Chapter 39
Biomedical Atlases: Systematics, Informatics and Analysis

Richard A. Baldock and Albert Burger

Abstract Biomedical imaging is ubiquitous in the Life Sciences. Technology advances, and the resulting multitude of imaging modalities, have led to a sharp rise in the quantity and quality of such images. In addition, computational models are increasingly used to study biological processes involving spatio-temporal changes from the cell to the organism level, e.g., the development of an embryo or the growth of a tumour, and models and images are extensively described in natural language, for example, in research publications and patient records. Together this leads to a major spatio-temporal data and model integration challenge. Biomedical atlases have emerged as a key technology in solving this integration problem. Such atlases typically include an image-based (2D and/or 3D) component as well as a conceptual representation (ontologies) of the organisms involved. In this chapter, we review the notion of atlases in the biomedical domain, how they can be created, how they provide an index to spatio-temporal experimental data, issues of atlas data integration and their use for the analysis of large volumes of biomedical data.

R.A. Baldock (✉)
MRC Human Genetics Unit, MRC Institute of Genetic and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK
e-mail: Richard.Baldock@hgu.mrc.ac.uk

A. Burger
MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK

Department of Computer Science, Heriot-Watt University, Edinburgh EH14 4AS, UK
e-mail: Albert.Burger@hgu.mrc.ac.uk; A.G.Burger@hw.ac.uk

I.I. Goryanin and A.B. Goryachev (eds.), *Advances in Systems Biology*, Advances in Experimental Medicine and Biology 736,
DOI 10.1007/978-1-4419-7210-1_39, © Springer Science+Business Media, LLC 2012

1 Introduction

Biomedical research has always relied on visual observation and imaging is a primary mechanism for recording data from the sub-cellular through to whole-organism level. In particular, imaging is used to capture the spatial organisation of biological entities, such as sub-cellular organelle and chromosomal organisation, cellular morphology, tissue organisation and organ histology and morphology. At the highest levels of resolution imaging is being used to capture molecular structures, synaptic organisation and molecular flux within the cytoplasm. Modern imaging techniques have been extended to capture 3D data not only at all ranges of resolution, but also to include the option of capture through time. Figure 39.1 shows a range of imaging modes that illustrate the nature of the spatio-temporal data produced for biomedical research.

In many cases image data are used to support simple observations. For example, gene X is expressed in the ventral half of the left ventricle, or the cells of the epithelial layer show an elongated appearance. As more data is collected, the trend is to use manual and automated means to extract numerical information from the

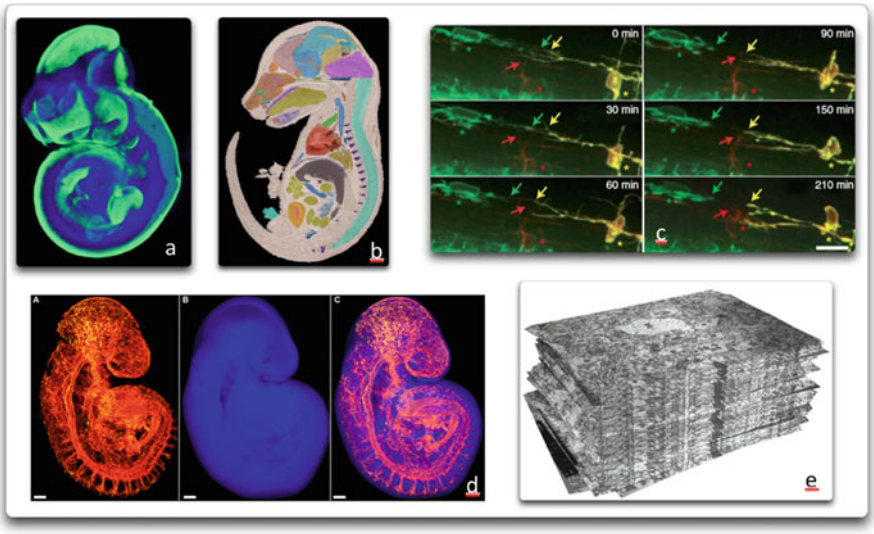

Fig. 39.1 (**a**) Optical Projection Tomography (OPT) image of mouse 10.5 dpc embryo in-situ hybridisation expression of Crabp1; (**b**) Caltech μMRI (Magnetic Resonance Imaging) image from the Caltech Mouse Atlas; (**c**) Time-lapse confocal images of oligodendrocyte development, adapted by permission from Macmillan Publishers Ltd., Nat Neurosci [21] copyright 2007; (**d**) Transgenic expression of vascular development in the mouse embryo [34]; (**e**) serial-section EM (Electron Microscopy) reconstruction of a neuropil structure courtesy of SynapseWeb, Kristen M. Harris, PI, http://synapses.clm.utexas.edu/

images to provide objective numerical analysis in terms of spatial patterns, signal intensities, shape and morphology, cell densities ablation recovery times, etc. As data is captured at higher rates and volumes, the requirements for image archiving and analysis are demanding far greater automation. The focus of this paper is image data captured to show information at the organ or whole-organism level of biological organisation. In our case this is with respect to embryo development and can include gene-expression patterns, lineage tracing, physiology and cellular activity, morphometric and mutant phenotype. At this level of biological organisation a key requirement is to be able to compare spatial and temporal patterning and to be able to collate information from across all the different imaging modalities. At the genomic and molecular biology level the natural framework for capturing data relationships is the genome, at the organ/organism level the appropriate framework is provided by explicit spatio-temporal atlases [9]. To some extent the spatial aspects of the information can be captured by annotation using an anatomy ontology, but this does not have the resolution or computational capability of an explicit coordinate framework provided by a digital atlas.

Atlases provide the integrating framework for spatial data of tissues, organs and whole organisms. For genomic and molecular level data, information between species can be compared by "mapping" of the sequence data. Such sequence mapping provides detailed comparative analysis of the evolution of the genome and enables the use of model organisms (e.g., mouse) to support research into human disease and abnormalities for translational purposes. By analogy the basic information captured at the whole-organ level can be compared across species including through to human for direct medical research and ultimately clinical application. If we take the "layer cake" view of biology passing from the lowest levels of organisation at the base through to tissue, organ and whole-organism level at the top, then the spatio-temporal data mapped to the atlases at the top serve as the target for a systems biology understanding of the high level biology. In addition, the basic research data captured, for example, for model organisms such as zebrafish and mouse, serve for comparative analysis and provide basic understanding to physiological and disease modelling applied for translational research into the human condition. This can extend through to medical and clinical data sets and ultimately through to individuals. At this end of the atlas range we envisage the use of a personal "myAtlas" to capture and record the clinical history of an individual and perhaps to support patient–doctor consultation. This view of the role of explicit spatio-temporal atlases in the context of biological and medical research and potentially clinical practice is illustrated in Fig. 39.2.

In this paper, we outline informatics aspects of atlasing frameworks in the context of biomedical research and illustrate these with procedures and examples from the eMouseAtlas project EMAP and EMAGE [4] (Edinburgh Mouse Atlas Project and Edinburgh Mouse Atlas Gene Expression database).

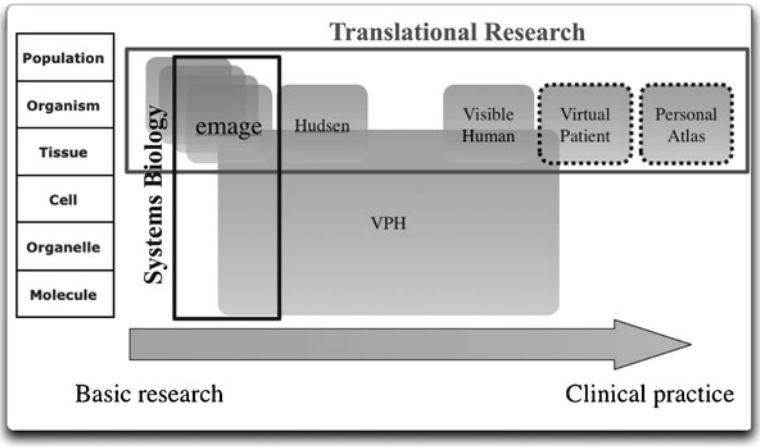

Fig. 39.2 Atlases in the context of systems biology and translation biomedical research. Hudsen is the human development atlas and data resource, the visible human indicates the adult level atlas and virtual patient and personal atlas indicate resources under development or envisaged. VPH refers to the international Virtual Physiological Human programme to develop computational and predictive models of adult human physiology

2 Atlas Systematics

In the scientific world, there often is a general understanding of the meaning of widely used key terms, such as, "gene" or "ontology", but a lack of agreement on their precise definition. This applies particularly to the use of the term "Atlas" in biomedical research. Here we develop a classification of resources that describe themselves as atlases and argue that a proper use of the term should imply an overt spatial representation used to express the spatial relationships in the data.

For most people the definition of an atlas relates back to the familiar geographic atlases and maps and is typically an overt depiction of a coordinate space, e.g., the surface of the earth. This is supplemented with the representation of features and regions which in the geographic example could be cities and countries. The Oxford English Dictionary (OED) defines the term *atlas*: *A collection of maps (or illustrative plates) in a volume*, where a *map* is defined as *A diagram or collection of data showing the spatial distribution of something or the relative positions of its components*. For us the equivalent of the collection of maps is a collection of 2D or 3D images, which define the space we need to represent for the mapping of data with spatial relationships. Some technologies, e.g., Optical Projection Tomography (OPT) [32] allow the generation of the 3D model directly, but from which 2D section images can be generated. Most atlases we know of use actual images, such as, generated by microscopy or MRI, instead of symbolic depictions (e.g., drawings). In either case, the visual representation in the form of sets of pixels and voxels, described in an appropriate coordinate framework, forms the first essential

component of our notion of a biomedical atlas. Although it is not stated so explicitly in the OED definition implying "spatial distribution" or "relative position" requires some sort of labelling or *mapping* of the artefacts in the context of the map. In geography we expect the regions of countries and cities on a map. Similarly, in the context of a biomedical atlas, we expect labels describing the components in the visual representation, e.g., the label "heart" refers to an image region depicting the heart in the image model. This implies that there is a mapping between the term and the image model.

The terms may simply consist of a controlled vocabulary such as the names of anatomical structures, or form a part of a formally specified ontology. This formalisation can be fairly lightweight, using languages such as SKOS [27], or rather detailed and precise, using languages such as OWL,[1] to describe it. The higher the level of formalisation, the more automated reasoning it will allow, but the more difficult it is to get widespread acceptance of the ontology as a standard within the biomedical community. This has implications for interoperability (see Sect. 5).
With this discussion we can identify components that could be part of an atlas:

Representation of space: a visual representation such as an image with the image coordinates allowing location of specific features or regions. For biomedical atlases this is typically a selected representative image or an averaged image over a number of samples.

Spatial reference terms: in biomedical atlases this is typically anatomy.
Mapping: locations or regions of the spatial reference terms in the context of the spatial representation.

Direction: definition of directions in the context of the underlying object. In a geographic atlas this is usually simple to identify North or to plot lines of latitude and longitude. In biomedicine it may require a much more complex mapping of left–right, dorsal–ventral and anterior–posterior axes at each location within the map.

Data: this is the association of data such as gene-expression or physiological state with different parts of the spatial representation.

Some "Atlas" resources only include the spatial representation in an implicit way by referring only to the anatomical terms. Examples of such atlases are the Human Protein Atlas [5] and Gene Expression Atlas [18], where data is annotated with anatomical tissue and cellular terms but there is no explicit spatial representation or mapping. All spatial association is via the spatial understanding of the user. At the other extreme are the full 3D, spatially mapped anatomical atlases, used as frameworks for capturing digital image data. Examples here are the mouse gene-expression databases eMouseAtlas [4, 7] and the Allen Brain Atlas [23]. Between these there are traditional "paper" atlases such as the Atlas of Mouse Development' by Kaufman [19] and those with digital content, e.g., the Paxinos Rat Brain Atlas [29].

[1] www.w3.org/TR/owl2-overview/.

In order to realise the power of a digital atlas to provide an objective spatial analysis and provide tools for spatial data mining an explicit spatial representation is essential, and we therefore define the minimal requirements for a biomedical atlas to be:

$$\text{Spatial Representation} + \text{Terms} + \text{Mappings} = \text{Atlas}$$

If an atlas also includes a specification of biological directions then more sophisticated query and analysis in biological terms becomes possible. The construction of biomedical atlases, the use of atlases to index spatio-temporal experimental data, the integration across atlases and other resources, and the use of atlases in the context of the analysis of large quantities of biomedical data are discussed in the following sections.

3 Atlas Construction and Spatial Annotation

Biomedical atlases that include an explicit coordinate framework can be constructed in many ways, including "simple" graphical modelling to depict the primary structures that are to be represented. In practice, atlases developed for biomedical research are typically based on one or more representative individuals using imaging that enables full 3D reconstruction. This can be a direct 3D imaging technique, such as, μMRI [11], μCT [1] (Computed Tomography), block-face imaging [35, 36] or OPT [32] or, if resolution and contrast are critical, then 2D imaging of microtome sections followed by 3D reconstruction. When the key requirement is to be able to capture spatial patterns for subsequent comparison and analysis, for example anatomical labelling or syn-expression grouping, then it is sufficient for the atlas to be a *representative* individual. Such an atlas can also be used to capture morphological variation of experimental sets by capturing both the mapped data and the spatial transformation from which variation in the original data set can be established [8]. If, however, the key purpose is to be able to assess the morphology of a new sample, it is more convenient to create a probabilistic atlas [13, 25].

For the mouse embryo models of the eMouseAtlas database we have used a combination of OPT 3D imaging to capture the original shape of the embryo followed by wax-embedding and microtome sectioning so that the individual section could be stained to reveal the cellular detail. These histological sections are imaged at high-resolution and reconstructed using the OPT image as a morphological template. When the 3D model has been reconstructed, it is then segmented into anatomical regions which provide the link between the spatial representation of the embryo (image coordinates) and the anatomy ontology. This process is illustrated in Fig. 39.3. The embryo atlas we have developed in Edinburgh is comprised of a series of 3D reconstructions, an anatomy ontology to describe the developing anatomy, plus a set of delineated regions or domains that link the ontology terms (at some level of resolution) to the 3D image model.

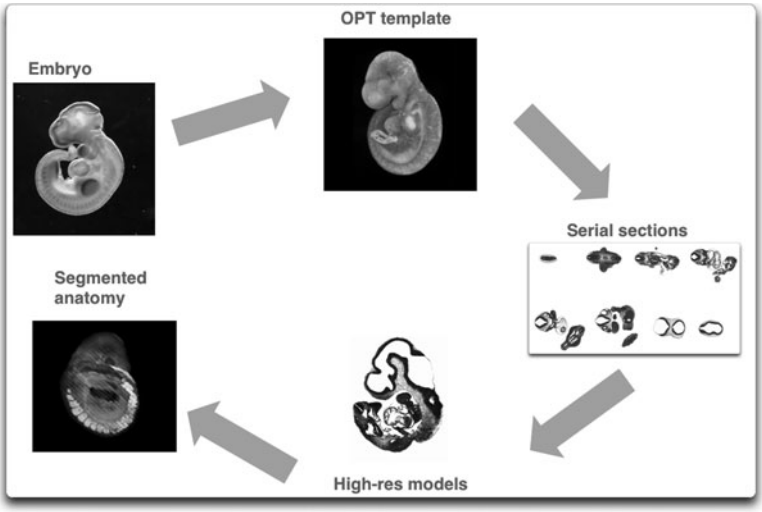

Fig. 39.3 Reconstruction process used to build the high-resolution 3D models of the mouse embryo for EMAP

The 3D image can be used directly as an atlas framework. In some cases it is possible to supplement the image coordinate frame with more biologically meaningful coordinates such as the stereotaxic coordinates used in neuroscience studies of the brain [16]. This is not always required and the key requirement for an atlas to be useful is a mechanism by which data can be spatially transformed or *mapped* into the atlas space. This is termed spatial registration or spatial mapping and in general is a complex non-linear transformation from the original (source) coordinate frame to the atlas coordinate frame (target). Image registration has been studied very thoroughly, especially for clinical imaging where comparison between modalities and for disease progression are important. Techniques that have been established typically define a deformation field across the volume of image space enclosing the source image of interest. This deformation field is established by manual definition of points of correspondence or automation and the full field defined via a mathematical model such as radial-basis function interpolation or physical modelling of the deformation. In either case we have found that the embryo presents special problems because of the extreme deformations that arise due to the flexibility and variability of presentation and pose. In this situation the standard warping techniques fail and we therefore have established warping based on the constrained distance transform [17], which is a combination of rapid manual alignment to correct the primary deformation due to pose followed by an automated process using the open-source software ANTs [2] to fine-tune the alignment. The WlzWarp software tool we use for the manual registration is illustrated in Fig. 39.4.

The procedure to transform experimental samples into the model space can be considered as *spatial annotation*. Each location within the sample acquires a

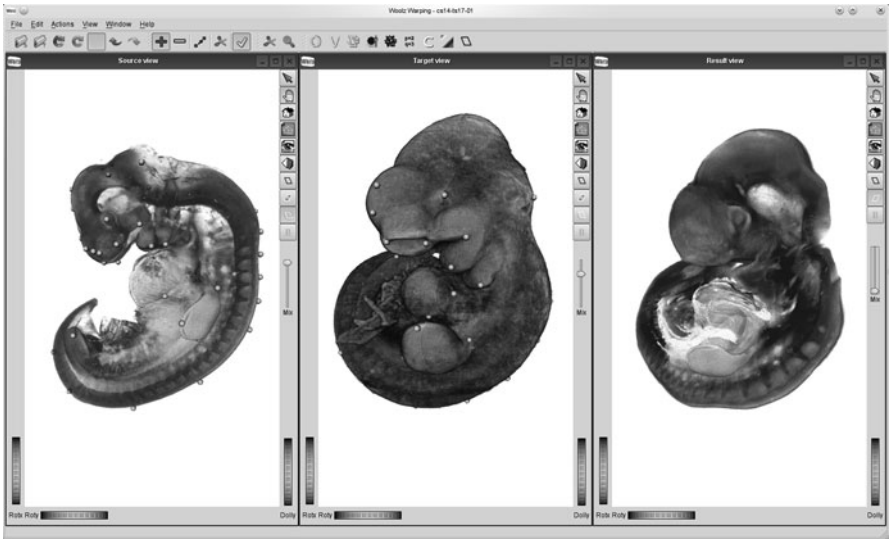

Fig. 39.4 Spatial mapping of a 3D image of a human Carnegie stage 14 embryo onto the EMAP Theiler stage 17 embryo model using the WlzWarp software tool developed to deal with the complex mappings required with embryos. *Left-hand frame* original human embryo; *middle frame* the target mouse embryo; *right-hand frame* the warped human embryo to match the mouse. The marked locations show locations of point-correspondences (Note: Carnegie and Theiler stages are classification systems for how far human and mouse embryos, respectively, have progressed in their development.)

mapping into the model. In this way, data from the model can be presented in the context of the sample, or data from the sample such as gene-expression signal intensity, can be transformed into the space of the atlas models. It is then possible to analyse the data either in terms of the atlas, e.g., to establish anatomical regions that show gene-expression, or to compare with other data directly, such as, other gene-expression patterns. In analogous fashion to a text-based annotation, spatial annotation enables search for patterns but directly in terms of the atlas space, e.g., queries, such as, "what genes are expressed at this locations?" or "what gene show expression in a similar pattern?" are now possible.

In the mouse atlas EMAP and associated gene-expression database EMAGE [7] the spatial annotation is a standardised procedure to map the source image, and to segment the signal into pre-defined strengths of expression.[2] The mapped signal patterns are held in the database and a query against the database results in direct comparison of image data. This is an image-processing operation and executed in

[2] http://www.emouseatlas.org/emage/about/data_annotation_methods.html.

an image server linked to the RDBMS (Relational Data Base Management System) which manages the metadata. For efficiency the spatial indexing and similarity calculations are encoded using the Woolz image-processing library.

4 Experimental Data and Atlases

Atlas frameworks can be used to capture, compare and analyse any spatial data, which can range from cellular signalling and gene-expression patterns through clonal distributions to long-range neuronal connectivity and physiological function. Here we will use the data captured in the context of the eMouseAtlas models to illustrate the issues of mapping and interoperability of atlas-based resources. The primary data for which the mouse embryo atlas was established is gene-expression patterns as revealed by in situ hybridisation to mRNA and immunohistochemistry with protein antibodies. In addition, we have mapped anatomy terms to the 3D space and explored direct mapping of cellular clonal data following lineage tracing experiments.

4.1 In situ Data

Transforming a gene-expression pattern from an in situ experiment involves two steps. The first is to establish the spatial transformation from the experimental data images to the atlas model. The second is to segment the signal in the context of the original data and to use the spatial mapping to transform the pattern to the atlas model context. Our experience with mouse embryo data indicates we need to deal with a number of different presentations of the information:

2D data: Intrinsically 2D data captured from the embryo. The prime example is a whole-mount view which is effectively a projection of the underlying 3D data onto 2D and in principle the original 3D location of the signal cannot be recovered. For this data we have adopted a simple approach of mapping the data to a projection of the atlas model, i.e., maintain the 2D character of the original data. Within EMAGE this implies that the data is segregated and a spatial query is currently against either the wholemeal data or the 3D data.

2D images of 3D data: These are microtome sections of the sample embryo which has been physically sectioned and stained. In principle, the section image can be mapped back into a 3D location within the atlas model. In practice this can be difficult, because distortion of the tissue section could mimic a re-location of the image within the 3D framework with a consequential ambiguity. This can be resolved by capturing more than one section and using the adjacent data to correctly align the sections that have been treated to show the in situ pattern. Most high-throughput data, such as generated for the Allen Brain Atlas [23] and Unexpress [12], is of this form – sparse 3D data.

In EMAGE we have adopted two strategies for the data. The first is simply to find the best matching section for each sample and to use a mapping tool such as Maxint to transform (warp) the image onto the atlas. The same tool then allows a segmentation of the signal into a series of domains to represent *strong*, *moderate*, *weak*, *possible* and *not-detected* expression strengths. An additional domain *not-observed* ensures that a null-return from the data base can distinguish data that shows no-detectable expression from no-data.

A second strategy is to project the 3D data onto 2D and to treat it in a similar fashion to wholemeal samples. This of course loses the 3D information and reduces the ability to discriminate patterns, but can be useful for a first-pass automated mapping to be followed up with a more detailed 3D mapping later.

Full 3D data: This is data from a 3D imaging technique such as OPT or confocal LAM (Laser Scanning Microscopy) or could be serial sectioning that can be reconstructed to a full 3D data set. This type of data provides the most complete view of the overall expression pattern, but is typically at a lower resolution and does not deliver the cellular detail of real histological sections. A benefit is that the process of 3D mapping is very much faster than the section-by-section mapping of sparse data, but does require sophisticated mapping tools such as the WlzWarp tool based on the constrained distance transform and potentially significant compute power for the automated fine-alignment phase.

4.2 Sparse Cell Data

In some experiments the observation may be a set of cells that exhibit a particular stain. A particular instance of this is a clonal set of cells arising from a single progenitor cell. This could be marking using a vital dye [22] or by a random recombination event in a tamoxifen-inducible cre-transgenic line [24]. The issue with this data is that the individual cells that comprise the clones cannot in general be identified in the target model. The mechanism for mapping is therefore to map by direct marking of the estimated best match for each clone cell. With serial section data this can be very time-consuming. This could be done by direct matching of a serial section series which encompasses the clone, but this is similarly time-consuming.

4.3 Anatomy and Physiology

Classically an atlas depicts the physical geography overlaid with coloured regions depicting countries and national boundaries. In biological atlases the closest analogy is anatomy overlaid on a histological image and rather like the geographic case the "country" boundaries are subject to disagreement and dispute! In the EMAP mouse

Fig. 39.5 The EMAP anatomy browser. The user can select arbitrary section views through the 3D model and show selected anatomical components overlaid on the histology. In this case we are showing the limb atlas material from DeLaurier et al. [10]

atlas the anatomy delineations are available for download and can be visualised in a number of applications. Figure 39.5 shows a screenshot of the anatomy viewer provided for visualisation in the context of a standard web-browser. In this case we are showing a view through the limb atlas of DeLaurier et al. [10].

The atlas can of course also capture physiological data such as calcium concentration and ion-channel status in the heart or functional imaging of the brain. This type of data will clearly include detailed temporal and behavioural information, but the spatial aspects of the observations can be mapped to the atlas and compared with other data. An example we have been exploring is discussed in the next section; it is the use of the atlas approach to integrate a detailed physiological model of the heart with a statistical model of dynamic heart morphology over the cardiac cycle. The basic concept is to map both models to a common atlas model which can then also bring in other data from for example the EMAGE gene-expression database into the same analysis.

5 Integration of Biomedical Atlases

Computational frameworks, such as the atlases described in this paper, are in the first instance mechanisms for the management of data and knowledge, initially simply to capture and store it, but subsequently also to query it and to perform complex analysis studies (see Sect. 6). Typically, we find more than one atlas covering the same or related data, and we usually want to link data in an atlas to that in a

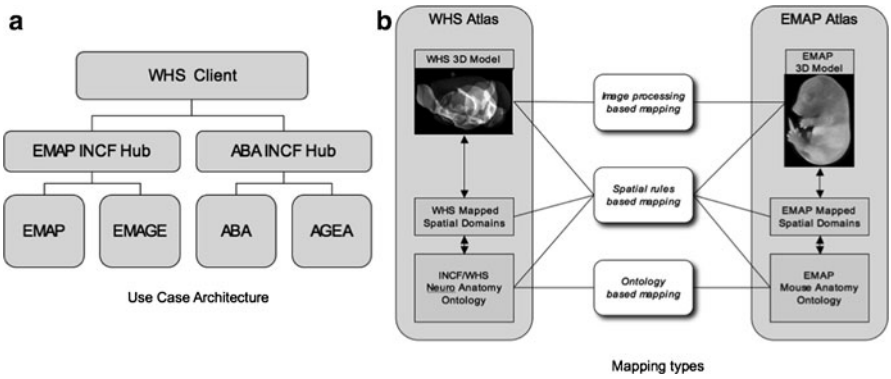

Fig. 39.6 (**a**) For the gene expression use case, a client application specifies a point in Waxholm Space (WHS) in order to access relevant gene-expression data in EMAGE and AGEA, (**b**) mapping between atlases can be achieved by applying image-processing techniques, ontology-based mappings and the specification of locations using spatial rules which are based on the 3D models as well as the ontological description of the atlas anatomies

non-atlas resource, e.g., entries for gene-expression experiments in EMAGE have links to Ensemble (www.ensemble.org) for further information about the gene under consideration. All this creates a challenging integration problem for biomedical atlases. As always, the desired interoperability between atlas and related resources depends to a large extent on agreed standards. In this section, we illustrate these interoperability issues, drawing on our experience of a use case study linking the EMAGE data set to the Allen Brain Atlas [28] using the emerging Waxholm Space standard [16] which is a 3D reference atlas for the adult mouse brain.

The basic architecture for this use case is shown in Fig. 39.6a. It is based on the INCF Digital Atlas Infrastructure (INCF-DAI), which is currently being developed by the INCF (International Neuroinformatics Coordination Facility, www.incf.org). The INCF hubs for EMAP and ABA (Allen Brain Atlas) are responsible for mapping the point of interest in Waxholm Space (WHS) into corresponding locations in their respective atlas spaces – an alternative approach where a central spatial transformation INCF hub will assume responsibility for all spatial transformations is being considered – and then retrieve the relevant gene-expression data for return to the client. At this stage, the hubs simply return URLs to html pages containing the gene-expression query results. The client displays these in two separate browser windows, but does not merge the results. To facilitate the latter, a standard for gene-expression query results needs to be agreed first. This is a key point, as it applies to many different types of data. Achieving interoperability between different spatio-temporal reference frameworks does not guarantee the interoperability of the data that is indexed by these frameworks. Standards, such as for gene-expression data, are required in addition, if the analysis of the data across multiple atlases is to be maximally supported by software.

As discussed in Sect. 2, although there is no single definition of what biomedical atlases should consist of, it is usually the case that they have an image component as

well as textual labels for identifiable regions of the image space. In some cases the textual labels are part of comprehensive anatomy ontologies. This leads us to the following three spatial mapping types: (1) based on image processing, (2) based on ontology mapping and (3) based on spatial rules; see Fig. 39.6b for an overview diagram in the context of our use case. The first of these uses image-processing algorithms to transform pixels and voxels from one space to another. In our examples we use a constrained distance transform to link between the WHS atlas and the EMAP atlas spaces. The second type is based on mapping anatomical concepts from one ontology to another, e.g., the concept `Cerebellum` in the ABA maps to the `Cerebellum` in EMAP. The third type uses spatial relations, such as, `contained_in` and `next_to`, known about identified regions in the atlas to describe a spatial location. Whilst the image-processing solution can potentially achieve very good accuracy, it does so only for atlases that are morphologically not too different. Ontology-based mappings deal with such differences more easily, but do not achieve the same level of precision. The use of spatial rules is a compromise solution that aims at reasonable accuracy in spite of some morphological differences.

We know that the level of formalisation of the terms used by atlases has a significant impact on their interoperability. In principle, a more detailed ontology leads to better integration possibilities, but only if this ontology is widely shared and used by the biomedical atlasing community. Herein, however, lies the dilemma, since the more detail one specifies in the ontology, the more difficult it becomes to obtain community acceptance. There exists, of course, a large body of work on the topic of biomedical ontologies, and a detailed discussion of it lies outside the scope of this paper. For a collection of relevant papers, we refer the reader to [6]. An area of biomedical ontologies that has not been explored very much thus far is their use in the context of spatial rules-based mappings, which will require, amongst other things, some level of standardisation of the meaning of directional terms, such as, “lateral to”, “close to”, etc. The challenge in biomedical atlasing is the lack of a single frame of reference, such as is available in the geo-sciences; there is only one Earth, but there are many instances of organisms such as human and mouse.

It is important to remember that in the context of integrated spatial queries, several mappings across different spaces may be involved. Figure 39.7 illustrates how we distinguish between four categories of spaces. Initially, experiments, such as for in-situ hybridisation gene-expression analysis, are carried out in the context of specific animal experiments resulting in 2D and 3D image data for their particular *experiments space*. These results are typically mapped into a standard spatial or spatio-temporal repository framework, the *respository space*, such as EMAP, through which they can then be queried. To integrate across two or more repositories may involve a *mediator space* such as Waxholm Space (WHS), and if the data that has triggered the original query of interest is based on a particular experiment, we require a mapping from the *query space* to the mediator space. Where labs produce their own data for their reference space, the distinction between repository and experiments space may not explicitly exist. So, an integrated query to EMAGE and AGEA brain gene-expression data, using WHS as the mediator, would involve at least 4 spatial mappings and potentially all three mapping types.

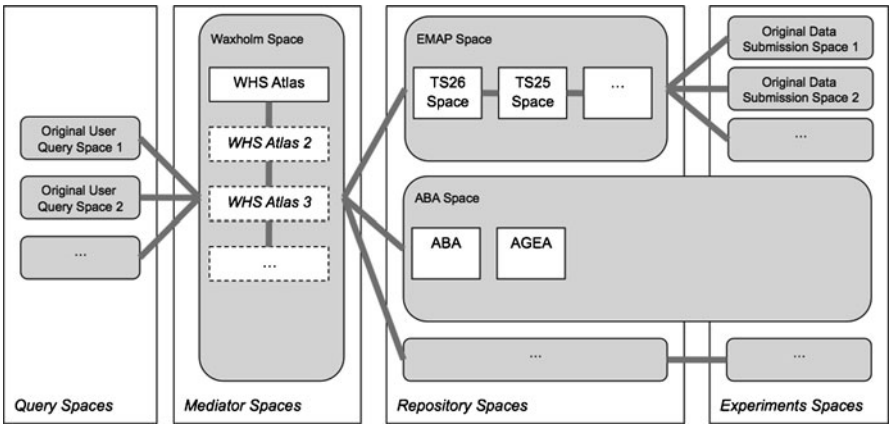

Fig. 39.7 Types of spaces

As the number of mappings across spaces increases, the accuracy of the results for a query is likely to diminish. Intuitively, one might argue to expect the overall accuracy to be determined by the "weakest link", i.e., the least accurate spatial mapping involved in the query. However, there may also be an accumulative effect resulting in even worse accuracy. There is also an issue of giving an end user the impression of high precision, for example, because his/her query space was very carefully mapped into the mediator space, but that the actual results are by far less accurate due to other mappings involved.

The above discussion has focused on the integration of atlases as spatio-temporal frameworks for experimental data, but as more and more such data becomes available, we also see an increase in computational models which firstly help explain the underlying biomedical processes resulting in this data, and secondly include predictive capabilities for scenarios that have not yet been studied experimentally. The integration of "data atlases" with the spatio-temporal frameworks of computational models is critical for the development and calibration, as well as the validation and verification, of the models. As part of the European Union's *Virtual Physiological Human* (VPH) research programme (www.vph-noe.eu), the RICORDO project (www.ricordo.eu) investigates this data model integration for volumetric data. Amongst other work, it has developed a spatial mapping from the computational heart developed at the University of Auckland to the EMAP atlas in Edinburgh. To the best of our knowledge this is the first example of mapping volumetric, computational VPH model sections to the corresponding location in a 3D framework for molecular data (gene expression). Although it is outwith the scope of this paper to discuss the technical details of this mapping, it illustrates one example of this extended requirement for atlas integration. Based on the increasing amounts of experimental data and related models, we predict that the importance of this type of integration will significantly increase over the next five years.

6 Using Atlases for Data Analysis

Atlas frameworks provide a straightforward context for spatial comparison and analysis. The types of analysis depend on the nature of the data collected and can be characterised by the nature of the input data and the output results. For example, if the input is a point or region defined within the atlas space, the result could be atlas-based, such as an image of the overall gene-expression intensity, or a numeric result, such as the similarity to another expression pattern, or just a list of assay results that match the query. Similarly the input could be a list of genes for which co-expression hot-spots are required in which case the output would be a heat-map type image with a gene-list associated with each point. In this section we illustrate the use of atlases for data analysis in the context of the embryo and atlas databases that we have integrated with the eMouseAtlas resource. These include the human embryo atlas and database Hudsen [20], GUDMAP [26], EurExpress [12] and Chick Atlas[3] databases.

6.1 Annotation and Query

Mapping data onto the atlas framework provides a means to specify a query on the database in graphical terms. This could be as simple as a single vertex or as an arbitrary point-set representing a complex region within the domain of the atlas model. In addition, an atlas within which the anatomical tissues have been delineated makes it possible to query using the anatomical terms. These provide two simple examples of data analysis. The first is annotation. By mapping the expression pattern onto the atlas model and comparing the mapped pattern with the anatomical domains delineated within the model, it is possible to generate an anatomical description of the gene-expression pattern. This is illustrated in Fig. 39.8 in the context of the EMAGE database. As well as establishing the list of tissues that show gene-expression, it is possible to calculate the relative proportion of each tissue that shows expression.

The second example is to use the spatial location or region as a means of finding genes expressed at the given location or area of interest. To process this query the given location is compared to each stored pattern in the database to establish if it is contained within the mapped region. In this case the spatial "index" of a mapped gene-expression pattern is represented internally as an image region or binary image. The query region as a single point or a second image region is compared with the expression pattern using a simple image operation of domain-intersection. This is equivalent to the intersection of two point-sets, but executed as an efficient image-processing algorithm. If the resulting intersection domain is non-empty then

[3] http://www.echickatlas.org/.

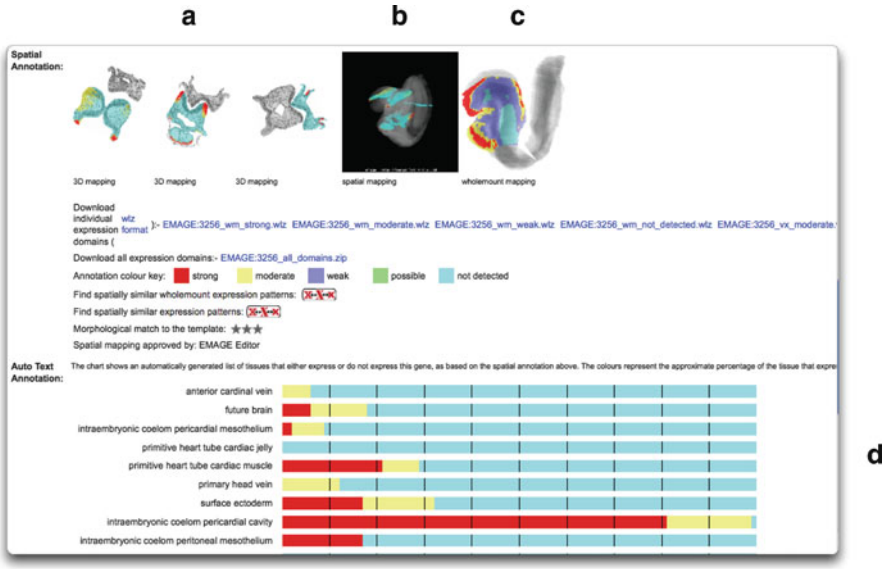

Fig. 39.8 Spatial analysis. The mapped expression pattern for the gene Bone morphogenic protein 5 (Bmp5) is mapped onto the Theiler stage 12 embryo atlas (**a**). The mapped section data is shown in 3D (**b**) and in the context of the wholemount embryo (**c**). The bar chart (**d**) shows the expression analysis in the terms of anatomical tissues that is automatically generated by comparing the expression domain with the delineated anatomy domains

the two patterns overlap. This is repeated for each pattern that could form a match. The result in the context of EMAGE is a list of assays that show overlap with the query region (see Fig. 39.9).

6.2 Similarity and Correlation

In addition to using the pattern to simply test spatial overlap or containment, the patterns can be compared for spatial similarity. For potentially dispersed and non-contiguous patterns we have discovered that the Jaccard index, which is a simple set-based measure of similarity, provides a suitable first-pass numerical value of similarity which is robust to the variation and noise found in typical gene-expression patterns. Here it is implemented in the context of a spatial region of interest defined by dilating the query pattern by the equivalent of about 300 μm. The tool is the LOcal Spatial Similarity Search Tool (LOSSST) and is described by Venkataraman et al. [33]. Figure 39.9 shows the result of using LOSSST to query the EMAGE database using the expression pattern of Bmp5 at Theiler stage 12. Using the option of mapping the query region through to other embryonic stages, it is possible to extend the query to return temporal data.

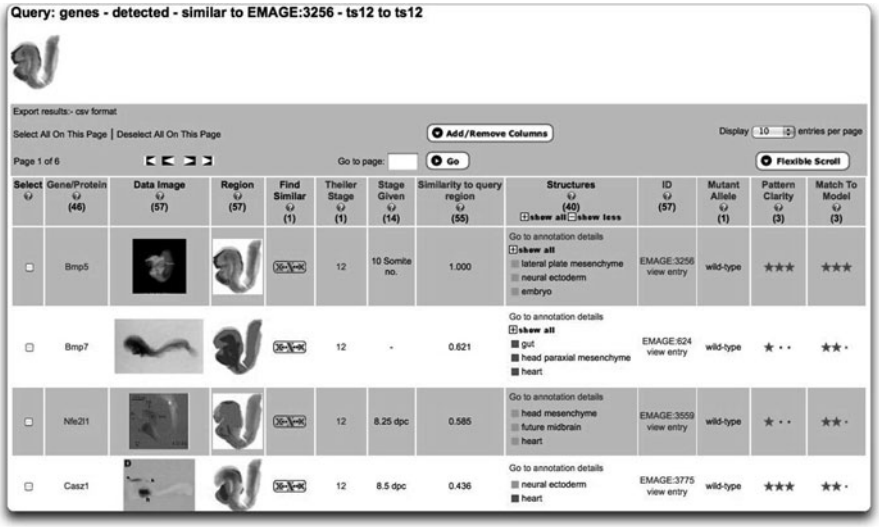

Fig. 39.9 Result of a spatial query on the EMAGE gene-expression database. In this screen shot the data has been sorted by similarity with the expression of Bmp5. The Bmp5 pattern is returned at the top if the list with a maximum similarity match of 1.0, the next most similar if Bmp7 from the same gene-family. Note real interface uses colour to show pattern strength

The use of similarity provides a sorted query result bringing to the top syn-expression patterns for any given gene. The same query can be posed on text-annotated data such as in the EurExpress database. The two annotation options are complementary. Text annotation can provide a more accurate and focussed return for very sparse and isolated tissue and cell-type specific expression patterns. Spatial annotation delivers accurate analysis of more complex and in particular near ubiquitous, but non-uniform, expression patterns.

A second measure that becomes available with spatially mapped data is expression correlation. With spatial similarity the query is to find genes with similar spatial expression patterns. It is also possible to test the expression correlation between spatial locations, i.e., to query for similar expression profiles between different locations. A good example of this is the interface provided by the Allen Brain Atlas [28] (see Fig. 39.10). For a given seed point, selected by "clicking" the required point on the screen, the system returns a correlation map which of course will have value 1 at the seed and typically includes the local region. Regions that have similar expression signatures are not always spatially connected and this may well indicate similar function or similar developmental lineage.

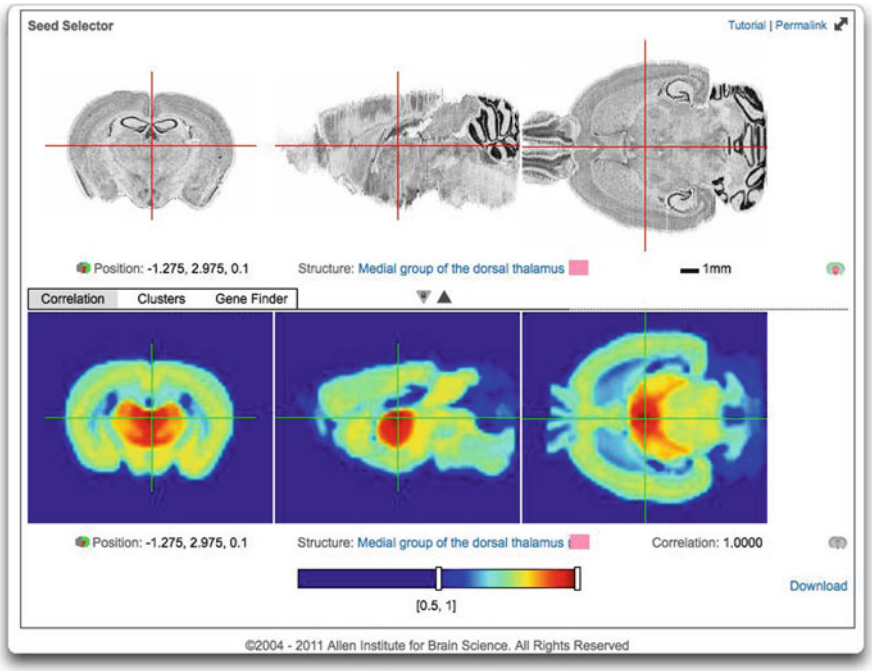

Fig. 39.10 Allen Brain atlas AGEA interface showing the correlation map for gene-expression with respect to the selected seed point

6.3 Data Mining

Atlas-based data with a mapping either onto the spatial framework of the atlas or the ontological framework, such as anatomy, becomes accessible for data mining. The simplest data mining approach is clustering based on a measure of spatial similarity or annotation. An example of this clustering is provided by the EurExpress data set [12], and the downloaded data can be visualized using standard cluster viewing packages, such as, TreeView and MeV. The results can be displayed in two ways: the first is as a set of gene-expression patterns that show similar spatial distributions and the orthogonal clustering will reveal the set of atlas regions that show similar expression profiles. These are related to the search options described above, but are not directed, and therefore provide a more objective overview of the structures implicit in the data.

Data mining can also be used to extract more detailed information from the data by using one set of data to train a classifier that can then be used to infer new relationships with a measure of confidence. An example here is the automated annotation of gene-expression patterns with anatomical terms by using an annotated set of images to train a classifier which can then be applied to new image samples. Han et al. [15] use this approach to develop tissue classifiers in the context of the EurExpress gene-expression data set.

6.4 Morphometric Variation

Atlases provide a natural framework in which to capture spatial patterns, such as, gene-expression, cell morphologies and behaviour. They can also be used to capture morphological variation even though the atlas may not be an "average" or "correct" model in the sense of representing the average size and shape for a given stage of development. In fact, in the context of mouse development a standard embryo is very difficult to define given the dynamic nature and heterochronicity of development even for pure strains. Nevertheless, if an experiment collects a standardised set of embryos with a protocol that will preserve size and shape, then morphological variation can be captured. The key to understanding this is that the data set that needs to be preserved is the non-linear mapping from each experimental instance as well as any data that may be mapped. If the mapping structures are available, i.e., the detail of how each point in the source experimental embryo is mapped, then it is straightforward to establish the average mapping from the experimental set to the atlas, and by applying the inverse of this transform to the atlas, an average embryo can be established. The meaning of this is that for each point within the average the sum of the displacements from each of the source experimental embryos will be zero. With this average in place other quantities that relate to morphological variation can be established. For example, the mean size and shapes of any given structure, say the heart, are the transformed version of the same structure in the original atlas. To establish the variation of any given feature, it is simply a matter of defining that feature in the atlas, e.g., the volume of the ventricular space in the brain, applying the inverse transform to the original sample and then re-measure the volume.

In addition, spatial patterns of variation, such as parts of tissues, that exhibit most volume variation can be displayed as a heat map in the context of the atlas and overlaid with other information. In this way it may be possible to associate morphological variation with gene-expression. Cleary et al. [8] showed how μMRI can be used to capture this type of data for late stage mouse embryos. Their methodology would not work for earlier stages of development because it lacks resolution and tissue contrast, but the principle is clear. Atlases deliver the necessary framework to associate complex morphological variation with other patterns and phenotypes in the biology.

7 Discussion and Conclusion

In the context of the developing mouse embryo, we have illustrated aspects of a new "bioinformatics" that can capture and manage data associated with higher levels of biological organisation. This approach in biology was pioneered by the Edinburgh Mouse Atlas Project (EMAP) [3,30] and demonstrates the use of explicit biomedical atlases to collate, compare and analyse spatially organised data. For the associated computer science infrastructure we coin the term "atlas informatics"

which covers the underlying theory and practice of using spatio-temporal atlases as the organising framework for spatial data. The geo-sciences have been working with geographic information systems for many years, but biology and medicine require significant extension of the techniques and capabilities, because of the variability of the underlying data sources and the complexity of the structures.

It is clear that atlases can provide the key integrating framework for data associated with individual model organisms and with the development of the methods and services for atlas integration many of the aspects of comparative analysis that are taken for granted at the genomic level become possible at the tissue and whole-organism level. This can be based on "simple" image-based mapping but a much richer semantic mapping is possible by using the underlying biology to define spatial location and direction. Developing this atlas semantic context and the logic and algebra that can use these *natural coordinate* systems is an immediate challenge for atlas informatics.

Finally, in a special issue of Science[4] dedicated to the "data deluge" a paper entitled "The disappearing third dimension", Rowe and Frank [31] discuss the difficulties of publishing 3D data, citing examples of tissue and palaeontology samples which may be difficult to replicate. They compare the image context with genomics which has a natural framework on which to associate data where re-use of experimental data is the norm. They conclude:

> Funding agencies can rejoice in the unexpected longevity and growing value in voxels they have already produced. But they must first secure the basic tenet of science by ensuring that researchers have the means to archive, disclose, validate and re-purpose their primary data.

Image repositories such as the Open Microscopy Environment [14] are essential to address part of this problem, but to retrieve, compare and analyse "re-purposed" data, a spatial framework is required, which is the role of atlases. Atlas frameworks are the key component of any informatics strategy to manage and analyse such data. In this paper we have illustrated some of the atlas informatics issues in the context of the mouse embryo but the underlying informatics model applies across biological, medical and natural sciences.

References

1. Aggarwal M, Zhang J, Miller MI, Sidman RL, Mori S (2009) Magnetic resonance imaging and micro-computed tomography combined atlas of developing and adult mouse brains for stereotaxic surgery. Neuroscience 162(4):1339–1350
2. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54(3): 2033–2044
3. Baldock RA, Bard J, Kaufman MH, Davidson D (1992) A real mouse for your computer. BioEssays 14:501–502

[4]Science 6 March 2009.

4. Baldock RA, Bard JBL, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, Sharpe J, Ross A, Stevenson P, Venkataraman S, Waterhouse A, Yang Y, Davidson DR (2003) EMAP and EMAGE: a framework for understanding spatially organized data. Neuroinformatics 1(4):309–325
5. Berglund L, Björling E, Oksvold P, Fagerberg L, Asplund A, Szigyarto CA-K, Persson A, Ottosson J, Wernérus H, Nilsson P, Lundberg E, Sivertsson A, Navani S, Wester K, Kampf C, Hober S, Pontén F, Uhlén M (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies. Mol Cell Proteom: MCP 7(10):2019–2027
6. Burger A, Davidson D, Baldock R (eds) (2008) Anatomy ontologies for bioinformatics: principles and practice. Springer, Dordrecht, The Netherlands
7. Christiansen JH, Yang Y, Venkataraman S, Richardson L, Stevenson P, Burton N, Baldock RA, Davidson DR (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. Nucl Acids Res 34(Database issue):D637–D641
8. Cleary JO, Modat M, Norris FC, Price AN, Jayakody SA, Martinez-Barbera JP, Greene NDE, Hawkes DJ, Ordidge RJ, Scambler PJ, Ourselin S, Lythgoe MF (2011) Magnetic resonance virtual histology for embryos: 3D atlases for automated high-throughput phenotyping. NeuroImage 54(2):769–778
9. Davidson D, Baldock R (2001) Bioinformatics beyond sequence: mapping gene function in the embryo. Nat Rev Genet 2:409–418
10. Delaurier A, Burton N, Bennett M, Baldock R, Davidson D, Mohun TJ, Logan MP (2008) The mouse limb anatomy atlas: an interactive 3D tool for studying embryonic limb patterning. BMC Dev Biol 8:83
11. Dhenain M, Ruffins SW, Jacobs RE (2001) Three-dimensional digital mouse atlas using high-resolution MRI. Dev Biol 232(2):458–470
12. Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, Magen A, Canidio E, Pagani M, Peluso I, Lin-Marq N, Koch M, Bilio M, Cantiello I, Verde R, Masi CD, Bianchi SA, Cicchini J, Perroud E, Mehmeti S, Dagand E, Schrinner S, Nürnberger A, Schmidt K, Metz K, Zwingmann C, Brieske N, Springer C, Hernandez AM, Herzog S, Grabbe F, Sieverding C, Fischer B, Schrader K, Brockmeyer M, Dettmer S, Helbig C, Alunni V, Battaini M-A, Mura C, Henrichsen CN, Garcia-Lopez R, Echevarria D, Puelles E, Garcia-Calero E, Kruse S, Uhr M, Kauck C, Feng G, Milyaev N, Ong CK, Kumar L, Lam MS, Semple CA, Gyenesei A, Mundlos S, Radelof U, Lehrach H, Sarmientos P, Reymond A, Davidson DR, Dollé P, Antonarakis SE, Yaspo M-L, Martinez S, Baldock RA, Eichele G, Ballabio A (2011) A high-resolution anatomical atlas of the transcriptome in the mouse embryo. PLoS Biol 9(1):e1000582
13. Duchateau N, Craene MD, Piella G, Silva E, Doltra A, Sitges M, Bijnens BH, Frangi AF (2011) A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. Med Image Anal 15(3):316–328
14. Goldberg IG, Allan C, Burel J-M, Creager D, Falconi A, Hochheiser H, Johnston J, Mellen J, Sorger PK, Swedlow JR (2005) The open microscopy environment (OME) data model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biol 6(5):R47
15. Han L, van Hemert JI, Baldock RA (2011) Automatically identifying and annotating mouse embryo gene expression patterns. Bioinformatics (Oxford, England) 27(8):1101–1107
16. Hawrylycz M, Baldock RA, Burger A, Hashikawa T, Johnson GA, Martone M, Ng L, Lau C, Larson SD, Larsen SD, Nissanov J, Puelles L, Ruffins S, Verbeek F, Zaslavsky I, Boline J (2011) Digital atlasing and standardization in the mouse brain. PLoS Comput Biol 7(2):e1001065
17. Hill B, Baldock RA (2006) The constrained distance transform: interactive atlas registration with large deformations through constrained distances. In: Proc DEFORM'06 – workshop on image registration in deformable environments, MRC Human Genetics Unit.
18. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A (2010) Gene expression atlas at the European bioinformatics institute. Nucl Acids Res 38(Database issue):D690–D698

19. Kaufman MH (1992) The atlas of mouse development. Elsevier Academic Press, London, revised 1995 edition
20. Kerwin J, Yang Y, Merchan P, Sarma S, Thompson J, Wang X, Sandoval J, Puelles L, Baldock R, Lindsay S (2010) The HUDSEN Atlas: a three-dimensional (3D) spatial framework for studying gene expression in the developing human brain. J Anat 217(4):289–299
21. Kirby BB, Takada N, Latimer AJ, Shin J, Carney TJ, Kelsh RN, Appel B (2006) In vivo time-lapse imaging shows dynamic oligodendrocyte progenitor behavior during zebrafish development. Nat Neurosci 9(12):1506–1511
22. Lawson KA, Meneses JJ, Pedersen RA (1986) Cell fate and cell lineage in the endoderm of the presomite mouse embryo, studied with an intracellular tracer. Dev Biol 115(2):325–339
23. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, Lin Chen, Li Chen, Chen T-M, Chin MC, Chong J, Crook BE, Czaplinska A, Dang CH, Datta S, Dee NR, Desaki AL, Desta T, Diep E, Dolbeare TA, Donelan MJ, Dong H-W, Dougherty JG, Duncan BJ, Ebbert AJ, Eichele G, Estin LK, Faber C, Facer BA, Fields R, Fischer SR, Fliss TP, Frensley C, Gates SN, Glattfelder KJ, Halverson KR, Hart MR, Hohmann JG, Howell MP, Jeung DP, Johnson RA, Karr PT, Kawal R, Kidney JM, Knapik RH, Kuan CL, Lake JH, Laramee AR, Larsen KD, Lau C, Lemon TA, Liang AJ, Liu Y, Luong LT, Michaels J, Morgan JJ, Morgan RJ, Mortrud MT, Mosqueda NF, Ng LL, Ng R, Orta GJ, Overly CC, Pak TH, Parry SE, Pathak SD, Pearson OC, Puchalski RB, Riley ZL, Rockett HR, Rowland SA, Royall JJ, Ruiz MJ, Sarno NR, Schaffnit K, Shapovalova NV, Sivisay T, Slaughterbeck CR, Smith SC, Smith KA, Smith BI, Sodt AJ, Stewart NN, Stumpf K-R, Sunkin SM, Sutram M, Tam A, Teemer CD, Thaller C, Thompson CL, Varnam LR, Visel A, Whitlock RM, Wohnoutka PE, Wolkey CK, Wong VY, Wood M, Yaylaoglu MB, Young RC, Youngstrom BL, Yuan XF, Zhang B, Zwingman TA, Jones AR (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445(7124):168–176
24. Marcon L, Arqués CG, Torres MS, Sharpe J (2011) A computational clonal analysis of the developing mouse limb bud. PLoS Comput Biol 7(2):e1001071
25. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L, Thompson P, MacDonald D, Iacoboni M, Schormann T, Amunts K, Palomero-Gallagher N, Geyer S, Parsons L, Narr K, Kabani N, Goualher GL, Feidler J, Smith K, Boomsma D, Pol HH, Cannon T, Kawashima R, Mazoyer B (2001) A four-dimensional probabilistic atlas of the human brain. J Am Med Inf Assoc (JAMIA) 8(5):401–430
26. McMahon AP, Aronow BJ, Davidson DR, Davies JA, Gaido KW, Grimmond S, Lessard JL, Little MH, Potter SS, Wilder EL, Zhang P, GUDMAP project (2008) GUDMAP: the genitourinary developmental molecular anatomy project. J Am Soc Nephrol (JASN) 19(4):667–671
27. Miles A, Bechhofer S (2008) SKOS simple knowledge organization system reference. W3C Recommendation
28. Ng L, Bernard A, Lau C, Overly CC, Dong H-W, Kuan C, Pathak S, Sunkin SM, Dang C, Bohland JW, Bokil H, Mitra PP, Puelles L, Hohmann J, Anderson DJ, Lein ES, Jones AR, Hawrylycz M (2009) An anatomic gene expression atlas of the adult mouse brain. Nat Neurosci 12(3):356–362
29. Paxinos G, Watson C (2004) The rat brain in stereotaxic coordinates – the new coronal set. 5th edition, Academic Press, Amsterdam, The Netherlands
30. Ringwald M, Baldock RA, Bard J, Kaufman MH, Eppig JT, Richardson JE, Nadeau JH, Davidson D (1994) A database for mouse development. Science (New York, NY) 265: 2033–2034
31. Rowe T, Frank LR (2011) The disappearing third dimension. Science (New York, NY) 331(6018):712–714
32. Sharpe J, Algren U, Perry P, Hill B, Ross A, Hecksher-Serensen J, Baldock R, Davidson D (2002) Optical projection tomography for 3D microscopy and gene expression studies. Science (New York, NY) 296:541–545 (The initial OPT paper).

33. Venkataraman S, Stevenson P, Yang Y, Richardson L, Burton N, Perry TP, Smith P, Baldock RA, Davidson DR, Christiansen JH (2008) EMAGE–Edinburgh mouse atlas of gene expression: 2008 update. Nucl Acids Res 36(Database issue):D860–D865
34. Walls JR, Coultas L, Rossant J, Henkelman RM (2008) Three-dimensional analysis of vascular development in the mouse embryo. PloS One 3(8):e2853
35. Weninger WJ, Geyer SH, Mohun TJ, Rasskin-Gutman D, Matsui T, Ribeiro I, Costa LDF, Izpisúa-Belmonte JC, Müller GB (2006) High-resolution episcopic microscopy: a rapid technique for high detailed 3D analysis of gene activity in the context of tissue architecture and morphology. Anat Embryol 211(3):213–221
36. Weninger WJ, Mohun T (2002) Phenotyping transgenic embryos: a rapid 3-D screening method based on episcopic fluorescence image capturing. Nat Genet 30(1):59–65