To gain further insight into the basis for the extended longevity and delayed aging of Snell dwarf (dw/dw) mice, we have measured levels of expression of 2352 genes in liver of mice at 6 months of age. We find 60 genes for which the Student's t statistic meets the arbitrary criterion of $p < .001$, and among these 17 meet the Bonferroni-adjusted significance criterion at $p < .05$, which corresponds to a nominal value of $p < .00002$. Using the Bonferroni criterion, we find that dwarf mice show increases in liver mRNA for two mannose-binding lectins, two DNA binding proteins, serum amyloid P component, corticosteroid-binding globulin, and insulin-like growth factor-binding protein 2, as well as decreases in a two phosphodiesterases, a pheromone-binding urinary protein, insulin-like growth factor-I (IGF-I), a calcium-binding protein calgranulin B, a deubiquitinating enzyme, a hydroxysteroid dehydrogenase, a DNA methyltransferase, a glycine transporter, and a placental lactogen. We also use this data set to compare the results of different suggested criteria for evaluating intergroup differences in gene expression. Of the 2352 genes examined, 524 (22%) showed a twofold difference between dwarf and normal mice, but most of these fail to meet the conventional significance criterion of $p < .05$, let alone criteria that have been adjusted to compensate for multiple comparison artifacts. The list of genes that show reliable differences between dwarf and control animals provides new insights into the range of changes induced by deficiencies in growth hormone, thyroid-stimulating hormone, and prolactin, and it will help to guide further studies of the pathways by which these hormone deficiencies contribute to delayed aging in these mutant mice.

As a step toward delineation of the mechanisms by which the dw/dw genotype leads to delayed aging, we have conducted a survey of liver gene expression in young mice of the mutant and control genotypes. Distinguishing real effects of the mutation from the many false positive results that are expected to occur whenever hundreds or thousands of genes are studied in parallel requires careful attention to significance criteria and mitigation of variation caused by technical factors, and we have therefore used this data set to compare a number of analytical strategies. The analysis has produced a list of 17 genes that meet rigorous significance criteria for differential expression in liver of dw/dw compared with control mice, and a further set of 43 other genes for which there is also strongly suggestive evidence for differential expression. This catalog of differentially expressed genes provides a number of interesting leads for further mechanistic exploration and also serves as the first step toward compiling a list of gene expression changes that are presumably shared in common by multiple models of decelerated or delayed aging in mice.

**METHODS**

**Mouse Husbandry**

DW/J-dw/+ females and C3H/HeJ-dw/+ male heterozygote breeders purchased from Jackson Labs were crossed to...
produce the (DW × C3H)F1 male mice used in this experiment. The (DW × C3H)F1 mice included three genotypic groups: dw/dw, dw/+ , and +/+ . Mice of the dw/dw genotype were identified by small body size (dwarfs), and the other two genotypes, which are not phenotypically distinguishable, were considered as nonmutant controls (+/?). The mice (all male) were housed in microisolator cages with 1/8 in. (0.31 cm) Bed-O-Cob, and they were given tap water and 5001 Rodent Chow ad lib; moist 5001 Rodent Chow was added to cages housing dw/dw mice. Nonmutant control mice were housed 2/cage. Dwarf mice were housed in cages that contained two nonmutant female mice (to avoid thermal stress) and two dwarf animals per cage.

Half of the mice used in this experiment received daily injections of porcine growth hormone (National Hormone Pituitary Program) at 50 µg/injection plus L-thyroxine (Sigma Catalog T0397, Sigma, St. Louis, MO) at 2 µg/injection in 100 µl of sterile saline, 5 days per week for an 11-week period beginning at approximately 30 days of age. The other half of the mice received saline injections according to the same schedule. Mice were euthanized at approximately 6 months of age, that is, 10 weeks after the last hormone injection.

Preparation of Labeled cDNA Targets and Hybridization to Immobilized Probe Sets

Total RNA was extracted from each liver by using the Atlas Pure Total RNA Isolation Kit (Clontech, Palo Alto, CA) following the vendor’s protocol. The RNA was digested with RNase-free DNase I to remove genomic DNA contamination. To prepare labeled cDNAs, we used the ATLAS cDNA Expression Array Kit (Clontech) following the recommended protocol in all steps for reverse transcription, 32P-labeling, and hybridization. Briefly, in each case 2–5 µg of total RNA were converted into 32P-labeled first strand cDNA by means of Moloney Murine Leukemia Virus reverse transcriptase. The purification of the labeled cDNA from unincorporated 32P-labeled nucleotides was achieved with CHROMA SPIN-200 column chromatography (Clontech, Palo Alto, CA). cDNA-fractions of highest activity were pooled and hybridized to the Mouse Atlas 1.2 Array and 1.2 Array II membranes, each of which contains 1176 spotted mouse cDNA fragments as well as control spots. After prehybridization (30 minutes at 68°C in ExpressHyb; Clontech) supplemented with 100 µg/ml of sheared salmon testes DNA (Sigma), the heat-denatured cDNA target preparation was added. Hybridization occurred overnight at 68°C with continuous rolling agitation. Membranes were washed 4 × 30 minutes in 2 × standard sodium citrate (SSC)/1% sodium dodecyl sulfate (SDS) at 68°C followed by two washes in 0.1 × SSC/0.5% SDS (30 minutes, 68°C). Membranes were sealed in sample bags (Wallac, Turku, Finland), exposed to storage phosphor screen for 1 to 3 days, and evaluated with a phosphorimager (Molecular Dynamics, Sunnyvale, CA).

Data Reduction

The digital images from the phosphorimager were processed by using the ArrayVision program (Imaging Research, St. Catharines, Ontario, Canada) to generate back-ground-subtracted pixel volumes for each of the 1176 spots on each membrane. For a small number of spots (about 2% of the total), the background subtraction produces a negative value because the spot in question is adjacent to another spot whose exceptionally high intensity produces an artifically high background reading; these spots are considered to be missing data and are for convenience arbitrarily assigned a value of 1 pixel unit. Because each set of labeled cDNA targets is exposed to two different sets of 1176 probes, the raw data set consisted of 2352 values for each of the 16 mice. Each value was then transformed to its common logarithm to avoid undue influence of the small number of very intense spots in the subsequent steps of analysis.

Normalization

We compared two normalization methods to reduce technical variation prior to significance testing.

Method 1.—This is adjustment by means. We calculated the median expression level across all 16 mice for each of the 1176 genes on the Clontech Atlas Mouse 1.2 membranes; this produces a “standard” set of 1176 median values. We then calculated the mean expression level for all 1176 genes for each of the 16 mice, as well as the mean expression level for the set of 1176 medians in the standard set. Then, for each mouse, we calculated the “ratio of means,” that is, the mean expression level for that mouse divided by the mean expression level for the standard data set. Lastly, we calculated the “adjusted” values for each data point (16 × 1176 points) as the unadjusted (observed) value divided by the ratio of means for the mouse in question. This process was repeated separately for the second membrane variety (Clontech Mouse 1.2 II membranes, with an additional 1176 probes), and the two sets of adjusted values (2352/mouse) were concatenated for analysis. This process has the effect of forcing the mean level of adjusted values for each mouse to the same level for each of the two probe sets, and it is similar to the method incorporated into many commercial software programs for comparing pairs of raw expression levels.

Method 2.—This is linear regression by ordinary least squares. “Standard” data sets were calculated for each of the two sets of 1176 probes, as in Method 1, using the median expression level across each 16 mice. Then for each probe set and for each mouse we calculated the intercept, a, and slope, b, by linear regression for the equation:

\[ Y = a + b X \]

where Y represents the standard data set and X represents the unadjusted values for an individual mouse. Lastly, for each mouse, we calculated the adjusted values for each data point by using the following formula:

\[ X\text{-adjusted} = a + b X \]

This process has the effect of minimizing the sum of the squared deviations between each individual mouse and the
standard data set derived from the same set of 1176 probes. In this context, however, the adjustment is an empirical one, because the actual data sets do not meet the assumptions underlying linear regression calculations; in particular, the residuals are not independent and are much higher for points with low intensity than for points of high intensity, because the relative level of measurement error is much higher for points that are close to background levels. Unless otherwise stated, all data shown and all significance tests presented used this method of normalization.

Assessment of Statistical Significance

For each gene we calculated the two-tailed Student’s t statistic comparing the eight normal to the eight dwarf mice by using an algorithm that assumes homoscedasticity, that is, equal variance between the two groups of mice. This method is slightly less conservative than a calculation that assumes heteroscedasticity, but it was chosen for consistency with the “false discovery rate” statistic (see point 3 in the following paragraphs) and with published methods for permutation-based significance thresholds, such as that used by Callow and colleagues (4); the assumption of homoscedasticity is also built into the SAS MULTTEST (SAS, Cary, NC) procedure used as described in the paragraphs that follow.

We considered four approaches to adjusting statistical significance to correct for the simultaneous testing of multiple hypotheses. These are presented in increasing order of conservatism (i.e., decreasing risk of Type I error and increasing risk of Type II error as observed in our data set).

1. Given adequate statistical power, a survey of 2352 genes should generate only a small number (2–3) of false positive results by using the arbitrary significance criterion that \( p(t) < .001 \). This criterion was previously adopted in studies of differential gene expression in breast cancer biopsy samples (5). This cannot be considered as an experimentwise error rate, however; in tests of a sufficiently large number of genes, some are likely to produce \( p < .001 \) by chance alone.

2. We also calculated the false discovery rate (FDR) \( p \) value for each gene (6). By definition, 5% of the genes that are found to have an FDR \( p \) value of \( < .05 \) are likely to represent Type I errors. Thus this criterion does include adjustments for the number of genes tested in parallel, but it does not provide a true experimentwise error rate. The FDR method controls the FDR only with independent \( p \) values that are uniformly distributed under null hypotheses of no gene effect.

3. We calculated an experimentwise \( p \) value for each gene by using the Bonferroni method. Because our survey initially included assessments of 2352 genes, this corresponds to using a threshold of 0.05/2352 = 0.00002 as significance criterion. This procedure produces a true experimentwise significance threshold, in that the entire collection of genes with adjusted \( p < .05 \) has only a 5% chance of containing a false positive result. The method is, however, very conservative (high Type II error rate), in part because it does not incorporate distributional characteristics or correlations among multiple variables (7), and in part because many individual genes were not evaluable in practice because of low signal strength.

4. We used a permutation approach (a component of the SAS MULTTEST procedure) to estimate an experimentwise significance criterion based on the empirical distribution of observed values. This method begins by generating a large series of permuted data sets in which the linkage between the sets of gene expression levels and the group identification (dwarf or control) is disrupted to correspond to the null hypothesis. For each permutation, levels of the \( t \) statistic are calculated for each of the 2352 genes, and the maximum value of the \( t \) statistic across the entire collection of genes is recorded. This procedure is repeated many times (20,000 permutations in the current case) to produce an empirical distribution of the \( t \) statistic given the null hypothesis. Each value of \( t \) calculated from the actual data set is then evaluated for significance by comparison with the empirical set of experimentwise maximum \( t \) statistics. This procedure is analogous to that used to evaluate significance thresholds for gene/trait linkage in gene mapping experiments (8).

Semiquantitative RT-PCR

cDNA synthesis and target amplification was accomplished in a single step, starting with 200 ng of total RNA by using the LightCycler-RNA Amplification Kit SYBR Green I (Roche, Indianapolis, IN) and the LightCycler System (Roche) following manufacturer’s protocols. Primers were at 0.4 \( \mu \)M each. The Taq polymerase was preincubated for 10 minutes at 4°C with anti-Taq antibody (Clontech). Target amplification reactions were cycled 50 times with a 95°C denaturation for 0 seconds, a 56–58°C annealing for 8 seconds, and a 72°C extension for 12 seconds; there were slopes of 20°C/s for all steps. The LightCycler analysis software was used for the relative quantification of data.

Results

A Catalog of Genes Differentially Expressed in Liver of Young Adult Dwarf and Control Mice

We measured liver gene expression levels in samples from eight normal and eight dwarf mice. Half of the animals in each group had been treated by 11 weeks of treatment with GH and thyroxine, followed by a 10 week wash-out period without hormone treatment. Because we found no evidence for significant effects of hormonal treatment on liver gene expression (see the paragraphs that follow), the data sets for treated and untreated mice were pooled for further analysis. After normalization by least-squares regression, a Student’s \( t \) statistic was calculated for each gene as a measure of the degree to which gene expression was altered by the \( dw/dw \) genotype. Table 1 presents a list of the 60 genes for which \( p(t) < .001 \). A complete list of results for each gene can be obtained upon request to the corresponding author.

There are several available methods for evaluating the statistical significance of these effects, that is, the probability that the apparent effect of the mutant gene is due to chance variations in random sampling alone. Unadjusted, comparisonwise \( p \) values less than .001 are expected to
Table 1. Genes That Show Altered Expression Levels in Snell Dwarf Mouse Liver at 6 Months of Age

<table>
<thead>
<tr>
<th>D*</th>
<th>N</th>
<th>Ratio (D/N)</th>
<th>p(t)‡</th>
<th>FDR</th>
<th>Description</th>
<th>GenBank</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>13</td>
<td>0.23</td>
<td>0.00000</td>
<td>0.000</td>
<td>phosphodiesterase 6D, cGMP-specific, rod, delta</td>
<td>AF0464000</td>
</tr>
<tr>
<td>69</td>
<td>1</td>
<td>48</td>
<td>0.00000</td>
<td>0.000</td>
<td>major urinary protein 1</td>
<td>M16355</td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>25</td>
<td>0.00000</td>
<td>0.000</td>
<td>zinc finger protein 162</td>
<td>X85802</td>
</tr>
<tr>
<td>39</td>
<td>445</td>
<td>0.09</td>
<td>0.00000</td>
<td>0.000</td>
<td>mannose-binding lectin, serum (C)</td>
<td>D11463</td>
</tr>
<tr>
<td>90</td>
<td>23262</td>
<td>0.00</td>
<td>0.00000</td>
<td>0.000</td>
<td>major urinary protein 1</td>
<td>M16355</td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>25</td>
<td>0.00000</td>
<td>0.000</td>
<td>zinc finger protein 162</td>
<td>X85802</td>
</tr>
<tr>
<td>39</td>
<td>445</td>
<td>0.09</td>
<td>0.00000</td>
<td>0.000</td>
<td>mannose-binding lectin, serum (C)</td>
<td>D11463</td>
</tr>
<tr>
<td>90</td>
<td>23262</td>
<td>0.00</td>
<td>0.00000</td>
<td>0.000</td>
<td>major urinary protein 1</td>
<td>M16355</td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>25</td>
<td>0.00000</td>
<td>0.000</td>
<td>zinc finger protein 162</td>
<td>X85802</td>
</tr>
<tr>
<td>39</td>
<td>445</td>
<td>0.09</td>
<td>0.00000</td>
<td>0.000</td>
<td>mannose-binding lectin, serum (C)</td>
<td>D11463</td>
</tr>
<tr>
<td>90</td>
<td>23262</td>
<td>0.00</td>
<td>0.00000</td>
<td>0.000</td>
<td>major urinary protein 1</td>
<td>M16355</td>
</tr>
</tbody>
</table>

Notes: D = dwarf; N = normal; FDR = false discovery rate. Boldface indicates genes where \( p < .00002 \), i.e., below the Bonferroni-adjusted significance criterion.

*Mean value for signal strength (in arbitrary units) in digitized images for the D mice for the gene indicated.
†Mean value for signal strength (in arbitrary units) in digitized images for the N control mice for the gene indicated.
‡Nominal, i.e., unadjusted, \( p \) value from \( t \) test for the gene indicated.
GENE EXPRESSION IN DWARF MICE

emerge by chance alone two to three times in a sufficiently high-powered analysis of 2352 genes. Our observations produce \( p < .001 \) for 60 of the genes surveyed, and it thus seems likely that very few of the genes in Table 1 have achieved this \( p \) value by chance alone. By contrast, a parallel calculation comparing the eight hormone-treated mice to the eight untreated mice yielded \( p(t) < .001 \) for only three genes (myelin basic protein expression factor 2 at \( p = .0003 \), carbonic anhydrase at \( p = .0002 \), and filensin at \( p = .0004 \)). Some or all of these three apparent hormone effects may well be the false positives expected in such a large series of comparisons.

A more conservative approach seeks to limit false positive results by calculating experimentwise \( p \) values. The conventional method is to base conclusions on the Bonferroni adjustment. In a survey of 2352 genes, the Bonferroni threshold is 0.05/2352 = .000021. Seventeen of the genes in Table 1 meet this experimentwise criterion, and one can thus conclude with 95% confidence that none of these 17 genes represents chance effects alone. Some groups prefer to eliminate from consideration genes for which the expression level is at or below background, prior to Bonferroni correction. Such an approach would have little effect in the current case: approximately half of the genes in our array are near the background value, and a revised significance criterion at \( p = .00004 \) would add only two genes to the list of those judged as significantly affected.

We also used a permutation-based method to generate an empirical distribution of the maximum values of the \( t \) statistic produced by permuted data sets in which the assignment of each mouse’s data set to control or dwarf groups is done at random. The observed values are then compared with this large set of maximum \( t \) values to determine how frequently the observed values are likely to have emerged by chance. For our current data set this permutation-based approach, with or without the step-down modification recommended by other groups (9), was no less conservative than the Bonferroni method (not shown).

Figure 1 shows the distribution of gene expression levels among individual mice for each of the 17 genes for which \( p(t) \leq .00002 \) in Table 1. For these genes there is little or no overlap between control and \( dw/dw \) mice in expression levels.

Table 1 also shows calculated values for the FDR \( p \) value. These values are calculated in such a way that an FDR < .05 for a specific gene supports the conclusion of differential expression with 95% confidence for that gene. Such a criterion is adjusted for the number of genes tested, but it does not provide an experimentwise confidence level, because a list of genes where FDR < .05 is likely to contain a proportion of false positives. There are in our data set 72 genes for which FDR < .05, of which the first 60 are shown in Table 1.

An analysis of 2352 genes is likely to produce approximately 24 false positive examples for which \( p < .01 \) by chance alone; our data set contains 139 examples (not shown) with nominal \( p < .01 \), prior to any adjustments for multiple comparisons. Thus approximately 83% (115) of these 139 genes, including the 60 of Table 1 and another 54 genes not shown, are likely to reflect true effects of the \( dw/dw \) genotype.

Confirmation of Selected Results by Semiquantitative PCR

To rule out systematic errors in array production or sample handling, we retested 11 of the genes listed in Table 1 by using an alternate method, real-time polymerase chain reaction (PCR) amplification of cDNA. In each case we tested samples from three normal and three dwarf mice, and we then calculated the ratio (dwarf/normal) of the mean values from the PCR data. Table 2 summarizes these data. In two cases the amplification step did not produce detectable product within the first 30 cycles; these are listed as “not detectable” in the table. In one other case, the PCR results were highly variable among the dwarf samples (values of 1.8 and 1500 arbitrary units; compared with a mean of 13 ± 7 units in normals). Among the remaining eight genes, seven produced ratios that were consistent with those seen in the array data set, and one (serum mannose-binding lectin) produced a ratio clearly inconsistent with the array result. Further tests, using other primer sets and/or other detection methods, will be needed to see which of the two methods for detecting the mannose-binding lectin mRNA is in error. These data suggest that most, though not all, of the ratios listed in Table 1 are likely to be reasonably good estimates of the real relative mRNA levels in the set of samples tested.

Comparison of Normalization Methods

Detection of differences in gene expression levels among groups is complicated by technical variations that increase variation among individual experiments and can therefore hide real differences that are due to the experimental conditions (in this case the \( dw/dw \) genotype). Normalization methods attempt to correct for differences in technical factors (such as target-specific activity, length of exposure, and quality of extracted mRNA) that could contribute to differ-

---

Figure 1. Differential liver gene expression in dwarf and nonmutant control mice: 17 examples. Each column represents a different gene, ranked (left to right) in the same order as the first 17 lines of Table 1. Each point shows the intensity for expression of that gene (normalized by least-squares regression) in an individual dwarf or control mouse. Note that for the third gene (major urinary protein) all values for normal mice are off scale, i.e., greater than 10,000 arbitrary units.
ences in apparent expression levels but do not reflect real differences among samples in mRNA abundance. The simplest and most common of these, incorporated into many statistical packages for analysis of array data, involves adjusting each array in a pair (or larger set of samples) to the same mean level of signal intensity. We compared the results of this mean-adjustment normalization with results obtained by using a least-squares regression approach. We found that the mean-adjustment procedure would have led to acceptance of significance (using the Bonferroni criterion) for 10 genes, instead of the 17 accepted by using the least-squares regression method. Although in each of these 10 cases the least-squares approach produced a value of \( p < 0.001 \), the rank order of genes produced by the two methods was somewhat different. We have also evaluated a series of other methods for normalization, including robust regression and regression based on genes expressed at highest levels of intensity, and we find that they do not offer a major improvement above the ordinary least-squares regression, at least for our current data set (not shown). Additional work, both empirical and theoretical, will be needed to develop optimal methods for normalization for data sets of this kind.

**Comparison of Selection Criteria**

The development of catalogs of differentially expressed genes is typically a first step in the elucidation of molecular pathways that discriminate among the groups of interest, in this case dwarf versus normal mice. Selection of genes that deserve further follow-up studies or deserve to be incorporated into theoretical schema can be based on a variety of criteria. We have argued elsewhere (10) that the use of selection criteria based on the ratio of expression levels alone, that is, without calculation of a test statistic that includes information about interanimal variance, can make it difficult to distinguish reliable findings from those that are due only to chance effects and random variation. To explore this idea further, we plotted the \( p \) value (testing differences caused by the \( dw/dw \) genotype) for each of our 2352 genes against the ratio of expression levels. The results are shown in Figure 2, in which ratios less than 1 have been converted to their reciprocals. We found 524 genes (22%) in which the expression ratio exceeded 2, among which the majority (266) would not have been considered statistically significant (\( p < 0.05 \)) even without corrections for multiple comparison artifacts. If a value of \( p < 0.001 \) is taken as the criterion for significance, then only 55 of the 524 genes with ratio > 2 would meet this criterion, as would six other genes with ratio < 2.

Figure 3 presents these data in a different form, showing what fraction of the genes with ratios above specific levels would be judged significant by using either conservative (\( p < 0.001 \)) or riskier (\( p < 0.01 \)) significance criteria. It is clear that only a small fraction of the genes that achieve ratios between 2 and 3 would meet either of these two significance criteria. Even for ratios between 3 and 5, over half of the genes fail to meet the least restrictive threshold (\( p < 0.01 \)). Ratio-based criteria would be expected to perform even less well in experiments based on tests of fewer than eight individuals per group.

![Figure 2. Scatterplot showing \( p \) value (comparison of \( dw/dw \) to normal) as a function of expression ratio for each of 2352 genes. The horizontal axis shows dwarf:normal ratio for those genes in which the ratio exceeds 1, and the normal:dwarf ratio for all other genes. Reference lines indicate genes in which the ratio > 2, and in which \( p(t) = 0.01 \) or 0.001.](image)

**Table 2. Semi-Quantitative PCR Results Compared With Array Results**

<table>
<thead>
<tr>
<th>GenBank</th>
<th>Description</th>
<th>Ratio (Array)*</th>
<th>Ratio (RT-PCR)†</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF046000</td>
<td>phosphodiesterase 6D, cGMP-specific, rod, delta</td>
<td>0.23</td>
<td>0.02</td>
<td>Confirmed</td>
</tr>
<tr>
<td>D11440</td>
<td>mannose-binding lectin, serum (C)</td>
<td>48</td>
<td>0.20</td>
<td>Not confirmed</td>
</tr>
<tr>
<td>M16355</td>
<td>major urinary protein 1</td>
<td>0.00</td>
<td>0.00</td>
<td>Confirmed</td>
</tr>
<tr>
<td>X885082</td>
<td>zinc finger protein 162</td>
<td>25</td>
<td>NA</td>
<td>Not detectable</td>
</tr>
<tr>
<td>X04480</td>
<td>insulin-like growth factor IA (IGF-IA); somatomedin</td>
<td>0.09</td>
<td>0.01</td>
<td>Confirmed</td>
</tr>
<tr>
<td>M83219</td>
<td>S100 calcium-binding protein A9 (calgranulin B)</td>
<td>0.13</td>
<td>NA</td>
<td>Not detectable</td>
</tr>
<tr>
<td>U29762</td>
<td>albumin D box-binding protein</td>
<td>3.1</td>
<td>2.0</td>
<td>Confirmed</td>
</tr>
<tr>
<td>L41519</td>
<td>hydroxysteroid dehydrogenase-5, delta&lt;5&gt;-3-beta</td>
<td>0.04</td>
<td>0.00</td>
<td>Confirmed</td>
</tr>
<tr>
<td>U11021</td>
<td>phosphodiesterase 7A</td>
<td>0.23</td>
<td>0.51</td>
<td>Confirmed</td>
</tr>
<tr>
<td>X67056</td>
<td>solute carrier family 6 member 9 (SLC6A9); sodium- and chloride-dependent glycine transporter 1 (GLYT1)</td>
<td>0.4</td>
<td>NA</td>
<td>High variance</td>
</tr>
<tr>
<td>M16355</td>
<td>major urinary protein 1</td>
<td>0.00</td>
<td>0.00</td>
<td>Confirmed</td>
</tr>
<tr>
<td>D11440</td>
<td>mannose-binding lectin, serum (C)</td>
<td>48</td>
<td>0.20</td>
<td>Not confirmed</td>
</tr>
<tr>
<td>M83219</td>
<td>S100 calcium-binding protein A9 (calgranulin B)</td>
<td>0.13</td>
<td>NA</td>
<td>Not detectable</td>
</tr>
<tr>
<td>U29762</td>
<td>albumin D box-binding protein</td>
<td>3.1</td>
<td>2.0</td>
<td>Confirmed</td>
</tr>
<tr>
<td>L41519</td>
<td>hydroxysteroid dehydrogenase-5, delta&lt;5&gt;-3-beta</td>
<td>0.04</td>
<td>0.00</td>
<td>Confirmed</td>
</tr>
<tr>
<td>U11021</td>
<td>phosphodiesterase 7A</td>
<td>0.23</td>
<td>0.51</td>
<td>Confirmed</td>
</tr>
<tr>
<td>X67056</td>
<td>solute carrier family 6 member 9 (SLC6A9); sodium- and chloride-dependent glycine transporter 1 (GLYT1)</td>
<td>0.4</td>
<td>NA</td>
<td>High variance</td>
</tr>
</tbody>
</table>

Notes: PCR = polymerase chain reaction; RT = reverse transcriptase.

* Dwarf/Normal ratio, from Table 1, for the indicated gene, using 8 samples from each genotype.

† Dwarf/Normal ratio from semi-quantitative real-time PCR reactions, using 3 samples from each genotype.
Design Considerations

Design of a study to discover differences in gene expression between two groups of subjects requires decisions about the number of samples to include in each group to generate a suitably large set of genes worth follow-up study. This question in turn depends on the variation among individual subjects, which will be higher for some genes than for others. Formal power calculations depend on prior information about the standard deviations among individual subjects tested for each of the genes in question. To provide some guidance for the design of future studies, we calculated the coefficient of variance (CV = standard deviation divided by the mean) for each gene in our data set, averaging these for the two groups of mice, and in Figure 4 present a plot of the CV as a function of intensity, with more intense spots toward the left of the figure. In this presentation the vertical axis shows CV as a rolling average of 50 genes; the first point, for example, shows the average CV of the genes ranked 1–50 in brightness, and the second point shows the average CV of genes ranked 2–51, and so on. It is clear from the comparison of the two lines that regression-based normalization generates lower CVs for genes throughout the range of intensities, compared with normalization based on mean expression levels alone. The data in the figure will be useful for estimating the number of samples needed to provide adequate statistical power for hypothesis testing. The brightest quarter of the genes in this experimental system, for example, typically have standard deviations that are between 7% and 30% of their mean expression levels. There is, however, substantial variation among genes even at the bright end of the scale; among the 100 genes with the highest median intensity, for example, CVs ranged from 4% to 61% (not shown). As expected, the average CV rises as the spot intensity falls, because for dimmer spots there is an increased ratio of measurement noise to biological information.

Discussion

Depending on the genetic background, gender, and vivarium setting, Snell and Ames dwarf mice live approximately 40–70% longer than nonmutant control animals (1,2). It is not yet clear which of the several endocrinological abnormalities that distinguish these mutants from controls are responsible, individually or jointly, for their extended survival or for the postponement of nonlethal diseases, immune changes, and age-dependent changes in collagen cross-linking (1) that together suggest their usefulness as a model of delayed or decelerated aging. The primary endocrine changes have been described in detail elsewhere (3,11), and they include diminished GH, TSH, and prolactin, with sec-

Figure 3. Stacked bar plot showing proportions of genes at differing significance levels for various levels of the dwarf/normal ratio. As in Figure 2, reciprocal ratios are used where expression of the gene in normal mice is greater than that in dwarf mice. Bars show, left to right, genes with varying ratios, given by the horizontal axis. The varying fill patterns indicate the number of genes (at the given ratio) for different p values.

Figure 4. Coefficients of variation (CV) increase as spot intensity diminishes. The CV is calculated by taking the average CV for the two groups of mice (i.e., CV for dwarfs plus CV for normals, divided by 2) and then computing a rolling mean over each set of 50 genes in rank order by intensity. One curve reflects the data set generated by the mean-adjustment procedure; the other shows the regression adjustment used throughout this paper.

Figure 5 illustrates the implications of the increase of noise with lower signal strength for this kind of experiment. Low-intensity spots, for which the signal-to-noise ratio is expected to be low, typically do not yield strong (p < .001) or even modest (p < .01) evidence for intergroup effects, presumably because variation is so high that authentic differences among mice are hard to detect. Spots of moderate intensity (for example, between 5 and 50 units in Figure 5) do generate convincing p values, but in most cases only when the expression ratio (dwarf/normal or vice versa) is quite high. For spots whose intensity is lower than 10 units, for example, achieving statistical significance at p < .001 was seen only for ratios above (and typically well above) 4:1. Only when intensity is quite high (e.g., above 50 units in Figure 5) do spots begin to generate significant results even for ratios as low as 2–3 (or occasionally even lower). This figure also illustrates the point that ratios above 2 are produced by many genes that do not meet significance criteria (see Figures 2 and 3), and that these potential false positives are more likely to emerge for genes with a low intensity and thus a higher measurement error.
ondary deficits in IGF-I and thyroid hormones. The extended longevity seen in the “little” mutant (lit/lit, defective for GH releasing hormone receptor Ghrhr) as well as in mice with loss of function mutations in the GH receptor (1,12) implies that deficits in GH-dependent pathways may contribute to extended survival in the dw/dw and df/df dwarfs. Such a model is consistent with the observations of diminished early life IGF-I levels (13) in CR mice, as well as with increased longevity in small breeds of dogs (14,15) that owe their diminutive stature to deficits in IGF-I production (16,17). Snell and Ames dwarf mice do, however, show a greater proportional increase in life span than do the lit/lit and GH-receptor knock-out mice, suggesting that the deficits in thyroid hormones and/or prolactin in the former mutants may also contribute to longevity extension in these animals.

The present set of results is the first stage in a plan to seek out alterations in gene expression that typically distinguish long-lived from closely related short-lived mice. Our ultimate goal is to construct lists of genes that discriminate dwarf from nondwarf mice, genes that discriminate CR mice from mice fed ad libitum, genes whose expression levels are altered in mice with longevity-extending alleles at quantitative trait loci, and so forth, and then to compare these lists to compile a catalog of genes whose expression levels are always higher or always lower in the longer-lived member of each comparison. Genes that are overexpressed in dwarf mice, in CR mice, in mice with favorable gene combinations, and in other long-lived models (18) will deserve special attention as potential regulators of the pace of the aging process, or at least as indicators of the authentic age-retarding processes. Along the way, the compilation of each list is likely to provide clues to the cellular and intercellular pathways that connect specific genes, diets, and allele combinations to postponed aging.

We chose to begin our analysis by using samples from young adult mice. We believe that samples from young adults are more likely to provide insights into the molecular controls of aging than are samples from older subjects. We conceive of aging as a process that turns young animals into old ones, and for this reason we suspect that differences in gene expression in early adulthood or in middle age are likely to prove responsible for the changes in tissue structure and function that are the result of aging. Studies of patterns of gene expression in aged animals are likely to reflect the effects of the many deleterious and compensatory processes that accompany old age (including the diseases of aging), and studies of exceptionally old animals will be further compounded by selection artifacts as less hardy individuals are removed from the cohort by lethal illnesses. Thus patterns of gene expression derived from old individuals are likely to be complex and multifactorial. We think that catalogues of age-dependent changes in gene expression, once they have been produced by using adequate numbers of subjects and appropriate statistical procedures, will provide valuable clues about the relationships connecting aging to illnesses, and help to test specific hypotheses about organ- or cell-specific pathophysiology. In contrast, information about early-life patterns of gene expression will be needed to clarify the ways in which specific mutant alleles, drugs, or dietary interventions create adult animals that age slowly.

The pursuit of such a scheme requires careful attention to analytical approaches that discriminate real—reliable and reproducible—differences in gene expression patterns from apparent differences that merely represent chance effects that lead to false positive (Type I) errors. Such errors are particularly likely to plague studies in which hundreds or thousands of genes are examined simultaneously, because conventional significance criteria (such as p values < .05) do not provide adequate protection in such a series of multiple comparisons. In an adequately powered study of 2352 genes, for example, about 118 genes would be expected to attain p < .05 even if the groups did not differ at all in gene expression patterns. Avoidance of such errors requires conservative adjustment of significance criteria, but high thresholds for acceptance of statistical significance in their turn lead to high levels of false negative conclusions unless the study design includes an appropriately high number of individual samples drawn from each of the tested populations.

Using a design with n = 8 mice in the control and experimental groups, we have compiled in Table 1 a listing of genes that seem most likely to be expressed at different levels in liver of young adult Snell dwarf and nonmutant control animals. Seventeen of these genes exceed the Bonferroni-adjusted significance level of p < .05 (i.e., nominal value of p < .00002 in a study of 2352 genes). A total of 60 genes meet a less stringent criterion (nominal p < .001) for which only two to three false positives are expected. Seventy-two genes meet a still less stringent criterion: for these genes, the FDR value has p < .05, suggesting that for each of these examples, considered individually, there is not more than a 5% chance that the result is due to chance alone. Other significance tests, based on permutation-based estimation of empirical significance thresholds, produced
results similar to those of the Bonferroni approach in this data set and have not been presented in detail.

It is important to confirm array results by using independent methods that are not subject to the same set of technical artifacts that might lead to systematic errors in estimation of mRNA levels, including, for example, the use of incorrectly identified arrays or features. Table 2 shows a selection of genes from Table 1, chosen to include some genes with either large or relatively small differential expression ratios, and to include ratios both above and below 1. Of the eight genes for which primers gave detectable and consistent signals in the PCR assays, seven showed ratios that were qualitatively in good accord with the values calculated from the array data. Perfect agreement would not be expected, in part because of differences in assay method but also because the PCR experiments used only 6 of the 16 samples used for the array experiments. Data of this kind do not provide independent support for the differential expression levels listed in Table 1, which would require evaluation of additional mice of both genotypes, but they do suggest that for most of the genes tested, the relative measures of RNA abundance estimated by the array method are not far from relative estimates that would have been derived by using alternate techniques.

Inferences based on the results shown in Table 1 (and in the supplementary results table that can be obtained from the authors) must be made with caution for several reasons. First, we have sought confirmatory evidence, based on real-time PCR, for only a small fraction of the mRNA species in the table. Although our limited data suggest that many of the ratios shown in Table 1 may well be confirmable by using other methods for mRNA quantitation, researchers wishing to pursue specific molecular leads will want to confirm the mRNA data in their own samples. Second, at present we do not know to what extent these differences in mRNA levels will be accompanied by parallel changes in the encoded proteins. Third, more work will be needed to see whether the genes in question show similar expression patterns in other tissues and in mice at older or younger ages. In addition, our initial survey was limited to the 2352 genes available in the commercial arrays we chose for this work, and an analysis of larger numbers of mouse genes will undoubtedly add new and equally provocative examples to the list.

We assume (and predict) that some of the particular genes listed in Table 1 will be found to show differential expression in multiple models of postponed aging, and that others will eventually be shown to reflect regulatory processes characteristic of **dw/dw** and closely related mutants but not of all slow-aging mice. A great deal of additional work will be needed to formulate and then test models that link these observations to one another either with respect to physiological sequences or with respect to common pathways of molecular controls. Some of the gene expression differences are easy to rationalize on the basis of known hormonal effects. The low levels of mRNA for IGF-I and of the pheromone-binding major urinary protein 1, for example, are both well-documented effects of low GH levels. The list of affected genes includes a number of DNA binding factors and DNA metabolizing enzymes (albumin D-box binding protein, zinc finger protein 162, DNA methyltransferase 3B, a subunit of DNA primase) that might be expected to regulate secondary cascades of altered gene expression. Several of the implicated mRNAs are thought to be involved in innate protection against bacterial pathogens (lipopolysaccharide-binding protein, and the two mannos-binding lectins). Some of the genes are best known for their roles in nonhepatic cells (such as the Ia invariant chain, tektin 1, an intestinal fatty acid binding protein, and the olfactory receptor gene), and further work will be needed to determine whether their apparent presence in mouse liver lysates reflects technical artifact, cellular heterogeneity, or authentic production by hepatocytes.

Possible changes in stress-resistance mechanisms would be of particular interest from a gerontologic perspective, but we note that of the 14 heat-shock proteins or cognates included in the data set, only one, GRP78 (also known as glucose-related protein, Hsce-70 and heat-shock 70 kD protein), shows an FDR p value of < .05, and GRP78 is expressed at twofold higher levels in normal than in dwarf mice. Among the other 13 members of this family, the 86 kD protein Hsp90 may also be expressed at twofold higher levels in normal mice (nominal p = .012), and Hsp60 (also known as GroEL) may be expressed at 2.7-fold higher levels in normal mice (nominal p = .012). Thus the data do not provide any evidence for the idea that elevation of heat-shock proteins, at least in liver, contributes to the longevity and disease resistance of the dwarf mice.

Producing a list of differentially expressed genes, such as that shown in Table 1, is intended to serve as a stimulus to further investigations in laboratories that specialize in the biochemical and cellular pathways—responses to intercellular signals, alterations in hormone responses, control of sets of coregulated genes, alterations in intermediary metabolism—in which the listed proteins are involved.

**Acknowledgments**

This work was supported by Nathan Shock Center Grant AG13283, Claude Pepper Center Grant AG08808, and Grant R01-AG040818.

Address correspondence to Dr. Richard A. Miller, University of Michigan School of Medicine, Box 0940, 5316 CCGCB, Ann Arbor, MI 48109-0940, E-mail: miller@umich.edu

**References**


Received August 21, 2001
Accepted October 28, 2001
Decision Editor: Edward Masoro, PhD