INNOVATION

Target discovery

Mark A. Lindsay

Target discovery, which involves the identification and early validation of disease-modifying targets, is an essential first step in the drug discovery pipeline. Indeed, the drive to determine protein function has been stimulated, both in industry and academia, by the completion of the human genome project. In this article, we critically examine the strategies and methodologies used for both the identification and validation of disease-relevant proteins. In particular, we will examine the likely impact of recent technological advances, including genomics, proteomics, small interfering RNA and mouse knockout models, and conclude by speculating on future trends.

In spite of increased spending on pharmaceutical research and development, which doubled in the United States between 1991 and 2001, the number of new drug approvals has remained relatively constant at 30 per year (new molecular entities). Furthermore, the emergence of molecular biology and the completion of the human genome project have failed to produce the expected flood of compounds targeting new, or as often termed 'novel', targets. So, with the licensing of only two to three compounds per year against novel targets, the majority of approvals continue to be those directed against therapeutically validated targets[1]. Moreover, these drugs are still predominately small molecules or protein/ antibody biopharmaceuticals that exert their action through modulation of protein activity and which are therefore restricted to the 'druggable' targets. An exception to this trend is the recent licensing of formivirsen (Vitravene; ISIS), an antisense treatment for cytomegalovirus retinitis[2], which indicates that the range of potential targets could in future be increased through modulation of protein expression at the translational level[3].

The reasons for the low numbers of successful drugs against novel targets are controversial, although it is undoubtedly related to their higher attrition rate during development, particularly from issues related to failures of on-target biological hypotheses and on- and off-target safety concerns. To address these problems, there has been a recent emphasis on the 'front-loading' of research to tackle these issues at an earlier stage in the drug discovery pathway. In particular, there has been increased interest in target discovery both for the identification of novel targets (target identification) and to reduce the subsequent failure from incorrect biological hypotheses through early validation (target validation).

This article aims to critically review the techniques and strategies used in target discovery. With this in mind, the initial discussion will be focused on the two main strategies used, which are the 'molecular' and 'systems' approaches. This will be preceded by an examination of the three areas of target discovery, the provision of disease models and the techniques used in target identification and target validation, before we speculate on future requirements and trends. In reviewing the ever-increasing number of novel and innovative technologies for target discovery, lack of space has meant concentrating on those in general use. Furthermore, as a result of the large differences between therapeutic areas, this will by necessity be approached in a broad way and we shall concentrate on common chronic and not infectious/acute disease.

Strategies in target discovery

The various techniques applied in target identification and validation can be grouped into two broad target discovery strategies: the 'molecular' (FIG. 1a) and 'systems' approach (FIG. 1b). In practice, however, both are used in varying proportions within different therapeutic areas. The 'systems' approach discussed here should not be confused with the recent emergence of 'systems biology', which is an attempt to construct models that explain biological responses using the vast amounts of information being produced from the molecular sciences[4].

In recent years there has been a significant shift towards the molecular approach in an attempt to identify new targets through an understanding of the cellular mechanisms underlying disease phenotypes of interest (FIG. 1a). For this reason, this approach is focused on the cells implicated in the disease and uses clinical samples and cell models. The molecular approach has been driven by the enormous experimental successes of molecular biology, and in particular genomics, the zenith of which so far is perhaps completion of the human genome project. In terms of target classes, the molecular approach is more likely to identify intracellular targets, such as regulatory, structural and metabolic proteins, and has been most extensively deployed in the area of oncology.

The systems approach is geared towards target discovery through the study of disease in whole organisms (FIG. 1b). In general, this information is derived from the clinical sciences and *in vivo* animal studies in the areas of physiology, pathology and epidemiology. The systems approach has traditionally been the main target-discovery strategy and this remains the case for many diseases, including obesity, atherosclerosis, heart failure, stroke, behavioural disorders, neurodegenerative diseases, hypertension and dislipidaemia, in which the relevant phenotype can only be detected at the organismal level. For these historical reasons, the majority of

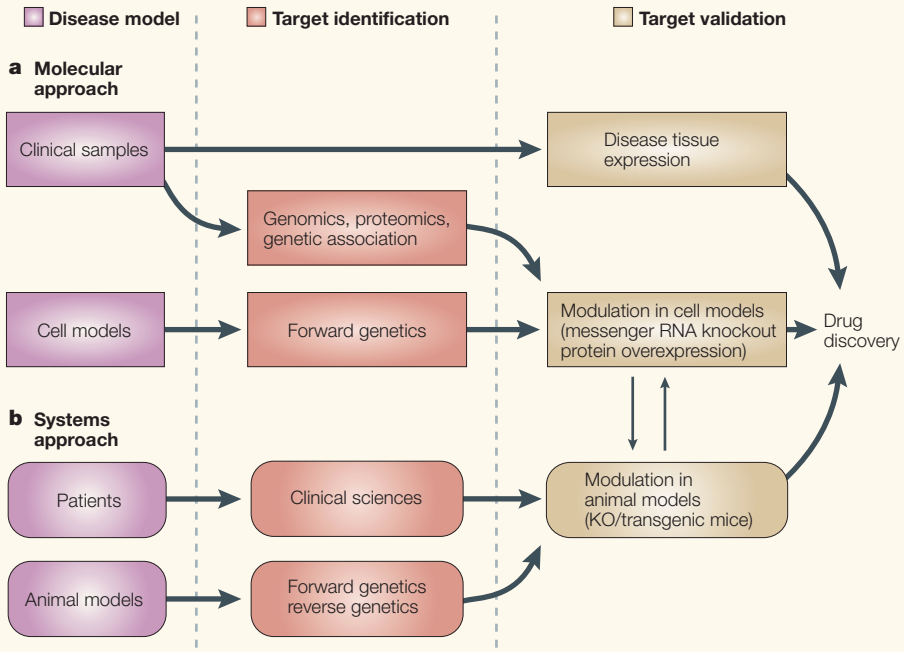

Figure 1 | **Overview of molecular- and system-based approaches to target discovery.** Target discovery is composed of three steps: the provision of disease models/tissues, target identification and target validation. The 'molecular' approach (**a**) uses techniques such as genomics, proteomics, genetic association and reverse genetics, whereas the 'systems' approach (**b**) uses clinical and *in vivo* studies to identify potential targets. During validation, modulation of gene expression and/or protein function in both cell and animal models is used to confirm the role of the target prior to passing into the drug discovery pipeline.

current drugs were identified through this strategy and include those that act against both disease phenotypes and intracellular/extracellular targets. Interestingly, because many of these drugs are directed against targets that were identified from physiological studies, rather than being directly implicated in the disease mechanism, they would probably not been identified by the molecular approach. For example, although changes in β_2-adrenoceptor expression/activity in airway smooth muscle has not been implicated in the mechanism of allergen hyper-reactivity that produces airway contraction in asthma, these symptoms are commonly treated with β_2-agonist[5].

Disease models

The incidence of many chronic diseases is strongly correlated with age, and such diseases are thought to be influenced by both genomic and the environmental factors. The overall contribution of genomic factors is still unknown, although it is believed that many diseases are influenced by the presence of susceptibility genes[6], and it is known that the development of many cancers results from alterations in the genetic material of somatic cells[7]. Similarly, with the exception of smoking[8], the role of environmental factors is controversial, although a number of studies have indicated the importance of infection/inflammation[9–11] and diet[12] in diseases such as atherosclerosis, CNS disease and cancer. The ultimate effect of these interactions is to produce an often permanent shift in phenotypic response with the corresponding pathophysiological and anatomical changes that are characteristic of the disease.

In undertaking target discovery, one would ideally perform clinical studies and obtain cell/tissue samples using normal and diseased human patients. In reality, this is normally unethical and/or impractical, which means that we must use cellular and/or animal models. These models, which are designed using information derived from clinical studies, normally attempt to reproduce one or more phenotype(s) that have been implicated in disease. Unfortunately, they often suffer from a number of significant problems that can make them poor predictors of human disease[13]. In the case of cell models, the central problem is in simulating the complexity of the *in vivo* biological interactions, particularly as many of these are unknown. This problem of complexity makes it increasingly difficult to predict the role of a protein as you proceed from the level of the cell to the tissue and organism. In addition, the use of immortalized lines to overcome the problems of availability begs the usual questions regarding their biochemical similarity with primary cells. To overcome the problems of complexity, we often use animal models (BOX 1). However, although these models can reproduce a particular disease phenotype, genomic differences (related to species and strain), and the difficulty of identifying and replicating the long-term environmental influences, means that underlying causes could be different[13].

Target identification

Target identification attempts to identify new targets, normally proteins, whose modulation might inhibit or reverse disease progression (FIG. 1). In recent years, the molecular strategy has predominated and led to the emergence of technologies that attempt to correlate changes in gene (genomics) and protein (proteomics) expression or genetic variation (genetic association) with human disease (FIG. 2). This approach is dependent on access to good clinical samples with supporting medical data, and requires a strong bioinformatics platform for data processing and interpretation. To date, genomics and proteomics have proved to be of limited utility in target identification, although they are increasingly being used in other areas of drug discovery, including toxicology (toxicogenomics) and the identification of disease biomarkers[14]. So, although these techniques identify large numbers of targets, the correlative nature of the data they generate means

Box 1 | **Apolipoprotein E knockout model for the study of atherosclerosis**

Atherosclerosis is a chronic disease that in its final form is characterized by cholesterol-rich lesions or plaques within the large and medium-sized arteries that are thought to contribute to acute manifestations of cardiovascular disease. These plaques start in the form of fatty streaks containing lipid-laden macrophages that develop, over a period of decades, into fibro-lipid lesions containing endothelial cells, monocytes/macrophages, smooth muscle cells and T cells. Although wild-type mice are generally resistant to the development of atherosclerosis, this can be induced in a number of inbred strains fed a diet that promotes hyperlipidaemia. However, apolipoprotein E knockout mice, which lack the major carrier of plasma cholesterol, spontaneously produce lesions on a normal diet. The progression of these lesions develops with age in a way that resembles that seen in humans and has made this an ideal model for the study of the development of atherosclerosis and the effects of diet. However, as with many animal models the genetic background influences susceptibility to atherosclerosis[66].

that it is not possible to determine whether changes in gene protein expression are a cause or effect of the disease. In these circumstances, time-consuming validation is often required and the problem becomes one of assigning priority among the large number of possible targets. An alternative strategy has been phenotype-orientated target identification that can be broadly divided into the areas of 'forward genetics' and 'reverse genetics' (FIG. 3). The former involves random modulation of the phenotype and the subsequent identification of the relevant affected gene (phenotype to gene), whereas the latter entails gene manipulation and examination of the phenotype (gene to phenotype).

Genomics

A genomics approach attempts to identify novel disease targets at the level of gene expression through the comparison of normal and diseased tissues (FIG. 2a)[15–17]. For these studies, gene microarray chips containing a collection of oligonucleotides are used for the rapid and parallel determination of messenger RNA expression in an RNA sample. Although it is possible to produce personal arrays through reference to genomic sequences or expressed sequence tags (EST), increasingly these arrays are obtained from commercial sources, because of increased quality and selection, and decreased costs. Significantly, the completion of the human genome project has resulted in the release of the Affymetrix U133 chips, which can be used to measure the expression of 32,000 human genes in a single experiment. Furthermore, the ongoing sequencing of the genomes of a range of other species will increase the availability of chips for the determination of gene expression in organism such as rats, mice, *Escherichia coli*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*.

As stated, these studies attempt to identify novel therapeutic targets by virtue of the fact that the gene(s) is differentially expressed in disease tissue. The number of 'hits' is dependent on the magnitude of change that is thought to constitute a significant biological effect, but is typically in the range of hundreds of genes. For example, a recent differential expression study of 9,183 unique complementary DNAs in normal and ductal breast cancer cells demonstrated a twofold increase or decrease in the expression of 303 genes[18]. This isfurther complicated by the presence of large numbers of false positives/negatives, which can be reduced (but never completely eliminated) through careful experiments designed to reduce background noise generated by both technological and biological

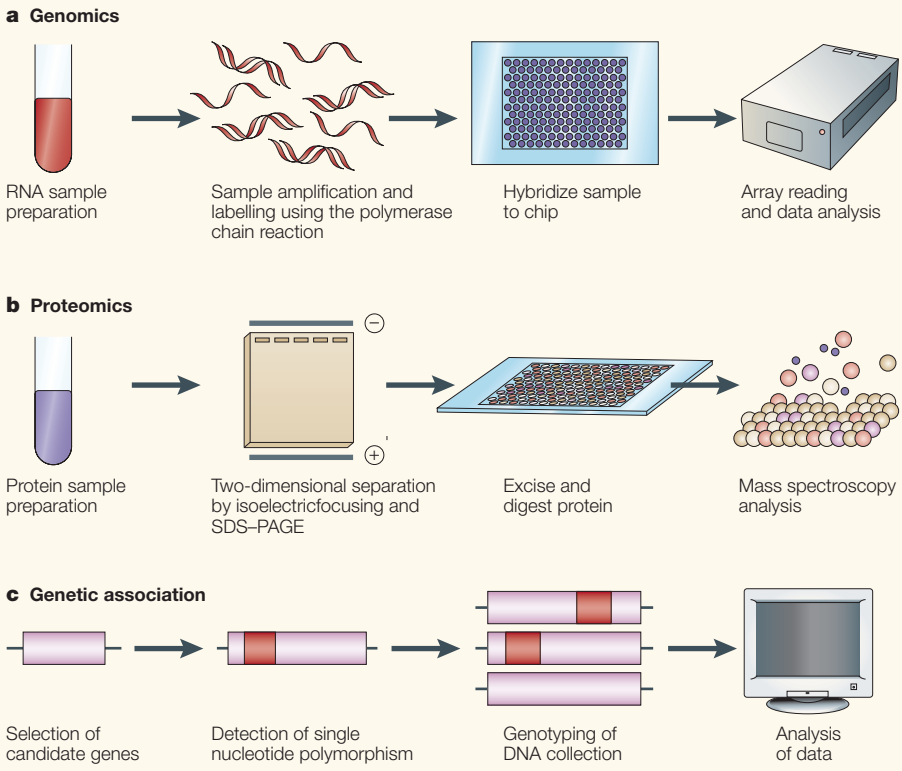

Figure 2 | **Correlative technologies used in target identification.** Techniques such as genomics (**a**), proteomics (**b**) and genetic association (**c**) attempt to identify novel targets through the measurement of the differential expression of messenger RNA, protein or genetic polymorphisms in diseased and normal tissue. SDS–PAGE, sodium dodecyl sulphate-polyacrylamide gel electrophoresis.

factors. However, as well as the correlative nature of these data, the most significant problem is that changes in mRNA level cannot always be automatically translated into corresponding alterations in protein expression/activity.

Proteomics

In its broadest interpretation, proteomics attempts to understand cellular function through the measurement of protein expression, activity and interaction with other biological macromolecules[19]. This has been traditionally undertaken using two-dimensional gel electrophoresis to separate the proteins, which are subsequently cut from the gel, enzymatically cleaved into fragments and possibly fractionated by liquid chromatography before identification using mass spectroscopy (FIG. 2b)[20]. To increase the throughput of this process, there has been limited recent movement towards the production of various protein array chips[21].

At present, there are significant technical and biological problems associated with the separation of proteins, which is a crucial step for both the identification of individual proteins and for the comparison of different samples[22]. So, a number of protein classes, such as those that are associated with membrane, positively charged and hydrophobic, are often difficult to separate using two-dimensional gels. Furthermore, the lack of an amplification step, such as the polymerase chain reaction used in genomic studies, means that the detection of low-abundance proteins is problematic. The separation characteristics of proteins are also affected by post-translational modifications, including phosphorylation, glycosylation, lipidation, acetylation and nitration. In light of these problems, the utility of proteomics in target discovery has so far been limited.

Genetic association

Mutations in just a single gene can cause severe disease, as exemplified by rare genetic diseases such as Huntington's disease[23], Duchenne muscular dystrophy[24] and cystic fibrosis[25]. Indeed, the identification of these mutations has been useful in determining disease mechanism and in aiding target identification. For instance, the potential role of increased plasma cholesterol in the development of atherosclerosis and coronary heart disease was first noted from studies of patients who were either homozygous or heterozygous for a defective allele for the receptor

for low-density lipoprotein (LDL), the principal carrier of plasma cholesterol (BOX 1)[26]. In general, the strict co-inheritance of specific mutations and disease has meant that the relevant genes were identified by linkage analysis and subsequent gene identification within this chromosomal region.

However, it is believed that most common diseases are complex disorders that are influenced by multiple genes[27]. Because mutations of no single gene segregates tightly with disease phenotype, this means that it is very difficult to localize such disease-related genes to specific chromosomal locations. Under such circumstances, the possible role of a gene in disease can be investigated using genetic association studies. These studies attempt to determine the relationship between a specific DNA variation (an allele), which is normally a single nucleotide polymorphism (SNP), and disease[28] (FIG. 2c). To embark on these studies, it is first necessary to select your gene(s) of interest, a decision that is normally based on a hypothesis about the biological function of the gene of interest. It is then necessary to identify one or more SNPs, which can be functional — such as those producing amino acid changes — or apparently non-functional but within the coding or non-coding region. In the latter case, it is assumed that the SNP might influence factors such as the rate of transcription of the gene, or affect mRNA processing and stability. When choosing these SNPs, it is normal to focus on common variants, many of which have recently been catalogued[29]. In an attempt to identify an association between the SNP and disease, comparisons are then made of the distribution between the genotyped control and diseased populations, either through examination of groups of case-controls, retrospective investigation of cohort studies or family-based association.

As with genomics and proteomics, genetic association shows only correlation between polymorphisms within a gene and disease. Furthermore, because SNPs in nearby genes are often inherited together, owing to linkage disequilibrium, it is often not possible to identify a single gene that is responsible for the observed association. In addition, the lack of reproducibility in genetic association studies often necessitates confirmation by several independent studies[28,30].

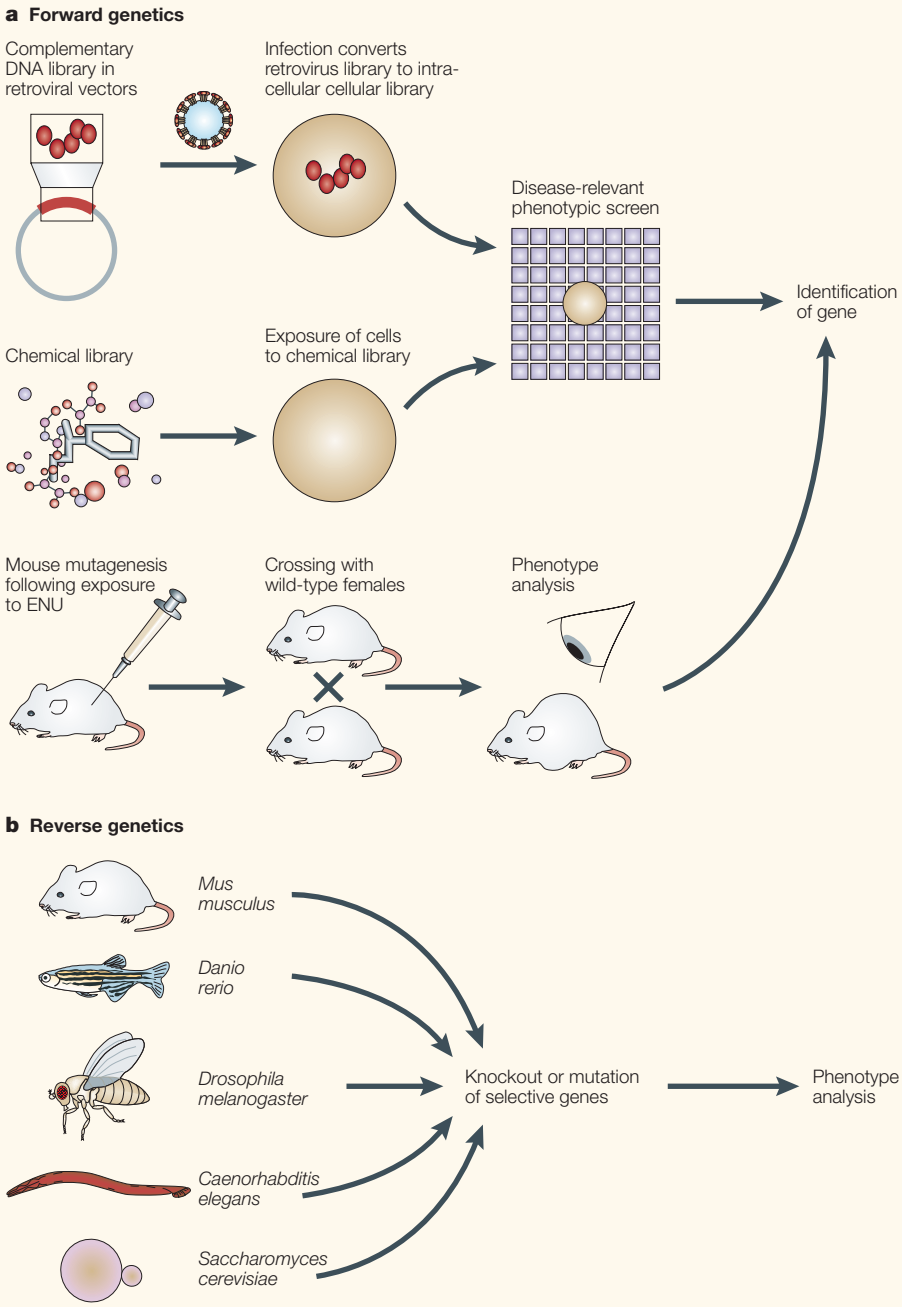

Figure 3 | **Phenotype-driven target identification.** Technologies such as forward (**a**) and reverse (**b**) genetics involve the identification of potential disease targets through modulation of a disease phenotype. Forward genetics involves random modulation of the phenotype using *in vitro* biological and chemical library screen or *in vivo* chemical mutagenesis and the subsequent identification of the gene (phenotype to gene), whereas reverse genetics entails gene manipulation and examination of the phenotype (gene to phenotype). ENU, ethylnitrosured.

Forward genetics

Forward genetics (that is, going from phenotype to gene) can be performed using both cell and animal models, and is dependent on the provision of a method to disrupt the phenotype, a reliable and normally high-throughput biological assay and a strategy for the identification of hits or novel targets (FIG. 3a).

Using cellular models, a library of 'chemical' or 'biological' agents is used to randomly modulate the phenotypic response. Forward 'chemical' genetics uses chemical libraries derived from a variety of sources, including commercial collections, natural products and diversity-orientated synthesis[31]. Although the process of hit identification is laborious, being traditionally performed with labelled compounds, this approach has the advantage of giving a lead compound for drug development. Biological screening has been most commonly performed using cDNA[32] libraries constructed from a variety of sources, including diseased tissues, and delivered and expressed from a viral vector. Because this

causes protein overexpression (gain-of-function), such vectors are normally used for the discovery of negative modulators of a phenotypic response and the identification of secreted proteins/modulators.

In animal models, forward genetics has been most extensively performed using the chemical mutagen ethylnitrosourea (ENU) to produce mouse mutations. Indeed, a number of consortiums have recently undertaken a systematic, genome-wide production and phenotypic analysis of these mutant mice to identify both disease targets and models[33,34].

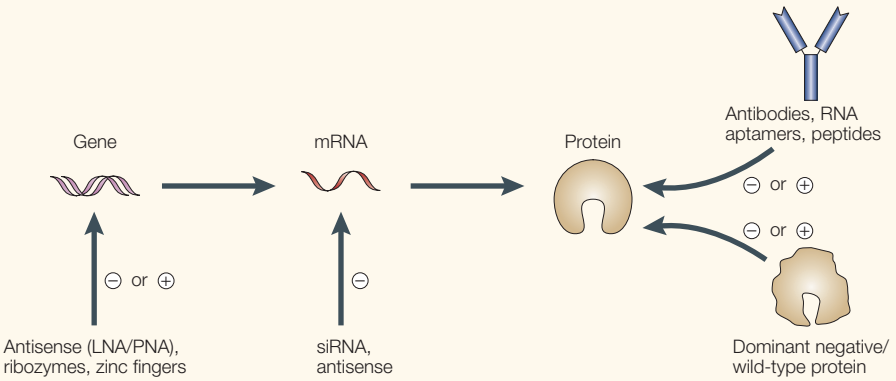

Figure 4 | **Overview of the techniques used in target validation.** The elucidation of target function can be undertaken using a variety of target-validation techniques. Expression of messenger RNA can be modulated either directly (small interfering RNA (siRNA) or antisense) or at the transcriptional level (peptide nucleic acids (PNA), locked nucleic acids (LNA), zinc fingers or ribozymes). Protein function can be modulated through expression of dominant negative and wild-type protein or using blocking/stimulating antibodies or aptamers.

Reverse genetics

The advent of sophisticated genomic tools applicable to a wide range of model organisms, combined with their typically short life cycles and the ease of generating large numbers of individuals for study, has inspired the use of organisms such as *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *Danio rerio* (zebrafish) in target identification[35] (FIG. 3b). Indeed, functional mutation studies, in conjunction with human homologue searching, has been successfully used to characterize the biochemical pathways that mediate cell-cycle regulation, p53 signalling, mitogen-activated protein (MAP) kinase signalling and apoptosis. In addition, these types of studies can be used to identify hits from forward chemical genetics. In this case, the initial studies must discover a phenotype that is affected by the compound; subsequently, the target is identified by screening a library of known mutants for resistance to the compound[35]. Although studies in lower organisms have clearly been useful for the elucidation of many important biochemical pathways, the principal issue is their relevance to human physiology and disease.

More recently, there has been a systematic attempt to apply reverse genetics approaches in mice. Of particular interest has been the attempt by companies such as Lexicon Inc. to construct and characterize 5,000 knockout animals, including all of the druggable classes (see Animal Validation Studies)[36].

Target validation

The role of target validation is to demonstrate the functional role of the potential target in the disease phenotype. Although this is ultimately dependent on human studies, validation in target discovery will normally require that the target is expressed in the disease-relevant cells/tissues, and that target modulation in cell and/or animal models ameliorates the relevant disease phenotype (FIG. 1). The first criterion involves the measurement of protein and/or mRNA expression in clinical samples using immunohistochemistry and *in situ* hybridization, respectively. Although protein detection is the preferred option, this technique is often limited by the availability of selective antibodies. For the second criterion, a range of strategies exist for modulating target expression at the protein and mRNA levels, although their utility is often limited by a lack of effective cellular delivery technologies (FIG. 4). Animal studies, which are normally the decisive factor in the decision to proceed to drug development, predominantly entail the use of knockout or transgenic mice.

Protein modulation

The modulation of protein function can be undertaken using either overexpression of wild-type dominant negative proteins or affinity reagents. Transient overexpression of dominant negative proteins from plasmids or viral vectors is commonly used, although its application is limited by the need for prior knowledge on how to disrupt the relevant functional domain. Furthermore, these studies can produce misleading results because of the occurrence of nonspecific biological effects at the elevated protein concentration typical of these studies. As an alternative, a number of commercial and academic groups have recently pioneered the use of affinity reagents — including antibodies[37], peptides[38] and RNA aptamers[39] — that can be used to activate or attenuate protein function. However, because their identification is time consuming and expensive, requiring repeated panning of the target against a library of affinity reagents, these are rarely used for target validation studies alone, but often provide the starting point for a biopharmaceutical.

mRNA modulation

The inhibition of mRNA expression, and subsequent protein knockdown (loss-of-function), is frequently used in cellular validation studies. This can be accomplished through the use of agents that target gene transcription, including peptide nucleic acids[40], locked nucleic acids[41], ribozymes[42] and zinc-finger proteins[43]. However, the most widely used strategy is the direct targeting of mRNA expression levels using antisense[44] or small interfering RNA (siRNA)[45]. The former approach uses DNA-based antisense sequences that hybridize to the target mRNA and induce RNase H/exonuclease-mediated degradation. Initially, antisense molecules suffered from a number of limitations, including their stability, toxicity and the difficulty in identifying active sequences, although these problems have been substantially overcome following the development of novel chemistries and predictive sequence algorithms. The RNA interference (RNAi) response is an innate cellular response that is involved in combating viral infection and in regulating mRNA expression[46]. RNAi is activated by siRNA molecules composed of double-stranded RNA of 21–23 nucleotides in length, and mediates the degradation of the corresponding single-stranded mRNA using the RNA-induced silencing complex (RISC). As it seems to offer a number of significant advantages, including low cost, increased selectivity and the availability of good algorithms for predicting effective sequences, siRNA is rapidly replacing antisense as the method of choice for mRNA modulation. Moreover, siRNA can be expressed with a hairpin loop from plasmid[47] and viral vectors[48,49], which offers the opportunity for both transient and permanent

cellular transfection. However, this initial enthusiasm has been tempered by recent observations of 'off-target' actions[50] (BOX 2).

The great advantage of protein knockdown through mRNA modulation is that only sequence information about the target gene is required. However, there are a number of problems associated with this methodology. For example, there is often no correlation between mRNA and protein levels, which is dependent on a number of factors including protein stability. Furthermore, even in the absence of a phenotypic change, it is difficult to completely eliminate a role for your target, because it is rarely possible to produce total protein knockdown. This can be particularly problematic with membrane targets, such as G-protein-coupled receptors (GPCRs) and cytokine receptors, which can have a significant receptor reserve.

Cellular delivery

One of the key problems in cell- and, in particular, animal-based studies is the availability of effective systems for the delivery of target validation tools. Typically, this is undertaken using cationic lipids[51] and polymers[52], which are not only able to package and neutralize the anionic oligonucleotides, but are also thought to facilitate endosomal release. However, these delivery systems exhibit a number of severe limitations, including the requirement for optimization with each cell type, low transfection levels, toxicity and, most importantly, the difficulty in delivering these systems to primary and non-dividing cells. For these reasons, considerable excitement was generated following the observation that short, cationic peptide sequences, called protein transduction domains (PTDs), seemed to mediate rapid *in vitro* and *in vivo* delivery of peptides, proteins, antisense molecules and plasmids, via a receptor- and endosomal-independent mechanism[53,54]. However, recent observations show that only small quantities of these PTD constructs reach the cytoplasm, with the majority being localized within the endosomes; this indicates that the initial observations were an artefact of the highly basic nature of the PTD[55,56].

An alternative to polycationic systems for delivery into primary and non-dividing cells is the use of viral vectors and electroporation. Viruses have been used for the expression of a range of target validation tools, including dominant negative/wild-type proteins and siRNA[57]. However, because infection is dependent on the expression of specific receptors, this means that viruses such as adenovirus[58], lentivirus[59] and herpes simplex virus[60] can only be used with selective cells/tissues and are often associated with unwanted cellular and immunological responses. Electroporation involves exposing cells and tissues to short, high-voltage pulses to produce a transient and reversible breakdown of the plasma membrane, and has been used for the *in vitro* and *in vivo* delivery of a variety of molecules, including drugs, antibodies and oligonucleotides[61]. In general, the *in vitro* electroporation conditions must be optimized for each cell type and target validation cargo, and its application is limited by high levels of cell death and damage.

Animal validation studies

Animal models continue to be the option of choice in target validation, because they represent the best available model for examining the complex interactions that underline pathophysiological responses. In target discovery, these studies are generally undertaken using knockout or transgenic mice, either in isolation or in conjunction with disease models. An interesting recent development has been the systematic attempt by Lexicon Inc. to produce knockout mice for all druggable genes, including those encoding GPCRs, kinases, phosphatases, nuclear hormone receptors, ion channels and proteases[36,62]. This has been made feasible through the use of homologous recombination combined with viral gene trapping for the production of the embryonic stem cells that are used to produce the knockout animals[63]. This approach not only promises to be of great use in target validation studies but, for the first time, opens up the opportunity of using knockout models for target identification. To this end, this company is undertaking long-term comprehensive phenotype screens to identify targets in areas such as cardiology, central nervous system/neurology, metabolism/obesity, osteoporosis, reproductive biology and oncology[64]. Moreover, this approach can used to identify new disease models (BOX 1) and can provide information with regard to possible toxicity/safety issues.

In addition to the time required to produce knockout and transgenic mice, the most significant problems this approach faces are embryonic lethality and the induction of compensatory mechanisms during development. However, these can be substantially overcome through the construction of conditional expression systems. With transgenic animals, this is achieved by the addition of inducible promoters, such as those for tetracycline or ecdysone, whereas for knockout mice these are produced by flanking the gene of interest between loxP sites that are excised in the presence of CRE recombinase. In the latter case, CRE recombinase expression can be both spatially and temporally regulated using tissue-specific and inducible promoters[65].

The future of target discovery

Of all the biological disciplines, pharmaceutical science is the most rigorous in the sense that the original biological hypothesis is ultimately tested in man using either a small-molecule inhibitor or a biopharmaceutical agent. The high failure rate of drug development, even among those drugs that have undergone extensive validation before entering clinical trials, demonstrates the enormous complexity of biological systems and the difficulties faced in making reliable predictions about the biological role of novel therapeutic targets. For these reasons, probably the most significant problem in target discovery, in particular with respect to target validation, is the provision of models that are truly predictive of disease. In the case of cellular models, molecular studies have been extremely powerful for the determination of biochemical pathways. Nonetheless, although these seem to be conserved, both between cells and species, the role these pathways subserve is strongly influenced by context. It is therefore necessary to have cell models that mimic the *in vivo* milieu and take into account factors such as intercellular interactions (both direct and indirect) and environmental factors. Similarly, although animal models permit the investigation of the disease as well as its development, it is important that they not only manifest the relevant phenotype(s), but that the underlying changes are similar to the human disease. The ability to perform this task will be dependent on an understanding of the causes of chronic disease achieved through longitudinal clinical studies that examine the changes at both the molecular and systems level in parallel. This will permit not only the identification of novel targets but also the development of better cell/animal models, and the identification of biological markers for the accurate and timely assessment of the effectiveness of new drugs in patients.

Following the completion of the human genome project it could be envisaged that target identification, in its broadest definition, has been completed. In these circumstances, the dilemma has now become one of identifying which of these many targets actually cause or modify disease. Although its impact on target discovery is difficult to assess, the initial optimism that molecular techniques such as genomics and proteomics would revolutionize target identification has largely proved to be groundless. This has resulted, in large part, from the problems of functional annotation and, in particular, in confidently eliminating the large number of potential targets that are identified by this approach. So, although it is likely that many of these targets contribute in varying degrees to disease phenotypes, it is the

Box 2 | **siRNA in target validation**

In eukaryotic cells, RNA interference (RNAi) is induced through the delivery of double-stranded RNA sequences of 21–23 nucleotides in length, called small interfering RNA (siRNA), that are complimentary to the target messenger RNA. The cleavage of the mRNA is achieved following binding to the multi-subunit complex termed the RNA-inducing complex (RISC), which is thought to use the antisense strand of the siRNA as a guide. Initial excitement was generated following early studies that showed that siRNA was able to produce relatively high RNA and protein knockdown (>50–90%) and, more importantly, that even single-nucleotide mismatches were able to significantly reduce this action, thereby indicating that this technique could offer a highly selective method of reducing gene expression. However, subsequent investigations have demonstrated that knockdown is not universal and that this is target dependent, with the specific parameters that determine this susceptibility being unknown[67,68]. Of more concern has been a recent study on the knockdown of insulin-like growth factor receptor and mitogen-activated protein kinase-14 (MAPK14, or p38 MAP kinase-α), which reports significant 'off-target' actions. This investigation showed that sequences of as few as eleven contiguous nucleotides sequences located within the 5′-end of the antisense or 3′-end of the sense strand of the siRNA were able to produce gene-silencing of non-targeted genes[50].

identification of the rate-limiting or therapeutically relevant step that is crucial. In this respect, the advent of siRNA for the selective modulation of gene expression would seem to offer a potential solution, although its applicability both *in vitro* and *in vivo* is limited by the availability of effective and non-toxic cellular delivery systems and the recent identification of 'off-target' effects[50].

Undoubtedly, the systems-based approach has provided a valuable source of drug targets and is likely to be boosted through large-scale studies of knockout mice and the subsequent identification of disease-causing targets and of new physiological pathways that can be exploited to modulate disease phenotypes. In fact, the usefulness of this approach is supported by a recent retrospective study of the knockout phenotypes for the 100 best-selling drugs, which showed that of the 34 available, 29 gave information that was indicative of their future therapeutic potential[64]. However, the fact that many of these phenotypic changes were subtle implies that this endeavour will require a large, detailed and multidisciplinary research effort. Nonetheless, these investigations look promising and would be greatly enhanced by the provision of conditional models to overcome the problems of redundancy and embryonic lethality.

In conclusion, we should address the question of why the enormous investment in biological research has failed to produce the expected flood of 'novel' treatments for disease. In addition to the many technical challenges that have been outlined in this review, probably the most significant problem in relation to target discovery is the multifactorial nature of the chronic diseases that are presently the focus of many pharmaceutical and biotechnology companies. In general, previous successes have been derived from a 'top-down' strategy, in which the relevant disease phenotypes were identified using the systems approach, with subsequent development of drugs that target either this phenotype or act through physiological regulatory systems. There are many examples of this process, which include the role of insulin in type 1 diabetes, acid secretion in peptic ulcers and blood pressure in stroke. Although the complexity of chronic disease makes this strategy more difficult, the identification of disease-relevant phenotypes at the system level would still seem to offer the optimum starting point. However, unlike the general trend towards the molecular approach being adopted by both academic and industry scientists, this will require increased emphasis on clinical studies, or a 'human-omics'-orientated approach. In this way, the information would then provide the basis for target identification using molecular and *in vivo* investigations and the optimization of disease models for target validation.

Mark A. Lindsay is at AstraZeneca Pharmaceuticals, 19F19 Alderley Park, Macclesfield, Cheshire SK10 4TG, UK. Honorary Senior Lecturer, Thoracic Medicine, National Heart and Lung Institute, Imperial College School of Medicine, Dovehouse Street, London SW3 6LY, UK.
e-mail: mark.lindsay@astrazeneca.com

doi:10.1038/nrd1202

1. Knowles, J. & Gromo, G. A guide to drug discovery: Target selection in drug discovery. *Nature Rev. Drug Discov.* **2**, 63–69 (2003).
2. Grillone, L. R. & Lanz, R. Fomivirsen. *Drugs Today* **37**, 245–255 (2001).
3. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
4. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
5. Larj, M. J. & Bleecker, E. R. Effects of β_2-agonists on airway tone and bronchial responsiveness. *J. Allergy Clin. Immunol.* **110**, S304–S312 (2002).
6. Venkitaraman, A. R. A growing network of cancer-susceptibility genes. *N. Engl. J. Med.* **348**, 1917–1919 (2003).
7. Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nature Genet.* **33**, 238–244 (2003).
8. Doll, R., Peto, R., Wheatley, K., Gray, R. & Sutherland, I. Mortality in relation to smoking: 40 years' observations on male British doctors. *BMJ* **309**, 901–911 (1994).
9. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).
10. Libby, P. Inflammation in atherosclerosis. *Nature* **420**, 868–874 (2002).
11. Weiner, H. L. & Selkoe, D. J. Inflammation and therapeutic vaccination in CNS diseases. *Nature* **420**, 879–884 (2002).
12. Weisburger, J. H. Eat to live, not live to eat. *Nutrition* **16**, 767–773 (2000).
13. Horrobin, D. F. Modern biomedical research: an internally self-consistent universe with little contact with medical reality? *Nature Rev. Drug Discov.* **2**, 151–154 (2003).
14. Gerhold, D. L., Jensen, R. V. & Gullans, S. R. Better therapeutics through microarrays. *Nature Genet.* **32**, 547–551 (2002).
15. Butte, A. The use and analysis of microarray data. *Nature Rev. Drug Discov.* **1**, 951–960 (2002).
16. Zhang, M. Q. Extracting functional information from microarrays: a challenge for functional genomics. *Proc. Natl Acad. Sci. USA* **99**, 12509–12511 (2002).
17. Slonim, D. K. From patterns to pathways: gene expression data analysis comes of age. *Nature Genet.* **32**, 502–508 (2002).
18. Seth, A. *et al.* Gene expression profiling of ductal carcinomas *in situ* and invasive breast tumors. *Anticancer Res.* **23**, 2043–2051 (2003).
19. Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M. & Fields, S. Protein analysis on a proteomic scale. *Nature* **422**, 208–215 (2003).
20. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
21. Templin, M. F. *et al.* Protein microarray technology. *Drug Discov. Today* **7**, 815–822 (2002).
22. Huber, L. A. Is proteomics heading in the wrong direction? *Nature Rev. Mol. Cell Biol.* **4**, 74–80 (2003).
23. Young, A. B. Huntingtin in health and disease. *J. Clin. Invest* **111**, 299–302 (2003).
24. Khurana, T. S. & Davies, K. E. Pharmacological strategies for muscular dystrophy. *Nature Rev. Drug Discov.* **2**, 379–390 (2003).
25. Ratjen, F. & Doring, G. Cystic fibrosis. *Lancet* **361**, 681–689 (2003).
26. Goldstein, J. L. & Brown, M. S. Molecular medicine. The cholesterol quartet. *Science* **292**, 1310–1312 (2001).
27. Tabor, H. K., Risch, N. J. & Myers, R. M. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Rev. Genet.* **3**, 391–397 (2002).
28. Romero, R., Kuivaniemi, H., Tromp, G. & Olson, J. The design, execution, and interpretation of genetic association studies to decipher complex diseases. *Am. J. Obstet. Gynecol.* **187**, 1299–1312 (2002).
29. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
30. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
31. Lokey, R. S. Forward chemical genetics: progress and obstacles on the path to a new pharmacopoeia. *Curr. Opin. Chem. Biol.* **7**, 91–96 (2003).
32. Lorens, J. B., Sousa, C., Bennett, M. K., Molineaux, S. M. & Payan, D. G. The use of retroviruses as pharmaceutical tools for target discovery and validation in the field of functional genomics. *Curr. Opin. Biotechnol.* **12**, 613–621 (2001).
33. Hrabe de Angelis, M. H. *et al.* Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature Genet.* **25**, 444–447 (2000).
34. Nolan, P. M. *et al.* A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genet.* **25**, 440–443 (2000).
35. Matthews, D. J. & Kopczynski, J. Using model-system genetics for drug-based target discovery. *Drug Discov. Today* **6**, 141–149 (2001).
36. Walke, D. W. *et al. In vivo* drug target discovery: identifying the best targets from the genome. *Curr. Opin. Biotechnol.* **12**, 626–631 (2001).
37. Pini, A. & Bracci, L. Phage display of antibody fragments. *Curr. Protein Pept. Sci.* **1**, 155–169 (2000).
38. Devlin, J. J., Panganiban, L. C. & Devlin, P. E. Random peptide libraries: a source of specific protein binding molecules. *Science* **249**, 404–406 (1990).
39. Burgstaller, P., Girod, A. & Blind, M. Aptamers as tools for target prioritization and lead identification. *Drug Discov. Today* **7**, 1221–1228 (2002).
40. Braasch, D. A. & Corey, D. R. Novel antisense and peptide nucleic acid strategies for controlling gene expression. *Biochemistry* **41**, 4503–4510 (2002).

41. Petersen, M. & Wengel, J. LNA: a versatile tool for therapeutics and genomics. *Trends Biotechnol.* **21**, 74–81 (2003).
42. Goodchild, J. Hammerhead ribozymes for target validation. *Expert. Opin. Ther. Targets* **6**, 235–247 (2002).
43. Urnov, F. D. & Rebar, E. J. Designed transcription factors as tools for therapeutics and functional genomics. *Biochem. Pharmacol.* **64**, 919–923 (2002).
44. Dean, N. M. Functional genomics and target validation approaches using antisense oligonucleotide technology. *Curr. Opin. Biotechnol.* **12**, 622–625 (2001).
45. Shuey, D. J., McCallus, D. E. & Giordano, T. RNAi: gene-silencing in therapeutic intervention. *Drug Discov. Today* **7**, 1040–1046 (2002).
46. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
47. Yu, J. Y., DeRuiter, S. L. & Turner, D. L. RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells. *Proc. Natl Acad. Sci. USA* **99**, 6047–6052 (2002).
48. Rubinson, D. A. *et al.* A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nature Genet.* **33**, 401–406 (2003).
49. Shen, C., Buck, A. K., Liu, X., Winkler, M. & Reske, S. N. Gene silencing by adenovirus-delivered siRNA. *FEBS Lett.* **539**, 111–114 (2003).
50. Jackson, A. L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnol.* **21**, 635–637 (2003).
51. Davis, M. E. Non-viral gene delivery systems. *Curr. Opin. Biotechnol.* **13**, 128–131 (2002).
52. Merdan, T., Kopecek, J. & Kissel, T. Prospects for cationic polymers in gene and oligonucleotide therapy against cancer. *Adv. Drug Deliv. Rev.* **54**, 715–758 (2002).
53. Wadia, J. S. & Dowdy, S. F. Protein transduction technology. *Curr. Opin. Biotechnol.* **13**, 52–56 (2002).
54. Lindsay, M. A. Peptide-mediated cell delivery: application in protein target validation. *Curr. Opin. Pharmacol.* **2**, 587–594 (2002).
55. Richard, J. P. *et al.* Cell-penetrating peptides. A reevaluation of the mechanism of cellular uptake. *J. Biol. Chem.* **278**, 585–590 (2003).
56. Green, I., Christison, R., Voyce, C. J., Bundell, K. R. & Lindsay, M. A. Protein transduction domains: are they delivering? *Trends Pharmacol. Sci.* **24**, 213–215 (2003).
57. Abbas-Terki, T., Blanco-Bose, W., Deglon, N., Pralong, W. & Aebischer, P. Lentiviral-mediated RNA interference. *Hum. Gene Ther.* **13**, 2197–2201 (2002).
58. Barnett, B. G., Crews, C. J. & Douglas, J. T. Targeted adenoviral vectors. *Biochim. Biophys. Acta* **1575**, 1–14 (2002).
59. Quinonez, R. & Sutton, R. E. Lentiviral vectors for gene delivery into cells. *DNA Cell Biol.* **21**, 937–951 (2002).
60. Burton, E. A., Fink, D. J. & Glorioso, J. C. Gene delivery using herpes simplex virus vectors. *DNA Cell Biol.* **21**, 915–936 (2002).
61. Gehl, J. Electroporation: theory and methods, perspectives for drug delivery, gene therapy and research. *Acta Physiol. Scand.* **177**, 437–447 (2003).
62. Harris, S. Transgenic knockouts as part of high-throughput, evidence-based target selection and validation strategies. *Drug Discov. Today* **6**, 628–636 (2001).
63. Abuin, A., Holt, K. H., Platt, K. A., Sands, A. T. & Zambrowicz, B. P. Full-speed mammalian genetics: *in vivo* target validation in the drug discovery process. *Trends Biotechnol.* **20**, 36–42 (2002).
64. Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs — will they model the next 100? *Nature Rev. Drug Discov.* **2**, 38–51 (2003).
65. Tornell, J. & Snaith, M. Transgenic systems in drug discovery: from target identification to humanized mice. *Drug Discov. Today* **7**, 461–470 (2002).
66. Grimsditch, D. C. *et al.* C3H apoE(-/-) mice have less atherosclerosis than C57BL apoE(-/-) mice despite having a more atherogenic serum lipid profile. *Atherosclerosis* **151**, 389–397 (2000).
67. Elbashir, S. M., Harborth, J., Weber, K. & Tuschl, T. Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods* **26**, 199–213 (2002).
68. Holen, T., Amarzguioui, M., Wiiger, M. T., Babaie, E. & Prydz, H. Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. *Nucleic Acids Res.* **30**, 1757–1766 (2002).

Online links

DATABASES

The following terms in this article are linked online to:
Online Mendelian Inheritance in Man:
http://www.ncbi.nlm.nih.gov/Omim/
cystic fibrosis | Duchenne muscular dystrophy | Huntington's disease

OPINION

Quest for Best

Bruce Booth & Rodney Zemmel

By combining innovative science with years of massive investment, drug makers seek to turn newly discovered chemicals into revolutionary blockbuster drugs that generate billions of dollars in revenue. So every year, they collectively spend tens of billions of dollars on the high-risk pursuit of the next Prozac or Viagra. But should they? This article analyses key themes around differentiation that we have found to be common among blockbusters, and examines the implications for creating future billion-dollar drugs.

"Pioneering don't pay" was a favourite phrase of Andrew Carnegie, the renowned nineteenth-century industrialist. Perhaps surprisingly, what was true then of steel mills seems to also apply today to the pharmaceutical industry. Although conventional wisdom in the pharmaceutical industry had often attributed disproportionate value to being first in class, we and others have recently found that being 'best', rather than first, has historically created more value[1–6]. Much of the industry's past value creation has come not from first-in-class drugs against completely new targets, but from follow-on drugs that improve the efficacy or reduce the side effects of existing compounds. Most of the industry's blockbuster drugs have been developed as best-in-class clinical innovations, and only rarely were they 'discovered' as first-in-class agents.

First-in-class is not best-in-class

Commercial success is, of course, largely the result of the discovery and development of clinically relevant and innovative products.

"Much of the industry's past value creation has come not from first-in-class drugs against completely new targets, but from follow-on drugs that improve the efficacy or reduce the side effects of existing compounds"

However, we found that first-in-class agents with new mechanisms of action have not driven the vast majority of value creation in the pharmaceutical industry during the past decade — most blockbusters from this period have had an established mechanism of action (FIG. 1a). Out of the thirty two blockbusters launched by fifteen of the top pharmaceutical companies during 1991–2000, we found that 75% were directed against clinically validated pharmacological targets. Several mechanisms have been the target of multiple blockbusters, including selective serotonin re-uptake inhibitors, 3-hydroxy-3-methylglutaryl coenzyme A (HMG–CoA) reductase inhibitors, histamine H_1 receptor antagonists and proton pump inhibitors.

Furthermore, our analyses indicate that innovations around clinically validated mechanisms aimed at being best in class have created more value for the industry than their first-in-class counterparts (FIG. 1b). We characterized the drugs launched by fifteen of the top pharmaceutical companies during the period 1991–2000 by their 'mechanistic' status at the time they were entering development. For instance, drugs that entered development against new pharmacological targets (those that had not yet been shown to provide therapeutic benefits in the clinic) were termed 'novel' agents. Those that targeted clinically validated mechanisms that had yet to reach the market (those, for example, that were still in Phase III trials) were classified as 'fast followers.' Last, those that entered development after their mechanisms were already targeted by marketed drugs were considered as 'differentiators' or even 'latecomers'. Novel drugs typically included those launched within zero to two years of the first-in-class mechanism. Fast followers were two to five years late to market, differentiators were five to fifteen years late and latecomers were launched more than fifteen years after a drug with a given mechanism of action was first brought to market. These time lags to market are approx-