Available online at www.sciencedirect.com
ScienceDirect
Developmental Biology 300 (2006) 476–484
DEVELOPMENTAL BIOLOGY
www.elsevier.com/locate/ydbio

# A database of mRNA expression patterns for the sea urchin embryo

Zheng Wei, Robert C. Angerer, Lynne M. Angerer *

*National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD 20892, USA*

Received for publication 5 June 2006; revised 10 August 2006; accepted 15 August 2006
Available online 22 August 2006

**Abstract**

We present an initial characterization of a database that contains temporal expression profiles of sequences found in 35,282 gene predictions within the sea urchin genome. The relative RNA abundance for each sequence was determined at 5 key stages of development using high-density oligonucleotide microarrays that were hybridized with populations of polyA+ RNA sequence. These stages were two-cell, which represents maternal RNA, early blastula, the time at which major tissue territories are specified, early and late gastrula, during which important morphogenetic events occur, and the pluteus larva, which marks the culmination of pre-feeding embryogenesis. We provide evidence that the microarray reliably reports the temporal profiles for the large majority of predicted genes, as shown by comparison to data for many genes with known expression patterns. The sensitivity of this assay allows detection of mRNAs whose concentration is only several hundred copies/embryo. The temporal expression profiles indicate that 5% of the gene predictions encode mRNAs that are found only in the maternal population while 24% are embryo-specific. Further, we find that the concentration of >80% of different mRNAs is modulated by more than a factor of 3 during development. Along with the annotated sea urchin genome sequence and the whole-genome tiling array (the transcriptome, Samanta, M., Tongprasit, W., Istrrail, S., Cameron, R., Tu, Q., Davidson, E., Stolc, V., in press. A high-resolution transcriptome map of the sea urchin embryo. Science), this database proves a valuable resource for designing experiments to test the function of specific genes during development.
Published by Elsevier Inc.

*Keywords:* Gene prediction; Microarray; Genscan

## Introduction

The sea urchin embryo provides a relatively simple and tractable system for analyzing early development of an invertebrate deuterostome, which is the closest outgroup to the chordates. Great progress has been made over the past decade in defining the molecular asymmetries that establish the embryonic axes and the genetic hard wiring, signaling pathways and cell–cell interactions that specify endoderm, mesoderm and ectoderm and some of their derivative structures. The recent annotation of the sea urchin genome sequence will stimulate this effort by facilitating the identification of additional gene regulatory and signaling molecules whose functions can be incorporated into existing gene regulatory networks (Angerer and Angerer, 2003; Davidson et al., 2002) and others being developed (e.g., Burke et al., 2006). Defining the gene regulatory networks underlying specification and subsequent differentiation of all major regions of the embryo is now a realistic goal.

To facilitate sorting among the many newly identified candidates for those likely to function at different points and in different processes during embryogenesis, we have used a DNA microarray approach to define the temporal patterns of expression for a set of mRNA sequences that includes the large majority of predicted protein coding genes encoded in the *Strongylocentrotus purpuratus* genome. We used the *ab initio* gene prediction program, Genscan (Stanford University), to identify putative genes together with BLAST searches of the NCBI non-redundant database to select within each prediction the most highly conserved sequence, since this has the highest probability of representing an authentic gene. In nearly all cases, each sequence was represented by duplicate sets of five oligomers (10 signals/prediction) on each of five microarrays.

This collection of oligomer probes was hybridized with labeled targets representing total polyA+ RNA from embryos at major stages of early development [2 h (maternal), 15 h (early blastula), 30 h (early gastrula), 48 h (late gastrula) and 72 h

* Corresponding author.
*E-mail address:* langerer@mail.nih.gov (L.M. Angerer).

0012-1606/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.ydbio.2006.08.034

(pluteus)]. RNA from 2-h embryos (first cleavage) was used as a surrogate for egg RNA because mRNA stored in the egg has very short polyA tracts, and therefore cannot be recovered efficiently (Wilt, 1977). This population consists almost exclusively of maternal transcripts because the amount of zygotic RNA produced during the first cell cycle from 2–4 gene copies is minute. At the early blastula stage (~200-cell; 8th cleavage; 15 h), the embryo is morphologically undifferentiated, but specification of the major tissue territories is well underway: the endomesoderm gene regulatory network (GRN) has been active since shortly after the 4th cleavage, and *nodal*, the gene currently most upstream in oral–aboral patterning of ectoderm has been activated. Between the early blastula and early gastrula stages (15 to 30 h), all the major tissue territories are specified and programs of cell-type-specific differentiation are initiated. Morphogenesis begins during this interval with ingression of primary mesenchyme cells and invagination of the archenteron. Between 30 h and 48 h, gastrulation is completed, the skeleton is elaborated, secondary mesenchyme cell types differentiate, the ectoderm develops to form oral and aboral epithelia separated by a neurogenic ciliary band, and the first neurons appear in the animal plate. Embryogenesis is complete by 72 h with the formation of a 2-armed pluteus larva with well-differentiated tissues containing approximately 15 cell types.

Here we evaluate the quality of the expression patterns provided by the microarray data. We show that mRNA sequences varying in prevalence from ~200 to ~150,000 copies per embryo are reliably detected using this resource and that the temporal profiles for the large majority of genes accurately reflect their modulations in expression during early development.

## Methods

### *Gene predictions from sea urchin genome sequence*

We used an early draft assembly (Spur20050415, http://www.hgsc.bcm.tmc.edu/projects/seaurchin/) of the 800-mbp genome of the sea urchin *S. purpuratus* (Human Genome Sequencing Center at Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu)). This was a collection of 188,642 contiguous sequences (scaffolds) that were assembled from 6× coverage whole-genome shotgun (WGS) sequences. To obtain gene predictions within this assembly, we used Genscan (Stanford University), because it is one of the most accurate *ab initio* gene prediction programs (Burge and Karlin, 1997; Guigo et al., 2000).

To estimate the fraction of genes that is included in the set of Genscan predictions, we searched it for 105 randomly selected *S. purpuratus* cDNAs, which had been determined experimentally and deposited in the Genbank non-redundant (nr) database. The representation is extremely high because all were present. While some exons are incorrectly predicted, it is important to recognize that, for the design of the microarray and the purposes of this analysis, completely accurate gene models are not required (see below and Results and discussion).

### *Probe design*

For each of the 35,282 predictions, we selected 5 oligomer probes from the most highly conserved sequence as determined by BLAST searches of the NCBI nr protein database. Two probes were selected with OligoArray (Rouillard et al., 2003). One set was designed to have a uniform predicted Tm among probes, which was achieved by varying lengths from 40 to 60 nt. The second set consisted of 60-mers with variable Tms. The remaining three probes were randomly selected 60-mers. In most cases, the different probes represent non-overlapping sequences, but for some very short predicted open reading frames, they were partially overlapping. In a few cases, 5 different probes could not be identified.

### *Microarray design*

Each microarray contained two identical blocks of 176,000 different probes (352,000 probes, total), each block representing 35,282 predictions. For negative controls, we included 3 different probe sequences that do not match any sequence in the sea urchin genome (BLAST E value >1). Each of these was represented 6 times on each array. As positive controls for constant signals at different stages, we also included 5 different probes complementary to the ubiquitin open reading frame. Each control probe was repeated 3 times within each block (30 determinations/array). Finally, each block contained concentration standards provided by Nimblegen. Five identical microarrays were produced using Nimblegen's photolithographic Maskless Array Synthesis (MAS) technology (http://www.nimblegen.com/technology/design.html).

### *PolyA+ mRNA preparation*

*S. purpuratus* embryos were cultured at 15°C in artificial sea water and collected at 2 h, 15 h, 30 h, 48 h or 72 h after fertilization. RNA was purified using Trizol (Invitrogen, Inc.) according to the manufacturer's instructions. Contaminating DNA was removed by incubation with DNase and the samples were further purified by extraction with phenol/chloroform and ethanol precipitation. PolyA+ RNA was purified by oligo dT affinity chromatography (Ambion, Inc.). The quality of the preparations was verified using formaldehyde gel electrophoresis and absorption spectral data.

### *Microarray processing*

Nimblegen, Inc. prepared labeled cRNAs from polyA+ RNA for hybridization to each microarray. Nimblegen Services carried out the hybridizations in a solution containing 50 mM MES buffer, pH 6.6, 0.5 M Na+, 10 mM EDTA, 0.005% Tween20 at 45°C, for 16–20 h and used 100 mM MES buffer, pH 6.6, 0.1 M Na+, 0.005% Tween20, 45°C as the most stringent wash. The scanned microarray signals were background-adjusted and quantile-normalized among the five arrays. The expression value is the robust multi-array average (RMA) (Irizarry et al., 2003) (http://www.nimblegen.com/technology/index.html).

### *Abundance estimates for different mRNAs*

The number of mRNA copies/embryo at 48 h (late gastrula) for 49 genes was determined by M. Howard-Ashby and posted at http://annotation.hgsc.bcm.tmc.edu/urchin/cgi-bin/login.cgi. The signal intensities from the microarray experiment were normalized values representing the expression levels of the same genes at 48 h.

### *Clustering and imaging of the temporal expression profiles*

Cluster and TreeView software (http://rana.lbl.gov/EisenSoftware.htm) was used to visualize the normalized temporal expression profiles of 268 annotated genes encoding transcription factors. The patterns were clustered with self-organizing maps (SOM; 100,000 iterations) (Eisen et al., 1998).

## Results and discussion

The set of gene predictions from which the oligomer probes were designed is described in Fig. 1. Our goal was to identify sequences that represent as many different putative genes as possible and to identify probes that most reliably detect authentic gene sequences. From an initial 97,000 Genscan predictions, 42,000 were selected because they showed some similarity ($E \leq$ e-3) sequences in the NCBI nr protein database. From this subset, most sequences encoding reverse transcriptase and other sequences associated with transposable elements as well as multiple copies of genes from repetitive gene families (e.g., early variant histones)

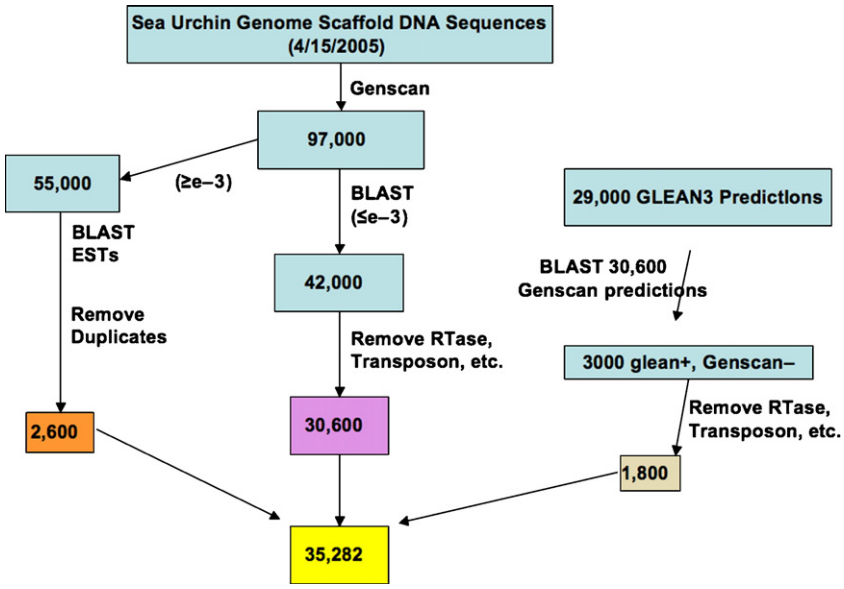


Fig. 1. Selection of gene predictions for representation on developmental microarrays. The draft assembly of the sea urchin genome (4/15/2005) was searched for exons by Genscan. These predictions were separated into conserved ($E \leq$ e-3) and non-conserved ($E \geq$ e-3) by BLAST searches of the non-redundant protein database at NCBI. Many duplicates as well as sequences encoding reverse transcriptase (RT) and other sequences associated with transposable elements and repetitive gene families were eliminated. The resulting set of 30,600 Genscan predictions was augmented with EST sequences found in the non-conserved set (orange box) and with Glean3 predictions not found in the conserved set (gray).

were removed, leaving 30,600 predictions (pink). To minimize the problem of including probes against incorrectly predicted exons, we extracted from each prediction the most conserved sequence based on alignments of BLAST results. The advantage of this approach is that it selects for authentic genes; the disadvantage is that it is more difficult to discriminate among genes containing similar highly conserved sequences. The remaining 55,000 predictions were used as queries in BLAST searches of the *S. purpuratus* EST database. After removing duplicates from the matches with greater than 97% identity, the remaining 2600 EST sequences were added to the set. When the GLEAN prediction list (Zhang et al., submitted for publication) became available, 3000 additional sequences were identified. After removing sequences associated with transposable elements and repetitive gene families, 1800 were added to the set. The final set contained a total of 35,282 predictions (yellow). This set contains some redundant sequences resulting from the fact that this early assembly contained some duplicates (alleles and assembly errors) and some partial gene sequences. There may also be some overlap between the EST and Glean3 sequences that were used to supplement the Genscan predictions.

### *Positive and negative controls*

The distribution of normalized signal intensities for the gene predictions (without background subtraction; Fig. 2, black line; 0×) spans approximately five orders of magnitude [~10 to $6 \times 10^4$ arbitrary units (AU)]. This distribution is similar at all stages of development (data not shown). A large fraction of the signal intensities clusters between 20 and 120 AU, most of which are attributable to background. Two factors contribute to the noise levels: one (general background) results from a variety of factors such as sample preparation and manufacture and processing of the arrays (labeling, hybridization and scanning). The second (cross-reaction) results from weak non-specific hybridization among marginally related sequences. To directly measure general background, we included 60-mers representing 3 different random sequences that do not match any sequence in the sea urchin genome as determined by BLAST searches ($E > 1$). The average signal (from 18 measurements/array; see Methods) for these true negative probes was very low (28 AU ± 2.1) and reproducible among the microarray slides (Fig. 3A). When 28 AU, termed the 1× background value, is subtracted from all the signals, the resulting distribution of signal intensities is as shown by the blue line in Fig. 2. To estimate the additional background resulting from weak cross-reaction, we monitored changes in the signal intensity distribution as a function of increasing levels of background subtraction (Fig. 2). At 2× subtraction (56 AU), the distribution changed dramatically to a more symmetric distribution (purple line), which is expected since the large majority of different mRNAs, i.e., the complex class of mRNAs (Galau et al., 1974), are present in the embryo at quite uniform concentration (1000–3000 copies/embryo). The distribution was further tightened at low signal intensities by increasing the subtraction to 3× (magenta line; 84 AU). This level of subtraction does not eliminate signals from transcripts known to be very rare, such as *pmar1b* (Oliveri et al., 2002), *foxy* (Samanta et al., submitted for publication) or *foxc* (Ransick et al., 2002). For all subsequent analyses this amount was subtracted from each signal value, unless otherwise indicated.

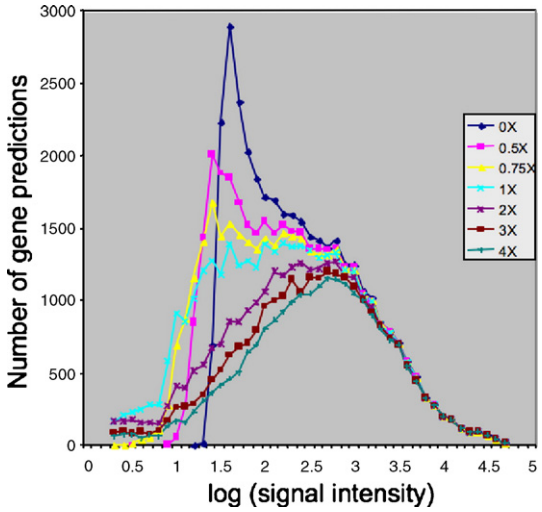


Fig. 2. Signal intensity distributions as a function of background subtraction. The average measured background with negative control probes of 28 AU is designated 1× subtraction. Shown is the distribution of signal intensities on a log scale versus the number of predictions. Each curve is the distribution that results from subtraction of a specific background value. The center of the distribution after 3× subtraction corresponds to signal intensities of ~800 AU, which in turn corresponds to 1000–3000 transcripts/embryo. See text for details.

The majority of authentic signals are found in the symmetric distribution, the mode of which (~800 AU) corresponds to the average concentration of the complex class of rare mRNAs (1000–3000 copies/embryo). The small skewing of the distribution between $1\times10^4$ and $6\times10^4$ AU reflects the contribution of abundant transcripts from a few genes that are present at concentrations up to approximately several hundred fold higher than those in the complex class (Shepherd and Nemer, 1980).

To verify that relative signals among the microarrays were accurately normalized, we used 5 different probes complementary to ubiquitin mRNA, which is widely employed to represent an mRNA present at approximately constant abundance throughout development. The normalized values from 30 readings from each microarray are essentially constant at all stages (Fig. 3B). The facts that the ubiquitin signals and the background levels are consistent among the 5 microarrays indicate that signals for a given mRNA can be reliably compared among stages.

### *Microarray signal intensities and transcript abundance*

To evaluate the extent to which microarray signals accurately reflect differences in the concentration of different mRNAs, we compared signal intensities with estimates of mRNA copies/embryo, as determined by quantitative PCR, for 49 cases where data were available for the same developmental stage and microarray probes were specific. A scatter diagram of the data is shown in Fig. 4 on $\log_2$ scale to better visualize the fold differences in expression level for rare mRNAs. This analysis indicates that while there is a correlation ($r=0.64$), some individual points vary from a strictly linear relationship by as much as tenfold as a result of the differences in probe sensitivity and the combined experimental errors of the different analyses. Most importantly, however, the relative signal intensities at different developmental stages for probes representing any one gene are reliable as demonstrated by constancy of ubiquitin and background signals among microarrays. These findings are in good agreement with previous studies (Chudin et al., 2002) showing that absolute signal intensities for different genes do not reliably reflect differences in transcript abundance levels because of differences in hybridization efficiency for different probe sets, whereas relative signals reporting the activity of the same gene under different circumstances do.

These data provide rough estimates of mRNA abundance that will be helpful in identifying interesting candidate genes and designing experiments. For example, the mRNAs present at ~1000–3000 molecules per embryo give signals around 500–1000 AU, as discussed above (Fig. 2), whereas more abundant mRNAs provide signal intensities between 25,000 and 200,000 AU. *SpHE* and *SpAN* mRNAs previously shown by RNA excess titration to be present at 150,000 and 25,000 copies/embryo at 12 h (Reynolds et al., 1992) gave signals of 42,200 and 5800 arbitrary units (AU), respectively. At the rare end of the distribution, values for a few known extremely rare mRNAs, such as *pmar1b* (250 copies/embryo at 15 h) (Oliveri et al., 2002) and FoxC, which is expressed in only a few cells during development (Ransick et al., 2002) were correspondingly lower (200–300 AU after 3× subtraction). Genes whose

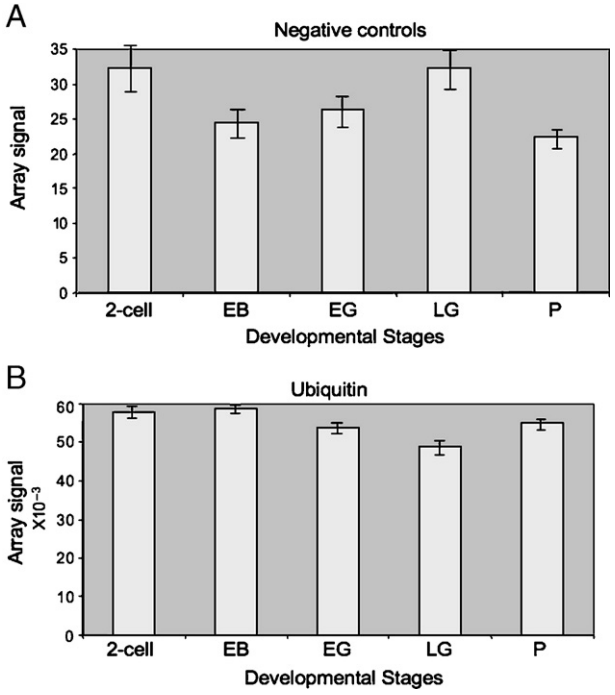


Fig. 3. Positive and negative controls. (A) Background signals at two-cell, early blastula (EB), early gastrula (EG), late gastrula (LG) and pluteus (P) stages were determined using three different probe sequences that were unrelated to any sequences in the sea urchin genome. These were replicated 6 times on each array. The signal intensities are the average of 18 values and the bars represent the standard error. (B) Relative levels of *ubiquitin* mRNA at the same developmental stages. The results shown here are the average of 30 *ubiquitin* signals/microarray and the bars represent the standard error.

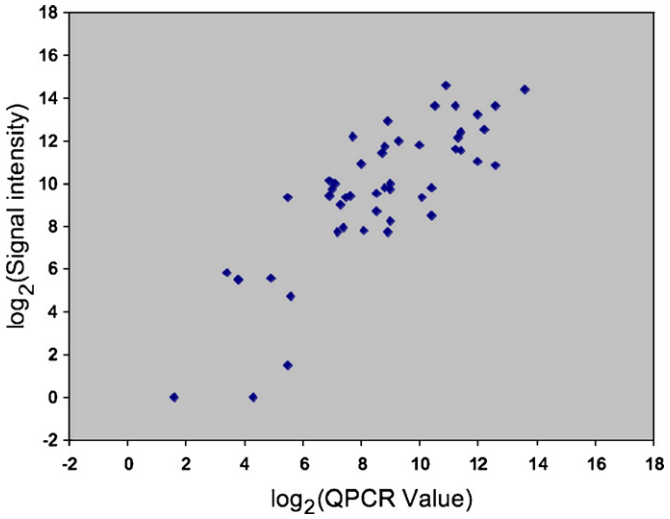


Fig. 4. Correlation of mRNA abundance values determined by microarray intensities and QPCR. The number of mRNA copies/embryo at 48 h (late gastrula) was determined for 49 genes by QPCR (see Methods). Each value is plotted versus the corresponding normalized microarray-generated value for the same gene at 48 h. The data are plotted on $\log_2$ scales.

expression during embryogenesis is not detectable by RNA excess titration or QPCR, such as *Hox 2*, *8* and *11/13a*, also gave no signals above background in the microarray analysis.

### *Comparison of signals for genes with known temporal expression patterns*

To more directly evaluate the accuracy of the temporal expression profiles, we compared the patterns of expression for 4 mRNAs that have been quantitated at the same stages as used in the microarray analysis (Lee and Davidson, 2004; Oliveri et al., 2002; Otim et al., 2004; Yuh et al., 2005). The mRNAs for each of these genes accumulate to very low (*pmar1b*), intermediate (*gatae, brn1/2/4*) or high (*endo16*) levels. As shown in Fig. 5, the microarray (red)- and QPCR (blue)-generated expression patterns are generally in good agreement. Minor differences in the shapes of the curves result from either differences in the developmental rate of the embryos used for these experiments and/or the combined experimental errors of the different analyses.

The reliability of the microarray-generated profiles was further substantiated qualitatively by examining the expression of genes that constitute the well-characterized endomesoderm gene regulatory network (Davidson et al., 2002) (Fig. 6). The genes have been clustered according to four modes of expression: (1) both maternal and zygotic; (2) zygotic, with the maximum at the early blastula stage (EB); (3) zygotic, with initial upregulation of expression at EB; and (4) zygotic, with initial upregulation of expression at the early gastrula stage (EG). To facilitate comparisons, mRNAs giving higher and lower signals within the same mode of expression have been plotted separately. Signals for two very rare mRNAs, *pmar1b* and *tbr*, have been multiplied by 10 to facilitate their visualization. In nearly all of these 32 cases, the patterns reliably reflect QPCR or in situ hybridization data published or posted at the Davidson web site (http://sugp.caltech.edu/endomes/). Maternal mRNAs that continue to be expressed throughout development include *SoxB1*, *hnf6*, *β-catenin*, *suppressor of hairless su*(*H*), *lef1* and *tbr*; early transient zygotic messages include *pmar1b* and *krl*; early mRNAs whose products are involved in stabilizing the network (*Otx*, *gatae* and *blimp1*) are either present continuously or appear during blastula stages and mRNAs encoding downstream differentiation products appear later (*endo16*, *sm50*, *sm30*, *apobec*, *ficolin*). The patterns of expression for several genes reflect their activity not only in endomesoderm, but also in other embryonic territories. For example, both *gsc* and *dri* are expressed at low levels early when they function in the PMC network, and at much higher levels later when they are expressed in the oral ectoderm. Several profiles differ at one of the stages from reported expression patterns. For example,

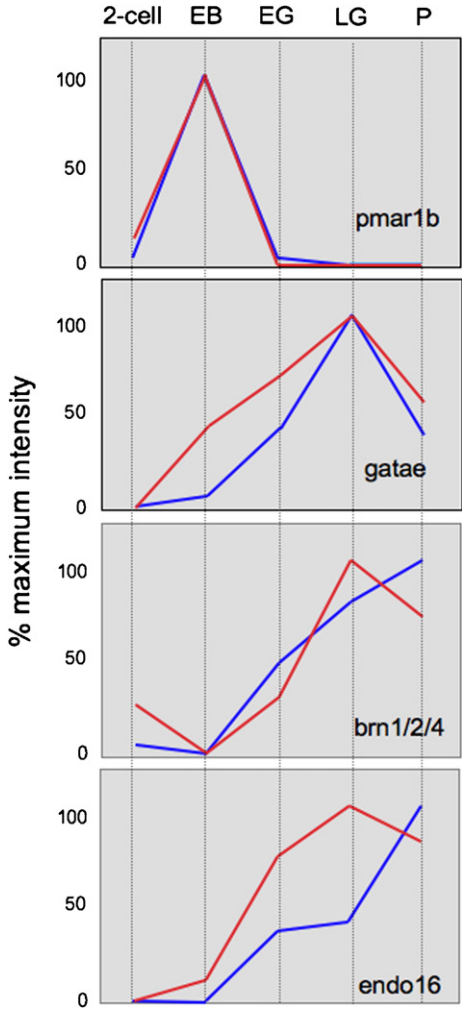


Fig. 5. Temporal expression array and quantitative RTPCR measurements generate similar developmental expression patterns. Data for each assay on the expression of each gene are plotted as % maximum intensity. The relative mRNA copies/embryo at different developmental stages were determined by QPCR for *pmar1b* (Oliveri et al., 2002), *gatae* (Lee and Davidson, 2004), *brn1/2/4* and *endo16* (Yuh et al., 2005) as shown by blue lines and scale. Normalized microarray signal intensities in arbitrary units (AU) are shown in red. Gene IDs are given in the Supplemental data.

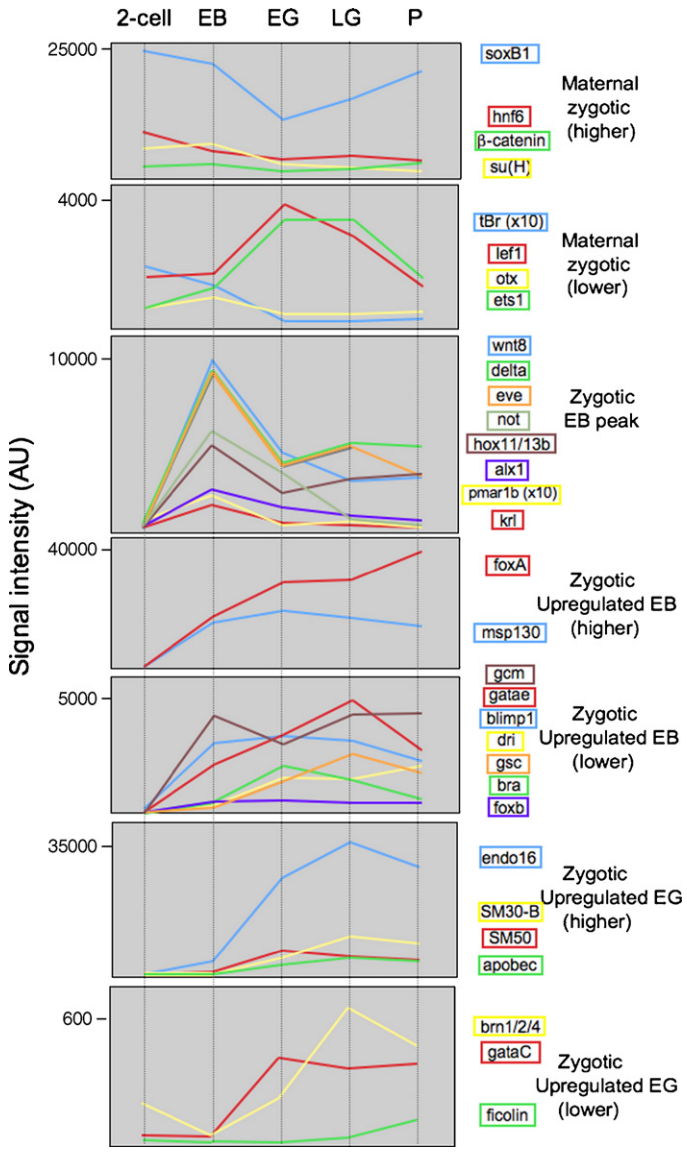


Fig. 6. Microarray data generates reliable temporal expression profiles. Shown are the well-characterized genes that constitute the sea urchin embryo endomesoderm gene regulatory network (Davidson et al., 2002). The two top panels show genes expressed both maternally and zygotically at either high (top) or low (second from top) levels. The bottom 5 panels show patterns for genes expressed during embryogenesis and grouped as described at right. Gene IDs are given in supplemental information. *Su(H)*, *suppressor of hairless*; *Krl*, *kruppel-like*; *dri*, *deadringer*; *gsc*, *goosecoid*; *otx*, *orthodenticle*; *alx*, *aristaless*; *tBr*, *t-brain*. See text for details.

the highest microarray signals for *Not* mRNA are at 15 h, but other reports based on in situ hybridization observations indicate that it first becomes detectable at the mesenchyme blastula stage (Peterson et al., 1999) between 21 and 24 h (http://sugp.caltech.edu/endomes/). The relative concentration of *msp130* mRNA is also higher than expected at 15 h, since Harkey et al. (1992) report detecting this mRNA by whole-mount hybridization just after hatching stage several hours later. In addition to these endomesodermal gene regulatory network genes, we have also examined many genes that are expressed in different ectodermal territories and verified that their temporal expression profiles are also in good agreement with expression patterns determined by standard molecular assays (data not shown).

In the process of validating expression profiles, we noticed a few cases in which microarray profiles were significantly different from published patterns. Further analysis showed that one or more of the individual probes showed some sequence similarity to closely related, but different, gene sequences in the Glean3 prediction set. Therefore, we examined all the microarray probe sequences for the potential to cross-react by examining whether closely related genes with different temporal patterns yielded composite profiles. For example, the temporal profiles for *SpAN* (SPU_004113) and the four *SpAN-like* genes (SPU_004114, 004115, 004116, 004117) are distinctly different, but the sequences are closely related (Angerer et al., 2006). When *SpAN* or *SpAN-like* probes were used as queries against the Glean3 sequences, perfectly matched 60-mer duplexes gave *E* values of e-27 while those for the most closely matched duplexes were e-22. Nevertheless, the temporal profile for the different *SpAN* probes is exactly as expected even though the *SpAN-like* signal intensities are high. Thus, the hybridization conditions are sufficiently stringent to eliminate this potential cross-reaction. Using the criterion that all closely related matches must have *E* values at least 5 orders of magnitude lower than that of a perfect match, we estimate that 93% of all probes monitor specific gene expression. The remaining 7% that may cross-react result primarily from the fact that microarray probes were derived from the most conserved sequences of a gene prediction, which, in a few cases, are shared by other proteins. To help investigators evaluate the reliability of individual probe sets, we have set up functions at our web site (http://urchin.nidcr.nih.gov/blast/exp.html) so that BLAST searches and determinations of the expression patterns for each can be executed rapidly.

### *Frequency of different types of gene expression profiles*

To determine the percentage of genes that are temporally regulated in the embryo, we imposed the conditions that modulations of signals among stages be greater than a factor of three and, in order to exclude randomly fluctuating background, the lowest value at any developmental stage be greater than 100 AU after 3× background subtraction (184 AU total, see Fig. 2). After this correction, very rare messages present at about 200–300 copies/embryo are still retained in the analysis. Using these criteria, we found that 66% of the genes are developmentally regulated between the early blastula and pluteus larva stages and that this value rises to 85% if maternal levels are also included. Similarly high fractions have been reported for genes expressed during 14 stages of *Drosophila* embryogenesis (75%) (Arbeitman et al., 2002) and during development of *C. elegans* from egg to adult (70%) (Jiang et al., 2001). The dominant quantitative regulation documented here is a corollary to the fact that most mRNAs in the sea urchin embryo are spatially regulated during development (Kingsley et al., 1993).

We also determined the fraction of putative mRNAs expressed in the embryo that are restricted to the maternal population (represented by two-cell polyA+ RNA) or that are

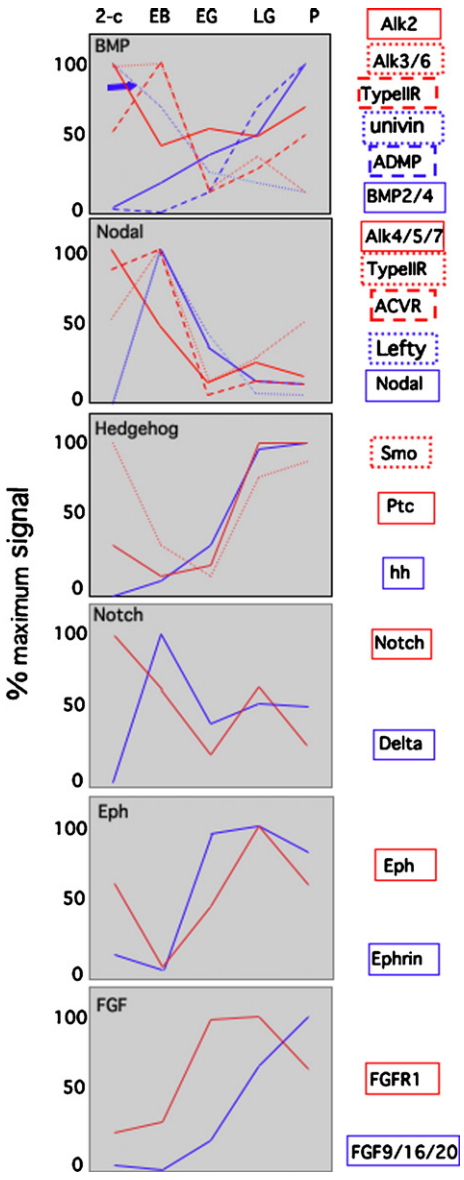


Fig. 7. Signaling pathways in development: genes encoding receptors are expressed before those encoding ligands. Shown are the temporal profiles for genes encoding ligands and receptors that are active during sea urchin development. All genes encoding ligands are indicated by blue lines and those for receptors by red. Different receptors or ligands within one pathway are indicated by dotted or dashed lines. BMP receptors include both Type I receptors, *Alk2* and *Alk3/6*, and the Type II receptor; the nodal Type I receptor is *Alk4/5/7*; the hedgehog (*hh*) receptors are *patched* (*Ptc*) and *smoothened* (*Smo*); gene IDs are given in the Supplemental information.

expressed only at later stages (15-h early blastula to 72-h pluteus). After subtracting 3× background, we found that only 5% are maternal-specific whereas 24% are strictly zygotic. The low fraction of maternal-only sequences is unexpected given that the complexity of egg RNA is twice that of gastrula RNA (Hough-Evans et al., 1977). It has been shown that the unfertilized oocyte harbors a poorly understood set of unusually long polyadenylated transcripts that are present at concentrations similar to those of authentic mRNAs in the complex class in the embryo but whose functions are unknown. These sequences are apparently not detected in significant numbers in the microarray experiment. Perhaps they lack conserved sequences or are not recovered efficiently because of short polyA tract lengths. Another possibility is that many maternal RNAs contain sequences also present in embryonic RNAs, but,

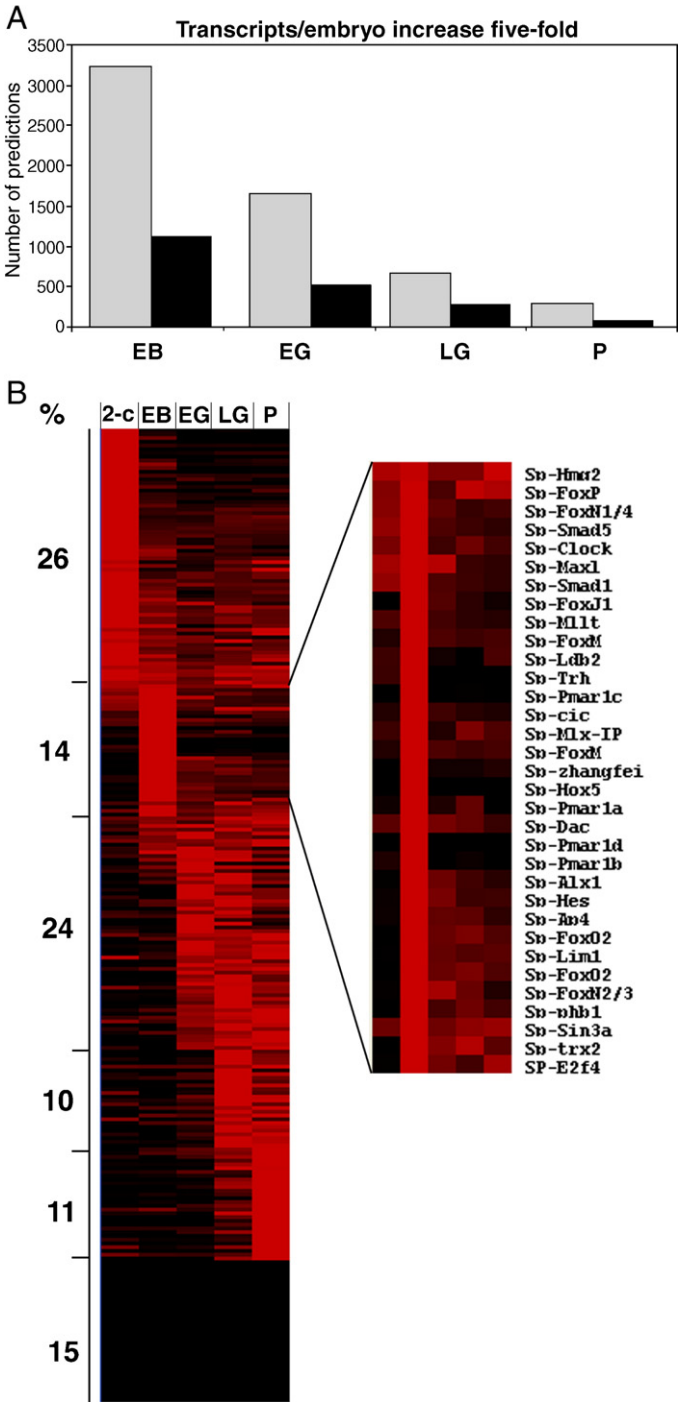


Fig. 8. (A) Each bar in the histogram represents the number of predictions at that stage that have fivefold higher signals than at all of the preceding stages (light gray) or those which are undetectable at earlier stages (black). (B) The spectrum of types of temporal expression profiles for 268 genes encoding transcription factors is shown. See Methods for details on the clustering analysis used. Expression levels were determined at the two-cell (2-c), early blastula (EB), early gastrula (EG), late gastrula (LG) and pluteus larva (P) stages.

unlike these RNAs, they have extra non-conserved sequences as a result of incomplete processing. Supporting this possibility are the observations that many different gastrula mRNAs have very high molecular weight counterparts in egg RNA (Kingsley et al., 1993) and that the structure of many egg RNAs resembles that of incompletely processed RNAs (for review, see Thomas et al., 1981).

The low fraction of maternal-specific transcripts reflects an interesting partitioning of the activities of genes involved in cell fate specification in this highly regulative embryo. Transcripts encoding most developmental gene regulatory factors are expressed only after fertilization; these include members of the endomesoderm, oral–aboral and left–right patterning networks. This is also the case for genes encoding signaling ligands, but not for the receptors to which they bind. This point is illustrated in Fig. 7, which shows a comparison of the temporal profiles for ligands and receptors in 6 signaling pathways known to be used during embryogenesis. Except for just one case (univin; top panel, arrow), the mRNAs encoding the ligands are not represented in the maternal population, but those encoding the receptors are. These results are consistent with the idea that the remarkable developmental plasticity of early blastomeres of the sea urchin embryo is because they are equipped to receive a large number of different signals. If this is the case, then specification and differentiation processes that are dependent on signaling pathways may be largely controlled by the production of spatially regulated signals.

We also estimated the fraction of genes that are activated at each of the 4 embryonic stages to support specification and differentiation of different cell types. The fractions of sequences that are either upregulated $\geq$5-fold (gray bars) or first detectable (black bars) at the indicated stages are shown in Fig. 8A. The largest number of different upregulated genes occurs at early blastula, which reflects not only activation of transcription but also the synthetic capacity of the increasing number of nuclei. Despite the increasing structural complexity and onset of cell-type-specific differentiation in embryos after the blastula stage, the number of newly upregulated genes declines substantially. The implication is that most of the proteins required for these processes begin to be produced far in advance of overt tissue differentiation and support the developmental programs of an increasing number of different cell types through their combinatorial activity.

How the gene regulatory capacity of the genome (the "regulome"; Consortium, in press) is used during embryonic development is of particular interest. We assembled temporal expression profiles for 268 genes encoding transcription factors that were annotated by members of the Sea Urchin Genome Sequencing Consortium [# genes in parentheses: M. Howard-Ashby (212), P. Oliveri (33), P. Martinez (22), S. Liang (6), L. Angerer (6), E. Chow (2), E. Arboleta (1), C. Flytzanis (1), P.Y. Lee (1)]. In order to compare the temporal patterns for all genes regardless of expression level, the five signal intensities for each gene were normalized to 100% maximum signal intensity and background corrections were applied as described above [after 3× background subtraction (84 AU), the maximum value at the five developmental stage must be greater than 100 AU]. Fig. 8B shows that 85% of the genes are expressed during embryogenesis between the two-cell and pluteus larva stages. This value is slightly higher than that reported for egg to the gastrula stage (Howard-Ashby et al., 2006) (80%), because it also includes genes expressed only at the pluteus stage. About a quarter of the genes are maximally represented in the maternal population. As development proceeds, the progressive deployment of new regulatory genes is strikingly distributed in successive waves of activation and repression; relatively few of these genes are uniformly represented throughout development. One of the critical points in development is the early blastula stage (EB) when initial specification of ectoderm, endoderm and mesoderm occurs. Some of the transcription factors involved in these processes are expressed primarily at this early time, and these are listed at the right. In some cases (e.g., the endomesoderm network) this clearly reflects the sequential activation of genes in a regulatory hierarchy; in others, it reflects differences in the time of specification and differentiation of different cell types (e.g., endomesoderm versus neural cell types). The initial clustering analysis of microarray-generated temporal profiles demonstrates that a wealth of information is now accessible on the readout of the sea urchin genome during sea urchin embryogenesis.

## Concluding remarks

We have developed this first array and database based on information from an early assembly of the *S. purpuratus* genome in order to make these resources available as soon as the annotation of the genome is complete. Further analysis using more recent versions of the assembly will permit refinement of the predictions to remove duplicates and false predictions. However, because each prediction can be tracked through successive refinements of the genome sequence, the arrays are usable now for analyzing expression patterns for most of the genes in the sea urchin genome. The database is searchable at http://urchin.nidcr.nih.gov/blast/exp.html. The complete set of probes is downloadable from the same web site.

## Acknowledgments

This work was supported by funds provided by the Intramural Research Program of the National Institute of Dental and Craniofacial Research.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ydbio.2006.08.034.

## References

Angerer, L., Angerer, R., 2003. Patterning the sea urchin embryo: gene regulatory networks, signaling pathways and cellular interactions. Curr. Top. Dev. Biol. 53, 159–198.

Angerer, L., Hussain, S., Wei, Z., Livingston, B., 2006. Sea urchin metalloproteases: a genomic survey of the BMP-1/tolloid-like, MMP and ADAM families. Dev. Biol. 300, 267–281.

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., White, K.P., 2002. Gene expression during the life cycle of *Drosophila melanogaster*. Science 297, 2270–2275.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94.

Burke, R., Angerer, L., Elphick, M., Humphrey, G., Yaguchi, S., Kiyama, T., Liang, S., Mu, X., Agca, C., Klein, W., Brandhorst, B., Rowe, M., Wilson, K., Churcher, A., Taylor, J., Chen, N., Murray, G., Wang, D., Mellot, D., Hallbook, F., Olinski, R., Thorndyke, M., 2006. A genomic view of the sea urchin nervous system. Dev. Biol. 300, 434–460.

Chudin, E., Walker, R., Kosaka, A., Wu, S.X., Rabert, D., Chang, T.K., Kreder, D.E., 2002. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. Genome Biol. 3 (research0005.1-0005.10).

Consortium, T.S.U.G.S. (in press). The genome of the sea urchin *Strongylocentrotus purpuratus*. Science.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Rust, A.G., Pan, Z., Schilstra, M.J., Clarke, P.J., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H., 2002. A genomic regulatory network for development. Science 295, 1669–1678.

Eisen, M., Spellman, P., Brown, P., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. 95, 14863–14868.

Galau, G.A., Britten, R.J., Davidson, E.H., 1974. A measurement of the sequence complexity of polysomal messenger RNA in sea urchin embryos. Cell 2, 9–20.

Guigo, R., Agarwal, P., Abril, J.F., Burset, M., Fickett, J.W., 2000. An assessment of gene prediction accuracy in large DNA sequences. Genome Res. 10, 1631–1642.

Harkey, M.A., Whiteley, H.R., Whiteley, A.H., 1992. Differential expression of the msp130 gene among skeletal lineage cells in the sea urchin embryo: a three dimensional in situ hybridization analysis. Mech. Dev. 37, 173–184.

Hough-Evans, B.R., Wold, B.J., Ernst, S.G., Britten, R.J., Davidson, E.H., 1977. Appearance and persistence of maternal RNA sequences in sea urchin development. Dev. Biol. 60, 258–277.

Howard-Ashby, M., Brown, C.T., Materna, S., Chen, L., 2006. Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus*. Dev. Biol. 300, 90–107.

Irizarry, R., Hobbs, B., Colllin, G., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., Kim, S.K., 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U. S. A. 98, 218–223.

Kingsley, P.D., Angerer, L.M., Angerer, R.C., 1993. Major temporal and spatial patterns of gene expression during differentiation of the sea urchin embryo. Dev. Biol. 155, 216–234.

Lee, P.Y., Davidson, E.H., 2004. Expression of Spgatae, the *Strongylocentrotus purpuratus* ortholog of vertebrate GATA4/5/6 factors. Gene Expression Patterns 5, 161–165.

Oliveri, P., Carrick, D.M., Davidson, E.H., 2002. A regulatory gene network that directs micromere specification in the sea urchin embryo. Dev. Biol. 246, 209–228.

Otim, O., Amore, G., Minokawa, T., McClay, D.R., Davidson, E.H., 2004. SpHnf6, a transcription factor that executes multiple functions in sea urchin embryogenesis. Dev. Biol. 273, 226–243.

Peterson, K.J., Harada, Y., Cameron, R.A., Davidson, E.H., 1999. Expression pattern of Brachyury and Not in the sea urchin: comparative implications for the origins of mesoderm in the basal deuterostomes. Dev. Biol. 207, 419–431.

Ransick, A., Rast, J.P., Minokawa, T., Calestani, C., Davidson, E.H., 2002. New early zygotic regulators expressed in endomesoderm of sea urchin embryos discovered by differential array hybridization. Dev. Biol. 246, 132–147.

Reynolds, S.D., Angerer, L.M., Palis, J., Nasir, A., Angerer, R.C., 1992. Early mRNAs, spatially restricted along the animal-vegetal axis of sea urchin embryos, include one encoding a protein related to tolloid and BMP-1. Development 114, 769–786.

Rouillard, J.-M., Zuker, M., Gulari, E., 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. 31, 3057–3062.

Samanta, M., Tongprasit, W., Istrail, S., Cameron, R., Tu, Q., Davidson, E., Stolc, V., submitted for publication. A high-resolution transcriptome map of the sea urchin embryo. Science.

Shepherd, G.W., Nemer, M., 1980. Developmental shifts in frequency distribution of polysomal mRNA and their posttranscriptional regulation in the sea urchin embryo. Proc. Natl. Acad. Sci. U. S. A. 77, 4653–4656.

Thomas, T.L., Posakony, J.W., Anderson, D.M., Britten, R.J., Davidson, E.H., 1981. Molecular structure of maternal RNA. Chromosoma 84, 319–335.

Wilt, F.H., 1977. The dynamics of maternal poly(A)-containing mRNA in fertilized sea urchin eggs. Cell 11, 673–681.

Yuh, C.H., Dorman, E.R., Davidson, E.H., 2005. Brn1/2/4, the predicted midgut regulator of the endo16 gene of the sea urchin embryo. Dev. Biol. 281, 286–298.

Zhang, L., Weinstock, G., Gibbs, R., Sodergren, E., submitted for publication. Annotation of the sea urchin genome. Science.