Topical Review

Analysing the developing brain transcriptome with the GenePaint platform

Gonzalo Alvarez-Bolado and Gregor Eichele

Max Planck Institute of Biophysical Chemistry, D-37077 Göttingen, Germany

We discuss technical means by which the complexity of gene and protein signalling cascades can be projected onto the complex structure of the mammalian brain. We argue that this requires both robotics and novel computational tools to register images of gene expression, annotate expression patterns and quantify gene expression. When sufficiently enriched and detailed, such gene expression/neuroanatomical atlases are hypothesis-generating tools and contain in themselves much of the information needed to investigate function in normal and genetically or otherwise modified brains. To be successful and useful, data-rich and comprehensive gene expression/neuroanatomical atlases have to be web accessible and structured in a way that allows the application of data exploration and mining tools.

(Received 3 May 2006; accepted after revision 4 July 2006; first published online 6 July 2006)

Corresponding author G. Eichele: Max Planck Institute of Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. Email: gregor.eichele@mpihan.mpg.de

Contributions of genomic and proteomic technologies to understanding the development and function of the mammalian brain

The task of the founders of neuroscience in the 19th century was to break down the 'relatively featureless' brain into understandable parts and subunits, to tease out the structure of it, to make it understandable first of all morphologically, as an object (Glickstein, 2006). This was immensely helped by the method of Golgi revealing neurons and their prolongations as individual cells of huge morphological complexity, allowing Cajal and his contemporaries to recognize in the formerly 'featureless' brain a highly organized anatomical structure, morphologically understandable, with defined parts and connections between them. Today, genomics is encountering technical and conceptual challenges reminiscent of the early trials of neuroanatomists. We aim here to spell out what some of these challenges are, particularly in relation to the brain, and discuss avenues to address them. What approaches are needed for comprehensive analysis of the interactivities between brain and genome? In the past decade, the 'relatively featureless' genome has become suddenly 'knowable' and 'explorable' through large-scale sequencing projects, expression studies and associated bioinformatics efforts, all helping in functional genome annotation (Lander *et al.* 2001; Waterston *et al.* 2002; Gibbs *et al.* 2004). The brain is by no means featureless from the genetic point of view either, but shows site-, time- and process-specific expression of many genes and proteins, both of which largely drive brain development, function and senescence in a concerted fashion.

How would one reveal the molecular order underlying the highly differentiated morphology of the brain? How can one map, in a coherent and standardized fashion, the expression of genes and proteins that constitute entire signalling cascades onto neurons and circuits? High-throughput technologies (e.g. microarray, sequencing, interactome screens) and data mining tools (gene ontology databases) have emerged to effectively address molecular questions at a genome-wide level. In model organisms such as *C. elegans*, complete descriptions of neuronal circuits are possible (Chalfie *et al.* 1985; Chao *et al.* 2004; Gray *et al.* 2005), but in species with highly complex brains less information of that nature is currently available. A likely avenue to enhance the knowledge of complex brains would be imaging molecular information onto the neuroanatomical substrate thus increasing the detail of the description of neurons and circuits by orders of magnitude. The first challenge for such a project is the creation of digital, searchable anatomical atlases of gene products (transcripts, proteins, modified proteins) on a genome-wide level for the normal brain (for a recent comprehensive review of gene expression atlases see Sunkin, 2006). These atlases need to be such that gene products are mapped onto individual neurons and circuits

DOI: 10.1113/jphysiol.2006.112763

with high accuracy so that gene products and neurons can clearly be linked to each other (Fig. 1).

Similar to Cajal's descriptive anatomy of the nervous system, gene expression atlases are hypothesis-generating tools. From histological data, Cajal was able to deduce organizational principles and functional circuits (arrows in Fig. 1) which guided later experimental approaches (Sherrington, 1906). Thus global gene expression maps will help generate specific hypotheses, which then will have to be tested by genetic and other types of interventions. For example, knocking out a gene in selected neurons could result in behavioural defects. To understand these defects it will be necessary to search for molecular changes on a genome-wide level and at cellular resolution using, e.g. *in situ* hybridization, immunohistochemistry or tagged reporter proteins. It could be argued that such a 'shotgun' strategy would create an excessive amount of irrelevant data. However, with efficient data production and mining tools at hand, casting the net widely will ensure that no aspect remains unexplored, as it is not clear *a priori* which factors are significant in a biological process.

Based on the above lines of reasoning, it seemed to us and to other investigators (see www.brain-map.org as well as Gong *et al.* 2003; Gray *et al.* 2004; Magdaleno *et al.* 2006) that genome-wide atlases of gene expression are very useful to have. Such atlases can be constructed by scaling-up conventional methods of gene expression analysis, but in the long term, automation of production of data and of data mining are indispensable. Although one can establish an atlas of gene expression by, e.g. ISH on a manual basis for a few thousand genes within a reasonable amount of time, the scientific benefit of such studies is probably much greater when multiple conditions (e.g. mutants, different developmental stages, etc.) can be sent through the pipeline in parallel and within a short period of time. This then does require a significant degree of automation of process.

This line of reasoning prompted us to initiate the development of robotic methods for gene expression analysis paralleled by the design of data formats that can be mined (Herzig *et al.* 2001; Visel *et al.* 2004). We refer to this approach as 'GenePaint', which is in fact a combination of automated or semiautomated procedures that allow the following: (1) determination and visualization of gene expression patterns by *in situ* hybridization on histological sections using solvent delivery robotics; (2) scanning of the *in situ* hybridization results with a semiautomated microscope; (3) automated uploading of images and metadata into a Web-based database using a laboratory information management system (LIMS); and (4) mining of manually or automatically annotated data.

Data generation and image acquisition using GenePaint technology

The data production components of GenePaint rely on instrumentation originally developed for robotic liquid handling (Fig. 2*A*). To take advantage of this off-the-shelf technology, *in situ* hybridization is carried out in flow-through chambers (Fig. 2*B*) making it possible to analyse ~1000 tissue sections in < 24 h. Instead of radioactive RNA probes GenePaint uses non-radioactive digoxigenin-tagged riboprobes (Yaylaoglu *et al.* 2005). Probes are detected on frozen sections by means of a two-step catalysed reporter deposition method comparable in sensitivity to radioactive protocols, and consistently giving high-quality results (see for instance Oldekamp *et al.* 2004; Yaylaoglu *et al.* 2005; Visel *et al.* 2006). We estimate that as few as five mRNA molecules per cell can be detected (Yaylaoglu *et al.* 2005). The benefit of using non-radioactive probes lies in the single cell resolution they can provide (Fig. 2*C–E*) and in the chemical stability of the probe making a long-term storage and use feasible.

Gene expression patterns are digitally recorded at a resolution adequate to reveal individual cells using a microscope equipped with a motorized scanning

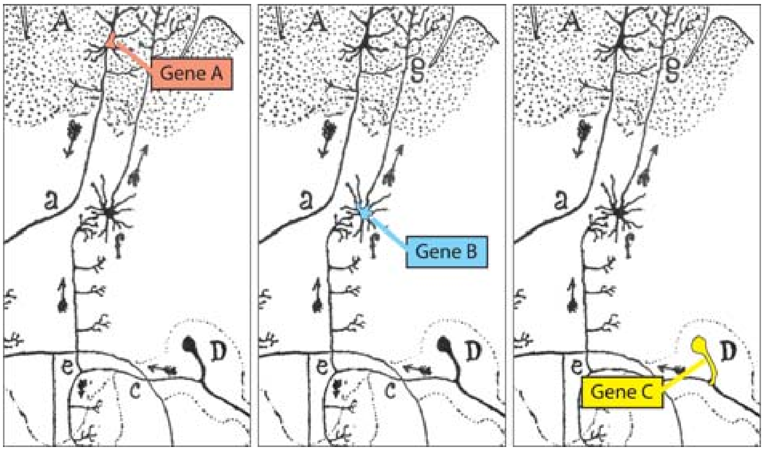

Figure 1. Major challenge in merging cellular and molecular data
A typical neuronal circuit involves multiple neurons (Ramón y Cajal, 1892). Shown is an idealized case where three different types of neurons each express a particular gene (A, B and C). While it is easy to determine which gene marker colocalizes with a particular neuron in a two-dimensional section using double-labelling strategies, this assignment becomes a formidable challenge when extended to the complete genome and to the entire three-dimensional brain. Moreover, it is unlikely that genetic markers exists that tag each and every type of neuron in the brain. Thus, many neurons probably can only be identified by a combination of gene markers requiring high quality image registration.

stage and a digital camera. The scanning process is automated and controlled by software (Carson *et al.* 2002; Visel *et al.* 2004) and also generates control files required for subsequent automated uploading of images into a web database. It should be noted that the GenePaint technology is also successfully used with paraffin sections (Fig. 2*E*) and can readily be extended to immunohistochemistry. Moreover, information handling tools (e.g. LIMS) are internet-compatible, thus allowing multisite data production with data storage in a single location.

Main content of GenePaint.org database

The data generated by the procedures summarized above are deposited on a web database termed GenePaint.org, to make expression patterns available to the scientific community. The database is a research tool in itself (see below) containing images and associated metadata (e.g. the sequence of the template used to prepare the RNA probe, hybridization conditions, etc.), as well as annotation of the expression patterns (Visel *et al.* 2004). As of Spring 2006, GenePaint.org holds the expression patterns of > 4000 genes the majority of which are for mouse embryonic stage E14.5. Here, for each gene a set of ∼24 serial sagittal 25 μm thick sections through the entire embryo body are deposited, unless expression of a gene is judged as 'not detected' in which case only a mid-sagittal section is provided. The amount of data for this stage present in GenePaint.org steadily grows since this database serves as one of the repositories of data generated by the Eurexpress Consortium (www.eurexpress.org), a group of investigators located in different European countries (*in situ* hybridization data are produced at: Telethon Institute of Genetics and Medicine, Naples, Italy; Max Planck Institute for Experimental Endocrinology, Hannover (now at the Max Planck Institute of Biophysical Chemistry in Göttingen), Germany; Max Planck Institute for Molecular Genetics, Berlin, Germany; Centre Européen pour la Recherche en Biologie et Medicine, Strasbourg, France; Division of Medical Genetics, University of Geneva, Geneva, Switzerland). By the end of 2008, the expression patterns of ∼20 000 genes will be housed in GenePaint.org. In addition to expression patterns of embryos, GenePaint.org also contains data from a pilot study for which the expression pattern of ∼300 genes for postnatal P7 mouse brains was determined.

Using GenePaint.org

Expression pattern of a particular gene. In the most simple type of data-mining strategy, the expression patterns of genes of interest can be retrieved through the 'Gene Directory' downloadable from the home page, or by entering into the query field either a gene name, gene symbol, LocusLink ID, GenBank accession number or GenePaint set ID. One can also enter at the advanced search page a DNA sequence, and have the BLAST search executed within GenePaint.org comparing the query sequence with the sequences of the templates used for RNA probe synthesis.

Expression pattern of all genes expressed in a certain structure. The search for expression in anatomical structures in E14.5 mouse embryos is a widely used

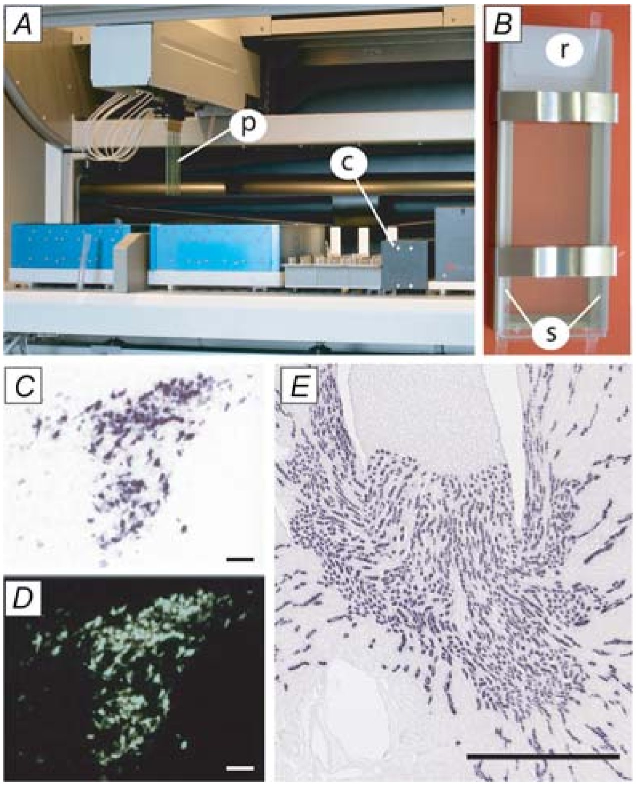

Figure 2. GenePaint robot and results
A, a solvent delivery robot whose cluster of pipettes (p) aspirate solutions from containers located on the platform and deliver them into flow-through hybridization chambers located in temperature-controlled chamber racks (c). *B*, a flow-through hybridization chamber in which a 5 mm thick glass plate is clamped together with a slide that carries sections. The thickness of this chamber is 80 μm which is achieved by placing two spacers (s) between the glass plate and the slide. A solvent reservoir (r) accommodates the solutions delivered by the pipettes to the chamber. *C*, expression of arginine vasopressin (*Avp*) in the paraventricular nucleus of the hypothalamus. Shown is a 25 μm fresh frozen mouse brain section. Note that despite this thickness the data provide single cell resolution of the staining as long as *Avp*-expressing cells are clearly spatially separated. *D*, a pseudo-darkfield of *C* generated by applying the 'invert' command of Adobe Photoshop. This operation generates an image similar to one that would be generated using a radioactive riboprobe and emulsion autoradiography. *E*, the expression of a solute carrier (*Slc12a1*) in an 8 μm thin paraffin section of a paraformaldehyde-fixed adult mouse kidney. Frozen sections (*C* and *D*) and paraffin section (*E*) both give expression data with strong signal and low background. Scale bars: *C* and *D*, 100 μm; *E*, 650 μm.

feature of GenePaint.org. To make expression patterns systematically searchable, genes with region-specific expression patterns (but not those expressed ubiquitously) are annotated in an ongoing effort. The annotation records expression level and local distribution pattern of a transcript in each of approximately one hundred hierarchically organized anatomical structures. In this way, all genes expressed in a certain structure can be retrieved using the 'structure selection tool' located in the advanced search page. For instance, we have used this powerful search tool to find genes in a very specific neuronal subpopulation, the Cajal-Retzius cells of the hippocampal fissure, which resulted in the intriguing finding that many gene products expressed in these cells have angiogenesis-related functions (Skutella *et al.* 2006; Zhao *et al.* 2006).

Phenotype analysis. Although the presence of a transcript in a particular cell or tissue does not necessarily imply the gene in question has a function at all sites of expression (Rodriguez-Trelles *et al.* 2005; Yanai *et al.* 2006), even fortuitous expression provides useful information in the form of cell markers. For example, certain cell populations may be absent in genetic mutants or may expand and cell-specific markers are commonly used to document such phenotypes. Hence a searchable atlas of gene expression patterns is an invaluable tool for mutant analysis. A related area of application is tying expression of a particular gene to a pathway. For example, targets of transcription factors are expected to be expressed in the same cells as the transcription factor itself. Moreover, in a mouse in which a transcription factor has been knocked-out, the expression of targets is expected to be affected.

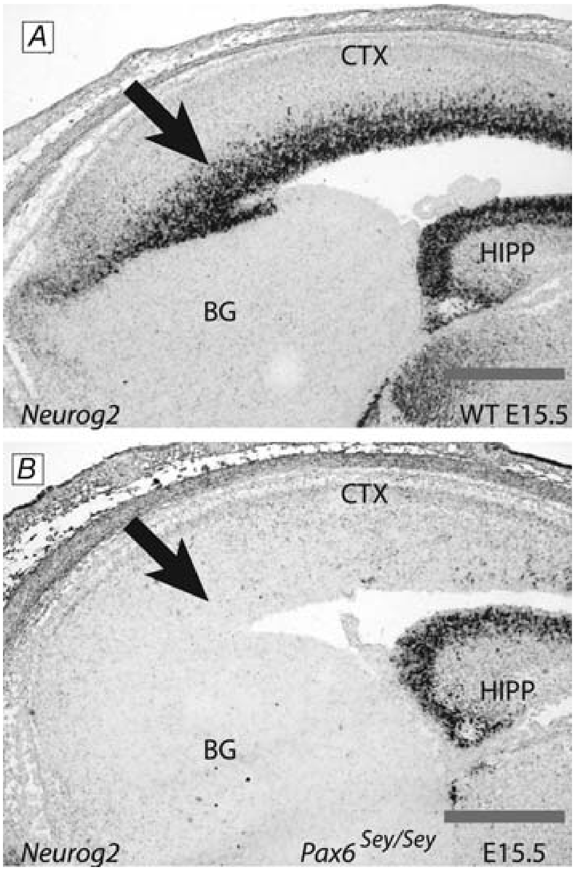

Figure 3. *Neurog2* expression in wild-type and *Pax6*$^{Sey/Sey}$ cortex at E15.5
Genepaint-ISH detection of *Neurog2* transcripts on sagittal sections of wild-type (*A*) and *Pax6*$^{Sey/Sey}$ (*B*) cortex at E15.5. Arrows indicate the region of expression in the germinal layers of the cortex. Neocortical expression is completely absent in the mutant (*B*). Abbreviations: BG, basal ganglia; CTX, neocortex; HIPP, hippocampus. Scale bars 500 μm.

The assortment of genes expressed in a particular anatomical structure at any given moment is characteristic for that structure and can be used as a screening tool in order to assess how that structure and/or processes within it change in a genetically or experimentally modified animal. In practical terms, the structure selection tool of GenePaint.org is used to identify transcripts that are regionally expressed in a tissue such as the developing cerebral cortex. At the present time a search for genes regionally and strongly expressed in the cerebral cortex at E14.5 results in ~400 hits. Next one can use the stable digoxigenin riboprobes used for atlas production to determine the expression pattern of these 400 genes in mutants defective for cortical development. *Small eye* (*Sey*) mice carry a mutation in the transcription factor gene *Pax6* (Hill *et al.* 1991) and are characterized by abnormal development of eye, pancreas and cortex (Jordan *et al.* 1992; Stoykova *et al.* 1996; St-Onge *et al.* 1997). Subjecting all 400 genes in the hit list to GenePaint expression analysis on E15.5 brain sections of *Small eye* and wild-type reveals that many of the 400 genes show changes in expression in mutant cortex suggesting that their expression is directly or indirectly controlled by *Pax6*. An example of this analysis is represented by *Neurogenin 2* (*Neurog2*), a cortical gene under the direct control of *Pax6* (Scardigli *et al.* 2003). *Neurog2* appears indeed among the ~400 genes of the hit list and, upon GenePaint expression analysis, shows dramatic expression changes in the *Pax6*$^{Sey/Sey}$ cortex (Fig. 3).

Analogous efforts are currently under way by other groups (T. Skutella, personal communication) in order to detect genes regionally expressed in the gut, and use them as markers to characterize mouse models of human chronic inflammatory conditions of the intestine.

The future: digital atlas plus celldetekt –the robotic neuroanatomist

The annotation of E14.5 expression patterns in GenePaint.org is currently being carried out by experts who look at every image and annotate the pattern according to ~100 anatomical structures. Although such text-based annotation data provide an effective method to search for regionally expressed genes, it will be in the long run impractical to annotate the huge datasets generated by robotic ISH. Besides, the inclusion of ~100 structures may be insufficient for fully encompassing an organ like the brain which contains > 1000 nuclei. A more powerful way of annotation of the brain uses software to overlay brain sections with an appropriate mesh representing brain subdivisions/nuclei, and then quantify expression in each subdivision by image analysis. Such a strategy allows for automatic identification of genes sharing any user-defined common expression pattern features. An example of such an effort is an approach using geometric modelling techniques to create a deformable digital atlas of the structures to be analysed (Carson *et al.* 2005*b*). The atlas can be adjusted to match for instance the major anatomical structures in neonatal (postnatal day 7) mouse brain tissue sections (the model on which it has been tested), accurately define the boundaries between brain nuclei, and provide a multiresolution coordinate representation of small structures. This technique has been combined with software able to estimate strength of gene expression (Celldetekt; (Carson *et al.* 2005*a*), in order to automatically annotate a large number of gene expression patterns in a way that allows queries and comparisons of expression patterns in user-defined regions of interest. This can be used for instance to identify candidate genes involved in regionalized biological or pathological processes (Carson *et al.* 2005*b*). This method can be extended to create 3-D atlases, allowing for more efficient alignments of expression patterns than a set of two-dimensional maps would permit (Ju *et al.* 2005, 2006).

Conclusions

Both experimental methods and computational strategies summarized here outline a path by which the complexities of gene expression and neuronal circuits can be brought together. We believe for this effort to be truly successful a high degree of automation and advanced computational tools dealing with the complex 3D geometry of the brain need to be developed and implemented. GenePaint hardware, GenePaint data handling tools, web databases, geometry-based annotation (Carson *et al.* 2005*b*) and search tools are first steps in a course that will eventually lead to deciphering how gene and protein cascades and neuronal circuits cooperate in the development and function of the mammalian brain.

References

Carson JP, Eichele G & Chiu W (2005*a*). A method for automated detection of gene expression required for the establishment of a digital transcriptome-wide gene expression atlas. *J Microsc* **217**, 275–281.

Carson JP, Ju T, Lu HC, Thaller C, Xu M, Pallas SL, Crair MC, Warren J, Chiu W & Eichele G (2005*b*). A digital atlas to characterize the mouse brain transcriptome. *PLoS Comput Biol* **1**, e41.

Carson JP, Thaller C & Eichele G (2002). A transcriptome atlas of the mouse brain at cellular resolution. *Curr Opin Neurobiol* **12**, 562–565.

Chalfie M, Sulston JE, White JG, Southgate E, Thomson JN & Brenner S (1985). The neural circuit for touch sensitivity in *Caenorhabditis elegans. J Neurosci* **5**, 956–964.

Chao MY, Komatsu H, Fukuto HS, Dionne HM & Hart AC (2004). Feeding status and serotonin rapidly and reversibly modulate a *Caenorhabditis elegans* chemosensory circuit. *Proc Natl Acad Sci U S A* **101**, 15512–15517.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521.

Glickstein M (2006). Golgi and Cajal: The neuron doctrine and the 100th anniversary of the 1906 Nobel Prize. *Curr Biol* **16**, R147–R151.

Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, Schambra UB, Nowak NJ, Joyner A, Leblanc G, Hatten ME & Heintz N (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917–925.

Gray PA, Fu H, Luo P, Zhao Q, Yu J, Ferrari A, Tenzen T, Yuk DI, Tsung EF, Cai Z, Alberta JA, Cheng LP, Liu Y, Stenman JM, Valerius MT, Billings N, Kim HA, Greenberg ME, McMahon AP, Rowitch DH, Stiles CD & Ma Q (2004). Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* **306**, 2255–2257.

Gray JM, Hill JJ & Bargmann CI (2005). A circuit for navigation in *Caenorhabditis elegans. Proc Natl Acad Sci U S A* **102**, 3184–3191.

Herzig U, Cadenas C, Sieckmann F, Sierralta W, Thaller C, Visel A & Eichele G (2001). Development of high-throughput tools to unravel the complexity of gene expression patterns in the mammalian brain. *Novartis Found Symp* **239**, 129–146; discussion 146–159.

Hill RE, Favor J, Hogan BL, Ton CC, Saunders GF, Hanson IM, Prosser J, Jordan T, Hastie ND & van Heyningen V (1991). Mouse small eye results from mutations in a paired-like homeobox-containing gene. *Nature* **354**, 522–525.

Jordan T, Hanson I, Zaletayev D, Hodgson S, Prosser J, Seawright A, Hastie N & van Heyningen V (1992). The human PAX6 gene is mutated in two patients with aniridia. *Nat Genet* **1**, 328–332.

Ju T, Warren J, Carson J, Bello M, Kakadiaris I, Chiu W, Thaller C & Eichele G (2006). 3D volume reconstruction of a mouse brain from histological sections using warp filtering. *J Neurosci Methods* (in press).

Ju T, Warren J, Carson J, Eichele G, Thaller C, Chiu W, Bello M & Kakadiaris I (2005). Building 3D surface networks from 2D curve networks with application to anatomical modeling. *Visual Comput* **21**, 764–773.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Magdaleno S, Jensen P, Brumwell CL, Seal A, Lehman K, Asbury A, Cheung T, Cornelius T, Batten DM, Eden C, Norland SM, Rice DS, Dosooye N, Shakya S, Mehta P & Curran T (2006). BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system. *PLoS Biol* **4**, e86.

Oldekamp J, Kramer N, Alvarez-Bolado G & Skutella T (2004). Expression pattern of the repulsive guidance molecules RGM A, B and C during mouse development. *Gene Expr Patterns* **4**, 283–288.

Ramón y Cajal S (1892). El nuevo concepto de la histología de los centros nerviosos. *Revista Ciencias Médicas* **18**, 457–476.

Rodriguez-Trelles F, Tarrio R & Ayala FJ (2005). Is ectopic expression caused by deregulatory mutations or due to gene-regulation leaks with evolutionary potential? *Bioessays* **27**, 592–601.

Scardigli R, Baumer N, Gruss P, Guillemot F & Le Roux I (2003). Direct and concentration-dependent regulation of the proneural gene Neurogenin2 by Pax6. *Development* **130**, 3269–3281.

Sherrington CS (1906). *The Integrative Action of the Nervous System*. Yale University Press, New Haven.

Skutella T, Conrad S, Bonin M, Hooge J & Alvarez-Bolado G (2006). Microarray analysis of the fetal hippocampus in the *Emx2* mutant. *Dev Neurosci* (in press).

St-Onge L, Sosa-Pineda B, Chowdhury K, Mansouri A & Gruss P (1997). Pax6 is required for differentiation of glucagon-producing α-cells in mouse pancreas. *Nature* **387**, 406–409.

Stoykova A, Fritsch R, Walther C & Gruss P (1996). Forebrain patterning defects in *Small eye* mutant mice. *Development* **122**, 3453–3465.

Sunkin SM (2006). Towards the integration of spatially and temporally resolved murine gene expression databases. *Trends Genet* **22**, 211–217.

Visel A, Alvarez-Bolado G, Thaller C & Eichele G (2006). Comprehensive analysis of the expression patterns of the adenylate cyclase gene family in the developing and adult mouse brain. *J Comp Neurol* **496**, 684–697.

Visel A, Thaller C & Eichele G (2004). GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucl Acids Res* **32**, D552–D556.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P & Lercher MJ (2006). Similar gene expression profiles do not imply similar tissue functions. *Trends Genet* **22**, 132–138.

Yaylaoglu MB, Titmus A, Visel A, Alvarez-Bolado G, Thaller C & Eichele G (2005). Comprehensive expression atlas of fibroblast growth factors and their receptors generated by a novel robotic in situ hybridization platform. *Dev Dyn* **234**, 371–386.

Zhao T, Kraemer N, Oldekamp J, Cankaya M, Szabó N, Conrad S, Skutella T & Alvarez-Bolado G (2006). *Emx2* in the developing hippocampal fissure. *Eur J Neurosci* (in press).

Acknowledgements

This work was supported by the Max Planck Society, the German Ministry of Research, the European Union and the Burroughs Wellcome Fund.