Global gene expression analysis of developing neocortex using SAGE

SEONG-SENG TAN*, JENNY GUNNERSEN and CHRISTOPHER JOB

Brain Development Laboratory, Howard Florey Institute, University of Melbourne, Victoria, Australia

ABSTRACT  The mammalian brain is estimated to contain about a hundred billion neurons, making it the most complex biological structure on earth. Trying to understand the assembly and function of this elaborate organ is a formidable task. Yet the information to build a brain is encoded by no more than a subset of the 80,000 genes present in the genome, a more manageable number. This review describes the use of SAGE technology (Serial Analysis of Gene Expression) to decode the genetic repertoire of genes that are differentially expressed in time and in space during development of the neocortex, the part of the mammalian brain responsible for complex traits. We demonstrate that SAGE is not only powerful for generating comprehensive molecular portraits from the developing cortex but can also assist in discovering new genes.

KEY WORDS: SAGE, differential gene expression, neocortex

Introduction

Current approaches to studying molecular control of cortical development are characterized by the traditional gene by gene approach. While this has provided occasional and tantalizing glimpses of key developmental processes, the nature of the problem demands a global rather than a gene by gene approach. So what's special about the developing cortex? The cortex is assembled from neurons with unusually complicated life histories, viz (1) long-distance contemporaneous migration from two separate germinal zones into specific layers in a stage-dependent fashion, (2) interactions with radial glial scaffold, other migrating neurons and the surrounding extracellular matrix, (3) reciprocal innervation with specific subcortical targets such as the thalamus and spinal cord, and (4) juggling of the above with a cell's differentiation program emanating from its lineage. In this context, a full understanding of cortical development is unlikely to advance without consideration of the multiple gene sets that are activated in response to each of the above morphogenetic events.

Global studies of gene expression are designed to meet this need by determining which genes are expressed and quantify the levels of expression. These molecular profiles provide "snapshots" of functional capabilities of cells and tissues under question; in addition, they also yield insights into the phenotypic differences between cortical tissues that differ spatially or temporally. Ideally, gene profiling outputs should meet two important goals. First, the output would provide a molecular inventory of genes that are expressed in developing cortical tissue in defined space and time. Data containing information on sequence abundance and complexity can then be analysed for higher order correlations using appropriately designed algorithms. Second, it should detect and quantify the expression of both known and unknown genes. The latter includes genes that have been neither cloned nor sequenced. Although microarrays and oligonucleotide chips are well suited for quantifying the expression of known genes, open-platform technologies such as SAGE are able to interrogate cells and tissues for both known and unknown genes.

In this review, we outline the use of SAGE to profile genes that are expressed in different regions of the developing neocortex, and during different stages of neocortical development. The unit of SAGE output is a "tag" consisting of a short (10 bp) cDNA sequence. Each tag represents a bar-code for an expressed gene in the mRNA pool. SAGE works by two principles (Velculescu et al., 1995). First, a short tag of about 10 bp of random nucleotide sequence contains sufficient information to distinguish between 1,048,576 transcripts (4^{10}) provided the tag is taken from a defined position in the mRNA (Fig. 4). Second, by joining the tags together for sequence analysis, SAGE provides a rapid and automatable procedure for highspeed data processing and analysis. Recent experiments from our laboratory and from other laboratories provide ample evidence that SAGE is a powerful method for comparing gene expression between different mRNA

Abbreviations used in this paper: SAGE, Serial analysis of gene expression.
pools (He et al., 1998; de Waard et al., 1999; Gunnersen et al., 2000, Augustine et al., 2001).

Overview of Neocortical Organisation

Mature neurons of the mammalian neocortex originate from dividing neuroepithelial cells that are situated at some distance away and are morphologically indistinguishable from each other (Bayer and Altman, 1991). Yet from this simple sheet of neuroepithelium, a complex structure, unique only to mammals, is assembled. To appreciate the developmental agenda, it is worth considering the organisation of this evolutionarily-selected organ. Parallel to the pial surface, the neocortex is divided into distinct areas serving different functional modalities such as visual, auditory, somatosensory and motor responses (Fig. 1). Although there are variations in complexity between different mammalian species, there is great consistency across and within species in the functional specialization of anatomically distinct areas. For instance, the map of areas from one individual to another is almost identical; areas devoted to vision are posterior in the cortex, whereas areas concerned with body movements are anterior.

Functional and Areal Specification of Cortex

From the beginning of this century, early investigators such as Brodmann have used comparative studies to propose that the cortex is composed of a patchwork of “cortical areas”, each with a specific and unique set of functions (Brodmann, 1909). Almost a century later, this view continues to form the backbone of neocortical research. While functional uniqueness of each cortical area (e.g. motor, visual, auditory) is generally accepted and understood, much less is known about the structural and genetic uniqueness of each area. Beyond the fact that neurons in each area may be wired to subcortical targets, and that they may exhibit local variations in cell density (e.g. visual cortex) or architectonic feature (e.g. barrel-field), it is generally accepted that neurons from different cortical areas possess common characteristics. Examples include (1) constant neuronal number per given volume of cortical tissue (Rockel et al., 1980), (2) constant quanta of neurons emanating from their progenitors (Tan et al., 1998) and (3) common intrinsic cellular microcircuitry (Sur et al., 1988; Metin and Frost, 1989).

A fundamental question is: what determines structural uniqueness during development? Starting from a uniform sheet of neuroepithelial cells, at least three kinds of diversity are generated in the developing cortex, viz neuronal phenotype, layer identity, and area specification. A major focus of current research is to address the question - what is responsible for partitioning the cortex into different tangential domains (area maps)? One view suggests that area identity is determined by interactions between the genetic endowment of the cells and specific environmental influences. No doubt, environmental influences (e.g. thalamocortical afferents, (O’Leary et al., 1994)) play key roles in determining area-specific features, but these are notoriously difficult to assay on an objective level. On the other hand, the expression of locale-specific genes provide the opportunity for molecular assays of regional identity.

The notion that functional and architectural specializations of the cortex are correlative of gene activity can be illustrated by the barrelfield example. In rodents, an area of the somatosensory cortex (SI) is devoted to processing sensory information from the vibrissae of the face through a series of intermediate relays (Woolsey and Van Der Loos, 1970). The cytoarchitecture of this part of the cortex is marked by discrete aggregates of neurons in layer 4 termed “barrels”. Each barrel is innervated by clusters of ventrobasal thalamic afferents, arranged in a pattern that mirrors the whisker field on the rodent’s snout (Woolsey and Van Der Loos, 1970). There is evidence that thalamic afferents to the cortex carry a blueprint of the vibrissae pattern, prior to the overt appearance of the barrels (Erzurumlu and Jhaveri, 1990). In addition, molecular correlates (e.g. proteoglycans, peanut lectin-binding proteins) of barrel field architecture are transiently present in SI before or during barrelfield differentiation (Cooper and Steindler, 1986; Crossin et al., 1989; Steindler et al., 1989), providing a compelling demonstration of gene activity in an area-specific manner (Cohen-Tannoudji et al., 1994).

Area-Specific Features in the Developing Cortex

What sort of evidence is available during embryonic stages to suggest the emergence of asymmetry in an otherwise uniform cerebral wall? Axon tracing experiments using DiI have revealed early regionalization of certain cortical regions that are wired to the thalamus. Among the earliest axons to emerge are those from the subplate, a transitory structure beneath the cortical plate. Even before corticogenesis is half-complete, subplate axons appear to traverse the internal capsule and invade specific regions of the thalamus (McConnell et al., 1989). Reciprocal connections from the thalamus to specific cortical areas suggest the presence of area-sensitive guidance cues for circuit formation (Barbe and Levitt, 1992; Ghosh and Shatz, 1993). Thus, area specification in the cortex needs to be considered over at least three stage-related levels: (1) acquisition of area identity before the subplate is formed, (2) specific recognition of subplate axons for thalamic targets, and (3) recognition of thalamic axons for specific cortical targets.

There are other examples of developmental asymmetry. The pattern of neurogenesis is not uniform, with gradual rostral to caudal, and lateral to medial gradients (Berry and Rogers, 1965). There appear to be pockets of the germinal zone where neurons are born earlier than their brethren in surrounding ar-

Fig. 1. Schematic representation of functional subdivisions of the rodent cortex viewed from the lateral aspect.
Molecular Markers of Regional Identity

This area of rich controversy continues to baffle those who look for molecules that specify cortical area maps. Unlike in the hindbrain or spinal cord, molecules that are shown to possess asymmetric distribution in the cortex appear to have broad expression patterns, overlapping multiple functional and anatomical domains (Fig. 2). This phenomenon, although disappointing for those looking for tight matching between specific gene expression and area identity, is not entirely surprising. The cortex is the product of at least 200 million years of evolution during which differential selection pressures have been brought to bear, in different animals over different eras (Northcutt and Kaas, 1995; Finlay and Darlington, 1995). Rather than a collection of smoothly integrated modules, the cortex is considered by some to be a mosaic of disparate areas, assembled in random fashion by opportunistic phylogenetic processes.

A number of markers with overt asymmetric expression in the developing cortex deserve comment (Fig. 2). The R197 line was created as part of an enhancer-trap experiment (Tan, 1991). Of all the markers under discussion, R197 exhibits the earliest genetic specification of a cortical region. As early as E 10.5, the lacZ reporter gene is expressed in medial but not lateral cerebral neuroepithelium (Fig. 3A). Although the identity of the “trapped” gene remains unknown, the strong and clear cut expression of the reporter in the ventricular zone of the cingulate and medial, but not lateral cortex at E10.5 suggests intrinsic neuroepithelial prepatternning (Fig. 3B). This medial versus lateral asymmetry persists throughout the entire period of corticogenesis into postnatal (Fig. 3C) and adult stages (Fig. 3D). There is no expression of the reporter in either lateral neocortex, or in lateral limbic cortex, suggesting that gene expression is segregated along medio-lateral lines, rather than the evolutionarily older limbic versus the more modern neocortex. Indeed, the pyramidal layer of the hippocampus (archicortex) exhibits expression of the reporter (Fig. 3D) as does the floor plate (Fig. 3B).

Efforts to clone the “trapped” gene have so far proved elusive. Nevertheless, the R197 mouse line provides a “proof of principle” demonstration of genetic specification of medial versus lateral regions long before the commencement of neurogenesis. Supporting this are recent experiments where Emx2 and Pax6 -/- mice show defects in regionalization across the caudal-lateral to rostral-medial axis of the forebrain (Bishop et al., 2000). Both these homebox transcription factors are expressed in gradients along with Pax6 having maximal expression at the rostral-medial pole and Emx2 at the caudal-lateral pole. In Emx2 -/- animals markers for anterior neocortex such as Cadherin 6 expand their zone of expression into the posterior region of the neocortex whereas in Pax6 -/- animals this expression contracts anteriorly. Cadherin 8 whose expression marks both the lateral somatosensory cortex and auditory cortex extends its expression medially in Pax6 -/- animals. These shifts cause anatomical changes as injection of Dil into the occipital neocortex will normally label the lateral geniculate nucleus but in Emx2 -/- animals, these projections connect with the ventroposterior nucleus whose afferents normally extend only so far as the somatosensory cortex. Both these results suggest that these factors purvey anterior (Pax6) or posterior (Emx2) character to the early neocortex by regulating position-dependent expression of downstream genes within the neocortex.

Fig. 3. Expression of the lacZ transgene in developing and adult R197 mouse cortex. (A,B) At E10.5, expression is restricted to the medial but not lateral domain of the cerebral wall. Expression is also seen in the floor plate. Boxed outlines in (B) represent tissue areas that may be used to generate region-specific SAGE libraries. (C) At postnatal day 1 (PN1), lacZ expression is clearly segregated in the medial (M) cortex and absent in the lateral (L) territory. (D) In adult cortex, lacZ expression is restricted to layers 2/3 of medial (arrow) and cingulate cortex. There is also expression in part of the CA1 and dentate gyrus of the hippocampus. Olf, olfactory bulb; Sup. Coll, superior colliculus.
It is unclear what the direct downstream targets of Emx2 and Pax6 are, but Table 1 lists known examples of molecular asymmetry in the developing cortex, in spatial and temporal terms. Of special relevance is the work by Levitt and colleagues (Barbe and Levitt, 1991) (Levitt et al., 1997). They showed that neurons from an older part of the cortex are committed to express LAMP (limbic-associated membrane protein) very early in development; this protein continues to be expressed even when pieces of limbic cortex are exchanged with sensorimotor neocortex (Barbe and Levitt, 1991). It has also been shown that explants of mouse somatosensory cortex, marked by a lacZ transgene, retained their gene expression pattern even after transplantation into other areas of the brain, including the cerebellum (Cohen-Tannoudji et al., 1994). The monoclonal antibody PC3.1 recognises a 29 kDa protein expressed specifically in neurons of parietal cortex and when VZ cells are isolated three weeks before the onset of PC3.1 expression, they develop immunoreactivity corresponding to the cortical source (Arimatsu et al., 1992). Otx1 and Otx2, two mammalian homologues of the Drosophila homeobox gene Orthodenticle, are expressed early in a “nested” pattern within the developing forebrain and later are exclusively found in layer 5 and 6 neurons (Simeone et al., 1992; Frantz et al., 1994a). Another POU-specific homeodomain gene, SCIP, is found early in the preplate, subventricular zone, and among migrating layer 5 neurons only (Frantz et al., 1994b).

As mentioned, several cadherins have expression patterns that overlap particular neocortical regions (Nakagawa et al., 1999). Cadherins are cell-cell adhesion molecules which interact with one another homophilically and have been shown to regulate synapse formation, suggesting that they may control the extent of local connections within region or regulate their connectivity to subcortical targets which express cadherins of the same sort (Fannon and Colman, 1996; Nakagawa et al., 1999). The Ephrin family of cell-surface signalling molecules have also been found to have asymmetric distribution within the neocortex. Of much interest is Ephrin A5 which has graded expression across the somatosensory cortex (Vanderhaeghen et al., 2000). The barrel field is distorted in Ephrin A5 -/- animals with individual whisker receptive fields being changed both in area and the extent to which they overlap with surrounding whisker representations (Vanderhaeghen et al., 2000; Prakash et al., 2000). More dramatically, introducing an additional source of FGF8 produced a duplication of the somatosensory barrel fields (Fukuchi-Shimogori and Grove, 2001). These results represent the first functional evidence that genes are capable of defining the boundaries for specific sensory representation and their topography.

### Table 1

<table>
<thead>
<tr>
<th>Marker</th>
<th>Spatial distribution</th>
<th>Temporal distribution</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMP</td>
<td>Limbic cortex</td>
<td>E14 (rat) to adult; fated by E12</td>
<td>(Fannon and Colman, 1996)</td>
</tr>
<tr>
<td>Otx-1</td>
<td>Layers V/VI – anterior neocortex</td>
<td>E13 (rat) to adult</td>
<td>(Frantz et al., 1994a)</td>
</tr>
<tr>
<td>Lатexin</td>
<td>Layers V/VI – lateral isocortex</td>
<td>P6 (rat) to adult; fated by E12</td>
<td>(Arimatsu et al., 1992)</td>
</tr>
<tr>
<td>Transgene (H-221 mouse)</td>
<td>Layer IV – somatosensory cortex</td>
<td>P2 (mouse) to adult; fated by E14</td>
<td>(Cohen-Tannoudji et al., 1994)</td>
</tr>
<tr>
<td>Transgene (R197 mouse)</td>
<td>Layers II/III – cingulate and medial neocortex</td>
<td>E10.5 (mouse) to adult</td>
<td>(Tan, 1991)</td>
</tr>
<tr>
<td>SCIP</td>
<td>Layer V – neocortex</td>
<td>E13 (rat) to adult</td>
<td>(Frantz et al., 1994b)</td>
</tr>
<tr>
<td>Cadherin 6</td>
<td>Somatosensory, Auditory</td>
<td>P0 (mouse) to adult</td>
<td>Suzuki et al., 1997</td>
</tr>
<tr>
<td>Cadherin 8</td>
<td>Frontal</td>
<td>P0 (mouse) to adult</td>
<td>Suzuki et al., 1997</td>
</tr>
<tr>
<td>Ephrin-A5</td>
<td>Motor, Primary</td>
<td>From P19 (rat)</td>
<td>Vanderhaeghen et al., 2000</td>
</tr>
<tr>
<td></td>
<td>Somatosensory</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Auditoriy, Gradient of expression in secondary Somatosensory</td>
<td>From P0 (mouse)</td>
<td></td>
</tr>
<tr>
<td>Ephrin-A7</td>
<td>Caudal Sensory cortex, Motor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emx-2</td>
<td>Gradient of expression with caudo-lateral-rostral-medial within VZ.</td>
<td>From E8.5 (mouse) through embryogenesis</td>
<td>Simeone et al., 1992; Boncinelli et al., 1995; Bishop et al., 2000</td>
</tr>
<tr>
<td>Pax-6</td>
<td>Gradient of expression with rostro-medial-caudo-lateral</td>
<td>From E8.5 (mouse) through embryogenesis</td>
<td>Stoykova &amp; Gruss, 1994; Bishop et al., 2000</td>
</tr>
</tbody>
</table>

### Stage-Specific Programs of Gene Expression

Proper brain function requires proper assembly of neurons and establishment of circuitry. These two processes occur during two critical epochs of brain development, before and after birth. Before birth, neurons are generated in large numbers in germinal zones from where they migrate to their proper positions. After birth, these neurons mature by connecting to distant targets and to other neurons. Defective generation of neurons in sufficient numbers, or impediments to neuron migration to proper destinations are certain recipes for neurological disorders, including epilepsy and mental retardation. Similarly, failure to establish proper neuronal arborization during the early postnatal period can lead to functional neurological disorders later on in life. We review here why we have selected developing cortices from embryonic day 15 and postnatal day 1 for systematic comparisons of expressed genes.

At embryonic day 15, the timepoint at which neocortical neurogenesis is midway to completion in the mouse, the peak period of exponential increase of neuronal progenitor cells in the ventricular zone (VZ) has passed (Takahashi et al., 1996). At this stage, about half of the cellular constituents in the ventricular epithelium are made up of post-mitotic, migratory neurons which will eventually settle and further mature to form the cortical plate (CP). Specifically, cells destined to form layer IV, otherwise known as the granular cell layer, are born (Bayer and Altman, 1991). This layer is of special functional importance as a subset of these cells, the spiny stellates, are the primary receivers of thalamic afferent input and hence are intimately involved in areal regionalization of the neocortex.

This association is further strengthened by observations that the cells of layer IV are directly coupled to those of the subplate (Ghosh, 1995b). The subplate is composed of a transitory cell population which appears to be crucial in guiding mature connections between the neocortex and the thalamus. Pioneer projections from the subplate to the internal capsule (IC) may act as a temporary scaffold for thalamocortical afferents (TAs) and corticothalamic efferents (Allendoerfer and Shatz, 1994) (Molnar et al., 1998). The interaction between layer IV and subplate cells may be necessary for this targeting.

By embryonic day 15, TAs have reached the subplate but they then ‘mark time’ forming temporary contacts with cells in the subplate prior to their invasion of the CP. Speculatively, these axons form short-lived circuits with subplate neurons so that they might target their point of ingress into the CP by actively querying...
the underlying subplate. This is corroborated from in vitro co-culture experiments indicating that accurate pathfinding between thalamus and cortex is not dependent upon diffusible factors (Molnar and Blakemore, 1991). More pertinent, inhibition of action potentials in embryonic feline TAs in vivo results in guidance failure (Catalano and Shatz, 1998). Analogous experiments in which subplate neuronal activity was inhibited, resulted in non-characteristic terminal branching of TAs in layer IV (Herrmann and Shatz, 1995).

These results do not mean that molecular addressing is unnecessary in the cortex but rather such address molecules are likely to be cell-cell adhesion or signalling proteins necessary for stimulating or inhibiting synapse formation between thalamic axons and target or non-target tissues. This is supported by experiments where explants of specific thalamic nuclei whose axons, in vivo, are destined for a particular cortical region will, in vitro, show improved outgrowth on target compared to non-target cortical membrane preparations (Bolz and Gotz, 1992). It follows that if such ‘access’ molecules do indeed exist then the crucial time for their expression, and hence for our identification of them, will be when TA axons are interfacing with the subplate, i.e., at around embryonic day in the mouse.

In contrast neurogenesis has all but finished by postnatal day 1 in the mouse with only a few cells still migrating to the upper layers of the CP (Bayer and Altman, 1991). Additionally, thalamocortical afferents ‘waiting’ in the underlying subplate invade the CP and begin to form functional synapses with layer IV neurons (O’Leary et al., 1994). Therefore the fundamental processes of neurogenesis, long-distance migration and axonal pathfinding have essentially been completed. By this time it is less likely that molecules specifying the positional aspects necessary for regionalization will still be required. Molecules reflecting regional differences in synaptic plasticity, neurotransmission and those that stabilise expression of these proteins will be more likely to be identified at this time.

**The Technology**

As mentioned, any attempt to understand developmental patterning and regionalization of the mammalian cortex would benefit from a global, rather than an individual gene, approach to obtain complete gene expression profiles of functional regions at different stages of development. In the adult mouse brain, about 12,000 genes are expressed (Hastie and Bishop, 1976), reflecting the complexity of cell types in this organ. What would be useful is a systematic quantitative analysis with the power to catalog not only the known genes expressed but also to discover new genes, measure the abundance of transcripts, and identify regional differences in expression.

SAGE has been developed specifically to meet such challenges. The technique has been used effectively to characterize the identity and abundance of the entire set of mRNA transcripts expressed from the yeast genome (designated the transcriptome), identifying genes that had not been predicted from sequence information alone (Velculescu et al., 1997). More importantly, phenotypic expression of the significantly more complex human genome has also proved amenable to SAGE analysis. New models for p53-induced apoptosis (Polyak et al., 1997) and the APC pathway (He et al., 1998) in colorectal cancer cell lines have been proposed based on insights gained from SAGE data. Genes shown to be induced or repressed in human gastrointestinal tumors relative to expression levels in normal tissues represent potential diagnostic or prognostic markers, or even therapeutic targets (Zhang et al., 1997). Use of SAGE as a comparative tool to investigate different physiological or pathological conditions complements the power of the technique to provide a quantitative overview of gene transcription in a particular tissue or cell type. This feature has been recognized by researchers in the Cancer Genome Anatomy (CGAP) project who are currently using SAGE as part of a large-scale multi-pronged approach to “delineate the molecular fingerprint of the cancer cell” ((Lal et al., 1999); www.ncbi.nlm.nih.gov/ncicgap).

**Advantages of SAGE over Other Methods**

Of the range of techniques available to assess and compare levels of gene expression, none is as suited to global expression profiling as SAGE. Techniques such as northern blotting (Alwine et al., 1977), RNase protection (Zinn et al., 1983) or reverse transcription-polymerase chain reaction (RT-PCR; (Veres et al., 1987)) are mainly useful for the study of one, or a small number of, known genes. Differential display (Liang and Pardee, 1992) or cDNA subtraction-based methods (Hedrick et al., 1984) (Diatchenko et al., 1996) utilize hybridization to uncover differential expression of unknown genes on a small to medium scale. These techniques suffer from the inability to provide information on transcript abundance.

Large scale expressed sequence tag (EST) sequencing is also an effective approach to gene discovery, however it is laborious due to the length of the clones and a high level of redundancy. More than 1.1 million human EST’s have been found to collapse by UniGene clustering to only approximately 63,000 unique genes (Schuler et al., 1996). In the case of the mouse ESTs, sequencing from the (variable) 5’-end of cDNAs has artificially enhanced the number of clusters, or genes, which currently stands at 15,275 (http://www/ncbi/nlm/nih.gov/UniGene). These clusters contain nearly 270,000 gene and EST sequences, of which approximately 214,000 are derived from 5’-EST reads. Thus it can be seen that the mouse EST deposition rate is lagging far behind that for the human ESTs.

Compared to other techniques, such as DNA microarrays or large-scale EST sequencing, SAGE is truly “global” by design. Using SAGE, all transcripts containing the 4 bp anchoring enzyme site (NlaIII sites occur, on average, every 211 bp - more frequently than the predicted 256 bp interval; V. Velculescu, pers. comm.) have a probability of detection which is directly proportional to the abundance of the mRNA in the cell. The number of SAGE tags sequenced can be adjusted to ensure that even very rare transcripts are represented and there is a high probability that these rare transcripts will correspond to previously unknown genes. While the use of DNA microarrays may be warranted if the sole aim is to rapidly acquire expression data for known genes, there is no facility for new gene discovery using array technology. In addition, arrays are necessarily limited to the size of the clone sets available for a particular organism. In the case of the mouse, clones covering slightly over 20,000 known genes and ESTs are currently available from Genome Systems (http://www.genomesystems.com/). The most comprehensive cDNA arrays commercially available are the Genechips® produced by Affymetrix which at present cover less
than 30,000 mouse genes and expressed ESTs due to redundancies. Thus the coverage of the mouse genome, which is estimated to contain between 30,000 and 100,000 genes, is low (Silver, 1995). By way of comparison, Kinzler and coworkers (Zhang et al., 1997) analyzed gastrointestinal SAGE tags for more than 300,000 transcripts derived from about 49,000 human genes, 548 of which were differentially expressed. Low abundance transcripts expressed at less than 5 copies per cell comprised 86% of the total transcripts and 49% of the tags in the low abundance category (with no match in the Genbank database) represent yet to be discovered genes.

Theoretically, SAGE is infinitely sensitive, since the number of tags sequenced can be progressively increased until no new tags are collected. Increasing the number of yeast tags from 40,000 to 50,000 resulted in accumulation of only approximately 300 new genes (Velculescu et al., 1997). After collecting a total of 700,000 human tags from numerous gastrointestinal tissues and cell lines, the rate of new tag accumulation per 50,000 tag increment is essentially zero (V. Velculescu, pers. comm.). In practice, SAGE and cDNA microarray hybridization are quantitative techniques of similar sensitivity (approximately 1/100,000 or several copies per cell; (Hastie and Bishop, 1976; Zhang et al., 1997; Bowtell, 1999; Duggan et al., 1999). SAGE has the advantage, however, of generating quantitative data in the format of transcript copy per cell, or percentage of the total transcript pool. Valid comparisons can be made not only between levels of a particular transcript in different samples but also between levels of different individual transcripts in a single sample. In a library of SAGE tags, two transcripts of equal abundance have an equal probability of being represented regardless of length or sequence composition (providing an anchoring enzyme recognition site is present). Two single SAGE tags are ligated together to form ditags prior to any amplification step. Thus, the potential for biased PCR amplification of particular ditags (although minimal since all ditags are the same length) can be eliminated during tag extraction by discounting repeated occurrences of particular ditag combinations (Velculescu et al., 1995).

In the case of microarrays, on the other hand, relative data is generated and the only meaningful comparison is between test and reference levels of fluorescent signal intensity at a single position in the array. Different lengths and sequence compositions of cDNA clones across the array will result in varying hybridization kinetics and signal output. Comparison between physically separated array components could be confounded further by artifacts arising from construction of the chip itself and from the small-volume hybridization.

A major strength of the SAGE technique is that information about absolute transcript abundance in particular tissues or cell types is cumulative and can be stored. This feature allows cross-referencing and mining of data from other SAGE projects via the unique tag identifier number assigned to each tag. Because the tag identifier number is calculated from the sequence of the tag, consistency of experimental design in selection of the anchoring enzyme (specifying position) and tagging enzyme (determining tag length) is important. The vast majority of SAGE libraries constructed to date have used the anchoring enzyme NlaIII and the tagging enzyme BsmF1. In addition, modification of the tagging enzyme allows tag length to be increased from 10 to 21 (long SAGE), significantly increasing accuracy of tag to gene mapping (SAGE 2001 meeting, September 2001, San Diego, USA).

Currently, the cost of setting up a facility for production and/or analysis of microarrays is beyond the budget of most laboratories (Bowtell, 1999). The combined cost of an arrayer and scanner currently amounts to between US$100,000 and US$300,000. Hardware aside, the purchase price of the entire available clone sets for microarray production is listed at US$40,000-US$48,000 for a mouse set and US$90,000 for the human set. Alternatively, a single pass over the Affymetrix 30K mouse gene/EST Genechip® microarrays could cost up to US$12,000. Without further techno-
logical advances, this price is not expected to diminish rapidly since a major disadvantage of the currently available high-density microarrays on glass supports is that they are not reusable. While running a large-scale SAGE project is by no means cheap, the fundamental instrumentation (thermocyclers for PCR and automated DNA sequencing) is already installed in the majority of laboratories making the technique far more accessible.

Comparison of Genes expressed at Different Stages and Different Regions

We have completed a comparison of the two SAGE libraries, each containing approximately 20,000 tags (from each time point) representing approximately 11,000 unique transcripts (Gunnersen et al., 2002). This analysis has identified genes that are differentially expressed in these discrete developmental epochs. Notably, genes more highly expressed at E15 included known cell division genes and transcription factors of the bHLH, Sox, zinc finger, ring finger and winged helix gene families. To further compare regionally-expressed genes, we generated 40,000 tags from anterior cortex and compared an equal pool from posterior cortex (from embryonic day 15). Preliminary results suggest differential expression of roughly 200 to 300 different genes. The readily identifiable genes fit into several different categories including transcription factors (homeobox, zinc finger), signalling molecules (Wnt receptor, phosphotases, kinases), and translational proteins. Future work will compare SAGE libraries obtained from medial versus lateral regions of the developing neocortex (Fig. 4).

Conclusions

The SAGE technology outlined in this review offers a powerful tool for systematic and comprehensive molecular description of the developing cortex. Genetic attributes from all developmental stages of the cortex can be plumbed, in addition to mutant cortices arising from genetic variation. Although we have chosen the cortex as the object of our enquiry, the same rationale may be applied to other parts of the nervous system. More importantly, this approach has the potential to re-define the current way of studying brain development in molecular terms. No longer will the traditional gene by gene approach be considered sufficient or sophisticated enough for studying the problem. Instead, simultaneous read out of genetic activity will provide new frameworks for generating new hypotheses. Just as the dawn of molecular biology led to radical shifts in methodologies and paradigms for studying evolutionary theory, genome scale analysis of gene expression is expected to provide new understandings on the operation of genetic networks during brain development.

References


