nature genetics

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

Comparing genomic expression patterns across species identifies shared transcriptional profile in aging

Steven A McCarroll[1], Coleen T Murphy[2], Sige Zou[3], Scott D Pletcher[4,11], Chen-Shan Chin[2], Yuh Nung Jan[1–3,5–7], Cynthia Kenyon[1,2,5,6], Cornelia I Bargmann[1,5–8] & Hao Li[2,9,10]

We developed a method for systematically comparing gene expression patterns across organisms using genome-wide comparative analysis of DNA microarray experiments. We identified analogous gene expression programs comprising shared patterns of regulation across orthologous genes. Biological features of these patterns could be identified as highly conserved subpatterns that correspond to Gene Ontology categories. Here, we demonstrate these methods by analyzing a specific biological process, aging, and show that similar analysis can be applied to a range of biological processes. We found that two highly diverged animals, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*, implement a shared adult-onset expression program of genes involved in mitochondrial metabolism, DNA repair, catabolism, peptidolysis and cellular transport. Most of these changes were implemented early in adulthood. Using this approach to search databases of gene expression data, we found conserved transcriptional signatures in larval development, embryogenesis, gametogenesis and mRNA degradation.

Gene expression profiling measures the expression levels of thousands of genes at once[1,2]. Most expression profiling studies have focused on the specific genes that respond to specific conditions, but another important direction in functional genomics is to derive insight from global patterns of gene expression. Genome-scale expression patterns have been used as physiological 'fingerprints' for classifying tumors[3,4] and assigning uncharacterized mutations and drugs to known pathways[5]. Because they use information from many genes at once, patterns have great discriminating power, even when the transcriptional effects on individual genes are small[5,6].

The patterns of changes in gene expression observed in microarray experiments can be extensive and complex. To try to analyze these patterns, we exploited the principle that important biological processes are often conserved between organisms. We present an approach to comparative functional genomics based on shared patterns of regulation across orthologous genes. We also present a method for identifying conserved biological components of those patterns that correspond to Gene Ontology categories. These methods can be used to search databases of microarray experiments to discover connections among biological processes in different organisms.

RESULTS

Comparing genomic expression patterns across species

We used phylogenetic analysis to systematically identify orthologous groups of genes for all pairwise comparisons between *C. elegans*, *D. melanogaster*, *Saccharomyces cerevisiae* and *Homo sapiens* (**Supplementary Tables 1–5** online). For *C. elegans* and *D. melanogaster*, we identified 3,851 most-conserved orthologous gene pairs (**Fig. 1a**).

We used DNA microarrays in each organism to compare gene expression under different conditions (**Fig. 1b**). We then used gene phylogenetic relationships to match systematically the measurements of differential expression between orthologous genes from the two organisms (**Fig. 1c**). We used the correlation of the log-transformed relative change in expression of orthologous genes to assess the extent of shared regulation.

Global similarity of transcriptional profiles of aging

Using this approach, we asked whether gene expression patterns in adult aging were shared by two highly diverged animals: the nematode *C. elegans* and the fruit fly *D. melanogaster*, whose last common ancestor existed about one billion years ago[7]. We used spotted-PCR-product microarrays[1] to compare gene expression in middle-aged adult (6 d adult) and young adult (0 d adult) sterile *C. elegans* hermaphrodites and used Affymetrix oligonucleotide microarrays[2] to compare expression in middle-aged adult (23 d old) and young adult (3 d old) female flies[8]. The cross-species Pearson correlation of the log-transformed relative change in expression of orthologous genes during aging was 0.144, which is significant at the 10^{-11} level. Sixteen comparisons of independent experimental replicates all had high significance values, with a mean

[1]Program in Neuroscience; and Departments of [2]Biochemistry and Biophysics and [3]Physiology; University of California, San Francisco, California 94143, USA. [4]Department of Biology, University College London, London, UK. Programs in [5]Genetics and [6]Developmental Biology; [7]Howard Hughes Medical Institute; [8]Department of Anatomy; and Programs in [9]Biophysics and [10]Biological and Medical Informatics, University of California, San Francisco, California 94143, USA. [11]Present address: Huffington Center on Aging, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. Correspondence should be addressed to H.L. (haoli@genome.ucsf.edu).

Published online 18 January 2004; doi:10.1038/ng1291

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

correlation of 0.155 ± 0.012 ($P < 10^{-35}$). These results indicate that most aging-related changes are species-specific, but the conserved component of these expression profiles could include several hundred *C. elegans–D. melanogaster* ortholog pairs. This result is highly statistically significant; it is not observed in one million randomized pairings of the expression results (**Fig. 2a**). Nonparametric tests confirmed the statistical significance of the shared regulation (Spearman rank correlation = 0.156, $P < 10^{-12}$; Kendall's Tau = 0.106, $P < 10^{-12}$).

We observed similarly correlated regulation during aging in microarray data sets from different tissues, laboratories and experimental platforms. We used Affymetrix microarrays to compare gene expression in heads of young and adult male flies and observed a similar correlation with aging *C. elegans* (0.148, $P < 10^{-11}$). These results suggest that the conserved regulation is present in *D. melanogaster* somatic tissue. A published profile of adult aging in *C. elegans* using Affymetrix microarrays[9] also showed highly significant correlations with profiles of aging from *D. melanogaster* heads ($R = 0.180$, $P < 10^{-6}$) and profiles of aging in whole female fruit flies ($R = 0.150$, $P < 10^{-6}$). Highly significant correlations in the change of transcript abundance with age were observed within two separate subsets of the *C. elegans–D. melanogaster* ortholog pairs: those ortholog pairs that have orthologs in the yeast *S. cerevisiae*, and those otholog pairs that have no homology to any yeast gene (**Fig. 2a**).

Biological features of conserved regulation

The statistical and explanatory power of gene expression analysis is greatly increased by grouping related genes into functional categories. The Gene Ontology annotation system[10] defines hundreds of groups of ortholog pairs with common molecular function, cellular localization or biological role. We searched the data sets for highly conserved subpatterns that corresponded to Gene Ontology categories by identifying those categories that contribute significantly to the observed correlation.

Fourteen Gene Ontology categories showed highly conserved patterns of regulation in aging *D. melanogaster* heads and aging *C. elegans* (**Fig. 2b** and **Supplementary Fig. 1** online), at a strict significance cutoff at which less than one false positive category would be expected by chance. No categories showed significant negative correlation. Similar comparisons using other published aging data sets[8,9] and other time points yielded a broadly overlapping set of Gene Ontology categories (**Supplementary Fig. 2** online), confirming the robustness of the result.

Aging in both *D. melanogaster* heads and *C. elegans* repressed genes in Gene Ontology categories for mitochondrial membrane and mitochondrial inner membrane (**Fig. 2b**), including many components of the mitochondrial respiratory chain, the ATP synthase complex and the citric acid cycle. Earlier studies identified individual oxidative metabolism genes that are repressed by aging in worms, flies or mammals[8,11,12]; our results suggest that these individual results are manifestations of a broad, conserved pattern that includes most oxidative metabolism genes. *C. elegans* and *D. melanogaster* also showed conserved patterns of regulation of genes encoding peptidases, and proteins for catabolism and DNA repair (**Fig. 2b**).

An unexpected shared feature of aging in *C. elegans* and *D. melanogaster* was the repression of orthologous genes involved in diverse ATP-using molecular transport functions, including primary active transporters, ion transporters and ABC transporters (**Fig. 2b**).

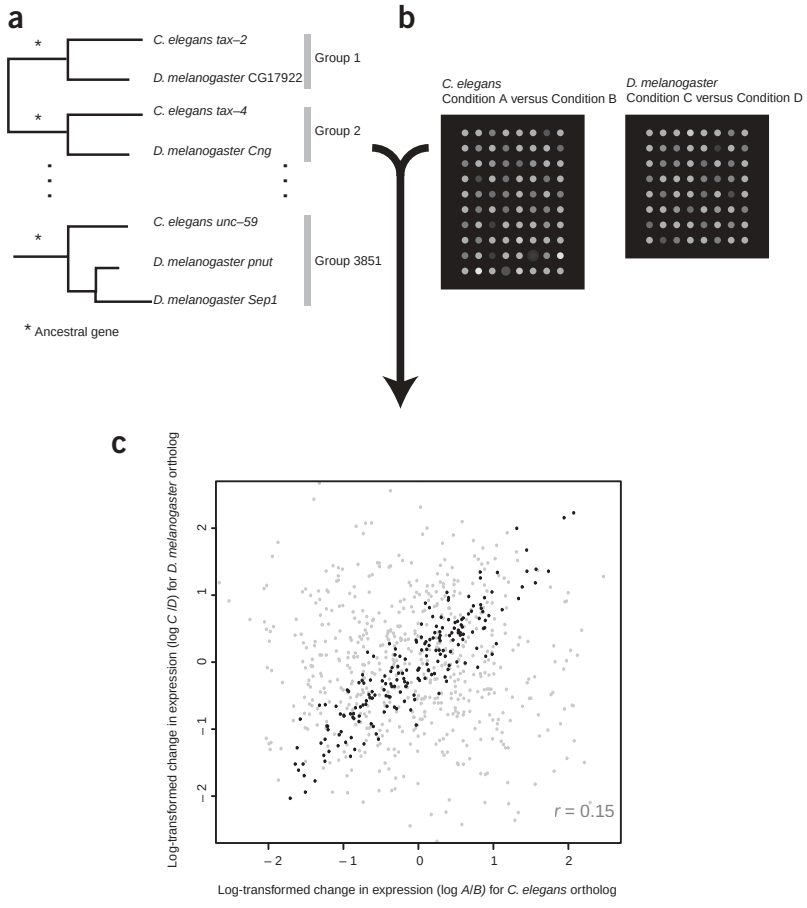


Figure 1 Comparative functional genomic analysis. (**a**) Phylogenetic analysis. Comparative sequence analysis was used to identify a complete set of candidate orthologs, genes related by vertical descent from an ancestral gene (asterisk) present in the last common ancestor of *C. elegans* and *D. melanogaster*. If an orthologous group has multiple genes from either species, the most-conserved orthologous gene pair was identified (for example, from group 3851, which contains the two *D. melanogaster* paralogs *Sep1* and *pnut*, *C. elegans unc-59* and *D. melanogaster pnut* were selected). (**b**) Expression profiling. For each organism, DNA microarrays were used to measure the relative expression of each gene under two conditions. (**c**) Phylogenetic integration of expression data. Measurements of log-transformed relative change in expression were systematically paired between orthologous genes from the two organisms. The correlation of the paired log-transformed relative change measurements was used to assess the similarity of the gene expression patterns in the two organisms. Hypothetical data was used here to illustrate the fact that even if most ortholog pairs (gray circles) lack any conserved regulation and contribute no correlation, a set of ortholog pairs (black circles) with partially conserved regulation can create a significant global correlation. For the data shown, in which conserved regulation contributes to expression of 25% of the ortholog pairs, the global Pearson correlation $r = 0.15$.

Aging seems to involve a decreased transcriptional commitment to active intracellular and intercellular movement of ions, nutrients and transmitters.

Most transcriptional changes were specific to worms or to flies. For example, in these experiments and other work[13], aging in *C. elegans* repressed genes encoding collagens and induced genes encoding histones, transposases and DNA and RNA helicases; these changes did not characterize *D. melanogaster* aging. Aging in *D. melanogaster* induced genes encoding cytochrome p450s, glycosylases and peptidoglycan receptors, but aging in *C. elegans* did not alter the expression of the orthologous genes.

Timing of conserved regulation

Two specific molecular features of aging, the repression of oxidative metabolism genes[12,14] and the correlation between transcriptional profiles of aging and stress[14], are widely assumed to represent responses to oxidative damage with advancing age. By profiling gene expression at time intervals throughout adulthood in *C. elegans* and *D. melanogaster*, we assessed how conserved gene expression programs were implemented over time. Both the conserved global pattern of change in gene expression (**Fig. 3a**,**b**) and the conserved repression of oxidative metabolism genes (**Fig. 3c**) were abruptly implemented early in adulthood. We profiled the transcriptional responses of worms and flies to heat and oxidative stress and found that stress responses were significantly correlated with early-adulthood transcriptional programs in both organisms (**Fig. 3d**). These results suggest that changes in gene expression with adult age are not solely implemented in response to cumulative damage. Instead, the timing of these conserved features of aging suggests developmentally timed transcriptional regulation in young adults.

Searching databases of genomic expression patterns

To increase the power and generality of comparative analysis, we developed methods for searching databases of gene expression profiles from different organisms, much as BLAST allows researchers to find related gene and protein sequences in different species. We assembled, from our own experiments and 300 published *C. elegans* experiments[9,15], a database of *C. elegans* expression profiles addressing larval development, sex differences, aging, environmental stress responses, neuronal signaling, organogenesis, dauer formation and developmental defects (**Supplementary Table 4** online). We then queried this database with the *D. melanogaster* aging data, by ranking the *C. elegans* expression profiles in this database according to their similarity to profiles of *D. melanogaster* aging.

Notably, the *C. elegans* profiles most similar to search profiles of *D. melanogaster* aging were profiles of *C. elegans* aging (**Table 1**).

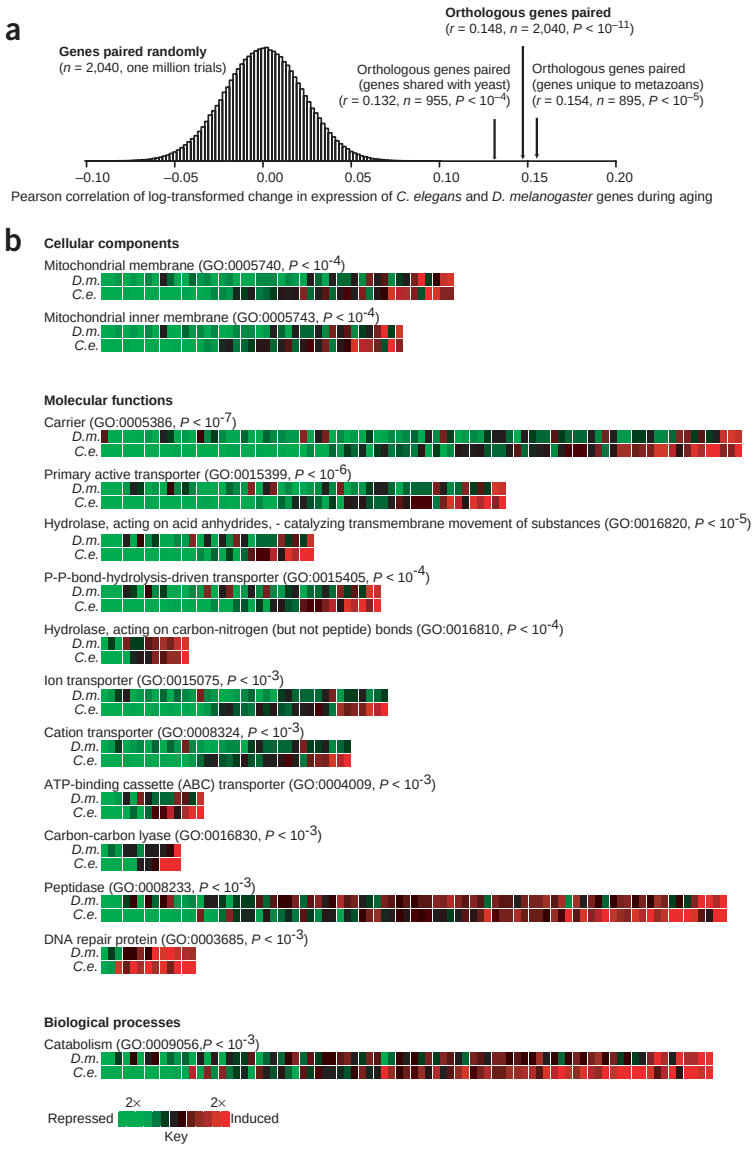


Figure 2 Correlated regulation of orthologous genes by aging in *C. elegans* and *D. melanogaster*. (**a**) Correlated effect of aging on expression of orthologous genes in *C. elegans* and *D. melanogaster*. Microarray measurements of log-transformed relative change in expression with age were paired for orthologous genes from *C. elegans* and *D. melanogaster*; this Pearson correlation for orthologous gene pairs (long arrow) was compared against a distribution of one million Pearson correlations (histogram) each obtained by pairing *C. elegans* and *D. melanogaster* genes randomly. (**b**) Shared transcriptional signature of aging in *D. melanogaster* (*D.m.*) heads and *C. elegans* (*C.e.*). Highly conserved patterns in the gene expression data sets that corresponded to Gene Ontology (GO) categories were identified (14 large blocks). For each Gene Ontology category, the measured change in expression of each gene (small colored rectangle within block) in that category is represented. Each *D. melanogaster* gene is shown above its *C. elegans* ortholog. Red indicates induction by aging; green indicates repression by aging. All ortholog pairs from the indicated Gene Ontology categories are shown; some Gene Ontology categories overlap, with some ortholog pairs belonging to more than one category. Statistical inferences were made at the Gene Ontology category level; most individual genes show small or statistically insignificant relative changes, but the broad pattern of these changes is conserved and highly significant. **Supplementary Figure 1** identifies the individual genes whose expression is represented here.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

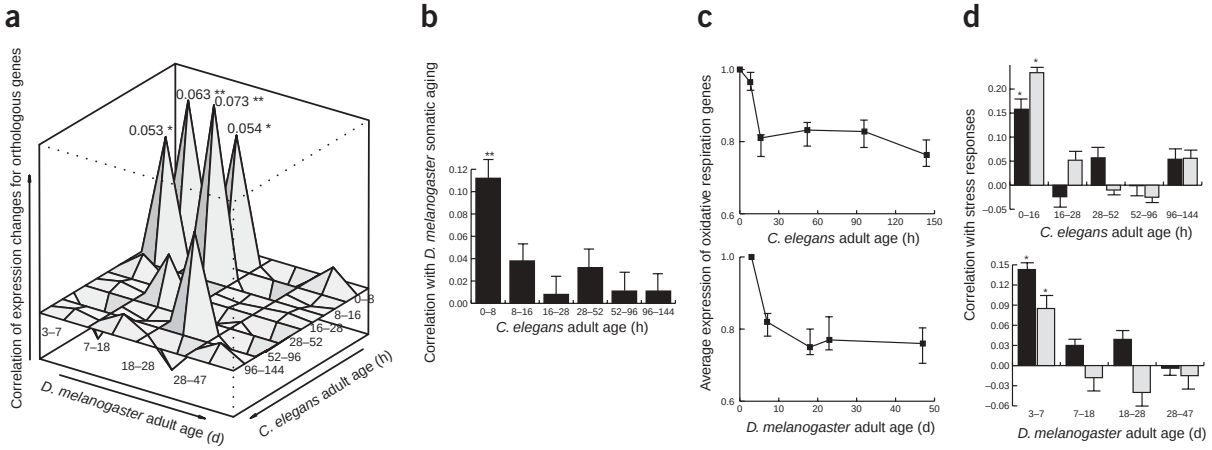


Figure 3 Temporal distribution of conserved gene regulation across adulthood in *C. elegans* and *D. melanogaster*. (**a**) Implementation of conserved gene expression changes over time. Genomic expression changes during six periods of *C. elegans* adulthood were compared with changes during four periods of *D. melanogaster* adulthood, by measuring the Pearson correlation of the log-transformed relative change in expression of orthologous genes. Statistical significance of correlations: $^{**}P < 0.001$, $^{*}P < 0.005$. (**b**) Global correlations in genomic expression changes during aging in transcripts from whole *C. elegans* and *D. melanogaster* heads. (**c**) Transcriptional repression of oxidative metabolism early in adulthood in *C. elegans* and *D. melanogaster*. Orthologous genes in the mitochondrial electron transport chain (GO:0005746) were identified. The expression of each gene, relative to its expression at the beginning of the time course, was obtained from the microarray data sets. The figure shows the median relative change (connected points) and two standard errors around the geometric mean relative change (bars) for this group of genes. (**d**) Implementation of a stress response pattern early in adulthood in *C. elegans* and *D. melanogaster*. Profiles of gene regulation during successive periods of *C. elegans* and *D. melanogaster* adulthood were compared with profiles of *C. elegans* heat stress (light bars) and *D. melanogaster* oxidative stress (dark bars), by measuring the Pearson correlation of the log-transformed relative change in expression of orthologous genes (for interspecies comparisons) or of the same genes (for intraspecies comparisons). Statistical significance of correlations: $^{*}P < 0.001$.

This cross-species similarity persisted across data sets from different *C. elegans* laboratories, *D. melanogaster* laboratories, specific experimental designs and microarray platforms (**Tables 1** and **2**). The next closest *C. elegans* matches to the *D. melanogaster* aging profiles were profiles of heat-stress responses. Aging and heat stress are related in *C. elegans*: many long-lived mutants are thermotolerant, and mild heat stress increases longevity[16,17]. The strongest negative correlation with expression profiles of *D. melanogaster* aging came from profiles of *daf-2(e1368)* mutants[18]. The gene *daf-2* encodes an insulin/IGF-1 receptor homolog; *daf-2* mutants age more slowly and live twice as long as wild-type worms[19,20]. *D. melanogaster* gene expression profiles thus seem to identify both analogous and related gene expression experiments in *C. elegans*.

To extend database searching to other biological questions, we searched the database of *C. elegans* gene expression profiles using published profiles of *D. melanogaster* larval development[21]. Among all *C. elegans* expression profiles, the closest matches to profiles of *D. melanogaster* larval development were profiles of *C. elegans* larval development[22] (**Table 2**). We observed shared patterns of regulation across Gene Ontology categories for protein processing,

Table 1 Closest matches among all *C. elegans* microarray experiments to an expression profile of aging in *D. melanogaster*

C. elegans samples compared	Correlation to profile of aging in *D. melanogaster*	Ortholog pairs	*P*
2 weeks old versus 2.5 d old[9]	0.180	408	8.34×10^{-5}
144 h adult versus 0 h adult	0.148	2040	9.24×10^{-12}
52 h adult versus 0 h adult	0.145	2037	2.45×10^{-11}
40 h adult versus 0 h adult	0.139	1902	5.67×10^{-10}
72 h adult versus 0 h adult	0.134	2029	6.74×10^{-10}
16 h adult versus 0 h adult	0.122	1934	3.74×10^{-8}
96 h adult versus 0 h adult	0.119	1993	5.18×10^{-8}
8 h adult versus 0 h adult	0.115	1947	1.77×10^{-7}
28 h adult versus 0 h adult	0.092	1965	2.24×10^{-5}
52 h adult versus 8 h adult	0.091	2000	2.34×10^{-5}
52 h adult versus 16 h adult	0.089	2055	2.66×10^{-5}
Heat-stressed 2 h versus control	0.089	2129	1.97×10^{-5}
144 h adult versus 8 h adult	0.086	2010	5.66×10^{-5}
72 h adult versus 8 h adult	0.085	1998	7.13×10^{-5}
Heat-stressed 4 h versus control	0.083	2118	6.57×10^{-5}
320 profiles of varied biological phenomena	–0.06 to 0.06		
daf-2(e1368) mutants versus wild-type (N2)[18]	–0.132	2156	3.67×10^{-10}

Closest matches in a database of *C. elegans* gene expression profiles to expression profile of aging in *D. melanogaster* (heads from 47-d-old versus heads from 3-d-old male flies). The database includes gene expression profiles addressing larval development, germline development, adult aging, environmental stress responses, neuronal signaling and cell fate defects. The database was made from about 340 microarray experiments, of which 40 are profiles of aging.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

protein transport, secretion and macromolecule catabolism (**Supplementary Fig. 3** online).

We used published profiles of embryonic development in *D. melanogaster*[21] to search the *C. elegans* gene expression experiments. The best matches were comparisons of gene expression in *C. elegans* embryos with expression in larvae[9,22] and comparisons of embryonic expression in different mutants[23] (**Table 2**). Shared patterns of change included Gene Ontology categories for cell cycle, DNA metabolism, cytoskeleton, microtubule-based processes and proteolysis (**Supplementary Fig. 4** online).

To assess whether database searching could make connections among more diverged organisms, we searched the *C. elegans* database with expression profiles of sporulation in the yeast *S. cerevisiae*[24]. The strongest matches to profiles of yeast sporulation came from profiles of germline formation in *C. elegans*[25] (**Table 2**). The database match seemed to recognize conserved transcriptional programs associated with meiosis; important matching Gene Ontology categories between yeast sporulation and nematode germline development included nucleoplasm, chromosome condensation and DNA strand elongation (**Supplementary Fig. 5** online).

The Stanford Microarray Database[15] contains 647 publicly available *S. cerevisiae* experiments and 2,247 *H. sapiens* experiments. We generated a table of ortholog pairs in yeast and humans to allow searches between these databases (**Supplementary Tables 5** and **6** online). Human mRNA degradation has been profiled in T-cells by blocking transcription with actinomycin D and then using microarrays to measure transcript abundance[26]. The strongest matches to this array experiment among all yeast experiments were profiles comparing *rpb1*, the RNA polymerase II mutant, with wild-type yeast[27] (**Table 2**). As both experiments represent a transcriptional block, the similarity of these profiles suggests that mRNA stability is conserved for orthologous genes in yeast and humans. Gene Ontology categories for kinases and transcription factors were among the most rapidly degraded mRNAs in both humans and yeast, and transcripts encoding ribosomal and core metabolic proteins were extremely stable in both organisms. Searching the human database with the yeast *rpb1* profiles yielded experiments that may correspond to transcriptional blockade: profiles of host responses to diverse pathogenic infections[28–31] and profiles from whole blood, which is dominated by mRNAs from erythrocytes, which lack nuclei and therefore do not carry out transcription[32].

DISCUSSION

We developed a method of identifying analogies among biological processes in diverse organisms by comparative analysis of gene expression patterns. These methods are freely available from our website.

We used this approach to identify a shared pattern of adult-onset gene regulation that is implemented by two highly diverged animals, *C. elegans* and *D. melanogaster*. An unexpected feature of this conserved program was the repression of genes encoding orthologous transporter-ATPases, which offers a candidate mechanistic connection

Table 2 Cross-species searches of DNA microarray databases

Query profile	Gene expression database searched	Matching profiles (best hits)	Correlation (R)	Ortholog pairs	*P*
D. melanogaster adult aging: 18 d versus 3 d (ref. 8)	*C. elegans*	Aging: 16 h adult versus 0 h adult	0.152	2400	4.47×10^{-15}
		Aging: 2 weeks old versus 2.5 d old[9]	0.150	422	1.04×10^{-4}
		Aging: 72 h adult versus 0 h adult	0.139	2505	1.64×10^{-12}
D. melanogaster larval development: 96 h versus 24 h (ref. 21)	*C. elegans*	Larval development: L2 versus L1 (ref. 22)	0.162	1334	1.27×10^{-9}
		Larval development: L3 versus L1 (ref. 22)	0.136	1334	3.25×10^{-7}
		Larval development: L4 versus L1 (ref. 22)	0.130	1334	8.83×10^{-7}
D. melanogaster embryonic development: 24 h versus 11 h (ref. 21)	*C. elegans*	Embryonic/larval development: 12 h versus egg[9]	0.306	624	3.04×10^{-14}
		Embryonic/larval development: L1 versus egg[22]	0.214	1284	2.00×10^{-12}
		Embryonic development: *skn-1* versus *par-1* embryos[23]	0.176	800	5.90×10^{-8}
S. cerevisiae sporulation: t = 2 h versus t = 0 h (ref. 24)	*C. elegans*	Germ line: N2 versus *glp-4*, young adults[25]	0.121	730	5.00×10^{-4}
		Germ line: N2 versus *glp-4*, L3s (ref. 25)	0.098	629	6.00×10^{-3}
		Germ line: N2 versus *glp-4*, L2s (ref. 25)	0.082	713	1.40×10^{-2}
H. sapiens mRNA decay: actinomycin D 45 min versus baseline[26]	*S. cerevisiae*	mRNA decay: *rpb1* versus wild-type, 10 min (ref. 27)	0.322	719	4.13×10^{-19}
		mRNA decay: *rpb1* versus wild-type, 5 min (ref. 27)	0.321	717	5.32×10^{-19}
		mRNA decay: *rpb1* versus wild-type, 15 min (ref. 27)	0.317	719	1.50×10^{-18}

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

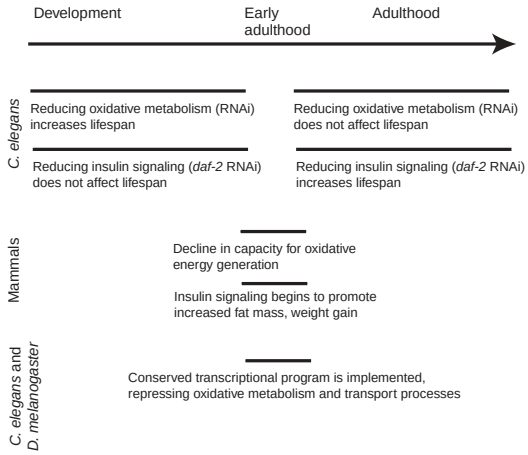


Figure 4 Aging, lifespan and conserved early-adult physiological change in metazoa. The insulin pathway begins to regulate lifespan on the first day of adulthood in *C. elegans*[34]. In mice, insulin signaling in adipose tissue starts to cause weight gain early in adulthood[36]. In *C. elegans*, oxidative respiration seems to limit lifespan until early adulthood, but not afterwards[33]. Mammals begin to lose capacity for oxidative energy generation early in adulthood[35]. *C. elegans* and *D. melanogaster* implement a conserved transcriptional program early in adulthood, one feature of which is the repression of oxidative metabolism genes.

between two known features of aging: reduction in ATP synthesis and decline in the physiological activity of neurons, muscle and excretory processes. An expected feature of this conserved program was the repression of genes with roles in mitochondrial oxidative respiration. Unexpectedly, however, we found that worms and flies both repressed these genes early in adulthood, before the onset of functional decline, and more abruptly than a damage-response model would predict.

In *C. elegans*, mitochondrial respiration before early adulthood limits its subsequent adult lifespan but later mitochondrial respiration does not (**Fig. 4**)[33]. At about the same time that this transition takes place, the insulin pathway begins to regulate lifespan[34]. Mammals also begin to lose oxidative capacity early in adulthood[35], and certain longevity-limiting effects of the insulin pathway on fat accumulation begin early in adulthood[36]. Our results show that the transformation of these relationships early in adulthood is accompanied by a conserved transcriptional program. An exciting direction in aging research will be to identify the signals that induce conserved physiological change early in adulthood.

Although these results suggest the potential of systematic comparative analysis in functional genomics, we expect that future work will improve our methods. For example, the development of methods to systematically assign genes to 'regulons'[37,38] may make possible regulon-based measures of correlation that could be more sensitive and specific in their identification of analogous biological programs. The integrative use of expression data from different species is an emerging area of research[39–43], and elements of these different approaches might be combined to develop additional tools. Our computational approach is also readily generalized to data on protein expression and modification[44].

Comparative functional genomics could be a powerful way to distinguish the essential from the species-specific features of biological processes, such as disease, stress and development. Aided by growing repositories for expression data[45,46] and conventions for reporting genomic experiments[47], measures of correlation in searchable databases could identify new analogies among disease states, mutant strains and drug responses in diverse organisms.

METHODS

Phylogenetic analysis. We obtained sequence data from Gadfly Release 2 of the Berkeley Drosophila Genome Project and Wormpep version 51 from the Sanger Centre. We merged these protein sets and subjected them to all-against-all BLASTP analysis using the BLOSUM62 substitution matrix, DUSTSEQ complexity filtering and a probability cutoff of 10^{-10}. We used the BLAST results to group *C. elegans* and *D. melanogaster* genes into clusters by means of an agglomerative clustering algorithm described elsewhere[48]. Agglomerative clustering yielded 5,042 clusters of 2–161 genes each.

We carried out multiple sequence alignment for each of the 5,042 clusters using CLUSTALW with default parameters. We used the sequence alignment for each cluster to generate a phylodendrogram by the neighbor-joining method, also using CLUSTALW. Points at which the resulting phylodendrograms branched into species-specific clades defined orthologous groups. For *C. elegans* and *D. melanogaster*, 3,851 groups were thus defined. If an orthologous group contained more than one gene for either species (1,290 cases, the result of additional branching after species divergence), we identified the most-conserved orthologous gene pair by comparing pairwise Smith-Waterman alignment scores. In the resulting ortholog table, each orthologous group was thus represented by a single pair of genes. An alternative method of identifying ortholog pairs, by identifying all mutual best hits directly from the BLAST results, yielded an ortholog table 90% identical to that yielded above, without significantly changing any of the subsequent statistical results. We used this approach to build ortholog tables for each pairing of *C. elegans*, *D. melanogaster*, *H. sapiens* and *S. cerevisiae*.

Strains, lifespans and culture conditions. For all *C. elegans* experiments, we used a CF512 fer-15(b26) II; fem-1(hc17) IV mutant strain, whose spermatids do not activate into spermatozoa at 25 °C. Culturing this strain at 25 °C prevented self-fertilization and therefore eliminated contributions from embryonic transcripts. Several *C. elegans* mutants do not develop a germ line, but given the important role of the germ line in regulating aging and life span[49], such strains do not offer a way to profile normal adult aging. The CF512 strain has normal germline stem cells and oocytes, ages at the wild-type rate and has the same lifespan as wild-type N2 worms. Adult age in *C. elegans* is measured from the first day of adulthood, after adult anatomy, adult behaviors and reproductive maturity are established. *C. elegans* has a median adult life span of about 10 d at 25 °C, with fecundity that peaks at 2 d of adulthood and is largely exhausted by 4 d. To yield synchronized *C. elegans* populations for analysis, we axenized eggs and then synchronized worms by L1 arrest, as described elsewhere[50]. For the aging experiments, we collected samples at 0, 8, 16, 28, 40, 52, 70, 96 and 144 h (6 d) after worms reached adulthood.

Experiments with tissue from whole fruit flies have been described elsewhere[8]. These experiments used the Dahomey strain, an outbred stock whose lifespan is similar to that of newly caught, wild populations. In *D. melanogaster*, adult age is measured from eclosion, when fully formed adults emerge from the pupal case. The median Dahomey female adult life span at 25 °C is 28 d; fecundity peaks at 10 d and is nearly exhausted by 21 d.

For experiments with tissue from *D. melanogaster* heads, we cultured the *w1118* strain in standard cornmeal agar medium. The *w1118* strain is an inbred lab stock widely used in genetic and transgenic studies. *w1118* males have a median lifespan of 35 d. We collected adult males within 24 h after eclosion. We maintained ~200 flies in constant darkness in each food bottle at 25 °C and 70% humidity and transferred them to fresh bottles every 3–4 d. We collected transcripts at 3 d and 47 d.

***C. elegans* expression profiling.** For *C. elegans*, we amplified 18,455 predicted *C. elegans* by PCR using oligos obtained from Research Genetics. The sequences of these oligos have been deposited in Wormbase. We printed the PCR products on glass slides using techniques described elsewhere[18]. We used the microarrays to survey gene expression by comparing mRNAs extracted from a sample at each time point to a common mixed reference mRNA pool by competitive hybridization. We extracted RNA with Trizol (GIBCO/BRL) and labeled it according to standard techniques.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

***D. melanogaster* expression profiling.** The experiments with tissue from whole *D. melanogaster* have been described elsewhere[8]. We hybridized at least four replicate Affymetrix roDROMEGa GeneChips for each sample point; the replicates were derived from independent RNA extractions of separate biological samples. The results for replicate GeneChips were consistent, with correlations (of log-transformed relative expression measurements) all exceeding 0.90.

We processed and analyzed samples from *D. melanogaster* heads in a different laboratory from that used for the whole–*D. melanogaster* experiments. To separate the head from the rest of the body, we froze flies and briefly vortexed them in liquid nitrogen. We collected fly heads using a sieve that retained fly bodies. We extracted total RNA using Trizol (GIBCO/BRL). We isolated poly(A)$^+$ RNA using Oligotex resin (Qiagen). We profiled samples with Affymetrix DrosGenome1 GeneChips using standard Affymetrix protocol.

Expression profiles of heat and oxidative stress. To profile the effect of heat stress on gene expression in *C. elegans*, we cultured and synchronized CF512 worms as described above. We then exposed CF512 adults to 30 °C (experimental condition) or maintained them at 25 °C (control condition) for 2, 4, 6, 8, 10 and 12 h. We compared corresponding experimental and control samples by competitive hybridization to DNA microarrays as described above.

To profile the effect of oxidative stress on gene expression in *D. melanogaster*, we cultured w118 male adult flies as described above. At 2 d, we fed the flies sucrose with 15 mM paraquat (experimental condition) or regular sucrose (control condition) for 30 h. We collected heads and extracted transcripts as described above and then profiled them using Affymetrix GeneChips.

Microarray data processing. Except where individual experimental replicates are discussed in the text, we generally used a composite gene expression profile that represented the average of experimental replicates. To construct such composite profiles, we averaged the log-transformed relative change measurements across replicates for each probe. We obtained profiles of differential gene expression (comparing two different samples or conditions from the same organism) in the following ways. In experiments using two-channel microarrays to compare two experimental samples directly, we used those measurements of log-transformed relative change directly. In experiments using multiple two-channel microarrays to compare multiple experimental samples to a common reference sample, we compared the experimental samples by calculating the difference between the log-transformed relative change measurements from different hybridizations, removing the effect of the reference sample. In experiments using multiple single-channel Affymetrix microarrays to profile multiple experimental samples, we compared the experimental samples by calculating the difference between normalized log-transformed relative change measurements from different hybridizations.

Calculation of interspecies correlations. We used the Pearson correlation (r) of the log-transformed relative change measurements for orthologous genes to measure global correlation between heterologous expression profiles ($r = \Sigma_{i=1..n} (x_i - \mu_x)(y_i - \mu_y)/n\sigma_x\sigma_y$, where $X = (x_1,x_2,...x_n)$ and $Y = (y_1,y_2,...y_n)$ are vectors of log-transformed relative change measurements for orthologous genes in *C. elegans* and *D. melanogaster*, respectively, and μ and σ are the mean and standard deviation of these measurements, respectively). We assessed the statistical significance of Pearson correlations using Student's t-test ($t = r\,((n-2)/(1-r^2))^{1/2}$) with $(n-2)$ degrees of freedom, where n is the number of ortholog pairs yielding gene induction measurements in both organisms.

Monte Carlo simulations. We found that 2,040 ortholog pairs yielded expression measurements in both organisms at both the old and young time points, allowing measurements of log-transformed relative induction with age. In each simulation, we randomly paired these 2,040 *C. elegans* and 2,040 *D. melanogaster* genes and then calculated the Pearson correlation of their respective experimental log-transformed relative change measurements. Across one million such simulations, the Pearson correlation was distributed in accordance with Student's t-test distribution. The distribution of simulated correlations had a mean of zero, a standard deviation of 0.022 and the following percentiles: 95th, 0.037; 99th, 0.053. The largest observation was 0.094.

Assessment of potential artifacts. Artifacts in the profiling process can introduce subtle trends which, if common to both profiles, could cause artifactual measured correlations. Although the cross-platform nature of interspecies comparisons makes such shared trends much less probable, an artifact of potential concern involves a potential relationship between measurements of differential hybridization and a gene's overall hybridization strength (a function of transcript abundance and GC content, both of which are correlated for orthologous genes). To assess the potential contribution of such effects, we repeated the Monte Carlo simulation, but rather than pairing genes randomly, we paired genes which were in the same quantile for overall hybridization intensity. The resulting distribution of correlations did not show a positive bias or a significantly greater variance.

Nonparametric statistical tests. For each Pearson correlation presented in this paper, we also calculated the Spearman rank correlation and Kendall's Tau. These three assessments of significance for global correlations were in broad agreement for all of the results discussed here.

Gene Ontology analysis. The Gene Ontology system organizes biological processes, biochemical functions and cellular compartments ('terms') on a directed graph that describes the relationships among these terms[10]. Each term on the Gene Ontology graph defines a subgraph, which consists of the term, its more specific subterms and the genes associated with those terms. For example, the subgraph for the term 'ion channel' includes genes associated with the 'potassium channel' and 'voltage-gated ion channel' terms. For each Gene Ontology term and its associated subgraph, we measured the contribution of associated ortholog pairs to the global Pearson correlation by the partial summation of the Pearson correlation $r_J = \Sigma_{i \in J} (x_i - \mu_x)(y_i - \mu_y) / n\sigma_x\sigma_y$, (where J is the set of ortholog pairs associated with the Gene Ontology category), using the global mean and variance from the entire gene induction profiles. The distribution of r_J is well approximated by a normal distribution with zero mean and standard deviation $n_J^{1/2}/n$, where n_J is the number of ortholog pairs in J. We assessed the significance of r_J using the z test with $z = r_J / (n_J^{1/2}/n)$. We analyzed only those subgraphs with expression data for a useful number (10–100) of gene pairs; there were about 250 such subgraphs for *C. elegans–D. melanogaster* comparisons, depending on the particular experiments compared. **Figure 2b** and **Supplementary Figures 1–5** online directly represent the results of this analysis for different microarray data sets, showing expression data for the ortholog pairs in those Gene Ontology categories that had significant z scores.

Statistical controls for Gene Ontology analysis. To bootstrap the false positive rate for these multiple, nonindependent hypothesis tests, we repeatedly shuffled the expression data and redid the analysis 10,000 times. Applying a test statistic cutoff of 3.0 in a two-sided test, the estimated false positive rate (average number of Gene Ontology categories with $|z| > 3.0$) from the randomized data was 0.73 ± 0.62, consistent with the false positive rate of 0.65 expected for the z test. To assess whether conserved gene regulation was significantly concentrated into Gene Ontology categories, rather than being randomly distributed across the genome, we carried out the following additional control. Starting with the correlated experimental data sets, we randomized the assignment of paired measurements to ortholog pairs and then redid the Gene Ontology analysis. The false-positive rate (average number of Gene Ontology categories with $|z| > 3.0$) was 1.40 ± 0.91. By contrast, the actual data sets had 14 significant Gene Ontology categories, a result that was not obtained in 10,000 simulations. Analogous results were obtained for the data in **Supplementary Figures 1–5** online.

Databases of microarray data. We downloaded all publicly available *C. elegans*, *S. cerevisiae* and *H. sapiens* microarray data from the Stanford Microarray Database[15]. We used the pixel regression correlation (cutoff = 0.6) to filter individual gene measurements and then obtained the log-ratio-of-medians for each probe for each experiment. We used only those profiles for which the identity of the original experiment was provided; this gave us about 300 *C. elegans*, 650 *S. cerevisiae* and 2,250 *H. sapiens* gene expression profiles. For the *C. elegans* database, we added our own aging and heat stress experiments (another 40 profiles) and carefully subjected all profiles to cross-replicate averaging and cross-reference differencing, as described above under microarray data processing. To search a database using a gene expression profile from one organism as a query, we ranked all the profiles in the database by their similarity to the query profile, using the similarity metric described above. In **Table 2**, we present the three closest matches from each database search.

Accession numbers. Microarray data sets have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus[45] with the following accession numbers: GSE832, GSM12883, GSM12884, GSM12885, GSM12886, GSM12887, GSM12888, GSM12889; GSE826, GSE827, GSM12770, GSM12772 and GSM12773. Data for the heat stress experiments have the following accession numbers: GSE946, GSM15008, GSM15009, GSM 15010, GSM15011, GSM15012, GSM15013 and GSM15014.

URL. Our website (http://worms.ucsf.edu/compare) allows users to analyze their own microarray data sets using the tools in this paper, dynamically explore the paper's database search results, identify significant Gene Ontology categories associated with each search result and browse the associated genes and measurements.

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

ACKNOWLEDGMENTS

We thank A. Malmberg, C. Patil, J. DeRisi and I. Herskowitz for discussions and comments on the manuscript; A. Dillin and J. Lehrer-Graiwer for assistance in building the *C. elegans* microarrays; S. Meadows for assistance with *D. melanogaster* expression profiling; and J. DeRisi and H. Bennett for advice, instruction and use of their equipment. This work was supported by a grant from the US National Institute on Deafness and Other Communication Disorders to C.I.B., a grant from the Ellison Foundation to C.K., and a Sandler Grant, Packard Fellowship and Life Sciences Informatics grant to H.L. S.A.M. was a Howard Hughes Medical Institute graduate research fellow; C.T.M. is a Life Sciences Research postdoctoral fellow; C.I.B. and Y.N.J. are investigators of the Howard Hughes Medical Institute.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 8 July; accepted 15 December 2003
Published online at http://www.nature.com/naturegenetics/

1. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
2. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**, 20–24 (1999).
3. Chung, C.H., Bernard, P.S. & Perou, C.M. Molecular portraits and the family tree of cancer. *Nat. Genet.* **32** Suppl, 533–540 (2002).
4. Ramaswamy, S., Ross, K.N., Lander, E.S. & Golub, T.R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).
5. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
6. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
7. Wang, D.Y., Kumar, S. & Hedges, S.B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266**, 163–171 (1999).
8. Pletcher, S.D. *et al.* Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster. *Curr. Biol.* **12**, 712–723 (2002).
9. Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G. & Brown, E.L. Genomic analysis of gene expression in C. elegans. *Science* **290**, 809–812 (2000).
10. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
11. Lee, C.K., Klopp, R.G., Weindruch, R. & Prolla, T.A. Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**, 1390–1393 (1999).
12. Kayo, T., Allison, D.B., Weindruch, R. & Prolla, T.A. Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proc. Natl. Acad. Sci. USA* **98**, 5093–5098 (2001).
13. Lund, J. *et al.* Transcriptional profile of aging in C. elegans. *Curr. Biol.* **12**, 1566–1573 (2002).
14. Zou, S., Meadows, S., Sharp, L., Jan, L.Y. & Jan, Y.N. Genome-wide study of aging and oxidative stress response in Drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* **97**, 13726–13731 (2000).
15. Sherlock, G. *et al.* The Stanford Microarray Database. *Nucleic Acids Res.* **29**, 152–155 (2001).
16. Finkel, T. & Holbrook, N.J. Oxidants, oxidative stress and the biology of ageing. *Nature* **408**, 239–247 (2000).
17. Lithgow, G.J., White, T.M., Melov, S. & Johnson, T.E. Thermotolerance and extended life-span conferred by single-gene mutations and induced by thermal stress. *Proc. Natl. Acad. Sci. USA* **92**, 7540–7544 (1995).
18. Murphy, C.T. *et al.* Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans. *Nature* **424**, 277–283 (2003).
19. Kimura, K.D., Tissenbaum, H.A., Liu, Y. & Ruvkun, G. daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans. *Science* **277**, 942–946 (1997).
20. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A C. elegans mutant that lives twice as long as wild type. *Nature* **366**, 461–464 (1993).
21. Arbeitman, M.N. *et al.* Gene expression during the life cycle of Drosophila melanogaster. *Science* **297**, 2270–2275 (2002).
22. Jiang, M. *et al.* Genome-wide analysis of developmental and sex-regulated gene expression profiles in Caenorhabditis elegans. *Proc. Natl. Acad. Sci. USA* **98**, 218–223 (2001).
23. Gaudet, J. & Mango, S.E. Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. *Science* **295**, 821–825 (2002).
24. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
25. Reinke, V. *et al.* A global profile of germline gene expression in C. elegans. *Mol. Cell* **6**, 605–616 (2000).
26. Raghavan, A. *et al.* Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.* **30**, 5529–5538 (2002).
27. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. USA* **99**, 5860–5865 (2002).
28. Boldrick, J.C. *et al.* Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl. Acad. Sci. USA* **99**, 972–977 (2002).
29. Detweiler, C.S., Cunanan, D.B. & Falkow, S. Host microarray analysis reveals a role for the Salmonella response regulator phoP in human macrophage cell death. *Proc. Natl. Acad. Sci. USA* **98**, 5850–5855 (2001).
30. Guillemin, K., Salama, N.R., Tompkins, L.S. & Falkow, S. Cag pathogenicity island-specific responses of gastric epithelial cells to Helicobacter pylori infection. *Proc. Natl. Acad. Sci. USA* **99**, 15136–15141 (2002).
31. Cuadras, M.A., Feigelstock, D.A., An, S. & Greenberg, H.B. Gene expression pattern in Caco-2 cells following rotavirus infection. *J. Virol.* **76**, 4467–4482 (2002).
32. Whitney, A.R. *et al.* Individuality and variation in gene expression patterns in human blood. *Proc. Natl. Acad. Sci. USA* **100**, 1896–1901 (2003).
33. Dillin, A. *et al.* Rates of behavior and aging specified by mitochondrial function during development. *Science* **298**, 2398–2401 (2002).
34. Dillin, A., Crawford, D.K. & Kenyon, C. Timing requirements for insulin/IGF-1 signaling in C. elegans. *Science* **298**, 830–834 (2002).
35. Somani, S.M. *et al.* Influence of age on caloric expenditure during exercise. *Int. J. Clin. Pharmacol. Ther. Toxicol.* **30**, 1–6 (1992).
36. Bluher, M., Kahn, B. & Kahn, C. Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science* **299**, 572–574 (2003).
37. Wang, W., Cherry, J.M., Botstein, D. & Li, H. A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae. *Proc. Natl. Acad. Sci. USA* **99**, 16893–16898 (2002).
38. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
39. Whitfield, M.L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell.* **13**, 1977–2000 (2002).
40. Teichmann, S.A. & Babu, M.M. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* **20**, 407–410 (2002).
41. Alter, O., Brown, P.O. & Botstein, D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci. USA* **100**, 3351–3356 (2003).
42. van Noort, V., Snel, B. & Huynen, M.A. Predicting gene function by conserved co-expression. *Trends Genet.* **19**, 238–242 (2003).
43. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
44. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
45. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
46. Brazma, A. *et al.* ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
47. Stoeckert, C.J., Jr., Causton, H.C. & Ball, C.A. Microarray databases: standards and ontologies. *Nat. Genet.* **32** Suppl, 469–473 (2002).
48. Rubin, G.M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
49. Hsin, H. & Kenyon, C. Signals from the reproductive system regulate the lifespan of C. elegans. *Nature* **399**, 362–366 (1999).
50. Lewis, J.A. & Fleming, J.T. Basic culture methods. in *Methods in Cell Biology, Volume 48: Caenorhabditis elegans: Modern Biological Analysis of an Organism* (eds. Epstein, H.F. & Shakes, D.C.) 4–30 (Academic Press, San Diego, California, 1995).

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics