23. M. R. Illies, M. T. Peeler, A. M. Dechtiaruk, C. A. Ettensohn, *Dev. Genes Evol.* **212**, 419 (2002).
24. P. Oliveri, E. H. Davidson, *Curr. Opin. Genet. Dev.* **14**, 351 (2004).
25. G. Amore, E. H. Davidson, *Dev. Biol.* **293**, 555 (2006).
26. V. F. Hinman, A. T. Nguyen, R. A. Cameron, E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13356 (2003).
27. D. H. Erwin, E. H. Davidson, *Development* **129**, 3021 (2002).
28. E. H. Davidson, D. H. Erwin, *Science* **311**, 796 (2006).
29. E. H. Davidson, *The Regulatory Genome. Gene Regulatory Networks in Development and Evolution* (Academic Press/Elsevier, San Diego, CA, 2006).
30. The Echinoid Directory (www.nhm.ac.uk/research-curation/projects/echinoid-directory).
31. G. Amore, E. H. Davidson, *Dev. Biol.* **293**, 555 (2006).
32. This work was partially supported by NSF grant IOB-0212869 (to R.A.C.), NIH grant RR-15044 (to E.H.D.), and the Caltech Beckman Institute. D.J.B. is supported by NASA, NSF, and the University of Southern California; K.J.P. is supported by NSF, NASA-Ames, and Dartmouth College.

10.1126/science.1132310

REPORT

The Transcriptome of the Sea Urchin Embryo

Manoj P. Samanta,[1] Waraporn Tongprasit,[2,3] Sorin Istrail,[4,5] R. Andrew Cameron,[5] Qiang Tu,[5] Eric H. Davidson,[5] Viktor Stolc[2]*

The sea urchin *Strongylocentrotus purpuratus* is a model organism for study of the genomic control circuitry underlying embryonic development. We examined the complete repertoire of genes expressed in the *S. purpuratus* embryo, up to late gastrula stage, by means of high-resolution custom tiling arrays covering the whole genome. We detected complete spliced structures even for genes known to be expressed at low levels in only a few cells. At least 11,000 to 12,000 genes are used in embryogenesis. These include most of the genes encoding transcription factors and signaling proteins, as well as some classes of general cytoskeletal and metabolic proteins, but only a minor fraction of genes encoding immune functions and sensory receptors. Thousands of small asymmetric transcripts of unknown function were also detected in intergenic regions throughout the genome. The tiling array data were used to correct and authenticate several thousand gene models during the genome annotation process.

Embryogenesis in the sea urchin occurs rapidly and is relatively simple in form (*1*). By 2 days after fertilization, when the embryo is in the late gastrula stage, there are about 800 cells and 10 to 15 cell types. Thus, genes expressed in individual cell types or territories represent a larger fraction of the total number of transcripts than do genes expressed in adult organs of vertebrates or in more complex embryos such as that of *Drosophila*. Earlier studies have provided extensive quantitative evidence on transcript prevalence for sea urchin embryos, both for populations of mRNA (and nuclear RNA) and for many individual transcripts, measured by quantitative polymerase chain reaction (QPCR) (*2–4*). The genome sequence of *Strongylocentrotus purpuratus* (*5*) enabled these advantages to be exploited for a whole-genome tiling array analysis of the embryonic transcriptome.

Transcriptome analysis by whole-genome tiling array (*6–9*) has three advantages relative to standard microarray analysis with oligonucleotide probes constructed on the basis of known or predicted protein-coding genes: (i) The genes identified are not limited a priori by the gene predictions used to design the probes and therefore are not biased in favor of more prevalent or more conserved sequences; (ii) the transcripts detected will include noncoding as well as protein-coding RNAs; and (iii) intron-exon boundaries plus untranslated regions (UTRs) are revealed. In comparison with expressed sequence tag (EST) or cDNA-based approaches, whole-genome tiling arrays offer an unbiased and complete view of the transcriptional activity of the genome in the developmental state examined and in addition display the intron and exon structures of expressed genes. In itself, tiling array data cannot assign a distant exon to its gene, but this shortcoming can be overcome by integrating tiling and EST/cDNA data for genome annotation.

Tiling array experiments have traditionally been performed only several years after genome sequencing (*9*). However, maskless array synthesizer technology permitted us to develop custom arrays from preliminarily assembled draft sequence. This initiative enhanced the genome project while it was still in process, by substantially reducing the gap between sequencing and comprehensive annotation of the genome.

To sample transcriptional activity throughout early sea urchin development on a single set of high-density microarrays, we prepared polyadenylated RNA from egg, early blastula (15 hours), early gastrula (30 hours), and late gastrula stage (45 hours) embryos. Samples were mixed in equal quantities, reverse transcribed, fluorescently labeled, and hybridized. The tiling array probes were designed from the initial draft assembled sequence, which at that time was based on 6× whole-genome shotgun sequence coverage (*5*). A total of 10,133,868 50-nucleotide (nt) probes were selected to uniformly represent the entire sea urchin genome, maintaining an average spacing of 10 nt between consecutive probes (table S1). Repetitive sequences and simple sequence tracts were excluded. The probes were synthesized on 27 glass-based microarrays. To avoid any potential bias due to cutoff selection based on unexpressed genomic probes, we also added a set of 1000 random sequences not represented anywhere in the genome to each array. The cutoff was such that only 1% of those random probes were falsely expressed. Additionally, each array included a small (2000) identical set of genomic control probes used for normalization purposes. After hybridization, data from all arrays were normalized according to the control probes, mapped back to the latest genome sequence assembly, and mounted on a genome browser together with the optimal set of computationally derived gene models [OGS set in (*5*); for visual presentation of all transcriptome results as in Fig. 1A, see www.systemix.org/sea-urchin]. Details of the methods used are available in the Supporting Online Material (*10*), and the microarray designs and experimental data have been deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo) under the accession code GSE6031.

Analysis of signals for 28 well-characterized genes (*11*) (table S2) showed that the array measurements were highly sensitive. When mapped against the known structure of these genes, it was apparent that transcribed regions were clearly distinguished from silent regions, and no intronic transcripts were detected. Intron-exon boundaries of expressed genes were thus clearly distinguishable (e.g., Fig. 1A, fig. S1). To establish a conservative statistical criterion of expression, we first established the background variance and chose a cutoff value about 2.5 times that of the mean background. At this value, about 1% of random control probes displayed apparently artifactual noise, e.g., single-point peaks over background surrounded by probes at the background level (as in the single-

[1]Systemix Institute, Los Altos, CA 94024, USA. [2]NASA Ames Genome Research Facility, Moffet Field, CA 94035, USA. [3]Eloret Corporation, Sunnyvale, CA 94086, USA. [4]Brown University, Providence, RI 02912, USA. [5]California Institute of Technology, Pasadena, CA 91125, USA.

*To whom correspondence should be addressed. E-mail: vstolc@arc.nasa.gov

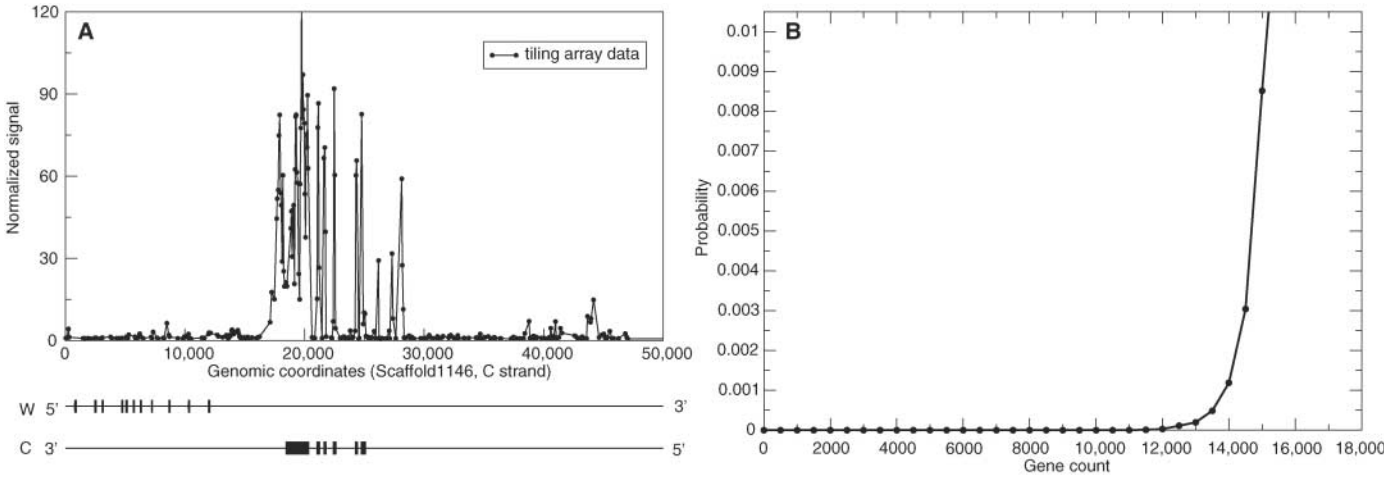

Fig. 1. Visualization of transcription profiles in protein-coding genes and probability of false-positives. (**A**) An active and an inactive gene. The protein-coding regions of the genes are indicated by the bars, and the orientation of the genes by the DNA strands (W, C) on which they are portrayed. Hybridization of each chip in the array is shown in arbitrary units (ordinate). The active gene is *Sp-gcm*, for which complete cDNA sequence is also available (*21*), transcribed from right to left. The activity profile includes the 3′-UTR, extending beyond the terminal codogenic region in the last exon. The inactive gene encodes an adenosine 5′-diphosphate ribosylation factor related protein. The peaks preceding the first exon of *gcm*, at 26,000 to 28,000, and at 45,000, are of the short noncoding RNA class. (**B**) Poisson probabilities of occurrence of falsely positive expression assignment, as a function of the total of number of apparently expressed genes. The probability that the expression profile of each gene could have been generated from adventitiously noisy probes was computed as described in the text.

probe intron peak of fig. S1A). We determined whether a gene is actually expressed in the 0- to 45-hour embryo by assessing the significance of transcriptional activity in the set of probes that lie within the predicted exons of that gene (*10*). For each gene in the OGS set (*5*), a Poisson calculation was performed, based on the number of probes in the array overlying the exons of the gene that score as active, to estimate the probability that the observed profile was artifactual. Above about 12,000 to 13,000 active gene models, the probability of false-positives rose rapidly (Fig. 1B, table S3). Some genuinely active genes are no doubt excluded by this cutoff—for example, genes that consist entirely of very small exons, or genes that are represented by very few probes (<3) because of sequence features that precluded choice of those sequence elements for representation in the probe set (*10*), or genes not represented in the genome assembly.

To estimate the number of genes expressed in the embryo up to the late gastrula stage, several corrections were required. Of the approximately 12,000 to 13,000 OGS gene models unequivocally scored as expressed (Fig. 1B), 1400 were duplicates, an artifact of high genomic polymorphism in the initial assembly process (*5*). A further 250 active gene models were excluded, because they are single-exon reverse transcriptase genes (mobile elements). On the other hand, this measurement detected a number of active open reading frames not represented in the gene model set used in this study (*5*). Where these were near one another, they were clustered, and the probability of accidental occurrence of these open reading frames in an 800-Mb genome was calculated. In total, ~1000 such putative genes were identified with a false-positive rate of <1% (table S4).

We may compare the end result, about 11,000 to 12,000 genes expressed, to the conclusion derived a quarter of a century ago from saturation single-copy sequence hybridization of embryo polysomal mRNA (*2*). This conclusion was that the same embryo uses about 8500 different genes (counting all the members of any given repetitive class of genes as 1) at the gastrula stage and, if other stages are added in (as they are here), about 10% more. Given that genes with high sequence similarity in large (more than 100-member) gene families would have been excluded from the earlier hybridization results, the two values are reasonably consistent. In any case, these measurements demonstrate that even by conservative estimates, a very large number of protein-coding informational units are required for the construction of this embryo, simple as it is, amounting to at least half of the total number of genes predicted in the *S. purpuratus* genome (12,000 out of 23,500) (*5*).

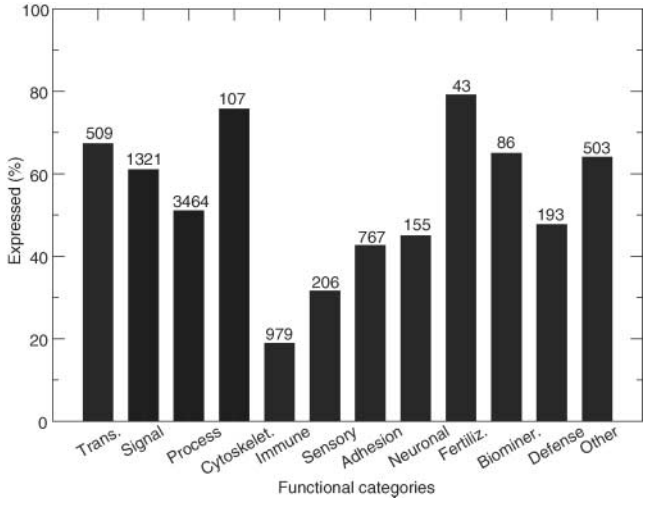

Fig. 2. Functional distribution of genes expressed in the embryo. The bar chart displays the percentage of annotated genes of different functional categories expressed in the sea urchin embryo. The functional categories are derived from a manual curation database (*5*) and are shown in table S3. The number at the top of each bar represents the total number of annotated genes in the corresponding category, including all expressed and unexpressed ones. Trans., transcription factors; Signal, signaling genes; Process, basic cellular processes such as metabolism; Cytoskelet., cytoskeletal; Fertiliz., fertilization; Biominer., biomineralization.

In *S. purpuratus*, the embryo gives rise to a larva after 3 days of development, within which the adult form develops during the successive weeks of larval feeding. By the late gastrula stage, only some small patches of undifferen-

tiated cells set aside from the processes of embryonic specification for adult body formation (*12*), and the midgut, will contribute to the adult body plan in the postembryonic period. The descendants of most of the 48-hour embryo cells will be jettisoned at metamorphosis. In contrast, in other embryos for which we have array-based transcriptome measurements, such as *Drosophila* (*9*) and *Caenorhabditis elegans* (*13*), the development of adult body parts begins immediately upon gastrulation, and there is no point after the very earliest stage at which embryonic gene use per se can be separated from gene use to construct the adult body plan.

More than 9200 OGS gene models were functionally annotated in the course of the genome project (*5*). In Fig. 2 we report the fractions of these genes expressed during embryogenesis, according to their functional classes (table S3). Most notable is the high embryonic usage of transcription factor and signaling genes. In other work (*14*), Howard-Ashby *et al.* showed by QPCR measurements that nearly 80% of all genes encoding transcription factors other than putative Zn-finger transcription factors are expressed by 48 hours [in Fig. 2, Zn finger proteins in the "Trans." category are probably not all transcription factors (*15*)]. Thus, it requires most of the "regulome" just to construct the single-cell-thick gastrula embryo. These same genes must, in general, be used repeatedly in the construction of the far more complex adult body plan. Genes related to basic cellular processes (e.g., intermediary metabolism) and cytoskeletal structure (e.g., actins and myosins) were also highly expressed; these would be expected to be required in cells of both embryo and adult tissues. This is true as well of detoxification and other xenobiotic defense molecules—the price of existence in the marine environment—and of biomineralization and neuronal molecules partially shared by the respective embryonic and adult differentiated cell types. By contrast, the immune genes (*5*, *16*) are largely expressed in the coelomocytes, which are the adult immune effector cell types. There is an elemental embryonic and larval immune defense system as well, mediated by certain embryonic mesenchymal cells, and this may account for the ~20% usage of immune genes in the embryo transcriptome (*16*). Sensory genes, such as G-coupled sensory receptors, are expressed in adult structures, the tubefeet and the pedecellaria (*17*), although again there is a rudimentary larval sensory system, about which very little is known. Little genomewide data are currently available on gene expression in adult bodies of the sea urchin. An experiment similar to this one is not meaningful for entire adult bodies made of large numbers of different tissues, because transcripts present in rare tissues will not be visible. Expression in each tissue will need to be measured individually.

Qualitatively, the transcription profiles enabled thousands of the >9200 gene models annotated by the Consortium (*5*) to be directly checked or corrected. Table S5 presents the results of our comparison between each predicted gene model and the transcribed regions derived from the tiling data. The gene models were mainly accurate, but missing exons were often identified by reference to these profiles. On average, the OGS genes expressed in the sea urchin embryo were 15.8 kb long and contained 9 exons, whereas the OGS genes on average were 11.9 kb long with 6.6 exons. Lack of tiling probes on short OGS genes with few exons may have contributed to the difference. The transcriptome data also indicated the dimensions of the 3′-UTR sequences (table S3), as well as the approximate transcription start sites. Many of the subgroups of sea urchin annotators used the high-resolution array data to manually curate their genes of interest (*5*). It was thus particularly useful for the subsequent analysis that the transcriptome measurements were carried out at a relatively early stage of the genome sequencing project as a whole, as soon as the initial assembly permitted.

Finally, as in all other whole-genome array hybridizations, many enigmatic transcripts were observed that are not included in protein-coding genes (table S6). A major class of these is composed of short ($\leq$ 200 nt) asymmetrically represented transcripts, of which some 51,000 were recorded (table S7). Only a small fraction (about 2000) represent sequences that also occur in active, protein-coding genes (including 3′-UTRs), and these repetitive sequences were excluded. Nor is any appreciable fraction complementary to other short noncoding sequences, again excluding the possibility that they are repetitive sequences transcribed elsewhere. Similarly, there was little homology to any known microRNAs (miRNAs), and the short transcripts are smaller than typical pre-miRNAs (>1 kb). Nearly 170 of the 51,000 transcribed regions are conserved in the human genome (BLAST cutoff 1×10^{-5}, table S7) and could potentially represent noncoding RNAs. Statistically, these 51,000 transcripts are exactly as likely to occur far from any active gene, in a distant intergenic domain, as near a gene or in its introns. To determine if these could in general be transcripts produced in cis-regulatory modules [e.g., (*18*)], we first manually examined a number of known examples, but in almost no instance observed the authenticated cis-regulatory modules we used as probes to be represented by these transcripts. We also compared the locations of the short transcripts in the vicinity of 28 well-characterized genes (table S2) to those of all interspecifically conserved sequence patches. There were about 500 such patches in introns or within 30 kb of these genes in either direction (*3*, *4*, *19*), and many have cis-regulatory activity [e.g., (*20*)]. Only 21 of these patches overlapped regions included in the short transcripts, a result not different from random expectation, and with one exception the termini of the patches and the transcripts were noncoincident. We therefore believe that these transcripts do not represent cis-regulatory modules, although experimental verification will be necessary.

The number of active genes sets in concrete terms the dimensions of the regulatory task of the genomic control apparatus driving embryogenesis. It cannot be said that the transcriptome is functionally understood until the individual roles and interactions of each component are revealed. Assessment of transcriptional activity across the whole genome represents the essential beginning of that process.

References and Notes

1. E. H. Davidson, A. Ransick, R. A. Cameron, *Development* **125**, 3269 (1998).
2. E. H. Davidson, *Gene Activity in Early Development* (Academic Press, San Diego, 1986).
3. H. Bolouri, E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9371 (2003).
4. For a large current compilation focused upon regulatory genes, see http://sugp.caltech.edu/endomes/.
5. Sea Urchin Genome Sequencing Consortium, *Science* **314**, 941 (2006).
6. P. Kapranov *et al.*, *Science* **296**, 916 (2002).
7. P. Bertone *et al.*, *Science* **306**, 2242 (2004).
8. V. Stolc *et al.*, *Science* **306**, 655 (2004).
9. F. Biemar *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12763 (2006).
10. Materials and Methods, figs. S1 and S2, and tables S1 to S7 are available as supporting materials on *Science* Online.
11. It is a set of regulatory genes for which full-length cDNA sequences are available. Their embryonic expressions were previously confirmed by QPCR measurement (*3*, *4*).
12. K. J. Peterson, R. A. Cameron, E. H. Davidson, *Bioessays* **19**, 623 (1997).
13. M. N. Arbeitman *et al.*, *Science* **297**, 2270 (2002).
14. M. Howard-Ashby *et al.*, *Dev. Biol.* 10.1016/j.ydbio.2006.10.016, in press.
15. S. C. Materna, M. Howard-Ashby, R. Gray, E. H. Davidson, *Dev. Biol.* 10.1016/j.ydbio.2006.08.32, in press.
16. T. Hibino *et al.*, *Dev. Biol.* 10.1016/j.ydbio.2006.08.065, in press.
17. F. Raible *et al.*, *Dev. Biol.* 10.1016/j.ydbio.2006.08.070, in press.
18. M. Ronshaugen, M. Levine, *Dev. Cell* **7**, 925 (2004).
19. Conserved patches were derived on the basis of sequence homology between two distant sea urchin species (*3*, *4*).
20. C.-H. Yuh *et al.*, *Dev. Biol.* **246**, 148 (2002).
21. A. Ransick *et al.*, *Dev. Biol.* **246**, 132 (2002).

22. This work was supported by grants to V.S. from the NASA Center for Nanotechnology, the NASA Fundamental Biology Program, the Computing, Information, and Communications Technology programs (contract NAS2-99092), NIH grant HD-37105 (to E.H.D.), and Brown University (to S.I.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/314/5801/960/DC1
Materials and Methods
Fig. S1
Tables S1 to S7
References

29 June 2006; accepted 23 October 2006
10.1126/science.1131898