Journal of Gerontology: BIOLOGICAL SCIENCES
2001, Vol. 56A, No. 2, B52–B57

Interpretation, Design, and Analysis of Gene Array Expression Experiments

Richard A. Miller,[1,2,3] Andrzej Galecki,[2] and Robert J. Shmookler-Reis[4]

[1]Department of Pathology, [2]Institute of Gerontology, and Geriatrics Center, University of Michigan, Ann Arbor.
[3]Ann Arbor DVA Medical Center, Michigan.
[4]Department of Geriatrics, University of Arkansas for Medical Sciences, Little Rock.

Experiments using arrays of cDNA targets to compare patterns of gene expression are beginning to play a prominent role in biogerontology, but drawing reliable conclusions from the resulting data sets requires careful application of statistical methods that discriminate chance events from those likely to reflect real differences among the samples under study. This essay discusses flaws in the logic of studies that base their conclusions on ratio calculations alone, reviews the multiple comparison traps inherent in high throughput systems that test a very large number of mRNAs simultaneously, and advocates a two-stage design in which significance testing applied to exploratory data is used to guide a second round of hypothesis-testing experiments conducted in a separate set of experimental samples.

NEW methods for the simultaneous assessment of the level of expression of hundreds or thousands of mRNA levels in individual cell or tissue samples have caught the attention of cell and molecular biologists, gerontologists among them. The lure is obvious: Instead of laborious, one-at-a-time assays for a handful of cytokines, cell cycle regulators, surface proteins, or transcription factors, high-throughput approaches seem to promise a cornucopia of quantitative gene expression data from which to select the most promising candidate genes for further analysis, as well as "expression fingerprints" that are as informative, and as detailed, as real fingerprints or DNA restriction fragment length polymorphism patterns. Articles presenting lists of mRNAs allegedly over- or underexpressed in the tissues of aged rodents (1) or in tissues derived from skin biopsies of young or aged human donors (2) have appeared in prominent peer-reviewed journals and are sure to be merely the vanguard of a flood of articles reporting the effects of age, species, mutations, diets, antioxidants, and disease states on patterns of gene expression in multiple tissue and cell sources. These articles have included, and will continue to include, lists of specific mRNAs found to be altered to a specific degree (e.g., 2-fold or 10-fold changes) by the factor of interest, accompanied by a discussion of patterns perceived within the data set: arguments that the list of altered genes includes many genes involved in antioxidant defenses, or cell cycle control, or responses to specific hormones, etc.

This essay presents the viewpoint that the design and interpretation of the most prominent gene expression studies published to date—as well as the majority of those now being presented at meetings or making their way through the review queue—are seriously flawed, that the data sets are filled with false-positive results, and that conclusions made on the basis of such fragile foundations are likely to prove misleading and premature.

The goal of such studies is usually to produce a listing of the genes whose expression distinguishes two samples of interest, for example, the muscle of young mice from that of old mice. For these lists to be useful as tests of specific ideas about aging and as guides for further work, it is important that most of the findings be reproducible (i.e., likely to produce equally large effects in replicate sets of samples). What criteria, then, should be used to ensure that such lists of age-sensitive genes contain only a small proportion of false positives (i.e., nonreproducible findings)? We will consider two sorts of criteria: (i) that the list should include all genes with a young/old ratio over some arbitrary value and (ii) that the list should include all genes where the two age groups meet some statistical significance test that compares effect size with its variance among subjects, such as the Student's *t* test. A recent review (3) provides a more comprehensive analysis of the statistical problems and opportunities involved in extracting biological insights from expression data sets and cites many useful articles describing alternate approaches to microarray-based data mining.

Ratio-Based Criteria (Young/Old [or Test/Control] Without Formal Significance Testing)

The two most prominent early reports of age-sensitive gene expression have eschewed significance testing and relied instead upon ratio-based criteria, so we will consider option (i) first. Presumably such a ratio-based criterion should be set high enough that few genes would reach the arbitrarily selected threshold by chance alone. Many investigators in this area, including those who have published their work in prominent journals, have selected an arbitrary criterion—typically a 2-fold but sometimes a 1.5-fold change—as the threshold for inclusion in their list of interesting results. We therefore performed a simulation study to see how often sets of random data would produce high ratios by chance. This simulation, based on the report of Lee

and colleagues (2), used a random number generator to produce fictional data for sets of 10,000 genes for each of six individuals, three of whom were considered the young cohort and three of whom were arbitrarily designated old. Each set of 10,000 random numbers followed the normal distribution with a mean of 100 units and a preselected *SD*. We then recorded the percentage of the genes for which the mean value of young samples was either twofold higher than old or less than 50% of the old value. The article by Lee and colleagues cited above in fact employed a somewhat different calculation, specifically the mean value of the young/old ratios taken across the set of nine possible pairs from three young and three old mice. We therefore tabulated the number of genes that exceeded various thresholds using this criterion as well.

Table 1 shows the results of this simulation study. For a data set in which, for example, all of the genes show a coefficient of variation (CV = 100 × *SD* divided by the mean) of 30%, the table shows that 58 genes would produce, by chance, a mean young/old ratio of 2× or higher. The criterion used in the study by Lee and colleagues (2) would increase the number of false positives from 58 to 212 (i.e., 2.1% of the genes examined). At this CV, only 1 gene per 10,000 would produce a fourfold change by chance alone, although the calculation adopted by Lee and colleagues would produce 16 false positives with fourfold changes. A less restrictive threshold (i.e., a 1.5-fold change) would produce false positives for 6.9% of the genes tested, or 10.1% using Lee and colleagues' criterion.

Table 1 also shows that the expected number of false positives depends greatly on the *SD* among individual test samples (individual mice in the example under consideration). This variation will combine the effects of technical factors (array quality, label activity, quality of the RNA preparation, etc.) with biological variance reflecting real differences among individual animals, donors, or cell cultures. There is at present little published data from which one can derive estimates of the variation to be expected in studies of gene expression. Estimates of this distribution will become more widely available when authors begin to publish variance levels in addition to mean values, and these will be useful in calculating estimates of statistical power prior to beginning a study. Biological traits can vary widely even among genetically identical inbred animals. The catalog compiled by Phelan (4) provides some indication of the variance to be expected. This review lists CV values for 49 miscellaneous traits taken from 24 different reports. The median CV value in these reports is 22%, and 30% of the traits show CV values in excess of 45%. Our own data on variance in liver gene expression in mouse liver (see Figure 1) found that 80% of the genes tested had CVs > 30% and that 56% of the genes had CVs > 50%. It thus seems very likely that other gene array experiments will frequently encounter genes where CVs exceed 30% to 50% for at least some fraction of the tested genes. The number of false positives, determined on the basis of a ratio threshold, can increase very rapidly if even a small proportion of the genes are highly variable among the animals or test samples. If, for example, a mere 10% of the genes in a given study have CVs = 40%, then Table 1 implies that one should expect to see 28 false positives per 10,000 genes using a 2× criterion, or 81 false positives using the calculation adopted by Lee and colleagues (1). Although the article by Lee and colleagues does not include the data on variation among samples that would be required to estimate false-positive rates, it is noteworthy that its comparison of muscle from 5-month and 30-month old mice reported 2× changes in 1.8% of the 6347 genes analyzed, approximately the number of false-positive results to be expected if all of the genes tested have CVs = 30% or if 10% of the genes have CVs = 60%.

The number of false positives to be expected depends greatly on the actual *SD* values in the set of genes under study, which will in turn vary with age, strain, cell type, intervention, etc. Each investigator should be able to use his or her own data set to produce an estimate of the number of ex-

Table 1. Simulation Study Results (Three Mice per Group)

	Mean Young/Mean Old				Mean of Nine Ratios			
Criterion	>1.5×	>2×	>3×	>4×	>1.5×	>2×	>3×	>4×
CV = 10	2	0	0	0	1	1	1	1
CV = 15	7	0	0	0	10	3	3	1
CV = 20	102	0	0	0	165	6	0	0
CV = 25	312	8	0	0	485	43	4	1
CV = 30	690	58	1	1	1005	212	49	16
CV = 35	976	138	7	1	1439	450	149	70
CV = 40	1337	287	19	3	1988	808	288	138
CV = 45	1668	472	54	15	2435	1208	565	291
CV = 50	1853	583	103	27	2694	1490	736	409
CV = 60	2434	989	268	96	3334	2131	1278	818
CV = 70	2737	1276	381	183	3820	2702	1662	1071
CV = 80	2930	1534	549	283	4061	3062	2084	1455
CV = 90	3073	1682	661	327	4312	3400	2401	1711
CV = 100	3150	1814	789	440	4491	3635	2666	1976

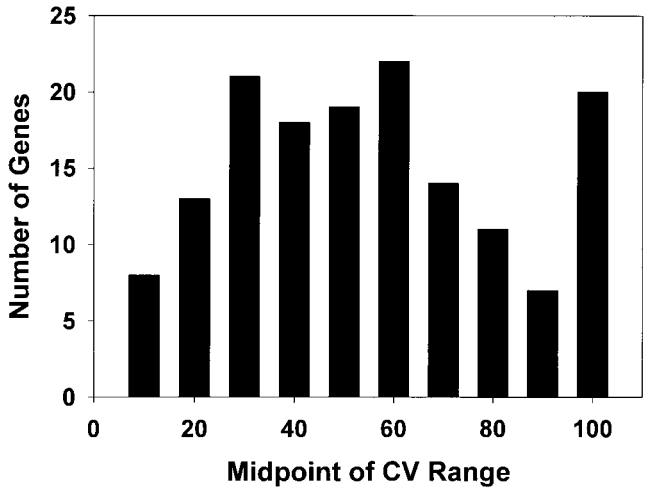

Figure 1. Distribution of coefficients of variation (CV = 100 × *SD* divided by the mean) for 153 genes expressed in the liver of Ames control stock mice aged 5 months, from data collected on $n = 4$ animals. CV values represented by midpoint of the range (e.g., CV = 20 represents 15 < CV < 25), except that CV = 100 assigned for all CVs > 95. (From Dozmorov, Bartke, and Miller [5]).

pected false-positive results and to compare these with the actual observations. Figure 1 shows an example of this approach from data generated during an analysis of liver samples from four mice, aged 5 months, of the Ames control stock (5). The histogram shows the distribution of CVs among the 153 genes expressed in the liver among the 588 target cDNAs arrayed on a Clontech (Palo Alto, CA) nylon membrane. This empirical distribution of CVs can be used, along with the simulation results collected in Table 1, to produce an estimate of the number of false positives to be expected. Table 2 shows an example of such a calculation and shows that a test comparing three young to three old mice, producing CVs distributed as in Figure 1, would produce approximately 28 false-positive results for 1.5-fold changes, 12 false positives with a 2× criterion, and 4 false positives for a 3× change. Thus, if analysis of the actual data shows that 23 genes produce changes of 2× (compared with 12 expected false positives), we would expect future work to show that about half of these would prove authentic but that the other half would represent merely chance effects.

The calculations shown in Tables 1 and 2 provide only rough approximations. For one thing, each simulation was based on a set of random numbers; therefore, replicate simulations yield slightly different results. Second, the simulations were based on the assumption that gene expression values are normally distributed. This assumption is incorrect because in many real data sets the distribution is skewed, with a preponderance of genes expressed at low levels. Third, in real data sets CVs tend to be higher for genes expressed at low levels, because for genes near the detection threshold measurement errors become progressively larger relative to biological variation. The calculation shown in Table 2 also makes the simplifying assumption that variation in samples from old mice is the same as in samples from young mice. Although these calculations provide only a rough guide, they do give an idea of the magnitude of the false-positive problem and of the likelihood that tests of additional animals will prove fruitful.

One violation of the normality assumption deserves special mention: instances in which the levels of expression of a specific gene turn out to be bimodal among individual subjects. Documentation of genuine bimodality requires fairly large sample sizes, but even in small samples such genes are associated with very large CVs. Of the 153 genes shown in Figure 1, 15 have CVs > 100. (Actually, Table 2 produces underestimates of the false-positive rate because it assigns CVs = 100 to all genes where CV > 100.) Most of these genes show high-level expression in only one or two of the four mice tested, with zero or near-zero expression in the other animals. Genes like these whose expression is sporadic among similar mice are particularly likely to give very high ratios in small series of this kind. If for a particular gene the distribution of expression among mice is truly bimodal or contains an occasional outlier—assumptions that cannot be tested without much higher numbers of animals—then assessment of the effects of age, treatment, or genotype on expression may be particularly difficult. Demonstration that genes with high ratios appear in two independent short series does not provide an adequate test against type I errors because genes with high variance will indeed frequently appear to differ, by ratio, between small groups of subjects even if there is no real effect of the diet or genotype under study.

The number of expected false-positive results depends on the number of mice (or other samples) tested in the study. Table 3 shows simulation results for varying *SD* (from 10% to 50% of the mean) for study designs utilizing two, three, four, or five animals in each of the two test groups. Using a criterion of a twofold change, for example, experiments in which CVs = 30% will yield 223 false positives if only two mice are used per group but will yield a mere six false positives if five mice are used per group. For CVs of 50%, and for designs with $n = 2$ mice per group, as many as 143 genes

Table 2. Calculation of Number of Expected False Positives for Various Effect Levels Using an Empirical Distribution of *SD*s

No. Genes ($N = 153$)	CV (Range)	Expected False Positives per 100 Genes				Expected False Positives per 153 Genes Expressed in Young Liver			
		1.5×	2×	3×	4×	1.5×	2×	3×	4×
8	5–15	2	0	0	0	0	0	0	0
13	15–25	1	0	0	0	0	0	0	0
21	25–35	7	1	0	0	1	0	0	0
18	35–45	13	3	0	0	2	1	0	0
19	45–55	19	6	1	0	4	1	0	0
22	55–65	24	10	3	1	5	2	1	0
14	65–75	27	13	4	2	4	2	1	0
11	75–85	29	15	5	3	3	2	1	0
7	85–95	31	17	7	3	2	1	0	0
20	95–105	32	18	8	4	6	4	2	1
Totals						28	12	4	2

Table 3. Numbers of False Positives Expected in Surveys of 10,000 Genes: Various Combinations of *SD* and Numbers of Mice per Group

SD	Mice per Group	Young/Old Ratio			
		>1.5×	>2×	>3×	>4×
10	2	5	3	0	0
10	3	2	0	0	0
10	4	1	0	0	0
10	5	0	0	0	0
20	2	278	9	0	0
20	3	102	0	0	0
20	4	35	0	0	0
20	5	14	0	0	0
30	2	1123	223	16	2
30	3	690	58	1	1
30	4	406	14	0	0
30	5	214	6	0	0
40	2	1921	668	135	48
40	3	1337	287	19	3
40	4	1015	146	5	0
40	5	731	66	4	0
50	2	2465	1071	310	143
50	3	1853	582	103	27
50	4	1479	373	34	8
50	5	1125	198	14	0

per 10,000 would by chance produce false-positive findings even when using a fourfold change as the criterion for acceptance; however, this number falls to near zero when $n = 5$ mice per group. Calculations similar to those shown in Tables 1, 2, and 3 can be used to estimate the number of false positives expected for any given empirical distribution of CVs. We recommend that those groups wishing to report gene array results without formal statistical evaluation of significance should accompany their reports of two- and threefold changes with a comparison table showing the numbers of false-positive results to be expected from their experimental design and observed distribution of CVs.

Criteria Based Upon Formal Significance Testing

An alternate approach is to base conclusions on formal significance testing using a conventional statistical criterion, an idea that is common outside the realm of gene-expression screening but has yet to make much headway among users of this cutting-edge technology. One plausible starting point would be to compute the Student's *t* test statistic for each gene in the set of interest as an index of how likely it would be to obtain the observed distribution of gene expression values by chance alone. Purists would object that it is not possible, for $n < 5$ or so, to check the assumptions on which the *t* test is based (normality and equality of variance), but even they may admit that a statistical test that includes information about interanimal variation is an improvement on ratio-based tests that ignore variance entirely. Genes with low interanimal variation will yield high values (i.e., low probabilities) of the *t* statistic given modest age or genotype effects (two- to fourfold, for example) and deserve more confidence than those in which large intersubject variation produces a nonsignificant p(*t*). Some laboratories specializing in array-based screening are beginning (6) to restrict their conclusions to genes where $p(t) < .05$, the conventional criterion for rejection of the null hypothesis of no effect.

A key problem with a *t*-test–based approach in the context of gene expression screening is that it ignores multiple comparison artifacts. Consider a hypothetical situation in which a postdoctoral scientist decides to measure expression levels of 10,000 genes in each of 20 young and 20 old mice and to make her biological interpretations on the basis of those genes where the age effect is large and consistent enough to reach $p(t) < .05$. Alas, unbeknownst to this researcher, a disgruntled technician has switched the identification codes on all the mice at random, so that the nominally "young" group actually contains an equal number of young and old animals. Among 10,000 genes, however, 1 in every 20 will, entirely by chance, reach $p(t) = .05$; the postdoc, not knowing of the deception, is pleased to find 500 genes that show "significant" age effects, and she makes her interpretation and conducts years of follow-up analyses on the basis of these entirely spurious and unreproducible findings. The problem, well described in most elementary statistics texts, is that a significance criteria of .05 does not protect against false-positive conclusions in a large series of tests.

The Bonferroni procedure is the accepted way to adjust significance criteria in such a situation. When testing 1000 hypotheses simultaneously, for example, one would use as criterion a *p* value of .05/1000 = .00005. Such a criterion is very conservative in the sense that it tends to produce large numbers of false negative conclusions; it tends, in other words, to make it hard to accept as proven hypotheses that are in fact true. If an experiment testing 1000 genes produces *p* values <.00005 for, say, 8 genes, one could confidently conclude that all eight genes are likely to distinguish old from young mice; there would be only 1 chance in 20 that *any* of the eight effects is due to chance alone. Producing such a high *p* value requires either very large numbers of animals or very small interanimal SDs—much smaller than are seen in practical cases. (Evidence that the experimental system in question gives very reproducible values for replicate aliquots of the same sample is not germane; the variation in weight among a set of laboratory members, for example, is not diminished by weighing them on a scale accurate at the microgram level.) If a survey of 10,000 genes shows that 20 of them reach $p(t) = .001$, it is likely that some of these 20 will prove reproducible in subsequent tests, but it is not possible to know which ones without further experimental data.

One way of dealing with this problem is to use a two-stage experimental design. The first stage is used for hypothesis generation: all genes are tested and ranked in order of statistical probability. In a typical case, few if any of the genes will show a sufficiently large age effect, with sufficiently low interanimal variance, to meet the Bonferroni criterion ($p = .000005$ for a set of 10,000 genes), but some are likely to provide suggestive evidence of a real effect, say $p < .001$. The second stage, then, involves testing a separate set of animals, using either the array method or some other convenient test (RT-PCR or RNAse protection assays, for example) for each of these genes that shows the most extreme probabilities in the initial survey. If, for example, the initial screen generates a list of 25 genes where $p < .001$, the second, hypothesis-testing phase of the study can employ a value of $p = .05/25 = .002$ as its criterion for hypothesis confirmation; any genes that reach this level in the second stage can be accepted as age-sensitive, at least in this organ, genotype, and age range.

This method—like any method using small number of animals to examine traits with high variance—is likely to suffer from a high false negative rate: those genes that show above-average interanimal variance will not produce significant *p* values at any stage of the analysis in tests that use only 5 to 10 mice per group. Investigators who have invested considerable effort in large-scale gene scanning surveys may therefore wish to make public—either in a formal report or in an associated electronic archive—lists of genes that show relatively large effects (say two- or threefold changes) even if these do not approach statistical significance; genes that show large effects, even with high interanimal variation, may still deserve further attention if the patterns of expression suggest or refute specific biological theories of interest.

If the cost of the animals (or human samples or cell lines) is relatively small compared with the overall cost of the testing program, it may be useful to carry out the initial first-stage survey using pools instead of individuals. If, for example, a group of 24 young mice can be tested as six pools

of 4 animals, the statistical analysis must treat this as $n = 6$ replicates, but the variation among the six pools is likely to be a good deal less than the variation expected from among six individual animals. Comparing six pools of young samples with six pools of old samples should increase the number of genes that achieve some arbitrary p value (e.g., $p < .001$) in the initial screen, and genes that appear promising in this initial survey can then be retested using fresh samples from individual animals of the age or treatment groups of interest. It may be possible to develop specialized methods for determining the optimal pooling strategy for the initial screening step, but it will be difficult to reduce these to simple rules because the decision will depend on the cost of each assay relative to the cost of preparing each sample and because the optimal pooling plan will differ for genes with different CVs.

Alternate Approaches to Data Interpretation

Expression data sets can be used to address other questions of interest beyond the issue of which specific genes are altered by aging, genotype, or intervention. The most attractive of these involve analysis of groups of genes defined either a priori by their known involvement in biological pathways of interest or a posteriori because they are observed to exhibit similar patterns of expression across sets of related samples. These two approaches are fundamentally different, and each presents its own set of pitfalls, but each deserves (and is receiving) intensive exploration. In the a priori approach, one could begin by assigning each gene to one or many overlapping sets: a gene for Ras, for example, might be placed in sets corresponding to oncogenes, to G proteins, to substrates for PK-C, to regulators of Raf, to genes responding to serum signals, and to many other categories. The expression data can then be used to seek examples of sets that show coordinate response to the intervention, diet, or age effect of interest. This approach, though promising, presents many difficulties, including complexities of feedback pathways, incomplete assessment of genes and pathways of interest, lack of consensus about pathway assignment, and the likelihood that the number of functionally defined gene sets will approach or exceed the number of individual genes and thus fail to reduce the number of statistical comparisons. Testing hypotheses on the basis of sets of functionally related genes requires that investigators specify their gene sets prior to examining their actual data to avoid circularity, and requires that hypotheses be tested not merely by listing genes within categories that do show parallel effects but instead by presentation of categorized genes that do or do not show the expected effects.

In addition, there is a repertoire of methods for defining clusters of genes based on similar (or, more generally, correlated) patterns of responses—for example after antigenic or nutrient stimulation, across different tissue types, or among sets of individual tumors. There is at present, however, no consensus as to which of the many alternative procedures are optimal for extracting biological information from these correlation matrices. Claverie (3) includes an excellent introduction to these approaches with an outline of the problems involved. At their best (7), clustering methods can reveal previously unsuspected relationships among genes not known to exhibit coordinated regulation and can provide new diagnostic tools for sorting individual tumors or individuals on the basis of expression patterns. Effective application of these clustering methods, however, requires not merely sophisticated selection among alternate clustering algorithms, but also very large numbers of tested individuals—numbers well beyond the small sample sizes so far tackled by experimental gerontologists.

Summary

The preceding discussion suggests several guidelines for the design, reporting, and evaluation of experiments that use gene array screening.

1. Bonferroni-adjusted significance criteria should be used to assess the likelihood that a particular effect of age, diet, drug, or mutation may have arisen by chance variation alone.
2. Studies reporting lists of gene expression differences that do not meet these adjusted significance thresholds (e.g., where $.05 > p > .0001$ for a study of 500 expressed genes) should be considered as hypothesis generating (i.e., as a prelude to hypothesis-testing studies in which a small subset of the original gene set is retested on additional samples not evaluated in the initial survey). Because two-stage analyses of this kind are expensive, groups of investigators pursuing similar questions may find it cost effective to combine their datasets, using the work of one group to test hypotheses generated by the other.
3. Publications that include lists of genes selected on the basis of ratio calculations (e.g., young/old ratios) should also report significance tests for each gene for which an effect is postulated. These publications should also include the outcome of a simulation study, such as the one presented in Tables 1, 2, and 3, that provides information about the number of expected false-positive results that would be generated using the numbers of samples (or pools) given the distribution of variance values observed during the study.
4. Reports of differential gene expression should not be published unless they contain either significance tests or, at least, calculated estimates of the number of expected false positives, because without this information it is not possible to evaluate the likelihood that the results represent chance effects alone.
5. Research proposals that include gene array approaches should include a formal power analysis so that reviewers can judge whether the number of samples to be tested is likely to produce statistically significant results for a useful proportion of the genes to be surveyed. Because such a power analysis depends critically on the distribution of variance values among the genes in the array, researchers who compile tables listing these variance levels should consider archiving them in publicly accessible (typically electronic) forms as a guide to others who are considering the use of similar technology in their own work.

Acknowledgments

We thank Phyllis Wise and the two anonymous reviewers for helpful suggestions and discussion, and Igor Dozmorov for the data set used for the

calculations shown in Figure 1 and Table 2. Fred Bookstein originally suggested that explicit simulations could assess the rate of false positive detections as a function of coefficient of variation and threshold signal magnitude and supplied a prototype for Table 1. Support for this work was provided by National Institute on Aging Grant AG13283.

Address correspondence to Richard Miller, Box 0940, University of Michigan, 5316 CCGCB, 1500 E. Medical Center Drive, Ann Arbor, MI 48109-0940. E-mail: millerr@umich.edu

References

1. Lee CK, Klopp RG, Weindruch R, Prolla TA. Gene expression profile of aging and its retardation by caloric restriction. *Science.* 1999;285: 1390–1393.
2. Ly DH, Lockhart DJ, Lerner RA, Schultz PG. Mitotic misregulation and human aging. *Science.* 2000;287:2486–2492.
3. Claverie JM. Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet.* 1999;8:1821–1832.
4. Phelan JP. Genetic variability and rodent models of human aging. *Exp Gerontol.* 1992;27:147–159.
5. Dozmorov I, Bartke A, Miller RA. Array-based expression analysis of mouse liver genes: effect of age and of the longevity mutant Prop1. *J Gerontol Biol Sci.* 2001;56A:B72–B80.
6. Tanaka TS, Jaradat SA, Lim MK, et al. Genome-wide expression profiling of mid-gestation placenta and embryo using at 15K mouse developmental cDNA microarray. *Proc Natl Acad Sci USA.* 2000;97: 9127–9132.
7. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 2000;403:503–511.

Received June 20, 2000
Accepted August 9, 2000
Decision Editor: Edward J. Masoro, PhD