

Applying genomics technologies to neural development

Seth Blackshaw* and Rick Livesey†

Genomic technologies have evolved from a minority interest to a set of generally applicable, powerful tools. Recent studies have demonstrated that such tools are of great use in studies of neural development, particularly when allied to advances in data analysis and methods for analyzing gene function.

Addresses

*Dept of Genetics and Howard Hughes Medical Institute, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA; e-mail: sblack@rascal.med.harvard.edu

†Wellcome/CRC Institute of Cancer and Developmental Biology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QR, UK; e-mail: rick@welc.cam.ac.uk

Current Opinion in Neurobiology 2002, 12:110–114

0959-4388/02/\$ – see front matter

© 2002 Elsevier Science Ltd. All rights reserved.

Abbreviations

2D	two-dimensional
GFP	green fluorescent protein
FACS	fluorescence activated cell sorting
IMAGE	integrated molecular analysis of genomes and their expression
MPSS	massively parallel signal sequencing
SAGE	serial analysis of gene expression

Introduction

Genomics commonly refers to any genome-scale approach to studying biological problems, most notably expression profiling, whereby one studies the expression of many, if not all, genes or proteins in a given cell or tissue type. Within developmental neurobiology, genomic technologies can be used as gene discovery tools — whether cell-specific, tissue-specific or stage-specific — or can be used to investigate the transcriptional or translational contributions to processes as diverse as neurogenesis, cell fate determination, differentiation, axon guidance and synapse formation. For instance, expression profiling can be used to classify cells, tissues or tumor types [1,2] and to investigate signal transduction pathways [3]. Genomics technologies have been used in invertebrate development to profile gene expression in the developing nematode [4,5] and in various sorted green fluorescent protein (GFP)-marked cell populations in *Drosophila melanogaster* [6,7]. In the developing mammalian nervous system, these approaches have been used to study maturation of the hippocampus [8], photoreceptor differentiation in the retina [9] and neural progenitor differentiation [10].

Genomics technologies have become widely available in recent years. The intellectual problems of interest in developmental biology have not changed dramatically over the same period, so why has there been a recently increased interest in genomic technologies? The main reason, perhaps, is that these technologies have proven to be robust, accurate and, perhaps most importantly, comprehensive

tools that can be used in any laboratory. In principle, with the availability of complete genome sequences from several organisms and the development of comprehensive expression tools, it is now possible to get definitive answers to the gene expression aspects of developmental neurobiology questions. However, powerful as these technologies are, we are just leaving the proof-of-principle stage of the application of these approaches to biological questions. As they become more widely used, there are both practical and scientific concerns that are useful to keep in mind when considering a study using these approaches. This review will attempt to shed light on some of these issues.

Choosing the right technology

Expression profiling can be divided into those technologies profiling mRNA levels and those profiling protein expression. The latter — proteomics — holds considerable promise, although is not currently in general use in many labs. However, powerful gel electrophoresis-based and microarray-based technologies have been developed and are becoming more widely used in non-specialist environments [11,12]. RNA expression profiling technologies are either sequencing-based, such as Serial Analysis of Gene Expression (SAGE) [13], or hybridization-based approaches, such as microarrays [14] (Table 1). The groundbreaking work carried out by Velculescu *et al.* [13] and Schena *et al.* [14] forms the basis of most genomic technologies in use today. In addition to those listed in Table 1, there are several profiling technologies under development that should dramatically increase throughput, such as the massively parallel signal sequencing (MPSS) technology [15,16]. Rather like SAGE, this emerging technology uses a novel approach to sequence short signature sequences from millions of cDNAs in a sample in parallel.

Choosing the most appropriate technology is a mixture of the pragmatic and the scientific. Making cDNA microarrays within a lab or institution can involve a significant degree of investment in infrastructure, including liquid handling systems for managing clones, PCR products or oligo-nucleotides, array printers and optical scanners, and personnel. The use of commercial microarrays removes many of these needs, although access to an optical scanner/reader is usually still necessary. SAGE does not require a large investment in new equipment, as it depends on library construction and DNA sequencing, but does require high-throughput DNA sequencing, which can be rate limiting and expensive, particularly for large-scale studies. There are now commercially available resources for all of these techniques (Table 1).

Scientifically, there may be a need to use sequencing-based or hybridization-based approaches, as each method has specific advantages and disadvantages. Sequencing-based

Table 1

Genomics technologies available for the study of neural development.

Method	Resources	Sources	Advantages	Disadvantages
Sequencing SAGE	Library construction, high-throughput sequencing	Library construction kit – Invitrogen	Few infrastructure needs; uses proven, common techniques; data are easily compared between studies, labs etc.	Labour intensive and relatively slow; sequencing costs
Hybridization cDNA microarrays	Printed arrays	Many companies and array sizes: examples include Mergen, NEN, Clontech, Stratagene, Agilent, Motorola.	No investment in equipment or array production required	Arrays may not include genes relevant to study; can be expensive; organism may not be represented
	Clone sets	Research Genetics (Invitrogen); National Institute on Aging; RIKEN; <i>Drosophila</i> (see www.bdgp.org)	Extensive; physical resource useful for further studies; organism or tissues may be represented	Inserts must be amplified and purified for printing; clone-tracking is error-prone
	Oligonucleotide sets	Operon, Compugen, MWG-Biotech	Ready to print; specific to genes of interest	Organism may not be represented; current sets are relatively small
Affymetrix chips	Affymetrix wash station and scanner	Affymetrix	Proven, reliable technology; data can be easily compared between studies, labs etc.	Requires specific equipment

approaches, such as SAGE, allow datasets generated at different times and places to be readily compared and distributed. Datasets generated by hybridization-based approaches, such as microarrays, have proven to be very difficult to compare between laboratories. Aside from technology differences, it is also likely that the comparisons made in a given experiment, although essential for answering a particular biological question, are not generally useful to the entire community. However, array experiments can be carried out very quickly, with many replicates, to provide robust statistical analyses under different experimental conditions. Sequencing-based technologies are limited, in terms of the number of replicate datasets and experiments, and by the speed and costs of library construction and sequencing. A final consideration for sequencing-based technologies is that large numbers of ESTs or genomic sequence must be available to allow ready identification of genes corresponding to the sequences or tags obtained, a factor that limits use of these approaches in non-model organisms. An alternative approach that we and others have successfully used for several non-model organisms, is to construct microarrays of random clones from cDNA libraries of interest (e.g. certain tissues or developmental stages) [17,18]. These are used in the same way as sequenced-clone arrays, but, following experiments and data analysis, a set of clones of interest is then sequenced to identify the corresponding genes.

Great expectations: study design and data analysis

Expression profiling studies that are well designed with appropriate controls generate data that can be successfully interpreted and yield the most useful information. Initially,

this basic principle has been somewhat overlooked, perhaps due to cost concerns and the limited availability of genomics technologies, although this is now changing. One of the most powerful applications of expression profiling in mammalian systems has been comparing gene expression between mutant and wild-type embryos or tissues [9,19]. This approach has several attractive features, including the existence of internal controls, consisting of littermates that are very closely matched to experimental animals in terms of developmental age, environment, genetic background, time of tissue harvesting, RNA extraction and so on. Minimizing as many variables as possible reduces the likelihood of detecting gene expression differences that are unrelated to the variable under study. This is generally true for all genomics studies, regardless of the nature of the starting material. In addition, replicated independent experiments are essential to allow critical statistical analysis of the data generated, particularly for genes expressed at low levels and for smaller changes in gene expression.

Aside from the availability of microarrays and the costs of sequencing, the most talked about issue surrounding expression profiling is data analysis. For labs with their origins in classical developmental and molecular biology, much of the rhetoric surrounding this topic may appear forbidding. Many different laboratories, institutions and companies have developed a variety of useful tools for analyzing large gene expression datasets [20,21]. These methods fall into two broad classes: exploratory statistical methods, such as hierarchical cluster analysis [20], and confirmatory statistics, such as the significance analysis of microarrays approach [22*]. Each approach is very useful for analyzing expression data and clustering methods are

particularly useful for identifying interesting features within large or complex datasets. For studies comparing a small number of different conditions, such as mutant–wild-type comparisons, there has been a noticeable shift away from studies using small numbers of experiments in a given dataset accompanied by cluster analysis, to replicated experiments (with several different control or reference samples or datasets) and critical statistical analyses [23*,24*]. Before beginning a study, researchers should establish a simple data-archiving system — with the help of publicly available databases, such as AMAD (Another Microarray Database; www.microarrays.org) — and design experiments with future statistical analysis in mind. There are several publicly available tools which allow statistical analysis of single and replicate experiments, paired experiments or across multiple experiments, including the statistical analysis of microarrays (SAM) and statistics for microarray analysis (SMA) packages [22*,25*,26*]. In addition to calculating the significance of observed differences and the reproducibility of experiments, these approaches also provide estimates of false discovery rates at different significance thresholds, a useful feature when setting criteria for follow-up studies.

Closing out: verification and function

Identification of large numbers of genes whose expressions change in correlation with a particular event or process, or that appear in a tissue-specific manner, is only the beginning of the process for a genomics study. It is imperative that researchers undertake a cycle of benchmarking to assess the rate of false positive discovery and how the gene expression measurements generated by their chosen technology correlate with those generated by a second, independent technology. As discussed above, estimates of false positives can be made from statistical analyses of replicated experiments. However, given that the focus of a single study is to identify a number of key genes that are central to a given process, it is necessary to independently confirm those findings most central to the conclusions of the study and to eliminate false positives.

Techniques used to verify expression data will vary depending on the process studied. Quantitative confirmation of relative gene expression can be carried out by real-time PCR, RNase protection or northern blotting, although the dynamic range of the latter is narrow and can overestimate or underestimate gene expression differences. Furthermore, the cellular heterogeneity of the nervous system clouds the interpretation of expression data obtained by the use of homogenized tissue preparations. The simplest way of doing this in a complex, multicellular tissue, such as the nervous system, is by high-throughput *in situ* hybridization to confirm cellular localization and obvious differences in gene expression. This method can distinguish between marked differences in gene expression in small subsets of cells and lower differences in gene expression across a large proportion of cells, a distinction lost when whole tissue samples are analyzed. The use of tissue

microarrays, in which samples from many tissues or developmental stages are arranged on the same slide and probed simultaneously, can considerably expand the power and scope of more traditional *in situ* approaches [27,28]. High throughput *in situ* hybridization is made increasingly straightforward by the availability of large numbers of sequence-verified ESTs, from the IMAGE (Integrated Molecular Analysis of Genomes and their Expression, <http://image.llnl.gov/>) and BMAP (Brain Molecular Anatomy Project, <http://brainest.eng.uiowa.edu>) projects [29]. However, because *in situ* hybridization is not quantitative, it is not appropriate for benchmarking changes observed with a given technology, or for confirming relatively small differences in gene expression.

Depending on the nature of the genomics study, a final consideration is the functional analysis of differentially expressed genes. The large number of differentially expressed genes observed in any given experiment is sometimes unanticipated, even in well controlled genomic studies. Depending on the organism or system used, this can overwhelm even a very large lab that uses conventional technologies. Thus, before designing an experiment, it is essential to develop a clear set of criteria for prioritizing genes, on the basis of function or expression pattern, and to choose as high-throughput as possible a system for carrying out gain or loss of function analysis. This is particularly true of studies in mammals, although gain of function studies in mammals have been made considerably more straightforward by the availability of full-length mammalian cDNAs via the Mammalian Gene Collection effort of the IMAGE consortium [29]. For vertebrate studies, depending on the question addressed, cell-based or explant-based studies and gene misexpression by electroporation, transfection or viral transduction can allow misexpression of many more genes in a given time than can whole animal studies. Emerging RNA interference-based approaches may also allow medium-throughput loss of function studies in these systems [30,31**]. A final consideration, again depending on the question studied, is for first-pass functional studies to be carried out in another organism, such as the zebrafish or *Caenorhabditis elegans*, so that the initial functional studies can take advantage of the powerful functional tools that are available in these systems.

How much is enough: sensitivity and starting material

Perhaps the most significant issue for genomics studies of neural development is the amount and purity of RNA or protein. Given the small amounts of tissue available for many neural development studies, this can be a real problem. Immunosorting or fluorescence-activated cell sorting (FACS) to increase cellular purity may be used if appropriate antibodies or marked strains are available, although care must be taken not to overly damage cells and degrade RNA during isolation [6,7*,32]. There are currently no methods available for protein amplification, although detection thresholds for signal, be it from a two-dimensional (2D) gel

or a mass spectrophotometer, are constantly dropping. Therefore, for proteomics, the technical limits of detection will dictate the nature of the study performed.

SAGE has been optimized for library production from less than 100,000 cells, with no RNA amplification needed [32]. There are currently several methods available for RNA/cDNA amplification from tens of thousands of cells and others that allow gene expression profiling studies using microarrays to achieve single cell resolution [33,34]. cDNA amplification methods depend on either PCR (exponential) or RNA polymerase (linear) amplification. However, there are significant limitations and shortcomings to RNA amplification technologies that should be kept in mind when designing studies that depend on these technologies. In both cases, the two concerns are: the representation of all transcripts present in the starting material in the final, amplified material (i.e. loss of transcripts); and the preservation of the relative abundances of the different transcripts. All of the available technologies compromise these features to some degree, most notably relative abundance. In our experience, many amplification technologies have a normalizing effect on transcript abundance that becomes particularly noticeable with increasing amplification. To correct for this in a study, it is imperative that data be generated by comparing material prepared using the same method and independently verified, as previously discussed.

Conclusions and future prospects

The near future will see the generation of many different datasets for a range of processes and regions of the nervous system, and the free sharing of data, obtained via SAGE or through use of common microarrays, throughout the neurobiological community. Anticipated developments include not only substantial improvements in current RNA-based technologies, but also the applications of proteomics and its allied technologies — such as large-scale 2D gel electrophoresis coupled with protein fingerprinting via mass spectrometry or the use of protein arrays — to the study of neural development. An exciting area is in the detection and profiling of small metabolites, or metabolomics, which will focus more attention on this very important category of molecules in neural development and function [35]. Finally, it is worth noting that these approaches are not simply a logical step forward in scale compared to previous tools in developmental biology, in which the expression of small numbers of genes were studied by *in situ* hybridization or northern blotting. Instead, they open up new ways of thinking about neural development. Most notably, they can enable a shift away from reductionism to a more systems-level approach, in which complex gene and protein expression networks are delineated and studied directly.

Acknowledgements

We thank C Cepko, in whose lab both of the authors have had the opportunity to apply genomics technologies to neural development. F Livesey thanks Vivian Cheung for introducing him to the practical aspects of microarray technologies.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al.*: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747-752.
 2. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer.** *Nature* 2001, **412**:822-826.
 3. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes *et al.*: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles.** *Science* 2000, **287**:873-880.
 4. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: **A gene expression map for *Caenorhabditis elegans*.** *Science* 2001, **293**:2087-2092.
 5. Jiang M, Ryu J, Kiraly M, Duke K, Reinke V, Kim SK: **Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*.** *Proc Natl Acad Sci USA* 2001, **98**:218-223.
 6. Bryant Z, Subrahmanyam L, Tworoger M, LaTray L, Liu CR, Li MJ, van den Engh G, Ruohola-Baker H: **Characterization of differentially expressed genes in purified *Drosophila* follicle cells: toward a general strategy for cell type-specific developmental analysis.** *Proc Natl Acad Sci USA* 1999, **96**:5559-5564.
 7. Furlong EE, Andersen, EC, Null B, White KP, Scott MP: **Patterns of gene expression during *Drosophila* mesoderm development.** *Science* 2001, **293**:1629-1633.
- The data presented in this paper and in [6] demonstrate the power of the use of FACS isolation of cells genetically marked with GFP, to obtain gene expression profiles from purified subsets of cells.
8. Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, Pham K, Schultz P, Welsh D, Tsien JZ *et al.*: **Genome-wide gene expression profiles of the developing mouse hippocampus.** *Proc Natl Acad Sci USA* 2001, **98**:8862-8867.
 9. Livesey FJ, Furukawa T, Steffen MA, Church GM, Cepko CL: **Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene *Crx*.** *Curr Biol* 2000, **10**:301-310.
 10. Geschwind DH, Ou J, Easterday MC, Dougherty JD, Jackson RL, Chen Z, Antoine H, Terskikh A, Weissman IL, Nelson SF, Kornblum HI *et al.*: **A genetic analysis of neural progenitor differentiation.** *Neuron* 2001, **29**:325-339.
- Geschwind *et al.* present an interesting study that combines cDNA subtraction with microarray screening to study neural differentiation.
11. Unlu M, Morgan ME, Minden JS: **Difference gel electrophoresis: a single gel method for detecting changes in protein extracts.** *Electrophoresis* 1997, **18**:2071-2077.
 12. MacBeath G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination.** *Science* 2000, **289**:1760-1763.
- The authors of this paper describe the construction of high-density arrays of recombinant proteins. These proteins are attached covalently to the slides, yet robustly retain their ability to interact with other proteins. The authors demonstrate applications for these high-density arrays in screening for protein-protein interactions, identifying protein kinase substrates, and elucidating the protein targets of small molecules.
13. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484-487.
 14. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467-470.
 15. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M *et al.*: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nat Biotechnol* 2000, **18**:630-634.

The technique described here results in the extraction of 16–20 base-pair signature sequences from millions of individual cDNAs in a single experiment. Given the fact that only 200,000–300,000 mRNA molecules are found per cell and that MPSS allows far more sequence tags than this to be

extracted in a given experiment, it may eventually prove the most sensitive and comprehensive of all cDNA-based measures of gene expression.

16. Mitra RD, Church GM: **In situ localized amplification and contact replication of many individual DNA molecules.** *Nucleic Acids Res* 1999, **27**:34.
17. Altmann CR, Bell E, Sczyrba A, Pun J, Bekiranov S, Gaasterland T, Brivanlou AH: **Microarray-based analysis of early development in *Xenopus laevis*.** *Dev Biol* 2001 **236**:64-75.
18. Zhu X, Mahairas G, Illies M, Cameron RA, Davidson EH, Etensohn CA: **A large-scale analysis of mRNAs expressed by primary mesenchyme cells of the sea urchin embryo.** *Development* 2001, **128**:2615-2627.
19. Luthi-Carter R, Strand A, Peters NL, Solano SM, Hollingsworth ZR, Menon AS, Frey AS, Spektor BS, Penney EB, Schilling G *et al.*: **Decreased expression of striatal signaling genes in a mouse model of Huntington's disease.** *Hum Mol Genet* 2000, **9**:1259-1271.
20. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
21. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96**:2907-2912.
22. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98**:5116-5121.
See annotation to [26*].
23. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17**:509-519.
See annotation to [26*].
24. Lee ML, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci USA* 2000, **97**:9834-9839.
See annotation to [26*].
25. Kerr MK, Churchill GA: **Statistical design and the analysis of gene expression microarray data.** *Genet Res* 2001, **77**:123-128.
See annotation to [26*].
26. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice.** *Genome Res* 2000, **10**:2022-2029.
The work described in [22*,23*,24*,25*,26*] allows the development of tools for robust statistical analysis of large gene expression datasets.
27. Kallioniemi OP, Wagner U, Kononen J, Sauter G: **Tissue microarray technology for high-throughput molecular profiling of cancer.** *Hum Mol Genet* 2001, **10**:657-662.
28. Rimm DL, Camp RL, Charette LA, Costa J, Olsen DA, Reiss M: **Tissue microarray: a new technology for amplification of tissue resources.** *Cancer J* 2001, **7**:24-31.
29. Strausberg RL, Feingold EA, Klausner RD, Collins FS: **The mammalian gene collection.** *Science* 1999, **286**:455-457.
30. Caplen NJ, Parrish S, Imani F, Fire A, Morgan RA: **Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems.** *Proc Natl Acad Sci USA* 2001, **98**:9742-9747.
31. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T: **Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells.** *Nature* 2001, **411**:494-498.
Both the authors of this paper and [30] describe the use of short double-stranded RNA duplexes to inhibit the expression of a number of different genes in cultured mammalian cells, in a similar fashion to what has been reported in invertebrates and preimplantation mammalian embryos. Should this technique prove generally applicable in vertebrates, it will revolutionize functional genomics.
32. St Croix B, Rago C, Velculescu VE, Traverso G, Romans KE, Montgomery E, Lal A, Riggins GJ, Lengauer C, Vogelstein B *et al.*: **Genes expressed in human tumor endothelium.** *Science* 2000, **289**:1197-1202.
33. Brady G, Iscove NN: **Construction of cDNA libraries from single cells.** *Methods Enzymol* 1993, **225**:611-623.
34. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P: **Analysis of gene expression in single live neurons.** *Proc Natl Acad Sci USA* 1992, **89**:3010-3014.
35. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ *et al.*: **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.** *Nat Biotechnol* 2001, **19**:45-50.