DNA microarrays and beyond: completing the journey from tissue to cell

Jason C. Mills, Kevin A. Roth, Ross L. Cagan and Jeffrey I. Gordon

For the cell biologist, identifying changes in gene expression using DNA microarrays is just the start of a long journey from tissue to cell. We discuss how chip users can first filter noise (false-positives) from daunting microarray datasets. Combining laser capture microdissection with real-time polymerase chain reaction and reverse transcription is a helpful follow-up step that allows expression of selected genes to be quantified in populations of recovered cells. The voyage from chip to single cell can be completed using sensitive new *in situ* hybridization and immunohistochemical methods based on tyramide signal amplification.

Gasp! Why another review about DNA microarrays? Haven't there been enough excellent descriptions of this technology published recently (see, for example, refs 1, 2)? Like many cell biologists, we have used commercial versions of this technology over several years to profile gene expression in our favourite tissue. We too have had to find practical ways of processing the resulting deluge of information. The promise of microarrays for the cell biologist is to provide a more complete molecular view of cellular states and responses in complex tissues[3]. The challenge is to translate this information to the level of individual cells. In this commentary, we discuss new approaches to aid the journey from tissue to cell. These approaches include: methods for separating signal from noise in vast microarray datasets, and for sorting lists of genes; the use of laser-capture microdissection and real-time polymerase chain reaction with reverse transcription (RT-PCR) to interpret changes in gene expression quantitatively; and the application of sensitive, rapid *in situ*-hybridization methods to identify cells that contain messenger RNAs of interest. We throw in some of our own solutions to problems you are likely to face, and provide some (free!) software to help you along the way.

Choosing your array

Microarrays come in two basic flavours: complementary DNA (cDNA) or oligonucleotide arrays. cDNA arrays can be obtained commercially on filters, or made on glass slides. To compare mRNA levels in two RNA preparations using glass-slide microarrays, probes are made from each preparation and one set is labelled with Cy3 fluorescent dye, and the other with Cy5. After hybridization to a single microarray, relative mRNA levels can be determined from the Cy3/Cy5 signal for each gene (see http://cmgm.stanford.edu/pbrown/mguide/index.html for information about array manufacture, probe generation and hybridization protocols, and for data-analysis software). Affymetrix GeneChips (http://www.affymetrix.com; ref. 4) are a popular type of oligonucleotide array. These chips offer sophisticated internal controls; each GeneChip contains up to 6,500 different genes. Each gene is represented by at least one set of ~20 different 'probe pairs'. A probe pair consists of a 25-base-pair (bp) 'perfect-match' oligonucleotide probe and a 25-bp 'mismatch' probe, in which the 13th position is designed not to match the target sequence (in Affymetrix-speak the oligonucleotide 'probe' sequences are on the chip, and the 'target' is the cRNA made from cellular RNA). The information across all 20 paired probes (the 'probe set') is integrated by proprietary GeneChip software. The software compares mRNA levels in two RNA preparations by analysing probe-set signals from two GeneChips — one hybridized with cRNA made from the first RNA preparation, the other hybridized with cRNA generated from the second.

Regardless of the system chosen, array experiments yield far more information than we are used to processing. So the first step in analysing these large data sets is to separate signal from noise.

Dealing with the data avalanche I: filtering noise

GeneChips are a good model for discussing noise: they are popular, mass-produced and, as discussed above, come with internal controls designed to reduce noise. In comparisons of two distinct populations of RNA, GeneChips yield a false-positive rate of only 1–2% (refs 4, 5). However, 1% means 100 false positives in a comparison of 10,000 genes[2,6]. As you may be dealing with 'only' a few hundred real changes, this false-positive rate can wreak havoc on an experiment's signal-to-noise ratio.

How do you distinguish real signals from noise in a chip-to-chip comparison? To date, most chip users have used two approaches. One is to use a form of replicate analysis (Fig. 1a, b). The other is to impose an arbitrary threshold of signal difference between experimental and control samples (fold-change), above which differences are considered to be real[1,7]. Each approach has its limitations: carrying out replicated chip experiments is expensive, whereas using an arbitrary threshold for fold-change (for example, defining an increase or decrease of at least twofold as significant) means that potentially important and reproducible biological changes could be masked. Also, fold-change is a ratio: probe intensities only reflect expression differences linearly within a limited range; if either probe-set has hybridization intensities outside this range, the ratio will be skewed[1,7].

We have developed an empirical approach for eliminating noise from Affymetrix mouse GeneChip data sets. To develop this method, duplicate cRNAs were generated from a single preparation of mouse-organ RNA, and independently hybridized to a pair of chips (Fig. 1c). One array was arbitrarily designated as the 'baseline chip', and the other as its 'partner'. Genes with expression levels designated as 'increased' or 'decreased' by GeneChip software in a comparison of baseline and partner chips were defined as false-positives (noise). A three-dimensional plot was then be prepared, with the signal intensities of false positives on the baseline chip plotted on one axis, and the intensities of the signals from the corresponding probe-sets on the partner chip on another axis. The third axis was used to express each combination of baseline and partner chip intensities as a fraction of the total number of false positives on the entire grid. The location of false-positive signals (determined on 9 grids generated from 18 paired comparisons of 9 independent RNA preparations) was used to rank combinations of signal intensities from the baseline and partner

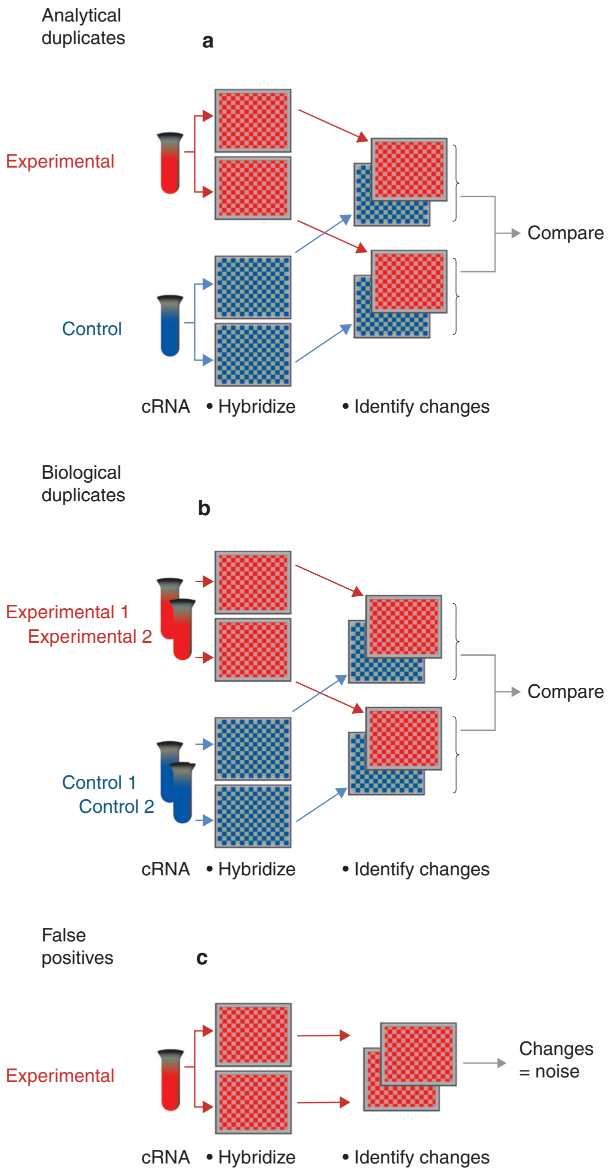

Figure 1 **Dealing with noise in microarray datasets.** a, b, **Strategies for replicated microarray analysis.** a, **Analytical duplication, in which two biologically distinct RNAs are compared. Duplicate cRNA targets are produced from each RNA sample; each cRNA is then hybridized to a chip. The expression profiles generated are compared as shown. Only replicated differences (such as a gene being designated as 'increased' in two separate comparisons) are selected for subsequent analysis.** b, **Biological duplication, in which a biological experiment is carried out twice. Each time, two RNA samples are obtained (for example, from an experimental and a control group). cRNA is generated from each RNA and each cRNA and is hybridized to a single chip. The comparison scheme shown is used to identify changes in gene expression that are exhibited in both experiments.** c, **Identification of the features of noise. A single RNA is used to produce duplicate cRNAs, which are then independently hybridized to a pair of microarrays. Transcripts designated as 'increased' or 'decreased' in a comparison of these paired chips are defined as false positives. By plotting the distribution of signal intensities across such comparisons, we have generated a series of look-up tables (LUTs, see text) for defining noise in Affymetrix GeneChip comparisons of biologically distinct mouse RNAs.**

chips that are most characteristic of noise. This ranking was expressed in the form of look-up tables (LUTs; see http://gordonlab.wustl.edu/mills), which allow noise (false positives) to be filtered from comparisons of biologically distinct RNAs. We have found that LUTs can eliminate 90% of noise, and that LUT-filtered changes in gene expression are more reproducible than those filtered by the customary (but arbitrary) requirement for a two- or threefold change in gene expression. For a good example of how to analyse noise in data sets from cDNA microarrays, see ref. 9.

Dealing with the data avalanche II: categorizing genes

Genes represented on arrays are not always well characterized (for example, they may be expressed sequence tags (ESTs)), and their functions are not necessarily apparent in the bioinformatic databases that are supplied with commercial arrays. Hence, once the noise has been filtered from a data set, the remaining genes that exhibit changes in expression must be accurately identified.

Fortunately, much of this gene 'data mining' can be done automatically. For example, lists of genes from array experiments can be imported into a spreadsheet application such as Microsoft Excel. 'Visual Basic for Applications' (VBA, the Excel macro language) can then be used to create a program that retrieves information from public databases about each gene on a list. VBA is straightforward: you don't need to be a software guru to design your own personalized search engine! Alternatively, we have expanded the VBA-based bioinformatics software developed in our laboratory to create a user-friendly package. The software takes lists of mouse, human or rat genes (identified by GenBank accession number) and queries the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene) to obtain the latest gene-specific information. Each gene is then hyperlinked to GenBank, TIGR, UniGene, and SwissProt for easy reference.

This software can be downloaded from our website (http://gordonlab.wustl.edu/mills). There are many other software options available for microarray analysis, both 'home-made' (like ours) and commercial[1] (see http://linkage.rockefeller.edu/wli/microarray).

Chip comparisons tend to flood the user with genes that span a large number of functional categories. Even nomenclature can be a problem: genes may be cloned independently and given two or more different names, complete with non-intersecting literature references. Fortunately, most key papers are now published online. In addition to PubMed and Medline, several sites are now devoted to collecting and organizing gene information. These include the ExPASy molecular-biology server (http://www.expasy.ch; see especially the 'NiceProt' summaries), Online Mendelian Inheritance for Man (OMIM; http://www.ncbi.nlm.nih.gov/Omim), the Munich Information Center for Protein Sequences (MIPS; http://www.mips.biochem.mpg.de), MedMiner (http://discover.nci.nih.gov/textmining/filters.html) and Genecards (http://bioinformatics.weizmann.ac.il/cards).

The principal aim is to assemble microarray hits into pathways and functional groups that provide an intelligible story of a cell's state, or its response to a stimulus. At present, there is no easy way to navigate through this stage of the analysis. Eventually, as increasing numbers of microarray data sets are collected and catalogued, certain combinations of gene-expression profiles will emerge as 'fingerprints' that identify particular pathways or cellular processes. For example, a stereotyped set of changes in a particular group of genes may identify a specific signal-transduction pathway, or suggest a stress response. Finding such modular responses[10]in data sets using this approach is not a pipe dream; there have been exciting recent steps towards establishing internet-searchable databases of compiled microarray results from different laboratories[11].

After a list of genes has been compiled and annotated, it is usual to select a subset of these genes to independently validate changes in their expression. We have found real-time, quantitative RT-PCR (qRT-PCR[12]) especially useful in this capacity. To reduce the cost of this expensive technique, SYBR green can be used to label PCR products[13].

Analysing the positives

The result of all the previous analysis should be a list of genes with validated changes in expression. As a cell biologist, now comes the fun part: determining the cellular basis of the altered gene expression in a particular tissue. Below, we outline some recently developed methods that should help.

Laser-capture microdissection. A useful way to define responses in quantitative terms within specific cell populations is laser-capture microdissection (LCM)[14,15] with real-time qRT-PCR. LCM uses a laser to capture small numbers of cells from frozen tissue sections[16]. The laser fuses each targeted cell to a film (http://www.arctur.com has information about a popular LCM system). The captured cells are then transferred to a test tube for isolation of RNA (see http://dir.nichd.nih.gov/lcm/lcm.htm for information about LCM, LCM protocols, and links to cDNA sequence databases generated from microdissected cell populations). An exciting direction for the future is the use of LCM as the starting point for proteomic analyses[17,18].

The precision of LCM allows the resolution of quantitative gene-expression analysis to be increased from the level of a tissue to one or more of its component cell populations. However, a significant hurdle is devising a way to mark those cell populations without degrading their RNA. One way is to cut serial sections and then label one section with cell-type-specific markers. An electronic image of the labelled section can then be used as a template to guide dissection of adjacent, unlabelled sections (we called this method 'navigated' LCM[19]).

Another challenge of LCM is to obtain sufficient amounts of RNA from captured cell populations. Fortunately, protocols are constantly improving (see http://dir.nichd.nih.gov/lcm/Protocol.htm for the latest). A typical good yield is 10–30 pg of RNA per cell. A few hours of laser capturing can isolate several thousand cells, more than enough for qRT-PCR validation of the level of a transcript of interest (we have found that 500 cell equivalents is enough for a single determination).

Sensitive methods for *in situ* hybridization. As a cell biologist, you will probably want to examine microarray-identified changes in gene expression at single-cell resolution.

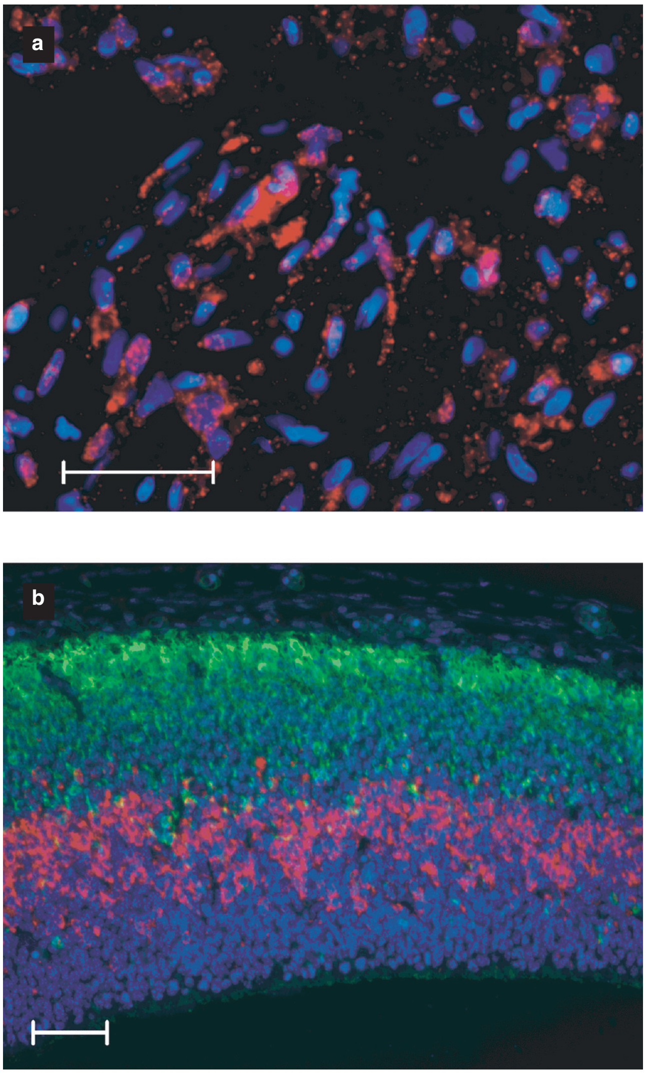

Figure 2 **Illustration of tyramide signal amplification (TSA)-based *in situ* hybridization (ISH).** a, **ISH detection of Bcl-X mRNA using a Dig-labelled cRNA probe and TSA Plus Direct Cy3 in a human brain tumour. Numerous Bcl-x expressing cells (red) are evident in this formalin-fixed, paraffin-embedded section of human glioblastoma multiforme. Cell nuclei are labelled with Hoechst 33,258 (blue).** b, **Dual ISH and immunohistochemical detection in embryonic mouse brain. Proliferating cells in the forebrain of a Bouin's fixed, paraffin-embedded mouse embryo were identified using a Dig-labelled H4 cRNA probe, and TSA Plus Direct Cy3 (red). Postmitotic neurons were identified using antibodies against microtubule-associated protein 2 (MAP2) and TSA Plus Direct fluorescein (green). Cell nuclei are labelled with Hoechst 33,258 (blue). Cells containing H4 mRNA are apparently restricted to the inner half of the section, in a region known as the ventricular zone, whereas MAP2-immunoreactive cells are largely confined to the outermost portion of the forebrain, which corresponds to the developing mantle zone. Scale bars, 50 μm.**

Fortunately, *in situ*-hybridization (ISH) techniques have become much simpler and quicker with the advent of detection using tyramide-signal amplification (TSA). TSA-based methods are faster and more sensitive than traditional ISH. Furthermore, they are compatible with a variety of tissue-fixation protocols and can be used in conjunction with immunohistochemical staining to identify cell populations accurately[20–22].

TSA is based on horseradish peroxidase (HRP)-catalysed deposition of labelled tyramine at sites of probe binding (see http://www.nen.com/products/tsa). HRP converts tyramine into a highly reactive, oxidized intermediate that binds rapidly and covalently to cell-associated proteins located at or near the HRP-linked probe. Signals can be detected at this stage by using a tyramine–fluorophore conjugate (this is known as the 'TSA direct method'). Alternatively, tyramine conjugated to non-fluorescent intermediates such as biotin can subsequently be detected using fluorescently labelled or enzyme-linked streptavidin (the 'TSA indirect method').

We have recently developed a simple protocol for ISH using second-generation TSA Plus reagents (NEN Life Sciences)[23]. This method can be carried out on either frozen or paraffin-embedded tissue sections (Fig. 2a), and the turn-around time can be as little as 8 h from probe application to probe detection. The protocol can also be modified for multi-label ISH or for dual ISH and immunohistochemical detection (Fig. 2b). Unlike antibody-dependent staining, once tissue fixation and processing are optimized, probes made from any gene should work equally well, facilitating the high-throughput that microarray users need. All in all, TSA–ISH seems poised to become the method of choice for interpreting tissue-based microarray results at single-cell resolution. Its role may be cemented over the next few years, as systems are being developed to automate hybridization and labelling.

Tomorrow

What about the future? First, Moore's 'law' — that technology doubles the capacity of microchips roughly every 18 months — may not apply to microarrays. In the case of microarrays, the rate of increase in the number of different sequences available on chips may double at a faster rate than this. We will probably have chip sets containing whole mammalian genomes within a few years, and costs should continue to tumble.

In the short term, methods will be developed so that LCM can be coupled to microarray-based analysis, allowing gene expression to be comprehensively profiled in small numbers of cells or even in single cells. The current limitation is low RNA yield, so new techniques must be developed to amplify sequences before hybridization to chips[24,25].

Once expression analyses reach this resolution, we will undoubtedly face new aspects of biological variation. For example, in simpler organisms such as *Caenorhabditis elegans* and *Drosophila*, seemingly equivalent neighboring cells can exhibit markedly different patterns of gene expression and developmental fates. Furthermore, in a comprehensive study of gene-expression profiles in yeast, untreated isogenic cells showed significant fluctuation in gene expression between 'identical' cultures[8]. Microarrays and associated technologies will allow us to explore further the biological variation within and between cell lineages. An even more dynamic view of what constitutes a cell is on the horizon. □

The authors are in the Department of Molecular Biology and Pharmacology (J.C.M. and K.A.R. are also in the Department of Pathology and Immunology), Washington University School of Medicine, St Louis, Missouri 63110, USA.
e-mail: cagan@molecool.wustl.edu

1. Bassett, D. E. Jr, Eisen, M. B. & Boguski, M. S. Gene expression informatics — it's all in your mine. *Nature Genet.* **21**, 51–55 (1999).
2. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
3. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 742–752 (2000).
4. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
5. Lee, C. K., Weindruch, R. & Prolla, T. A. Gene-expression profile of the ageing brain in mice. *Nature Genet.* **25**, 294–297 (2000).
6. Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **8**, 1821–1832 (1999).
7. Der, S. D., Zhou, A., Williams, B. R. & Silverman, R. H. Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 15623–15628 (1998).
8. Mills, J. C. & Gordon, J. I. A new approach for filtering noise from high density oligonucleotide microarray datasets. *Nucleic Acids Res.* **29**, e72 (2001).
9. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
10. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
11. Brazma, A., Robinson, A., Cameron, G. & Ashburner, M. One-stop shop for microarray data. *Nature* **403**, 699–700 (2000).
12. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).
13. Steuerwald, N., Cohen, J., Herrera, R. J. & Brenner, C. A. Analysis of gene expression in single oocytes and embryos by real-time rapid cycle fluorescence monitored RT-PCR. *Mol. Hum. Reprod.* **5**, 1034–1039 (1999).
14. Emmert-Buck, M. R. *et al.* Laser capture microdissection. *Science* **274**, 998–1001 (1996).
15. Goldsworthy, S. M., Stockton, P. S., Trempus, C. S., Foley, J. F. & Maronpot, R. R. Effects of fixation on RNA extraction and amplification from laser capture microdissected tissue. *Mol. Carcinog.* **25**, 86–91 (1999).
16. Jin, L. *et al.* Analysis of anterior pituitary hormone mRNA expression in immunophenotypically characterized single cells after laser capture microdissection. *Lab. Invest.* **79**, 511–512 (1999).
17. Banks, R. E. *et al.* The potential use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis — preliminary findings. *Electrophoresis* **20**, 689–700 (1999).
18. Simone, N. L. *et al.* Sensitive immunoassay of tissue cell proteins procured by laser capture microdissection. *Am. J. Pathol.* **156**, 445–452 (2000).
19. Wong, M. H., Saam, J. R., Stappenbeck, T. S., Rexer, C. H. & Gordon, J. I. Genetic mosaic analysis based on Cre recombinase and navigated laser capture microdissection. *Proc. Natl Acad. Sci. USA* (in the press).
20. van Gijlswijk, R. P. *et al.* Fluorochrome-labeled tyramides: use in immunocytochemistry and fluorescence *in situ* hybridization. *J. Histochem. Cytochem.* **45**, 375–382 (1997).
21. Wiedorn, K. H., Kuhl, H., Galle, J., Caselitz, J. & Vollmer, E. Comparison of *in situ* hybridization, direct and indirect *in situ* PCR as well as tyramide signal amplification for the detection of HPV. *Histochem. Cytochem.* **111**, 89–95 (1999).
22. St Croix, B. *et al.* Genes expressed in human tumor endothelium. *Science* **289**, 1197–1202 (2000).
23. Zaidi, A. U., Enomoto, H., Milbrandt, J. & Roth, K. A. Dual fluorescent *in situ* hybridization and immunohistochemical detection with tyramide signal amplification. *J. Histochem. Cytochem.* **48**, 1369–1376 (2000).
24. Ohyama, H. *et al.* Laser capture microdissection-generated target sample for high-density oligonucleotide array hybridization. *Biotechniques* **29**, 530–536 (2000).
25. Baugh, L. R., Hill, A. A., Brown, E. L. & Hunter, C. P. Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Res.* **29**, E29 (2001).

ACKNOWLEDGEMENTS
We thank L. Hooper, D. Syder, T. Stappenbeck, M. Wong and P. Sullivan for helpful comments. Work in our laboratories cited here was supported by grants from the NIH and from NEN Life Science Products (to K.A.R.). J.C.M. is a postdoctoral fellow of the Howard Hughes Medical Institute.