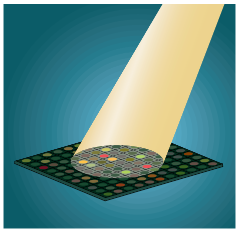
Array of hope

Eric S. Lander

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. e-mail: lander@genome.wi.mit.edu



Bob Crimi

Genomics aims to provide biologists with the equivalent of chemistry's Periodic Table[1]—an inventory of all genes used to assemble a living creature, together with an insightful system for classifying these building blocks. A short decade ago, the task of enumeration alone appeared to many to be a quixotic quest. Whereas chemical matter is composed of a mere hundred or so elements, organismal parts lists are huge—running into the thousands for bacteria and hundreds of thousands for mammals. Genomic mapping and sequencing, however, has steadily extended its dominion: it has domesticated the Megabase and will tame the Gigabase in the not-too-distant future.

The next great challenge is to discern the underlying order. The Periodic Table summarized chemical propensities in its rows and columns, and thereby foreshadowed the secrets of subatomic structure. Understanding biological systems with 100,000 genes will similarly require organizing the parts by their properties. The Biological Periodic Table will not be two-dimensional, but will reflect similarities at diverse levels: primary DNA sequence in coding and regulatory regions; polymorphic variation within a species or subgroup; time and place of expression of RNAs during development, physiological response and disease; and subcellular localization and intermolecular interaction of protein products. The traditional gene-by-gene approach will not suffice to meet the sheer magnitude of the problem. It will be necessary to take 'global views' of biological processes: simultaneous readouts of all components.

Arrays offer the first great hope for such global views by providing a systematic way to survey DNA and RNA variation. They seem likely to become a standard tool of both molecular biology research and clinical diagnostics. These prospects have attracted great interest and investment from both the public and private sectors. The reviews in this supplement describe important issues in this fast-moving area[2–12].

Technical foundations

The field has evolved from Ed Southern's key insight, one-quarter of a century ago, that labelled nucleic acid molecules could be used to interrogate nucleic acid molecules attached to a solid support[13]. The Southern blot was the first array. It was only a small step to filter-based screening of clone libraries, which introduced a one-to-one correspondence between clones and hybridization signals. The next advance was the use of gridded libraries, stored in microtitre plates and stamped onto filters in fixed positions; each clone could be uniquely identified and information about it accumulated. Powerful applications were soon envisaged; several groups explored expression analysis by hybridizing mRNA to cDNA libraries gridded on nylon filters. The ideas were sound, but the implementation still clumsy.

The explosion of interest in array technologies has been sparked by two key innovations. The first was the use of non-porous solid supports, such as glass, which has facilitated miniaturization and fluorescence-based detection. About 10,000 cDNAs can be robotically spotted onto a microscope slide and hybridized with a double-labelled probe, using protocols pioneered by Pat Brown and colleagues[14]. The second was the development of methods for high-density spatial synthesis of oligonucleotides. Steve Fodor and colleagues have adapted photolithographic masking techniques used in semiconductor manufacture to produce arrays with 400,000 distinct oligonucleotides, each in its own 20 µm^2 region[15]. Other companies are developing *in situ* synthesis with reagents delivered by ink-jet printer devices.

The new generation of array technologies is still in its infancy. As one reviewer wryly notes[8], the scientific literature contains more reviews about arrays than primary research papers applying them. The techniques have become established in only a few places. The tools remain prohibitively expensive for many laboratories (owing to the actual capital cost of setting up an arraying facility or the amortized capital costs reflected in the purchase price of arrays). Still, these problems are likely to be solved by economies of scale, free-market competition and time—just as they are for new generations of computer microprocessors.

RNA expression

The focus of most current array-based studies is the monitoring of RNA expression levels. The tools are most comprehensive for the yeast *Saccharomyces cerevisiae*. It is possible to create both oligonucleotide and cDNA arrays containing detectors for all yeast genes, thanks to the availability of the complete yeast genome sequence, and it is easy to detect expression levels considerably below one message per cell owing to the relatively low complexity of the yeast genome. Yeast geneticists have recently begun reporting global expression studies of such fundamental processes as mitosis and meiosis[11].

The tools are also quite powerful for mammalian genomes, albeit with room for improvement. Arrays containing 5,000–10,000 genes are already in common use, and current protocols allow reliable detection of messages present at several copies per mammalian cell. Incremental advances are likely to improve both fabrication and sensitivity. It seems safe to predict that, not long after the turn of the century, researchers will be able to purchase standardized oligonucleotide and cDNA arrays containing the complete sets of 100,000 human and mouse genes.

The challenge is no longer in the expression arrays themselves, but in developing experimental designs to exploit the full power of a global perspective. The issues are both technical and conceptual.

Some of the most important biological applications involve studying very small target tissues—for example, the apical ectodermal ridge in a developing limb; tumour cells embedded in a sea of surrounding stromal cells; or a particular class of neuron in the cortex. A reliable quantitative amplification procedure is very much needed.

Experimental manipulations will also need to be rigorously controlled. Responses to microenvironment (for example, the position of a culture dish in an incubator or the time of day at which an assay is performed) pose a special risk of misleading global expression studies, in which one is fishing through 100,000 genes to find the small subset that vary. It is well known among *aficionados* that comparison of the 'same' experiment performed a few weeks apart reveals considerably wider variation than seen when a single sample is tested by repeated hybridization.

The greatest challenge, however, is analytical. The first expression profiling experiments involved comparing just two samples, with the aim of identifying those genes whose expression levels differed (for example, in metastatic versus nonmetastatic derivatives of a tumour cell line). Deeper biological insight is likely to emerge from examining datasets with scores of samples—for example, multiple time points from multiple cell lines treated independently with multiple growth factors. Each gene defines a point in k-dimensional space (where k is the number of samples studied), and functional similarities are likely to reveal themselves as 'clusters' in this space. Computational scientists working in the field of 'data mining' have devised a dizzying assortment of techniques for clustering, predicting and visualizing patterns in high-dimensional space—most based on inherent assumptions about the types of patterns to be found. Empirical exploration will be needed to flesh out which types of datasets and analytical tools will be most fruitful for biology.

How well can causation be inferred from correlation? The problem is akin to inferring the design of a microprocessor based on the readout of its transistors in response to a variety of inputs. The task is impossible in a strict mathematical sense, in that the microprocessor layout could be arbitrarily complicated, but is likely to prove at least somewhat tractable in a more constrained biological setting, especially when combined with ways to cut specific wires in biological circuits using antisense and related techniques. The great opportunities ahead would well justify an influx of bright young computational scientists and technologists into biology.

DNA variation

Arrays can also be used to study DNA, with the primary application being identification and genotyping of mutations and polymorphisms. These applications pose rather different challenges than RNA expression monitoring, and many issues remain to be worked out.

Identification of novel DNA variants has largely been the province of oligonucleotide, as opposed to spotted, arrays[7,9]. Exploiting the ability to perform custom synthesis at high density, one can construct a 'tiling' array to scan a target sequence for mutations. Each overlapping 25-mer in the sequence is covered by four complementary oligonucleotide probes that differ only by having A, T, C or G substituted at the central position. An amplified product containing the expected sequence will hybridize best to the expected probe, whereas a sequence variation will typically alter the hybridization pattern. Such tiling arrays have been used to detect variants in such targets as the HIV genome, human mitochondria and the gene encoding p53. In such specific settings, the process can be optimized to have high specificity and sensitivity.

The approach has also been used for much larger surveys—for example, a set of more than 100 tiling arrays were used to scan for single nucleotide polymorphisms (SNPs) in 21,000 sequence tagged sites (STSs) covering more than 2 Mb of genomic DNA[16]. Such tiling arrays are powerful tools, but still imperfect. In high-throughput applications, homozygous variants are readily detected, but heterozygotes may be missed. (The expected pattern is abolished in the first case, but is still present at one-half the intensity in the second case and sometimes dominates the alternative pattern.) Single-base substitutions will result in a perfect match to one of the four alternative probes at the position, but deletions and insertions will not be specifically recognized unless a corresponding probe is included in the array.

In theory, vast regions could be surveyed for DNA variation. If today's feature size could be reduced 20-fold to 1 μm, 100 Mb could be surveyed on a single 2 cm×2 cm array—and an entire human genome on 30 arrays. Before such fantasies can be realized, major hurdles need to be overcome. First, the requirement for specific target amplification must be circumvented. Present techniques allow hybridization of total mammalian RNA, but not genomic DNA, which has 100-fold greater complexity. Each target locus requires developing a specific PCR assay. (Notably, genome-wide surveys of DNA variation are feasible in yeast, owing to small genome size and ability to work with haploids[17].) Second, extreme miniaturization will require the development of more sensitive labelling and detection techniques.

Genotyping of large sets of known DNA variants is a rather different issue. Oligonucleotide arrays have already been synthesized containing specific detectors for each allele at many loci[16]. A better approach, however, may be to fabricate a generic array containing 'tag sequences', as originally described for certain yeast applications[18]. If primers for each locus are tailed with a unique tag sequence, genotyping reactions can be performed and then hybridized to the generic array so that each assay anneals to its corresponding address. Our group has been developing such an approach using fluorescently labelled single-base extension reactions (P. Sklar & J. Hirschhorn, unpublished data) and others have been similarly employing oligo-ligation assays. Such generic detectors can be created with both oligonucleotide arrays and spotted arrays.

Tremendous hype has surrounded potential applications of human variants such as SNPs. Several biotech firms seem prepared to spend $100 m or more to create private databases of SNPs in the hope that pharmaceutical firms will feel compelled to pay huge licence fees to gain access. Yet, there remain fundamental open questions about human population genetics[12]—including the role of common genetic variants in causing human disease, the extent of linkage disequilibrium (ancestral segments) across the human genome and the nature of variation within and between populations—that must be resolved before the real utility of SNPs in distinctive settings (for example, basic research versus clinical trials) becomes clear.

As nucleic acid arrays begin to penetrate the research community, technologists are already entertaining visions of protein arrays, antibody arrays and cell arrays as well as non-array–based global readouts. Molecular biology is rising to the challenge of exploiting the comprehensive description in biology's Periodic Table. Still, it is worth remembering that this only brings biology to the point that chemistry reached at the start of the twentieth century. As chemical phenomena such as buckyballs and high-temperature superconductors remind us, there will still be wonderous surprises not even hinted at in the Periodic Table.

1. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
2. Southern, E. Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nature Genet*. **21**, 5–9 (1999).
3. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. Expression profiling using cDNA microarrays. *Nature Genet.* **21**, 10–14 (1999).
4. Cheung, V.G. *et al.* Making and reading microarrays. *Nature Genet.* **21**, 15–19 (1999).
5. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
6. Bowtell, D.L. Options available—from start to finish—for obtaining expression data by microarray. *Nature Genet.* **21**, 25–32 (1999).
7. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
8. Cole, K.A., Krizman, D.B. & Emmert-Buck, M.R. The genetics of cancer—a 3D model. *Nature Genet.* **21**, 38–41 (1999).
9. Hacia, J. Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genet.* **21**, 42–47 (1999).
10. Debouck, C. & Goodfellow, P. DNA microarrays in drug discovery and development. *Nature Genet.* **21**, 48–50 (1999).
11. 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).
12. Chakravarti, A. Population genetics—making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
13. Southern, E.M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol*. **98**, 503–517(1975).
14. Schena, M. *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
15. Fodor, S.P.A. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).
16. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
17. Winzeler, E.A. *et al.* Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197 (1998).
18. Shoemaker, D.D. *et al.* Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genet.* **14**, 450–456 (1996).