© 1997 Nature Publishing Group http://www.nature.com/naturegenetics

A new dimension for the human genome project: towards comprehensive expression maps

Tom Strachan[1], Marc Abitbol[2], Duncan Davidson[3] & Jacques S. Beckmann[4]

The current Human Genome Project is largely devoted to structural characterisation of our genome. We now need international co-ordination of a second phase of genome analysis, the systematic construction of expression maps using both basic and high-resolution expression assays. Databases recording different types of expression pattern for a variety of human cell types need to be established and co-ordinated. There is a compelling need for a database of gene expression in early human development, but the scarcity of human material for study requires optimisation of research strategies and co-ordination of expression studies in early human and mouse development.

The Human Genome Project has entered its final phase: several large-scale DNA sequencing projects are underway and destined to yield a wealth of sequence data that will occupy bioinformaticians for years to come. The current effort is in *structural genomics*, devoted largely to determining the exact chromosomal positions and structural organisations of our 80,000 or so genes. In the post-sequencing era[1], efforts in *functional genomics* are being contemplated[2]. This is the route now being pursued by yeast geneticists following the successful completion of the *Saccharomyces cerevisiae* genome project[3,4] and it is understandable that we should seek insights from model organisms regarding gene function[5].

With regard to our own genome, we can expect a progressive expansion in functional studies of genes using human and animal cultured cell lines and transgenic animals. Knowledge of detailed gene structure will not, however, be sufficient to mount such functional studies: there will also be a need to expand our limited knowledge of human gene *expression*, systematically documenting expression patterns in various human organs, tissues and cell types. This is particularly important in early human development, where access to suitable samples for study has been limited, consequently placing heavy reliance on animal models. Our advocacy of this comparatively neglected field of human genome investigation is motivated by several factors, notably the growing awareness of the limitations involved in extrapolating from gene expression studies in animal models. We therefore advocate co-ordination of an international effort to document systematically human gene expression using various methods including surveys of high-resolution expression patterns in cells and tissues, and to store the data in publicly accessible electronic databases.

Human gene expression mapping

How should we prepare for a new phase of genome study devoted to gene expression? Many different expression assays are possible, and one is already being comprehensively undertaken — the characterisation of cDNA sequences from a variety of different cell types. In addition to providing a human gene catalogue, a major rationale of such projects is to place genes on physical maps[6,7] before large-scale sequencing. These data will be invaluable in clarifying the structure of genes and their transcripts, and eventually in identifying expressed polymorphic variants. But such approaches are not suited as rapid general-expression screens: the relative abundance of individual gene transcripts in different cDNA libraries provides only a crude measure of comparative transcriptional activity[8,9]. In contrast, powerful technologies have been devised which may provide rapid and more meaningful genome-wide expression screens. In addition, studies of gene expression *in situ* offer high-resolution expression maps.

Towards rapid whole-genome expression screens. Routine methods for surveying human gene expression have usually been limited in capacity, being used typically to track the expression of one or a few genes at a time in isolated RNA or protein samples. RNA blot hybridization, ribonuclease protection assays, RT-PCR and western blotting are well-established examples and provide valuable information on the relative abundance of expression products and often on other characteristics such as product size. Methods designed to survey the expression patterns of many genes simultaneously have become popular. Two-dimensional (2D) gel electrophoresis for assaying polypeptide expression has a long history[10], but at the transcript level, powerful methods of surveying differential expression of many genes have only recently been developed, including differential display[11], the SAGE method[12,13], and multiplex hybridization assays using complex microarrays of cDNA clones[14–18] or of gene-specific oligonucleotides[18–20].

Differential display and derivative techniques are valuable in identifying novel genes and in probing expression of complex, but not fully characterized, genomes. When essentially all human genes have been identified, however, they will be less powerful than alternative methods which can survey simultaneously the transcriptional profiles of all genes. The SAGE method has a high throughput but may be disadvantaged by certain features, such as the reliance on extremely short sequence tags to identify individual genes, and may not be fully amenable to automation. In contrast, the microarray technologies, in particular, are very well suited to automation. They involve depositing complex arrays of indi-

[1]Department of Human Genetics, University of Newcastle upon Tyne, UK, [2]Laboratoire d'Embryologie Humaine, Université René Descartes, Paris, France, [3]MRC Human Genetics Unit, Western General Hospital, Edinburgh, UK, [4]Généthon, Evry, France and Fondation Jean Dausset-CEPH, Paris, France. Correspondence should be addressed to T.S. or J.S.B. e-mail: tom.strachan@ncl.ac.uk. e-mail:beckmann@genethon.fr

© 1997 Nature Publishing Group http://www.nature.com/naturegenetics

vidual cDNAs or gene-specific oligonucleotides onto glass 'chips' in a miniaturized format, in the former case by high-speed robotic printing and in the latter, by *in situ* synthesis using a combination of photolithography and oligonucleotide chemistry that permits independent specification of individual oligonucleotides[20]. The resulting 'DNA chips' permit simultaneous monitoring of the expression of thousands of human genes in normal or disease cells[15,16,19] (Fig. 1), and numerous applications can be envisaged. For example, there is considerable diagnostic potential in assaying for inappropriate gene expression associated with the onset of cancers. In addition, by comparing the expression profiles of genes in disease tissues with those in equivalent normal tissues, they may help to identify strong candidate disease genes, or novel pharmaceutical targets.

Once a comprehensive human gene catalogue is in place, the availability of sequence information for each gene will enable rapid identification of gene products. At the level of 2D gel electrophoresis, for example, individual proteins may be identified by comparing observed and expected radiolabelling patterns following incorporation of specific amino acids labelled with different radioisotopes[21]. The application of this general approach to studying human proteins is, however, handicapped by technical limitations: the method is difficult to standardize and automate, and the resolution is limited to only a few thousand of the more abundant proteins in the starting sample. Instead, recent major technological improvements in mass spectroscopy may provide a more profitable approach. Total protein fractions can be separated chromatographically, individual fractions proteolytically cleaved and the cleaved peptides analysed by mass spectroscopy, in combination with database searching[22].

High-resolution *in situ* gene expression studies. Such studies attempt to record the original spatial distribution of gene products in tissues and cells. Both the transcriptional and translational expression status can be monitored by using labelled nucleic acid probes or antibodies, respectively. Although normally represented in two dimensions, the expression images can be electronically digitised and the digitised profiles from serial tissue sections can be integrated to generate three-dimensional expression maps[23,24]. By studying equivalent tissue sections at different times and stages of development, detailed spatiotemporal expression maps can be electronically documented. Using appropriate oligonucleotide or antibody probes, it is also

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Fig. 1 Monitoring human gene expression using complex microarrays of cDNA clones or oligonucleotides. ***a***, cDNA clone array. A 1.0-cm^2 DNA microarray containing 1,046 cDNA clones from a whole human blood cell cDNA library was hybridized with a fluorescent labelled cDNA probe prepared by reverse transcription from human bone marrow mRNA. Confocal laser scans of the fluorescence label are represented in a pseudocolour rainbow scale to indicate rough quantitation of expression levels: background levels being violet, then progressing through indigo, blue green, yellow, orange to red, the most abundant expression. The parallel format of the array allows precise comparisons and differential expression measurements[14]. (Image courtesy of Mark Schena, Stanford.) ***b***, Oligonucleotide array. A fluorescence image of a 1.28-cm x 1.28-cm array containing over 65,000 different oligonucleotide probes for 1,641 unique genes (each gene is represented by 20 oligonucleotide pairs, with each pair including a faithful gene sequence oligonucleotide an a related mismatched oligonucleotide). The array also includes RNA standards that are used as quantitative references (β-actin, glyceraldehyde-3-phosphate dehydrogenase, transferrin receptor and one phage and three bacterial RNAs). The image was obtained following overnight hybridization of a sample made directly from human tissue, and was from one of four such arrays, collectively sampling 6,500 different human genes. (Image courtesy of David Lockhart, Affymetrix Inc.)

© 1997 Nature Publishing Group http://www.nature.com/naturegenetics

possible to track individual isoforms, gaining information on their temporal order of appearance, tissue-specificity and so on. Overlapping spatiotemporal expression domains for specific isoforms of non-allelic genes may also identify functionally important interactions that could otherwise elude detection.

High-resolution expression profiles can afford new insights into gene function, which could be especially valuable in the case of recently identified disease genes. They will not be able to provide unequivocal demonstrations of functional interactions between genes, but comprehensive expression studies will eventually indicate sequential hierarchies or patterns of gene activity. When combined with other functional information on the genes being investigated, such data will help unravel the multilayered complexity of developmental or physiological *expression programmes* (including not just the traditional A1→A2 →A3 vertical pathways of gene activation, but complex circuits resulting from additional horizontal interactions between members of different vertical pathways).

Additional important applications will be available for some disease genes. Comparative expression studies of plausible candidate genes may be valuable in identifying particular promising genes to study, especially in cases where the phenotype often results from microdeletions. Species differences in gene expression may also be uncovered, offering explanations as to why the phenotypes of animal disease models often do not resemble their human counterparts, and this knowledge may lead to new insights into molecular pathogenesis, as in Usher's syndrome 1B (see below). In addition, by identifying previously unforeseen expression domains, high-resolution gene expression studies may be useful for the development of new therapeutic strategies and in assessing potential side effects of new therapeutic reagents. Finally, comparative high-resolution expression studies may provide detailed insights into species differences, and may lead to a form of *comparative molecular anatomy*.

Recording human expression data and establishing priorities. There are several gene-expression databases for a variety of model organisms, including *in situ* gene expression databases[25]. In contrast, electronic documentation of human gene expression patterns is at a primitive stage: although some databases have been initiated to document 2D protein expression patterns in individual tissue types and cell lines[26,27], databases of transcriptional profiles are currently limited to EST/cDNA collections[6,8]. There is a need, therefore, for additional databases, systematically documenting basic and high-resolution gene expression patterns in a variety of human tissues and cell types. Most likely, different databases will be established in different centres, reflecting local expertise in particular types of expression assay or target tissues. Provided that the expression data are compatible and that suitable electronic links are established between the different expression databases and with other genome/gene databases (including future gene-function databases), the diversity of databases should not prove a major obstacle to efficient data retrieval.

What are the priorities in documenting human gene expression? Which tissues should be sampled first, and which expression assays will provide the most valuable information? Except in the case of early human development (see below), there will be no major difficulties in obtaining suitable material for study, and future databases can be expected to document expression in a variety of adult tissues and cell types. Expression patterns will also be documented for certain disease states, notably some cancers. Already, a pilot survey of gene expression using DNA microarray technology has been reported in a melanoma cell line[16], and the recently proposed Cancer Genome Anatomy Project envisages similar surveys of gene expression in a variety of common cancers at the precancerous and malignant stages in addition to cDNA sequencing projects[28,29]. Equivalent surveys may be conducted for other disease states, and will be most informative in cases where the differences map to only a very few genes. In contrast, if there are many differences (for example, resulting from pleiotropic effects of a single mutation, complex gene–gene/ or gene– environment interactions), interpretation could be difficult because of our lack of understanding of detailed expression programmes.

Basic expression screens vs. high-resolution assays. Rapid whole genome expression screens offer a powerful 'first pass' expression screen, albeit at low resolution. At the protein level, it is difficult to envisage how the current technologies could be extended to whole genome expression screening, but transcriptional profiling is quite different. In particular, there have been spectacular recent advances in microarray technology whereby scaling up is both simple and inexpensive. Now, using information in the available EST databases, researchers at Affymetrix are making arrays to monitor the expression of some 50,000 human genes. One approach uses a single 1.28-cm^2 array containing 400,000 oligonucleotides. Further technological improvements could permit simultaneous sampling of every human gene on even smaller arrays with a quantitative sensitivity level of less than one transcript per cell. The extraordinary power of such methods will ensure numerous applications particularly in comparative expression studies. For example, in addition to normal/disease tissue comparisons, expression differences could be charted during cell differentiation, apoptosis or tissue aging, and in response to changed environments, such as following exposure to a new drug.

Rapid whole genome expression screens, however, also have inherent limitations: because the expression screen is of low resolution, the cell-sampling procedure may not be sufficiently discriminatory to answer some questions. Given that the data they provide will be an aggregate expression profile of all the cells in the sample under study, such methods will often be most suited to surveying expression in cultured cell lines or comparatively homogeneous tissues. If, however, expression is to be surveyed in original tissues and in cells that are restricted spatially (to a few cell layers for example), or at the other extreme, if one wishes to study simultaneous expression in physically distant cells or even subcellular components (as in the case of extremely long neurons), then low-resolution expression screens are inadequate. However, the sheer size of adult/infant tissues makes comprehensive high-resolution *in situ* gene expression screens and 3D reconstruction of gene expression impractical. Information obtained by low-resolution screens may enable limited high-resolution screens to be targeted to selected genes in promising cells or tissuea. But to obtain more comprehensive *in situ* expression profiles, it will be most efficient to survey tissues in early human development.

Human gene expression maps and human development

The documentation of gene expression in early human development is a promising area for study, justified on both medical and scientific grounds. Developmental abnormalities are common and often severe, whereas from a purely scientific perspective, there is a clear need to improve our sketchy understanding of early human development. A large fraction of human genes are thought to be expressed in early development, notably in the brain, and ought to be studied in the relevant human embryonic and fetal tissues. Such studies should contribute to our understanding of gene function and developmental expression programmes, and also illuminate our understanding of the anatomy and morphology of early human development.

Rapid low-resolution expression screens may profitably be conducted on selected embryonic and fetal tissues, or derived cell

© 1997 Nature Publishing Group http://www.nature.com/naturegenetics

lines. But there will also be a need for high-resolution *in situ* expression studies: the molecular basis of important developmental processes may often feature very few cells or cell layers at a time, and the results obtained will enable comparison with equivalent data in model organism studies. Both RNA and protein levels can be monitored in *in situ* expression studies and, because of the biologically imposed miniaturisation, individual experiments can survey expression in a large range of tissues at a time. As the great majority of human organogenesis occurs during the embryonic period, studying *in situ* gene expression throughout the embryonic period enables insights into the development of adult tissues, which cannot easily be studied in a comprehensive way using the same techniques. For processes of organogenesis that continue into later prenatal development, fetal sections will also be important, as in the case of understanding brain development.

Our meagre knowledge of early human development is directly related to the general difficulty in accessing suitable samples for study. In cases where maternal consent and ethical approval have been obtained, post-implantation embryonic and fetal material can be retrieved for research studies following first- or second-trimester termination of pregnancy, respectively[30]. Much of our current knowledge of early human development has been founded on histological studies using archived embryo sections, notably those of the Carnegie collection in Washington[31,32]. But the widely accepted Carnegie system for classifying human embryonic development is limited in resolution (see below) and there are considerable gaps in our knowledge of many aspects of early development, such as fetal brain development. Recently, several centres in North America, Europe and Japan have engaged in attempts at 3D reconstruction of the anatomy and morphology of human embryos and fetuses and constituent tissues[32–38]. Companion studies of *in situ* gene expression patterns in human embryos will also be useful because they can be expected to illuminate our anatomical knowledge. In particular, 3D reconstruction of 2D human gene expression images could be important in providing added insights into the morphology of complex organs such as brain, heart and sense organs.

The study of gene expression in early human development has hitherto been a mostly uncharted area: high-resolution *in situ* gene expression studies on human embryo sections were first reported only a few years ago[30]. Instead, we have relied on studies in animal models, notably the mouse. Extrapolation from such models has been justified by the observations that many developmental processes and developmental control genes have been strongly conserved during evolution, in many cases from fruit flies to humans, suggesting that mammalian early development has been essentially conserved. However, the first studies of gene expression in human embryos are revealing distinct species differences for some expression domains.

An important example concerns the gene for myosin VIIA, which is mutated in patients with Usher's syndrome type 1B (congenital deafness, vestibular dysfunction and retinitis pigmentosa) and also the *shaker-1* mouse (cochlear and vestibular dysfunction, but no detectable retinal dysfunction)[39,40]. Myosin VIIA expression in the developing retina in mice and other rodents is confined to the retinal pigmentary epithelium, but in humans and other primates, the fetal retina shows an additional myosin VIIA expression domain in the photoreceptor cells[41]. These results suggest that the retinal pathogenesis in humans results from abnormal myosin expression in photoreceptor cells rather than a defect of retinal pigmentary epithelium, as was formerly supposed. Other recent examples include the calpain-3 gene (a strong expression domain in human embryonic and fetal heart, but not in mouse heart from E9.5 to birth; F. Fougerousse *et al.*, manuscript in preparation) and some Wnt genes such as *Wnt7A* (expression in the midbrain of mouse embryos is confined to the ventral and lateral regions, but in human embryos midbrain expression is restricted to the dorsolateral regions and there is an additional expression domain in the telencephalon; P. Bullen *et al.*, manuscript in preparation).

Comparable studies in sibling species of *Drosophila* have shown that even in the case of genes that are critically important during development, novel expression domains can arise very rapidly in evolution[42]. Thus, although the coding sequences of such genes are very highly conserved, relatively rapid alteration of *cis*-acting regulatory sequences can result in novel spatiotemporal domains, which, in turn, may ultimately lead to the acquisition of novel functions[43,44]. Once a gene has been identified as the locus for a human developmental disorder; therefore, it is desirable to proceed immediately to study its human embryonic expression patterns[45].

A gene expression database of early human development

Data production and database format. A systematic, electronically documented study of gene expression in early human development is both desirable and ethically acceptable[30,46]. The stored data should encompass both basic expression data from established techniques and rapid whole genome expression screens, as well as the output of high-resolution *in situ* expression studies.

Is such a goal attainable, and if so, how should it be achieved? The scale of such a venture is undeniably vast. Take the question of embryonic samples: the samples that have been stored in archived collections have been treated in a way that makes subsequent gene expression studies impractical and so new embryo collections are needed. At the moment only a very few research centres, notably in the UK and France, are systematically collecting and storing such material, and an individual institution may typically expect to obtain only a very few hundred embryos per year. In the case of *in situ* expression studies, it will be most efficient to concentrate on transverse sections. Even then, however, existing technologies would only permit reasonably comprehensive screens for about 5–10 genes per embryo. Clearly, every effort must be made to maximise the use of such precious embryonic material (see below).

In order to cope with the scale of the proposed endeavour, it will be most effective to develop an international collaboration involving a network of research centres, mostly drawn from Europe because of the current difficulty in conducting such research in the United States[30]. The financial investment required will be substantial but could be sponsored by a combination of international funding agencies in addition to allocations from various national human genome projects. Individual centres should have the capacity for systematically collecting, storing and processing suitable embryonic/fetal material, and for subsequently studying gene expression directly, and, where appropriate, making material available to other collaborating labs for gene expression studies. Indeed, it will be desirable to establish laboratories that will provide services to a wider scientific community, given the difficulties in carrying out this type of research because of restricted access to the material.

In order to integrate the data effectively, the results need to be stored in a common public database or series of linked databases. But there will also be a clear need for a consensus standard by which to measure the quality of the input data, helping data integration and facilitating peer review. In time, it should be possible to have Internet access to complex gene expression images, including 3D reconstruction images, in the database. New Web software, such as Java, can be developed to provide users of the database with interactive programmes for image browsing and manipulation that will maximise interpretation of the data. The database also needs to have active links to established genome and sequence databases, and importantly to other gene expression databases and

© 1997 Nature Publishing Group http://www.nature.com/naturegenetics

future gene function databases. The format of the Mouse Gene Expression Information Resource[25,26] provides a good example on which to model the proposed database. In an effort to facilitate gene expression comparisons between human and mouse, the Newcastle group is working closely with the Edinburgh Mouse Atlas group[47] and the first 3D reconstruction images of gene expression in human embryos are being obtained (Fig. 2).

Variation and landmarks of expression in development. Interpretation of the results from the expression studies is subject to experimental errors, and to uncertainties regarding the precise developmental age of the material studied and of the co-ordinates in the examined tissue sections. As a result, a statistical appraisal of the results will need to be implemented. In addition, some natural, and environmentally induced, variability should be expected between individuals, with differences in the exact timing of expression and the temporal order of expression for many genes. Finally, another important source of variation in expression patterns is the maternal effects that will influence the expression of some genes. It will therefore be necessary to provide users of the database with tools to evaluate the data, which take into account the various factors that contribute to variability of expression pattern. Because of the limited number of human embryos, however, there is no clear solution as to how the variation could be measured, although the study of animal model systems could possibly provide some answers.

Individual variation resulting from pathogenic mutations will also be encountered occasionally. Therefore, karyotype analysis of placental samples (preferably from multiple sites in order to address the problem of confined placental mosaicism) should routinely be used to screen for cytogenetic abnormalities. But small mutations will go undetected, although their presence may be indicated in some cases that have a clearly abnormal morphology. Clearly, analysis of samples obtained from therapeutic abortions will provide valuable information regarding gene expression in abnormal development[48].

Notwithstanding the problems of uncontrollable fluctuations in expression pattern, comprehensive documentation of expression profiles for numerous genes should provide valuable tissue-specific *expression landmarks* of early development. Such landmarks could be of great value in enabling more accurate staging in human development where the uncertainty regarding the exact time of conception means that we have to rely heavily on appreciation of developmental stages in order to estimate age. The Carnegie staging system is based on anatomical/morphological features and is limited both by low resolution (most of the individual stages span several days of development) and by being restricted to the embryonic period, up to the end of the eighth week of development[31,49]. By choosing genes with expression patterns that are sharply defined in time and space, and which are comparatively independent of natural or artefactual variability, a high-resolution expression-based staging system could eventually be constructed. Expression landmarks of development will also be important in species comparisons, since the relative rates of development of particular systems can vary between species — as in the case of prenatal neurological development, which is comparatively retarded in rodents. Finally, expression landmarks of development will provide a detailed framework for topographical referencing of other signals during development, and should facilitate the superimposition of the reconstructed 3D gene expression images from multiple embryos that are adjudged to be at equivalent developmental stages.

Optimisation of research strategies and co-ordination of gene expression studies. Because of the limiting amounts of human material available for study, comprehensive *in situ* gene expression studies of early human development cannot be carried out in the short term. It will therefore be important to prioritise expression studies. In addition, there is a great need to increase efficiency, both at the level of the technologies being used (which are currently greatly rate-limiting) and by optimising research strategies. Using current technologies over a period of, say, ten years, it would be feasible for about ten research centres to carry out reasonably comprehensive *in situ* embryonic expression studies on potentially 1,000 important genes (with expression studied in at least 5–7 Carnegie stages chosen to be representative of the embryonic period). The general efficiency of the embryonic gene expression studies can be expected to increase with increased automation and the use of more sensitive detection methods. For example, multiplex probing using several probes that are labelled with different fluorophores may increase the amount of expression information obtained per tissue section. But even then we will need to prioritise the choice of

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Fig. 2 Three-dimensional reconstruction of *WNT5A* expression patterns in part of the brain of a Carnegie stage 17 human embryo. Transverse sections were digitised and then reconstructed in 3-D using programmes developed by R. Baldock and M. Stark (MRC Human Genetics Unit, Edinburgh). A proportion of these sections have been hybridised with an antisense *WNT5A* probe. Expression patterns were digitised and then superimposed on the tissue reconstruction, and are visualised as a pink signal, or in the case of signal falling on the plane of cutting, as a red signal. The neural tissue is visualised as gray. Expression is seen in the floor of the midbrain and hindbrain (lower part of image), and in the caudal diencephalon (dorsal and ventral thalamic nuclei at the cut plane in the upper part of the image). The reconstruction was carried out at the MRC Human Genetics Unit in Edinburgh as a collaboration between the Newcastle and Edinburgh groups. (Photo courtesy of Phillip Bullen, University of Newcastle upon Tyne.)

© 1997 Nature Publishing Group http://www.nature.com/naturegenetics

Table 1 • Issues in developing optimal and synergistic strategies for studying gene expression in early human and mouse development

Priorities for selecting genes for studying expression in human development
- disease related genes
- developmentally important genes, or possibly downstream genes in well-defined pathways
- genes that are potential therapeutic targets

Genes where human expression studies are essential or at least as important as mouse studies
- primate-specific genes
- genes known to show species-specific expression patterns (including expression differences resulting from species differences in alternative RNA processing and imprinting)
- genes involved in highly specialised organs or organs showing significant species differences in anatomy/morphology (such as brain)
- genes involved in disease processes affecting complex brain functions or in diseases that are not easily modelled in mice (neurodegenerative, sensory, behavioural, psychiatric, late-onset)

Research areas where expression studies should be conducted in mouse but not human embryos
- rodent-specific genes
- whole mount embryo studies

Research areas where mouse expression studies should be a general 'first pass' screen
- identifying developmental/physiological pathway(s) containing the target gene
- identifying position of the target gene in a developmental/physiological pathway(s)
- validation of putative coding sequences (e.g. from genomic DNA analysis and computer programme prediction) and anonymous coding sequences (human/mouse EST programmes)
- genes belonging to a 'conserved' developmental pathway

Research areas where mouse expression studies are conducted after human studies
- validation or confirmation of results obtained on human material

Technological improvements
- increased automation of probing technologies
- multiplex FISH
- quantitative data measurements
- subcellular resolution

Standardization of procedures
- definition of molecular expression landmarks of development
- standardisation of nomenclature (anatomical, gene names)

Structure of human gene expression databases

Optimization of coordination between databases
- links between human and mouse expression databases, requiring compatibility of data and interoperability of databases
- links to databases listing gene and gene product structure and to databases with details of gene function (protein-protein interactions, epistatic relationships)

Public access to databases

genes for study, and to develop strategies that maximise the use of human material (Table 1). There is also currently a great need to improve the capacity for recording 3D expression images. Again, however, one could anticipate great improvements in software and hardware development in the near future, making this approach less laborious than at present.

Which genes should be studied first? A major priority will be genes implicated in (inherited) disorders and cancers, and genes suspected of being critically important in development. Genes known or suspected to show significant species differences in embryonic gene expression patterns would also be strong candidates for study. In addition, it will be important to investigate genes which are expressed in developing organs that show significant species differences, or which are of crucial importance to unique human functions, such as the developing brain. Finally, it may also be very profitable to study downstream effector genes, given that they may show up more significant species differences in expression pattern than the master regulator genes.

To pursue optimal strategies that maximise the use of human material, the choices of questions to be asked, the experimental approaches used and the material to be studied require careful consideration. Some questions can be best addressed using basic expression assays such as the microarray hybridization assays. In the case of high resolution *in situ* expression studies it will be particularly important to classify proposed studies on the basis of the perceived relative merits of studies in human material or in animal models. The mouse is the favourite animal model of mammalian development, partly due to the power of gene targeting in ES cells as a vehicle for generating developmental mutants. Because of the comparatively plentiful supply of mouse embryos effective global strategies would involve complementary approaches in the two species, often using mouse material as a first screen to identify studies that would be particularly profitable to carry out in human material (Table 1).

Concluding remarks

We believe that there is a strong need for international co-ordination of a second phase of the human genome project, expression mapping. Different approaches exist, necessitating the integration of the results from these complementary methods and an increase in efficiency of high resolution *in situ* human gene expression studies. It will also be important to develop methods of integrating the data with that from different animal models, and so to develop the important new discipline of comparative molecular anatomy, which could be particularly valuable in understanding early brain development, for example.

© 1997 Nature Publishing Group http://www.nature.com/naturegenetics

Several years of discussion and planning were required before the Human Genome Project was launched. The momentum and international dimension to the structural genomics phase developed when efforts were made to share common resources and establish publically available electronic databases. The time is now ripe to develop a similar international effort to co-ordinate a systematic study of human gene expression. There is a very considerable need for large scale studies of gene expression in early human development, and for co-ordination of human and animal model studies. The items we have highlighted by no means represent a comprehensive list of the areas that require consideration, but we hope that they will stimulate further discussion.

Acknowledgements

We gratefully acknowledge financial support from the Wellcome Trust (TS), the UK Medical Research Council (DD) and the Association Française contre les Myopathies (M.A. & J.S.B.). We are particularly indebted to P. Bullen, F. Fougerousse, M. Schena and D. Lockhart for conveying unpublished results. We are also grateful for comments on the manuscript from various colleagues.

1. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
2. Fields, S. The future is function. *Nature Genet.* **15**, 325–327 (1997).
3. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567 (1996).
4. Oliver, S. From DNA sequence to biological function. *Nature* **379**, 597–600 (1996).
5. Miklos, G.L.G. & Rubin, G.M. The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**, 521–529 (1996).
6. Boguski, M.S. & Schuler, G.D. ESTablishing a human transcript map. *Nature Genet.* **10**, 369–371 (1995).
7. Schuler, G.D. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
8. The BodyMap Anatomical Expression Database of Human Genes at http://www.imcb.osaka-u.ac.jp/bodymap.
9. Bains, W. Virtually sequenced: the next generation. *Nature Biotechnol.* **14**, 711–713 (1996).
10. Anderson, N.G. & Anderson, N.L. Twenty two years of two-dimensional electrophoresis: past, present and future. *Electrophoresis* **17**, 443–453 (1996).
11. Laing, P. & Pardee, A.B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 274–280 (1992).
12. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
13. Adams, M.D. Serial analysis of gene expression: ESTs get smaller. *Bioessays* **18**, 261–262 (1996).
14. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
15. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619 (1996).
16. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
17. Schena, M. Genome analysis with gene expression microarrays. *Bioessays* **18**, 427–431 (1996).
18. Goffeau, A. Molecular fish on chips. *Nature* **385**, 202–203 (1997).
19. Lockhart, D. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
20. Editorial. To affinity ... and beyond. *Nature Genet.* **14**, 367–370 (1996).
21. Maillet, I., Lagniel, G., Perrot, M., Boucherie, H. & Labarre, J. Rapid identification of yeast proteins on 2-dimensional gels. *J. Biol. Chem.* **271**, 10263–10270 (1996).
22. Chait, B.T. Trawling for proteins in the post-genome era. *Nature Biotechnol.* **14**, 1544 (1996).
23. Ringwald, M. *et al.* A database for mouse development. *Science* **265**, 2033–2034 (1994).
24. Davidson, D., Baldock, R. & Bard, J. A 3-D atlas and gene expression database of mouse development: implications for a database of human development. In *Molecular Genetics of Early Human Development* (eds Strachan, T., Lindsay, S., Wilson, D.) (BIOS Scientific Publishers Ltd., Oxford, in the press).
25. Davidson, D. *et al.* Gene expression databases. In *In situ Hybridization. A Practical Approach* (ed. D. Wilkinson) (Oxford University Press, Oxford, in the press).
26. Appel, R.D. *et al.* Federated 2-dimensional electrophoresis database - simple means of publishing 2-dimensional electrophoresis data. *Electrophoresis* **17**, 540–546 (1996).
27. http://expasy.hcuge.ch/ch2d/2d-index.html/
28. Kuska, B. Cancer Genome Anatomy Project set to take off. *J. Natl. Cancer Inst.* **88**, 1801–1803 (1996).
29. Nelson, N. Microarrays pave the way to 21st century medicine. *J. Natl. Cancer Inst.* **88**, 1803–1805 (1996).
30. Burn, J. & Strachan, T. Human embryo use in developmental research. *Nature Genet.* **11**, 3–6 (1995).
31. O'Rahilly, R. & Muller, F. *Developmental stages in human embryos.* Carnegie Institution of Washington, publication 637 (1987).
32. http://magenta.afip.mil/embryo
33. http://wwwcivm.mc.duke.edu/civmPeople/SmithBR/documents/HumanEmbryo.html
34. http://www,st-and.ac.uk/~www_sbms/terrapin/frontpage.html
35. http://www.ncsa.uiuc.edu/evl/SHOWCASE/cheong/embryo.html
36. Verbeek, F.J., Huijsmans, D.P., Baeten, R.J.A.M., Schoutsen, N.J.C. & Lamers, W.H. Design and implementation of a database and program for 3D reconstruction from serial sections: a data-driven approach. *Microscopy Res. Tech.* **30**, 496–512 (1995).
37. Machin, G.A. Computerized graphic imaging for three-dimensional representation: general principles and application to embryo/fetal development. *Int. Rev. Exp. Pathol.* **36**, 1–30 (1996).
38. Shiota, K., Nakatsu, T. & Ire, H. Computerised three-dimensional reconstruction of the brain of normal and holoprosencephalic human embryos. *Birth Defects: Original Article Series* **29**, 261–271 (1993).
39. Weil, D. *et al.* Defective myosin VIIA gene responsible for Usher syndrome type 1B. *Nature* **374**, 60–61 (1995).
40. Gibson, F. *et al.* A type VII myosin encoded by the mouse deafness gene *shaker-1*. *Nature* **374**, 62–64 (1995).
41. El-Amraoui, A. *et al.* Human usher 1B/mouse shaker-1: the retinal phenotype discrepancy explained by the presence/absence of mysoin VIIA in the photoreceptor cells. *Hum. Molec. Genet.* **5**, 1171–1178 (1996).
42. Wang, D., Marsh, J.L. & Ayala, F.J. Evolutionary changes in the expression pattern of a developmentally essential gene in three Drosophila species. *Proc. Natl. Acad. Sci.* **93**, 7103–7107 (1996).
43. Li, X. & Noll, M. Evolution of distinct developmental functions by three Drosophila genes by acquisition of different cis-regulatory regions. *Nature* **367**, 83–87 (1994).
44. Strachan, T. & Lindsay, S. Why study human embryos: the imperfect mouse model. In *Molecular Genetics of Early Human Development* (eds Strachan, T., Lindsay, S., Wilson, D.). (BIOS Scientific Publishers Ltd., Oxford, in the press).
45. Li, Q.Y. *et al.* Holt-Oram syndrome is caused by mutations in *TBX5*, a member of the *Brachyury (T)* gene family. *Nature Genet.* **15**, 21–29 (1997).
46. Bennett, R. & Harris, J. The ethics of human embryo studies. In *Molecular Genetics of Early Human Development* (eds Strachan, T., Lindsay, S. & Wilson, D.I.) (BIOS Scientific Publishers Ltd., Oxford, in the press).
47. Baldock, R., Bard, J., Kaufman, M. & Davidson, D. A real mouse for your computer. *Bioessays* **14**, 501–502 (1992).
48. Abitbol, M. *et al.* Nucleus basalis magnocellularis and hippocampus are the major sites of FMR-1 expression in the human fetal brain. *Nature Genet.* **4**, 147–153 (1993).
49. Bullen, P., Wilson, D., Lindsay, S., Robson, S.C. & Strachan, T. The Carnegie system for staging human embryos: a photo-essay. In *Molecular Genetics of Early Human Development* (eds. Strachan T, Lindsay S, Wilson D). (BIOS Scientific Publishers Ltd., Oxford, in the press).