Development 126, 3857-3867 (1999)
Printed in Great Britain © The Company of Biologists Limited 1999
DEV5305



# EST analysis of gene expression in early cleavage-stage sea urchin embryos*

**Youn-Ho Lee[1,‡], Guyang Matthew Huang[2,§], R. Andrew Cameron[1], Geoffrey Graham[1,3], Eric H. Davidson[1,¶], Leroy Hood[2] and Roy J. Britten[1,3]**

[1]Division of Biology, California Institute of Technology, Pasadena, CA 91124, USA
[2]Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195, USA
[3]Kerckhoff Marine Laboratory, Corona del Mar, CA 92625, USA
*This work is dedicated to the memory of James G. Moore, who contributed greatly in its early stages
‡Present address: Korea Ocean Research and Development Institute, Polar Research Center, Seoul, Korea
§Present address: Pangea Systems, Oakland, CA 94612, USA
¶Author for correspondence (e-mail: davidson@mirsky.caltech.edu)

*Accepted 6 June; published on WWW 5 August 1999*

## SUMMARY

**A set of 956 expressed sequence tags derived from 7-hour (mid-cleavage) sea urchin embryos was analyzed to assess biosynthetic functions and to illuminate the structure of the message population at this stage. About a quarter of the expressed sequence tags represented repetitive sequence transcripts typical of early embryos, or ribosomal and mitochondrial RNAs, while a majority of the remainder contained significant open reading frames. A total of 232 sequences, including 153 different proteins, produced significant matches when compared against GenBank. The majority of these identified sequences represented 'housekeeping' proteins, i.e., cytoskeletal proteins, metabolic enzymes, transporters and proteins involved in cell division. The most interesting finds were components of signaling systems and transcription factors not previously reported in early sea urchin embryos, including components of Notch and TGF signal transduction pathways. As expected from earlier kinetic analyses of the embryo mRNA populations, no very prevalent protein-coding species were encountered; the most highly represented such sequences were cDNAs encoding cyclins A and B. The frequency of occurrence of all sequences within the database was used to construct a sequence prevalence distribution. The result, confirming earlier mRNA population analyses, indicated that the poly(A) RNA of the early embryo consists mainly of a very complex set of low-copy-number transcripts.**

Key words: Genome, Gene expression, mRNA, Transcript complexity, Sea urchin

## INTRODUCTION

The sea urchin embryo has proved particularly useful for studies of the regulatory mechanisms that underlie the early processes of embryogenesis. Recent studies have focused on spatial control of differential gene expression and on specification functions that depend on signaling between blastomeres. In sea urchin embryos, these processes begin early in cleavage (reviewed by Davidson et al., 1998). Here we describe an initial expressed sequence tag (EST) analysis of cleavage-stage (7-hour postfertilization) mRNA populations from *Strongylocentrotus purpuratus* embryos. This study was undertaken to provide a qualitative exploration, by a relatively unbiased method, of the nature of the biosynthetic program by then established. This is an interesting moment in early development: at 7 hours, the embryos are in 6th cleavage and all of the major early lineage compartments have segregated from one another. Blastomere specification processes are well underway, as evinced by the onset of regional programs of spatial gene expression. The embryo genomes are already operating at maximal rates of transcription, relative to any later stages. Essentially all maternal mRNA has been loaded onto the polysomes by this point, but newly synthesized zygotic messages are also being translated (reviewed by Davidson, 1986; Davidson et al., 1998).

The mRNA populations of *S. purpuratus* embryos are relatively well known. Early in development, there are on the order of $10^4$ different polysomal mRNA sequences, as established by mRNA excess hybridization against single-copy DNA tracers (reviewed by Davidson, 1986). Most of the mRNA (i.e., ~90% by mass) consists of relatively low prevalence transcripts, though there is a rising population of more highly represented early blastula-specific zygotic mRNA species (Reynolds et al., 1992), and also some other relatively prevalent mRNAs encoding proteins required for cell division, such as the cyclins (Evans et al., 1983; Pines and Hunt, 1987; Kelsowine-Miller et al., 1993). As expected, these classes of message are well represented in the EST database that we here report. The only mRNA species that are very highly prevalent in the 7-hour embryo encode zygotically expressed early histones, translation of which accounts for about 8% of total protein synthesis in this species at 7 hours postfertilization (Goustin, 1981). However,

the early histone messages are not polyadenylated, and would not be expected to be well represented in the cDNA library sampled in this study. Though the maternal 'cleavage-stage histone' mRNAs are polyadenylated, they are much more rare (reviewed by Davidson, 1986). There is a total of about $5\times10^7$ molecules of poly(A) RNA per embryo. If 90% of these consist of relatively low prevalence messages (Lasky et al., 1980; Flytzanis et al., 1982) and the complexity of the latter is about $10^4$ species, then each species will be represented by only about 0.01% of the mRNA molecules. Put another way, when the embryo has been divided up into several hundred cells at the late blastula stage, i.e., as it approaches its final cytoplasm-to-nucleus ratios, there are about $10^5$ mRNA molecules/cell, and the typical low prevalence RNA of the complex class of message will be present in the average population at less than ten copies per cell (see Davidson, 1986 for calculations). In reality there is of course a continuum of prevalence; many species exist at <3 copies per average cell; others at 3-10; others at 10-30, etc. (Lasky et al., 1980). In a $10^3$ EST database, transcripts present at 10 copies per average cell or less would be expected to be encountered only once if at all; or, if they are found twice or a few times, this would imply the order of $10^2$ copies per cell.

In the following, we consider the poly(A) RNA population of the 7-hour embryo, both qualitatively and quantitatively. Those EST sequences that produce significant matches with known protein-coding sequences in the GenBank database provide a qualitative snapshot of the biosynthetic functions in which the early embryo is engaged. We have also reconstructed the prevalence distribution of sequences in the embryonic poly(A) RNA population from that of the EST sequences. By this means, we confirm that the mRNA population of the early embryo consists mainly of a low prevalence, high complexity sequence set, as deduced earlier from its complexity and hybridization kinetics, its synthesis kinetics, and from measurements of cDNA hybridization to randomly selected cDNA clones (Galau et al., 1976, 1977; Flytzanis et al., 1982; Lasky et al., 1980; Duncan and Humphreys, 1981; Xin et al., 1982; Davidson, 1986).

## MATERIALS AND METHODS

### cDNA libraries

cDNA libraries were constructed in the Gibco P-Sport vector, following manufacturer's instructions, and using N6 random primers. The libraries were arrayed in 384-well plates in a Genetix Q-Bot robot. The ESTs determined in this work were obtained from eight different plates of a library made from 7-hour early cleavage-stage *S. purpuratus* embryos. mRNA was extracted from these embryos as described (Lee et al., 1986). The screening experiments described below were carried out on high-density filters prepared from the arrayed library using the 4×4 format described by Maier et al. (1994). On each 22×22 $cm^2$ filter, 18,432 clones are spotted in pairs oriented by a preassigned program loaded into the Q-Bot (Maier et al., 1994).

### EST sequencing

Clones were withdrawn from areas of the 384-well plates, the DNA prepared in an Autogen robot and their sequence obtained by conventional procedures in an automated ABI 377 sequencer, using dideoxy chain terminators or a T3 dye primer. The sequences were accessioned in GenBank under the identifiers: AF122056 to AF122818.

### Sequence searches

Each of the nucleotide sequences was translated in all six reading frames and the resulting six amino acid sequences were assembled into a single sequence which was used to search GenBank DNA sequences with TFASTA and TBLASTN. This is an effective method of identifying weak sequence similarities. We have required that the TFASTA and TBLASTN results agree for a sequence to be listed, except in marginal cases.

### Screening of high-density filters

High-density colony blots of arrayed cDNA libraries from 7-hour, 20-hour and 40-hour-old sea urchin embryos were utilized for this work. An oligonucleotide Sox probe was generated from the overlapping regions of two Sox EST clones: 5′-GCAAGAGGTTAG-GAGCCGAATGGAAGTTGCTTTCTG-3′. The probe was end-labeled with T4 polynucleotide kinase. The membranes bearing the cDNA libraries were pre-wet first with water and then with hybridization solution, and placed in hybridization bottles. Prehybridization (for 2 hours) and hybridization (for 16 hours) were carried out at 37°C in 6× SET, 5× Denhardt's solution, 50 mM PBS, pH 7.4, 0.25% SDS, 100 μg/ml sonicated denatured salmon sperm DNA. After hybridization, the membranes were washed for 10 minutes in 2× SSC, 0.2% SDS and for 15 minutes in 1× SSC, 0.2% SDS at room temperature and then for 30 minutes at 37°C in TMACl mix, which is 3 M tetramethylammonium chloride, 50 mM Tris (pH 8.0), 2 mM EDTA, 0.1% SDS. They were then washed for 30 minutes in TMAC1 mix at 45°C. The membranes were wrapped in plastic wrap and exposed to films for autoradiography. Positive spot pairs were identified by reference to the spotting template (cf. Maier et al., 1994). To obtain transcript prevalence estimates, the number of spot pairs reacting with probes for that transcript (N) was created. Relative prevalence was taken as P=N/(18,432F) where 18,432 is the number of spot pairs per filter, and F the number of filters used in the analysis. Absolute prevalence is P×T where T is the number of mRNAs for egg or embryo (see text).

## RESULTS AND DISCUSSION

### Overall distribution of sequence categories

The 956 ESTs that are the subject of this report were obtained from a directionally cloned, arrayed cDNA library prepared from 7-hour embryo poly(A) RNA. Insert lengths in this library lay mainly in the range 1.2-2.5 kb, and the average readable EST length on which the following analysis is based was about 500 nucleotides. These sequences were obtained using the 5′ vector primer with respect to insert orientation, and thus they preferentially sample the 5′ ends of the inserts. The library was directionally cloned from cDNA initiated on random primers to avoid a bias towards 3′ trailer sequences, which are generally very long on sea urchin embryonic transcripts. This strategy, plus the use of the 5′ sequencing primers, resulted in a large majority of ESTs containing either identified or putative protein-coding sequences, as we show below. Those ESTs that do not contain open reading frames (ORFs) derive either from 3′ trailer sequences of bona fide messenger RNAs, or from the interspersed repeat sequence class of poly(A) RNAs that constitute about 60% of the total cytoplasmic poly(A) RNA mass of *S. purpuratus* eggs and early embryos (Costantini et al., 1980; Davidson, 1986). These are long, non-translatable, apparently unprocessed

transcripts resembling nuclear pre-mRNA in structure and consisting of covalently linked single copy and interspersed repetitive sequence transcripts (Calzone et al., 1988). Most of the repetitive sequence elements recognized in the EST database are probably of this origin.

The translated sequences (see Materials and Methods) were compared to GenBank and significant matches collected. The initial EST set contained two classes of sequence which were subtracted from the database prior to analysis, namely 12 ESTs consisting of vector sequences and 33 sequences derived from an ETS class transcription factor mRNA. As described in footnote 1 of Table 1, the ETS sequences were artifactually over-represented because they were cloned at a natural *Not*I site, which vastly increases the probability of recovery relative to other sequences. After removal of the vector sequences and all but one ETS EST from the data set, there remained 956 sequences. Table 1 indicates the general categories, defined by these sequence similarity searches, into which the total set of ESTs are divided. There are three general categories: (1) recognized protein-coding sequences for which there is a high probability that the identifications are meaningful (see footnote 2 of Table 1), plus a few ESTs displaying significant similarities to unidentified ESTs of other organisms, together amounting to about 24% of the 956 ESTs; (2) sequences representing known classes of transcript other than mRNAs encoded in the nuclear genome, namely, mitochondrial RNAs, rRNAs and interspersed repeat containing poly(A) RNA transcripts, totaling about 25% of the 956 ESTs; and (3) sequences belonging to neither of the above classes, totaling about 51% of the 956 ESTs. These could represent unidentified mRNAs.

**Table 1. Summary of EST sequence categories***

| Category | No. of sequences | %‡ |
|---|---|---|
| Recognized protein-coding sequences§ | 215 | 22.5 |
| Similar to unidentified ESTs or cosmid sequences | 17 | 1.8 |
| Σ protein-coding sequences | 232 | 24.3 |
| No significant similarity to known sequences | 487 | 50.9 |
| Σ possible protein-coding sequences | 719 | 75.2 |
| Mitochondrial sequences | 36 | 3.8 |
| Ribosomal sequences | 136 | 14.2 |
| Interspersed repetitive sequence transcripts¶ | 65 | 6.8 |
| Σ other sequences | 237 | 24.8 |
| Total | 956 | 100 |

*A total of 1001 sequences were examined. 45 of these were rejected because they consisted of vector sequences or were sequences representing a portion of the mRNA for an ETS transcription factor. These sequences were present at an artificially high frequency due to occurrence of a natural *Not*I site within the ETS message. The *Not*I digestion step during preparation of the library endows these molecules with a much higher probability of cloning than have other cDNA molecules. Clones originating at endogenous insert *Not*I sites are revealed by the absence of linker nucleotides at the vector insert junction if the sequence covers that junction and this tell-tale feature was discovered in about half the ETS clones. However, only half of clones originating at an endogenous *Not*I site would be expected to be oriented so that this junction is at the 5′ end of the clone represented in the EST. Since their high frequency is the result of an artifact, all of the 34 ETS clones recovered were removed from the database (except for one, since ETS mRNA is certainly present, as was already known; Chen et al., 1988).

‡Per cent of total sequences (956), i.e., after removal of vector and ETS sequences (see * above).

§Probability of match to GenBank protein-coding sequence occurring by chance is $<10^{-6}$; see Table 2 for values.

¶Early embryo poly(A) RNA includes many apparently unprocessed, non-translatable RNAs containing interspersed repetitive sequences (reviewed by Davidson, 1986; Calzone et al., 1988) and repetitive sequences are also occasionally found in the 3′ trailer of bona fide mRNAs. The sequences included in this category display significant similarity to known repeat elements (by comparison with an unpublished catalogue of repeat sequences derived from the *S. purpuratus* Genome Project).

## EST identification by matches with GenBank

Table 2 lists all of the 215 significant matches discovered by comparing the EST sequences against GenBank (i.e., the initial row of Table 1). These matches are for the most part evidence of membership in the same family of proteins rather than specific identifications of particular proteins. Of course, where the echinoderm mRNA has earlier been sequenced, as with the cyclins, arylsulfatase, the metallothioneins, dynein, kinesin, Spec2A, actins, histones, fibropellin and some other well-studied proteins, the identifications are indeed exact. But in general they merely indicate the probable family of proteins to which belongs the sequence that the EST fragment encodes. Detailed examination of the matches leads to an important caveat. Only when the probability of chance occurrence is less than about $10^{-12}$ are the matches likely to identify long sequence overlaps, thus strongly implying membership in a given family of proteins. Probabilities of chance occurrence higher than this, down to our limit of $10^{-6}$, may only signify the presence of recognized protein motifs, which may or may not be shared amongst families of proteins other than that listed for the given entry in the Table.

The classification system that we have employed in Table 2 essentially divides the identified ESTs into four major categories. Category A can loosely be described as structural and enzymatic housekeeping proteins that are required in dividing, metabolizing cells. Category B consists of signaling and intercellular communication proteins. Category C includes transcription factors and other nuclear proteins that affect gene regulation, and category D consists of specialized products. In Table 3 the number of diverse mRNA species represented in each of these categories is listed, irrespective of their prevalence, that is, irrespective of the number of occurrences of each sequence match in Table 2. There are a total of 153 different proteins recognized. The majority of the diversity recovered (~60%) is in category A, consisting of housekeeping proteins of every variety. About 20% of all the different proteins are putatively involved in intercellular communication and/or signaling, and about 11% are nuclear proteins, probably utilized in the regulation of gene expression directly or indirectly. However, these values cannot be used to estimate the fraction of all the mRNA complexity devoted to these respective functional categories because the recognized sequence class of the EST database is biased towards more prevalent mRNAs compared either to the non-recognized ESTs (Table 1) or to the total mRNA population, as we discuss below. This is because current knowledge is itself biased toward more prominent proteins (i.e., more prevalent mRNAs), and because rare mRNAs, which constitute the vast majority of the total mRNA complexity, are grossly underrepresented in the EST database just because of its small size. However, the probability that a

**Table 2. EST sequence similarities, with grid position, gene description and probability of occurrence by chance**

A. Functions that many kinds of cells use

AI. Transportation and binding proteins for ions and other small molecules

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 2_A16 | 176 | RATBAND32E Rat band 3 $Cl^-/HCO_3^-$ exchanger | 1.6e-23 | F |
| 2_B15 | 135 | RATBAND32E Rat band 3 $Cl^-/HCO_3^-$ exchanger | 2.4e-20 | F |
| 6_H21 | 166 | RATBAND32E Rat band 3 $Cl^-/HCO_3^-$ exchanger | 1.4e-25 | B |
| 9_A08 | 171 | RATBAND32E Rat band 3 $Cl^-/HCO_3^-$ exchanger | 8.6e-36 | B |
| 2_G01 | 169 | AF011354 Fruit fly EF-hand calcium-binding protein | 1.6e-15 | B |
| 8_N16 | 146 | S70609 Human glycine transporter type 1b | 8.5e-35 | B |
| 8_P22 | 191 | MUSGT1A Mouse glucose transporter 1 | 1.0e-10* | B |
| 5_L24 | 105 | HSKCC Human K-Cl cotransporter | 7.3e-12* | F |
| 6_A09 | 274 | HSKCC Human K-Cl cotransporter | 1.0e-38 | B |
| 3_E18 | 222 | OAATPMR Sheep ($Na^+$, $K^+$) ATPase catalytic subunit alpha | 0 | F |
| 2_M22 | 276 | PFATP1 Plasmodium falciparum cation transporter | 1.8e-24 | F |

AII. RNA processing, polymerizing, splicing, and binding proteins and enzymes

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 3_K09 | 235 | MUSCIRPB Mouse cold-inducible RNA-binding protein | 1.1e-16 | B |
| 2_J13 | 172 | HSBAT1MR Human BAT1 nuclear RNA helicase (DEAD family) | 0 | F |
| 3_B23 | 147 | HSNP68M Human nuclear p68 protein | 1.3e-29 | B |
| 3_E12 | 210 | HSNP68M Human nuclear p68 protein | 4.3e-65 | B |
| 8_G10 | 136 | HSNP68M Human nuclear p68 protein | 1.5e-42 | B |
| 6_H23 | 168 | BTNDNAHII Cow nuclear DNA and RNA helicase II | 1.6e-57 | B |
| 5_M15 | 159 | CEU24123 Caenorhabditis elegans polyA-binding protein | 1.3e-33 | B |
| 3_C12 | 189 | HSRPIILS Human RNA polymerase II largest subunit | 4e-27 | F |
| 5_H01 | 174 | HUMRPOLAA Human RNA polymerase subunit hRPB 33 | 2.4e-40 | F |
| 5_B01 | 120 | HSU2AF Human large subunit of splicing factor U2AF | 5.8e-36 | B |
| 5_I23 | 330 | HUMSRP20 Human SR protein family | 8.2e-22 | B |
| 6_K13 | 226 | HUM9G8SF Human 9G8 splicing factor | 2.2e-27 | B |
| 9_A10 | 154 | HUMHCCA Human splicing factor (CC1.4) | 1.6e-47 | B |
| 6_H10 | 249 | HSHNRNPU Human U21.1 protein | 1.9e-17 | B |
| 4_F08 | 175 | BFZ83273 Branchiostoma floridae snRNP-like protein | 0 | F |
| 2_C18 | 107 | DMMUSASH Fruit fly musashi neural RNA-binding protein | 1.8e-08* | B |
| 8_N22 | 170 | HSPMRNACF Human pre-mRNA cleavage factor I | 0 | F |

AIII. Cell replication: Histones, cyclins and allied kinases, DNA polymerases topoisomerases, DNA modification

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 2_K14 | 129 | CFCDC42 Dog CDC42 GTP-binding protein | 6.8e-28 | B |
| 5_G21 | 184 | CFCDC42 Dog CDC42 GTP-binding protein | 4.2e-42 | B |
| 6_J16 | 186 | CFCDC42 Dog CDC42 GTP-binding protein | 0 | F |
| 5_N01 | 117 | XLU66558 Xenopus laevis cell division control, Cdc6 | 3.9e-26 | B |
| 2_D01 | 152 | AB008364 Hemicentrotus pulcherrimus cyclin A | 1.2e-61 | B |
| 2_P15 | 203 | AB008364 Hemicentrotus pulcherrimus cyclin A | 0 | F |
| 3_B24 | 162 | AB008364 Hemicentrotus pulcherrimus cyclin A | 0 | F |
| 3_J06 | 140 | AB008364 Hemicentrotus pulcherrimus cyclin A | 1.9e-50 | B |
| 4_D12 | 226 | AB008364 Hemicentrotus pulcherrimus cyclin A | 0 | F |
| 4_O11 | 334 | AB008364 Hemicentrotus pulcherrimus cyclin A | 0 | F |
| 5_O24 | 115 | AB008364 Hemicentrotus pulcherrimus cyclin A | 1.2e-40 | B |
| 6_C02 | 256 | AB008364 Hemicentrotus pulcherrimus cyclin A | 1e-27 | B |
| 8_G16 | 104 | AB008364 Hemicentrotus pulcherrimus cyclin A | 6.8e-39 | F |
| 9_B17 | 163 | AB008364 Hemicentrotus pulcherrimus cyclin A | 0 | F |
| 2_D02 | 122 | SGCYCLINB Sphaerechinus granularis cyclin B | 0 | F |
| 3_F19 | 141 | SGCYCBSPV Sphaerechinus granularis cyclin B | 1.7e-61 | B |
| 3_G18 | 147 | SGCYCLINB Sphaerechinus granularis cyclin B | 1.9e-36 | B |
| 5_A04 | 198 | SGCYCBSPV Sphaerechinus granularis cyclin B | 0 | F |
| 6_C21 | 168 | SGCYCBSPV Sphaerechinus granularis cyclin B | 0 | F |
| 6_P05 | 226 | SGCYCLINB Sphaerechinus granularis cyclin B | 2.9e-14 | F |
| 8_B20 | 305 | SGCYCBSPV Sphaerechinus granularis cyclin B | 0 | F |
| 9_D07 | 152 | SGCYCBSPV Sphaerechinus granularis cyclin B | 0 | F |
| 9_E06 | 128 | SGCYCBSPV Sphaerechinus granularis cyclin B | 1.1e-47 | B |
| 2_G21 | 177 | PMU84113 Psammechinus miliaris histone H1 (cleavage) | 1.1e-37 | B |
| 3_C11 | 140 | PMU84114 Psammechinus miliaris histone H2A (cleavage) | 9.5e-42 | B |
| 3_D12 | 84 | PMU84114 Psammechinus miliaris histone H2A (cleavage) | 4.2e-29 | B |
| 6_H20 | 46 | PMU84114 Psammechinus miliaris histone H2A (cleavage) | 2.1e-11* | F |
| 6_J04 | 133 | PMU84114 Psammechinus miliaris histone H2A (cleavage) | 1.1e-47 | B |
| 9_E07 | 87 | PMU84114 Psammechinus miliaris histone H2A (cleavage) | 6.6e-39 | B |
| 3_F15 | | SUSHISTONE S. purp histone H2A.F/Z | 6.6e-43 | B |
| 3_H24 | 70 | MUSH2AX1X Mouse histone H3.2 | 6.6e-21 | F |
| 5_H16 | 140 | MUSH2AX1X Mouse histone H3.2 | 0 | F |
| 6_A21 | 151 | PLH33H Paracentrotus lividus histone H3.3 | 0 | F |
| 2_H03 | 140 | CMHIST34 Duck H4 histone gene | 8.7e-46 | B |

**Table 2. Continued**

| | | | | |
|---|---|---|---|---|
| 2_H06 | 58 | PMHISH4 Psammechinus miliaris histone H4 | 9.6e-07* | F |
| 6_H01 | 280 | HSLON Human Lon protease-like protein, ATP-dependent | 0 | F |
| 3_D21 | 86 | XLRANBP1 Xenopus laevis Ran binding protein 1 | 7.6e-22 | F |
| 3_K06 | 127 | AF044588 Human spindle protein regulating cytokinesis | 1.2e-16 | B |
| 8_J17 | 324 | MMU28168 Mouse GP106 adenomatous polyposis locus homolog | 0 | F |
| 8_E15 | 146 | AF045581 Human BRCA1 associated protein | 1.0e-26 | B |
| AIV. Cytoskeleton and membrane proteins | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 6_D04 | 129 | GGACTL Chicken actin-like protein | 0 | F |
| 5_D18 | 146 | DMLETHAL Drosophila alpha actinin | 5.8e-38 | B |
| 3_C21 | 156 | SPU47278 S. purp axonemal dynein light chain p33 | 0 | F |
| 8_K16 | 153 | ANDDIC3 Sea urchin dynein intermediate chain 3 (IC3) | 0 | F |
| 3_E09 | 254 | HUMHMP2 Human motor protein | 8.1e-35 | F |
| 6_C15 | 146 | RNU60416 Rat myr6 myosin | 1.3e-20 | F |
| 3_D15 | 105 | S59344 Xenopus nuclear pore complex glycoprotein | 2.7e-38 | B |
| 2_C04 | 115 | HSU95735 Human SNARE protein Ykt6 | 9.1e-16 | F |
| 5_H15 | 202 | SPU38523 S. purp tektin | 2.2e-30 | B |
| 2_F04 | 115 | GLU92645 Gecarcinus lateralis alpha-1-tubulin | 5.9e-47 | B |
| 4_H12 | 333 | CRUTUBAA Chinese hamster alpha-tubulin I | 0 | F |
| 6_G02 | 150 | CRUTUBAC Chinese hamster alpha-tubulin II | 2.9e-41 | B |
| 8_M16 | 167 | PVATUB2 Patella vulgata alpha-2 tubulin | 2.3e-13 | F |
| 6_L16 | 339 | AF022655 Human cep250 centrosome | 2.5e-09* | F |
| AV. Protein synthesis cofactors, tRNA synthetases, ribosomal proteins | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 3_B19 | 169 | CELEFT2A Caenorhabditis elegans elongation factor 2 | 0 | F |
| 8_J18 | 137 | CELEFT2A Caenorhabditis elegans elongation factor 2 | 4.4e-23 | B |
| 8_N21 | 229 | CELEFT2A Caenorhabditis elegans elongation factor 2 | 9.9e-66 | B |
| 4_G08 | 173 | AF012088 Human initiation factor eIF4G1 | 3.6e-24 | B |
| 6_G23 | 85 | HSU76111 Human translation repressor NAT1 | 1.8e-12* | B |
| 9_B15 | 194 | HUMSUIISO Human sui1iso1 initiation factor | 2.2e-32 | F |
| 3_J02 | 216 | TGU02371 Tripneustes gratilla laminin binding protein | 0 | F |
| 5_I13 | 214 | TGU02371 Tripneustes gratilla laminin binding protein | 4.9e-45 | B |
| 5_I02 | 301 | HSSTRNAS Human seryl-tRNA synthetase | 0 | F |
| 9_B07 | 200 | CLRRS1 Long-tailed hamster arginyl-tRNA synthetase | 0 | F |
| AVI. Intermediary synthesis and catabolism enzymes | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 6_K04 | 205 | S81092 Mouse acyl-coenzyme A:cholesterol acyltransferase | 5.1e-11* | F |
| 9_D17 | 124 | HSRNALAGA Human L-arginine: glycine amidinotransferase | 0 | F |
| 9_D19 | 124 | HSRNALAGA Human L-arginine: glycine amidinotransferase | 0 | F |
| 9_A11 | 122 | SQUCARPSYN Dogfish shark carbamyl phosphate synthetase | 3.2e-44 | B |
| 6_L11 | 235 | GDCARANH Chicken carbonic anhydrase | 1.8e-40 | F |
| 9_D12 | 87 | AF028609 Pig LCHYD-HAD precursor | 8.3e-24 | F |
| 6_H22 | 183 | MMU87147 Mouse flavin-containing monooxygenase 3 | 2.7e-26 | F |
| 5_M03 | 164 | HSETFBS Human electron transfer flavoprotein beta subunit | 7e-25 | F |
| 4_A08 | 175 | HUMMDMCSF Human trifunctional folate enzyme | 0 | F |
| 5_D15 | 189 | PALGLUSYN Paracentrotus lividus glutamine synthetase | 9.6e-43 | B |
| 6_A22 | 197 | PALGLUSYN Paracentrotus lividus glutamine synthetase | 0 | F |
| 6_F01 | 151 | PALGLUSYN Paracentrotus lividus glutamine synthetase | 3.4e-21 | B |
| 8_C13 | 244 | PALGLUSYN Paracentrotus lividus glutamine synthetase | 0 | F |
| 2_E09 | 133 | CHKPECM Chicken phosphoenolpyruvate carboxykinase | 1.7e-31 | B |
| 5_C08 | 105 | CHKPECM Chicken phosphoenolpyruvate carboxykinase | 1.6e-38 | F |
| 3_E06 | 128 | HSPNP Human purine nucleoside phosphorylase | 6.8e-22 | F |
| 5_B22 | 174 | SSP41 Clam ribonucleotide reductase small subunit | 2.7e-10* | F |
| 6_B16 | 77 | SSP41 Clam ribonucleotide reductase small subunit | 5.8e-09* | F |
| 3_B16 | 130 | XELAHH Xenopus laevis adenine homocysteine hydrolase | 0 | F |
| 4_N08 | 174 | XELAHH Xenopus laevis adenine homocysteine hydrolase | 0 | F |
| 5_G20 | 211 | DMAHCYGEN Fruit fly S-adenosyl-L-homocysteine hydrolase | 0 | F |
| 6_C23 | 217 | RATA26S Rat alpha 2,6-sialyltransferase | 3.9e-22 | F |
| 2_C16 | 105 | MMU34883 Mouse ATP sulfurylase | 9.5e-47 | B |
| 2_F15 | 113 | MMU34883 Mouse ATP sulfurylase | 6.6e-34 | F |
| 4_N12 | 84 | S81373 Rat sulfated glycoprotein 1 | 1.6e-07* | B |
| 2_F08 | 116 | MUSTHSM Mouse thymidylate synthase | 0 | F |
| AVII. Stress response, detoxification and cell defense proteins | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 3_N07 | 168 | D83971 Cowpea CPRD14 drought resistance protein | 5.1e-24 | F |
| 3_A21 | 189 | RNU18729 Rat cytochrome b558 alpha subunit | 1.3e-27 | B |
| 8_F14 | 208 | HUMFERC Human ferrochelatase | 0 | F |

*Table 2 continued overleaf*

**Table 2. Continued**

| | | | | |
|---|---|---|---|---|
| 2_G15 | 127 | BRPHSP70 Brugia pahangi heat shock protein 70 | 1.4e-35 | F |
| 2_J15 | 88 | SUSMETA S.purp metallothionein | 0 | F |
| 4_M12 | 131 | SUSMETA S.purp metallothionein | 0 | F |
| 9_E02 | 102 | GGCYP1A5 Chicken cytochrome P450 1A | 1.2e-09* | F |
| 9_F15 | | BOVCYPC21A Cow cytochrome P450-c21 | 1.9e-18 | F |
| AVIII. Protein degradation and processing, proteases | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 6_H14 | 276 | BTCATHEPL Cow cathepsin L protease | 2.2e-23 | F |
| 2_C13 | 170 | DMU60591 Fruit fly kuzbanian metalloprotease | 2.1e-20 | B |
| 6_B15 | 236 | RNU50194 Rat tripeptidylpeptidase II | 0 | F |
| 2_F20 | 195 | HSU39318 Human E2 ubiquitin conjugating enzyme | 0 | F |
| 3_B21 | 145 | AF032456 Human ubiquitin conjugating enzyme | 1.3e-67 | B |
| AIX. Apoptosis-related | | | | |
| 6_L09 | 359 | AB008449 Bombyx mori BmP109, Bcl-2 family | 3.5e-11* | B |
| 8_G17 | 167 | MMU35846 Mouse AAC-11 apoptosis inhibitor | 5.3e-09* | B |
| B. Cell-cell communication | | | | |
| BI. Signaling receptors, including cytokine and hormone receptors, and signaling ligands | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 2_D03 | 66 | S78731 Malacosoma disstria hormone receptor 2 | 5.4e-07* | F |
| 3_C09 | 249 | HSU41745 Human PDGF-associated protein. | 7.6e-15 | F |
| 3_I01 | 214 | HSHPS12 Human pHS1-2 receptor homolog | 7e-14 | F |
| 4_E11 | 322 | CVTRP Calliphora transient receptor potential protein | 1e-17 | F |
| 2_B09 | 115 | AF034606 Danio rerio (zebrafish) chordin | 9.4e-08* | B |
| 2_K13 | 189 | AF034606 Danio rerio (zebrafish) chordin | 1.6e-07* | B |
| 2_A09 | 141 | AF027208 Human AC133 5-transmembrane type receptor | 2.3e-13 | F |
| 2_H02 | 65 | AB001106 Human glia maturation factor gamma | 1.9e-09* | B |
| 3_G12 | 243 | XLU77640 Xenopus laevis lunatic fringe signaling protein | 4.7e-17 | F |
| 6_A20 | 150 | D50646 Mouse stromal cell-derived factor-2, secreted | 1.6e-29 | F |
| 9_F17 | 177 | CGU48852 Hamster HT protein | 1.3e-13 | F |
| BII. Intracellular signal transduction pathway molecules including kinases and signal intermediates such as beta-catenin | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 6_I08 | 250 | DROCNO Fruit fly canoe protein | 5.8e-09* | B |
| 5_K14 | 110 | AF042862 Chicken casein kinase 1 | 2.7e-24 | B |
| 2_E02 | 115 | SUTCATN Hawaiian sea urchin beta-catenin | 2.7e-53 | B |
| 2_G08 | | HS21GARP Human 21-glutamic acid-rich protein | 1.9e-10* | B |
| 2_G09 | | HS21GARP Human 21-glutamic acid-rich protein | 9.6e-10* | B |
| 4_I11 | | HS21GARP Human 21-glutamic acid-rich protein | 1.6e-08* | B |
| 3_E19 | 109 | DMGTPBP Fruit fly GTP-binding protein | 3.7e-28 | B |
| 4_B12 | 116 | APGPASMR Starfish G protein (alpha subunit) | 0 | F |
| 6_B10 | 106 | APGPASMR Starfish G protein (alpha subunit) | 1.1e-19 | F |
| 6_J09 | 307 | APGPASMR starfish G protein (alpha subunit) | 3.8e-31 | F |
| 2_B08 | 157 | RNGSK3B Rat glycogen synthase kinase 3 β | 0 | F |
| 2_D13 | 178 | RNGSK3B Rat glycogen synthase kinase 3 β | 0 | F |
| 8_B13 | 187 | DMMGN Drosophila melanogaster mago-nashi protein | 0 | F |
| 5_C13 | 260 | LPU02967 Lytechinus pictus protein kinase C | 0 | F |
| 6_L14 | 165 | S55223 Rat 14-3-3 beta type, pKC regulator | 9.1e-09* | F |
| 6_L15 | 189 | HSU42390 Human Trio multidomain protein | 1.7e-34 | B |
| BIII. Extracellular matrix proteins and cell adhesion, e.g., integrins and integrin receptors, and cadherins | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 2_H08 | 116 | SUSEGFI S. purp fibropellin Ia | 3.0e-43 | B |
| 3_B22 | 173 | SUSEGFI S. purp fibropellin Ia | 0 | F |
| 3_D20 | 83 | SUSEGFI S. purp fibropellin Ia | 7.8e-65 | B |
| 3_F23 | 54 | SUSEGFI S. purp fibropellin Ia | 1.1e-34 | B |
| 9_C06 | 107 | SUSEGFI S. purp fibropellin Ia | 0 | F |
| 5_M17 | 228 | SUSEGFI S. purp fibropellin Ia | 2.0e-11 | B |
| 8_A14 | 163 | SUSEGFI S. purp fibropellin Ia | 3.6e-54 | B |
| 8_P20 | 174 | SUSEGFI S. purp fibropellin Ia | 0 | F |
| 9_B16 | 200 | MMRNAASFA Mouse arylsulfatase A | 1e-39 | B |
| 6_M05 | 220 | HSGLYPIC Human heparan sulfate proteaglycan | 1.8e-13 | F |
| 9_C08 | 216 | LVU40065 Lytechinus variegatus extracellular matrix | 1.5e-20 | B |
| 6_E20 | 135 | S73803 HLC-32 hyaline layer component | 3.0e-33 | B |
| 8_H15 | 245 | MSLAMA Mouse laminin A, C-terminal fragment | 4.8e-14 | F |
| 8_L21 | 101 | SPU65432 S. purp laminin alpha chain | 7.8e-13 | B |
| 8_B12 | 267 | HS5T4OA Human 5T4 Oncofetal antigen | 3e-10* | F |
| 4_D11 | 246 | DMMAS1OV Fruit fly alpha 1,2 mannosidase | 0 | F |

**Table 2. Continued**

C. Transcription factors and other gene regulatory proteins
CI. Sequence-specific DNA-binding proteins

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 6_A23 | 168 | HSU25435 Human transcriptional repressor, CTCF | 5.1e-18 | B |
| 5_F24 | 116 | SUSETS S. purp ETS | 2.6e-107 | B |
| 2_O14 | 116 | XELXMAD Xenopus laevis Mad2 protein | 8.7e-11* | F |
| 9_F12 | 159 | AF016886 S. purp paired box protein (suPaxB) | 0 | F |
| 2_B06 | 135 | AF040250 Human DNA binding protein, PO-GA | 4.2e-26 | B |
| 1_A01 | 134 | LFSOXLF2 L.fuscus sox protein | 9.1e-23 | F |
| 8_L16 | 136 | LFSOXLF2 L.fuscus sox protein | 4.9e-14 | F |
| 5_J02 | 161 | RSU08214 Rat DNA binding protein URE-B1 | 3.1e-16 | F |
| 6_J06 | | SPU38281 S. purp orphan steroid hormone receptor | 6.7e-17 | F |

CII. Non-DNA binding proteins that perform positive or negative roles, e.g., EIA or CBP

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 6_L09 | 359 | AB008449 Bombyx mori BmP109, Bcl-2 family | 3.5e-11* | B |
| 2_F10 | 207 | DMU19269 Fruit fly Dachshund protein | 4.4e-55 | B |
| 2_E10 | 110 | RNU83883 Rattus norvegicus p105 coactivator | 1.9e-21 | B |
| 6_D10 | 130 | HSU22055 Human 100 kDa coactivator | 1.9e-30 | F |
| 6_J14 | 124 | HSMI2218 Human 218 kDa Mi-2 presumed helicase | 2.2e-52 | B |
| 8_J10 | 273 | MUSPVZ3A Mouse mSin3A corepressor | 7.4e-20 | B |
| 9_D08 | 172 | HSU75308 Human TBP-associated factor (hTAFII130) | 2.4e-38 | B |
| 9_C15 | 122 | AF016270 Human thyroid hormone receptor coactivator | 2.9e-22 | B |

Column 1, clone position, i.e., plate and well number in frozen arrayed library (384-well plates)

Column 2, length in amino acids of open reading frame (ORF) matching GenBank entry. Where there is no entry the ORF cannot be determined from the GenBank data.

Column 3, GenBank entry name and gene description. Where searches led to similar sequences but from different animals only one name is used in this Table.

Column 4, Probability ($P$) of occurrence of similarity by chance. Matches that are clearly secure are those in which $P \leq 10^{-12}$; matches in which $P > 10^{-6}$ are considered likely to be meaningless; matches for which $10^{-6} > P > 10^{-12}$ are possibly meaningful and are marked by an asterisk.

Column 5, Probabilities listed in column 5 are from B, TBLASTIN search; or F, TFASTA search result. The lowest probability of accidental match was chosen.

given prevalent mRNA species will appear in the 956 clone database is much higher. Thus only some very general conclusions can be drawn. It does seem clear that, at least in the moderate prevalence class of 7-hour embryo mRNAs, a large fraction of species will belong to category A, consisting of messages encoding transporters, cytoskeletal proteins, cell division proteins, enzymatic machinery, etc., and this will probably be true of the great mass of the yet unknown rare messages in the early embryo as well since many of the category A transcripts occur only once in the database and thus are probably of low abundance. Though long suspected, this observation provides the first relevant and direct evidence for the high fraction of early embryo RNAs encoding housekeeping machinery.

Among the housekeeping proteins of category A are some potentially interesting finds. Examples include a considerable number of ion transporters (subcategory AI); RNA splicing factors, RNA polymerase and RNA mobilizing enzymes (AII); a protein strongly related to a spindle protein that regulates cytokinesis (AIII); a nuclear pore complex protein and a centrosomal protein (AIV); and two proteins that in other systems control apoptosis (AIX).

For the present authors, as for most developmental biologists, the most interesting categories are (B) signaling and (C) transcription control, and the most useful aspect of any EST project is the new probes for interesting genes that it affords. Several such discoveries are included in parts B and C of Table 2, i.e., sequences not previously isolated from echinoderm material. In category B, these include sequences related to chordin and lunatic fringe, a secreted activator of the Notch signaling receptor (Johnston et al., 1997) (subcategory BI), and various G proteins and casein kinase. In category C, we found sequences similar to five different transcription factors to our knowledge not previously recovered from these embryos, including factors of the Sox and MAD families. All of these finds have new implications for the functional activities of the 7-hour embryo. For example, though mRNAs encoding two putative ligands of TGF-β family have been reported, namely univin (Stenzel et al., 1994) and an orthologue of human BMP5-8 (Ponce et al., 1999), the presence of chordin and of a MAD class transcription factor suggests that signaling mediated by one or the other of these ligands is occurring or will soon occur in cleavage-stage embryos, and that it could be involved in the early specification functions. Similarly, though the early embryo contains maternal Notch mRNA and protein (Sherwood and McClay, 1997), the presence of a fringe family mRNA at 7 hours postfertilization suggests a current regulation of this pathway, for which there is yet no role assigned in the cleavage-stage embryo. In addition, we note that one of the unsolved problems in the regulatory molecular biology of the early sea urchin embryo is the mechanism by which maternal transcription factors are modified so that they become active in the appropriate embryonic territories (Davidson et al., 1998).

In quantitative terms, this is a very small EST project, and of all the sequences obtained only about a quarter generated significant matches against the GenBank database. Prima facie, it is remarkable that the recognized sequences include such interesting examples and provide such potentially useful probes.

## ESTs displaying no significant similarity to known genes

About half the total EST set displays no significant match

**Table 3. The number of different genes expressed by class**

| Class | | No. |
|---|---|---|
| | **A. Functions that many kinds of cells use** | |
| I. | Transportation and binding proteins for ions and other small molecules | 8 |
| II. | RNA processing, polymerizing, splicing and binding proteins and enzymes | 15 |
| III. | Cell replication, histones, cyclins and allied kinases, DNA polymerases, topoisomerases, DNA modification | 16 |
| IV. | Cytoskeleton and membrane proteins | 14 |
| V. | Protein synthesis cofactors, tRNA synthetases, ribosomal proteins | 7 |
| VI. | Intermediary synthesis and catabolism enzymes | 18 |
| VII. | Stress response, detoxification and cell defense proteins | 7 |
| VIII. | Protein degradation and processing, proteases | 5 |
| IX. | Apoptosis-related proteins | 2 |
| | **Total** | **92** |
| | **B. Cell-cell communication** | |
| I. | Signaling receptors, including cytokine and hormone receptors, and signaling ligands | 10 |
| II. | Intracellular signal transduction pathway molecules including kinases and signal intermediates such as beta-catenin | 11 |
| III. | Extracellular matrix proteins and cell adhesion, e.g., integrins and integrin receptors, and cadherins | 8 |
| | **Total** | **29** |
| | **C. Transcription factors and other gene regulatory proteins** | |
| I. | Sequence-specific DNA-binding proteins | 8 |
| II. | Non-DNA binding proteins that perform positive or negative roles, e.g., EIA or CBP | 7 |
| III. | Chromatin proteins other than AIII with regulatory function, e.g., polycomb class proteins | 3 |
| | **Total** | **18** |
| | **D. Specialized terminal differentiation products** | |
| I. | Secreted | 2 |
| II. | Intracellular | 4 |
| | **Total** | **6** |
| | **E. Not enough information to classify** | **8** |
| | **Total number of different proteins recognized** | **153** |

with any sequence in GenBank (Table 1). In order to determine whether these clones are likely to represent the non-translatable poly(A) RNAs of the early embryo, or in contrast, are bona fide mRNAs too divergent for recognition using the (mainly mammalian) GenBank database, we carried out a statistical search for open reading frames (ORFs). This was done by determining the distribution of lengths between stop codons, compared to the lengths that occur in randomly generated sequences of a matched sequence length distribution. Results are shown in Fig. 1; details can be found in the legend. Briefly, the observed distribution of ORF lengths in the unidentified ESTs was compared with that expected to occur on a chance basis in a set of sequences of the length distribution of these ESTs. Fig. 1 shows clearly that the ORF length distribution observed (histogram) cannot be accounted for on a random basis. Most of the unidentified ESTs in the analysis thus contain protein-coding sequences, i.e., by our estimate, 65-80%. It follows that if we consider all the possible protein-coding sequences, in the identified plus unidentified clones (75.2% of the total in Table 1), 73-84% of these indeed include codogenic mRNA sequence. The remainder consist of 3′ trailer sequences of mRNAs, or are interspersed repeat-containing transcripts, probably mainly the latter. On this basis, the total number of ESTs representing the interspersed repeat transcript class would be the sum of those in which repeats were recognized (65 sequences; Table 1) plus the non-coding sequences inferred from the analyses of Fig. 1. The fraction of the whole library (as sampled in the ESTs) representing interspersed transcripts would thus lie in the range of 25-40%. This is consistent with the estimate that, in eggs and early embryos, this class of poly(A) RNA by mass constitutes over 50% of the total (Costantini et al., 1980), if the probability of reverse transcription is about the same per molecule, since the interspersed poly(A) RNAs are on the average at least 5× as long as are the mRNAs.

**Table 4. Sox mRNA prevalence during development**

| Stage | No. positive clones/ no. filters screened* | Fraction | Estimated molecules/embryo‡ |
|---|---|---|---|
| 7 hour | 41/3 | $7.4\times10^{-4}$ | $3.7\times10^{4}$ |
| 20 hour | 16/5 | $1.7\times10^{-4}$ | $8.5\times10^{3}$ |
| 40 hour | 68/3 | $12.3\times10^{-4}$ | $6.2\times10^{4}$ |

*Each 22×22 cm2 filter contains 18,432 clones.

‡At the 500-cell stage an ‘average cell’ or cell equivalent contains about $10^5$ mRNAs; i.e., the whole embryo has about $5\times10^7$ mRNAs throughout embryogenesis. The embryo has this many cells in the late blastula-early gastrula stage. For comparative purposes, the number of Sox mRNA molecules per average cell is thus about 120 at this stage.

## Prevalence distribution

As discussed above, the more prevalent mRNAs are expected to be overrepresented in the category of ESTs recognized by sequence comparison with GenBank and, by the same token, they will be underrepresented in the unrecognized category. Thus to obtain a balanced image of sequence prevalence distribution in the early embryo poly(A) RNA from the EST database, all EST sequences excluding ribosomal, mitochondrial and repetitive sequences were compared against one another to detect multiple occurrences, and the results pooled with those shown in Table 2 for the identified

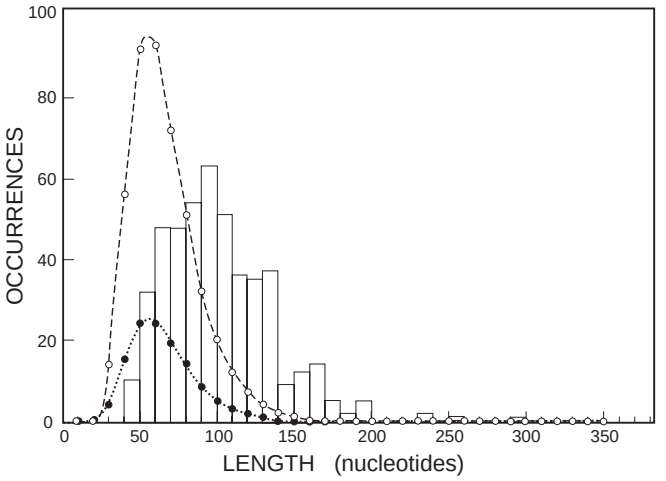


**Fig. 1.** Statistical estimation of ORFs in sequences without significant GenBank matches. The observed length distribution of ORFs in ESTs without significant GenBank matches is compared with that expected on the basis of chance for sequences of the same length and composition. All sequences with matches to coding regions for proteins, ribosomal RNAs, mitochondrial RNAs and known repeated sequences, were removed from the list. Undetermined nucleotide positions in the sequences (N's) were deleted and the 19 sequences containing the largest number of N's were not considered. The remaining 468 sequences were translated and the locations of stop codons determined. The maximum ORF lengths were then identified. The maximum lengths were classified into 10-nucleotide bins (abscissa) and the number of sequences in each bin is indicated as a histogram. The taller curve, which has the same total area as the histogram, shows the ORF length distribution expected by chance. This distribution was calculated individually by making a random sequence of a length and composition that matches each EST, and then determining the longest distance in this random sequence between stop codons or between a stop codon and a terminus. To obtain the curve shown, this calculation was repeated 100 times to reduce statistical fluctuation, these randomly occurring ORF lengths were placed in 10-nucleotide bins and a curve was drawn through the points delimiting the lower bound of these bins. To estimate the fraction of the observed ORFs that is likely to be due to randomly occurring ORFs, we reduced the amplitude of the curve describing the random calculation so that it matches as closely as possible that portion of the histogram likely to be due exclusively to randomly occurring ORFs (i.e., 40-60 nucleotide ORFs). This is shown by the lower curve, which includes 27% of the total observed ORF distribution.

ESTs. The summed prevalence distribution is shown in Fig. 2. As expected, no sequence occurred with the frequency of cyclins A and B (i.e., 10 and 9 occurrences, respectively) in the unrecognized EST set and, in fact, the maximum multiplicity was three occurrences of one of the unidentified sequences. Fig. 2 presents the prevalence distribution in both number and mass terms. Respectively, the number representations give the frequency of molecules occurring at each prevalence as percent of total, and the mass representations give the frequency of total mass represented by the sum of molecules at each prevalence. These distributions can be compared with earlier number (e.g., Flytzanis et al., 1982) and mass (e.g., Lasky et al., 1980) distributions for *S. purpuratus* embryo poly(A) RNA, as deduced from hybridization of large random sets of cDNA clones with labeled cDNA. The present results led to remarkably similar conclusions. They confirm that the very large majority of embryo poly(A) RNAs are of the rare sequence class, represented here by molecular species appearing only once in the EST database (i.e., ~80% of all ESTs included in Fig. 2). As noted above, because of the small size of this database we cannot estimate the actual frequency of these rare mRNAs, except that it is significantly <100 per average cell equivalent (i.e., at a 500-cell stage where there are about $10^5$ mRNA molecules/cell). As the lower abscissa of Fig. 2 shows, for more prevalent mRNAs, the expected frequencies are on the order of several hundred mRNA molecules per average cell. All told, the prevalent transcripts, here represented as those occurring more than once, constitute about 12% of the total sum of poly(A) RNA sequences in the analysis.

## Tracking prevalence changes

The ESTs considered here represent an early stage of embryonic development and it is often important to examine change in representation of given transcripts as embryogenesis proceeds. Availability of a comparable set of arrayed library grids in a high-density filter format renders this an easy measurement. As one example, we take the Sox class transcription factor mRNAs uncovered in this project. Two Sox mRNAs were identified, namely those in plate 4 position B08 and plate 8 position L16 (Table 2). A probe was designed from the overlapping region of these clones and used to screen (1) several filters each containing 18,432 randomly selected clones from the 7-hour library that was the source of the ESTs, (2) equivalent filters from a 20-hour mesenchyme blastula stage library, and (3) filters from a 40-hour late gastrula stage library. As Table 4 shows, Sox mRNAs are probably present as modestly prevalent maternal mRNAs ($3.7\times10^4$/egg). A majority of these have disappeared by the mesenchyme blastula stage, but the prevalence again increases by late gastrula, so that the prevalence per average cell increases from about 70 to about 120 molecules, undoubtedly as result of zygotic transcription. This pattern is fairly typical, as found earlier for a large set of unidentified messages (Flytzanis et al., 1982). Note that these values confirm the inference in the lower abscissa of Fig. 1, i.e., that linear extrapolation of prevalence in the EST data set can be used to provide a thumbnail estimate of prevalence for the whole embryo (or for the average cell), for sequences that occur more than once. This approach to prevalence determination, which does not depend at all on accurate measurement of the amounts of probe hybridized to given clones, is not only quick and easy but is also relatively robust. That is, since the library is arrayed, and large amounts of plasmid DNA are present in each spot, every spot pair represented by a given probe will hybridize similarly every time the array is screened, in contrast to λ plaque screening, in which under the usual conditions plaque size for a given recombinant varies greatly at each plating. In our experience, the amount of hybridization to each member of a spot pair is almost always less than a factor of 2.5, a level of variability that does not affect detectability. The proportion of spot pairs representing a given sequence is directly related to the prevalence of that sequence in the parental mRNA used to make the library, except for the possibility that it is under- or

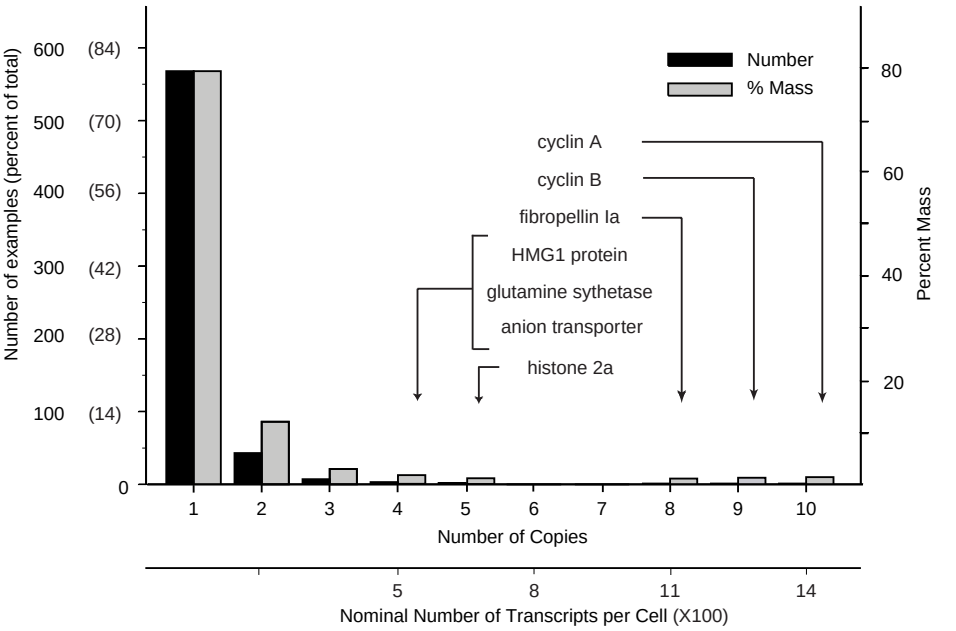


**Fig. 2.** Sequence prevalence distributions for 7-hour embryo ESTs of known and putative mRNA classes. The number of sequences occurring in the frequency classes indicated on the abscissa was summed in the combined pool of ESTs that contain recognized protein-coding sequences and ESTs which could derive from mRNAs (i.e., 719 sequences; see Table 1). The solid bars indicate the frequency distribution expressed in terms of number of sequences in each prevalence class, and as percent of total sequences in the analysis (left ordinates). The gray bars indicate the prevalence distribution in terms of per cent of total mass in the analysis, where per cent mass is calculated as the product of the number of occurrences and the number of sequences per occurrence class, divided by 719. The histograms may underestimate the actual prevalence because the clones were derived from random primed cDNAs, which usually but not always extend to within several hundred nucleotides of the 5′-end of the message (i.e., the length of the ESTs). Thus some non-overlapping sequences in the unrecognized sequence set may come from the same message. Such would normally be the case for the minor fraction of poly(A) RNAs belonging to the interspersed repeat RNA sequence class (see text), but since they are in general low prevalence transcripts (reviewed in Davidson, 1986), this will have little overall effect on the distribution. The identity of the sequences in the higher prevalence classes is indicated (see Table 2), and their extrapolated prevalence, i.e., copies per average embryonic cell (C) is indicated on the lower abscissa. This is calculated as $C = (10^5/719)\times P$ where $10^5$ is the number of mRNAs per average cell at a 500-cell stage (see Davidson, 1986) and P is the number of occurrences of the sequence in the 719 EST set.

overtranscribed by the reverse transcriptase employed to generate the cloned cDNA. This occasional inaccuracy, however, equally affects methods based on quantitative measurement of the amount of a cDNA hybridized to a given clone (or a synthetic DNA sequence).

## Conclusions

Though it is of relatively small size, analysis of this set of ESTs has yielded several kinds of useful information pertaining to the mid-cleavage-stage sea urchin embryo. Among the main results are the following.

(1) We have obtained an overall, quantitative classification of the various types of transcript represented in the arrayed 7-hour embryo cDNA library and in the embryo itself.

(2) A representation of the population structure of the embryo poly(A) RNA has emerged that strongly supports the earlier conclusions that most transcripts are rare in the egg, while a few species, most of which are already known, occur at a modestly higher prevalence. The major complexity of early embryo RNA, i.e., the greatest diversity of genes represented in its transcript populations, is in the rare sequence class. A practical implication is that since a relatively small fraction of the ESTs consist of prevalent mRNA species (or of mitochondrial, ribosomal and interspersed RNAs), it is not particularly advantageous to normalize libraries of this stage, or to go to great effort to remove or identify very prevalent transcripts prior to other analyses. Furthermore, this result emphasizes the importance of methods (including EST analysis) to which rare mRNAs are accessible: this is of course where most the expressed genetic information is to be found.

(3) The prevalence distribution also provides a reasonable estimation of the actual frequency of occurrence in embryos of the more highly represented transcripts, those present at a few hundred molecules per average cell. Furthermore, it is easy to track developmental changes in representation using array library prints once the desired sequences have been identified.

(4) Sea urchin embryos express many protein-coding sequences that are too divergent from those at present in GenBank to provide identification; we found at least as many unidentified protein-coding sequences as identified ones. Eventually knowledge of protein-folding motifs will permit educated guesses as to function from most protein-coding sequences, and we may expect the ‘unidentified’ category to shrink.

(5) Finally, we have uncovered a number of recognized molecules previously not known to be expressed in cleavage-stage embryos. Among these are Notch and TGF-β family signaling components, a function for which is now implied in the mid-cleavage embryo.

A *Strongylocentrotus purpuratus* genome project has been initiated (funded by the Stowers Institute for Medical Research), from which there is now emerging a high-resolution BAC-end sequence map of the whole genome and a large collection of arrayed libraries representing various embryonic and larval stages and cell types. The EST analysis described in this paper illustrates the illuminating informational returns that can accrue when genomic approaches are applied to a developmental system that is relatively well characterized at the molecular level.

We thank Dr Hans Lehrach, Director, Max Planck Institute for Molecular Genetics, Berlin, for his help with the construction of the arrayed library used in this study. This work was supported by the Stowers Institute for Medical Research.

## REFERENCES

**Calzone, F. J., Lee, J. J., Le, N., Britten, R. J. and Davidson, E. H.** (1988). A long, nontranslatable poly(A) RNA stored in the egg of the sea urchin *Strongylocentrotus purpuratus*. *Genes Dev.* **2**, 305-318.

**Chen, Z. Q., Kan, N. C., Pribyl, L., Lautenberger, J. A., Moudrianakis, E. and Papas, T. S.** (1988). Molecular cloning of the EST-protooncogene of the sea urchin and analysis of its developmental expression. *Dev. Biol*. **125**, 432-440.

**Costantini, F. D., Britten, R. J. and Davidson, E. H.** (1980). Message sequences and short repetitive sequences are interspersed in sea urchin egg poly(A)$^+$ RNAs. *Nature* **287,** 111-117.

**Davidson, E. H.** (1986). *Gene Activity in Early Development*. Third Edition. Orlando, Florida: Academic Press.

Davidson, E. H., Cameron, R. A. and Ransick, A. **(1998). Specification of cell fate in the sea urchin embryo: Summary and some proposed mechanisms.** ***Development*** 125**, 3269-3290.**

**Duncan, R. and Humphreys, T.** (1981). Most sea urchin maternal mRNA sequences in every abundance class appear in both polyadenylated and nonpolyadenylated molecules. *Dev. Biol*. **88**, 201-210.

**Evans, T., Rosenthal, E., Youngblom, J., Distel, D. and Hunt, T.** (1983). Cyclin: A protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell* **33**, 389-396.

**Flytzanis, C. N., Brandhorst, B. P., Britten, R. J. and Davidson, E. H.** (1982). Developmental patterns of cytoplasmic transcript prevalence in sea urchin embryos. *Dev. Biol*. **91,** 27-35.

**Galau, G. A., Klein, W. H., Davis, M. M., Wold, B. J., Britten, R. J. and Davidson, E. H.** (1976). Structural gene sets active in embryos and adult tissues of the sea urchin. *Cell* **7,** 487-505.

**Galau, G. A., Lipson, E. D., Britten, R. J. and Davidson, E. H.** (1977). Synthesis and turnover of polysomal mRNAs in sea urchin embryos. *Cell* **10**, 415-432.

**Goustin, A. S.** (1981). Two temporal phases for the control of histone gene activity in cleaving sea urchin embryos (*S. purpurtus*). *Dev. Biol*. **87**, 163.

**Johnston, S. H., Rauskolb, C., Wilson, R., Prabhakaran, B., Irvine, K. D. and Vogt, T. F.** (1997). A family of mammalian *Fringe* genes implicated in boundary determination and the Notch pathway. *Development* **124**, 2245-2254.

**Kelsowine-Miller, L., Yoon, J., Peeler, M. T. and Winkler, M. M.** (1993). Sea urchin maternal messenger RNA classes with distinct developmental regulation. *Dev. Genet.* **14**, 397-406.

**Lasky, L. A., Lev, Z., Xin, J.-H., Britten, R. J. and Davidson, E. H.** (1980). Messenger RNA prevalence in sea urchin embryos measured with cloned cDNAs. *Proc. Natl. Acad. Sci. USA* **77,** 5317-5321.

**Lee, J. J., Calzone, F. J., Britten, R. J., Angerer, R. C. and Davidson, E. H.** (1986). Activation of sea urchin actin genes during embryogenesis. Measurement of transcript accumulation from five different genes in *Strongylocentrotus purpuratus. J. Mol. Biol*. **188**, 173-183.

**Maier, E., Meier-Ewert, S., Ahmadi, A. R., Curtis, J. and Lehrach, H.** (1994). Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridization. *J. Biotech.* **35**, 191-203.

**Pines, J. and Hunt, T.** (1987). Molecular cloning and characterization of the messenger RNA for cyclin from sea urchin eggs. *EMBO J*. **6**, 2987-2995.

**Ponce, M. R., Micol, J. L., Peterson, K. J. and Davidson, E. H.** (1999). SpBMP5-7, a new member of the transforming growth factor-β superfamily expressed in sea urchin embryo. *Mol. Biol. Evol.*, **16**, 634-645.

**Reynolds, S. D., Angerer, L. M., Palis, J., Nasir, A. and 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.

**Sherwood, D. R. and McClay, D. R.** (1997). Identification and localization of sea urchin Notch homologue: Insights into vegetal plate regionalization and Notch receptor regulation. *Development* **124**, 3363-3374.

**Stenzel, P., Angerer, L. M., Smith, B. J., Angerer, R. C. and Vale, W. W.** (1994). The univin gene encodes a member of the transforming growth factor-beta superfamily with restricted expression in the sea urchin embryo. *Dev. Biol*. **166**, 149-158.

**Xin, J.-H., Brandhorst, B. P., Britten, R. J. and Davidson, E. H.** (1982). Cloned embryo mRNAs not detectably expressed in adult sea urchin coelomocytes. *Dev. Biol*. **89,** 527-531.