using expression databases for drug discovery

Reference databases are being built cataloguing the amount of RNA each gene makes under different conditions and in different individuals. These databases can be mined easily and effectively to aid the drug discovery process.

By Dr Greg Lennon

The success of the Human Genome Project provokes the question: what are the implications for drug discovery? And which are immediate, versus quite long term? This article will discuss one of most fruitful – and most immediate – implications on the drug discovery process of knowing the number and position of all 100,000 genes in the genome. Specifically, this article focuses on what might be called the 'Human Genome *Expression* Projects', which aim to help understand the conditions under which each of these genes is activated or suppressed, whether by biological processes such as ageing and disease, or more pertinently, pharmaceutical intervention.

You presently have more than 100 billion cells in your body, categorised into around 200 tissue types. In each cell type, a different set of 20,000 genes out of the possible 100,000 are likely to be active (or expressed) right now. (The other 80,000+ are basically shut off, meaning no messenger RNAs are being transcribed.) Each of the active genes can be 'on' to a different level, in terms of the number of corresponding RNA or protein molecules being made from that gene. Cells of similar types have many active genes in common, and furthermore, are likely to have similar amounts of RNA (or protein) from each of them. One of your neurons, on the other hand, can be readily distinguished from one of your hepatocytes not only by which genes are active, but even by the amounts of RNA derived from each gene. Now think about the person located nearest you at the moment. While their genome holds the same total number of genes as you have, and they have the same number of cells, their levels of expression (and even the genes active at all) form a pattern of gene expression distinct from yours. Now imagine having access to the levels of expression of the genes in tissues from normal and diseased patients, from humans, animals and cell lines undergoing drug treatments, indeed, from the entire spectrum of samples relevant to your research. The collection of many such differences in gene expression, and the correlation with corresponding medical, biological, and pharmaceutical information, forms nothing less than a profound basis for accelerating the drug discovery process. The collection of such data into intelligently designed databases will form a lasting repository dwarfing in size and significance the genomic sequence databases.

This type of data can be used to complement your ongoing pharmaceutical development processes. Questions this data can help answer include:

● Is the target of my lead compound directly or indirectly inhibited by any known drug?
● Which of my lead compounds for a certain indication have the highest ratio of 'good' changes to bad (toxic or otherwise unwanted) changes?
● Which transcripts (or pathways) are the best markers for disease X, and thus might be the best ones to use in a multi-gene HT screen for new lead compounds?
● Which pathways (or transcripts) show similar changes in many different diseases, and thus might be non-specific markers of common cellular processes?

- How does the drug I am about to launch into trials differ from drugs already on the market for this indication, as measured by all the genes affected? And how is it similar?
- What genes are most commonly aberrantly expressed in people with a certain disease compared to comparable people without such a disease?
- What are the clinical parameters most indicative of a person who does not respond to the class of drugs of which I am about to launch a new clinical trial?
- Are there other indications that might benefit from treatment with one of my drugs?
- Which of many protein targets are likely to be the best to put into a traditional single target high throughput compound screen, based on expression data from over-expressing and/or inhibiting each of the candidates?

Notice that the overall approach is almost equally applicable to all organisms, from the study of genes active in small pathogens, or activated by them in their hosts, through all branches of the tree of life. As long as an organism has genes, the measurement of the degree to which each of its genes is expressed under different conditions can be studied.

So what is being done now?

Efforts are under way in parallel to encompass at least three aspects of these types of projects: the technical methods of generating such data, the informatic challenges in keeping the experimental, sample, and gene data organised and accessible, and the bioinformatics required to extract meaningful information from such combined data.

The technical methods for generating such data comprise both open methods, where the gene need not have been identified in advance, and closed methods, which parlay a gene or sequence already in hand with higher throughput assays. The history and current status of such methods has been recently reviewed[1]. Currently, most of the higher throughput gene expression assays require 1-10μg of total RNA (about the amount isolated from 1-10 million cells), and in just a few days can generate mRNA abundance information for all genes available on their platforms that give rise to transcripts above certain thresholds of sensitivity. The most widely used standardised platform is the Affymetrix GeneChip®, which currently has the potential to measure 60,000 human genes simultaneously. GeneChips® representing genes from a wide variety of other species are also available[2].

Throughout private and public sectors, a wide variety of techniques are being used to measure mRNA abundances from patient and animal tissues as well as from cell lines. The basic strategy is simply to collect detailed sample information, perform the gene expression measurements, and then place all the information into a database. The key questions to ask of all such efforts include (1) are quality control measures in place throughout the process such that the data generated could be reproduced? and (2) are the units of measurement primarily valid relative only to other measurements made on the same system at the same

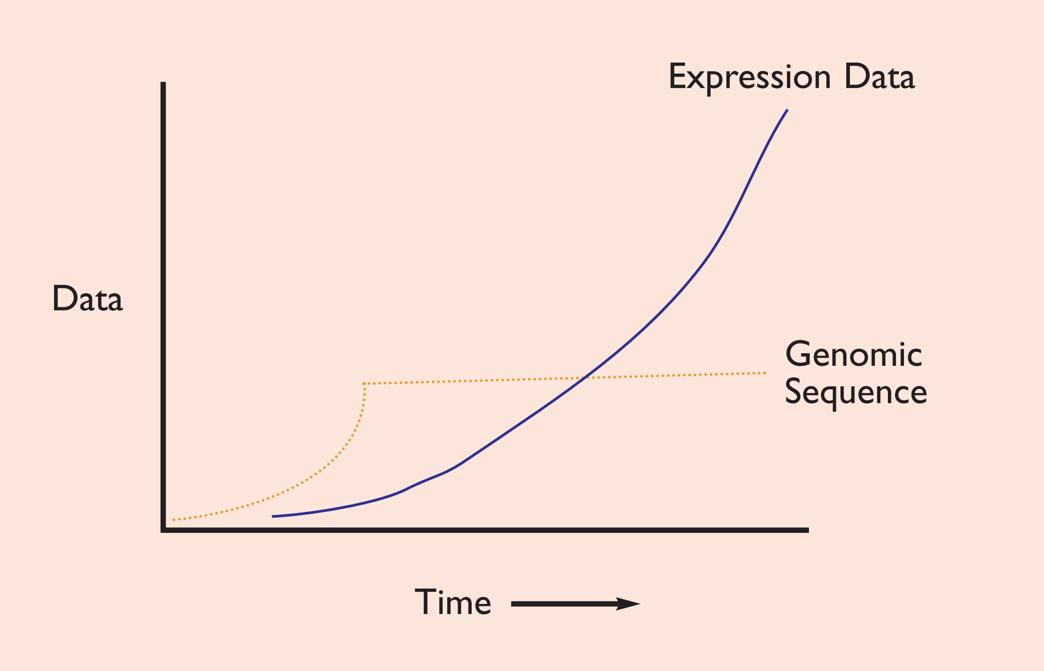


Figure 1
The growth of the number of expression datapoints (one datapoint = one measurement of the number of transcripts from one gene in one sample) will swiftly exceed the total number of sequenced base-pairs in a genome

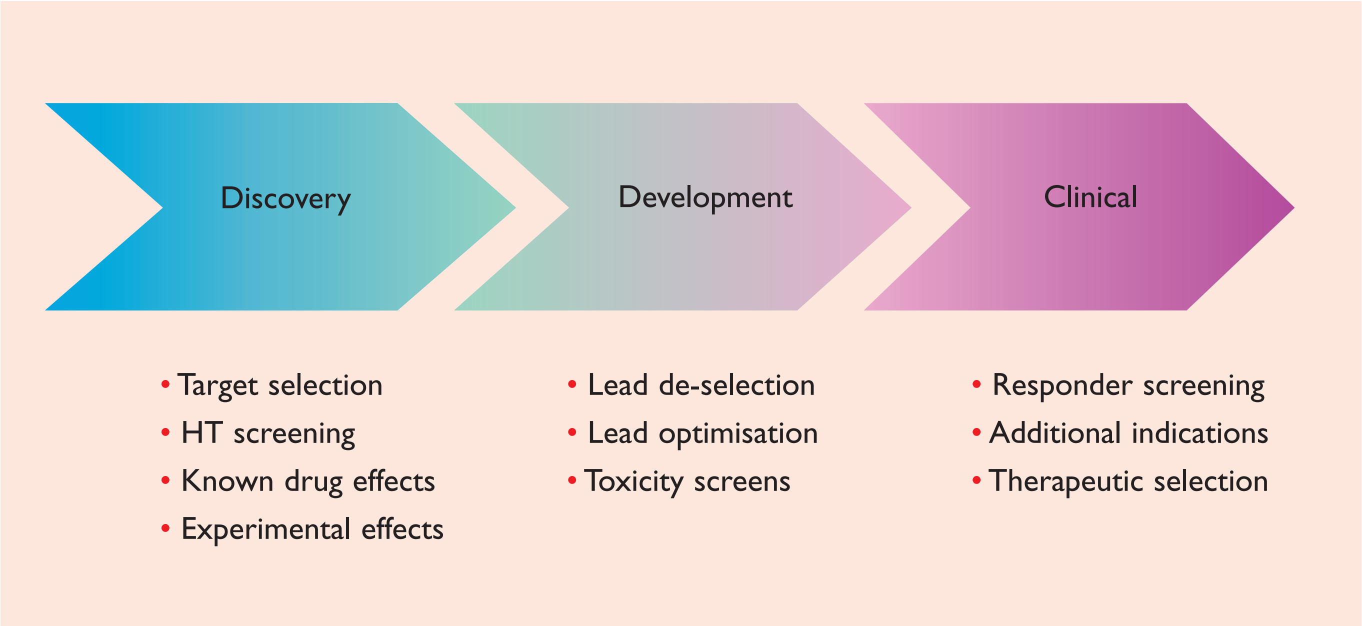


Figure 2
mRNA expression data can be applied to numerous steps in the drug discovery process

time, or are they more absolute and thus comparable to data sets derived in other labs at other times by other methods?

Many well-designed experiments are intended to produce only relative data; the key to long-term success in a genomics context is to also produce data with greater longer-term value, such as genomic sequence itself. Since today's 'best' technology is highly unlikely to be tomorrow's, reference databases in this field should both aim to use today's platforms to create near-absolute data, and to plan for the inevitable need to transition from one platform to the next. This is little different from creating a reference music collection – how would you go about planning for the platform changes from albums to cassettes to compact discs to whatever the future holds?

Since the field in many ways is still in its infancy, many labs are currently making their own microarrays. Over time, the number of labs continuing to do so will approach the number of labs still making their own restriction enzymes.

Where's the data?

Individual laboratories conducting expression experiments have had little choice but to develop ad hoc databases to hold the large amounts of data these experiments typically generate. Furthermore, while some of the data can be presented in a journal article, the vast amount of information – which includes which transcripts did not change as well as those that did – cannot. Some groups have established publicly accessible websites containing the (unpublished) data. Electronic publications are likely to increase the suitability of linking articles to such websites.

Analogous to public sequence databases, public expression databases are being established. The largest public efforts are the ArrayExpress[3], Gene Expression Omnibus (GEO)[4], and GeneX databases[5], being developed by EMBL/EBI, NIH/NCBI and the NCGR, respectively. (An interesting and quite diverse list[6]of more than 50 expression databases or software tools can be found on the NCGR site.) The key challenges to such databases will be enabling meaningful cross-laboratory, and cross-platform, analyses. In distinct contrast to sequence databases, which collect information that can be 99.99% reproduced by any lab sequencing the same genomic clone, the world of expression information contains little that can be readily reproduced between labs. Uniqueness of samples as well as experimental conditions make cross-platform and cross-lab analyses difficult. Public databases will be best used to study self-consistent data from single sources until further standardisation evolves.

Private database efforts have the luxury of being in control of how all their data is produced, and can institute quality assurance measures that might be seen as too costly for academic laboratories. The largest such effort is Gene Logic's GeneExpress™ suite of databases[7]. These databases contain expression data generated using Affymetrix GeneChips® for thousands of clinical (human patient) and experimental samples. For example, for the 30,000+ genes represented by the Affymetrix Hu42K GeneChips®, these databases can already tell you which tissues each gene is normally active in, and which diseases show aberrant regulation. As GeneChips® surveying even larger numbers of genes become available, either commercially or from Gene Logic's custom GeneChip® development efforts, the databases get updated. Analysis programs allow the user to mine not only the database on its own, but to also compare data in the database with any data the user may have privately generated. Publicly available expression (and sequence) data is incorporated along with privately generated gene indices to annotate all human genomic sequence with expression information.

Ultimately, adequate quality assurance measures will be developed across both academic and private expression laboratories as they have for laboratories generating sequence data. Platform-specific data must be converted to units such as parts per million, or even copies per cell. Some groups have started such efforts. One lab has converted data from yeast cell cycle experiments into 'ERA's, or estimates of relative abundance[8]. Another effort[9] uses competitive template polymerase chain reaction (CT-PCR) templates to quantitate the abundance of any transcript as the number of molecules per million transcripts of the ubiquitous beta actin gene. Much research focus is just starting to be applied in a serious manner to the methods for normalising data within and between platforms.

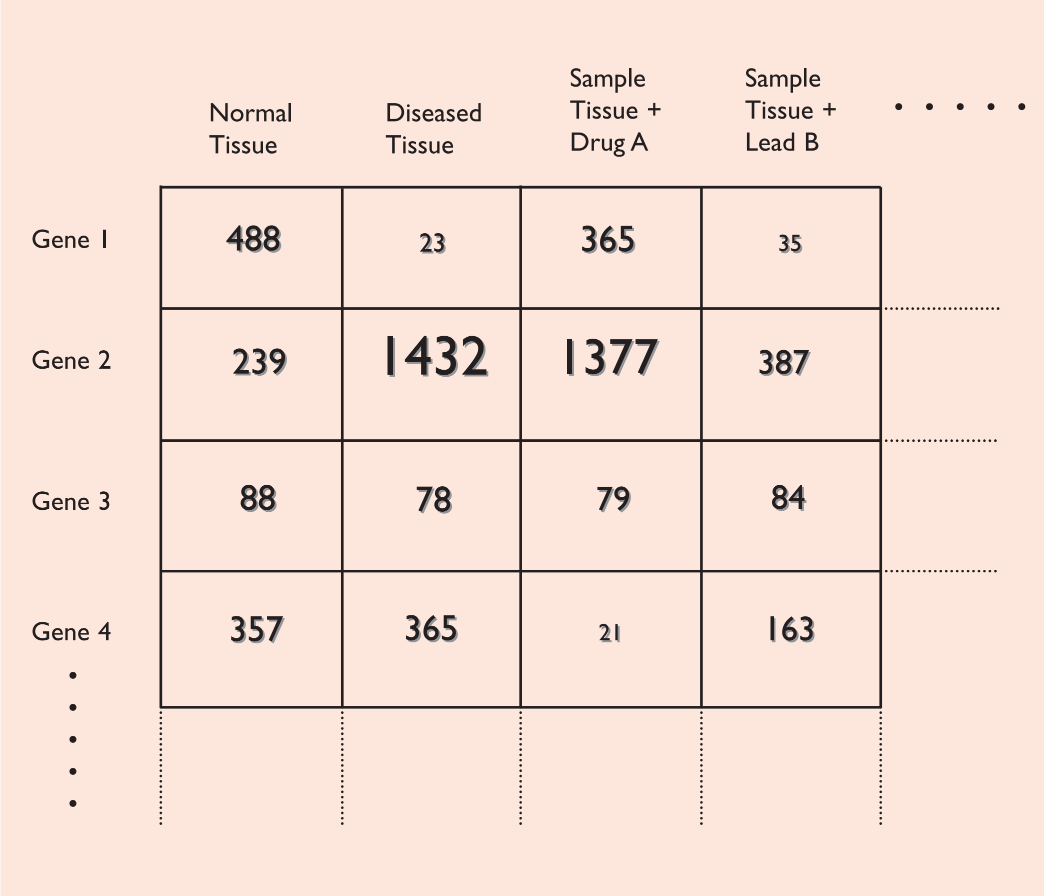
How would I deploy this internally?

Assuming scientists with sufficient bioinformatic skills are at hand, and are supported by a reasonable information technology group, the first question to ask is which project is most in need of breakthroughs of the type that expression data can help foster.

Do you need to find more targets? Do you need to choose 50 genes to form the basis of a multi-gene HTS screen for lead compounds? Do you need to broadly analyse the potential mechanism of action of some drug candidates? Differentiate responders from non-responders? Or would you like to rank the potential toxicity of a series of lead compounds? The best way to start using expression data, which can certainly aid in these and many other efforts, is to choose the most pressing problem.

Then decide on whether to outsource some or all of the data generation experiments. Over the next few years, you will certainly want to have at least some core competency in-house, both for scientific strategy (whether or not aided by consultants), experiments themselves, and the associated bioinformatic analyses.

Decide on whether to subscribe to one or more commercial reference databases, and if not, how to



	Normal Tissue	Diseased Tissue	Sample Tissue + Drug A	Sample Tissue + Lead B	• • • • •
Gene 1	488	23	365	35	
Gene 2	239	1432	1377	387	
Gene 3	88	78	79	84	
Gene 4	357	365	21	163	
• • • • •					

Figure 3
Simplified schematic of how expression databases store information. The axes consists of all genes being surveyed, and, all samples. Measurements are ideally available in units allowing cross-platform comparisons, such as parts per million (ppm)

References

1 Lennon, G. (2000) High-throughput gene expression analysis for drug discovery. Drug Discovery Today, 5(2), 59-66.

2 Details on Affymetrix systems can be found through their website at URL www.affymetrix.com/

3 www.ebi.ac.uk/arrayexpress/

4 www.ncbi.nlm.nih.gov/geo/

5 www.ncgr.org/research/genex/

6 www.ncgr.org/research/genex/other_tools.html

7 www.genelogic.com

8 Aach, J, Rindone, W & Church, GM. (2000) Systematic Management and Analysis of Yeast Gene Expression Data. Genome Res. 10(4), 431-445.

9 Willey, JC, Crawford, EL, Jackson, CM, Weaver, DA, Hoban, JC, Khuder, SA, & DeMuth, JP. (1998) Expression measurement of many genes simultaneously by quantitative RT-PCR using standardized mixtures of competitive templates. Am. J. Respir. Cell Mol. Biol. 19, 6-17. See also the GENE website at URL www.genexnat.com

use the public data. Note that there will certainly be an increased bioinformatics cost involved with putting public data into a framework of highest utility to your own scientists; commercial databases generally incorporate all public data in with their private data. The most useful commercial efforts create private databases from your own data as well, so that your scientists can seamlessly query across all the data, regardless of whether it was generated in-house, at the vendor, or in academia. The informatic infrastructure to allow you to smoothly handle large amounts of expression data can also be developed from scratch or purchased, independently of data content.

Plan on gaining experience from the first experiments, and assume that their main function is to allow you to properly set up follow-on experiments. A lesson being learned at numerous pharmaceutical companies currently is that there is insufficient attention to sample collection. At the minimum, blood should be collected from patients in clinical trials in a manner that will not subsequently prevent you – should you wish – from checking both patient genotypes (eg extract DNA and type it for mutations) as well as RNA expression patterns (eg transcript profiles in the types of cell collected in a routine blood draw). Biopsies when collected should also be handled in a manner conducive to later DNA and RNA profiling. The additional costs are minuscule relative to overall clinical trial costs and may form the basis for very valuable expression information.

Coming attractions

In the initial rush to survey and better understand the human genome, expression information will be the primary source for cataloguing the conditions and tissues in which genes are active. As increasingly rich clinical information is tied to this expression data, the overall utility will grow even more rapidly, since the statistical significance of changes can be refined with increasing numbers of samples.

Complementing these efforts will be similar efforts analysing protein abundance, and the development of increasingly sophisticated bioinformatics to QC/curate the data, automatically mine the data on its own, and integrate it with relevant experimental, clinical and medical data.

Improvements are certainly still needed, in at least three aspects: cost, standardisation and integration. Cost per expression data point will continue to drop, and concomitantly there will need to be integration across platforms, especially across the academic realm. Integration with most common medical measurements (such as clinical chemistry values and physiological read-outs) will allow for the most effective molecular-medical bridges, just as technological advancements will mean less clinical material will be needed and measurements will increasingly be made real-time and non-invasively.

Similarly, the need for accurate, comprehensive, electronically accessible medical records is overwhelming. The correlations between expression data and clinical outcomes will demand it. As an example: this is not just having information about which drugs each patient from whom you have collected a sample was supposedly taking, but when the dosages were taken, down to at least the hour level, and over how long a period, in the context of all other potentially relevant physiological and pharmaceutical data.

The overall need for the handling of variation at DNA, RNA, and protein levels also becomes evident to experienced practitioners in these fields. Variation at the genotype level (such as single nucleotide polymorphism profiles), and alternative forms of mRNAs and proteins must be captured to analyse along with the expression information.

Ultimately, we will need to discriminate between individual cells of ostensibly the same type. Molecular characterisation will undoubtedly reveal cellular subtypes. Within properly defined sub-types, individual cells will vary from one another based on their history and current environment. Furthermore, we will need to take genotypes into account too, so we will need to genotype each individual at each of the estimated 100,000 nucleotide positions where variation is likely to be most meaningful.

Conclusions

Even more powerfully than genomic sequence, large-scale expression information is becoming ready to apply to many aspects of the drug discovery process. While the field is in many ways still in its infancy, and thus subject to the inevitable growing pains associated with technologies that are improving daily, it is useful now. And although predictions – especially about the future – are always difficult, it is easy to predict that the sooner you start applying this type of data in your drug discovery efforts, the sooner you will appreciate how expression can help unlock some of the many secrets of the human genome. DDW

__Dr Greg Lennon__ is a founder of the IMAGE Consortium, the largest collection of public cDNA clones, and served as CSO for Gene Logic, Inc from 1997-2000 prior to becoming a genomics consultant. He can be reached by e-mail at greg_lennon@yahoo.com.