ELSEVIER



# Transcriptional data: a new gateway to drug repositioning?

**Francesco Iorio[1,4], Timothy Rittman[2,5], Hong Ge[3,5], Michael Menden[1] and Julio Saez-Rodriguez[1]**

[1] EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
[2] Dept of Clinical Neurosciences, Herchel Smith Building, Forvie Site, Addenbrooke's Hospital, Robinson Way, Cambridge CB2 0SZ, UK
[3] Dept of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, UK
[4] Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

**Recent advances in computational biology suggest that any perturbation to the transcriptional programme of the cell can be summarised by a proper 'signature': a set of genes combined with a pattern of expression. Therefore, it should be possible to generate proxies of clinicopathological phenotypes and drug effects through signatures acquired via DNA microarray technology.**

**Gene expression signatures have recently been assembled and compared through genome-wide metrics, unveiling unexpected drug–disease and drug–drug 'connections' by matching corresponding signatures. Consequently, novel applications for existing drugs have been predicted and experimentally validated.**

**Here, we describe related methods, case studies and resources while discussing challenges and benefits of exploiting existing repositories of microarray data that could serve as a search space for systematic drug repositioning.**

## Introduction

During past decades the main strategy of drug development has been high-throughput screening of different molecules to identify lead compounds showing activity against single therapeutic targets and pathways. However, the ratio of successfully identified drugs to screened molecules has decreased dramatically over the years [1]. Furthermore, targeting individual elements of pathogenic pathways is not always a successful approach for tackling the complexities of the disease state; even when a target pathway is identified, a suitable drug might not be found. For example, in Alzheimer's disease the 'amyloid hypothesis' has driven the search for drugs that stop aggregation of pathogenic beta-amyloid, which generates potentially toxic oligomers and plaques, but so far these efforts have not led to a successful disease-modifying treatment [2]. In addition, the cost of bringing an effective drug to the market is large and growing with a significant portion of investment needed in the research and development phase [3]. Many promising molecules never come into clinical use because they show unfavourable pharmacokinetic properties or cause adverse reactions in humans. As a consequence there is a pressing need to identify successful treatments for many diseases in innovative ways that could overcome these drawbacks.

Drug repositioning [4] is a potential alternative to new drug discovery that promises to address some of these issues by identifying new therapeutic applications for existing drugs. One of the advantages of reconsidering established drugs is that they have already been approved and, hence, they can potentially be re-marketed in a faster and more cost-efficient way – by skipping Phase I clinical trials [5]. Moreover, pharma company pipelines already include many drug candidates that have passed Phase I trials but were not successful in Phase II or III (i.e. being safe but not sufficiently effective in treating the condition they were originally designed for). This implies that the search basin for repositionable drugs is vast and much larger than the set of approved drugs [6].

*Corresponding author:*. Saez-Rodriguez, J. (saezrodriguez@ebi.ac.uk)
[5] These authors contributed equally to this work.

1359-6446/06/$ - see front matter  http://dx.doi.org/10.1016/j.drudis.2012.07.014

Please cite this article in press as: F.. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014

Most cases of successfully repositioned drugs can be linked to serendipity, such as the classic example of sildenafil which is used to treat erectile dysfunction but was originally developed as a cardiovascular drug [7]. However, systematic approaches have recently been proposed. Most of these are based on the principle that shared properties between compounds could hint at similar efficacy or commonality in their mode of action (MoA). Successful strategies based on this assumption have been devised and published in different areas of computational drug discovery: from chemoinformatics [8] and structural bioinformatics [9] to text mining and meta-data analyses [10] and, recently, genome-wide association studies [11]. Many of these strategies benefit from recent advances in data integration and systems biology [12] and among them a new trend has emerged over the past few years that is based solely on the analysis of gene expression data [13].

The traditional 'central dogma' of molecular biology is the principle of genes encoding mRNA that is translated into proteins. This defines a biological information flow that, moving through levels of increasing complexity and emerging properties, links the underlying genetic make-up of the cell to its clinicopathological state [13]. In such a context, transcriptional profiling enables the capture of a multidimensional view of this complexity at an intermediate level, reflecting genomic and environmental effects.

So far in computational drug discovery, drug response and disease phenotypes have been correlated with underlying pathological processes through 'back-tracking' approaches that can infer primary causes of transcriptional changes but require the integration of heterogeneous data sources and *a priori* known signalling and regulatory models [14–16]. Transcriptional profiles have also been used as a single data layer to dissect drug MoA through reverse-engineering techniques [17]. By contrast, recent studies suggest that purely data-driven approaches making use of gene expression data alone are well suited to identifying new drug repositioning opportunities. The leading idea is that comparing the expression profile of a cell before and after exposure can quantitatively assess the changes brought about by active compounds on the transcriptional programme. The corresponding signature of differential gene expression (SDE) can be considered as the summary of the compound's effect. Furthermore, a drug-induced SDE can then be compared with a disease-associated SDE similarly obtained through differential expression analysis of diseased versus healthy conditions. If they are sufficiently negatively correlated (i.e. the genes upregulated in the disease SDE are down-regulated in the drug SDE and vice versa) then it is reasonable to hypothesise that the effect of the drug on transcription is opposite to the effect of the disease (Fig. 1a). As a consequence, the drug might be able to revert the disease SDE and hence the disease phenotype itself [18–20]. Alternatively, from a shared SDE it can be hypothesised that two drugs could share a therapeutic application, regardless of the similarity in their chemical structure and that they impinge on different intracellular targets or pathways [21–24] (Fig. 1b).

Despite the relative simplicity of these ideas, recent applications have shown that they could serve as the basis for identifying drug repositioning opportunities in different therapeutic areas to treat heterogeneous diseases from cancer [25,26] to Alzheimer's disease [24] and Crohn's disease [27].

In the following sections we examine how gene transcription profiles have been analysed in single case studies and we will describe several publicly available resources; finally we discuss challenges and future directions.

## Matching gene expression signatures to 'connect' phenotypes

Pioneering studies have shown that collections of gene sets (i.e. groups of genes sharing a common biological function, chromosomal location or regulation) can be used to interpret and extract biological insights from genome-wide expression profiles, by using parametric [28] or non-parametric statistical methods [29].

A genetic signature is defined by associating a gene set with a specific pattern of expression [30]. Gene expression profiling has been widely used as phenotype proxies [31], to build phenotype taxonomies [30,32], for systematic functional discovery [33] and for classification and/or cataloguing purposes [30,34]. Most importantly, gene expression signatures have been effective in recovering 'connections' between genes, drugs and diseases involving (or involved in) the same biological process, by combining a large collection of gene expression data following drug treatment with a pattern-matching method [35]. A seminal example of this is given by the Connectivity Map (cMap) [18,35], which is the first large public database of genome-wide gene expression profiles from five different human cancer cell lines treated with more than 1000 bioactive small molecules.

The aim of the cMap project was to generate a 'map' that can be searched for 'connections' between gene expression profiles associated with disease states and those following treatment with a large set of existing drugs. To query this map, the authors devised a pattern-matching tool based on Gene Set Enrichment Analysis (GSEA) [29] through which these connections can be inferred and statistically assessed.

The effectiveness of this method for *in silico* drug discovery and drug repositioning has been demonstrated already by its very first applications [36,37], and it highlights the potential of gene transcription profiling to serve as the common language to link chemistry, biology and the clinic, by inferring genome-wide similarities or differences [35]. Numerous studies have been published using the cMap dataset and the cMap tool, with different aims (a comprehensive list is provided on the cMap website). This underscores the power of gene expression profiles and gene signatures in characterising biological states and acting as a surrogate phenotype, despite the difficulty in interpreting the meaning of predicted associations, let alone the precise part played by individual genes in these signatures [31]. Subsequent achievements have been to characterise the whole landscape of human gene expression [38], to establish large repositories of transcriptional data [39,40] and to make publically available a large amount of gene expression data that could be mined to compose drug and disease signatures (Fig. 2). Moreover, the robustness of these signatures has been shown across tissue types and experiments [41] and, during the past two years, the use of transcriptional data for drug repositioning has emerged as a useful and effective strategy [13,42], bringing about a new dawn for the vast quantities of DNA microarray data already in the public domain.

Although numerous approaches for *in silico* drug repositioning based on gene expression data have been published [19,20,22,24,


Please cite this article in press as: F.. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014

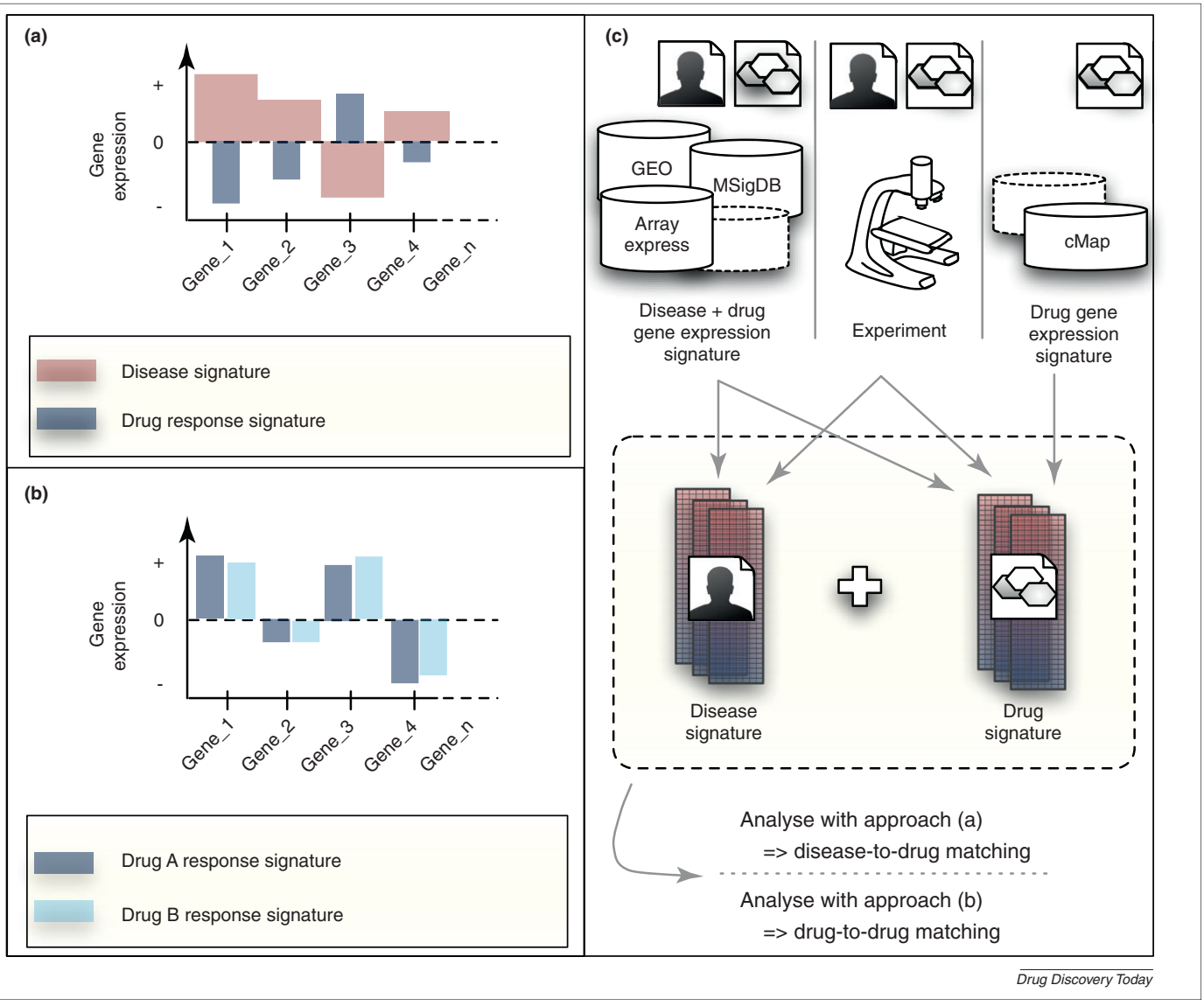


**FIGURE 1**

Signature reversion **(a)** and guilt-by-association **(b)** approaches in gene-expression-based drug repositioning. In (a) the aim is to identify a drug where the effect on transcription is opposite to a disease signature. In (b) drugs eliciting similar gene expression signatures are sought and hypothesised to share a common mode of action. Many publicly available repositories can be queried to generate drug and disease signatures that can be compared to each other and integrated with newly generated experimental data **(c)**.

25,27,36,37,43,44], all of them are methodologically similar and make use of the cMap SDEs as a reference database of drug responses in combination with signature-matching strategies. The majority of these methods can be subdivided into two main classes (features of which are summarised in Fig. 1). Methods in the first class aim to identify novel 'drug–disease' connection, whereas those in the second class aim to infer 'drug–drug' connections. In both cases gene expression profiles are used to summarise drug responses and disease states; and comparison between the two are based on the following simple but powerful assumptions:

(i) If an SDE summarising the response to a given approved drug is sufficiently negatively correlated to the SDE characterising a disease state, then that drug might be able to 'revert' the disease signature, hence the drug might be able to treat the disease phenotype. If already approved for other uses, the drug could be repositioned to treat that disease (Fig. 1a).

(ii) If two drugs elicit similar SDEs, even if acting on different intracellular targets, they could share a common MoA. In this case, the first drug could be repositioned to treat conditions for which the second drug has already been approved, or vice versa (Fig. 1b).

In the following section we will review case studies for methods in both classes.

### *Reverting phenotype signatures to revert phenotypes*

In this section we review methods based on the assumption that a drug that can revert a disease SDE might revert the disease phenotype itself. Building on this idea, several successful studies (methodologically similar to each other) identified new drugs for hepatocellular carcinoma [26], and were able to show the efficacy of vorinostat (currently used to treat cutaneous T-cell lymphoma) in treating gastric cancer [43] and also to predict


Please cite this article in press as: F.. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014

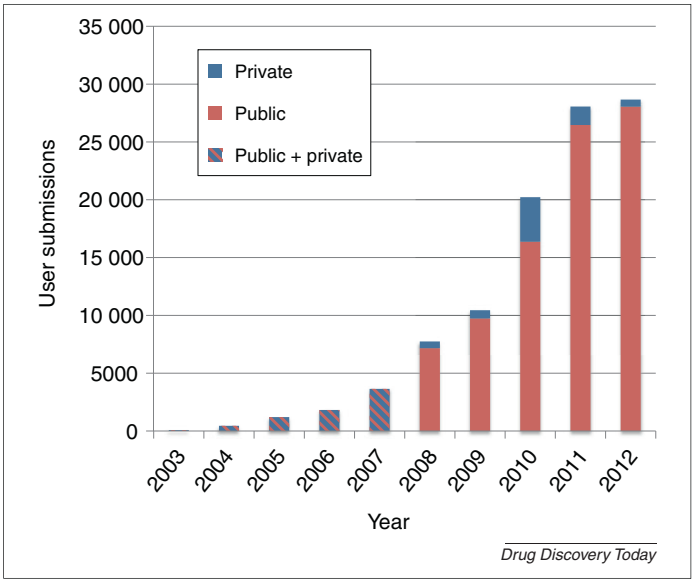


**FIGURE 2**

Rate of growth of ArrayExpress data in terms of experiments (i.e. user submission). This trend is set to increase further in the future, as new high-throughput sequencing-based transcriptomic applications result in the generation of huge amounts of data.

several candidate therapeutics for cancer in a systematic manner [20,25].

Owing to their robust experimental validation and the methodological similarity with other cited works, here we concentrate on two representative case studies. The first is presented in a publication by Dudley *et al.* [27]. As a first step, the authors assembled an SDE for inflammatory bowel disease (IBD), which is a chronic inflammatory gastrointestinal disorder for which only few safe and effective drugs exist, from public gene expression data [40]. Then they mined the cMap dataset to identify drugs that had SDEs that opposed those of IBD. They then developed an algorithm to generate a 'therapeutic score' for each of the drugs in the cMap, applying a significance threshold value to determine when a drug SDE was opposite to the disease SDE. Among the top-ranked therapeutic predictions, the authors found not only the corticosteroid prednisolone (for which the efficacy in treating IBD has already been established and therefore defined as a positive control) but also, interestingly, a second candidate topiramate, an anticonvulsant drug approved for epilepsy, never linked before to IBD, which had a predicted therapeutic score higher than that of prednisolone. The authors used a trinitrobenzenesulfonic (TNBS) acid induced rodent model of IBD to validate their prediction *in vivo*. They showed that topiramate treatment improved damage in colon tissue which was one of the most severe symptoms of the induced phenotype. They therefore suggested that, given its safety profile, topiramate could indeed be repositioned to treat IBD in humans.

In a similar study, Kunkel *et al.* [45] generated SDEs of skeletal muscle atrophy, a condition currently lacking pharmacological therapy, by chronic fasting in human patients and mouse models. They used the resulting two SDEs to search the cMap database and both their queries returned ursolic acid as the only compound with an SDE opposite to that of the disease state. The authors went on to verify experimentally that ursolic acid reduced muscle atrophy and stimulated muscle hypertrophy in mice. They identified the MoA to be enhancement of skeletal muscle insulin/insulin-like growth factor-1 (IGF-1) signalling and inhibition of atrophy-associated skeletal muscle mRNA expression. Moreover, they observed additional effects on the characteristics of muscle following treatment with ursolic acid, including reductions in adiposity, fasting blood glucose, plasma cholesterol and certain triglycerides. These findings suggest a potential use of ursolic acid in muscle atrophy and other metabolic myopathies. With respect to the study by Dudley *et al.* [27], the methodological difference is that here the authors used a partial SDE composed only of genes with significant differential expression in skeletal muscle atrophy, rather than using a genome-wide SDE. Moreover, the authors used the cMap query tool for matching these partial SDEs to compounds rather than designing their own therapeutic score.

In the remainder of this section we describe how the signature reversion strategy has been used successfully to predict synergistic drug combinations when matching drug SDEs with SDEs characterising biological states other than diseases, thus highlighting the generality of such a method.

Motivated by the aim of reducing drug resistance of a cancer in a pharmacological way, Wei *et al.* successfully identified rapamycin as a modulator of glucocorticoid resistance in acute lymphoblastic leukaemia [46]. As in the previous cases, the first step was to identify an SDE representing a biological state to be 'reverted' by a drug. However, rather than using an SDE derived from a generic disease state, the authors derived a gene expression signature that differentiated acute lymphoblastic leukaemia samples sensitive to glucocorticoids from glucocorticoid-resistant samples, hence generating a drug resistance SDE rather than the SDE of a disease. Furthermore, the authors searched the cMap dataset for drugs with an SDE matching the signature they computed in an opposite way, identifying several potential active compounds. The top ranked drug in this list was rapamycin. Further analysis found that rapamycin elicits a sensitising action to glucocorticoids by acting on the antiapoptotic factor MCL1 (induced myeloid leukaemia cell differentiation protein).

In a similar study, Hassane *et al.* identified drugs that enhanced the antileukaemic effect of partenolide, a drug effective at reducing the survival and leukemogenic activity of primary human acute myeloid leukaemia stem cells [47]. However, partenolide induces cellular protective responses that reduce its cytotoxicity. As the starting point the authors selected a previously published SDE of response to partenolide. With this signature they queried the cMap database and they identified compounds acting along the phosphatidylinositol-3-kinase and mammalian target of rapamycin (mTOR) pathways among those eliciting an SDE similar to that of partenolide. Finally, they verified that treating acute myeloid leukaemia cells with a combination of partenolide and phosphatidylinositol-3-kinase/mTOR inhibitors was more effective than treating with partenolide alone at decreasing the viability of cells and tumour burden *in vitro* and in murine xenotransplantation models.

Taken together, these studies clearly show the potential of 'signature reversion' in identifying new uses for existing drugs as well as to predict novel chemosensitising effect and synergistic drug combinations.


Please cite this article in press as: F.. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014

### *Mining similarity of transcriptional responses to drugs to identify a shared MoA*

In contrast to the examples in the previous section, here we review approaches based on the assumption that if two drugs elicit similar transcriptional responses then they could share a MoA and hence could be applied to the same pathological condition.

Inferring drug target binding by comparing the molecular similarity of sets of candidate drug compounds has been a traditional approach in drug discovery. This is ligand-based drug design and has been most often applied when structural information regarding the target proteins and their binding sites are absent. Candidate drug compounds known to inhibit the same target protein can be compared using their chemophoric (interaction) patterns. However, only when the 3D geometries of their interaction patterns match can a pharmacophore (the complementary set of binding interactions) be inferred representing the possible shared binding site of the target protein. Comparisons of interaction geometries can be seen in the literature reporting QSAR and comparative molecular field analysis (CoMFA) studies of two or more drug compounds that share a common target or equivalent binding sites in homologous proteins [48]. Conversely, a comparison of binding sites known to be targeted by one set of inhibitors and drugs can be used to infer equivalent binding sites in new targets. Consequentially, the target of a new drug can, in principle, be deduced by looking at the targets of the drugs most similar to it.

This 'guilt-by-association' principle has been successfully applied to identify new targets for existing drugs [8], by defining the corresponding set of ligands for a large number of known targets and then computing chemical similarities between drugs and ligand sets. In addition to structural similarity, the same principle has been applied to exploit other kinds of drug similarity in MoA discovery and repositioning in structural bioinformatics [9] where proteins with similar binding sites are targeted by the same drug; text mining [10], where two drugs sharing a semantic concept are assumed to share a therapeutic application; recently, 'modulatory profiling' [23], measuring changes in efficacy of lethal compounds when used in combination with a second cell-death-modulating agent (here drugs with similar modulatory profiles could have the same MoA); finally, as mentioned above, gene expression data, where two drugs elicit a similar SDE and could have a common MoA even if they act on different intracellular targets [22].

Based on the premise of shared genome-wide molecular activity, Iorio *et al.* [22] systematically compared all the cMap drugs in a pairwise fashion, rather than searching for drugs eliciting an SDE similar or opposite to an input signature. By doing this they identified a large number of drug–drug 'associations' based on the extent of similarity between the corresponding SDEs. By making use of a novel similarity score, they constructed a network representation in which each node is a drug and each edge (connection) indicates a significant similarity between the SDEs of the connected nodes. They divided the drug network into groups of densely interconnected nodes termed 'communities', containing drugs eliciting similar SDEs. Communities were strikingly populated by drugs with similar known MoAs or sharing a therapeutic application.

The authors demonstrated the power of their method to identify the MoA of novel drugs by analysing their neighbouring communities once they were integrated in the network. In a similar way, they showed how the drug network could be used to infer new applications for already existing drugs by searching subnetworks surrounding a drug with a desired MoA for other compounds never linked before to that MoA. By doing this, they were able to predict and experimentally verify that fasudil, a safe Rho-kinase inhibitor approved in Japan to reverse blood vessel obstructions after ischemic stroke, can enhance cellular autophagy [21], a metabolic process implicated in several neurodegenerative disorders.

A related method was proposed by Hu and Agarwal [24], who inferred a drug–disease network in which two nodes were connected by an edge if the corresponding SDEs were significantly similar (in the case of drug–drug connections) or significantly negatively correlated (in the case of drug–disease connections). To achieve this, the authors integrated the SDE of the cMap drugs with a large number of disease SDEs assembled by mining the Gene Expression Omnibus (GEO) repository [40]. Connections representing anticorrelations were predictive of new indications for existing drugs, such as the potential use of some antimalarial drugs for Crohn's disease, and the possible repositioning of several existing drugs as therapeutic options for Huntington's disease. This approach can be seen as a precursor hybrid method, mixing together the two types of approaches of disease signature reversion and guilt-by-association. Moreover, the authors hypothesise that drug side effects could be predicted by the analysis of similarity between drug and disease SDEs.

In conclusion, the results presented show how large collections of gene expression data following drug treatment could be exploited through a guilt-by-association approach with the aim of identifying drug repositioning opportunities.

## Resources for computational expression-based drug repositioning

Several resources support computational drug repositioning based on transcriptional data and the functional characterisation of gene sets and signatures. Some freely available tools and database are listed in Table 1.

ArrayExpress [39], GEO [40] and the cMap [18,35] are large public repositories of gene expression data from where disease and drug-response signatures can be assembled. DAVID [49], MsigDB [29] and GeneSigDB [50] are useful tools for functionally characterising large gene lists by using pre-defined functional terms, or pre-defined gene signatures representing different biological entities and processes from public repositories. These signatures can also be used to characterise with regard to function large sets of differentially expressed genes from microarray studies through non-parametric statistical methods that can also provide complementary information, such as the GSEA tool [29] or Expression Analysis System Explorer (EASE) and regulatory motif analysis [51,52].

The cMap query tool has two extensions: sscMap [53,54] and the MoA by network analysis (MANTRA) tool [22]. sscMAP is a free-to-download java implementation of the cMap algorithm bundled with the reference dataset, enabling the integration of user-defined data. MANTRA makes use of a post-processed version of the cMap dataset, where compounds are catalogued into a drug similarity network. In this network two drugs are connected if they elicit a similar transcriptional response in human cell lines. With MANTRA users can integrate a drug under investigation into the network and deduct its MoA by analysing the surrounding subnetwork. Moreover, it is possible to identify drug repositioning opportunities by


Please cite this article in press as: F.. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014

Please cite this article in press as: F. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014

**TABLE 1**

**Publicly available resources to derive, compare and functionally characterise gene expression signatures**

| *Resource* | Short description | Minable for partial and genome-wide signatures of drug responses and disease states | Tool for signature matching and classification of microarray data | Functional characterisation of gene sets/ signatures | Oriented to drug-discovery and repositioning | Website |
|---|---|---|---|---|---|---|
| **ArrayExpress Gene Expression Omnibus – GEO** | Public repositories of gene expression data | ✓ | | | | http://www.ebi.ac.uk/arrayexpress/ http://www.ncbi.nlm.nih.gov/geo/ |
| **Database for Annotation, Visualization and Integrated Discovery – DAVID** | Functional annotation tools to associate biological meaning to list of genes through analysis of over-represented terms | | | ✓ | | http://david.abcc.ncifcrf.gov/ |
| **Gene Expression Atlas** | Subset of ArrayExpress archive, servicing queries for condition-specific gene expression patterns | ✓ | ✓ | | | http://www.ebi.ac.uk/gxa/ |
| **Molecular Signature Database – MsigDB Gene Signature Database – GeneSigDB** | Collections of annotated gene signatures from different sources | ✓ | | ✓ | | http://www.broadinstitute.org/gsea/msigdb http://compbio.dfci.harvard.edu/genesigdb/ |
| **Gene Set Enrichment Analysis – GSEA** | Tool able to determine if an *a priori* defined gene signature shows statistically significant, concordant differences between two biological states | | ✓ | ✓ | | http://www.broadinstitute.org/gsea |
| **ProfileChaser MicroArray Rank Query – MarQ** | Tools to search the GEO repository for experiments whose differential expression looks similar or opposite to a gene expression signature or a query experiment | ✓ | ✓ | | | http://profilechaser.stanford.edu/ http://marq.dacya.ucm.es/ |
| **Connectivity Map – cMap** | Large collection of gene expression data following drug treatment that can be queried with an integrated pattern-matching tool, based on GSEA, to find drugs eliciting a response similar or opposite to a given gene signature | ✓ | ✓ | | ✓ | http://www.broadinstitute.org/cmap/ |
| **Statistically significant connections' map – sscMap** | Java implementation of the cMap tool bundled with the corresponding dataset and making it extendable with adding custom collections of reference profiles | | ✓ | ✓ | | http://purl.oclc.org/NET/sscMap |
| **Mode of Action by NeTwoRk Analysis – MANTRA** | Tool for the analysis of the mode of action of novel drugs and the identification of drug repositioning opportunities, based on network theory and GSEA and making use of a post-processed version of the cMap database | ✓ | ✓ | | ✓ | http://mantra.tigem.it |
| **Drug versus Disease – DvD** | Computational pipeline for comparing disease and drug-response gene expression signatures from publicly available resources | ✓ | ✓ | | ✓ | www.ebi.ac.uk/saezrodriguez/dvd |

searching the neighbourhood of a 'seed' compound with a desired MoA for 'safe' compounds never linked before to that MoA.

Several tools are freely available for mining gene expression data repositories based on similarity with an input signature in a similar manner to cMap. ProfileChaser [55] searches microarray repositories based on genome-wide patterns of differential expression, and MARQ [56] mines GEO for experiments that generate a differential expression profile that is similar or anti-correlated to an input gene expression signature. Finally, DvD is a recently developed tool providing a pipeline for the comparison of drug and disease gene expression profiles from public microarray repositories.

## Challenges of signature-matching methods

A potential major problem affecting the methods described here is the challenge of integrating independent microarray studies. Microarrays do not measure gene expression in absolute units. As a consequence, an improper handling of multiple gene expression profiles obtained in different experimental settings would capture similarities in these settings rather than in the represented biological states (a phenomenon known as the 'batch effect' [57]). By contrast, cells in different pathological conditions or with different genomic backgrounds respond very differently to the same drug treatment. Consequently, classic microarray analysis approaches might not produce optimal results, because they tend to discriminate gene expression profiles on the basis of the experimental settings in which they have been produced rather than on the basis of the stimuli they are responding to (for example a drug treatment).

In most of the methods we describe in this review, these problems are partially addressed by making use of non-parametric statistics [29], genome-wide ranked lists of genes [18] and 'consensual responses' to drugs [22] rather than classic similarity metrics applied to individual profiles of expression values or fold-change-derived significance scores. However, a potential drawback of these techniques is that they might dilute cell-specific 'gene expression signals' by pooling together the transcriptional response to the same drug but from different experimental settings (i.e. different cell lines, dosages or observation times). These problems have been tackled by designing *ad hoc* similarity scores [58] and genome-wide metrics [44,59,60].

RNA-seq technology might overcome many of these limitations, because it can detect amounts of RNA over a wider dynamic range. In the long run, RNA-seq could replace microarrays for SDE analysis; meanwhile use of microarray data remains attractive, being not only a simpler and more cost-effective technology but also one with a vast collection of already publicly available data.

## Concluding remarks

We have reviewed approaches using microarray data to assist in the elucidation of compound MoA with the specific goal of identifying new potential applications for existing drugs. A significant number of published results show that microarray technology provides a unique opportunity to identify repositionable drugs by exploiting the vast amount of existing publicly available data where the potential has not yet been fully capitalised.

The methods we described do not consider mechanistic aspects, but simply use transcriptional signatures as readouts from the 'black-box' of cellular mechanisms. Therefore, they cannot provide any information about cell signalling pathways where a deregulation can result in an observed expression signature.

It could be argued that as long as the drug works the mechanism is a secondary consideration. But at the same time, it is reasonable to expect that additional insight into a MoA for a given drug can be obtained by integrating expression data with knowledge of (and ideally data from) the systems in which the drugs operate, known regulatory relationships between genes and signalling pathway maps.

The challenge for the future will be to take current analyses to a higher level, integrating signatures and mechanistic insights inferred by other recently developed approaches. This will require repositories of comprehensive gene expression data for disease states and compound effects, and integration with prior knowledge of cellular networks on which drugs operate, and further development of computational methods to translate this data into effective medicines.

So far, recent results encouragingly illustrate that computational approaches using public gene expression microarray data can be successfully employed to infer new potential drug therapies. We argue that this can (and probably will) be further exploited in the near future.

## Acknowledgements

TR and HG started writing this review, based on a lecture given by FI, as an assignment for the graduate course: Reviews in Computational Biology, at the Cambridge Computational Biology Institute, organised by Christoph Dessimoz and James Smith – whom we thank for their helpful comments.
HG is funded by the Wellcome Trust.
TR is funded by the Medical Research Council and a Raymond and Beverly Sackler Scholarship.
MM is an EMBL internal pre-doctoral fellow.
FI is a fellow of the joint EMBL – EBI & Wellcome Trust Sanger Institute post-doctoral (ESPOD) programme.
We thank Gabriella Rustici and Ibrahim Emam for kindly providing Fig. 2 and data describing ArrayExpress. We thank Clare Pacini, Mathew Garnett, Silvano Squizzato, Pedro Ballester, Thomas Klabunde, Sonja Eidorn, Constantine Alifrangis, Aidan Macnamara, Niccolò Bolli, Annalisa Buniello and Annalisa Mupo for kindly reading the manuscript and providing helpful comments, and John P. Overington for his suggestions.

## References

1 Booth, B. and Zemmel, R. (2004) Prospects for productivity. *Nat. Rev. Drug Discov.* 3, 451–456
2 Mangialasche, F. *et al.* (2010) Alzheimer's disease: clinical trials and drug development. *Lancet Neurol.* 9, 702–716
3 Paul, S.M. *et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* 9, 203–214
4 Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673–683


Please cite this article in press as: F.. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014

5 Chong, C.R. and Sullivan, D.J. (2007) New uses for old drugs. *Nature* 448, 645–646
6 Tartaglia, L.A. (2006) Complementary new approaches enable repositioning of failed drug candidates. *Expert Opin. Investig. Drugs* 15, 1295–1298
7 Renaud, R.C. and Xuereb, H. (2002) From the analyst's couch: erectile-dysfunction therapies. *Nat. Rev. Drug Discov.* 1, 663–664
8 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181
9 Haupt, V.J. and Schroeder, M. (2011) Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief. Bioinform.* 12, 312–326
10 Andronis, C. *et al.* (2011) Literature mining, ontologies and information visualization for drug repurposing. *Brief. Bioinform.* 12, 357–368
11 Sanseau, P. *et al.* (2012) Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* 30, 317–320
12 Iskar, M. *et al.* (2012) Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr. Opin. Biotechnol.* 23, 609–616
13 Lussier, Y.A. and Chen, J.L. (2011) The emergence of genome-based drug repositioning. *Sci. Transl. Med.* 3, 96ps35
14 Chindelevitch, L. *et al.* (2011) Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics* 28, 1114–1121
15 Kasarskis, A. *et al.* (2011) Integrative genomics strategies to elucidate the complexity of drug response. *Pharmacogenomics* 12, 1695–1715
16 Pham, L. *et al.* (2011) Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proc. Natl. Acad. Sci. U. S. A.* 108, 13347–13352
17 di Bernardo, D. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered networks. *Nat. Biotechnol.* 23, 377–383
18 Lamb, J. (2007) The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer* 7, 54–60
19 Sirota, M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3, 96ra77
20 Mcart, D.G. and Zhang, S-D. (2011) Identification of candidate small-molecule therapeutics to cancer by gene-signature perturbation in connectivity mapping. *PLoS ONE* 6, e16382
21 Iorio, F. *et al.* (2010) Identification of small molecules enhancing autophagic function from drug network analysis. *Autophagy* 6, 1204–1205
22 Iorio, F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14621–14626
23 Wolpaw, A.J. *et al.* (2011) Modulatory profiling identifies mechanisms of small molecule-induced cell death. *Proc. Natl. Acad. Sci. U. S. A.* 108, E771–E780
24 Hu, G. and Agarwal, P. (2009) Human disease-drug network based on genomic expression profiles. *PLoS ONE* 4, e6536
25 Shigemizu, D. *et al.* (2012) Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer. *PLoS Comput. Biol.* 8, e1002347
26 Chen, M-H. *et al.* (2011) Gene expression-based chemical genomics identifies potential therapeutic drugs in hepatocellular carcinoma. *PLoS ONE* 6, e27186
27 Dudley, J.T. *et al.* (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Trans. Med.* 3, 96ra76
28 Khatri, P. *et al.* (2002) Profiling gene expression using Onto-Express. *Genomics* 79, 266–270
29 Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550
30 Califano, A. *et al.* (2000) Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8, 75–85
31 Nevins, J.R. and Potti, A. (2007) Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat. Rev. Genet.* 8, 601–609
32 Rhodes, D.R. and Chinnaiyan, A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37, S31–S37
33 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
34 Ramaswamy, S. *et al.* (2002) A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* 33, 49–54
35 Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935
36 Hieronymus, H. *et al.* (2006) Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* 10, 321–330
37 Hassane, D.C. *et al.* (2008) Discovery of agents that eradicate leukemia stem cells using an *in silico* screen of public gene expression data. *Blood* 111, 5654–5662
38 Lukk, M. *et al.* (2010) A global map of human gene expression. *Nat. Biotechnol.* 28, 322–324
39 Parkinson, H. *et al.* (2011) ArrayExpress update – an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39, D1002–D1004
40 Barrett, T. (2004) NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Res.* 33, D562–D566
41 Dudley, J.T. *et al.* (2009) Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.* 5, 1–8
42 Harrison, C. (2011) Translational genetics: signatures for drug repositioning. *Nat. Rev. Genet.* 12, 668–669
43 Claerhout, S. *et al.* (2011) Gene expression signature analysis identifies vorinostat as a candidate therapy for gastric cancer. *PLoS ONE* 6, e24662
44 Jin, G. *et al.* (2011) A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy. *Cancer Res.* 72, 33–34
45 Kunkel, S.D. *et al.* (2011) mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell Metab.* 13, 627–638
46 Wei, G. *et al.* (2006) Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* 10, 331–342
47 Hassane, D.C. *et al.* (2010) Chemical genomic screening reveals synergism between parthenolide and inhibitors of the PI-3 kinase and mTOR pathways. *Blood* 116, 5983–5990
48 Dean, P.M. and Lewis, R.A., eds) (1999) *Molecular Diversity in Drug Design*, Kluwer Academic Publishers
49 Huang, D.W. *et al.* (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57
50 Culhane, A.C. *et al.* (2009) GeneSigDB – a curated database of gene expression signatures. *Nucleic Acids Res.* 38, D716–D725
51 Hosack, D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4, R70
52 Liu, Y. and Ringnér, M. (2007) Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis. *Genome Biol.* 8, R77
53 Zhang, S. and Gant, T. (2008) A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics* 9, 258
54 Zhang, S. and Gant, T. (2009) sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics* 10, 236
55 Engreitz, J.M. *et al.* (2011) ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics* 27, 3317–3318
56 Vazquez, M. *et al.* (2010) MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures. *Nucleic Acids Res.* 38, W228–W232
57 Lander, E.S. (1999) Array of hope. *Nat. Genet.* 21, 3–4
58 Iskar, M. *et al.* (2010) Drug-induced regulation of target expression. *PLoS Comput. Biol.* 6, 1929–1935
59 Plaisier, S.B. *et al.* (2010) Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res.* 38, e169
60 Licamele, L. and Getoor, L. (2010) Indirect two-sided relative ranking: a robust similarity measure for gene expression data. *BMC Bioinformatics* 11, 137


Please cite this article in press as: F.. Iorio, et al., Transcriptional data: a new gateway to drug repositioning?, Drug Discov Today (2012), http://dx.doi.org/10.1016/j.drudis.2012.07.014