Supplemental data

Embryo staging and collection

Bristol N2 worms were grown on *E. coli* strain OP50 at 25°C. For the dissection, staging and aging that follow, everything was performed in a climate-controlled room at 22°C. Young gravid hermaphrodites were cut in 100 µl water in a depression well slide. Bleach solution (10 µl) [4:1 NaOCl (6% available chlorine), 0.5 M KOH] was added and worms were triturated by pipet for ~10 seconds before adding 10 µl 20% BSA and triturating for an additional ~10 seconds. For staging embryos at the four-cell stage, one- and two-cell embryos were collected and washed by serial transfer via mouth pipet through a series of five 100 µl drops of water in 1 cm^2 hydrophobic barrier wells on the surface of a microscope slide previously treated with Sigmacote (Sigma). Four-cell and older embryos were set aside, and two-cell embryos were pooled. Embryos were pulled from the pool and put into a new pool as they reached the four-cell stage. The pool of four-cell embryos was expanded in this way for approximately 5 minutes or until any of the oldest members of the pool reached the six-cell stage. The pool was then double-checked and six-cell or abnormal-looking embryos were eliminated. The pool was then transferred to another 100 µl drop of water on a separate slide and a stopwatch was started so that time zero is relatively late in the four-cell stage. After the appropriate amount of time aging, embryos were examined and any that did not appear to have developed normally were eliminated and all others were transferred to the lid of a 0.6 ml eppendorf tube (in 1-3 µl) and frozen in liquid nitrogen. Nuclei counts were made in controls that were staged and aged as above but frozen on poly-L-lysine treated slides then fixed and stained with DAPI, in order to calibrate aging time with the published lineage and to measure temporal dispersion (data not shown).

For staging embryos at pseudocleavage (PC), mothers were cut as above, except only those embryos with a single partial cleavage furrow were collected and for a period of exactly 3 minutes. A stopwatch was started at the end of the 3 minute collection (PC plus 0 minutes), and the embryos were transferred to the first wash of five serial washes. By the fifth wash (PC plus ~4 minutes) true cleavages had resulted in two-cell embryos, while PC furrows had relaxed resulting in one-cell embryos. One-cell embryos were collected and either frozen at PC plus 6 minutes or aged until PC plus 32 minutes, by which time they had all made it to the early four-cell stage. Because PC is more transient than the four-cell stage, such staging results in smaller cohorts of embryos with less temporal dispersion.

RNA isolation and amplification

For a detailed protocol of the RNA isolation, amplification and labeling procedures please see http://www.mcb.harvard.edu/hunter/. Briefly, RNA was isolated by adding 100 μl TRIzol reagent (Invitrogen), vortexing briefly, pipetting up and down eight to ten times, adding 7 μl water and 1 μl linear polyacrylamide (5 μg/μl; GenElute LPA, Sigma), vortexing for 10 seconds, adding 20 μl $CHCl_3$, vortexing for 30 seconds, spinning at 13,000 for 5 minutes, transferring the aqueous phase, adding 60 μl isopropanol and incubating overnight at –20°C. RNA was pelleted by spinning at 16,000 ***g*** for 25 minutes, washed once with 75% ethanol and resuspended in 4 μl DEPC-treated dH_2O (including 20 ng of the (dT)-T7 primer). In some cases embryo collections were pooled at the TRIzol step in order to obtain either 10 embryos per RNA prep (PC6, PC32, 0 minutes) or 15 (all other samples) in a final volume of 100 μl TRIzol.

mRNA was amplified and labeled as described elsewhere (Baugh, 2001), with the notable exception that 20 ng (dT)-T7 primer was used in a 2 μl reverse transcription reaction in the first round of amplification (as opposed to 10 ng and 1 μl). Amplified RNA was quantified by UV absorbance at 260 nm and analyzed by electrophoresis. Yields for the samples used ranged between 1.7 and 18.7 μg.

Array hybridizations

Hybridizations were performed essentially as described in the Affymetrix Expression Analysis Technical Manual, except 1 μg amplified RNA was used per hybridization. Arrays were stained using the Affymetrix recommended antibody amplification method, and scanned with the Affymetrix GeneChip scanner. Four replicate scans were averaged for each array; this averaging improved the signal-to-noise ratio for the arrays, compared with an average of two scans.

Data reduction

Array images were reduced to probe intensity values and stored in .cel file format using Affymetrix GeneChip 3.1. Data in .cel files were normalized and converted to average difference values using the dChip software (β-test version 2001) (Li and Wong, 2001). Average difference values were converted to transcript abundance estimates, in units of ppm, by reference to a standard curve of 11 spiked in vitro transcripts as described elsewhere (Hill, 2001).

Absolute decisions (present/absent/marginal calls) were computed by GeneChip 3.1. The absolute decision is based on the magnitude of the difference between hybridization intensity and array background and on the fraction of probe pairs with fluorescence above background and noise (see the Affymetrix GeneChip analysis suite user guide for details).

Moving average

For the purposes of plotting gene expression profiles, clustering and phasing, the data were transformed by computing the moving average of means over two time points. Ten averages of adjacent timepoints that were part of distinct series were computed, starting with PC32 (about 4 minutes younger than 0 minutes) and 0 minutes. As a result, the first moving average time point is –2 minutes relative to the four-cell stage and the last time point is 165 minutes (the average of 122 minutes and 186 minutes). The purpose of the moving average was to reduce systematic gene-specific differences between series 1 and series 2. Hence, PC6 and PC32, both part of series 2, were not averaged. Moving average transformed data was not used for statistics or developmental classification.

ANOVA

A modified Welch F statistic (Zar, 1999) was used for all hypothesis testing. Individual replicate data was $\log_e$ transformed as the first step of all statistical analyses. The calculation of the modified Welch F statistic for each gene was as described [see Eqns 10.22-10.27 by Zar (Zar, 1999)], except that 'regressed' error estimates r^2 were substituted for the s^2 error terms in the equations. For each gene, these regressed error estimates were abundance-dependent pooled error estimates that represented a median error estimate from a window of genes of similar abundance to the gene of interest.

Regressed error estimates were computed as follows. Replicate data (containing K timepoints and G genes) was log-transformed, and the (KG) means u_{kg} and variances s_{kg}^2 (k=1..K, g=1..G) were computed for all genes on a given array design (A, B or C). Regressed error estimates r_{kg}^2 were windowed medians of the observed variances s_{kg}^2, using a window size of W=0.01KG. To reduce computation time, we applied a 'jumping', not a 'running' median. That is, all r_{ij}^2 within the first window were assigned the median of that window, the window was shifted by W and the process repeated. Based on empirical testing of windowed medians to improve the median fit to s_{kg}^2, we applied two constraints to the windowed median estimates to make the fit robust and consistent with a simple two component (additive background + multiplicative sampling error) noise model: (1) r_{ij}^2 was constrained to be a decreasing function of the mean frequency in log space, i.e $r_{ij}^2 \leq r_{kl}^2$, for $u_{ij} > u_{kl}$; (2) as the windowed median simply assigned the median of a window of s_{ij}^2 values to each u_{ij}, a strict functional relationship was not guaranteed by this fit alone. Therefore, in rare cases when the median windows assigned multiple r_{ij}^2 values to a single u_{ij} value, we re-assigned to that mean u_{ij} the largest regressed error r_{ij}^2 that was associated with that mean in the dataset.

A randomization test was used to compute the probability P_g of the observed F statistic for gene *g* under the null hypothesis that developmental time had no effect on expression. As the number of experimental replicates was different at some timepoints on some array designs, each array design (A, B or C) was randomized independently.

The randomization test was carried out as follows. For each array design, the log-transformed (G×N) data matrix was assembled, where G was the number of genes on the array, and N the total number of observations (for example, for the A array design, G=6617 and N=50). For each of the G genes the F statistic was computed, within series 1 (K=5 timepoints), within series 2 (K=7 timepoints), and for each paired-timepoint contrast of interest (K=2 for each contrast). The N timepoint labels were then randomly shuffled, and all F statistics recomputed. The random permutation was repeated N_P=200 times to generate one $G \times N_P$ matrix containing the null distribution of F for each of the two within-series ANOVAs, and equivalent

$G \times N_P$ matrices for each paired-timepoint contrast. Each of the G gene-specific F statistics from the observed data were referred to their corresponding null distribution, and the p-value for each gene *g* was computed as:

P_g=(count of $F_{null} \geq F_{obs}$)/GN_P

In the null distribution we included all genes, as opposed to referring each gene to the null distribution arising from random shuffling of the observations of that gene only. Thus, each null distribution contained $G \times N_P \sim 6 \times 10^5$ observations of F. To validate this approach, we examined the null distribution of the F statistic for 22 probesets corresponding to 11 cRNAs spiked into the A array hybridizations at levels from ~3-1000 transcripts per million. The null distribution of F was not correlated with expression level for these spiked messages, i.e. F was pivotal in the sense of Westfall and Young (Westfall and Young, 1993).

Phasing

Moving average transformed data of the 3157 RD genes with *P*<0.001 in either within-series ANOVA was used for phasing. Each gene was normalized by its mean over all ten moving average time points. Normalized abundances were $\log_2$ transformed. The ten moving average time points were subdivided into four time windows: –2 minutes; 12 minutes, 32 minutes and 47 minutes; 60 minutes, 75 minutes and 92 minutes; and 112 minutes, 133 minutes and 165 minutes. A mean value was calculated for each of the four windows and the values were ranked 1-4 for each gene. The genes were then sorted in an iterative, nested fashion. First, they were sorted according to earliest window rank. Genes ranking highest in the earliest window were set aside and the remaining genes were sorted according to second window rank. Again, genes ranking highest in the second window were set aside and the remaining genes were sorted according to third window rank. The process was repeated until four groups of genes had been defined. Each group was sorted again, independent of the other three groups, according to the mean value for the time window preceding the highest rank of the group. For the earliest time window the latest time window was used as the preceding window. This second sort, performed four independent times, makes transitions between and down each of the four groups of genes smooth. Breakpoints marking the boundaries of the four groups are nevertheless apparent and should not be misinterpreted. The phasegram was plotted using TreeView (Eisen, 1998).

Cluster analysis

We desired a clustering algorithm that is insensitive to experimental noise, does not force all input genes into a cluster, and does not require an a priori determination of the number of output clusters. Clusters were generated by the QT clustering algorithm (Heyer, 1999). The algorithm assembles a series of clusters ordered by size, largest first, with no limit on the number of clusters other than the coherence (cluster diameter) defined a priori (0.7 in our clusters). To ensure robust clusters the distance metric used for clustering was 1-R_{avg}, where R_{avg} was the average Pearson correlation coefficient between moving average profiles over 20 realizations of the data plus simulated noise. Noise was generated by a two-component model consisting of an additive Gaussian background with standard deviation 2 ppm, and a multiplicative Gaussian sampling error with s.d.=0.1. Simulated data were floored at 1 ppm.

Expression pattern classification

Paired timepoint ANOVA tests serve as the primary basis for classification, though within-series ANOVA tests as well as present calls in the first time point (PC6) are also considered. A cutoff of $P<0.01$ was used with all statistical tests unless otherwise noted. Paired timepoint tests used include ten spanning roughly one cell cycle (PC6×PC32, PC32×23 minutes, 0×41 minutes, 23×53 minutes, 41×66 minutes, 53×83 minutes, 66×101 minutes, 83×122 minutes, 101×143 minutes, 122×186 minutes) as well as eight more spanning roughly two cell cycles (PC6×23 minutes, PC32×53 minutes, 0×66 minutes, 23×83 minutes, 41×101 minutes, 53×122 minutes, 66×143 minutes, 83×186 minutes). All of these tests are within only one of the two time series, the former consisting of adjacent timepoints within a series and the latter consisting of alternate timepoints within a series. Significant increases and decreases in abundance observed within defined time domains were used to classify genes (see below). Time domains were selected following visual inspection of the clustered data and were defined as follows: 'maternal degradation' domain equals PC6 to 83 minutes; 'embryonic' domain equals PC6 to 186 minutes, ‘induction following degradation’ domain equals 53-186 minutes.

The definition of each class is as follows: ‘maternal’ genes are called present in at least one of the three PC6 replicates; ‘embryonic’ genes increase significantly during either time course. Specifically, among the genes flagged as dynamic in either of the two within-series ANOVAs, embryonic genes are the subset that also significantly increase in at least two of the eighteen total paired timepoint tests or significantly increase in either the 122×186 minutes or 83×186 minutes comparison. ‘Maternal degradation’ (MD) genes are the subset of maternal genes that decrease without first increasing in abundance. Specifically, among the genes flagged as dynamic in either of the two within-series ANOVAs, MD genes decrease significantly in at least two of the ten total paired timepoint tests, but do not significantly increase ($P<0.05$) in any paired timepoint test before the earliest significant decrease ($P<0.01$). ‘Embryonic transient’ genes are the subset of embryonic genes in which the latest significant increase is earlier than their latest significant decrease. ‘Maternal-embryonic’ genes are in the intersection of the maternal and embryonic classes. ‘Maternal degradation-embryonic’ genes are the subset of maternal degradation genes that significantly increase in at least two of the eight total paired timepoint tests in the ‘induction following degradation’ time domain. ‘Maternal-embryonic transient’ genes are in the intersection of the maternal and embryonic transient classes. ‘Maternal degradation-embryonic transient’ genes are in the intersection of the Maternal degradation-embryonic and embryonic transient classes. ‘Strictly maternal’ genes are the subset of maternal genes that are not also classified as embryonic. ‘Strictly embryonic’ genes are the subset of embryonic genes that are not also classified as maternal. ‘Strictly maternal degradation’ genes are the subset of maternal degradation genes that are not also classified as embryonic. ‘Strictly embryonic transient’ genes are the subset of embryonic transient genes that are not also classified as maternal.

Select classes were subclassed by the defining timepoint in the expression profile of each gene; there is no overlap between the subclasses of a particular class. Maternal degradation subclasses are based on the earliest significant decrease (abbreviated ‘pd’ for primary decrease). Embryonic and strictly embryonic subclasses are based on the earliest significant increase (abbreviated ‘pi’ for primary increase). Embryonic transient subclasses are based on the time of max expression (abbreviated ‘max’).

REFERENCES

Baugh, L. R., Hill, A. A., Brown E. L. and Hunter C. P. (2001). Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res.* **29**, E29.

Eisen, M., Spellman, P., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868.

Hill, A. A., Brown, E. L., Whitley, M. Z., Tucker-Kellog, G., Hunter, C. P. and Slonim, D. K. (2001). Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.* **2**, 0055.1-0055.13.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31-36.

Westfall, P. H., and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. New York: Wiley.

Zar, J. H. (1999). *Biostatistical Analysis*. Upper Saddle River, NJ: Prentice-Hall.