INNOVATION

The Connectivity Map: a new tool for biomedical research

Justin Lamb

Abstract | The ultimate objective of biomedical research is to connect human diseases with the genes that underlie them and drugs that treat them. But this remains a daunting task, and even the most inspired researchers still have to resort to laborious screens of genetic or chemical libraries. What if at least some parts of this screening process could be systematized and centralized? And hits found and hypotheses generated with something resembling an internet search engine? These are the questions the Connectivity Map project set out to answer.

With the benefit of hindsight, the great successes of biomedicine can be viewed as the simple act of connecting a disease with a disease-modifying gene product and a chemical modulator of that protein. From folklore (fever–aspirin–cyclooxygenase[1]) to rationality (chronic myeloid leukaemia–BCR-ABL–imatinib[2]) to design (hypertension–β-adrenergic receptors–propranolol[3]), and regardless of the order in which the dots are joined, the same principle holds. Whatever the specific research focus, populating this 'connectivity map' (FIG. 1) is the core business of biomedicine.

The problem is that it is very difficult to make these disease–gene–drug connections. The fundamental challenge is that clinical medicine, molecular genetics and chemistry are very different disciplines. Physicians talk about the clinical presentation of a patient, biologists about the phenotype of a fibroblast from a knockout mouse, and chemists about the binding of a small molecule to purified proteins. What is needed is to translate diseases, gene function and drug action into the same language. This is, we believe, now possible through the common vocabulary of genome-wide expression profiling. Once disparate biologies have been represented in this common global analytical space (FIG. 2), finding connections is a relatively simple numbers problem.

With this basic notion we set out to develop a generic solution for the identification of functional connections between drugs, genes and diseases with the aspiration of revolutionizing the approach to, and thereby accelerating the pace of, biomedical discovery. The results of our pilot study were published recently[4], together with two papers that show how this new approach can be applied to real-world problems in cancer research. In the first case, we were able to identify the mechanism of action of two poorly characterized natural products found to have a desirable inhibitory activity against androgen-receptor (AR) signalling[5]. And in the second, we found a drug capable of reversing glucocorticoid resistance in acute lymphoblastic leukaemia (ALL) cells[6].

Our solution is based on the creation of a large reference catalogue of gene-expression data from cultured human cells perturbed with many chemicals and genetic reagents (which we collectively refer to as 'perturbagens'). The general notion that a compendium of gene-expression profiles could be used for discovery was first proven by Hughes *et al*[7]. With a large collection of genome-wide expression data from yeast deletion mutants and drug-treated wild-type yeast cells, they showed that functions could be assigned to previously uncharacterized genes, and the targets of small molecules identified. But what sets the Connectivity Map resource apart from the gene-expression compendia that have come before (BOX 1) are the choices we made in an attempt to make it useful and approachable for the bench researcher.

Design principles

Four principles guided the Connectivity Map project. These are detailed below.

Centralize, systematize and compromise. An increasing appreciation of chemical genetics and the resulting desire to find bioactive 'tool' compounds to probe and annotate biologies of interest[8], together with the advent of RNA interference (RNAi) and the availability of genome-wide ablation libraries, has made the emergent theme in academic biology the same as it has been in industrial pharmaceutical discovery for decades: screening. The same libraries of chemicals and genetic constructs are screened over and over again, just with a different assay each time. The burden of screening ever-expanding chemical and genetic libraries against many individual readouts in serial and diffuse fashion is enormous, not just in terms of replication of the infrastructure and the perturbagen collections, but also the time invested by individual researchers on this extremely tedious yet highly demanding task. Therefore, a public reference catalogue of the global effects of all bioactive small molecules and genetic manipulations captured in the universal language of gene-expression profiling is now one of the most useful resources we can imagine.

However, gene-expression profiling remains an expensive pursuit, and is still highly susceptible to operator and facility-dependent idiosyncrasies. Linking together expression data from different centres, let alone different technologies, is frankly an unnecessary complication. The development of high-throughput high-density profiling platforms — such as the Affymetrix High Throughput Array (HTA), which enables automated amplification, labelling, hybridization and scanning of 96 samples in parallel a day — minimizes variability at the same time as providing substantial economies of scale. This technology, in a centralized production environment, therefore seems to be the best solution for the generation of the reference data. On the other hand, even this much firepower is insufficient to enable the analysis of every one of the estimated 200 different cell types exposed to every known perturbagen at every possible concentration for every possible duration. Compromises are therefore required. We selected a small number of established cell lines, chose a single exposure time and used one or occasionally a handful of doses in our pilot study[4]. As expected, being off by one log of concentration turned a strong signal into a barely detectable one, as had been shown before[9]. And, for example, one needs a cell that expresses peroxisome proliferator-activated receptor-γ (PPARγ) to see the effects of PPARγ agonists. But many effects were not unduly sensitive to cellular context or concentration, and even the bigger issue of the potential mismatch between *in vitro* responses and *in vivo* biology was not as pronounced as one might have imagined. So we are comfortable with the

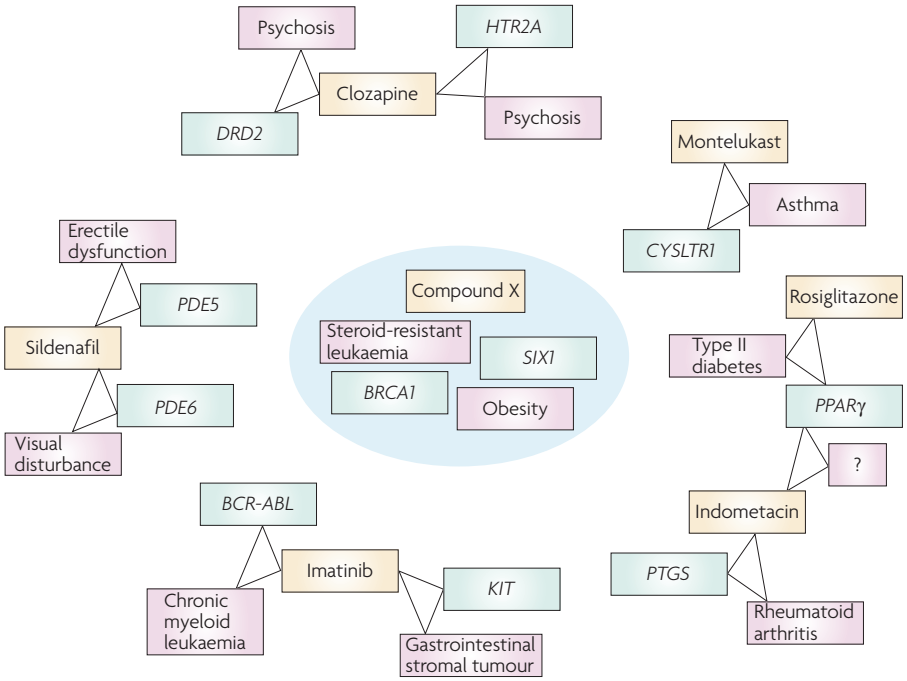

Figure 1 | **A connectivity map.** Functional relationships between a drug (yellow), a gene (green) and a disease (pink) constitute the nodes in this map. The connections between BCR-ABL, imatinib and chronic myeloid leukaemia (CML) provide the pristine example: the BCR-ABL fusion oncoprotein causes CML; imatinib inhibits BCR-ABL; imatinib thereby reverses CML[2]. However, most nodes feature disease-modifying rather than disease-causing proteins: cyclooxygenases (also known as prostaglandin-endoperoxide synthases; PTGS) do not cause rheumatoid arthritis, but indometacin, as an inhibitor of these enzymes, does provide an effective therapy for the condition[57]. Similarly, thiazolidinedione derivatives such as rosiglitazone treat type II diabetes through the activation of peroxisome proliferator-activated receptor-γ (PPARγ)[58], the LTD4 cysteinyl leukotriene receptor (CYSLTR1) antagonist montelukast is used as an asthma therapy[59], and the inhibition of phosphodiesterase 5 (PDE5) by sildenafil treats erectile dysfunction[60]. Connections between nodes are prevalent, reflecting wide-spread compound promiscuity. This can be within a protein class (imatinib also inhibits the KIT receptor tyrosine kinase and sildenafil inhibits PDE6), representing either therapeutic or off-targets and additional indications or side effects (gastrointestinal stromal tumour[61] and visual disturbance[62], respectively), or between entirely unrelated proteins (indometacin also binds and activates PPARγ[63]). Effects on many targets might also underlie a single indication: the antipsychotic effects of clozapine are mediated by affinity for the L5-hydroxytryptamine (serotinin) receptor 2A (HTR2A), dopamine receptor D_2 (DRD2) and probably many other receptors[33]. Note that this particular map is populated with connections discovered by traditional approaches, and therefore represents the concerted efforts of many thousands of researchers over many years. The aim of the Connectivity Map project is to provide a systematic solution for the discovery of the type of functional connections between drugs, genes and diseases shown here. Many 'orphan' diseases, genes and drugs still exist; a few of these are shown in the centre of this map.

compromises that were made, especially because the cost of 'screening' our data — literally a few minutes at the computer with no experimental overhead — makes finding a hit worth the price of many misses.

Empower the user. The problem with traditional gene-expression compendia (BOX 1) is that they are not much use if they do not contain the biological condition you are interested in. You could try to convince the curator to include your perturbagen of interest in his or her collection, but even then, it is unlikely that their production condition (such as cell line, treatment duration or concentration) is what you would have chosen. Furthermore, perhaps someone else has done the perfect experiment already or used a set of particularly rare clinical specimens, analysed the data and published only a list of differentially expressed genes. Or you might already have expression profiles from your own finely-tuned experiment. You would like the option to use these hard-won data. And although it is a universal language, there are many dialects (of content and technology) of gene-expression data.

We addressed these problems with a bipartite format of internal reference and external query data. Internal reference data are genome-wide expression profiles — made with our preferred technology and produced systematically — whereas external query data is provided by users in the form of a gene-expression signature of their biological state of interest (BOX 2). Queries can therefore be derived from any transcriptional profiling platform, because being composed of a relatively small number of genes with just a sign for their direction of change they can easily be mapped directly against the reference data. This enables the integration of legacy data and even experiments from non-human species. But more importantly, rather than being beholden to the curators of the reference collection, individual investigators can construct their queries remotely, exercising their specific biological expertise. Therefore, our approach divests responsibility for how the platform is ultimately used to the research community.

Actionable intelligence. With the obvious exception of the sequencing of the human and model organism genomes, genomics has generally not significantly affected the way bench research in biomedicine is conducted, and the bounty of new drug targets that the pharmaceutical industry thought could be found through these approaches have not materialized[10]. We have all been seduced by the undoubted power of gene-expression profiling. But the infatuation quickly fades when confronted with a long list of differentially-expressed genes, many of which are unfamiliar. Of course, the problem is that these — like the barcodes on groceries — are simply not interpretable by humans, and do not, therefore, help in the design of a real experiment. Our resource aims to be something like the scanner at the checkout, decoding lists of genes into the language of chemistry, and thereby providing hypotheses that are meaningful and actionable within the context of a traditional research project (FIG. 3). The two published Connectivity Map 'success stories', which exemplify two very different applications of our tool in cancer research[5,6], illustrate this point.

Motivated by the importance of AR signalling in prostate cancer, and specifically, the paucity of effective therapeutic approaches to hormone-refractory disease[11], Hieronymus *et al.*[5] used an innovative gene-expression-based high-throughput small-molecule screening

method (BOX 3) in an attempt to discover new inhibitors of AR-mediated transcription. However, having successfully identified celastrol and gedunin — two uncharacterized natural products — they were confronted by the all-too-common problem of figuring out the molecular basis for their activity. A gene-expression signature of the effects of these compounds was generated and used to query the Connectivity Map database. This analysis showed a strong connection between celastrol and gedunin and various well-known heat shock protein 90 (HSP90) inhibitors, immediately suggesting the hypothesis that these compounds were acting through this molecular chaperone. Subsequent wet laboratory experiments showed that this was indeed the case, defining a new family of AR modulators and a new structural solution to the targeting of HSP90 (REF. 5).

Wei *et al.*[6] were struggling to solve the serious clinical problem of resistance to the synthetic glucocorticoid dexamethasone in ALL. They had produced gene-expression profiles from a carefully annotated collection of 29 primary patient samples, representing both glucocorticoid sensitivity and resistance. But, as is often the case, the list of differentially-expressed genes was unenlightening. However, querying the Connectivity Map database with this gene-expression signature showed a strong connection between glucocorticoid sensitivity and the mammalian target of rapamycin (mTOR) inhibitor sirolimus, a US Food and Drug Administration (FDA)-approved immunosuppressant. This suggested that this compound might, therefore, reverse dexamethasone resistance. Indeed, testing this hypothesis showed that dexamethasone and sirolimus together decreased the viability of an otherwise dexamethasone-resistant cell line so well that a clinical trial of this new combination therapy in ALL now seems warranted[6]. This then represents an example of how our tool can begin to expand the connectivity map shown in FIG. 1.

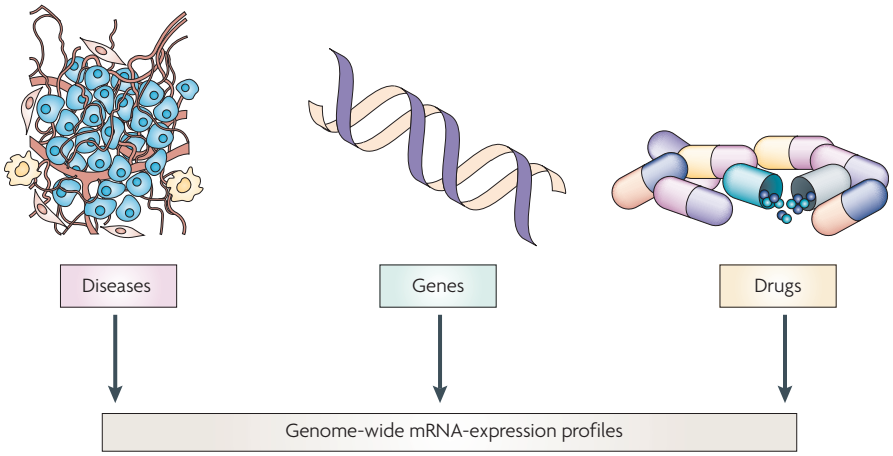

Figure 2 | **A universal functional bioassay.** As all of the transcripts are now known, and robust technologies for their simultaneous measurement are available, it is possible to capture objective high-dimensional depictions of all induced or organic biological conditions in a common global analytical space, and thereby readily appreciate similarities between them.

Box 1 | **Gene-expression compendia**

The first large-scale gene-expression compendium was generated by Hughes *et al.*[7], and contained data from 276 yeast deletion mutants, 11 conditional alleles and wild-type cells treated with 13 drugs (data available from Rosetta Inpharmatics). The Global Cancer Map[38] — populated by data from 218 human tumour samples, spanning 14 histological classes and 90 normal human tissues — was the original expression catalogue of organic states. This was followed by collections of expression profiles from large numbers of normal human and mouse tissues, perhaps most notably the Gene Expression Atlas[39,40]. In addition to catalogues of induced and organic states, a third type of large-scale expression database is exemplified by a collection of profiles from the livers of 111 mice from a cross of two inbred strains that has been used in several highly innovative integrative genomic analyses to identify key drivers of complex disease traits[41–43].

Several commercial gene-expression databases exist. Iconix's DrugMatrix contains data for the effects of systemic administration of hundreds of drugs on many rat tissues [44], and Gene Logic's BioExpress has thousands of profiles from normal, diseased and drug-treated human and animal tissues and cell lines. These seem to have use for the prediction of toxic liabilities and elucidation of the mechanisms of action of new chemical entities, and also for the repurposing of failed clinical candidates.

ArrayExpress[45,46] and Gene Expression Omnibus[47,48] (GEO) are not compendia in the strictest sense, as they contain data from many different experiments and technologies, and so are best described as repositories. The problem of unifying data of this sort into a super-compendium has not yet been solved. A notable exception to this, albeit in a focused sample space, is the Oncomine Cancer Profiling Database[49]. This contains thousands of individual gene-expression profiles from over 200 published cancer-expression datasets, but these data are pre-normalized, and complex meta-analyses across its entire content can be conducted using provided tools.

Accessibility. The availability of the data and data-mining tools, how easy they are to use and compatibility of the output with external resources are all issues of accessibility. The first is easily solved with a publicly-accessible website from where we serve all the Connectivity Map data, meta-data (such as perturbagen name, concentration, cell line and batch) and analytical tools through a simple, intuitive user interface (FIG. 4). Useability is a thornier issue. Flexibility is the guiding principle for most gene-expression analysis packages. For example, our local solution — GenePattern[12] —provides dozens of algorithms, each with lots of tunable parameters. However, for many researchers, this author included, there are often just too many choices. Our objective, therefore, is to keep all of the computation in the background and provide a single analysis tool with no 'knobs' whatsoever; queries are executed from our website in real-time with a single click. We are under no illusions that the current methods are necessarily the best or the only way to find connections, or even that finding connections is the only application for our data. Consequently, the raw data are also freely available for download by anyone with the inclination to explore alternatives. But our primary aim is to deliver testable hypotheses even to those without any special skills in gene-expression data analysis.

If a query executed at our website indicates that the effects of your favourite small molecule are connected with those of 'compound X', you would like to learn something more about it: what else is

it known to do? What is its chemical structure? Where can I buy some? There are many excellent public chemical annotation resources, and we do not attempt to replicate these, preferring to link out to them instead. Indeed, a fully integrated network of structure, activity, annotation and supplier databases — of which our resource would be just one node — should be an ultimate community objective. To perhaps start this ball rolling, the Connectivity Map has arranged for direct links to ChemBank[13,14] for structures and synonyms, and displays ATC (Anatomical Therapeutic Chemical) classifications — the hierarchical, human-readable alpha-numeric codes that are assigned to thousands of drug substances through a WHO (World Health Organization) initiative — for functional annotation. And to enable access to the actual chemical matter itself, vendor names and catalogue numbers are provided for all perturbagens in our dataset. However, an essential prerequisite for seamless cross-referencing between resources is standardized nomenclature. Many small molecules are known by dozens of different names (such as systematic, IUPAC (International Union of Pure and Applied Chemistry), trivial, USAN (United States Adopted Name), proprietary, and so on), often resulting in great confusion. The Connectivity Map has therefore exclusively adopted recommended international non-proprietary names (rINN), and we urge others to do the same. These names are provided as a unifying community resource by the WHO, and are chosen not only to be distinctive but also to convey information about function and similarity. For example, one can deduce that lumiracoxib is both a COX inhibitor and similar to celecoxib and rofecoxib from rINN alone, something that can not be done with their proprietary names (Prexige, Celebrex and Vioxx, respectively).

Box 2 | **Gene-expression signatures**

A gene-expression signature is, put simply, the list of genes differentially expressed in any biological distinction. However, the cardinal feature of a gene-expression signature is that it provides a rich representation of biology with a small number of genes, and this demands a more exacting definition: a list of genes rationally distilled from an unbiased global scan of gene-expression changes observed across a carefully selected sample set designed by an expert in the biology in question.

The notion of signatures came from early work on cancer classification, at a time when whether or not gene-expression profiles contained any useful information was still an open question. Golub *et al.*[50] profiled 38 bone-marrow samples — 27 from patients with acute lymphoblastic leukaemia (ALL) and 11 with acute myeloid leukaemia (AML), as defined by established clinical criteria (the carefully selected sample set) — and asked if these data could be used to predict the class of another 34 leukaemia samples. Using a scoring scheme based on the 50 genes most strongly statistically correlated with the ALL/AML distinction (the rational distillation) they showed that indeed they could, and the gene-expression signature was born.

Deriving signatures from perturbational experiments comes with the problems of selecting appropriate controls, doses and treatment durations. However, these have also been found to be useful. For example, we used a 21-gene signature derived from a cell line that ectopically expressed an oncogene to validate this *in vitro* model by showing enrichment of its signature in profiles from human cancer samples that naturally overexpressed this gene[51]. This paper also introduced a rank-based scoring metric — later refined into Gene Set Enrichment Analysis[52] — that enabled the use of signatures for which the component genes were flagged only as upregulated or downregulated rather than with the magnitude of their differential expression. These even simpler non-parametric gene sets are now commonly used. However, it is worth noting that one brand of commercial gene-expression signature (Drug Signatures) does use weighting factors[53].

Limitations and uncertainties

The Connectivity Map approach itself, as well as our current implementation of it, are not without limitations. Some of the shortcomings are obvious, some less so, and there are a few that might best be characterized as 'known unknowns'. Some discussion of the most important of these is warranted.

Cells that grow on plastic in a laboratory are very different from tissues in a whole organism. Our reliance on cell cultures means that effects modulated by specific microenvironments or that involve more than one cell type are simply inaccessible. This is perhaps the most fundamental limitation of our approach, as much of physiology and pharmacology is non-cell-autonomous. One example from the realm of cancer therapeutics is the aromatase inhibitors. Although they ultimately affect breast epithelial cells, they do so by blocking oestrogen synthesis in the stroma[15]. Another limitation is that the collection of cells we use does not express every protein, even in aggregate. So to continue the breast cancer theme, our methods would never appreciate the actions of the selective oestrogen-receptor modulators (SERMs) unless cells that express the oestrogen receptor (ER) were included in our panel. Further, if more subtle, the action of a pure ER antagonist (such as fulvestrant) could only be detected in those cells grown in the presence of the corresponding agonist. Indeed, we have shown exactly this[4]. What is not quite so obvious is the extent to which cells from one tissue differ from cells from another, or how established cell lines differ from primary cells, at least in their responses to perturbagens that affect ubiquitous cellular processes. The experience so far with therapeutically-relevant biologies of this type, such as histone modification, molecular chaperones and mTOR signalling, is that they are remarkably similar[4]. The downside of this is that gratuitous expansion of the dataset in the cell dimension is unlikely to efficiently deliver access to very much more biology. But it remains to be seen if judicious cell line choices will produce high concentrations of valuable idiosyncratic and nonredundant responses.

How best to interpret the result of a Connectivity Map query is still an open question. We have provided several graphical and statistical tools in our web interface to assist the user (such as, 'barview' visualizer, combined scores and permuted *P* values), but we do not yet have enough experience to know with certainty what constitutes a 'good' score, the number of instances of a particular perturbagen that are required to have confidence in a connection, or how statistical significance relates to biological significance. Part of the problem is that the answers to these questions are inextricably linked to the query itself. For example, a gene-expression signature produced from the treatment of cultured cells with a small molecule is different from one derived from a comparison of diseased and normal tissues. By definition, the first signature can be modulated in its entirety by the action of a chemical, whereas there is no reason to suppose *a priori* that the second example is druggable, either in whole or in part, not to mention the background noise that occurs in signatures from tissue specimens. On the other hand, the value

of finding a potential therapy for a human disease is much higher than the value of annotating an existing small molecule. The difficulty and expense involved in testing a particular hypothesis is also relevant. Consequently, the criteria for determining what constitutes a 'hit' might be different depending on the query. Unfortunately, only repeated use and more experiments can refine the process.

Perhaps the biggest shortcoming of the current Connectivity Map resource is the complexity of the core dataset. Both the number and diversity of the perturbagens in our collection is extremely low, meaning that the fraction of all possible induced cellular states represented is probably quite small. Fortunately, this is something we can address immediately. Our top priority is to aggressively pursue the expansion of our resource in the perturbagen dimension.

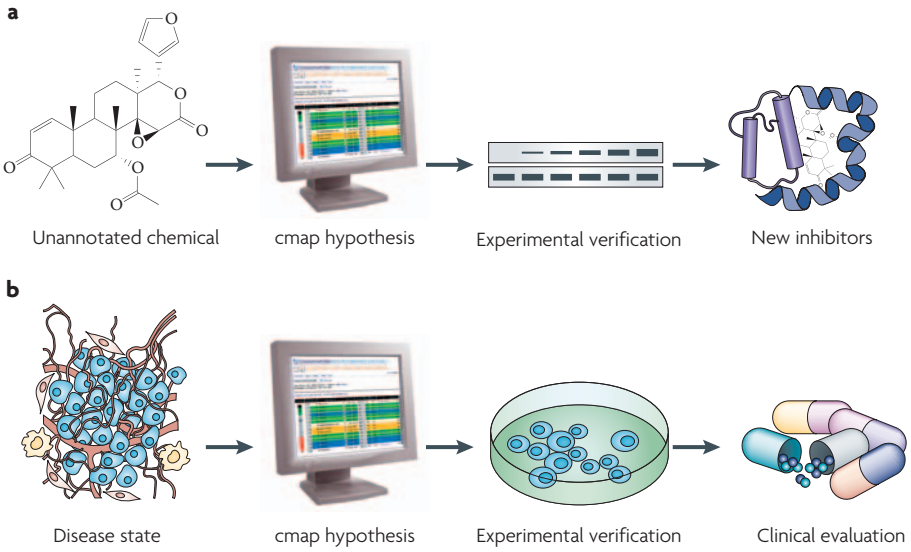

Figure 3 | **The Connectivity Map is a tool for the bench researcher.** The paths to showing that (**a**) an uncharacterized small molecule is a heat shock protein 90 (HSP90) inhibitor[5] and (**b**) sirolimus can reverse glucocorticoid resistance in acute lymphoblastic leukaemia[6] (see text for details) illustrate that the Connectivity Map (cmap) resource is best used in the context of a traditional research project. Indeed, its ultimate value relies on detailed experimental validation and follow-up.

Future content

Our pilot Connectivity Map resource[4] contains gene-expression profiles from a total of 453 individual treatments of 4 human cell lines with 164 discrete bioactive small molecules, that we call build 01. Emboldened by the apparent value of this small dataset, we are now generating profiles for the next public release. But there is a large gap between where we are now and the ultimate objective of a dataset approaching saturation of all chemical and genetic perturbagens, and prioritization (and constant evaluation of the benefits of continuing) is essential.

Our initial focus is the approximately 1,500 small-molecule drugs ever licensed for human use by the FDA. The proven safety and tolerability of these compounds greatly shortens the path to their clinical evaluation for any new indication found. This, taken together with the high frequency with which these agents are shown to have additional desirable activities (see for example REFS 16–18, as well as our own example with sirolimus[6] (see above)) explains why many academic programs[19] and even some commercial entities[20] seek to maximize the opportunities for so-called drug repurposing by screening exclusively in this space.

The Connectivity Map dataset will be extended to include genetic perturbagens. We will continue the focus on content with value for therapeutics discovery and use RNAi[21] to individually ablate each of the 600 of the approximately 3,000 human proteins that are predicted to bind drug-like small molecules[22,23] expressed in our panel of cell lines. Should one of these perturbagens be connected with a disease signature, there is a reasonable likelihood that a small-molecule lead compound either already exists or could be rationally designed. Of course, the equivalence of ablation with pharmacological modulation is still an open question, and a lead is far from a drug. But the protein products of the druggable genome[22] also constitute the entire universe of small-molecule off-targets. Connectivity Map queries with signatures of new potential drugs could therefore provide comprehensive selectivity assessments either as a strategy for flagging potential side effects or the identification of desirable polypharmacology profiles (FIG. 1).

Knowledge discovery is sometimes just as important as drug discovery. Bioactive small molecules that do not possess the exacting physicochemical characteristics (or toxicity profile) to imbue them with at least the potential to become medicines are nevertheless of great value as 'chemical tools' in the context of the Connectivity Map, as is shown by the connection we found between uncharacterized natural products and several HSP90 inhibitors (see above)[5], and we will continue to add these so-called non-drug bioactives to our dataset. However, it is probably impossible, and certainly impractical, to find a small molecule to modulate every protein encoded in the human genome[24]. Indeed, a recent global analysis of conventional

Box 3 | **Signature screening**

The concept of signature screening was introduced by Stegmaier *et al.*[54], who realized that if a gene-expression signature (BOX 2) really was the proxy for a phenotype of interest, it could be used to find small molecules that effect that phenotype without knowledge of a validated drug target. In the first example of what has become known as GE-HTS (gene-expression-based high-throughput screening) they derived a five-gene signature of leukaemia differentiation from a careful microarray analysis of primary cells from patients and unaffected individuals, and screened a library of 1,739 bioactive small molecules for compounds that induced the expression of the signature genes in a leukaemia cell line. Their results subsequently led these workers to propose a clinical trial of a new leukaemia therapy[55], impressively proving the principle. This success prompted the development of a technology capable of measuring the levels of up to 100 transcripts in thousands of samples at low cost and high throughput[56]. And this, in turn, has both enabled GE-HTS with larger signatures and larger libraries[5], and provided a practical solution for gene-expression signature-based clinical tests (Benjamin L. Ebert, personal communication). The Connectivity Map can be thought of as a resource for signature screening *in silico*.

screening data for nearly five million nonredundant structures collected over 25 years in both industry and academia found that the total number of human proteins for which chemical tools have been discovered so far is just 836 (REF. 25). This leaves 'biological tools' as the only practical way to gain access to all corners of biological space[24]. We will, therefore, exploit genome-wide open-reading frame (ORF)[26] and short hairpin RNA (shRNA)[27] collections to extend the Connectivity Map dataset to ectopic expression and ablation of all human genes. Of course, there are special challenges to producing high-quality gene-expression profiles with RNAi; including the absence of appropriate null controls, off-target effects[28] and the induction of the interferon response[29,30]. Our favoured solution is knockdown followed by wild-type reconstitution, and although more complicated, this also provides a generic approach to studying mutant alleles. Reconstitution with disease-associated, constitutively-active or interaction-defective (such as mutant proteins capable of physical interaction with only a subset of their usual binding partners)[31,32] alleles represent what we call 'sub-atomic' perturbagens, and using genetic tools of this sort we hope that the Connectivity Map will ultimately contain data for all of biological space at a granularity less than that of individual proteins.

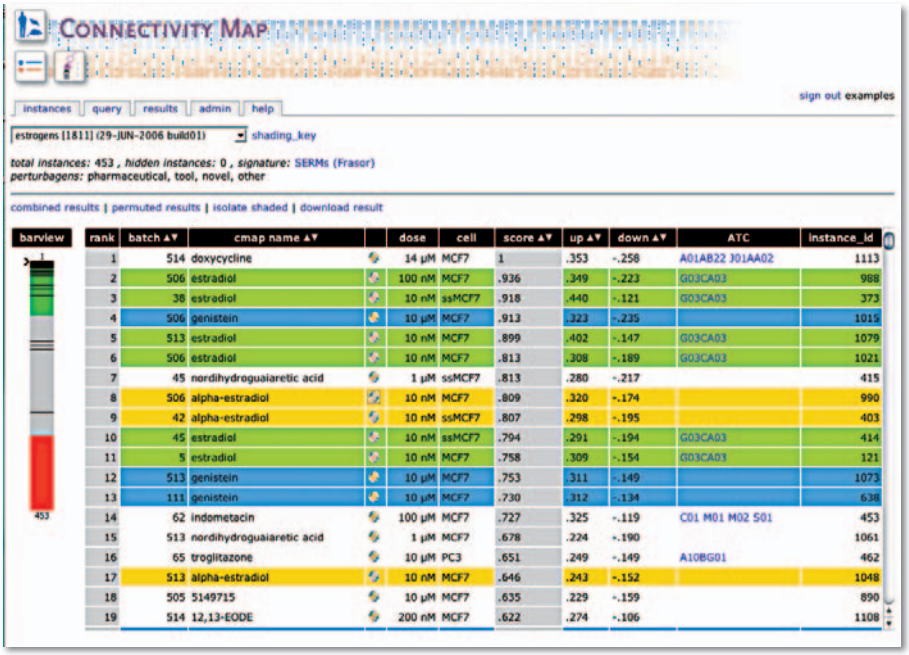

Figure 4 | **The Connectivity Map web interface.** A page from the Connectivity Map web site showing the results from a typical query.

Complexity simplified

Biological complexity and, in a desire to understand it, systems biology have become hot topics of late. Our closing remarks therefore highlight one reason why the complexity of drug action and human disease processes is now considered so important, and how the Connectivity Map offers a pragmatic 'complexity-ready' solution available today.

Despite a steady increase in the number of new targets with potent chemical matter reported each year, the rate at which drugs with a new target reach the market has remained much the same[22,25]. One sobering interpretation is that the prevailing model in which a complex disease can be reversed by modulating the activity of a single protein is now played out. More extreme still, given that more than a third of compounds in a representative collection of more than quarter of a million bioactive small molecules were found to have more than one target[25], is the idea that even the successful pharmaceuticals we already have are perhaps so because of unappreciated polypharmacology.

This is not so radical. For example, clozapine — arguably the most successful antipsychotic drug of all time — has high affinity for more than a dozen biogenic amine receptors, and the therapeutic relevance of this target promiscuity is shown by the failure of the agents developed to ligate only a subset of these proteins to deliver substantially better clinical performance[33]. Impinging on a single pathological process at many discrete points presumably provides a broader response profile and some multiplication of, or capacitance in, efficacy[34]. There is also a grander rendering of this notion: biology has evolved in such a way that a great many interdependencies are effectively hard wired. The consequence of this, despite the fact that there is an almost infinite universe of possible molecular circumstances, is that only relatively few of these are 'allowed'; all the others will rapidly revert to the nearest stable one. This is homeostasis at work. But clearly, there are many viable and stable disease states (such as cancer). Therefore, attempts to defeat these by breaking one thread within this web of interdependency with a single targeted intervention are probably futile. What is really needed is to switch the entire state to a more favourable one. And for this, a coherent set of effects on many targets is probably necessary. But however one chooses to look at it, the exploitation of promiscuity is an emerging theme for drug discovery[35].

The problem that arises then is that the conventional target-based biochemical assay is unable to appreciate the synergistic or even emergent effects of simultaneously affecting multiple proteins. A comprehensive network view of disease processes might ultimately enable the nodes, which when coordinately targeted will effect the desired outcome, to be predicted *in silico*[36], and structural biology and chemoinformatics might combine for the rational design of the corresponding selectively promiscuous small-molecule modulators[37], but both are still some way off. The immediate solution for polypharmacology drug discovery is, we believe, gene-expression-based screening approaches like the Connectivity Map. These are indifferent to the identity and number of proteins initially ligated by a small molecule (as they are to the precipitating lesions of a disease), being based instead on matching high-dimensional representations of the resultant biological states. These approaches are therefore ideally suited to the identification of the 'state-switching' perturbagens alluded to earlier, regardless of which of the myriad different possible combinatorial solutions they present. That being said, the number of compounds one would need to screen to find one is probably still very large. Therefore, the next challenge for the Connectivity Map is to increase throughput

and reduce unit cost to the extent necessary to enable the comprehensive profiling of true industrial-scale libraries of drug-like small molecules.

Justin Lamb is at the Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA.
e-mail: justin@broad.mit.edu

doi:10.1038/nrc2044

1. Vane, J. R. & Botting, R. M. The mechanism of action of aspirin. *Thromb. Res.* **110**, 255–258 (2003).
2. Druker, B. J. *et al*. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1031–1037 (2001).
3. Black, J. Drugs from emasculated hormones: the principle of syntopic antagonism. *Science* **245**, 486–493 (1989).
4. Lamb, J. *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
5. Hieronymus, H. *et al*. Gene expression signature-based chemical genomic prediction identifies novel class of HSP90 pathway modulators. *Cancer Cell* **10**, 321–330 (2006).
6. Wei, G. *et al*. Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL-1 and glucocorticoid resistance. *Cancer Cell* **10**, 331–342 (2006).
7. Hughes, T. R. *et al*. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
8. Austin, C. P. The completed human genome: implications for chemical biology. *Curr. Opin. Chem. Biol.* **7**, 511–515 (2003).
9. Cornwell, P. D., De Souza, A. T. & Ulrich, R. G. Profiling of hepatic gene expression in rats treated with fibric acid analogs. *Mutat. Res.* **549**, 131–145 (2004).
10. Lindsay, M. A. Target discovery. *Nature Rev. Drug Discov.* **2**, 831–838 (2003).
11. Feldman, B. J. & Feldman, D. The development of androgen-independent prostate cancer. *Nature Rev. Cancer* **1**, 34–45 (2001).
12. Reich, M. *et al*. GenePattern 2. 0. *Nature Genet.* **38**, 500–501 (2006).
13. Strausberg, R. L. & Schreiber, S. L. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **300**, 294–295 (2003).
14. Tolliday, N. *et al*. Small molecules, big players: the National Cancer Institute's initiative for chemical genetics. *Cancer Res.* **66**, 8935–8942 (2006).
15. Smith, I. E. & Dowsett, M. Aromatase inhibitors in breast cancer. *N. Engl. J. Med.* **348**, 2431–2442 (2003).
16. Rogers, J. T. *et al*. Alzheimer's disease drug discovery targeted to the APP mRNA 5′untranslated region. *J. Mol. Neurosci.* **19**, 77–82 (2002).
17. Stavrovskaya, I. G. *et al*. Clinically approved heterocyclics act on a mitochondrial target and reduce stroke-induced pathology. *J. Exp. Med.* **200**, 211–222 (2004).
18. Rothstein, J. D. *et al*. β-lactam antibiotics offer neuroprotection by increasing glutamate transporter expression. *Nature* **433**, 73–77 (2005).
19. Miller, T. M. & Cleveland, D. W. Treating neurodegenerative diseases with antibiotics. *Science* **307**, 361–362 (2005).
20. Borisy, A. A. *et al*. Systematic discovery of multicomponent therapeutics. *Proc. Natl Acad. Sci. USA* **100**, 7977–7982 (2003).
21. Elbashir, S. M. *et al*. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
22. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
23. Orth, A. P., Batalov, S., Perrone, M. & Chanda, S. K. The promise of genomics to identify novel therapeutic targets. *Expert Opin. Ther. Targets* **8**, 587–596 (2004).
24. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855–861 (2004).
25. Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S. & Hopkins, A. L. Global mapping of pharmacological space. *Nature Biotechnol.* **24**, 805–815 (2006).
26. Rual, J. F. *et al*. Human ORFeome version 1. 1: a platform for reverse proteomics. *Genome Res.* **14**, 2128–2135 (2004).
27. Root, D. E., Hacohen, N., Hahn, W. C., Lander, E. S. & Sabatini, D. M. Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nature Methods* **3**, 715–719 (2006).
28. Jackson, A. L. *et al*. Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnol.* **21**, 635–637 (2003).
29. Bridge, A. J., Pebernard, S., Ducraux, A., Nicoulaz, A. L. & Iggo, R. Induction of an interferon response by RNAi vectors in mammalian cells. *Nature Genet.* **34**, 263–264 (2003).
30. Sledz, C. A., Holko, M., de Veer, M. J., Silverman, R. H. & Williams, B. R. Activation of the interferon system by short-interfering RNAs. *Nature Cell Biol.* **5**, 834–839 (2003).
31. Vidal, M., Brachmann, R. K., Fattaey, A., Harlow, E. & Boeke, J. D. Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proc. Natl Acad. Sci.* **93**, 10315–10320 (1996).
32. Milstein, S. & Vidal, M. Perturbing interactions. *Nature Methods* **2**, 412–414 (2005).
33. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Rev. Drug Discov.* **3**, 353–359 (2004).
34. Keith, C. T., Borisy, A. A. & Stockwell, B. R. Multicomponent therapeutics for networked systems. *Nature Rev. Drug Discov.* **4**, 71–78 (2005).
35. Frantz, S. Drug discovery: playing dirty. *Nature* **437**, 942–943 (2005).
36. Rual, J. F. *et al*. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
37. Hopkins, A. L., Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127–136 (2006).
38. Ramaswamy, S. *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
39. Su, A. I. *et al*. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA* **99**, 4465–4470 (2002).
40. Su, A. I. *et al*. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067 (2004).
41. Schadt, E. E. *et al*. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
42. Mehrabian, M. *et al*. Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genet.* **37**, 1224–1233 (2005).
43. Schadt, E. E. *et al*. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37**, 710–717 (2005).
44. Ganter, B. *et al*. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.* **119**, 219–244 (2005).
45. Brazma, A. *et al*. ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
46. Parkinson, H. *et al*. ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **33**, D553–D555 (2005).
47. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
48. Barrett, T. *et al*. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res.* **33**, D562–D566 (2005).
49. Rhodes, D. R. *et al*. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1–6 (2004).
50. Golub, T. R. *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
51. Lamb, J. *et al*. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell* **114**, 323–334 (2003).
52. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
53. Natsoulis, G. *et al*. Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.* **15**, 724–736 (2005).
54. Stegmaier, K. *et al*. Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genet.* **36**, 257–263 (2004).
55. Stegmaier, K. *et al*. Gefitinib induces myeloid differentiation of acute myeloid leukemia. *Blood* **106**, 2841–2848 (2005).
56. Peck, D. *et al*. A method for high-throughput gene expression signature analysis. *Genome Biol.* **7**, R61 (2006).
57. Burke, A., Smyth, E. & FitzGerald, G. A. in *Goodman and Gilman's the pharmacological basis of therapeutics* (ed. Brunton, L. L.) 671–715 (McGraw-Hill, New York, 2006).
58. Lehmann, J. M. *et al*. An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPARγ). *J. Biol. Chem.* **270**, 12953–12956 (1995).
59. Reiss, T. F. *et al*. Effects of montelukast (MK-0476), a new potent cysteinyl leukotriene (LTD4) receptor antagonist, in patients with chronic asthma. *J. Allergy Clin. Immunol.* **98**, 528–534 (1996).
60. Boolell, M. *et al*. Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. *Int. J. Impot. Res.* **8**, 47–52 (1996).
61. Heinrich, M. C. *et al*. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *J. Clin. Oncol.* **21**, 4342–4349 (2003).
62. Marmor, M. F. & Kessler, R. Sildenafil (Viagra) and ophthalmology. *Surv. Opthamol.* **44**, 153–162 (1999).
63. Lehmann, J. M., Lenhard, J. M., Oliver, B. B., Ringold, G. M. & Kliewer, S. A. Peroxisome proliferator-activated receptors α and γ are activated by indomethacin and other non-steroidal anti-inflammatory drugs. *J. Biol. Chem.* **272**, 3406–3410 (1997).

Acknowledgements
Thanks to T. Golub, the Connectivity Map team and members of the Broad Institute Cancer and Chemical Biology Programs.

Competing interests statement
The author declares no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
AR | HSP90 | mTOR

FURTHER INFORMATION
ArrayExpress: http://www.ebi.ac.uk/arrayexpress/
BioExpress: http://www.genelogic.com
ChemBank: http://chembank.broad.harvard.edu/
Connectivity Map web site: http://www.broad.mit.edu/cmap
DrugMatrix: http://www.iconixpharm.com
Gene Expression Atlas: http://symatlas.gnf.org
Gene Expression Omnibus: http://www.ncbi.nlm.nih.gov/geo/
GenePattern web site: http://www.broad.mit.edu/cancer/software/genepattern/
Global Cancer Map: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Oncomine Cancer Profiling Database: http://www.oncomine.org
Rosetta Inpharmatics: http://www.rii.com/publications/2000/cell_hughes.html
WHO Collaborating Centre for Drug Statistics Methodology: http://www.whocc.no/atcddd/
WHO MedNet: http://mednet.who.int/
Access to this interactive links box is free online.