seminars in **CELL & DEVELOPMENTAL BIOLOGY,** Vol 8, 1997: pp 509–517

The mouse atlas and graphical gene-expression database

D. Davidson, J. Bard†, R. Brune†, A. Burger‡, C. Dubreuil*, W. Hill*, M. Kaufman†, J. Quinn*, M. Stark* and R. Baldock**

The large amounts of gene-expression data on mouse development are now too extensive to be stored in any format other than that of a database. Furthermore, as this data is intrinsically graphical and as, in the early developmental stages at least, its boundaries do not map directly to those of anatomical tissues, the natural way to store it is in graphical format. We are therefore constructing a database able to handle such graphical gene-expression data by mapping it onto 3-D reconstructions of mouse embryos whose tissues have been delineated. This article reviews the progress that has been made in this project and describes its two major components, CD-ROMs of the 3-D reconstructions to be held on the user's computer and a gene-expression database that will be maintained at a host site, the two being linked over the internet by a complex Java-based interface for submitting data and querying the database.

Key words: reconstruction / gene expression / graphical databases / mouse development

©1997 Academic Press Ltd

GENE-EXPRESSION PATTERNS, together with the phenotypes of mutant embryos, are the primary source of information for the study of gene function in the embryo. Overlapping or complementary patterns suggest possible interactions between gene products. The association of gene expression with forming structures, or with particular cellular activities, suggests the involvement of particular groups of genes in local developmental processes, pointing the way to subsequent studies of gene function.

Extracting this kind of information from the literature by scanning for related gene expression patterns is becoming impractical due to the complexity and number of patterns involved. Moreover, the sheer quantity of information that is needed to describe properly the spatially and temporally complex expression of even a single gene creates serious problems for conventional publishing methods. The result is that much of the original data remains unpublished, with authors only reporting the 'main features' of gene expression and, in many cases, restricting their description to particular regions of interest.

The solution to these limitations is to build a database designed to hold gene-expression data.[1] Such a database would need to integrate information from a wide range of sources, including assays on dissected and homogenized parts of the embryo (e.g. RT-PCR, Northern-, and Western-blots, RNase protection assays, etc.) as well as high-resolution in-situ studies (in-situ hybridization, immunohistochemistry and histochemistry including reporter-gene experiments). It would also be able to store expression data at any resolution, from incomplete accounts at low-resolution to complete descriptions, and would allow information to be shared across the biomedical community. Above all, such a database would enable a researcher to quickly search gene-expression patterns, and rapidly receive details of biologically significant relationships among these patterns.

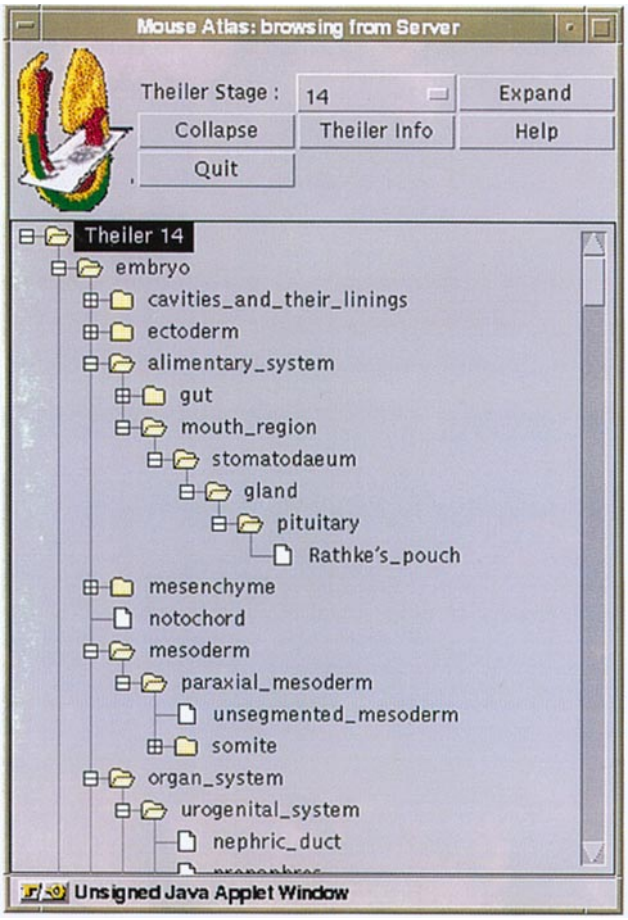
The question arises as to the most appropriate format in which to record the salient features of gene-expression patterns so that this information can be used optimally. One way is to record gene expression in textual format, using the names of the anatomically-defined regions or structures where the gene is expressed. Indeed, when structures and cell types have begun to differentiate, gene expression can generally be assigned to named components.[2] An additional advantage is that a hierarchical textual description can be used to record low-resolution as well as high-resolution data (e.g. 'gut, mouth region' as a low-resolution alternative to the more specific 'Rathke's pouch', see Figure 1).

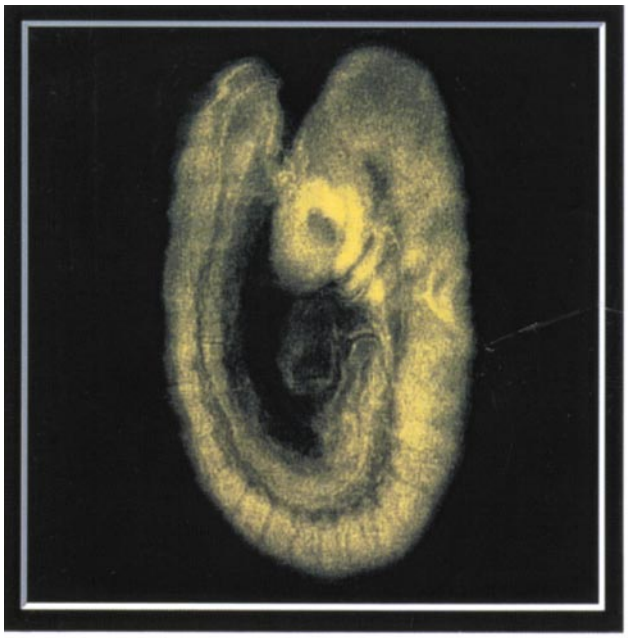
Text alone cannot, however, adequately describe all gene-expression patterns. In early development and

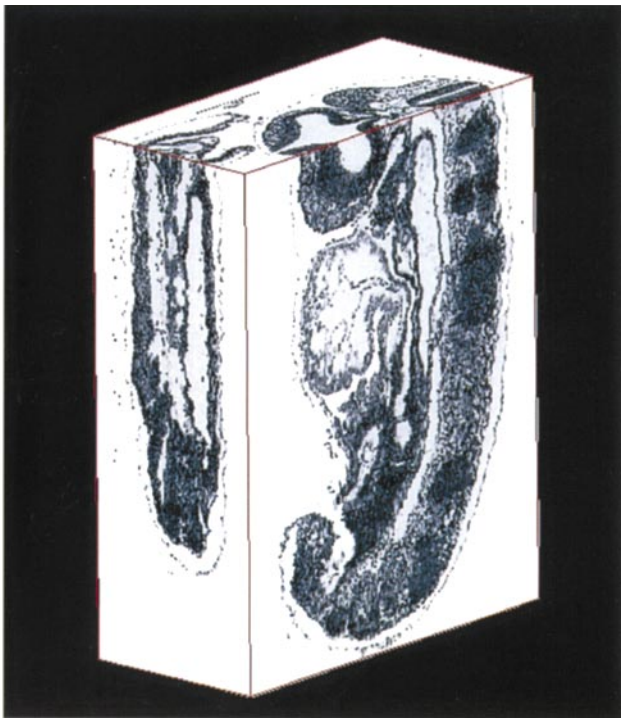
*From the *MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK, †Anatomy Department, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK and ‡Department of Computer Science, Heriot Watt University Riccarton Campus, Edinburgh, UK*

©1997 Academic Press Ltd

1084-9521/97/050509 + 09 $25.00/0/sr970174


Mouse Atlas: browsing from Server
Theiler Stage :
14
Expand
Collapse
Theiler Info
Help
Quit
Theiler 14
embryo
cavities_and_their_linings
ectoderm
alimentary_system
gut
mouth_region
stomatodaeum
gland
pituitary
Rathke's_pouch
mesenchyme
notochord
mesoderm
paraxial_mesoderm
unsegmented_mesoderm
somite
organ_system
urogenital_system
nephric_duct
Unsigned Java Applet Window





during the processes of pattern formation and organogenesis, gene expression domains frequently do not match those structures that have been recognized and named by anatomists. This non-conformity is not merely an inconvenience but reflects a fundamental principle of development. The internal, genetic control of pattern formation is often mediated by interacting signal systems that operate across space and often independently of anatomically-defined boundaries. Although these signals determine the formation of recognizable structures, the relationship between gene expression and tissue structure is often not obvious; a complex system of interactions with numerous players, including signal receptors, intracellular signalling pathways and interacting transcription factors lies between the primary pattern of signals and the ultimate anatomical pattern. In general, it seems that the outcome of these interactions is the deployment in space of feedback loops and threshold responses that establish differences between adjacent cells and lead to morphogenesis and anatomical differentiation.[3-5] The spatial discontinuity between early gene-expression patterns and anatomical development thus represents one important component in the gap between genotype and phenotype and the processes that bridge this discontinuity are the focus for many studies of gene function. It is therefore important to be able to record gene expression patterns independently of anatomical structure.

A database that is used to investigate gene function must therefore bring together, in the same system, independent descriptions of the anatomy and the spatial patterns of gene expression.

The mouse atlas

In order to provide a framework for incorporating the required textual and spatial formats, we are building a digital atlas of mouse development that is a key component of a larger gene-expression graphical database. The atlas comprises two parts; first, a database of named anatomical components that allows textual assignment of gene expression, second, a series of digital, 3-D model embryos at successive stages of development that forms the basis for purely spatial descriptions. In order to integrate textual and spatial descriptions, we are delineating the major named anatomical structures in the 3-D models of embryos at defined stages of development.[6]

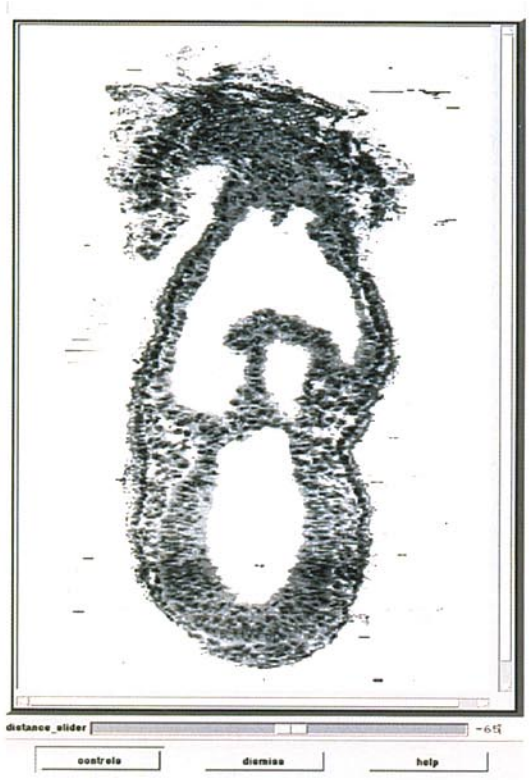
The mouse embryo anatomy database

This database includes all the named anatomical tissues that can be recognized at each Theiler Stage[6] of mouse development. The list of components derives from collating all the structures identified in *The Atlas of Mouse Development*,[7] supplemented with some additional features (e.g. each somite has its own entry). In order to provide a sensible view of the anatomy of the embryo and, most importantly, to enable gene expression to be recorded at any level of resolution, the list is organized in a spatial hierarchy (Figure 1).

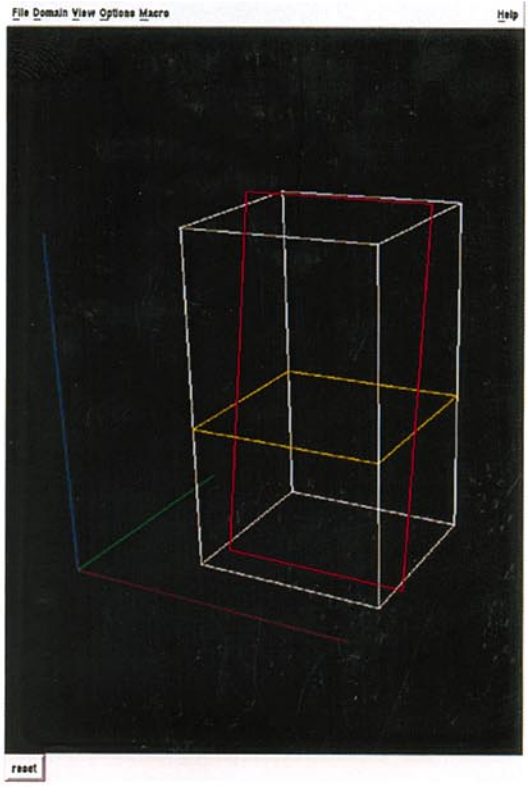
The resulting anatomical vocabulary will allow gene expression to be described down to the cellular level, with the list of tissues representing a collection of pigeon holes in which data can be stored. While any particular tissue can naturally only be represented by a single pigeon hole, the user will be able to group the components in different ways so that, for example, the *tibia* can be included under both the hindlimb and the skeleton. This anatomy database provides the common language that links the *MRC Graphical Gene Expression Database* (GGED) described here with the *Gene-Expression database* (GXD[2]). In addition, the *Mouse Embryo Anatomy Database* will enable mouse gene expression data to be linked to any other mouse embryo databases that use the same anatomical vocabulary.

Figure 1. Part of the mouse embryo anatomy database illustrating the general organization of the database as a spatial ('part of') hierarchy. Components of the hierarchical tree can be expanded or collapsed to enable chosen parts of the anatomy to be viewed, as shown here for part of the mouth region of the gut. Clicking on a name will display a box with further data (not shown) that includes synonyms, information on tissue derivation/fate and classification of tissue architecture (e.g. pavement epithelium or pseudostratified epithelium, etc.). The database is illustrated here as viewed via the World Wide Web using a Java-enabled browser.

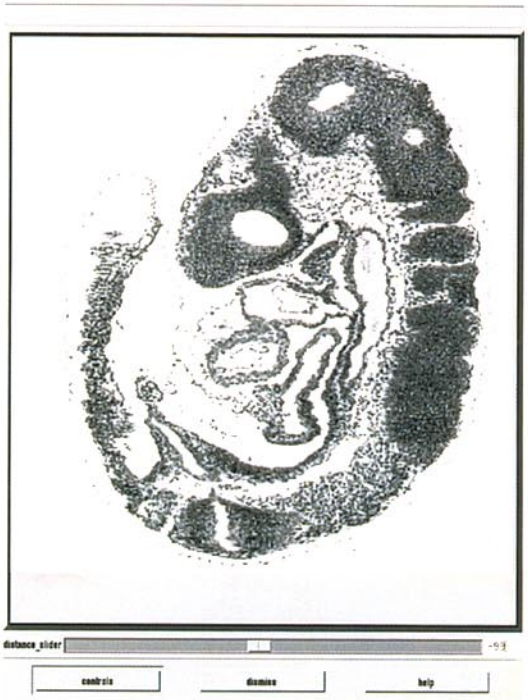
Figure 2. An example of the voxel model of the E9 mouse embryo (Theiler Stage 14) reconstructed from images of 307 transverse histological sections. (A) The model displayed as if it were a whole-mount preparation. This image was obtained using the ray-tracing software, *VolVis* (http://www.cs.sunysb.edu/~vislab). (B) A view of the same model embryo as a block of grey-level image data cut in the transverse, frontal and sagittal planes to display the histological structure at the faces of the block. This image was obtained using *Sunvision* software (Sunmicrosystems. http://www.sun.com/).


distance_slider
controls
dismiss
help

A


File Domain View Options Macro
Help
reset

B


distance_slider
controls
dismiss
help

C


distance_slider
controls
dismiss
help

D

The 3-D model embryos

Each model embryo in the mouse atlas is being built by digitizing images of serial sections and reconstructing them to produce a block of grey-level data that shows the histological structure of the embryo at each stage[8] (Figure 2). These grey-level models are to be preferred over models built from contours of anatomical components which can only include the outlines of those structures selected by the originator and not the cellular detail. An additional advantage of this approach is that the block of data can be digitally resectioned in any plane to display the histological view that corresponds to the viewer's own material (Figure 3A).

The reconstruction is composed of a regular array of volume picture elements (voxels). Delineating a set of voxels defines a 3-D domain to which any data can be attached (e.g. gene expression data, phenotypic data, or anatomical names[9]). The embryo models thus provide the spatial context for a potentially very large amount of phenotypic and molecular data. The idea of spatial mapping of data is central to the concept of the Mouse Atlas and distinguishes it from the attachment of data to text and the accumulation of images of original data.

Embryo reconstructions are being built for the principal stages from fertilization to approximately 17.5 days of development (E17.5), but we intend to take an 'open-ended' approach and will include additional stages and higher resolution models of selected organs as required. For embryos up to E8 (Theiler Stage 12), models are being reconstructed from 2 μm-thick plastic sections stained with Alcian blue to display details of tissue structure at the cellular level (Figure 3B); models of older embryos are being reconstructed from 6–8 μm-thick wax sections, stained with haematoxylin and eosin, that will show the main tissue components, but not cellular detail (Figure 3C). Embryos used for the older stages will, where practicable, be those specimens used to illustrate the text *The Atlas of Mouse Development*,[7] thereby ensuring that they correspond to the standard reference material.

Figure 3. Model embryos reconstructed from serial transverse sections and digitally re-sectioned in alternative planes using software developed at the MRC Human Genetics Unit (http://genex.hgu.mrc.ac.uk/). (A) An arbitrary section through a model of an embryo at the egg cylinder stage (approx. 7.5 days of development, Theiler Stage 10). (B) The relative angle of section through the reconstruction illustrated in A. The embryo was reconstructed from 2 μm-thick sections cut in the transverse plane (yellow); the section illustrated in A lies in the plane shown in red. (C) An approximately para-sagittal section through the model of the E9 embryo shown in Figure 2. (D) All of the major components of the embryo have been delineated in the model; the illustration shows some of these identified in colour and displayed in the same section as shown in (C): surface epithelium blue-green; Rathke's pouch, orange; neural tube + brain, yellow; otic placode, light green; gut, brown, sinus venosus, red, left atrial chamber, blue; somites, green.

Translating between text and spatial mapping

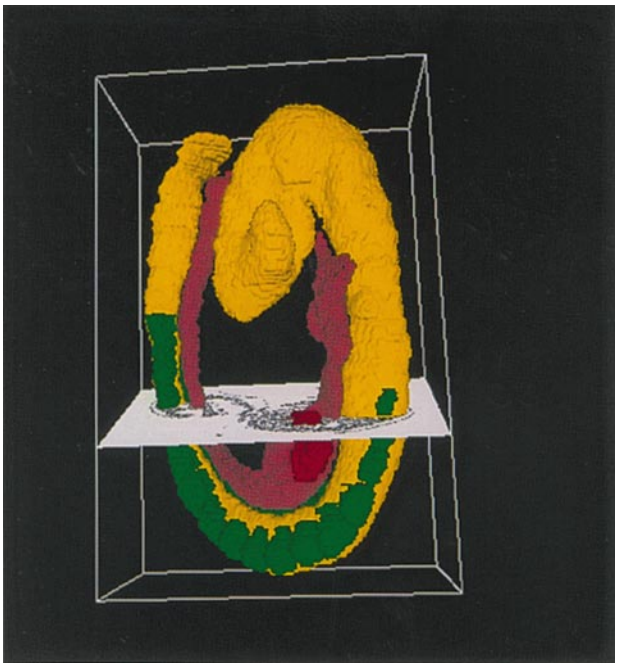
The major named anatomical components are being delineated in each model embryo (Figure 3D). This allows data originally recorded in one format to be combined with data from the other. It also enables the database to describe and compare anatomically and spatially restricted gene-expression patterns and thus provides the means to investigate those processes by which spatially deployed gene expression gives rise to recognizable tissue structure. Delineating the anatomy serves three additional functions: it provides a spatial definition of the anatomical terms, it allows the structures present in the embryo to be identified on the screen, and it permits individual anatomical components to be displayed in 3-D view so indicating spatial relationships during development (Figure 4). The use of this latter aspect of the mouse atlas as a teaching aid is discussed elsewhere.[10]

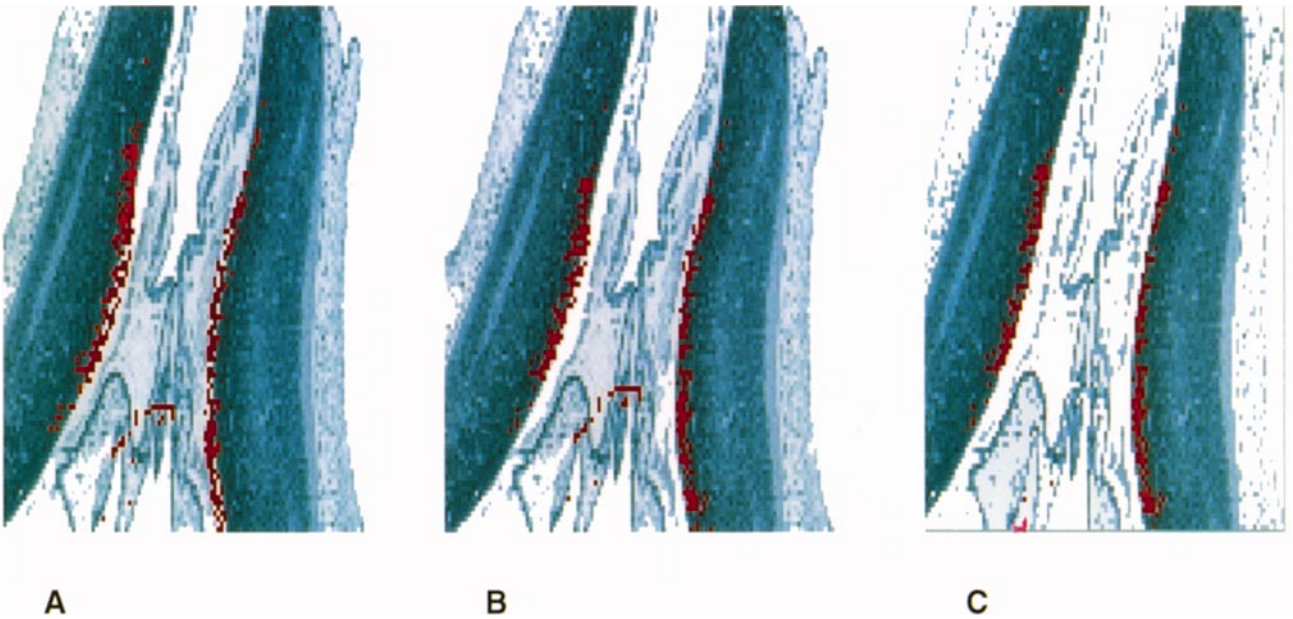
The MRC graphical gene expression database as part of the mouse gene expression information resource

The mouse atlas will be closely linked with the two parts of a gene expression database, the *GXD* and the *MRC Graphical Gene Expression Database* (the GGED). The GXD, being developed at The Jackson Laboratory,[2] will record gene-expression domains in text format supplemented with non-spatially mapped images of the original data. The GXD will use the Mouse Embryo Anatomy Database to provide the names of structures to describe gene-expression domains. The GGED will use the GXD to record ancillary data (gene name, experimental details, etc.) and will record gene-expression domains spatially mapped to the embryo models of the mouse atlas.

The GXD, MRC Database and Mouse Atlas, will each operate as modules in an integrated database, the *Mouse Gene Expression Information Resource*

(MGEIR[11]). This arrangement will enable the integration of a very wide range of information including data from extensive, large-scale screens and intensive high-resolution in-situ studies. The GXD is also integrated with the *Mouse Genome Database* (MGD) which contains genetic and phenotypic information on mouse strains and mutants. The practical considerations involved in using the MGEIR and other





gene-expression databases have been reviewed elsewhere.[1] Here we confine our discussion mainly to the spatial aspects of the gene expression database.

Operation of the MRC gene expression database

Operationally, the mouse atlas and MRC graphical gene expression database will function on two levels, local and remote. This is necessary as the 3-D mouse models are so large (up to several hundred Mbytes) that it is impracticable to send complete models across the internet in an acceptable time. Communication between user and host will be from the user's desktop computer employing a graphical interface linked via the internet to the home of the database.

At the local level, the embryo models will be held on CD-ROM and used as an environment for laboratory analysis of experimental results as well as a context in which to construct submissions and queries to the database. The mouse atlas contains, for example, tools that will help users with relatively little anatomical experience to determine the developmental stage of their experimental material, to recognize the plane in which it has been sectioned, and to identify and name the principal anatomical components. As phenotypic information about each part of the embryo (for example, on cellular proliferation, apoptosis, etc.) is added to the atlas,[1] comparison of gene-expression domains with this information may also aid the interpretation of results. Operating at the local level, the database may also help users during the course of a project to organize and store their laboratory results and to share data between collaborating sites in order to rapidly communicate results and make joint interpretations on the basis of combined data.

The key function of the atlas, however, is to submit data and query the database. The MGEIR will hold both spatial and textual information (in an object-oriented database) and entries will be made via a dual text and image interface. At least two alternative methods will be available for entering spatially-mapped data into the database, manual segmentation (analogous to painting) and semi-automatic image transformation (a process by which original images of an in-situ gene-expression pattern are mapped directly onto the appropriate regions of the reference model). Both of these processes may be combined with textual input. Consider, for example, a gene expressed in the dorsal part of the neural tube. This pattern may first be entered as 'neural tube', calling up the already delineated anatomical domain; the pattern can then be restricted to the dorsal region by manual image segmentation. The data will be stored in text as expression in the neural tube, and spatially as the refined domain that occupies only part of the neural tube. The combination of text description, manual segmentation, and semi-automatic image transformation will be used to ensure that entering data and querying the database will be as user-friendly as possible (see, for example, Figure 5). Submissions to the database will be monitored jointly by editorial teams at the Jackson Laboratory and the MRC.

Queries will be made in the context of the MGEIR using a dual text and image interface. The local atlas, with its delineated anatomical domains, will function as the context for queries that take advantage of textual and spatial constraints and will be able to access data entered originally as either text or spatially-mapped images. It will, for example, be possible to ask *What genes expressed in the cerebellum have the same pattern of expression as gene X in the forebrain?, or What TGF-β-related genes are expressed at stage 15 in*

Figure 4. Selected anatomical components of the 9-day-old embryo displayed in 3-D mode using *AVS* software (http://www.avs.com/). The user can select which parts of the model embryo (complete with delineated anatomical components) are displayed. Here, only a single transverse section though the grey-level part of the image is displayed, together with selected anatomical components (neural tissue, gut, left part of the sinus venosus, and somites, all in the same colours as used in Figure 3D). In this way the precise spatial relationships between different domains in the embryo can readily be perceived.

Figure 5. Mapping gene-expression patterns onto a model embryo. (A) Gene-expression pattern (red) is projected onto an equivalent section in the model embryo. The expression does not match precisely the histology of the section in the standard model. (B) Here, semi-automatic image transformation has been used to improve the mapping by identifying equivalent histological features in the experimental and model sections: an image transformation is applied that matches the histological structure of the two (not shown); the same transformation is then applied to the gene-expression signal to leave only small discrepancies between the signal and the standard section. (C) The remaining discrepancies have been removed by manual image segmentation guided by examination of the experimental material under a microscope so as to match the mapped data as closely as possible to the observed pattern (a spurious digital signal resulting from a small piece of refractile material in the section has been removed). This illustration indicates the flexibility of data input methods as they will apply to the mouse database. The material illustrated here is the hindbrain region of a human embryo (work carried out in collaboration with Professor Tom Strachan, Department of Human Genetics, University of Newcastle-upon-Tyne).

mesenchyme within 200 μm of a defined region of the epithelium?

Progress to date

Thus far, the object-oriented anatomical database is complete to Theiler Stage 22; stages to birth (Stage 26) will be available later in 1998. We have reconstructed embryos at the 2-cell, 14-cell stages and at Theiler stages 9, 10, 12, 13 and 14 and we expect that a representative series of 3-D model embryos from fertilization to Theiler Stage 14 (E9), each with delineated anatomical domains, will be available on CD-ROM by the beginning of 1998. Later stages will be produced as they are completed. We hope to be able to produce the CD-ROM at a cost that every laboratory will be able to afford. The *MRC Graphical Gene Expression Database* will take longer to produce, but we expect that a basic test version will be available late in 1998, with a public database running before 2000.

Links to other data

Understanding gene function is, of course, more than describing expression patterns. In addition to the links between the gene-expression database and the MGD and mouse atlas, it will be important to establish links to data relating directly to gene function. In particular, it may be possible to establish links to databases that may be built in the future such as, for example, databases containing information from genetic and biochemical studies on the interactions between the products of different genes. It may also eventually be possible to link to databases documenting the development of mutant embryos (for example, TBASE) and to those for species other than mouse. One obvious possibility is to link to data from molecular genetic studies of human development;[9] but it may be possible to link data from more widely separated species such as zebrafish[12] and *Drosophila*.[13]

Data compatibility

To create links outside the MGEIR two things will be necessary, database interoperability (e.g. via CORBA interfaces[14]) and data compatibility, with this latter requirement being of particular importance. Where there are homologous structures and genes and the same techniques, etc. are used, then the same terms can clearly be used to refer to them in different databases. The use of a common vocabulary does, however, raise problems where homologies are debatable, and progress here will depend on concerted action across the community.

The purpose of all these databases is, of course, to facilitate communication across the world of developmental biology, and a key to the success of the enterprise is that those constructing these databases make linking them a priority. It is therefore important that forums be established where this can happen.

Acknowledgements

We thank Martin Ringwald, Janan Eppig and their colleagues at the Jackson Laboratory for many discussions. We are grateful to our colleagues Professor Tom Strachan and Dr Philip Bullen for allowing us to illustrate gene-expression mapping using images of their material. We express our gratitude to the European Science Foundation for their support of the Utrecht meeting. This provided a first opportunity for all those working on gene-expression databases to meet and to consider ways in which their separate databases might communicate.

References

1. Davidson D, Baldock RA, Bard JBL, Kaufman MH, Richardson JE, Eppig J, Ringwald M (1997) Gene-expression databases, in In situ hybridization. A practical approach (Wilkinson D, ed). IRL Press, Oxford in press
2. Ringwald M, Davis GL, Smith AG, Trepanier LE, Begley DA, Richardson JE, Eppig JT (1997) The mouse gene expression database GXD. Semin Cell Dev Biol 8:489-497
3. Cohn MJ, Tickle C (1996) Limbs: a model for pattern formation within the vertebrate body plan. Trends Genet 12:253-257
4. Lawrence PA, Struhl G (1996) Morphogens, compartments, and pattern: lessons from Drosophila?. Cell 85:951-961
5. Hammerschmidt M, Brook A, McMahon AP (1997) The world according to *hedgehog*. Trends Genet 13:14-21
6. Theiler K (1989) The House Mouse: Atlas of Embryonic Development. Springer, New York
7. Kaufman MH (1994) The Atlas of Mouse Development 2nd Ed. Academic Press, London
8. Baldock RA, Verbeek F, Vonesch J-L (1997) 3D reconstructions for graphical databases of gene expression. Semin Cell Dev Biol 8:499-507
9. Davidson D, Baldock RA (1997) A 3-D atlas and gene expression database of mouse development: implications for a database of human development, in Molecular Genetics of Early Human Development (Strachan T, Lindsay S, Wilson D, eds). BIOS Scientific Publishers Ltd, Oxford in press
10. Kaufman MH, Brune RM, Baldock RA, Bard JBL, Davidson D (1997) Computer-aided 3-D reconstruction of serially sectioned

mouse embryos: its use in integrating anatomical organization. Int J Dev Biol 41:223-233
11. Ringwald M, Baldock R, Bard J, Kaufman M, Eppig JT, Richardson JE, Nadeau JH, Davidson D (1994). Science 265:2033-2034
12. Westerfield M, Doerry E, Kirkpatrick AE, Driever W, Douglas SA (1997) An on-line database for zebrafish development and genetics research. Semin Cell Dev Biol 8:477-488
13. Janning W (1997) FlyView, a *Drosophila* image database, and other *Drosophila* databases. Semin Cell Dev Biol 8:469-475
14. CORBA Object Management Group http://www.omg.org/