High throughput genomics and drug discovery – *parallel universes or a continuum?*

HTS has been in place for approximately 10-12 years and has achieved a three-order of magnitude scale up. Genomics has been in a 'high throughput' mode for approximately four years for areas such as genotyping, is an emerging field governed by sporadic technology leaps and data generation leaps. However, despite these approaches, there are many discrepancies and parallels that exist between the two areas: sample management/assay assembly/parallel processing/ data analysis. However, some differences exist with respect to regulatory issues, public perception and ethical consent. The article will compare and contrast these fields and highlight where both disciplines may learn from each other.

By Dr Mark Beggs and Audrey Long

High throughput screening (HTS) is a new discipline that has established itself in the field of discovery lead identification in the last decade. The objective was to extend the capabilities of traditional empirical screening approaches to a model in which, in the extreme, the entire corporate compound library could be screened against a biological target within a time scale of less than a year. The underpinning rationale was to provide a tool by which a wide range of structurally diverse chemical structures could be screened for activity against biological targets of interest. By increasing the capacity of the screening organisation, it was anticipated that a substantially larger number of novel, unpredicted structure activity relationships could be identified.

Within the pharmaceutical world, (pharmaco)genomics promises the potential of improved clinical trial design through the use of patient stratification using SNP genotyping; the ability to assess the effects of genetic variability thereby enhancing decision making in R&D; and the development of personalised medicines in conjunction with diagnostic suitability testing. The recent approval of Herceptin as a breast cancer treatment reflects the start of this trend[1]. Further, genomics is being employed to identify novel discovery targets through use of relevant markers. Suitable targets identified in this manner will enter the classic lead identification route.

By embracing genomics as a part of the R&D process, savings on drug development costs are estimated to achieve an average saving of $300 million and two years development time per drug. This represents savings of 35% and 15% in cost and time respectively[2].

At its inception, at the start of the last decade, HTS throughputs were typically in the order of 200-300 compounds screened per target per day. This rose to a few thousand per day by the early 1990s to be followed by reports of 10,000 compounds per day by the middle of the decade. Three years later peak throughputs of 100,000 compounds per day were being reported. Current peak capabilities are now in the region of 200-300,000 compounds per day in the hands of leading practitioners (**Figure 1**).

Considerable parallels exist with the genomics world. Here progress was limited by the technology available in areas such as automated sequencing[3]. The Human Genome Project is a good case in point. At its inception in October 1991, it was

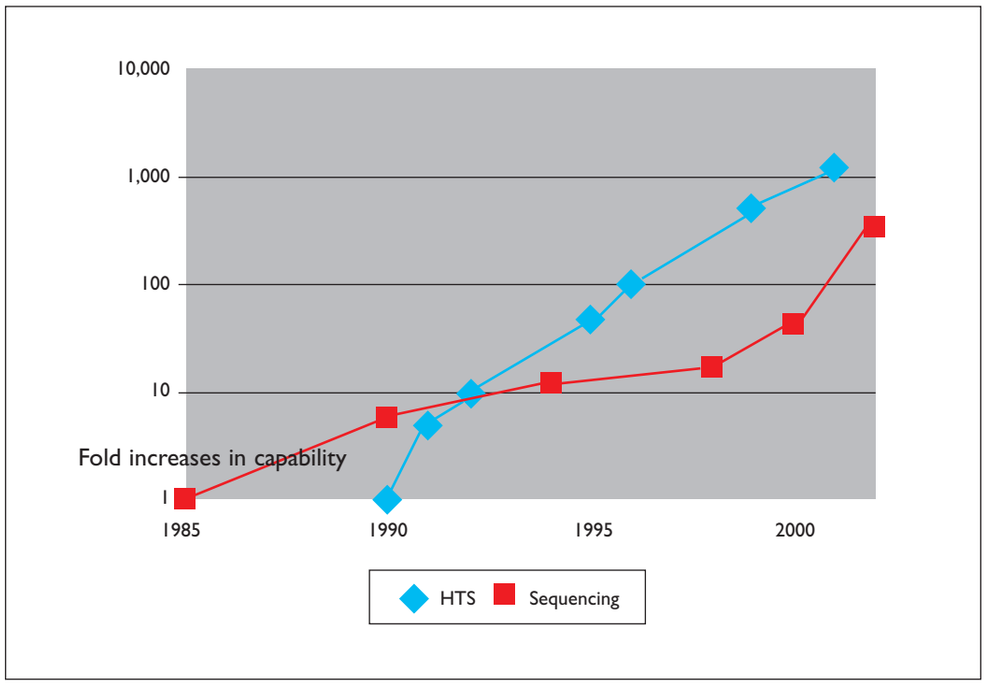

Figure 1
Figure shows the relative increases in capability from initial start points for DNA sequencing (1985; 85 years to sequence human genome) and HTS (1990; 200 wells per week per screen). Based on improvements up to the 2001/02 performance levels, the human genome could now be sequenced in three months and a 1 million compound file screened in four weeks respectively

projected to last 15 years[4]. As a result of significant technological improvements, the project was officially completed in 2000, six years ahead of the predicted end date. In a wider context, the availability of new sequencing technology has allowed what would have been an 80-100 year project using the manual sequencing techniques of the 1980s to be reduced to a three-month timescale (**Figure 1**).

Sample access and regulatory issues

The pace of progress within the HT genomics arena has been limited by factors that are, in the main, non-science related. Significant ethical and regulatory constraints govern the way in which blood, tissue and DNA samples have to be managed (**Figure 2**). Public perception of the use of genetic information has required safeguards to protect the individual and prevent potential misuse of an individual's genetic information[5,6]. Such regulation requires full sample traceability through Good Laboratory Practice (GLP). GLP is a regulatory constraint which, until recently, pharmaceutical research environments had not been required to work within – it is an area which many genomics research groups within pharmaceutical companies are now rapidly implementing.

Acquisition of patient samples with relevant clinical data for both healthy and disease state subjects has required extremely high levels of co-ordination and careful explanation of the patient consent issues to the donor. Specifically, it requires informed patient consent for the intended use of donated samples and any subsequently derived biological materials, as well as the derived data associated with each sample. The Pharmacogenetics Working Group has set out a categorisation procedure[7] for both samples and data for genetic research (**Table 1**). Patient sample anonymisation is achieved by using a lock and key system where the key is held by a neutral party, the patient identity protected, the end user of the biological materials and data arising out of its use can freely use data. The understanding is that all processes and data arising out of its use is available for scrutiny and back traceability in accordance with strict regulatory and ethical guidelines.

These operating procedures require interdisciplinary groups to work together often with no direct benefit to themselves. For example, collection of healthy samples by a nurse or doctor, together with associated phenotypic and medical data, is not a high priority as it detracts from the 'day job'. This notwithstanding, properly collected and anonymised samples represent a valuable and very finite asset, as re-access to the patient is a significant problem. Existing patient samples are therefore being carefully preserved for future, but as of yet, undefined analyses. In this respect, the sample preservation issues affecting genomics parallel

those in the HTS field where resynthesis of chemicals contained within the corporate collection is in practice unrealistic.

Sample access, management and integrity

The increased requirement for the rapid provision of large numbers of chemically synthesised test samples for HTS has required a radical rethink of the storage and retrieval algorithms employed. The early years of HTS were supported from entirely manual stores where aliquots were weighed out by hand to supply to the screening laboratory. Microplates containing the subsequently solubilised compounds were prepared locally. Previously, sample data was provided via hand-typed labels affixed to sample bottles, the data being manually transposed by the screening laboratory. The microplates thus produced were tracked using locally applied and frequently degenerate identities. This approach did not bode well for the overall integrity of sample data.

This position has been replaced by the deployment of modern automated sample stores that have built on automated warehousing principles employing sophisticated tracking software and process hardware[8,9]. Such stores have automated that entire manual process beginning with sample registration, storage and retrieval, through to preparation and supply of assay plates to the relevant screening facility, while providing full sample identification through use of sophisticated process and environmental control with full audit capability. No regulatory constraints have been applied in this area to date.

The management of genomics samples presents a key difference from that in Pharma LI in that the geographic distribution of sample processing, storage/access and utilisation is complex (**Figure 2**). Samples may be collected, processed, stored, accessed and analysed and interpreted in multiple distinct locations, thus making the whole process of sample acquisition, processing, registration and data tracking extremely complex and labour intensive.

Sample tracking and integrity

A second difference in the processing of samples from a biorepository compared to that of a

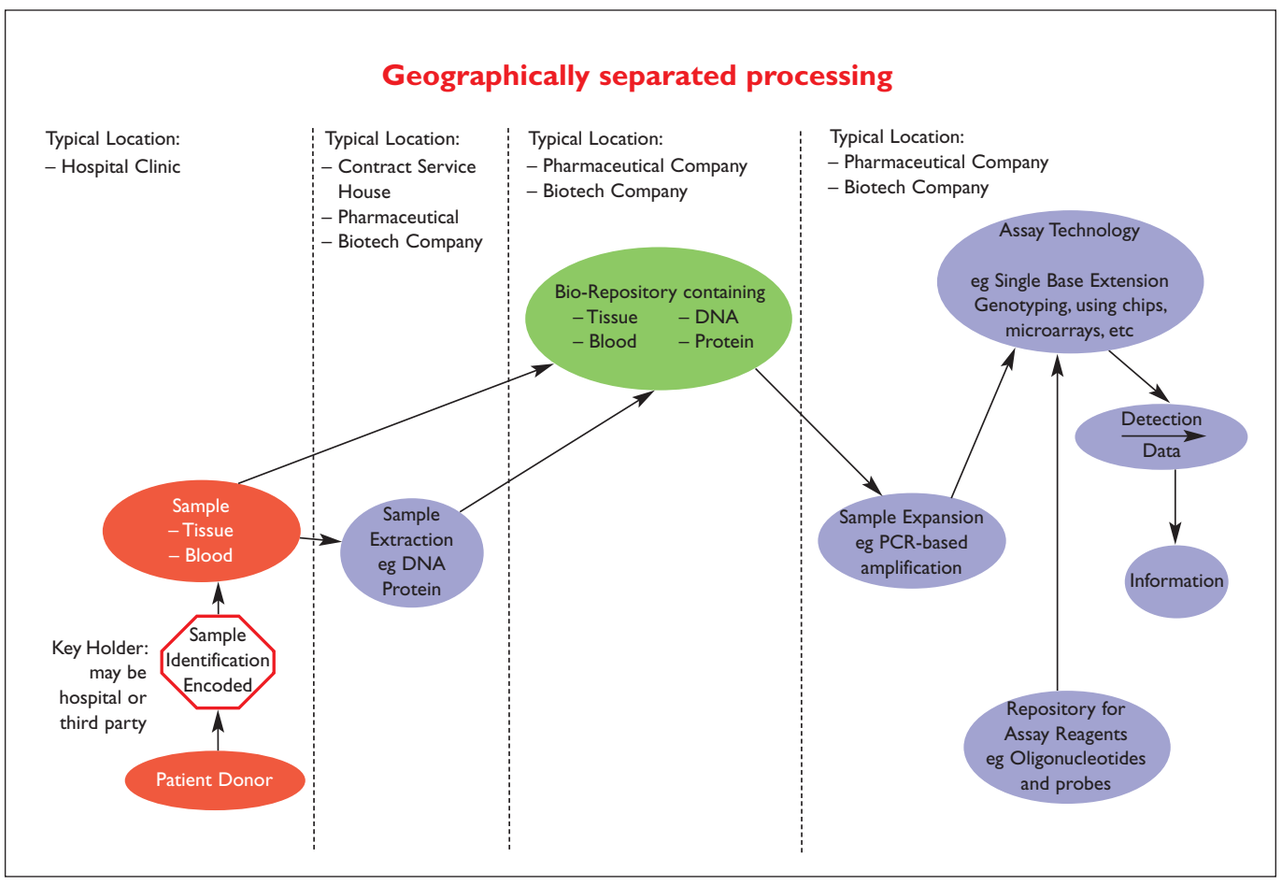

Figure 2
A comparison of the business rules, process needs, ethical considerations and LIMS/data handling requirements for genotyping and HTS applications

chemical compound repository, is that of cross contamination. The pharmacological potency of library file compounds is such at a 1 in 1,000 level of contamination will pass unnoticed in a biological assay. In contrast many of the technologies employed in a genotyping context require the use of the polymerase chain reaction technique to enrich the sample prior to analysis. Enrichment systems may allow non-specific amplification of a contaminating DNA, thus making the whole area of traceability and the minimisation and/or measurement of cross contamination a key issue for applications such as genotyping. The use of automated biorespositories is becoming a key area to ensure full traceability and regulatory compliance. An example of a typical biorepository configuration provided by The Automation Partnership is shown in **Figure 3**.

Assay technology

Major advances have been made in the field of HTS assay technology to support the operation of screening at increasing throughputs. An example of such improvements has been the reduction in the number of process steps to avoid separation steps by elimination processes such as centrifugation, phase extraction or filtration. Particularly noteworthy in this respect was the introduction by Amersham of scintillation proximity assay technology that allowed receptor binding and immunoassays to be run at high throughput for the first time. Other similar enabling technologies including HTRF, TRF, and FCS followed[10]. The second round of improvement was that of assay miniaturisation and the accompanying improvements in micro-plate design and manufacture. HTS assays are now routinely run in 384-well format as opposed to the original 96-well format. Higher density plate formats (1536 and 2300) are also available though these are currently employed to a much smaller extent.

Within the genomics world, the ability to genotype at throughput levels of 20,000 genotypes per day upward has been enabled by those technologies that have provided a fully supported approach. Such technologies have provided integrated packages scoping the input of DNA sample, assay assembly, result detection and interpretation in conjunction with a fully traceable software package. Examples of fully supported technologies include Sequenom's Mass Array systems[11] and Orchid's SNPatron system[12]. The next significant step will be to allow sample conservation by using multiplexing and automatic data interpretation

Table 1

Categories for genetic research samples and data

IDENTIFIED SAMPLES
Data are those labelled with personal identifiers such as Name or Social Security Number. Use of a clinical trial subject number does not make the sample/data identified.

CODED SAMPLES
Data are those labelled with a clinical trial subject number that can be traced or linked back to the subject only by the investigator. Samples do not carry any personal identifiers.

DE-IDENTIFIED SAMPLES
Data are double-coded and labelled with the unique second number. The link between the clinical study subject number and the unique second number is maintained, but unknown to investigators and patients. Samples do not carry any personal identifiers.

ANONYMISED SAMPLES
Data are double-coded and labelled with the unique second number. The link between the clinical study subject number and the unique second number is deleted. Samples do not carry any personal identifiers.

ANONYMOUS SAMPLES
Data are those that do not have any personal identifiers and identification of the subject is unknown. Anonymous samples may have population information (eg, the samples may come from patients with diabetes, but no additional individual clinical data).

Source: The Pharmacogenetics Working Group Working Paper 1

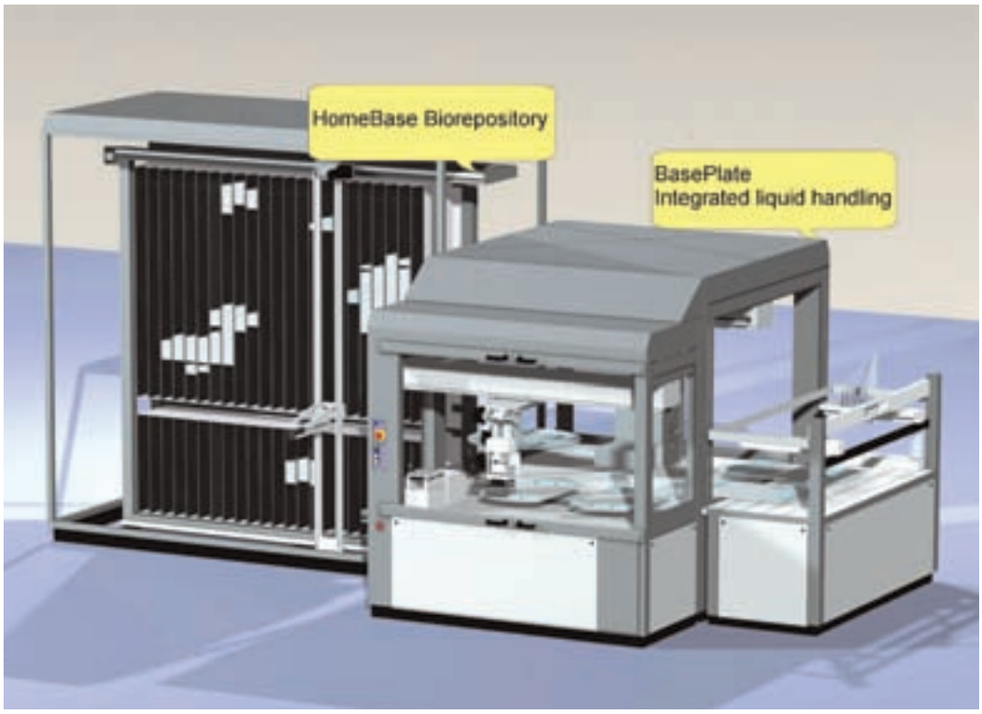

Figure 3
An example of an automated biorepository showing the automated storage, retrieval and management elements necessary for biological samples (including DNA). Samples are contained in the HomeBase store under appropriate environmental control. These are retrieved for replication, reformatting and issue using the BasePlate liquid handling platform

techniques. Future replacement strategies for the current microplate-based technologies utilise alternative physical platforms such as microarrays, chips, CDs and 'lab on a chip'. Although the uptake and utilisation of these alternative approaches is currently patchy, the desires to decrease assay costs and decrease sample usage are likely to ensure the emergence of one or more of these technologies.

Data processing

The requirements for processing HTS data have been relatively straightforward. Instrument data is associated with the micro plate from which the assay signal was determined and a link made to the identity of the test compounds contained in that particular micro plate. Once this link has been established, the percentage effect of each compound is calculated relative to internal assay control samples within each screen run. Screen operatives examine predefined quality control parameters that allow easy detection of systematic human or robot errors. Any such data are eliminated from the approved data set. Large volumes of summary data for each screen run are subsequently stored on corporate databases thereby making it available for interrogation by chemi- and bio-informaticians. The majority of HTS data is negative, showing no correlation between compound structure and biological activity. A small number, typically 0.1-0.5%, of compounds show a positive effect and are scheduled by the HTS software for a more detailed pharmacological examination in subsequent runs of the screen in question. Few rules, outside those intra-company compound exchange agreements, govern the release of information within the organisation that sponsors LI screens.

In contrast, the access to public data and the 'fee for access' to SNP and sequence data requires that data mining is possible as disparate and remote locations throughout the genomics world. However, the plethora of different formats of data has meant that the ability to compare data from differing platforms, for example between microarray-based systems, microplate-based systems and chip-based systems, is extremely difficult. It has resulted in a need for standardisation and inter-technology compatibility in order to allow genomics to move forward.

Process automation

Accompanying the improvements in HTS assay technology was a concomitant increase in available process automation. These range from discrete workstations, able to process one micro plate at a time to modern uHTS platforms with capacities in excess of 300 microplates per run.

LI teams have deployed process automation to

References:

1 Herceptin: www.nice.org.uk
2 A Revolution in R&D. The Boston Consulting Group, November 2001.
3 www.appliedbiosystems.com
4 www.hgp.com
5 The Pharmacogenomics Journal 2001 1: 23-26.
6 Regulatory Requirements for Inclusion of Pharmacogenetic Testing in Applications for Clinical Trials in Europe. J Regulatory Affairs, Feb 2001.
7 Pharmacogenetics Working Group Working Paper 1.
8 Harrison, WJ (1997). The importance of automated sample management systems in realising the potential of large compound libraries in drug discovery. J. Biomol. Screen 2.203.
9 Holland, S (1998). Breaking the bottleneck in the lead discovery process: the development of the automated store (ALS) at Glaxo Wellcome. In Eurolab Automation 98 Meeting proceedings. Oxford, UK p.31.
10 Seethala, R (2001). Homogeneous assays for high throughput and ultra high throughput screening. Handbook of Drug Screening, eds Seethala, R& Fernandes, P Marcel Dekker, NY.
11 www.sequenom.com
12 www.orchid.com
13 Archer, R (1999). Faculty or Factory? Why Industrialised Drug Discovery is Inevitable. J. Biomol. Screening 4, 235-237.
14 Beggs, M (2000). HTS – Where next? Drug Discovery World 2000 25-30.

varying extents depending on the desired scale of operation and the availability of a work force prepared to accept a high level of repetitive work. Arguably, the current suite of leading edge HTS automation has evolved to a point where it is a mature technology and where further increases in throughput are unlikely. If throughputs of 200-300,000 wells per day are being realised by some practitioner, an average-sized compound file can, in principle, be screened in less than a week. The real issue now to be addressed is the one of improving the sustainability of operation.

Multi-site vs single-site approaches

This paper has outlined the objectives of both HTS in the context of Discovery LI and genotyping. Pharma organisations have aspirations to increase the scale of both sets of activities very significantly. Despite a very significant installed HTS capacity, the number of sample wells screened per annum represents less than 10% of the installed capacity. Indeed it is rare to find, even within large Pharma houses, organisations that actually deliver more than 25 million wells screened per annum. An average return of 10-15 million wells screened per annum is the norm. This contrasts with organisational objectives of delivering in excess of 150 million wells screened per annum. In other words an increase of some 10-15 fold is still required. In a similar vein, the increases in scales of genotyping approaches requires moving from 50,000 SNP assays per day to 500,000-2 million SNP assays per day over the next 2-3 years. This represents an increase of between 10 to 40 times the currently achieved performance.

It is the contention of this article that, to date, both sets of activities have been operating in a pioneering mode in which the objective was to surmount the major technological hurdles. It is a tribute to all those involved that so much progress has been made in this direction. However, to realise the desired end states, both disciplines will have to undergo a final transition in which the high skills base activities can be reduced to practice and can be successfully operated 'at scale' on a sustainable basis. This will require careful attention to the non-technological aspects of both operations. Such facets include pro-active management of all logistic 'feed' lines, the development of appropriate organisational models and skill sets for the sustained operation and careful attention to facility design and logistical process. Some leading thinkers in the HTS world are already constructing the first 'drug discovery factories' along these lines[13,14].

The nature of the industrialised genomics operation will differ from that envisaged for HTS. The design must reflect the necessary segregation of parts of the process from sample collection, which takes place at clinics and hospitals, through to extraction of DNA from biologically derived materials. The likely approach envisages discrete parts of the process being established in separate geographic locations, but the design of the functional units will be horizontally transparent with respect to sample tracking, LIMS integration and process control and co-ordination. This environment will support full audit trails and regulatory compliance.

Ultimately, the output from both HTS and genotyping activities is the same: quality data. The ultimate winners in both fields will be those organisations able to generate large volumes of data and understand how to interpret the resulting complex chemi/bio-informatic or genetic linkage data sets. The real benefit accruing from 'process industrialisation' will be an improvement in data quality and consistency that aids comprehension. DDW

Dr Mark Beggs *is Manager – Integrated Discovery at TAP in which role he provides consultancy on complex screening system design. Mark is a biochemist/enzymologist and was formally Director of HTS at Janssen, and prior to that was involved in HTS system operation and assay development at Zeneca and GW. Mark is a Council Member of the Society for Biomolecular Screening. He has published and spoken widely on the subject of Pharma productivity.*

Audrey Long *is Genomics Marketing Manager at TAP. Audrey is a biochemist/molecular biologist with experience in the pharmaceutical, biotech and genomics fields. Prior to joining TAP she was involved in setting up the European arm of Lark Technologies, a molecular biology contract research organisation (CRO). In addition, previous positions included roles within Pharmacia Biotech's industrial division (now Amersham Biosciences) working with pharmaceutical and biotechnology companies for FDA regulated protein and nucleic acid production and purification.*