A Database for Mouse Development

Martin Ringwald,* Richard Baldock, Jonathan Bard, Matthew Kaufman, Janan T. Eppig, Joel E. Richardson, Joseph H. Nadeau, Duncan Davidson

Genetic information is expressed in complex and ever-changing patterns throughout the developing mammalian embryo. A description of these patterns that can be accessed through information resources is crucial for our understanding of the network of genetic interactions that underlies the processes of normal development, disease, and evolution.

We now have numerous methods for measuring gene expression, and data on expression patterns are accumulating rapidly. An important problem remains—that is, how to acquire, manage, analyze, interpret, and disseminate these data. We need to be able to answer both simple queries, such as when and where a particular gene is expressed, and more complex queries that require a deeper understanding of developmental processes and access to diverse information resources [sequence (*1*), gene mapping, and disease description databases (*2*)]. Here we describe an ongoing project between the Jackson Laboratory, the Medical Research Council (MRC), and the University of Edinburgh to establish a gene expression information resource for mouse development that would contain such information organized so that it can be readily queried. [Efforts are also ongoing in other model systems (*3*).]

Genes can encode alternative RNAs that in turn give rise to proteins with distinct expression patterns and biological activities. To be useful, a gene expression information resource should document which RNA and protein species are encoded by a given gene, and should include hybridization data, immunofluorescence analysis, Northern (RNA) and protein immunoblot analysis, ribonuclease protection, and polymerase chain reaction and complementary DNA (cDNA) sequence analysis, each of which provides unique insights into expression patterns. The information from each of these assays must be stored and integrated. Because expression data are highly dependent on the molecular assays and reagents used, experimental conditions must be documented so that accurate interpretation of empirical observations is possible.

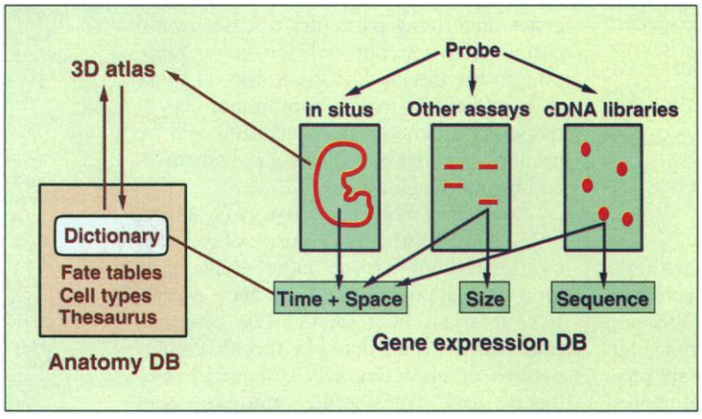

Fig. 1. The gene expression information resource. A 3D atlas of mouse development will be linked to an anatomy database (DB) through labeling of the anatomical structures. The gene expression database stores and integrates the data from the different expression assays together with the experimental conditions. Description of the probe is critical as it constitutes the molecular window for each assay. In "other assays" the size and the expression pattern is described separately for each detected band because each band represents a distinct molecular species. Sequence data will be included by cross referencing to other databases.

For these reasons, "raw" rather than "processed" expression data must be stored (Fig. 1). In this format, new raw data can be readily added and integrated with other information in the database. Moreover, novel insights resulting from new data can be represented, because higher order information can be synthesized from the available raw information.

The gene expression database must describe the time and space of gene expression in standardized ways. To achieve this goal, our database will be coupled to a three-dimensional (3D) atlas of mouse development and to a text database of mouse anatomy. A query system links these three components. The 3D atlas is a computer representation of mouse anatomy that will provide a standardized coordinate system for the storage and analysis of visual expression data (Fig. 2). The technical problems are challenging—at a cellular level, the digital images of serial sections at all representative stages of mouse development must be aligned (*4*), each anatomical structure must be labeled, and in situ data must be overlaid. The ultimate goal, and challenge, is to transfer electronically digitized in situ data from different embryos into the 3D atlas (*5*).

A key component of the anatomy database is a dictionary of anatomical terms derived from the *Atlas of Mouse Development* (*6*) that names the tissues and structures for each developmental stage. Terms from the dictionary will be assigned to each tissue in the 3D atlas and used by the gene expression database as the standard nomenclature for data entry and database queries. This assignment is the critical link between the graphical and text-based approaches. The anatomy is modeled hierarchically from body region to tissue to substructure in order to fit with the different degrees of resolution in data capture and analysis. Additional information associated with each anatomical structure (cell type and fate, synonymous names, definitions, and references) provides the opportunity for queries that relate expression to developmental anatomy. Fate tables, for example, enable analysis of differentiation pathways, whereas synonyms establish a thesaurus of alternative terminology, adding flexibility to database queries (*7*).

In spite of the fact that "a picture is worth a thousand words," text remains a prerequisite for database queries. Textual information is essential for the description and integration of expression data that are not derived from in situ studies. Text provides the basis for integrating this database with other information resources. Thus, although more complex approaches such as image querying are under development, a text database enables research progress with existing technology.

Textual descriptions of in situ studies will be complemented with digitized images of original expression data, and these will be indexed with anatomical terms from the dictionary. Ultimately, the 3D atlas will enable display and analysis of 3D expression domains, because these patterns can be viewed from any angle, arbitrarily "sectioned," selectively displayed, and overlaid with each other and with anatomical structures. It is difficult to visualize complex expression patterns in our mind. So, the 3D

M. Ringwald, J. T. Eppig, J. E. Richardson, and J. H. Nadeau are at the Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609 USA. R. Baldock and D. Davidson are in the Medical Research Council Human Genetics Unit, Western General Hospital, Edinburgh, UK. J. Bard and M. Kaufman are in the Department of Anatomy, University of Edinburgh, Edinburgh, UK.

*To whom correspondence should be addressed.

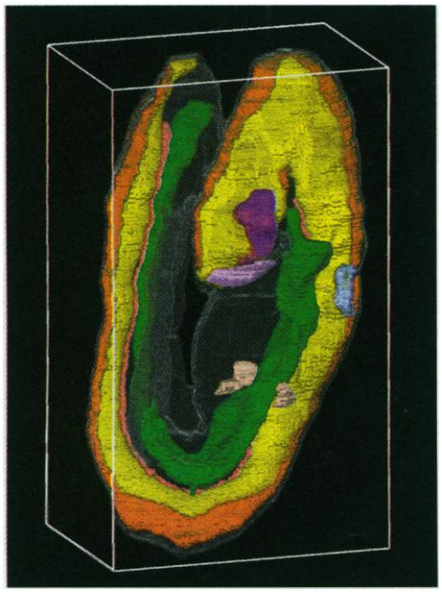

Fig. 2. A reconstruction of the 9-day mouse embryo. Anatomical areas are color-coded: green, gut; pale blue, rudimentary ear; purple, rudimentary eye; pink, rudimentary nose; salmon, notochord; and yellow, neural tube. The small beige areas are where Wtl is first expressed and the orange areas indicate the parts of the neural tube where the gene *MSX1* is expressed. The reconstructions were made at the MRC Human Genetics Unit.

atlas will fundamentally extend our thinking about the relation between gene expression and tissue differentiation. Because expression domains generally do not align with anatomical structures as they are presently known, 3D representations of expression patterns will give molecular definitions for developmental components. New developments in magnetic resonance microscopy will strongly reinforce this research direction (*8*).

Standard publication methods are inadequate for making gene expression data freely and widely available. For example, most raw data for in situ studies are never published but are merely referred to in brief text summaries. Electronic data submission from contributing laboratories will therefore be a key requirement for the success of the gene expression information resource. This resource should be seen as a tool for the biomedical research community to place gene expression data into the proper biological and analytical context, so that the pieces of the developmental puzzle can successively be put together. Electronic communication tools like the World Wide Web (*9*) can support this collaborative effort and will allow links with complementary efforts related to gene expression in other species (*3*) and to large-scale characterization of cDNA sequences (*10*).

A gene expression information resource can support a variety of complex biological queries. A user could, for example, ask: What gene products are found at a specific location and developmental stage? What genes are expressed along specific differentiation pathways, and in which order do the gene products appear or disappear? To which functional classes do those genes belong? Is there evidence for direct molecular interactions, for hierarchical or combinatorial control, for signaling pathways?

Because this database would be integrated with others containing DNA and protein sequence data (*1*), genetic and physical mapping data, and descriptions of disease states and mutant mice (*2*), the combined information resource will enable complex queries. Access to genetic mapping databases, for example, could reveal correlations between gene expression patterns and chromosomal localization that might exist because of alternative chromatin states. The gene expression information resource could facilitate identification of genes underlying particular diseases by focusing attention not only on genes mapping to the disease locus, but also on those with expression in the appropriate cell or tissue. The benefits of integration will be primarily limited by technological advances and by our imagination.

Analysis of molecular networks resulting from combinatorial mechanisms of gene activation is one of the more challenging long-term applications of the gene expression information resource. For example, transcription is regulated by specific combinations of cis-acting DNA sequence elements and trans-acting transcriptional regulatory factors. Queries investigating gene expression, DNA sequence, and transcription factor databases (*11*) could answer questions such as: Do promoters of genes with overlapping expression patterns share regulatory sequence elements? And, are there transcription factors that have the appropriate binding specificity and expression pattern to allow interaction with certain promoters at a particular developmental stage? Because interacting molecules must be present at the same time and location during development, integration of the gene expression information resource with other databases will promote research beyond simple biochemical affinities to networks of regulatory interactions.

The Human Genome Project has focused on structural aspects of the genome, namely, genetic and physical maps, gene identification, and DNA and protein sequences. Numerous databases exist to manage these data for the genomics and biomedical research communities. Similarly, numerous databases describe disease states and mutant phenotypes. The gene expression information resource fills an important gap in our knowledge of the relation between structural information and normal and abnormal phenotypes. Once established, the gene expression information resource will provide fundamental information for elaborating the function of our 50,000 to 100,000 genes and for determining how heritable mutations and epigenetic modifications lead to genetic disease.

References and Notes

1. D. Benson, D. J. Lipman, J. Ostell, *Nucleic Acids Res.* **21**, 2936 (1993); C. M. Rice, R. Fuchs, D. G. Higgins, P. J. Stoehr, G. N. Cameron, *ibid.*, p. 2967; W. C. Barker, D. G. George, H.-W. Mewes, F. Pfeiffer, A. Tsugita, *ibid.*, p. 3089; A. Bairoch and B. Boeckmann, *ibid.*, p. 3093.
2. Mouse Genome Database and Mouse Locus Catalogue, The Jackson Laboratory, Bar Harbor, ME. An overview of currently available mouse genome informatic resources can be found in J. S. Takahashi, L. H. Pinto, M. H. Vitaterna, *Science* **264**, 1724 (1994); M. T. Davidson, T. H. Roderick, D. P. Doolittle, in *Genetic Variants and Strains of the Laboratory Mouse*, M. F. Lyon and A. G. Searle, Eds. (Oxford Univ. Press, Oxford, 1989), p. 432; A. J. Cuticchia, K. H. Fasman, D. T. Kingsbury, R. J. Robbins, P. L. Pearson, *Nucleic Acids Res.* **21**, 3003 (1993); M. Baraitser and R. Winter, *Dysmorphology Database and London Neurogenetics Database* (Oxford Univ. Press, Oxford, 1993); M. L. Buyse and C. N. Edwards, *Am. J. Perinatol.* **4**, 8 (1987).
3. Controlled anatomical vocabularies are now being used to describe gene expression patterns in the nematode and in the fruit fly [R. Durbin and W. Gelbart, personal communication; S. Veretnik and B. R. Schatz, in *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, L. Hunter, D. Searls, J. Shavlik, Eds. (AAAI Press, Menlo Park, CA, 1993), p. 411; M. Ashburner and R. Drysdale, *Development* **120**, 2077 (1994)]. In addition, there is a 3D atlas of *Drosophila* development that could be used for graphical display and analysis of expression patterns [V. Hartenstein, in *Development of* Drosophila, A. Martinez-Arias and M. Bate, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1993)].
4. Magnetic resonance microscopy allows reconstruction of mouse embryos without sectioning. However, because of limited resolution, analysis is not yet possible at the cellular level. See references in (*6*) for further information.
5. R. Baldock, J. Bard, M. Kaufman, D. Davidson, *BioEssays* **14**, 501 (1992).
6. M. H. Kaufman, *The Atlas of Mouse Development* (Academic Press, London 1992).
7. The text and graphical database of gene expression is being developed jointly by the Jackson Laboratory and MRC Human Genetics Unit (and collaborating laboratories in Europe and North America) and extends the Jackson Laboratory Mouse Genome Database. The 3D atlas and anatomy database are being developed primarily by the MRC and the University of Edinburgh.
8. B. R. Smith, G. A. Johnson, E. V. Groman, E. Linney, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 3530 (1994); R. E. Jacobs and S. E. Fraser, *Science* **263**, 681 (1994).
9. B. R. Schatz and J. B. Hardin, *Science* **265**, 895 (1994).
10. J. M. Sikela and C. Auffray, *Nat. Genet.* **3**, 189 (1993).
11. P. Bucher, *The Eukaryotic Promoter Database* (EMBL Data Library, Heidelberg, 1993); D. Gosh, *Nucleic Acids Res.* **21**, 3117 (1993).

12. Development of the gene expression information resource at the Jackson Laboratory is supported by a grant from the Keck Foundation. Work at Edinburgh is supported by the MRC, the BBSRC, and by the European Science Foundation. M.R. is supported by the Deutsche Forschungsgemeinschaft.