References and Notes

1. J. A. Goodrich, T. Hoey, C. J. Thut, A. Admon, R. Tjian, *Cell* **75**, 519 (1993).
2. C. P. Verrijzer, K. Yokomori, J. L. Chen, R. Tjian, *Science* **264**, 933 (1994).
3. R. Dikstein, S. Zhou, R. Tjian, *Cell* **87**, 137 (1996).
4. N. Tanese, D. Saluja, M. F. Vassallo, J. L. Chen, A. Admon, *Proc. Natl. Acad. Sci. U.S.A* **93**, 13611 (1996).
5. G. Mengus, M. May, L. Carré, P. Chambon, I. Davidson, *Genes Dev.* **11**, 1381 (1997).
6. T. Hoey *et al.*, *Cell* **72**, 247 (1993).
7. S. L. Sanders, P. A. Weil, *J. Biol. Chem.* **275**, 13895 (2000).
8. A. Yamit-Hezi, R. Dikstein, *EMBO J.* **17**, 5161 (1998).
9. O. Wolstein, A. Silkov, M. Revach, R. Dikstein, *J. Biol. Chem.* **275**, 16459 (2000).
10. Total protein extracts were prepared from mouse ovaries of 8-week-old mice that were hormonally synchronized for 48 hours by injection of 7 IU of PMS-G. Ovaries were harvested, washed in phosphate-buffered saline, dounced in extraction buffer [200 mM KCl, 100 mM tris (pH 7.9), 0.2 mM EDTA, 10% glycerol, and 0.1% NP-40], and incubated for 1 hour on ice. Either 20 μl of anti-Myc negative control or 20 μl of anti-TBP/protein A–Sepharose CL-4B (Amersham Pharmacia) beads were incubated with 0.3 ml of ovary extract for 4 hours at 4°C. Immune complexes were washed twice in 0.2 M KCL, twice in 0.5 M KCl, and then twice in 0.2 M KCl on ice. Beads were boiled in sample buffer and separated on either a 7.5 or 10% SDS–polyacrylamide gel electrophoresis (SDS-PAGE) gel. Western blot analysis was carried out with rabbit polyclonal antisera diluted to 1:5000 for anti-$TAF_{II}105$ or to 1:2500 for anti-$TAF_{II}250$ and TBP. The secondary antibody to immunoglobulin G–horseradish peroxidase (Pierce, Rockford, IL) was diluted to 1:5000, and proteins were visualized with enhanced chemiluminesence (Amersham Pharmacia). For the protein samples in Fig. 2D, splenocytes derived from mice of each genotype were activated with lipopolysaccharide (LPS), lysed in sample buffer, and separated by a 6% SDS-PAGE gel.
11. Overlapping phage clones containing a portion of the $TAF_{II}105$ gene were isolated from a mouse 129Sv genomic library and subcloned to generate pBT3. Not I–linearized pBT3 was electroporated into R1 ES cells. The targeting event removed half of exon h located in the 3′ end of the $TAF_{II}105$ gene, replacing it with a *neo* cassette in the reverse orientation. Genomic DNA was isolated from neomycin-resistant ES clones, digested with Bam HI, and analyzed by Southern blot analysis against a ^{32}P-labeled genomic fragment. Targeting efficiency was about 1.6%. Heterozygous ES cells were injected into blastocysts to establish founder chimeric mice, and chimeric mice that transmitted to the germ line were bred to homozygosity. To genotype mice, tail genomic DNA was analyzed in a single PCR reaction containing three primers (105G2, AACATGTAATGGATTTTCT; 105G4, GGTCTGTATTTCCTTAATGG; and Neo2, CTAATTCATCAGAAGCTGAC), which amplifies a 200–base pair (bp) fragment and a 150-bp fragment from the wild-type and mutant $TAF_{II}105$ alleles, respectively. Western blot analysis was used to confirm the loss of full-length $TAF_{II}105$ and the absence of COOH-terminally truncated forms of $TAF_{II}105$. Heterozygous $TAF_{II}105$ mice were back-crossed to the inbred C57BL/6J strain (Jackson Laboratories, West Grove, PA).
12. R. N. Freiman, R. Tjian, unpublished data.
13. R. L. Brinster, H. Y. Chen, M. E. Trumbauer, M. K. Yagle, R. D. Palmiter, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4438 (1985). $TAF_{II}105^{-/-}$ females were bred with vasectomized males, and the presence of a copulatory plug indicated a successful mating. On day 0.5 postcoitum, WT B6SJL F_2 hybrid two-cell eggs were transferred into one oviduct of each –/– pseudopregnant female as previously described. Superovulation was induced by a standard hormone regimen, and ovulated eggs were collected 13 hours after injection of human chorionic gonadotropin as previously described.
14. Supplementary Web material is available on *Science* Online at www.sciencemag.org/cgi/content/full/293/5537/2084/DC1.
15. J. A. Elvin, M. M. Matzuk, *Rev. Reprod.* **3**, 183 (1998).
16. P. G. Knight, *Front. Neuroendocrinol.* **17**, 476 (1996).
17. P. Sicinski *et al.*, *Nature* **384**, 470 (1996).
18. C. A. Gross *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **63**, 141 (1998).
19. M. A. Hiller, T.-Y. Lin, C. Wood, M. T. Fuller, *Genes Dev.* **15**, 1021 (2001).
20. L. Kaltenbach, M. A. Horner, J. H. Rothman, S. E. Mango, *Mol. Cell* **6**, 705 (2000).
21. J. C. Dantonel, S. Quintin, L. Lakatos, M. Labouesse, L. Tora, *Mol. Cell* **6**, 715 (2000).
22. G. J. Veenstra, D. L. Weeks, A. P. Wolffe, *Science* **290**, 2312 (2000).
23. I. Martianov *et al.*, *Mol. Cell* **7**, 509 (2001).
24. D. Zhang, T. L. Penttila, P. L. Morris, M. Teichmann, R. G. Roeder, *Science* **292**, 1153 (2001).
25. Antisense riboprobes labeled with ^{32}P–uridine 5′-triphosphate were synthesized with Maxiscript (Ambion, Austin, TX). Plasmid pAA12 was linearized with Acc65I and transcribed with T7 RNA polymerase to synthesize an antisense $TAF_{II}105$ probe; plasmid pZ33 was linearized with Pst I and transcribed with T3 RNA polymerase to synthesize an antisense $TAF_{II}130$ probe. Hybridizations were carried out with RPA III (Ambion). To detect $TAF_{II}105$ and $TAF_{II}130$ levels, we used 30 μg of total RNA (Ambion), and to detect 18*S* rRNA levels, we used 1 μg of total RNA. For assays in Fig. 4, 3 μg of total RNA derived from +/– and –/– ovaries was used.
26. Ovaries from 6-week-old females were removed from hormonally primed mice, dissected, weighed, and fixed in 10% neutral buffered formalin. Specimens were embedded in paraffin, sectioned at 5 μm, stained with hematoxylin and eosin, and examined and photographed at the identical magnification with a Leica DMR microscope equipped with a MetaMorph Imaging System. For in situ hybridizations, ovaries were dissected from HET and KO females and fixed in a 6:3:1 cocktail of ethanol, formaldehyde (37%), and acetic acid overnight at 4°C. Ovaries were dehydrated in an ethanol series and Histoclear (National Diagnostics, Atlanta, GA) and then embedded in paraffin. Sections of 14 μm were prepared and probed with digoxygenin (DIG)–labeled antisense RNA probes (Roche). An alkaline phosphatase-conjugated antibody to DIG (Roche) was used at a 1:2000 dilution to detect specific RNAs with the substrates NBT and BCIP (Gibco-BRL), and RNAs were visualized by the formation of a blue/purple precipitate.

27. We thank M. Haggart for technical assistance; L. Chu for contributions to this project; D. Bhattacharya and L. Liang for advice with the immunology assays; O. Kelly, B. Martin, and M. Dionne for instruction with in situ hybridizations; N. Hernandez for providing the anti-TBP; A. Ladurner for providing control and $TAF_{II}250$ antibodies; R. Henry for advice with immunoprecipitations; D. Schichnes and S. Ruzin of the CNR biological imaging facility; A. Winoto for RI ES cells; K. Thomas for a 129Sv genomic library; W. Skarnes for mouse blastocyst injections; R. Harland, A. Hochheimer, M. Holmes, C. Inouye, Y. Isogai, B. Lemon, and M. Levine for comments; and J. Lim for preparation of the manuscript. R.N.F. is a fellow of the Leukemia and Lymphoma Society. Supported in part by a grant from NIH to R.T.

24 April 2001; accepted 16 July 2001

A Gene Expression Map for *Caenorhabditis elegans*

Stuart K. Kim,[1]* Jim Lund,[1] Moni Kiraly,[1] Kyle Duke,[1] Min Jiang,[1] Joshua M. Stuart,[2] Andreas Eizinger,[1] Brian N. Wylie,[3] George S. Davidson[3]

We have assembled data from *Caenorhabditis elegans* DNA microarray experiments involving many growth conditions, developmental stages, and varieties of mutants. Co-regulated genes were grouped together and visualized in a three-dimensional expression map that displays correlations of gene expression profiles as distances in two dimensions and gene density in the third dimension. The gene expression map can be used as a gene discovery tool to identify genes that are co-regulated with known sets of genes (such as heat shock, growth control genes, germ line genes, and so forth) or to uncover previously unknown genetic functions (such as genomic instability in males and sperm caused by specific transposons).

The completion of the *C. elegans* genome sequence has identified nearly all of the genes in the genome (19,282 genes) (*1*), but the function for most of these genes remains mysterious. A scant 6% of them have been studied with the use of classical genetic or biochemical approaches (1135 genes), and only about 53% show homology to genes in other organisms (10,303 genes) (*2*). The current challenge is to develop high-throughput functional genomics procedures to study many genes in parallel in order to elucidate gene function on a global scale (*3*–*8*). In one approach, a compendium of gene expression profiles was assembled from a large number of yeast DNA microarray experiments (*9*), which made it possible to ascribe potential functions to previously unknown genes by comparing their expression results to those of genes with known functions. Here, we have established a compendium of gene expression profiles for an animal, *C. elegans*.

We combined data from many DNA mi-

[1]Department of Developmental Biology and Genetics, Stanford University Medical School, Stanford, CA 94305, USA. [2]Stanford Medical Informatics, 251 Campus Drive, MSOB X-215, Stanford, CA 94305, USA. [3]Computation, Computers and Mathematics Center, Sandia National Laboratories, Albuquerque, NM 87185–0318, USA.

*To whom correspondence should be addressed. E-mail: kim@cmgm.stanford.edu

croarray experiments in order to identify sets of co-regulated genes. In each experiment, RNA from one sample was used to generate Cy3-labeled cDNA, and RNA from another sample was used to prepare Cy5-labeled cDNA. The two cDNA probes were simultaneously hybridized to a single DNA microarray and the ratio of the Cy3 to Cy5 hybridization intensities was measured. We have combined data from 553 experiments performed in collaboration with 30 different laboratories (*10*), including 179 experiments with microarrays containing 11,917 genes (63% of the genome) and 374 experiments using microarrays that have 17,817 genes (94% of the genome). The experiments compare RNA between mutant and wild-type strains or between worms grown under different conditions. Figure 1A shows the types of experiments that have been done to date, including experiments on wild-type development, heat shock, Ras signaling, aging, the dauer stage, sex regulation, and germ line gene expression (*6*, *7*, *10*).

To find out which genes are co-expressed, we first assembled a gene expression matrix in which each row represents a different gene (17,817 genes) and each column corresponds to a different microarray experiment (553 experiments) (Fig. 1B). The matrix contains the relative expression level for each gene in each experiment (expressed as $\log_2$ of the normalized Cy3/Cy5 ratios). We calculated the Pearson correlation coefficient between every pair of genes. For each gene, the similarity between it and the 20 genes with the strongest (positive) correlations were used to assign that gene to an *x*-*y* coordinate in a two-dimensional scatter plot with the use of force-directed placement. In this *x*-*y* ordination step, genes are positioned relative to each other under the influence of attractive and repulsive forces. Each gene is attracted to other genes with a force proportional to their similarity in gene expression, but a constant force also repels each gene from groups of other genes. We then used a computer program called VxInsight to visualize the spatial distribution of the genes, resulting in a display in which genes with a high correlation are placed near to each other on a two-dimensional scatter plot. [Force-directed placement and data mining with VxInsight are described in (*11*, *12*), available Online at www.cs.sandia.gov/projects/VxInsight.html, and Link 1 at *Science* Online (*13*)]. As a further visual cue, the two-dimensional scatter plot is converted into a three-dimensional terrain map in which the *z* axis denotes the density of genes within an area (Fig. 2A).

The gene expression map shows gene expression clusters for nearly all of the genes (17,661 genes, 93% of the genome) formed

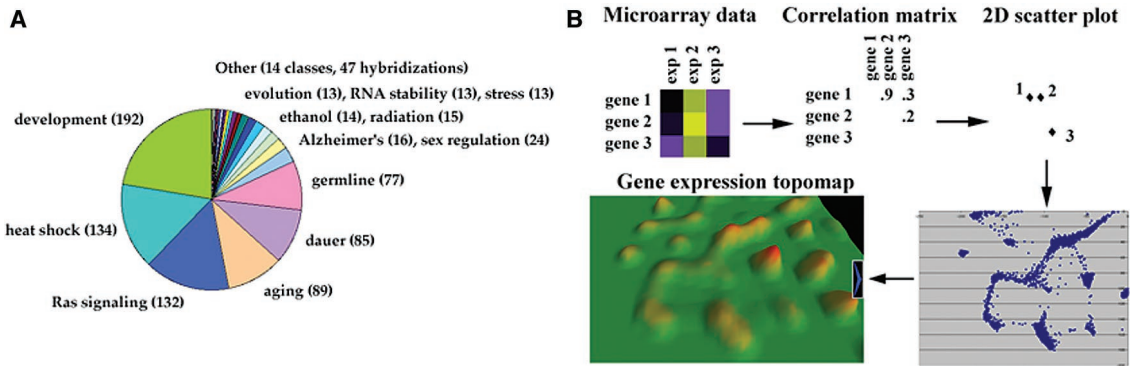

Fig. 1. (**A**) Pie chart shows types of experiments used to generate the gene expression terrain map (*10*). Numbers in parentheses refer to the number of microarray hybridizations done for that experiment class, out of a total of 553 different microarray hybridizations. Some microarray hybridizations fall into multiple classes. (**B**) Construction of the gene expression terrain map by VxInsight. Expression data involving 17,661 genes and 553 experiments are shown. In the expression matrix, yellow denotes increased relative gene expression and blue denotes decreased gene expression. Only three genes and three experiments are shown for simplicity. The expression data are used to calculate Pearson correlations between every pair-wise combination of genes. The most correlated genes in the correlation matrix are used to construct a two-dimensional scatter plot. The scatter plot is converted to a gene expression terrain map showing the gene correlations in three dimensions, where the altitude of a mountain corresponds to density of the genes, denoted by red, yellow, and green.

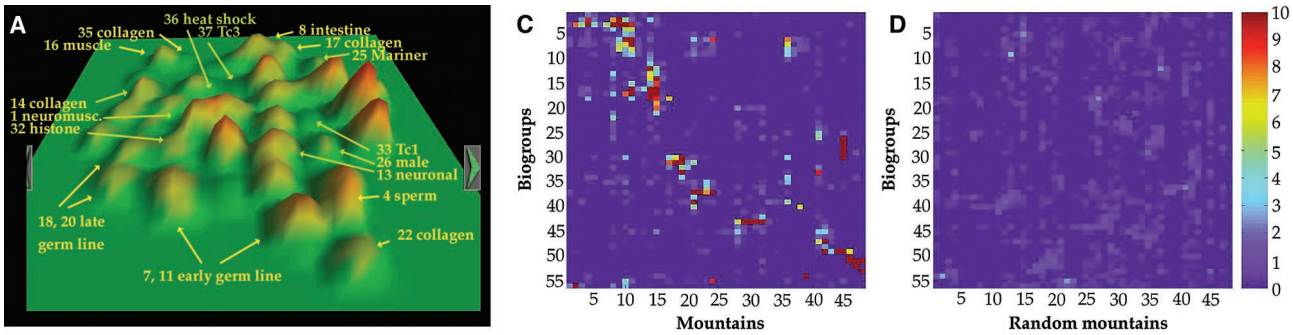

Fig. 2. (**A**) *Caenorhabditis elegans* gene expression terrain map created by VxInsight at lowest resolution, showing three-dimensional representation of 44 gene mountains derived from 553 microarray hybridizations and consisting of 17,661 genes (representing 98.6% of the genes present on the DNA microarrays) (*31*). Selected gene classes that are enriched in specific mountains are shown. (**B**) Terrain map derived from randomized data. (**C** and **D**) We created 56 lists of genes with similar biological function (biogroup), such as genes involved in meiosis, mitosis, translation, DNA synthesis, etc. We then counted the number of genes that overlap in the biogroup with that of the gene expression mountain. We calculated the probability of seeing the observed number of overlaps or more by chance (*P* value) for each biogroup-mountain pair assuming a hypergeometric distribution. Overlap *P* values for each biogroup with each mountain (C) and with randomly constructed mountains of the same size as the original mountain (D) are shown. Scale shows the $\log_{10}$ (*P* value). The list of biogroups and the mountains are shown in Web table 2 and Web table 3 (*13*), respectively. The biogroups and mountains are ordered so that neighbors have similar mountain profiles.

by numerous, diverse microarray experiments (Fig. 2A) (*14*). The raw *C. elegans* expression data can be downloaded from (*13*), and copies of VxInsight can be downloaded from http://cmgm.stanford.edu/~kimlab/topomap/vxinsight.htm. Genes were assigned to individual gene expression clusters (terrain map mountains), and each cluster was numbered according to size, from mount 0 (2703 genes) to mount 43 (5 genes) (Table 1). Each mountain contains sets of highly correlated genes, and the mountain width denotes the overall level of correlation of the genes in that mountain. Mountain altitude signifies the number of genes present in that mountain. It is not yet clear how well gene expression correlations between genes in different mountains can guide the relative placement of one mountain to other mountains on the map.

To assess the significance of the topographical patterns shown in Fig. 2A, we first randomized the expression table by shuffling the values within each row and then reclustered the genes. We observed no appreciable structure in the randomized terrain map (Fig. 2B), suggesting that the geography observed in the actual expression map (Fig. 2A) has biological significance. Then, to assess the stability of the gene expression terrain map, we either rederived the map from random starting positions or added a small amount of noise to the data and noted that there was a high degree of overlap between the various derived maps [Web Links 2 and 3 (*13*)]. To determine which correlations are dependent on specific sets of experiments, we split the experiments into two nonoverlapping sets, formed two new expression maps, and compared gene correlations on one map with those on the other. We observed that many genes have similar neighbors in both maps [Web Link 4 (*13*)]. Lastly, we showed that the observed overlaps between clusters on the gene expression terrain map and groups of genes with similar biological functions are much higher than would be expected by random chance (Fig. 2, C and D) (*13*, *15*). This demonstrates that there are strong biological patterns embedded in the expression data and that the clustering produced by VxInsight has biological relevance. A wide variety of other algorithms [such as hierarchical clustering (*16*)] could have been used in addition to VxInsight to cluster genes on the basis of their expression profiles. We chose to use VxInsight because depicting gene correlation data in three dimensions is extremely useful to visualize patterns of gene expression in large data sets.

We studied the genes in each mountain to find patterns suggesting the underlying biological property for that group of genes. We also looked through 56 sets of genes that were previously known to function together (Web table 1) and found that 46 showed enrichment in one or more of the gene expression mountains (Fig. 2C). Some of the gene expression mountains grouped genes together that were expressed in similar tissues (such as muscle, neuron, germ line), whereas other mountains grouped genes that had similar cellular functions (for example, histones, ribosomal genes, collagens). Overall, we were able to infer a potential physiological importance for 30 of the 44 mountains by showing that specific mountains were enriched for particular sets of genes. The functional interactions suggested by the gene expression terrain map are based entirely on expression data. Thus, in addition to biochemistry and genetics, one could now infer gene func-

Table 1. The R value is a measure of the correlation of the expression patterns of the genes in a mountain. For each mountain, the Pearson correlation between each gene and every other gene in that mountain was calculated. R is the median of all of these Pearson correlations. Large mountains tend to have lower R because genes on opposite sides of the mountain have lower correlations. Unless otherwise noted, representation factors are significant at $P < 0.001$ (*17*). The probability was determined using either the exact hypergeometric probability or using the normal distribution approximation, when appropriate.

Mount	No. of genes	R	Functional groups (representation factor)
0	2703	0.11	
1	1818	0.15	Muscle (4.0×); neuronal (2.7×); PDZ genes (2.9×)
2	1465	0.15	Germ line–enriched (3.8×); oocyte (4.6×)
3	1363	0.13	Reverse transcriptase (3.0×)
4	1195	0.41	Sperm-enriched genes (21×); protein kinases (6.8×); protein phosphatases (15×); major sperm proteins (13×)
5	978	0.22	
6	909	0.21	Neuronal genes (6.5×)
7	810	0.43	Germ line–enriched (12×); oocyte (9.0×); meiosis (11×); mitosis (4.4×)
8	803	0.21	Intestine (13×); *Entemeba histolytica N*-acetylmuraminidase (12×); protease (6.4×); carboxylesterase (7.3×); lipases (10×); antibacterial proteins (17×); UGT (2.8×)
9	786	0.16	
10	635	0.19	
11	587	0.38	Germ line–enriched (13×); oocyte (13×); meiosis (8×); mitosis (10×); histone H1 (18×); retinoblastoma complex (26×)
12	462	0.29	
13	396	0.10	Neuronal genes (3.1×; $P < 0.006$); reverse transcriptase (4.0×)
14	353	0.38	Collagen (2.6×; $P < 0.005$)
15	247	0.37	
16	230	0.40	Muscle (24); collagen (29×)
17	210	0.37	Collagen (9.6×)
18	190	0.38	Germ line (2.4×); oocyte (4.1×); biosynthesis (2.6×); protein synthesis (9.7×)
19	189	0.29	Amino acid metabolism (5.5×); lipid metabolism (5.0×); cytochrome P450 (12×)
20	160	0.46	Germ line–enriched (7.5×); biosynthesis (10×); protein expression (16×); heat shock (10×)
21	154	0.30	Lipid metabolism (10×)
22	151	0.58	Collagen (8×)
23	143	0.53	Protein expression (19×); energy generation (8.6×)
24	133	0.37	Amino acid metabolism (3.9×); lipid metabolism (8.5×); fatty acid oxidation (22×)
25	102	0.44	Mariner transposases (173×)
26	95	0.43	Male-enriched genes (9.5×)
27	87	0.48	Amino acid metabolism (8×); energy generation (8.8×)
28	61	0.28	
29	40	0.53	
30	36	0.41	Protein expression (7.7×)
31	25	0.36	
32	24	0.47	Nucleosomal histones (226×)
33	27	0.43	Tc1 transposon (538×)
34	17	0.44	
35	15	0.59	Collagen (60×)
36	10	0.71	Heat shock (337×)
37	11	0.77	Tc3 transposon (1600×)
38	8	0.44	
39	8	0.42	
40	8	0.43	Protein expression (23×)
41	7	0.45	Protein expression (26×)
42	6	0.33	
43	5	0.69	

tions with the use of gene expression data.

Several mountains were highly enriched for genes from particular tissues or organs. For example, previous microarray experiments identified a total of 650 sperm-enriched genes (*6*). Of these, 583 genes (89%) are present in mount 4, which is 21 times (21×) more than the number of genes expected due to random chance [defined as the representation factor (*17*)] (Fig. 3A and Web table 1).

The sperm-enriched genes were defined using microarrays containing only 63% of the genome, and 848 of the genes in mount 4 were present on these microarrays (and, thus, were available to be identified as sperm-enriched). Thus, highly sperm-enriched genes (99.9% confidence level) composed about 69% of mount 4. Much of the remainder of mount 4 consisted of genes that are sperm-enriched but at a lower level; 775 genes in mount 4 were sperm-enriched at the 95% confidence level (88% of mount 4 out of 848 genes).

The major sperm protein (MSP) genes, which are genes encoding proteins that bind each other in forming the sperm cytoskeleton and are required for sperm motility (Fig. 3, A and B) [see movie (*13*)] (*18*), clustered together at one end of mount 4. As noted previously, protein kinases and phosphatases are enriched in sperm (*6*). These gene classes were also highly enriched in mount 4; specifically, 103 of 361 protein kinase genes (6.8× higher than random chance) and 67 of 106 protein phosphatases (15×) are present in mount 4 (Web table 1). Because sperm are unusual cells in that they are transcriptionally and translationally inactive, the high abundance of protein kinases and phosphatases in mount 4 suggests that sperm commonly use protein phosphorylation to regulate protein activity.

Previous microarray experiments identified 258 oocyte-enriched genes and 508 genes enriched in both sperm and oocytes (germ line–intrinsic genes) (*6*). The germ line–enriched and oocyte-enriched genes were concentrated in three mountains: mount 7 (12× and 9×, respectively), mount 11 (13× and 13×), and mount 18 (2.4× and 4.1×). In addition, germ line–enriched genes were concentrated in mount 20 (7.5×) [Fig. 3C and movies at (*13*)]. These four mountains contain 483 of the 766 germ line– and oocyte-enriched genes (63%). Of the remaining 283 germ line–enriched genes, 161 (21%) were found in mount 2, which is a large mountain containing many genes involved in diverse biosynthetic pathways.

These four mountains segregate the germ line genes according to their different biological roles. For example, the first two (mount 7 and mount 11) were highly enriched for meiosis and mitosis genes and, therefore, may reflect genes expressed in the early germ line. We identified a set of 23 genes known to be involved in meiosis; 12 are in mount 7 (11× representation factor) and six are in mount 11 (8×) (Web table 1). The list of meiosis genes contains six involved in forming the synaptonemal complex, and all are contained in mount 7 (*19*). We identified a set of 80 genes known to be involved in mitosis (Web table 1). Of these, 16 are in mount 7 (4.4×) and 26 are in mount 11 (10×). The list of mitosis genes contains five that are orthologs of components of the mammalian retinoblastoma (Rb) tumor suppressor complex. The Rb tumor suppressor complex regulates cell growth and division by controlling gene expression throughout the cell cycle (*20*). In *C. elegans*, this complex consists of LIN-35 (Rb), HDA-1 (histone deacetylase), and RBA-1/RBA-2 (both RbAP48) (*21*). All four genes encoding proteins in the Rb tumor suppressor complex were present in mount 11. In addition to these four genes, *lin-9* is implicated in Rb complex formation as *lin-9* mutants have a similar phenotype to *lin-35*, *hda-1* and *rba-2* mutants (synthetic multivulva) (*22*). We observed that *lin-9* was clustered with the Rb complex genes in mount 11. Thus, both mutant phenotype and microarray expression data indicate that *lin-9* may play a functional role in the Rb complex.

Mount 18 and mount 20 were both enriched for protein expression and biosynthesis genes, respectively. We identified 478 genes involved in various biosynthetic pathways, such as energy generation, nucleotide synthesis, carbohydrate metabolism, fatty acid oxidation, and amino acid synthesis (Web table 1). The biosynthesis genes were mildly enriched in mount 18 (2.6×) and strongly concentrated in mount 20 (10×). Then, we identified 390 genes involved in protein synthesis, such as genes encoding tRNA synthetases, ribosomal proteins, chaperones, heat shock proteins, protein translocation components, and RNA processing proteins (Web table 1). These protein synthesis genes are enriched in mount 18 (9.7×) and mount 20 (16×). Biosynthesis and protein expression are highly active during oogenesis, as small germ line cells enlarge into enormous oocytes ready to begin growth of the new embryo. Thus, genes clustered in mount 18 and 20 may correspond to late germ line genes.

Eight genes are known to be expressed primarily in the intestine (Web table 1). Five of the intestinal genes were expressed in mount 8, which is 13× the number expected given the size of this mountain (803 genes) (Fig. 3D). Additional genes in mount 8 are likely to be expressed in the intestine because they encode proteins involved in digestion or protection from bacterial infection. Mount 8 contained five genes that are similar to *Entemeba histolytica N*-acetylmuraminidase (a bacterial lysozyme, 12× enriched), suggesting that these genes may be expressed in the *C. elegans* intestine to digest bacterial cell walls. There were 32 protease genes in mount 8 (out of 116 proteases in the genome, 6.4× enriched) that could be expressed in the intestine to break down bacterial proteins. Carboxylesterases are enzymes used by the intestine to metabolize carbohydrates and sugars; 12 (out of a total of 36 carboxylesterases in the genome, 7.3× enriched) are expressed in mount 8 including *ges-1*, which is known to be expressed in the intestine (*23*). Lipases are enzymes used by the intestine to digest lipids; 15 of the 32 lipases in the *C. elegans* genome are contained in mount 8 (10× enriched). Mount 8 contained the gene *nuc-1*, which encodes a deoxyribonuclease (DNase) expressed by the intestine for digestion of bacterial DNA (*24*). Two genes

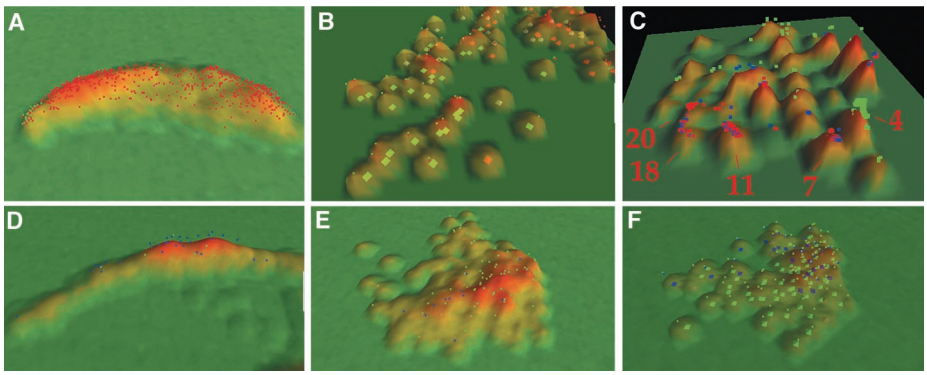

Fig. 3. (**A**) Mount 4 (sperm). Sperm-enriched and MSP genes are shown in red and green, respectively. (**B**) Enlarged view of MSP genes (green) and sperm-enriched genes (red) in mount 4. (**C**) Germ line genes in mounts 7, 11, 18, and 20. Sperm-enriched (green), oocyte-enriched (blue) and germ line–enriched genes (red) from (*6*) are shown. Numbers refer to mountains. (**D**) Mount 8 (intestine). Intestinal (green) and protease (blue) genes are shown. (**E**) Mount 16 (muscle). Muscle (blue) and collagen (green) genes are shown. (**F**) Mount 26 (male). Male-enriched (green) and lectins (blue) are shown.

encoding proteins similar to the mammalian low-density lipoprotein (LDL) receptor were present in mount 8 and could function in the intestine to bind sterols in the lumen and internalize them into intestinal cells. Mount 8 contained two genes that encode insulin-related peptides that might be expressed in the intestine to regulate uptake of nutrients.

Another function of the intestine is that it protects against bacterial infection and from ingestion of harmful chemicals. Mount 8 contained seven out of nine genes that encode antibacterial proteins similar to granulysin of cytotoxic T cells (17× enrichment). These genes may be expressed in the intestine to protect the worm from bacterial infections. Mount 8 contained a metallothionein gene (*mtl-2*), which is known to be expressed in the intestine and function to bind and inactivate heavy metals (*25*). Mount 8 contained eight genes encoding UDP-*N*-acetylglucosamine:alpha-3-D-mannoside beta-1, 2-*N*-acetylglucosaminyltransferase I (where UDP is uridine 5′-diphosphate) out of a total of 64 such genes in the genome (2.8-fold enrichment), including *gly-14*, which is known to be expressed in the intestine (*26*). These genes encode enzymes that are of major importance in the modification and subsequent inactivation of toxic compounds. They could be expressed in the intestine to protect the worm from harmful chemicals.

Thirty-nine genes are known to be expressed primarily in muscle (Web table 1). These genes were enriched in mount 1 (4.1×) and mount 16 (24×). Mount 1 is a large mountain with diverse types of genes, and it was also enriched for many neuronal proteins. In mount 1, the known muscle genes included primarily receptors, extracellular proteins, or receptor-associated proteins such as *egl-19* (which encodes a voltage-dependent calcium channel), *unc-52* (which encodes a component of the basement membrane), or *egl-30* (which encodes a G_{alpha} protein) (Fig. 3F) (*27–29*). Mount 16 included genes that make the muscle filaments themselves, such as those encoding myosin light chain, myosin heavy chain, paramyosin, and two types of troponin (Fig. 3E).

We examined 88 genes that are known to be enriched in neuronal cells. These neuronal genes were clustered in mount 1 (2.7×), mount 6 (6.5×), and mount 13 (3.1×). Both muscle and neuronal genes are clustered in mount 1, and the known muscle or neuronal genes in mount 1 tended to encode receptors or receptor-associated proteins. One possibility is that these genes function in synaptic transmission at neuromuscular junctions. For example, PDZ-containing proteins are expressed in synapses and appear to have a role in clustering or localizing neurotransmitter receptors in both the pre- and postsynaptic densities (*30*). There are 58 genes with PDZ domains in *C. elegans*, and 17 of these were concentrated in mount 1 along with other neuromuscular genes (2.9× enriched). In addition to neuronal genes, mount 13 was enriched for retrotransposons (4.0×), suggesting that retrotransposons might be active in worm neurons.

Previous microarray experiments comparing adult males with adult hermaphrodites identified 1651 male-enriched genes, consisting not only of the sperm genes (enriched in mount 4) but also genes expressed in the soma such as in the male copulatory organ or in male-specific neurons (*7*). Many of the male-enriched genes were clustered in mount 4, corresponding to sperm-enriched genes. The male-enriched genes were also enriched in mount 26 (9.5×) (Fig. 3F). Of the 95 genes in mount 26, 83 are male-enriched (87%) and are likely expressed in the male soma. Mount 26 contained 15 genes that encode cell surface markers (C-type lectins), suggesting that these genes may function to distinguish the extracellular surfaces of male and hermaphrodite cells.

The second general pattern of gene clusters observed in the gene expression terrain map corresponds to sets of genes that form functional modules, such as genes that act in one biochemical pathway or encode similar types of proteins. For example, mount 20 and mount 36 were both enriched for heat shock genes. In particular, 7 of the 10 genes in mount 36 encode heat shock proteins (337× enriched). The remaining three genes (F26H11.3, F58E10.4, and Y43F8B.2A) were not previously known to be involved in the heat shock response. We performed another set of heat shock microarray experiments and found that all three are heat shock–regulated at the 99% confidence level (Table 2). Thus, direct experimental evidence confirmed the genetic relation suggested by the juxtaposition of three unknown genes with known heat shock protein genes.

Mount 32 is highly enriched for histone genes (226×); of the 24 genes in this mountain, 22 are histone genes that comprise the nucleosomal core (H2A, H2B, H3, and H4). The other type of histone (H1) is not part of the nucleosome itself but serves as a linker between nucleosomal subunits on chromatin. There are five histone H1 genes, and three of these are in mount 11 (18×) along with early germ line genes.

The 99 transposons in the *C. elegans* genome consist mainly of Mariner elements, Tc1, Tc3, Tc4, and Tc5 (Web table 1). In most cases,

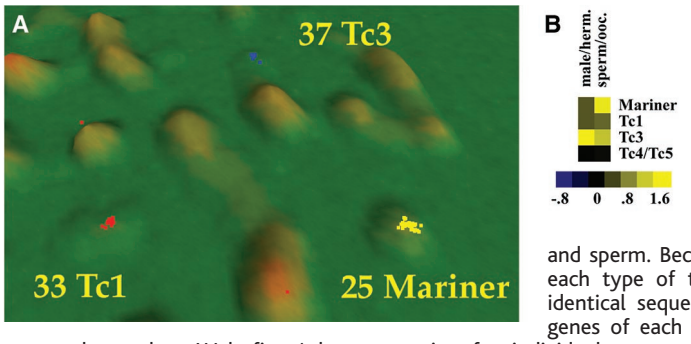

Fig. 4. (**A**) Transposon clusters in the gene expression terrain map. Tc1 (red), Tc3 (blue), and Mariner (yellow) transposons are indicated. Numbers refer to mountains. (**B**) Transposon expression in males and sperm. Because different copies of each type of transposon have nearly identical sequences, expression for all genes of each type of transposon are averaged together. Web fig. 4 has expression for individual transposon copies. Male/herm., experiments comparing adult male to adult hermaphrodite RNAs (*7*); sperm/oocyte, experiments comparing *fem-3(gf)* to *fem-1(lf)* worms (*6*). Yellow and blue denote high- and low-expression levels, respectively.

Table 2. Heat shock induction levels for 10 genes in mount 36. Heat shocks were for 15 min at 33°C, and RNA expression levels were measured 30 min after heat shock. Results show average expression levels (± SE) from four independent experiments. HSP, heat shock protein.

Gene	Induction ± SE	Protein
C12C8.1	65.3 ± 19.3	HSP70
F44E5.4	82.5 ± 24.8	HSP70
F44E5.5	109.7 ± 41.8	HSP70
hsp-16.11	39.7 ± 14.7	HSP-16
hsp-16.1	52.3 ± 15.0	HSP-16
hsp16-2	68.7 ± 22.6	HSP-16
hsp16-41	39.0 ± 5.0	HSP-16
F26H11.3	11.1 ± 2.3	Bromodomain protein
F58E10.4	5.1 ± 1.4	Similar to *S. cerevisiae* YNL155W
Y43F8B.2A	10.5 ± 1.6	Similar to Y43F8B.M

transposons of the same type fell into the same cluster, as was expected because different members of each transposon type have nearly identical sequences and would be expected to cross-hybridize. The Mariner transposons fell into mount 25, most Tc1 copies were in mount 33, and Tc3 copies were in mount 37 (Fig. 4A). Tc4 and Tc5 show more sequence heterogeneity and were spread out in mounts 0, 1, 3, and 9. The expression map showed that the Tc1, Tc3, and Mariner transposon families were expressed differently from each other, suggesting different types of developmental regulation. To begin to elucidate this developmental control, we examined the expression profiles for the transposons in the published microarray data (*6*, *7*). We found that average expression of Mariner transposons was high in sperm relative to oocytes, suggesting that this transposon may have a higher mobilization rate in the male compared with the hermaphrodite germ line (Fig. 4B). We also found that the average expression of Tc3 was high in the male soma, as it is enriched in males versus hermaphrodites but not in sperm versus oocytes.

Additional sets of genes that cluster in the same mountain on the gene expression terrain map are shown in Table 1 and listed in Web table 1. Further investigation is likely to reveal many more clusters of genes on the terrain map.

The gene expression database provides higher resolution than individual microarray experiments because the expression patterns of particular groups of genes are refined by a multitude of experiments. For example, the germ line microarray experiments (*6*) identified 758 genes that are enriched in the hermaphrodite germ line, but the gene expression terrain map was able to subdivide these genes into four mountains (mounts 7, 11, 18, and 20) enriched for genes with distinct biological roles. Furthermore, the position of genes within a mountain in the terrain map often provides information about its function, as we frequently observed that genes with similar function were placed close to each other in a section of one mountain. This level of detail was not observed in microarray experiments comparing only two worm samples (*31*).

The ability to identify candidate genes whose function can subsequently be confirmed by experimental testing depends greatly on the resolution of the terrain map. Some sets of genes (such as the heat shock genes, sperm-enriched genes, nucleosomal histone genes, and ribosomal genes) show tight clustering in which genes that are known to be functionally related are adjacent to each other on the gene expression map. Other groups of genes (such as the retinoblastoma complex genes) may be loosely clustered together in the same expression mountain.

Although the sperm versus oocyte experiments were specifically designed to identify sperm and oocyte genes (hypothesis testing), the terrain map also grouped genes even when they were not specific targets of any of the experiments in the database (undirected knowledge discovery). For example, none of the experiments were specifically designed to reveal expression in muscle, intestine, or neurons, or to show expression by the histone, collagen, or transposon genes (Fig. 1A). Nevertheless, these genes form discrete clusters or mountains on the terrain map, most likely because they showed serendipitous co-regulation in one or more of the experiments in the large database. In many cases, mountains on the gene expression terrain map reveal unexpected interactions between genes. These types of unexpected gene clusters are best revealed using undirected data mining of a global gene expression database rather than testing specific hypotheses.

Caenorhabditis elegans is a powerful model system to analyze biological processes with the use of functional genomics approaches. In addition to global expression studies, efforts are under way to determine the mutant phenotype of most *C. elegans* genes using RNA interference and to identify protein binding interactions on a whole genome level using a high-throughput, yeast two-hybrid approach (*32*–*36*). Thus, there is a rapid accumulation of expression data, mutant phenotypes, and protein binding interactions, making it possible to begin to elucidate cellular, developmental, and organismic processes on a global scale.

References and Notes

1. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
2. M. C. Costanzo *et al.*, *Nucleic Acids Res.* **28**, 73 (2000).
3. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
4. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* **6**, 639 (1996).
5. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).
6. V. Reinke *et al.*, *Mol. Cell* **6**, 605 (2000).
7. M. Jiang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 218 (2001).
8. A. A. Hill, C. P. Hunter, B. T. Tsung, G. Tucker-Kellogg, E. L. Brown, *Science* **290**, 809 (2000).
9. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).
10. S. K. Kim, unpublished data. Personal communications from colleagues are as follows: V. Ambros (Dartmouth College), P. Anderson (Univ. of Wisonsin), I. Callard (Boston Univ.), C. Conley (NASA-Ames), D. Eisenmann (Univ. of Maryland), S. Emmons (Albert Einstein Univ.), A. Fire (Carnegie Institute), M. Hengartner (Univ. of Zurich, Switzerland), T. Johnson (Univ. of Colorado), J. Kimble (Univ. of Wisconsin), J. Lee (Yonsei Univ., Korea), P. Larsen (Univ. of Los Angeles), C. Link (Univ. of Colorado), G. Lithgow (Univ. of Manchester, England), S. Mango (Univ. of Utah), S. Mcintire (Univ. of California, San Francisco), W. Shafer (Univ. of California, San Diego), R. Menzel (Free Univ., Berlin), R. Padgett (Rutgers Univ.), J. Thomas (Univ. of Washington), K. Thomas (Univ. of Missouri), L. Vassilieva (Univ. of Utah), and D. Zarkower (Univ. of Minnesota).
11. G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, B. N. Wylie, *J. Intelligent Inform. Syst.* **11**, 259 (1998).
12. G. S. Davidson, B. N. Wylie, K. W. Boyack, *Cluster Stability and the Use of Noise in Interpretation of Clustering*, IEEE Symposium on Information Visualization, in press.
13. Web figures, tables, movies, and text are available on *Science* Online at www.sciencemag.org/cgi/content/full/293/5537/2087/DC1.
14. We compared the gene clustering results with the use of VxInsight to those using hierarchical clustering, which is a standard method to cluster genes based on Pearson correlation coefficients (*27*). We obtained similar results using the two methods and found that there was strong overlap between mountains formed using VxInsight and gene clusters using hierarchical clustering.
15. Using a conservative Bonferroni correction, the probability of observing one of the red dots in Fig. 2C is approximately 10^{-6}. The actual significance of the entire result is much more than this because there are 64 different overlaps with this level of significance, whereas the random solution contains no overlaps at this significance level.
16. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
17. The representation factor shows whether genes from one list (list A) are enriched in another list (list B), assuming that genes behave independently. The representation factor is defined as: (number of genes in common between both lists)(number of genes in the genome)/(number of genes in list A)(number of genes in list B).
18. S. W. L'Hernault, T. M. Roberts, *Methods Cell Biol.* **48**, 273 (1995).
19. A. M. Villeneuve, personal communication.
20. N. Dyson, *Genes Dev.* **12**, 2245 (1998).
21. X. Lu, H. R. Horvitz, *Cell* **95**, 981 (1998).
22. E. L. Ferguson, H. R. Horvitz, *Genetics* **123**, 109 (1989).
23. L. G. Edgar, J. D. McGhee, *Dev. Biol.* **114**, 109 (1986).
24. C. J. Lyon, C. J. Evans, B. R. Bill, A. J. Otsuka, R. J. Aguilera, *Gene* **252**, 147 (2000).
25. J. H. Freedman, L. W. Slice, D. Dixon, A. Fire, C. S. Rubin, *J. Biol. Chem.* **268**, 2554 (1993).
26. S. Chen, S. Zhou, M. Sarkar, A. M. Spence, H. Schachter, *J. Biol. Chem.* **274**, 288 (1999).
27. R. Y. Lee, L. Lobel, M. Hengartner, H. R. Horvitz, L. Avery, *EMBO J.* **16**, 6066 (1997).
28. L. Brundage *et al.*, *Neuron* **16**, 999 (1996).
29. T. M. Rogalski, B. D. Williams, G. P. Mullen, D. G. Moerman, *Genes Dev.* **7**, 1471 (1993).
30. S. K. Kim, *Curr. Opin. Cell Biol.* **9**, 853 (1997).
31. There are 156 genes that are present on the DNA microarrays but not represented on the gene expression terrain map, either because there is a large amount of missing data or they show almost no variation across experiments.
32. A. G. Fraser *et al.*, *Nature* **408**, 325 (2000).
33. P. Gonczy *et al.*, *Nature* **408**, 331 (2000).
34. F. Piano, A. J. Schetterdagger, M. Mangone, L. Stein, K. J. Kemphues, *Curr. Biol.* **10**, 1619 (2000).
35. I. Maeda, Y. Kohara, M. Yamamoto, A. Sugimoto, *Curr. Biol.* **11**, 171 (2001).
36. A. J. M. Walhout *et al.*, *Science* **287**, 116 (2000).

37. We would like to especially thank S. Scherer (Acacia Biosciences) for guidance and advice on this project, M. Werner-Washburne for help in applying VxInsight to microarray analysis, and A. Owen and L. Lazzeroni for helpful advice on statistics. We thank J. Ryu, P. Roy, and J. Shaw for critical comments on the manuscript. We thank the programmers at the Stanford Microarray Database for their help in the microarray analyses, and Proteome for annotation of *C. elegans* genes. Supported by grants from the National Institute for General Medical Sciences, National Center for Research Resources, Merck Genome Research Institute, Aventis, and by Laboratory Directed Research and Development, Sandia National Laboratories, U.S. Department of Energy (DE-AC04-94AL85000).

13 April 2001; accepted 20 August 2001