Encircling the regions of the pharmacogenomic landscape that determine drug response

Adrià Fernández-Torras[1], Miquel Duran-Frigola[1,*] and Patrick Aloy[1,2,*]

1. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

* Corresponding authors: miquel.duran@irbbarcelona.org; patrick.aloy@irbbarcelona.org

Abstract

The integration of large-scale drug sensitivity screens and genome-wide experiments is changing the field of pharmacogenomics, revealing molecular determinants of drug response without the need for *a priori*, hypothesis-driven assumptions about drug action. In particular, transcriptomic signatures of drug sensitivity may guide drug repositioning, the discovery of synergistic drug combinations and suggest new therapeutic biomarkers. However, the inherent complexity of transcriptomic signatures, with thousands of genes differentially expressed, makes them hard to interpret, giving poor mechanistic insights and hampering translation to the clinics. Here we show how network biology can help simplify transcriptomic drug signatures, filtering out irrelevant genes, accounting for tissue-specific biases and ultimately yielding functionally-coherent, less noisy drug *modules*. We successfully analyzed 170 drugs tested in 637 cancer cell lines, proving a broad applicability of our approach and evincing an intimate relationship between modules' gene expression levels and drugs' mechanisms of action. Further, we have characterized multiple aspects of our transcriptomic modules. As a result, the drugs included in this study are now annotated well beyond the reductionist (target-centered) view.

Author Summary

Large scale pharmacogenomics studies performed with hundreds of cell lines offer a means to link the molecular features of the cells to their response to drug treatments. Unfortunately, simple drug-gene correlations are usually not enough to consistently identify what gene expression patterns will determine drug sensitivity, as the tissue of origin of the cells, together with the expression of e.g. membrane transporters can greatly confound the analysis. To ameliorate these biases, we have devised a network-based strategy that selects genes that are both well correlated to drug response and closeby in the human protein interaction network. Reassuringly, we have confirmed that our identified drug sensitivity *modules* are tightly connected to the mechanisms of action of the drugs. Moreover, while our modules have no more than 100 genes, they retain the predictive power of the much larger gene *signatures* that are typically obtained by drug-gene correlations alone. Here, we release the characteristic modules for almost 200 drugs, in a format that is suitable for downstream bioinformatics analyses such as gene-set enrichment analysis.

Introduction

Gene expression profiling has become a mainstay approach to characterize cell properties and status, unveiling links between gene activities and disease phenotypes. Early efforts were put into discovering transcriptional signatures that are specific of a disease state, involving the comparison of a relatively small number of disease and healthy samples [1]. Even though these studies have provided a rich account of disease biology, they have failed to yield better drug therapies, as causality and response to perturbations cannot be directly inferred from two-state (disease-vs-healthy) differential gene expression analysis [2, 3]. To address this issue, initiatives have flourished that profile only the basal gene expression levels of hundreds of cell lines, treat them with drug molecules and then measure a simple readout such as growth rate [4-7]. Provided that the panel of cell lines is large enough, this allows for a new type of gene expression analysis where basal gene expression levels are correlated to drug response phenotypes, paving the way for target identification, biomarker discovery, and elucidation of mechanisms of action and resistance [8-13].

The largest cell panels available today are derived from cancerous tissues, since a crucial step towards personalized cancer medicine is the identification of transcriptional signatures that can aid drug prescription. However, these signatures are composed of several hundreds of genes, making them difficult to interpret and translate to the clinics. Recent assessment of sensitivity signatures for over 200 drugs [9] revealed that among key genes are those involved in drug metabolism and transport, with intended therapeutic targets only detected in a fraction of signatures, and identifying the cell line tissue of origin as an important confounding factor. In practice, the length of the signatures largely exceeds the number of sensitive cell lines available for each drug, which often yields inconsistent results between cell panels from different laboratories [14]. The current challenge is to filter and characterize transcriptional signatures so that they become robust, informative and more homogeneous while still retaining the complexity (hence the predictive power) of the original profiles.

Systems biology, and in particular network biology, offers a means to integrate large amounts of omics data [15]. Most of network biology capitalizes on the observation that genes whose function is altered in a particular phenotype tend to be co-expressed in common pathways and, therefore, co-localized in specific network regions [16]. Following this principle, it has been possible to convert intricate genome-wide signatures to network signatures, or *modules*, that are less noisy and easier to interpret [17]. Raphael and co-workers, for instance, pioneered an algorithm to map cancer mutations on biological networks and retrieve ”hot“ regions affected by these mutations, distinguishing functional (driver) mutations from sporadic (passenger) ones [18]. Califano’s group, on the other hand, combined gene expression data with regulatory cellular networks to infer protein activity [19]. Overall, network-based methods come in many flavors and offer an effective framework to organize the massive amount of omics data that has been released in the last two decades [20].

While many genes and proteins have enjoyed such a network-based annotation (being them circumspect in well-defined modules such as pathways and biological processes), this did not translate into a deeper understanding of drug activity. For a number of drugs the mechanism of action remains unclear [3] and off-targets are often discovered [21]. The recent publication of drug screens against cancer cell line panels, and the transcriptomics signatures that can be derived therefrom, provide a broader view of drug activity, enabling the full implementation of network biology techniques. Here, we undertake the task of obtaining and annotating

transcriptional modules related to over 170 drugs. We show how our modules are able to capture meaningful aspects of drug's biology, being them robust to inherent biases caused by e.g. the cell's tissue of origin, and having a tight relationship to mechanisms of action and transportation events occurring at the membrane. This, we believe, yields a novel, rich portrait of drug activity, well beyond the more limited and commonly adopted target-centered view.

Results and discussion

The Genomics of Drug Sensitivity in Cancer (GDSC) is the largest cancer cell line (CCL) panel available to date [8]. This dataset contains drug sensitivity data (growth-inhibition) for 265 drugs screened against 1001 cell lines derived from 29 tissues, together with *basal* transcriptomics profiles of the cells (among other omics data). Aware of the work by Rees et al [9], we first looked for artifactual associations between drug responses and CCLs, depending on their tissue of origin. We found that CCLs derived from neuroblastoma, hematopoietic, bone and small cell lung cancer tissues might bias global studies of drug-gene associations, due to their unspecific sensitivity to drugs (Figure S1), and removed them from further analyses. We also excluded genes whose expression levels were low or constant across the CCL panel, and drugs tested against fewer than 400 CCLs (see Methods for details). As a result, we obtained a pharmacogenomic dataset composed of 224, 15,123 genes and 637 CCLs.

Following the conventional strategy to analyze pharmacogenomic datasets, we calculated *independent* drug-gene associations, simply by correlating the expression level of each gene to the potency of the drug (area under the growth-inhibition curve; AUC) across the CCL panel. We used a z-transformed version of the Pearson's ρ, as recommended elsewhere [22]. Figure 1A shows the pair-wise distribution of the z-correlation (zc) measures between the 15,123 genes and the 224 drugs. To identify significant drug-gene associations, we set a cutoff of +/- 3 zc, based on an empirical null distribution derived from randomized data. For each drug, we obtained a median (Med) of 323 *positively* correlated genes [first quartile (Q1): 161, third quartile (Q3): 682], and Med: 242 *negatively* correlated genes [Q1: 95, Q3: 594] (Figure 1B). Some drugs, like the BRAF inhibitor Dabrafenib or the EGFR inhibitor Afatinib, had over 1,500 *positively* and *negatively* correlated genes, while others, like the antiandrogen Bicalutamide or the p38 MAPK inhibitor Doramapimod, hardly had a dozen of them. We observed that the

number of genes that correlate with drug response strongly depends on the drug class (Figure 1C), being EGFR and ERK-MAPK signaling inhibitors the classes with the largest number of associated genes, and bromodomain and chromatin histone acetylation inhibitors the ones with the fewest correlations. This variation may be partially explained by the range of drug potency across the CCL panel, since it is *easier* to detect drug-gene correlations if the drug has a wide sensitivity spectrum (Figure S2).

Similarly, the analysis of independent drug-gene correlations suggests that some genes are *positively* correlated to many drugs, sometimes with apparently unrelated mechanisms of action (MoA). For instance, we found 5% of the genes associated to more than 10% of the drugs (Figures 1B and S3). The transcripts of these "frequently-correlated genes" are enriched in membrane processes, specifically focal adhesion ($P < 1.72 \times 10^{-18}$), and extracellular matrix (ECM) organization ($P < 1.46 \times 10^{-17}$), including subunits of integrin, caveolin and platelet-derived growth factors (PDGFs). These genes determine, among others, the activation of Src kinases [23-26]. Overall, ECM proteins are known to play an important role in tumor proliferation, invasion and angiogenesis [27, 28], and are often involved in the upstream regulation of cancer pathways [29] such as PI3K/mTOR [23-25], MAPK [24] and Wnt signaling [30], cell cycle and cytoskeleton regulation [31]. It is thus not surprising that ECM genes determine drug response in a rather unspecific manner.

On the other hand, frequent *negatively* correlated genes were associated to the metabolism of small molecules (xenobiotic metabolic processes, $P < 9 \times 10^{-4}$). We found, among others, the cytochrome CYP2J2 and the GSTK1 and MGST glutathione transferases, which are highly expressed in cancers and known to confer drug resistance through their conjugating activity [32-35]. Following other studies that reported similar results [9], we checked for the presence of multidrug transporters (MDTs). Reassuringly, we found the efflux pump transporter ABCC3 and a total of 32 different solute carriers (SLC) to be *negatively* correlated to the potency of many drugs. Of note, we also found the ABCA1 transporter and other 10 SLCs to be among the frequent *positively* correlated genes, highlighting the key role of transporters and carriers in determining drug potency.

All of the above suggests that the systematic analysis of independent drug-gene correlations is sufficient to highlight *unspecific* determinants of drug sensitivity and resistance (i.e. frequent

positively and *negatively* correlated genes, respectively). However, while these determinants are recognized to play a crucial role, they do not inform targeted therapies, as they are usually unrelated to the MoA of the drug. Hence, we assessed whether measuring drug-gene correlations would also be sufficient to elucidate drug targets, i.e. we checked if the expression level of the target does correlate with the potency of the drug. Since most drugs had more than one (*n*) annotated targets, we conservatively kept the *maximum* value of the *n* corresponding drug-gene correlations. To standardize this measure and test its significance, we randomly sampled 1,000 sets of *n* genes, kept the maximum correlations, and derived an empirical Z-score. Figure 1D shows that the expression level of most drug targets did *not* correlate with drug response. Indeed, only 26 of the 207 drugs with an annotated target had "positively-correlated targets" (Z-score > 1.65, P-value ~ 0.05). Remarkably, 6 of the 7 EGF pathway inhibitors in our dataset were among these drugs, as were 3 of the 4 IGF pathway and 5 of the 25 RTK pathway inhibitors. We noted that the molecular targets for these pathways are usually cell surface receptors, e.g. EGFR, IGFR, ALK, KIT, MET and PDGFRA. Overall, of the 26 drugs with positively-correlated targets, 17 bind with cell surface receptors, showing a propensity of drug-gene correlations to capture membrane targets (odds ratio = 14.39, P-value = 9.7×10^{-9}).

The poor amount of positively-correlated targets illustrates how independent drug-gene correlations are generally unable to reveal MoAs, especially if the drug target is located downstream in signaling pathways (i.e. not in the cell surface). Some authors have suggested that the cell tissue of origin plays an important and confounding role in defining drug response signatures. To assess this possibility, we repeated the calculation of Pearson's zc correlations, separately for each of the 13 tissues represented in our dataset. We found that, in general, the trends observed at the tissue-specific level were consistent with the global trends, although tissue-specific correlations were milder due to low statistical power (i.e. few cell lines per tissue) (Figure S4A). Accordingly, we verified that none of the tissues had a dominant effect on the measures of drug-gene correlations (Figure S4B). Nonetheless, some drug-gene pairs had a rather heterogeneous pattern of correlations. For instance, going back to the targeting of EGFR (which was *positively* correlated with Afatinib and Gefinitib) we can see in Figure 1E that the "global" correlation can be mostly attributed to non-small cell lung carcinoma (NSCLC) cells (zc > 1.96, P-value < 0.05). Indeed, NSCLC is an approved indication for both drugs. Moreover, and consistent with recent studies [36-39], Gefitinib has a significant correlation to EGFR in breast CCLs, whereas Afatinib correlates with this target in pancreas.

From drug-gene correlations to drug modules

The previous analyses demonstrate that conventional drug-gene correlations do *not* directly testify drug targets, and suggest that standard transcriptomic drug signatures contain unspecific and indirect correlations that may mislead mechanistic interpretation. Recent advances in network biology precisely tackle these problems, as they can (i) filter signatures so that they become more functionally homogeneous, and (ii) they allow for the measurement of network distances so that genes that are proximal to the target can be captured and connected to it, even if the expression of the target itself is *not* statistically associated to the drug.

Hence, we set to mapping drug-gene correlations onto a large protein-protein interaction (PPI) network, retaining only genes that could be grouped in network *modules* (i.e. strongly interconnected regions of the network). The Methods section details the procedure followed to detect drug modules. In brief, we first filtered out those genes whose gene expression was frequently, and unspecifically, correlated to the potency of many drugs (Figure S3). This slightly reduced the number of associations from a median of 323 [Q1:161, Q3:682] *positively* and 242 [Q1:95, Q3:594] *negatively* correlated genes per drug to 243 [Q1:124, Q3:455] and 165 [Q1:63, Q3:421], respectively. Next, to identify genes acting in coordination (i.e. participating in enriched Reactome pathways [40, 41]), we adapted the gene set enrichment analysis (GSEA) [42] to handle drug-gene correlations instead of fold-changes, and we ran it for each drug (Figure 2B). The resulting GSEA-filtered list of genes kept Med:52 [Q1:22, Q3:110] *positive* and Med:44 [Q1:17, Q3:102] *negative* correlations per drug. After this stringent filtering, we submitted the GSEA-filtered list of genes to HotNet2 [43], a module detection algorithm that was originally developed for the identification of recurrently mutated subnetworks in cancer patients (Figure 2C). As underlying network, we chose a high-confidence version of STRING [44], which contains 14,725 proteins and 300,686 interactions. HotNet2 further filtered the list of genes correlated to each drug, keeping only those that were part of the same network neighborhood. Finally, we used a module-expansion algorithm called DIAMOnD [40] to recover strong drug-gene correlations that had been discarded along the process, mainly due to the limited coverage of the Reactome database (Figure 2D).

Our pipeline yielded at least one "positively correlated module" (PCM) for 171 of the 224 drugs, composed of Med:40 genes [Q1:23, Q3:70]). Similarly, we obtained "negatively correlated modules" (NCM) for 167 of the drugs (Med: 35 genes [Q1:18, Q3:64]). Thus, compared to the original signatures, drug modules have a considerably smaller size, achieving an 80% reduction in the number of genes (Figure 3A), commensurate with the size of manually annotated pathways in popular databases (Figure S5). For roughly two thirds of the drugs, we obtained only one PCM and one NCM and, for the remaining drugs, a second, usually smaller module was also identified. The complete list of drug modules, and functional annotations, is available in the Supplementary Data 1 file.

Drug modules are tightly related to mechanisms of action

To assess the mechanistic relevance of our drug modules, we measured their distance to the corresponding drug targets, i.e. we formulated the hypothesis that targets are proximal to dysregulated network regions. For this, we used again the DIAMOnD algorithm [40], this time to retrieve, for each drug, a list of genes ranked by their proximity to the corresponding drug module(s) (see Methods). Figure 3B shows that intended drug targets were remarkably up-ranked in these lists, being them closer to the drug modules than any other set of random proteins, including druggable and pharmacological receptors [45], which usually have prominent positions in the PPI network due to the abundant knowledge available for them. In 82% of the PCMs, the corresponding targets were among the top decile of proximal proteins, which implies a dramatic increase in mechanistic interpretability compared to the scarce 12.25% of drugs that could be connected to their targets via conventional analysis of drug-gene correlations.

A unique quality of drug modules is that network-based distances can be natively measured between them [46]. Profiting from this property, we measured the distance between all drug modules pair-wise (Supplementary Data 2) and grouped them by drug class (Figure 3C) (see Methods and alternative statistical treatments in Figure S6). It is apparent in the diagonal of Figure 3C that drugs belonging to the same category tend to have proximal modules, some of them in a very specific manner, as it is the case for the inhibitors of the ERK-MAPK signaling cascade. On the other hand, drugs targeting RTK signaling and genome integrity had modules nearby a broad range of drug classes, and they were in turn notoriously proximal among them,

in good agreement with the recently reported cross-talk between these two processes [47, 48]. More importantly, we noticed that drug modules could recapitulate non-trivial links between drug classes. IGFR, for instance, is a growth factor receptor that is known to trigger cell replication and proliferation events. Consistently, modules of drugs targeting this pathway were proximal to cell replication processes such as mitosis, cell cycle, DNA replication and other nucleus-related events.

Drug modules retain the ability to predict drug response

We have shown that drug modules are neatly related to the MoA of the drug, but the question remains as to how much they retain the predictive power of the, more complex, original transcriptomic signatures. In the CCL setting, gene expression profiles have proven to be a valuable proxy for drug response [5, 11, 49], crucially contributing to state-of-the-art pharmacogenomic predictors. To test whether our much smaller drug modules have comparable predictive power to the standard signatures, we devised a very simple predictor based on the GSEA score (see Methods). In brief, given a drug we tested whether *sensitive* cell lines were enriched in the corresponding drug modules. We expect that genes in PCMs are *over*-expressed in sensitive cell lines, whereas those in NCMs should be *under*-expressed. Analogously, we took the *positively* and *negatively* correlated genes from the original drug-gene associations (signatures), and performed the GSEA-based prediction as well. To nominate a cell "sensitive" to a certain drug, we ranked CCLs by their sensitivity and kept the top *n* CCLs, *n* being 25, 50 or 100. This simple binarization is, in practice, proportional to more sophisticated sensitivity/resistance categorizations such as the waterfall analysis [14], and has the advantage to yield prediction performance metrics comparable between drugs.

Figure S7 demonstrates that, when applied to the GDSC, drug module enrichment analysis is able to classify sensitive cell lines with high accuracy, especially for the top 25 sensitive cell lines (area under the ROC curve (AUROC): 0.76). Indeed, the predictive power of the modules has surpassed that of the original signatures (AUROC: 0.65), which is a remarkable achievement considering that drug modules are 80% smaller than the original signatures. Note, however, that such high AUROC scores are somewhat expected, since signatures and modules are themselves derived from the GDSC sensitivity data. Moreover, it is out of the scope of this study to assess whether, compared to the original signatures, drug modules

could achieve superior predictions in more complicated machine learning settings, and we do not make claims in this regard. What matters in this baseline exercise is that the predictive potential is maintained when using modules instead of signatures.

To test whether drug modules are still predictive in settings other than the GDSC, we collected gene expression and drug potency data from the Cancer Therapeutic Response Portal (CTRP), another major CCL panel. About 37% of our drugs were also available in this panel. In CTRP, drug sensitivity data is measured independently to GDSC, which poses an additional challenge to the prediction task due experimental inconsistencies and the use of different growth-inhibition metrics [14]. 403 of the CCLs are shared between GDSC and CTRP, and gene expression data is also measured independently. We performed the GSEA-based sensitivity prediction for *all* CTRP CCLs. Figure 3D shows the distribution of prediction performances for the 70 drugs, together with illustrative ROC curves belonging to drugs for which we achieved good (AUROC > 0.8, top), acceptable (0.8 > AUROC > 0.6, bottom-left) and bad (AUROC < 0.6, bottom-right) predictions. As it can be seen, when focusing on the top 25 sensitive CCLs, over a fourth of the drugs had good predictions, including the nicotinamide phosphoribosyltransferase inhibitor Daporinad. Acceptable predictions could be reached for half of the cases (e.g. I-BET-762), which is a comparable result to recent attempts to translate sensitivity predictors between different CCL panels [50]. Unfortunately, for the remaining 24 of the drugs, the predictive performance was not different to the random expectation (e.g. Docetaxel). Noticeably, performance scores did not decline when focusing *only* on CCLs that were exclusive to CTRP (i.e. not part of the GDSC dataset) (Figure 3D, dashed lines).

Encouraged by the robust predictive power of the drug modules, and as a means to assess their potential clinical applicability, we tested our collection of modules against The Cancer Genome Atlas (TCGA) cohort, containing gene expression data for 6,276 patients, and 19 tumor types. After running our pipeline, we could assign at least one drug module to almost every patient, suggesting that our results may be of clinical relevance. When considering only drugs for which modules achieve a good prediction of the response in GDSC (AUROC > 0.8), we found that half of them had at least one module (PCM or NCM) enriched in more than 10% of the patients (FDR < 0.01), indicating that our cell-based modules are also detected in clinical samples (Figure 3E). However, further analyses and, especially, consistent clinical drug response data [51] are needed to strengthen our claims in this regard.

Module-based characterization of drugs

Given that, by definition, drug modules relate to certain vicinities in biological networks, they are expected to be (at least to some extent) functionally homogeneous. This feature makes modules easy to interpret. To illustrate this feature, we tested drug modules for enrichment against a collection of high-order biological processes (the Hallmark gene sets) available from the Molecular Signatures Database (MSigDB) [52]. As illustrative example, we focus on three drug classes, namely drugs targeting mitosis, RTK and ERK-MAPK signaling. Figure 4A shows a heatmap with the enrichment (odds ratio) of drug modules across the Hallmark biological processes.

It can be readily seen that drugs targeting mitosis have modules enriched in processes related to cell cycle and replication (Figure 4A, top). Specifically, processes associated to the Myc transcription factor were enriched in all of the drugs, suggesting that the efficiency of mitosis targeting depends on the status of this proto-oncogene. The MPS1 inhibitor MPS-1-IN-1, on the other hand, showed a specific enrichment for the Sonic hedgehog (Shh) signaling pathway. Dysregulation of the Shh pathway is associated to one of the major subgroups of medulloblastoma, a malignant form of brain cancer. Indeed, MPS1 is highly expressed in medulloblastomas and recent studies highlight it as a promising therapeutic target [53, 54].

In contrast to mitosis inhibitors, drugs targeting the RTK pathway are enriched in biological processes outside the nucleus (Figure 4A, middle). Among them we found hypoxia and epithelial-mesenchymal transition (EMT). Both mechanisms are known to be associated to tyrosine kinases [55, 56]. Interestingly, a subgroup of RTK inhibitors (namely ACC220, CEP-701, NVP-BHG712 and MP470) were characteristically associated to the PI3K-AKT-mTOR signaling cascade. With the exception of NVP-BHG712, these inhibitors have the tyrosine kinase FLT3 as a common target, in good agreement with recent studies showing a strong connection between FLT3 and the PI3K-mTOR pathway and reporting a critical role of the latter in conferring drug resistance [57, 58].

Compared to the two previous drug classes, the modules of ERK-MAPK pathway inhibitors are enriched in a wide range of biological processes (Figure 4A, bottom). However, while some

processes like apoptosis are detected in most of the 15 drugs in this category, others are rather MoA-specific. Oxidative phosphorylation (OXPHOS), for example, is represented in 3 of the 4 BRAF inhibitors. It is known that, while BRAF inhibitors boost OXPHOS (leading to oncogene-induced senescence), activation of glycolytic metabolism followed by OXPHOS inactivation yields drug resistance [59, 60]. Similarly, VX11e, the only drug in our dataset targeting ERK2, shows a distinctive enrichment in Myc-regulated proteins, while FMK, the only drug targeting the Ribosomal S6 kinase, is enriched in p53 signaling pathway and inflammatory response processes. All these observations are consistent with previous studies [61-64].

To assess the added value of drug response modules with respect to their biological interpretability, we compare the above results with the Hallmark genes enrichments obtained directly from the *positively* and *negatively* correlated genes (Figure S8). When plotting the distributions of enrichment scores, we observe a significant shift towards higher enrichments in the drug response modules (Figure 4B), with over 1,250 Hallmark-drug pairs showing values >5, compared to only about 250 pairs for the direct correlations (Figure 4C). We also provide the complete enrichment analysis for the *full* collection of drugs (Supplementary Data 3), and a description of drug modules from different perspectives, such as their size, coverage with respect to the original drug signatures, network distance to drug targets, subcellular localization of the gene products and the proportion of disease-related and druggable genes (Supplementary Data 4). The comparison of these features allows for a rapid assessment of drug modules. For example, we noticed that drugs targeting cell replication yield modules that are far apart from the drug targets and, more generally, we found a clear relationship between modules´ distances to the corresponding drug targets and the fraction of membrane and nuclear proteins in the module, being membrane-rich modules usually closer to the drug targets (Figure S9).

Concluding remarks

In the last decade we have witnessed the rise of CCL pharmacogenomic studies. However, the use of CCL panel screens has been controversial from the early days, mainly because results are difficult to reproduce across platforms, and it remains unclear whether immortalized cells are still able to model patient samples. Another limitation is the overwhelming amount of drug-

gene correlations that can be derived from these experiments, yielding genetic patterns of drug sensitivity that are impossible to interpret. We have shown, for example, that (i) the number of correlated genes is highly variable across drugs, that (ii) some genes are unspecifically correlated to many drugs and that (iii) not all drug-gene pairs are equally correlated in every tissue. We propose that converting transcriptomic signatures to network modules may simplify the analysis, giving smaller, more robust and functionally-coherent sets of drug-gene correlations. We have extensively validated this strategy, proving that the drug response modules have a tight connection to the mechanism of action, being them enriched in biological processes of pharmacological relevance, and exhibiting comparable predictive power to the original, more complicated signatures.

Despite being comprehensive and predictive, our approach inherits some of the limitations of ordinary transcriptomics analysis. Expression levels of mRNA do not perfectly match protein abundance, nor they are able to capture post-translational modifications such as phosphorylation events, which are key to some of the pathways studied here. Moreover, notorious variability in gene expression and drug sensitivity data are necessary for drug-gene correlations to be captured, requiring in practice considerably large panels of CCLs, which limits the throughput of the technique to a few hundred drugs. In particular, one cannot precisely measure correlations within poorly represented tissues, which in turn makes it difficult to disentangle tissue-specific transcriptomic traits that are irrelevant to drug response. Our module-based approach partially corrects for this confounding factor, although integration of other CCL omics data (such as mutations, copy number variants and chromatin modifications) may further ameliorate these issues and also provide new mechanistic insights. For this, systems biology tools that can learn the relationships between different layers of biology are paramount. In this line, the release of CCL screens with readouts other than growth inhibition or proliferation rate [65, 66] will help unveil the connections between the genetic background of the cells and the phenotypic outcome of drug treatment.

All in all, though, transcriptomics is likely to remain the dominant genome-wide data type for drug discovery, as recent technical and statistical developments have drastically reduced its cost [67]. The L1000 Next-generation Connectivity Map, for instance, contains about one million post-treatment gene expression signatures for about twenty thousand molecules [65]. These signatures await to be interpreted and annotated and, more importantly, they have to be

connected to pre-treatment signatures in order to identify therapeutic opportunities. We believe that network biology blueprints like the one presented here will enable this connection, encircling the relevant parts of the signatures and measuring the distances between them.

Methods

Data preparation and drug-gene correlations. We collected gene expression and drug response data from the GDSC web resource (https://www.cancerrxgene.org). In order to consider only genes suitable for correlation analysis, we discarded those whose expression levels were low or stable across cell lines. For this, we analyzed the distribution of basal gene expression of each gene in every CCL, and filtered out the ones with an expression level below 5 (log2 units) across the panel. Regarding drug response data, GDSC provides measurements of cell survival at different drug concentrations (Area Under the dose-response Curve, AUC). As this measure is inversely proportional to the drug sensitivity (i.e. the more sensitive the cell, the less it survives), we used the 1-AUC as our measure of potency. Hence, *positive* correlations denote drug sensitivity due to gene over-expression, while *negative* correlations denote that sensitivity is associated to gene under-expression.

Recent studies reported a confounding effect linked to certain tissues in the global analysis of drug-gene correlations. To identify these potential biases in our dataset, we performed a principal component analysis (PCA) on the matrix of raw drug-gene correlations (Pearson's ρ between 1-AUC and gene expression units). Then, we correlated the loadings of the first PC of the correlation matrix with gene expression values of each CCL, as recommended by Rees et al. [9]. Finally, we filtered out CCLs belonging to tissues that were distinctively correlated to the drug-gene correlations profiles (Figure S1), i.e. we removed leukemia, myeloma, lymphoma, neuroblastoma, small-cell lung and bone CCLs. In addition, we only kept drugs with sensitivity measurements available for at least 400 CCLs.

After this filtering process, we calculated again, for each drug-gene pair, the Pearson's correlations between basal gene expression and 1-AUC drug potencies across CCLs. We applied the Fisher's z-transformation to the correlation coefficients in order to account for variation in the number of CCLs available for each drug [22]. Overall, we obtained *positive* and

negative drug-gene correlations for 224 drugs and 15,123 genes across a total of 637 cancer cell lines. We considered significant drug-gene correlations (zc) beyond +/-3 (Figure 2A).

Frequently correlated genes. We counted, for each gene, the number of drugs for which the gene was found to be significantly correlated (zc beyond +/- 3), and analyzed the resulting cumulative distribution (Figure S3). Looking at the distribution's plateau (starting at count 24), we identified a portion of genes frequently correlated to many drugs. From this plateau, we retrieved a total of 787 *positive* and 760 *negative* frequently correlated genes (FCG) (~5% of the total genes). Finally, we performed a Fisher's exact test to identify biological processes enriched in those FCGs, using the Gene Ontology database [68] and the DAVID toolbox (https://david.ncifcrf.gov/summary.jsp).

Tissue correlations. We first split the CCL panel into sets of CCLs belonging to the same tissue. Then, we calculated drug-gene correlations (zc) separately for each of the 13 tissues represented in our dataset. In order to verify that the measures of *positively* correlated genes (PCG) and *negatively* correlated genes (NCG) were consistent across tissues, we calculated, for each drug-PCG/NCG pair, the median of zc across tissues. Reassuringly, tissue-specific correlations did, in general, have the same direction (same sign of zc) than the global correlation used throughout this study (Figure S4B).

Drug-target correlations. We obtained drug targets from the GDSC resource (disambiguating them with DrugBank [69], when necessary). We could assign at least one target to 211 of the total 224 drugs. We focused on the zc correlation of the targets to assess whether targets expression values were positively correlated to drug sensitivity. As not all drug targets will necessarily be correlated to drug sensitivity, we kept the maximum correlation value across the targets. Then, to validate the statistical significance of this measure, we performed a bootstrap analysis. For each drug target, we sampled a random drug-gene pair from our distribution. Here again, from all the sampled pairs, we kept the maximum correlation. This process was repeated 1,000 times for each drug. This allowed us to apply a Z-score normalization to the selected target correlations, thus adjusting for variation in drug target number and making results comparable between drugs.

Drug module detection. We first removed frequently correlated genes from the significant drug-gene correlations, reducing the number of associations to Med:243 [Q1:124, Q3:455] *positively* and Med:165 [Q1:63, Q3:421] *negatively* correlated genes per drug. Further, we used the correlation values (zc) to run, for each drug, a gene-set enrichment analysis (GSEA) [42] in order to identify what genes participate in enriched Reactome pathways [40, 41]. We only considered Reactome pathways composed of at least 5 genes. Then, for each drug, we kept the significantly correlated genes found in any of the enriched pathways (P-value < 0.01), regardless of the sign of the enrichment (Figure 2B). The resulting GSEA-filtered list of genes kept Med:52 [Q1:22, Q3:110] *positive* and Med:44 [Q1:17, Q3:102] *negative* correlations per drug. Then, taking the zc values as input scores, we submitted the GSEA-filtered list of genes to HotNet2 [43], using a high-confidence version of STRING [44] (confidence score > 700, containing 14,725 proteins and 300,688 interactions) (Figure 2C). We ran HotNet2 iteratively, keeping the largest module and removing its genes for the next iteration, until the modules had less than 5 genes or were not statistically significant (P-value > 0.05). To recover strong drug-gene correlations proximal to the drug modules, we used the module-expansion algorithm DIAMOnD [40] (Figure 2D). For each drug, we recovered those significantly correlated genes that fulfilled any of the following two criteria: (i) genes that were neither present in any of the Reactome pathways nor frequently-correlated to many drugs but were in the top 200 closest genes to the module or (ii) frequently-correlated genes in the top 100 closest genes to the module. Hence, we obtained at least one *positively* correlated module for 171 of the drugs (Med:40 genes [Q1:23, Q3:70]) and one *negatively* correlated module for 167 of the drugs (Med:35 genes [Q1:18, Q3:64]).

Drug target-module distances. The module-expansion algorithm DIAMOnD [40] provides a list of genes sorted by their network-based proximity to the module. Accordingly, we retrieved, from our filtered version of STRING (see details above), the top closest 1,450 genes (~10% of the largest connected component of the network) for every drug module. Then, we annotated the position of the drug targets in the resulting DIAMOnD lists, taking the median value when having more than one drug target. Since our DIAMOnD lists contain the top decile of proximal proteins to each drug module, we transformed those ranking positions measures to network percentile measures. To put drug target distances in perspective, we measured distances to three different sets of random proteins. The first consisted of random proteins retrieved from the STRING network. For the second, we collected all genes defined as *Tclin* or *Tchem* in the

Target Central Resource Database [45], named "Druggable proteins". Finally, the third random set integrates all the pharmacologically active drug targets covered by DrugBank (https://www.drugbank.ca/).

Distances between modules. We calculated, separately, distances between *positively* and *negatively* correlated modules using a network distance based on percolation theory [46]. As this distance measure is sensible to the size of the modules, we normalized the values. We first grouped our modules by their number of genes (size) and, for each module, we calculated the distribution of shortest distances from the gene with highest closest centrality [70] to the rest. Using the obtained distribution, we randomly selected one gene from the network and, by following this distribution, we retrieved neighbours at the given distances until we obtained a new module. If the distribution constraint could not be met (often due to lack of neighbours at specific distance) but the resulting module had at least half of the genes, we used the DIAMOnD algorithm [40] to retrieve the remaining genes (alpha=10). We repeated this process until we obtained 10 random modules of each size. Next, we distributed the random modules into ranges (intervals) of 5 (i.e. from 10 to 14, from 15 to 19...), obtaining 50 random modules in each interval. Then, for each group, we randomly retrieved 30 modules (without replacement) and calculated the network-based separation distance between them. As a result, we obtained a random distribution of specific distances at each interval, from which we calculated their mean and standard deviation. Finally, we used the corresponding mean and standard deviation to derive a Z-score for every module-module distance. Hence, we show the distances as Z-score values, where the more negative the value the closer the distance between modules.

Drug response prediction using drug modules. We performed drug response predictions in the GDSC dataset using the drug modules and the *positively/negatively* correlated genes (PCG/NCG), beyond +/-3 zc (we considered only those PCG/NCG from drugs having at least one module). We devised a simple GSEA-based predictor in which we evaluated which CCLs were enriched in the corresponding PCG, NCG or drug modules. To do so, we first normalized the gene expression of each gene across the different CCLs (i.e. calculating standard scores for each CCL-gene pair). Further, we calculated the enrichment scores of each CCL for every drug's PCG, NCG and modules. Then, to nominate a CCL "sensitive" to a certain drug, we ranked the CCLs by their sensitivity and kept the top *n* CCLs (*n* being 25, 50 or 100). Finally, we built Receiver Operating Characteristic (ROC) curves for each drug using the enrichment

scores (ES) and each of the *n* CCL sensitive binarization. To obtain a single prediction for each drug, we kept the highest ES among the modules and the significantly correlated genes. To make these results comparable, we only considered the first *positively* and *negatively* correlated module of each drug.

To compare whether the modules derived from GDSC data can be transferred to other settings, we applied our prediction procedure to the Cancer Therapeutics Response Portal (CTRP) data, with. In total, 70 of the GDSC drugs were also present in CTRP. In this case, we collected the gene expression data from the CTRP web resource (https://ocg.cancer.gov/programs/ctd2/data-portal). In accordance to the GDSC prefiltering, we removed all CCLs derived from neuroblastomas, hematopoietic, bone and small cell lung cancer tissues, leaving a total of 636 CCLs, 403 in common with our GDSC panel. Drug response predictions for CTRP were performed as detailed above. In this case, however, we used the best ES among all the modules associated to the drug. Additionally, we repeated this analysis using only CCLs exclusive to CTRP (i.e. not shared with the GDSC panel).

Drug module enrichments in TCGA patients. We downloaded gene expression data ("median Z-scores") for 6,276 patients and 19 cancer tissues from the cBioPortal resource (http://www.cbioportal.org). Then, we performed a Gene Set Enrichment Analysis (GSEA) for each patient and drug module, ensuring that the direction (up/down) of the GSEA result corresponded to the direction of the module. Finally, we applied the Benjamini-Hochberg correction to select patient-drug module pairs (FDR < 0.01). In the Figure 3E we considered only the enrichments of the first drug modules. For a complete list of the enrichment results, please see Supplementary Data 4.

Modules enrichment in Hallmark gene sets. We downloaded the Hallmark gene set collection from the Molecular Signature Database (MSigDB) of the Broad Institute http://software.broadinstitute.org/gsea/index.jsp). From each gene set, we removed those genes that were not among our original 15,123 GDSC transcripts. Further, we combined all the second modules with the first, obtaining a maximum of two modules per drug. Finally, we performed, for each drug, a Fisher's exact test enrichment analysis using separately the *positively* and *negatively* correlated modules. In Figure 4A, we analyzed three different drug

classes by plotting the enrichment scores (odds ratio) in a heatmap. All the Hallmark enrichments are provided in Supplementary Data 3.

Building the features table. We collected data from our original correlations, several network-based measures and different biological collections found in the literature. Then, we manually integrated all the data, calculated informative scores for each module and scaled the values if necessary. All these efforts led to the creation of a features table containing 21 different components. In brief, we can divide these components in the following groups: (i) **General features**, which illustrate basic and general features derived from the omics panel. (ii) **Network features**, where we detail different features involving network measurements. (iii) **Biological features**, showing a collection of biological features related to drug's biology. For a detailed description of each feature, please see Supplementary Data 4 file and the corresponding legend.

Acknowledgements

The authors would like to thank Prof. Ben Raphael and Dr. Matthew Reyna (Princeton University) for guidance in the use of HotNet2. This works has been partially supported by the Spanish Ministerio de Economía y Competitividad (BIO2016-77038-R) and the European Research Council (SysPharmAD: 614944). AF-T received a fellowship of the IRB Barcelona Math4Life programme.

References

1. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. Nat Rev Genet. 2007;8(8):601. doi: 10.1038/nrg2137.
2. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov. 2011;10(6):428-38. doi: 10.1038/nrd3405.
3. Swinney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov. 2011;10(7):507-19. doi: 10.1038/nrd3480.
4. Monks A, Scudiero D, Skehan P, Shoemaker R, Paull K, Vistica D, et al. Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. J Natl Cancer Inst. 1991;83(11):757-66.
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. doi: 10.1038/nature11003.
6. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(D1):D955-D61. doi: 10.1093/nar/gks1111.
7. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015;5(11):1210-23. doi: 10.1158/2159-8290.cd-15-0235.
8. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016;166(3):740-54. doi: 10.1016/j.cell.2016.06.017.
9. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12(2):109-16. doi: 10.1038/nchembio.1986.
10. Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. Genome Res. 2017;27(10):1743-51. doi: 10.1101/gr.221077.117.
11. Lee S-I, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat Commun. 2018;9(1):42. doi: 10.1038/s41467-017-02465-5.
12. Liu X, Yang J, Zhang Y, Fang Y, Wang F, Wang J, et al. A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia. Sci Rep. 2016;6:22811. doi: 10.1038/srep22811.

13. Niepel M, Hafner M, Duan Q, Wang Z, Paull EO, Chung M, et al. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. Nat Commun. 2017;8(1):1186. doi: 10.1038/s41467-017-01383-w.

14. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, et al. Inconsistency in large pharmacogenomic studies. Nature. 2013;504(7480):389-93. doi: 10.1038/nature12831.

15. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009;461(7261):218-23. doi: 10.1038/nature08454.

16. Kim Y-A, Przytycka TM. Bridging the Gap between Genotype and Phenotype via Network Approaches. Front Genet. 2012;3:227. doi: 10.3389/fgene.2012.00227.

17. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761supp):C47-C52. doi: 10.1038/35011540.

18. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18(3):507-22. doi: 10.1089/cmb.2010.0265.

19. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet. 2016;48(8):838-47. doi: 10.1038/ng.3593.

20. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56-68. doi: 10.1038/nrg2918.

21. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25(2):197-206. doi: 10.1038/nbt1284.

22. Dančík V, Carrel H, Bodycombe NE, Seiler KP, Fomina-Yadlin D, Kubicek ST, et al. Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. J Biomol Screen. 2014;19(5):771-81. doi: 10.1177/1087057113520226.

23. Gilcrease MZ. Integrin signaling in epithelial cells. Cancer Lett. 2007;247(1):1-25. doi: 10.1016/j.canlet.2006.03.031.

24. Givant-Horwitz V, Davidson B, Reich R. Laminin-induced signaling in tumor cells. Cancer Lett. 2005;223(1):1-10. doi: 10.1016/j.canlet.2004.08.030.

25. Keely PJ, Westwick JK, Whitehead IP, Der CJ, Parise LV. Cdc42 and Rac1 induce integrin-mediated cell motility and invasiveness through PI(3)K. Nature. 1997;390(6660):632-6. doi: 10.1038/37656.

26. Gelderloos JA, Rosenkranz S, Bazenet C, Kazlauskas A. A Role for Src in Signal Relay by the Platelet-derived Growth Factor α Receptor. J Biol Chem. 1998;273(10):5908-15. doi: 10.1074/jbc.273.10.5908.

27. Patarroyo M, Tryggvason K, Virtanen I. Laminin isoforms in tumor invasion, angiogenesis and metastasis. Semin Cancer Biol. 2002;12(3):197-207. doi: 10.1016/S1044-579X(02)00023-8.

28. Keely P, Parise L, Juliano R. Integrins and GTPases in tumour cell growth, motility and invasion. Trends Cell Biol. 1998;8(3):101-6. doi: 10.1016/s0962-8924(97)01219-1.

29. Kim S-H, Turnbull J, Guimond S. Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. J Endocrinol. 2011;209(2):139-51. doi: 10.1530/JOE-10-0377.

30. Du J, Zu Y, Li J, Du S, Xu Y, Zhang L, et al. Extracellular matrix stiffness dictates Wnt expression through integrin pathway. Sci Rep. 2016;6:20395. doi: 10.1038/srep20395.

31. Moreno-Layseca P, Streuli CH. Signalling pathways linking integrins with cell cycle progression. Matrix Biol. 2014;34:144-53. doi: 10.1016/j.matbio.2013.10.011.

32. Lee CA, Neul D, Clouser-Roche A, Dalvie D, Wester MR, Jiang Y, et al. Identification of novel substrates for human cytochrome P450 2J2. Drug Metab Dispos. 2010;38(2):347-56. doi: 10.1124/dmd.109.030270.

33. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013;138(1):103-41. doi: 10.1016/j.pharmthera.2012.12.007.

34. Allocati N, Masulli M, Di Ilio C, Federici L. Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases. Oncogenesis. 2018;7(1):8. doi: 10.1038/s41389-017-0025-3.

35. Hayes JD, Flanagan JU, Jowsey IR. Glutathione transferases. Annu Rev Pharmacol Toxicol. 2005;45:51-88. doi: 10.1146/annurev.pharmtox.45.120403.095857.

36. Green MD, Francis PA, Gebski V, Harvey V, Karapetis C, Chan A, et al. Gefitinib treatment in hormone-resistant and hormone receptor-negative advanced breast cancer. Ann Oncol. 2009;20(11):1813-7. doi: 10.1093/annonc/mdp202.

37. Zhang X, Zhang B, Liu J, Liu J, Li C, Dong W, et al. Mechanisms of Gefitinib-mediated reversal of tamoxifen resistance in MCF-7 breast cancer cells by inducing ERα re-expression. Sci Rep. 2015;5:7835. doi: 10.1038/srep07835.

38. Huguet F, Fernet M, Giocanti N, Favaudon V, Larsen AK. Afatinib, an Irreversible EGFR Family Inhibitor, Shows Activity Toward Pancreatic Cancer Cells, Alone and in Combination with Radiotherapy, Independent of KRAS Status. Target Oncol. 2016;11(3):371-81. doi: 10.1007/s11523-015-0403-8.

39. Ioannou N, Dalgleish AG, Seddon AM, Mackintosh D, Guertler U, Solca F, et al. Anti-tumour activity of afatinib, an irreversible ErbB family blocker, in human pancreatic tumour cells. Br J Cancer. 2011;105(10):1554-62. doi: 10.1038/bjc.2011.396.

40. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015;11(4):e1004120. doi: 10.1371/journal.pcbi.1004120.

41. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649-D55. doi: 10.1093/nar/gkx1132.

42. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50. doi: 10.1073/pnas.0506580102.

43. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106-14. doi: 10.1038/ng.3168.

44. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 2016;45(D1):D362-D8. doi: 10.1093/nar/gkw937.

45. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, et al. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov. 2018. doi: 10.1038/nrd.2018.14.

46. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601. doi: 10.1126/science.1257601.

47. Mahajan K, Mahajan NP. Cross talk of tyrosine kinases with the DNA damage signaling pathways. Nucleic Acids Res. 2015;43(22):10588-601. doi: 10.1093/nar/gkv1166.

48. Chen M-K, Hung M-C. Regulation of therapeutic resistance in cancers by receptor tyrosine kinases. Am J Cancer Res. 2016;6(4):827-42.

49. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol. 2014;15(3):R47. doi: 10.1186/gb-2014-15-3-r47.

50. Juan-Blanco T, Duran-Frigola M, Aloy P. Rationalizing Drug Response in Cancer Cell Lines. J Mol Biol. 2018. doi: 10.1016/j.jmb.2018.03.021.

51. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173(2):400-16.e11. doi: 10.1016/j.cell.2018.02.052.

52. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems. 2015;1(6):417-25. doi: 10.1016/j.cels.2015.12.004.

53. Alimova I, Ng J, Harris P, Birks D, Donson A, Taylor MD, et al. MPS1 kinase as a potential therapeutic target in medulloblastoma. Oncol Rep. 2016;36(5):2633-40. doi: 10.3892/or.2016.5085.

54. Maachani UB, Kramp T, Hanson R, Zhao S, Celiku O, Shankavaram U, et al. Targeting MPS1 Enhances Radiosensitization of Human Glioblastoma by Modulating DNA Repair Proteins. Mol Cancer Res. 2015;13(5):852-62. doi: 10.1158/1541-7786.MCR-14-0462-T.

55. Glück AA, Aebersold DM, Zimmer Y, Medová M. Interplay between receptor tyrosine kinases and hypoxia signaling in cancer. Int J Biochem Cell Biol. 2015;62:101-14. doi: 10.1016/j.biocel.2015.02.018.

56. Thiery JP. Epithelial–mesenchymal transitions in development and pathologies. Curr Opin Cell Biol. 2003;15(6):740-6. doi: 10.1016/j.ceb.2003.10.006.

57. Lindblad O, Cordero E, Puissant A, Macaulay L, Ramos A, Kabir NN, et al. Aberrant activation of the PI3K/mTOR pathway promotes resistance to sorafenib in AML. Oncogene. 2016;35(39):5119-31. doi: 10.1038/onc.2016.41.

58. Nogami A, Oshikawa G, Okada K, Fukutake S, Umezawa Y, Nagao T, et al. FLT3-ITD confers resistance to the PI3K/Akt pathway inhibitors by protecting the mTOR/4EBP1/Mcl-1 pathway through STAT5 activation in acute myeloid leukemia. Oncotarget. 2015;6(11):9189-205. doi: 10.18632/oncotarget.3279.

59. Haq R, Shoag J, Andreu-Perez P, Yokoyama S, Edelman H, Rowe GC, et al. Oncogenic BRAF regulates oxidative metabolism via PGC1α and MITF. Cancer Cell. 2013;23(3):302-15. doi: 10.1016/j.ccr.2013.02.003.

60. Haq R, Fisher DE, Widlund HR. Molecular pathways: BRAF induces bioenergetic adaptation by attenuating oxidative phosphorylation. Clin Cancer Res. 2014;20(9):2257-63. doi: 10.1158/1078-0432.CCR-13-0898.

61. Amatangelo MD, Goodyear S, Varma D, Stearns ME. c-Myc expression and MEK1-induced Erk2 nuclear localization are required for TGF-beta induced epithelial–mesenchymal transition and invasion in prostate cancer. Carcinogenesis. 2012;33(10):1965-75. doi: 10.1093/carcin/bgs227.

62. Marampon F, Ciccarelli C, Zani BM. Down-regulation of c-Myc following MEK/ERK inhibition halts the expression of malignant phenotype in rhabdomyosarcoma and in non muscle-derived human tumors. Mol Cancer. 2006;5:31. doi: 10.1186/1476-4598-5-31.

63. Moens U, Kostenko S, Sveinbjørnsson B. The Role of Mitogen-Activated Protein Kinase-Activated Protein Kinases (MAPKAPKs) in Inflammation. Genes. 2013;4(2):101-33. doi: 10.3390/genes4020101.

64. Cho Y-Y, He Z, Zhang Y, Choi HS, Zhu F, Choi BY, et al. The p53 Protein Is a Novel Substrate of Ribosomal S6 Kinase 2 and a Critical Intermediary for Ribosomal S6 Kinase 2 and Histone H3 Interaction. Cancer Res. 2005;65(9):3596-603. doi: 10.1158/0008-5472.can-04-3935.

65. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-52.e17. doi: 10.1016/j.cell.2017.10.049.

66. Wawer MJ, Li K, Gustafsdottir SM, Ljosa V, Bodycombe NE, Marton MA, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed

high-dimensional profiling. Proc Natl Acad Sci U S A. 2014;111(30):10911-6. doi: 10.1073/pnas.1410933111.

67. Frumkin I, Schirman D, Rotman A, Li F, Zahavi L, Mordret E, et al. Gene Architectures that Minimize Cost of Gene Expression. Mol Cell. 2017;65(1):142-53. doi: 10.1016/j.molcel.2016.11.007.

68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9. doi: 10.1038/75556.

69. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668-72. doi: 10.1093/nar/gkj067.

70. Freeman LC. Centrality in social networks conceptual clarification. Soc Networks. 1978;1(3):215-39. doi: 10.1016/0378-8733(78)90021-7.

Figures

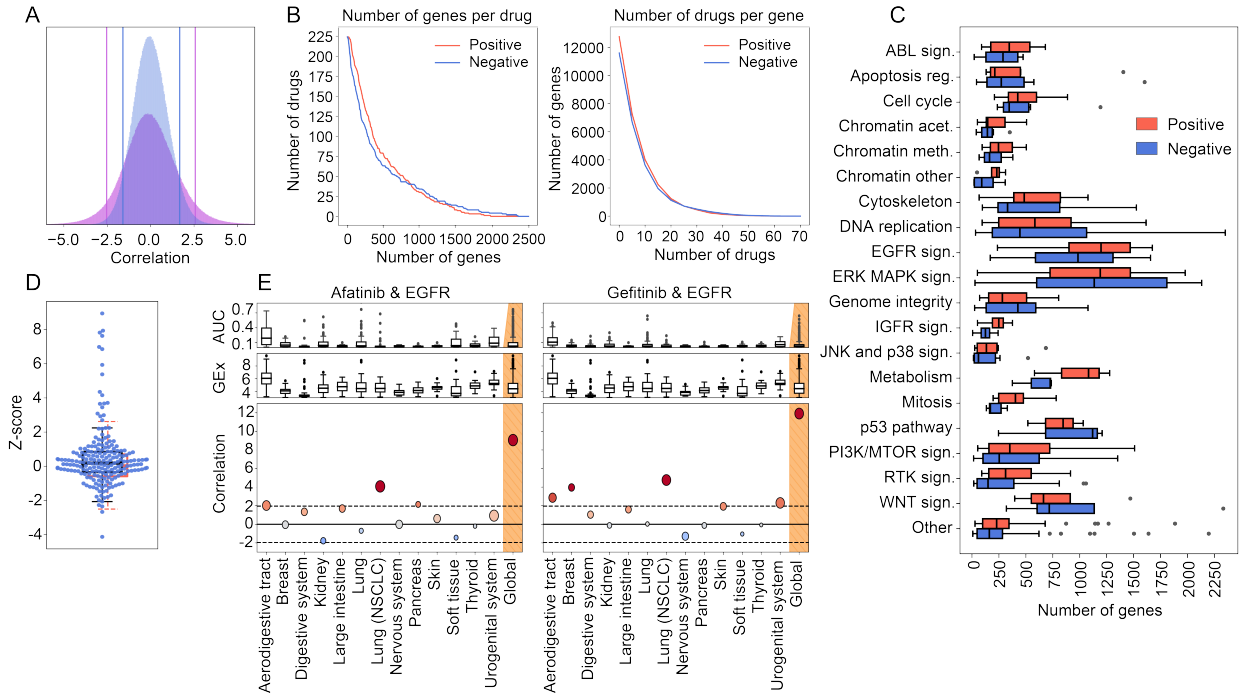

Figure 1. Analysis of drug-gene correlations. (A) Original drug-gene correlation distribution (purple) and random drug-gene correlation distribution (blue) in which we randomly permuted gene expression values across cell lines. The vertical lines show the percentiles 5 and 95 of each distribution. (B) (Left) First, the number of correlated genes is counted for each drug. Then, we plotted the number of drugs versus the number of genes per drug. (Right) In this case, the number of drugs is counted for each gene. Accordingly, we plotted the number of genes versus the number of drugs per gene. (C) Number of positively (red) and negatively (blue) correlated genes across the different drug classes. (D) Maximum positively-correlated targets for each drug after applying a bootstrap normalization (see methods). Each dot represents one drug-target correlation. The red boxplot shows the background (random) distribution of this approach. (E) Z-transformed correlations (zc) between the drugs Afatinib (left) and Gefenitib (right) and the Epidermal Growth Factor Receptor (EGFR) gene expression in each of the tissues. In the first row, we show the drug sensitivity (AUC) across tissues (measured in the corresponding cancer cell lines (CCLs)). In the second, we provide the basal gene expression (GEx) of EGFR across tissues. The rightmost column (background in orange) shows the correlation considering all the tissues (Global). The size of the bubbles is proportional to the number of CCLs in each tissue while the colors show the degree of correlation (i.e. the redder the more positive correlation, while the bluer the more negative correlation).

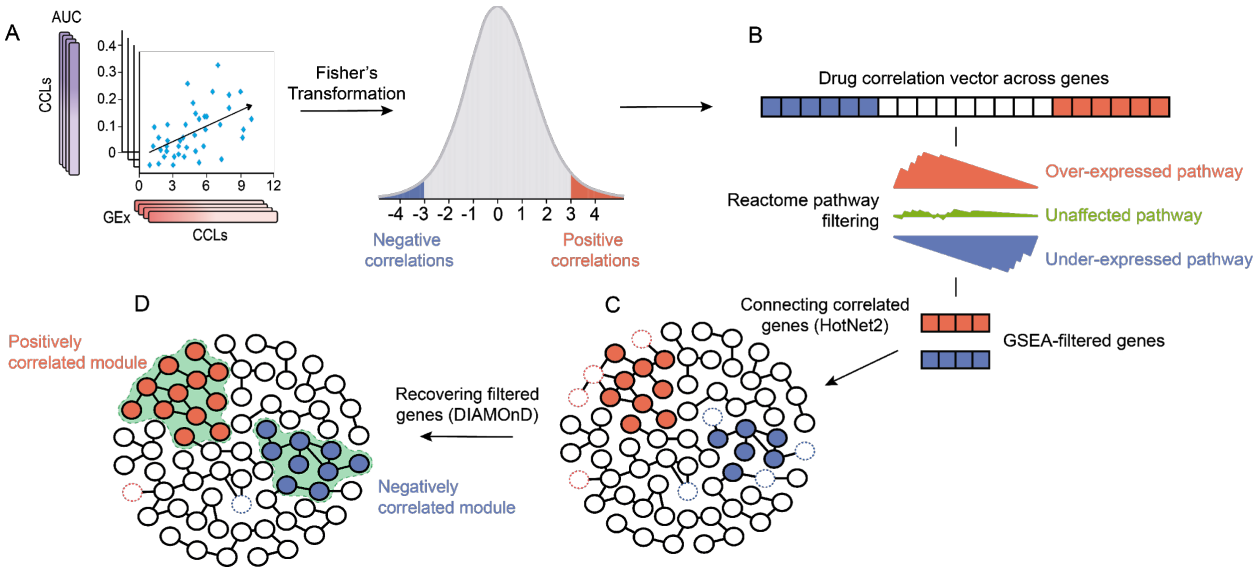

Figure 2. Methodological pipeline to identify drug modules. (A) The process of obtaining modules starts with the calculation of z-transformed Pearson's correlation (zc) between gene expression and drug sensitivity data of each drug-gene pair. From the zc distribution, we considered a gene significantly correlated to the drug when its zc was beyond +/- 3. (B) Then, after removing those "frequently-correlated genes" to the potency of drugs, we ran, for each drug, a gene-set enrichment analysis (GSEA) in order to identify what genes participated in enriched Reactome Pathways. By only keeping those significantly correlated genes found in any enriched Reactome pathway, we further filtered the number of genes correlated to the drug (GSEA-filtered genes). (C) At this point, we submitted the GSEA-filtered list of genes to HotNet2, obtaining characteristic drug modules from the STRING network. (D) However, the stringent filtering of this approach can constraint the number of significantly correlated genes in the modules (dashed circles). To address this issue, we used the module-expansion algorithm DIAMOnD to recover those pre-filtered genes that were close to the module.

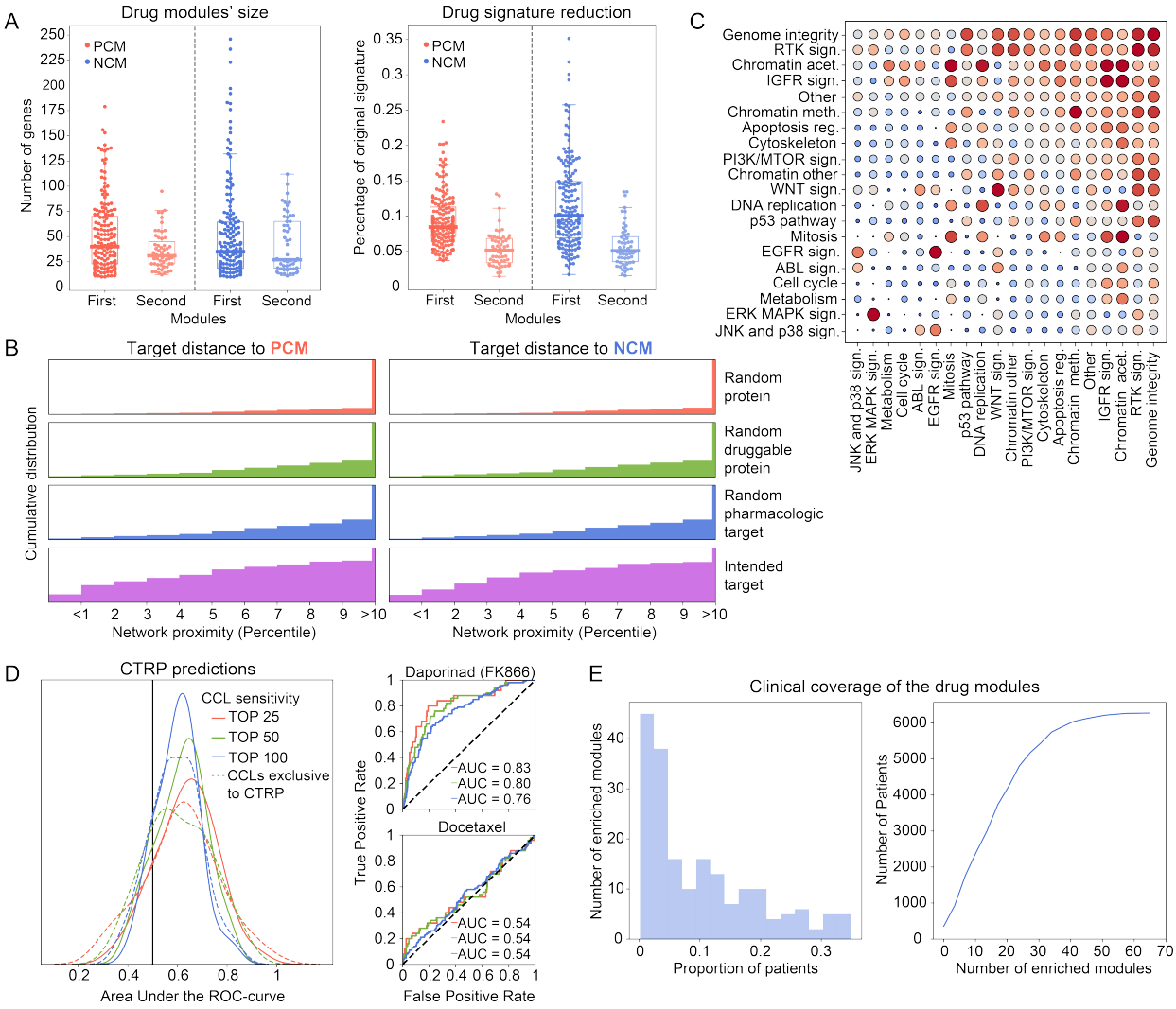

Figure 3. Drug modules' analysis. (A) Number of genes (left) and proportion of the original correlation (right) of each positively correlated module (PCM) and negatively correlated module (NCM). (B) Cumulative distribution of distances between the indented drug targets (purple) and the drugs' PCMs (left) and NCMs (right). The distances comprehend the top decile of proximal proteins. The smaller percentile the closer to the module (being 0 if the protein is already in the drug module). Besides, we provide module distances to three different sets of proteins: random proteins from the STRING network (red), random proteins defined as *Tclin* or *Tchem* in the Target Central Resourse Database (see Methods) (green) and random pharmacologic drug targets retrieved from DrugBank (blue). (C) Network-based distances between drug classes. Each dot represents a proximity measure. The bigger and redder the dot, the closer the distance between drug classes. Conversely, the smaller and bluer the dot, the further the distance. Drug classes are sorted from more to less specific proximities. The network-based distance calculation is detailed in Methods. (D) (Left) Distribution of Area Under the ROC curve (AUROC) from the CTRP drug predictions using GDSC drug modules. The different colors show the different cutoff used to nominate a cell "sensitive" to a certain drug. The dashed lines show the predictions using only CCLs not shared with the GDSC panel. The black line indicates the 0.5 AUROC value. (Right) ROC curve predictions for 2 different drugs: Daporinad (FMK866) and Docetaxel. (E) (Left) Number of enriched drug modules with respect to the fraction of TCGA-patients where they are found. (Right) Cumulative distribution of the number of TCGA-patients across the enriched drug modules.

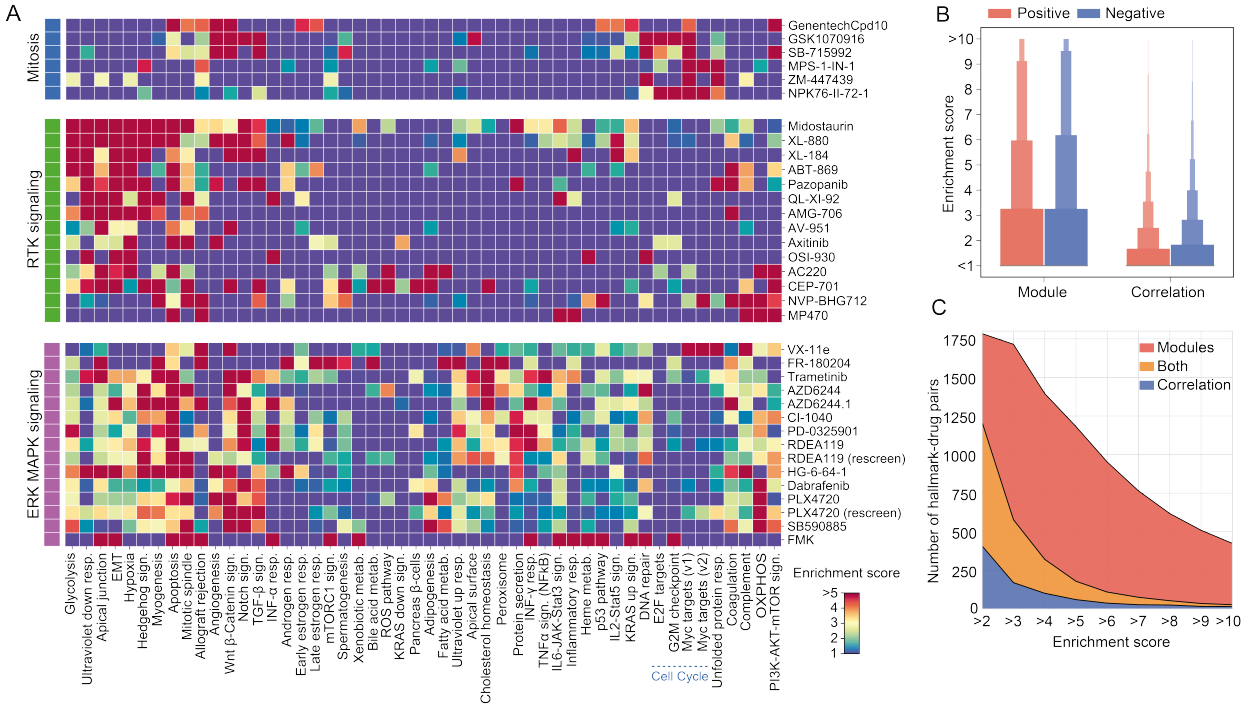

Figure 4. Module enrichment in the Hallmark collection. (A) Heatmap showing the enrichment scores (odds ratio) of the positively correlated modules (y-axis) in the Hallmark gene set collection (x-axis) for 3 different drug classes: Mitosis (blue), RTK signaling (green) and ERK MAPK signaling (purple). Redder colors show higher enrichments. Please note that the maximum enrichment score was set at 5. (B) Distribution of the enrichment scores in the Hallmark collection gene sets. (C) Number of hallmark-drug pairs at different enrichment scores. We show the pairs found only in the modules (red), only in the correlations (blue) or in both (orange).