steady-state expression level? Does it rise and fall with some measurable amplitude? Does it peak at middle age and fall steadily back to “young” expression levels? These are the sorts of questions that can only be answered by high-resolution microarray studies of aging. One of the first such studies using this type of approach was published by Pletcher, Partridge, and colleagues in 2002, using Drosophila melanogaster as a model system (26). The study compared patterns of gene expression from...
calorically restricted *Drosophila* to ad libitum-fed *Drosophila*, coupled with replication and a high density of sampling over the course of the life spans (Fig. 2). Sampling from a large population was carried out at six time points over the course of the ad libitum-fed flies’ life span, and at eight time points over the course of the calorically restricted flies’ life span. Fifty flies were collected at each time point and divided into five groups of 10, and the RNA from each group of 10 flies was pooled and used for microarray studies with the Affymetrix platform. For each time point, three independent pools of flies (biological replicates) and two technical replicates were carried out.

In microarray parlance, a technical replicate means that a duplicate array (using the same biological sample) is made to ensure that there is a high level of technical precision of the microarray measurements and that there will be little contribution of “noise” as a result of the array assay itself. Typically, the results from technical replicates are averaged before formal statistical analysis is carried out. For example, if you have five biological replicates and you carry out two technical replicates per sample, this would give you 10 arrays. However, this does not mean that *N* (the number of independent samples) is 10. It is still 5, and one approach to gain precision might be averaging each pair of technical replicates per biological replicate before analysis (other approaches, such as random effects models, are also available).

Given the fixed total number of arrays reported in the paper by Pletcher *et al.*, the effective *N* per time point is reduced from a possible five biological replicates to three, as these researchers carried out two technical replicates for two of the pooled biological samples. As they implicitly employ a form of averaging by pooling 10 individuals per biological sample, and thereby reduce the variance of expression between individuals as a function of age, it is possible that the most efficient allocation of arrays would have been to increase technical replication to reduce the variation between arrays on the same biological sample. Although physical pooling of samples has its costs (for instance, it results in the inability to look for the source of variance between individual samples), for particular questions, it can also increase the efficiency of the experimental design. When designing experiments with a fixed number of arrays, if one has estimates of variation (both biological and technical) and the question of interest is well defined, then it is a reasonably straightforward task to determine the optimal allocation of arrays (technical versus biological) and thereby employ pooling to maximize the power of the experiment (minimize the variance of the estimated parameter(s) of interest; for example, see (27)). It is important to note that one can always statistically pool data later “in silico” (via computer-based analysis), but obviously one can never go back to look at expression from an individual sample if the sample itself has been physically pooled.

In their analysis, Pletcher *et al.* used a multiple testing approach (this time, controlling the FWER using the conservative Bonferroni method) to avoid excessive numbers of false positives. In addition, they used an exploratory hierarchical clustering algorithm to identify groups of genes with similar expression profiles. A natural step after identifying potentially interesting genes is to find sets of potentially related genes (those undergoing similar changes in expression patterns) using a clustering (grouping) algorithm, of which many are available. Like many others who analyze microarray data, Pletcher *et al.* used a free statistical package, R (http://cran.r-project.org), which has an extensive set of statistical analysis functions. An important step in the reporting of microarray data is the validation of results via some alternate methodology. The reason for doing this is not so much to validate the technology per se, because many studies have shown a high concordance between the abundance of certain transcripts as detected by microarray and that detected by real-time reverse-transcriptase polymerase chain reaction (PCR) when done correctly [for example, see (28)]. (In real-time PCR, the concentration of product is monitored during each PCR cycle rather than only at the end of the reaction. This technique provides a more accurate measure of the initial template concentration than other PCR-based methods.) The logic of carrying out such validation studies is more to give a measure of confidence that the authors have correctly implemented the microarray technology “in-house,” and that therefore the study has merit and the conclusions are likely to be reproducible if carried out in another laboratory. This is, of course, assuming that the real-time PCR is carried out correctly, because this method also has a great deal of variability in its implementation and the consequent veracity of its results (29).

Pletcher *et al.* acknowledge that it would be beneficial to validate some of their findings with real-time PCR in individual flies, rather than in pools of flies, but do not report any alternative studies to validate their findings. In summary, the principal conclusions of this report are that (i) calorific restriction “slows down” the transcriptional profile of aging and (ii) there appears to be a strong increase in the transcriptional response to microbial infection in aged versus young flies. This report is arguably an example of ambitious interpretation from limited data. With approximately 70 chips available, could they identify genes that differentiated flies undergoing calorific restriction from those that were ad libitum fed? Could they identify functional clusters of genes that changed with time? Could they identify novel age-related expression profiles? All these points are worth addressing in their own right but are moot in the absence of an appropriate experimental design. By deciding on a high resolution of sampling (six to eight time points), the authors compromised their ability to make robust conclusions about the time points themselves by only sampling three groups of animals per time point (*N* = 3). They added two technical replicates on top of this (which were each carried out with different methodology), which provides less variable estimates of whatever gene expression is present in the three pools from which they sampled. Technical replicates will always give higher measures of reproducibility than biological replicates, and it is formally not correct to combine technical and biological replicates to arrive at *N* (that is, one should use statistical techniques that can allow for repeated measures on the same statistical unit).

This was the first aging-related study using microarrays to refer to and use the statistical language R in the analyses. Bioconductor (http://www.bioconductor.org) is an add-on package to R and has functions specifically designed for microarray analyses, from processing the raw data to quality control, multiple testing (multtest), graphic displays, and clustering. Bioconductor is also free and constantly updated with both improvements to existing functions and new analysis tools. The utility of the resources made available at the Bioconductor project facilitate reanalysis of other investigators’ data, reconstruction of chip “images” (to see how well hybridizations were carried out, for example) from raw data and, best of all, a community of statistically enlightened researchers who are actively developing...
new tools to analyze microarray data on an almost weekly basis. Commercial packages for analysis often have good support, but the analytical backbone of such software is usually 1 to 2 years behind the sorts of solutions developed and made freely available in the Bioconductor project.

**Studies by Johnson, Kim, and coworkers**

This brings us to the next longitudinal study on aging published in the same year, that of Johnson, Kim, and colleagues (30). Like Pletcher and Partridge, Johnson’s group used pools of synchronously aging animals (*C. elegans*) to obtain the RNA for a transcriptional profile of aging. In this case, they synchronized large populations of *C. elegans* on plates (using in excess of 100,000 worms per time point). The worms carried temperature-sensitive fertility mutations (which cause sterility at high temperature), a feature that prevented the synchronously aging population from being contaminated with younger offspring. Three different sterile strains were used to generate the RNA used in the studies, *fer-13, spe-9; fer-13, and spe-9, emb-27*. Table 1 of Lund et al. (30) nicely describes the experimental design of the study, making it easy for the reader to understand what was done, and how. This is not a trivial point; many of the papers we discuss here are very readable in the sections where the implications of the results are discussed (which we have not attempted to cover in any detail here) but are difficult to comprehend in their study design or basic methodology because of a lack of clarity in the methods section. In some cases, it is entirely ambiguous as to how many replicates were carried out at different ages. This problem is likely a result of the rapidly developing set of standards in this technology in recent years and the comparative novelty of these papers when they were initially reviewed.

Johnson’s team essentially profiled aging across six time points over the course of the life span of *C. elegans*. The number of biological replicates used at each time point (combined from the three independent mutant strains listed above) varied from three to six. From the perspective of capturing the biological variance inherent in aging populations, the approach offers some advantages over that used by Pletcher et al., as Lund et al. were using more than three (between four and six) independent biological replicates for most of their time points and creating a vast pool of averaged RNA from many thousands of individuals, which presumably would result in the detection of the most robust changes with age across thousands of individuals. However, an implicit assumption is made in their experimental design that each of the strains used ages in the same way and can therefore be treated as functionally identical. No data are provided showing this to be true at the transcriptional level. The exact makeup of the pool at each time point also varied with respect to the strains that were included in a nonsystematic way (from one to five), and the potential influence of strain variability on gene expression at each time point is unclear.

These authors rank the genes based on the statistical significance of the association of age with expression [instead of assuming a model for either an increase or a decrease in expression with age, they divided the life span into a number of arbitrary groups (categories) based on age and used a standard statistical technique (one-way ANOVA) that tested whether there were any differences in mean expression among the categories]. They take into account multiple testing, but simply set a more conservative *P* value (<0.001) rather than formally controlling for an experiment-wise error rate (such as FWER or FDR). Because different strains are used, the authors do worry about whether this could impact the conclusions about which genes undergo the greatest age-related changes in expression, as discussed above. However, they do not formally test for the influence of age on each strain but simply point out that for almost all of the genes selected, age was a stronger predictor of expression level than was the strain identity (that is, age explained more variance in expression than did the strain). If one is interested in estimating the average association of age and expression among these strains, then the design suffices and no apologies are necessary for differences among strains (that is, the associations with age are not confounded by strain). However, the question as to whether the expression profiles by age differ in an important way by strain could have been approached in a more straightforward manner by examining the statistical interaction of age and strain on expression.

Lund et al. conclude that there are coordinated changes in the expression of functionally related groups of genes with age, including decreased expression of heat-shock genes and increased expression of dauer-regulated genes. [The dauer is a stress-resistant form of the worm that is induced by starvation (see “Dauer Power” at http://sageke.sciencemag.org/cgi/content/abstract/2002/31/nw110].] However, like most of the prior studies discussed above, no independent validation of differential expression with age was carried out.

**Meta-analysis**

If we turn now to the current year (2004) in our selective review of microarrays and aging, we find the first example of a meta-study report on investigations that use microarrays as a tool to investigate the biology of aging (31). This is a purely bioinformatics approach, in that the authors take a lot of the previous published microarray data that they deem relevant to aging and do a “meta-analysis” in the hope of identifying meaningful trends across a number of studies such as those discussed above, or across studies that look at early stages in the life span through gene expression profiling (32). They compare gene expression patterns across organisms by using comparative analysis of microarray experiments. Specifically, the researchers attempt to find similarities in the relative expression of the same genes in the old versus the young among different species. First, they use a phylogenetic analysis to narrow the focus to a subset of orthologous genes. They use simple correlation of the orthologous genes and compare this estimated correlation in relative expression to randomly drawn sets of genes of equivalent number. However, they appear to have made an error when reporting the *P* value of these associations: The authors treat the genes as independent observations when deriving statistical inference. The mistake made here has been described in the ecological literature as pseudo-replication (33), defined roughly as the treatment of statistically dependent observations as independent experimental trials. Typically, this results in an overstatement of statistical significance.

To identify pseudo-replication, one must first define the target population and the units of statistical independence. In this case, one might argue that the target population is all metazoa, and the units of independence are the species. If that is the case, then one possible experiment would be a comparison of data derived from microarray analyses performed on a random sample of metazoan life paired by different ages. However, for the purposes of this discussion, we will agree to narrow the target...
population to a clone of worms and a corresponding clone of fruit flies. Thus, each correlation estimate is a single measurement of a random variable, which is itself a complicated function of many dependent random variables. More generally, estimates derived from the same microarray data on a sample should be assumed to be statistically dependent. When estimates are made from independent random variables (such as measures of relative gene expression obtained from comparing two microarrays on which samples from two individuals of the same clone at different ages were analyzed), then deriving variability estimates and inference can be much more complex. If one can identify sets of priori independent genes, then standard repeated measures techniques (for example, random effects models) can be used. However, if this approach is not defensible, then one should repeat the experiment with independent draws from the target population and make proper inferences either by assuming a (justifiable) model for the dependencies among genes or by using a more nonparametric approach and a sufficiently large sample size.

Studies by Melov and coworkers

Turning now to our own recently published study (S.M. and colleagues), we were interested in temporally profiling transcriptional changes during aging at a high resolution at the individual level (34). Our main reason for using microarrays was that we were interested in comparing gene expression profiles of discrete individuals from a clonal population of wild-type C. elegans (strain N2) to those from a long-lived mutant [(daf-2 (http://sageke.sciencemag.org/cgi/genedata/sagekeGdbGene;38)] (Lee Perspective at http://sageke.sciencemag.org/cgi/content/full/2004/18/pe18). Gene expression profiling of individual nematodes was our goal, in addition to the comparison of the two strains of C. elegans. Recently, amplification of RNA from small amounts of total RNA has become commonplace, and the distribution of transcriptional abundances is largely unaltered by the amplification process. This procedure makes possible the microarray analysis of individual worms, which generally contain nanogram quantities of protein. This is desirable for the following reasons:

(i) Despite the fact that a population of C. elegans is clonally identical, it follows a mortality curve similar to that of outbred species. This observation implies that there are some random or stochastic effects on mortality that act in combination with the environment and genetics that could be uncovered through expression profiling at the individual level. Typically, thousands of worms are pooled for each time point to obtain sufficient RNA for analyzing expression changes that occur with age, and pooling worms at this level will obscure changes that occur in some individuals but not others with age. Therefore, it is advantageous to look at individual worms if one wants to understand individual variation with age.

(ii) Individual worms have characteristic morphologies with increasing age, and their behavior can even be classified into different types with age [for example, as fast or slow moving (35)]. Therefore, analysis of individual worm expression profiles might reveal transcriptional differences that are associated with specific age-related behaviors or other individual phenotypic markers.

(iii) Many investigators use fertility mutants (such as glp or spe mutants) or drugs (such as inhibitors of DNA replication) to block reproduction to facilitate survival studies. However, the influence of such treatments on transcription is largely unknown, and conclusions about gene expression in relation to "normal" aging based on such studies should be treated with caution, because the transcriptional profile that is obtained by using such methods might be unduly influenced by the mutations or chemicals used to facilitate the studies. Therefore, it is important to study gene expression during aging under standard survival conditions using N2, the most commonly used wild-type strain, as a benchmark for comparison to longevity mutants.

(iv) Mass culture of worms is known to cause the induction of the stress response at the transcriptional level (36). Therefore, to study expression changes that occur with age that are not influenced by mass-culture conditions, we studied profiles of individual worms under "standard" survival conditions, which the vast majority of investigators use to infer mortality kinetics when studying longevity mutants and aging.

Many of these principles apply equally well to protein studies, in that when one is doing biochemistry on aging populations of worms, assays are generally done on pools of thousands of individuals, so individual aspects of aging are not observed. However, no technology yet exists for proteomic profiling of individual worms with age, so this potential area of research remains fallow until technology breakthroughs facilitate the quantification and identification of nanogram/picogram quantities of protein.

It can be argued that as worms or flies age, they tend to become dysregulated with regard to synchronous aging, so that even though the individuals in a population are all chronologically the same age, they are actually different physiological ages. This phenomenon is clearly seen in the shape of typical survival curves. If such changes in physiological age are mediated in part by transcription, then gene expression profiling of large numbers of individual worms or flies may detect these different classes of aging. As a first step toward carrying this out to a high level of confidence, we performed a small pilot experiment (37 arrays) that profiled gene expression with age in four to five individuals at four time points through the life span of strain N2 compared with daf-2 (34), using methods developed in the Bioconductor project, to detect differential expression with age (Fig. 2). Although our sample size was too small to discern different subclasses of aging, our studies demonstrated (i) the feasibility of profiling the gene expression of discrete individuals and (ii) that the daf-2 mutant had a different gene expression profile as compared with the wild-type strain, which was previously unreported. However, as in the other studies we discuss above, no validation of the gene expression output was presented.

Studies by Tower and coworkers

In a recent study of aging in Drosophila, Landis et al. used gene expression profiling of young versus old flies and compared these profiles to those of flies that had been stressed by exposing them to high concentrations of oxygen (37). The central idea here is that oxidative damage is a major cause of the aging process (see “The Two Faces of Oxygen” at http://sageke.sciencemag.org/cgi/content/full/2001/1/oa5) and that comparison of the stressed young flies to normal aged flies might reveal an overlap in their transcriptional profiles. Although quite a modest microarray study in scope (there were 17 chips in the study), the study is conceptually clear with respect to what the authors were testing. There is an appropriate level of replication for each type of treatment, with at least five biological replicates for each treatment and a validation by Northern blot for selected
genes identified through differential expression analysis (which, as in the studies described above, was done using SAM). Apart from confirming the group’s previous work with regard to changes in the expression of genes encoding heat-shock proteins during aging, the authors found a strong correlation in expression profiles between oxidatively stressed young animals and aged flies. One problem with this paper is that the reporting of statistical methods and results with regard to the selection of differentially expressed genes is very cryptic, and at the least requires an intimate familiarity with the particular software used in order to understand how the analysis was carried out [SAM (37)]. Optimally, authors should explicitly describe the statistic used to order genes, the criterion used to select differentially expressed genes, and a resulting error rate (FDR, FWER). These additions would facilitate analysis of the raw data as well as help to clarify how the conclusions were reached.

Like a number of other groups, the authors generally observed that the transcriptional profile of flies undergoing oxidative stress appeared to be similar to that of aging flies. Likewise, both groups of flies displayed an induction of the immune response and purine biosynthetic pathways at the transcriptional level. One novel feature of the study was the tagging of genes identified by differential expression analysis with the gene for green fluorescent protein, which allowed the expression of the encoded proteins to be followed over the course of the life span. This experiment confirmed the differential expression detected by SAM at the protein level. Furthermore, this result implies an ability to predict survival in individual flies on the basis of the expression of certain genes at a given time in life, which conceptually would be relevant to humans. The extension of this concept to ourselves is one of the many promises of microarray technology. It is not difficult to visualize that with appropriate sampling, experimental design, and analysis, it would be feasible to obtain a “snapshot” of physiological age via microarray, which could then be used to predict life span and potential risk factors for age-related disease. This is clearly where the technology is heading, and there is considerable excitement over the potential use of such methods.

Studies by Yankner and coworkers

We turn now to the final study discussed in this Review. Bruce Yankner and colleagues recently published an impressive body of work that was the first reported gene expression study on human brain aging over the life span (38). There were a number of novel aspects to the study: (i) The authors used a reasonable number of patients to look at transcriptional changes that occur with age (N = 30, age range 26 to 106 years of age). (ii) They not only identified genes that were differentially expressed with age within a cohort, but they also carried out some additional experiments that might shed light on why they detected differential expression with age in certain genes. (iii) They carried out a number of controls that to some extent validated their observed results.

When carrying out a study on human aging, it is critical to have clinically well-defined samples. One of the accepted characteristic features of aging-related experiments in animals (be they worms, flies, mice, or other species) is that the population is generally well defined and controlled. Individual laboratory organisms, for example, all grow up in a homogenous environment and are all fed the same diet. This feature minimizes environmental differences between individuals and maximizes the likelihood that one will really be examining the consequences of age, independent of environment. This situation is definitely not the case for humans, particularly near the end of life, because prolonged or even short stays in the hospital or the emergency ward can result in a variety of pharmacologies being rapidly administered to counter the increased frequency of age-related conditions and/or emergency situations that unfortunately arise as an all-too-predictable consequence of increasing age. The impact of such pharmacologies or treatments on gene-expression profiles is almost completely unknown. Although studies can be designed to abrogate such issues (for example, biopsies can be taken from muscles of different aged, healthy, clinically well-characterized individuals), most studies rely on post-mortem tissue obtained through collaboration with physicians in a hospital setting, or alternatively, specimens are obtained through tissue banks.

In the study by Lu et al. (38), samples were all derived from the frontal pole of brains from 30 different individuals, who had been screened and ascertained to be normal, both neuropathologically and cognitively. The only additional data presented in the report germane to the clinical status of the individuals was the sex and the post-mortem interval (this information is in the supplemental materials). No information is provided with regard to cause of death or types of drugs administered to the patients as a result of whatever caused them to be in the hospital in the first place. These factors alone could have an important bearing on the gene expression profiles. This is particularly true with regard to the DNA damage aspects of the study (see below), as many drugs used in a clinical setting are known to be genotoxic and can result in oxidative DNA damage.

This point brings us to one of the novel features of the report, which was the identification of genes that exhibit reduced expression with a corresponding increased level of oxidative damage to their promoters with age. The authors conclude that the increased level of oxidation of these promoters is driving down the expression of such genes with age, and imply that this might be an important characteristic in aging itself. The levels of DNA damage they detect are unusually high, and quite unprecedented [well above reported levels in previous controlled experiments (39, 40)]. Assays for oxidative damage to DNA have been mired in controversy in recent years (40). This has resulted in recommendations as to what is the “right way” to measure such damage, versus the “wrong way(s)” (39). Unfortunately, it is likely that the levels of damage reported by Lu et al. are highly inflated, as they used methods for isolating DNA from the samples that are known to cause artifact (39). Specifically, they left out the recommended metal chelator desferrioxamine from their isolation buffer, which will result in increases in the levels of oxidative damage caused by transition metal-mediated oxidation. It is also well known that the concentrations of certain metals increase with age in the brain in multiple species (41). One possible explanation for the authors’ observations of increased oxidative DNA damage with age in specific genes is that the authors are indirectly detecting increased levels of iron or other transition metals in the brains of older individuals. From their results, this would imply that there are certain genes that are predisposed to this type of damage, given that not all genes were equally affected in the study. This is a fascinating and intriguing observation and is worth additional investigation.

The authors also report a high level of concordance be-
tween their real-time PCR validations of differential expression across 18 genes that is quite puzzling. In general, Lu et al. report extremely tight agreement between results from the two methods (arrays versus real-time PCR), even in the magnitude of the expression level. That is, if the real-time PCR data for a given gene showed that its expression level in old individuals was 90% of the level in young individuals, the array showed exactly the same result. Other studies have shown much less concordance between the two techniques, and it is not clear why the data in Lu et al. showed exactly the same result. Other studies have shown that the middle-aged group from their study (40 to 70 years of age, N = 11) displays the most variability in levels of gene expression as detected by microarray analysis, as well as the most variability in levels of DNA damage. Their method of analysis is an association of interest between two continuous variables (genes and age), and so simple correlation (and its statistical significance) was used to select genes potentially related to aging. Although they use a fixed P value cut-off (<0.005), they could have used formal multiple testing criterion (such as FDR) to choose those genes “significantly” related to aging, as they did for their categori
cal age group comparisons (between individuals who were ≥73 years old versus those who were ≤42 years old). Bootstrap methods have been proposed to control error rates of interest more accurately for more complicated tests and parameters beyond simple two-sample comparisons (20–22).

Overall, there are some intriguing trends in the data from this paper, but clearly this sort of study indicates the need for well-defined clinical samples of different ages and a very large N to overcome the variability that is undoubtedly present as a result of environmental and genetic factors beyond the control of the investigator.

Conclusions

In summary, microarrays as a tool to investigate the biology of aging hold much promise. Microarray-based analysis demands some rigor with regard to technique, experimental design, and attention to well-established statistical principles. In the coming years, we will see microarray studies using hundreds of chips; indeed, this is already common in the cancer field. What is needed with regard to aging-related studies using microarrays is (i) an increase in the resolution of sampling with age; (ii) sufficient replication and validation to be sure of the results; (iii) a definition of the gene expression variance of any given population or sample; and (iv) an informed bioinformatics analysis that will do more than generate lists of genes that are differentially expressed. The potential of microarrays to uncover mechanisms of aging through gene discovery has yet to be shown, but at the very least, microarrays may prove quite useful in providing an accurate biomarker of physiological age.

References and Notes


10. C. Kihara, T. Tsunoda, T. Tanaka, H. Yamana, Y. Furukawa, K. Ono, O. Kita


16. T. A. Prolla, D. B. Allison, R. Weinreich, Response to “Interpretation, de


42. S.M. is supported by a National Institutes of Health Grant AG18679, and a senior scholar Ellison award from the Ellison Medical Foundation. S.M. would like to thank K. Beckman and T. Golden for helpful comments.