bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Background: 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 previous knowledge about drug action. In particular, transcriptional signatures of drug sensitivity may guide drug repositioning, prioritize drug combinations and point to new therapeutic biomarkers. However, the inherent complexity of transcriptional signatures, with thousands of differentially expressed genes, makes them hard to interpret, thus giving poor mechanistic insights and hampering translation to clinics.

Methods: To simplify drug signatures, we have developed a network-based methodology to identify functionally coherent gene modules. Our strategy starts with the calculation of drug-gene correlations and is followed by a pathway-oriented filtering and a network-diffusion analysis across the interactome.

Results: We apply our approach to 189 drugs tested in 671 cancer cell lines and observe a connection between gene expression levels of the modules and mechanisms of action of the drugs. Further, we characterize multiple aspects of the modules, including their functional categories, tissue-specificity and prevalence in clinics. Finally, we prove the predictive capability of the modules and demonstrate how they can be used as gene sets in conventional enrichment analyses.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Conclusions: Network biology strategies like module detection are able to digest the outcome of large-scale pharmacogenomic initiatives, thereby contributing to their interpretability and improving the characterization of the drugs screened.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

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 channeled into discovering transcriptional signatures that are specific to a disease state. This work involved the comparison of a relatively small number of diseased and healthy samples [1]. Although providing a rich account of disease biology, these studies have failed to yield better drug therapies, as causality and response to drug perturbations cannot be inferred directly from two-state (diseased vs. healthy) differential gene expression analysis [2, 3]. To address this issue, initiatives have flourished to profile the basal gene expression levels of hundreds of cell lines, together with their response to treatment over an array of drug molecules using a simple readout such as growth rate [4-7]. Provided that the panel of cell lines is large enough, this approach allows for a new type of gene expression analysis where basal expression levels are *correlated* to drug response phenotypes. A series of recent studies demonstrate the value of this strategy for target identification, biomarker discovery, and elucidation of mechanisms of action (MoA) 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 guide drug prescription. However, current signatures are composed of several hundred genes, thereby making them difficult to interpret, harmonize across platforms and translate to clinical practice [14-16]. Recent assessment of sensitivity signatures for over two hundred drugs [9] revealed that key genes include those involved in drug metabolism and transport. Intended therapeutic targets, though important, are detected in only a fraction of signatures, and cell line tissue of origin has been identified as a confounding factor throughout the signature detection procedure. 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 [17].

Network biology offers means to integrate a large amount of omics data [18]. Most 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 [19]. Following this principle, it has been possible to convert genome-wide signatures to network signatures, or *modules*, that are less noisy and easier to interpret [20]. Raphael and co-workers,

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for instance, developed an algorithm to map cancer mutations on biological networks and identify "hot" regions that distinguish functional (driver) mutations from sporadic (passenger) ones [21]. Califano's group combined gene expression data with regulatory cellular networks to infer protein activity [22]. Overall, network-based methods come in many flavors and offer an effective framework to organize the results of omics experiments [23].

While many genes and proteins have enjoyed such a network-based annotation (being circumscribed within well-defined modules such as pathways and biological processes), drug molecules remain mostly uncharacterized in this regard. For a number of drugs the mechanism of action is unclear [3] and off-targets are often discovered [24]. Recent publications of drug screens against cancer cell line panels, and the transcriptional signatures that can be derived from there, provide a broader view of drug activity and enable the full implementation of network biology techniques. Here we undertake the task of obtaining and annotating transcriptional modules related to 189 drugs. We show how these modules are able to capture meaningful aspects of drug biology, being robust to inherent biases caused by, for example, the cell's tissue of origin, and having a tight relationship to mechanisms of action and transportation events occurring at the membrane. Finally, we perform a series of functional enrichment analyses, which contribute to a better understanding of the molecular determinants of drug activity.

Methods

Data preparation and drug-gene correlations. We collected gene expression and drug response data from the GDSC resource (https://www.cancerrxgene.org). We first discarded those genes whose expression levels were low or stable across cell lines (Additional file 1: Fig. S1A). To this end, we analyzed the distribution of basal expression of each gene in every CCL and filtered out those with an expression level below 4.4 (log2 units) across the panel (see Additional file 1: Fig.S1B for a robustness analysis). Regarding drug response data, GDSC provides measurements of cell survival at a range of drug concentrations (Area Under the dose-response Curve, AUC). Since this measure is inversely proportional to drug sensitivity (i.e. the more sensitive the cell, the shorter its survival), we used the 1-AUC as a measure of potency. Thus, *positive* correlations denote drug sensitivity cause by gene over-expression while *negative* correlations indicate that sensitivity is associated with gene under-expression.

Recent studies report a confounding effect of certain tissues in the global analysis of drug-gene correlations [9]. In order to identify these potential biases in our dataset, we performed a principal

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 with gene expression values for each CCL. Finally, we filtered out CCLs belonging to tissues that were strongly correlated to the drug-gene correlation profiles (Additional file 1: Fig. S2A). We removed leukemia, myeloma, lymphoma, neuroblastoma, small cell lung cancer (SCLC) and bone CCLs. In addition, we considered only drugs with sensitivity measurements available for at least 400 CCLs, as recommended by Rees et al. [9].

After this filtering process, we recalculated, for each drug-gene pair, the Pearson's correlation between basal gene expression and 1-AUC drug potencies across CCLs. We applied Fisher's z-transformation to the correlation coefficients in order to account for variation in the number of CCLs available for each drug [25]. Overall, we obtained *positive* and *negative* drug-gene correlations for 217 drugs and 15,944 genes across a total of 671 CCLs. Drug gene correlations (z_{cor}) beyond +/- 3.2 were considered to be significant (Additional file 1: Fig. S1C; Additional file 1: Fig. S1D shows that this cutoff is a robust choice).

Frequently correlated genes. For each gene, we counted the number of correlated drugs (z_{cor} beyond +/- 3.2) and inspected the resulting cumulative distribution (Additional file 1: Fig. S3). Genes at the 5% end of the distribution were considered to be "frequently correlated genes" (FCGs). We found 869 *positive* and 799 *negative* FCGs, which were removed from further analyses. Finally, we performed enrichment analyses on those genes using the Gene Ontology database [26] and the DAVID toolbox (https://david.ncifcrf.gov/summary.jsp) (hypergeometric tests).

Tissue-specific correlations. First, we split the CCL panel into sets of CCLs belonging to the same tissue. We then calculated drug-gene correlations (z_{cor}) separately for each of the 13 tissues represented in our dataset. In order to verify that measures of *positively* correlated genes (PCGs) and *negatively* correlated genes (NCGs) were consistent across tissues, we calculated the median z_{cor} across tissues for each drug-PCG/NCG pair. In general, tissue-specific correlations had the same "direction" (i.e. same sign of z_{cor}) as the global correlation used throughout the study (Additional file 1: Fig. S4A, left panel).

Drug-target correlations. We obtained drug targets from the GDSC resource (disambiguating them with DrugBank [27], when necessary). We assigned at least one target to 202 of the 217 drugs. We focused on the z_{cor} correlation of the targets to check whether target expression (positively) correlates with drug sensitivity. When more than one target was annotated per drug, we kept the maximum correlation. To validate the statistical significance of this measure, we

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

randomly sampled genes (corresponding to the number of known targets per drugs; here again we kept the maximum correlation). This process was repeated 1,000 times for each drug. The mean and the standard deviation of this null distribution were used to derive a z-score, making results comparable between drugs.

Drug module detection. After removing frequently correlated genes from the list of drug-gene correlations, we kept 182 [median; Q1: 84, Q3: 372] *positively* and 122 [median; Q1: 41, Q3: 337] *negatively* correlated genes (PCGs, NCGs) per drug. Further, we used correlation values (z_{cor}) to run a gene-set enrichment analysis (GSEA) [28] for each drug and identify the genes that participate in enriched Reactome pathways [29, 30]. 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). The resulting GSEA-filtered list of genes retained 100 [median; Q1: 49, Q3: 277] *positive* and 77 [median; Q1: 30, Q3: 221] *negative* correlations per drug. Then, taking the z_{cor} values as input scores, we submitted the GSEA-filtered list of genes to HotNet2 [31], using a high-confidence version of STRING [32] (confidence score > 700). We ran HotNet2 iteratively, keeping the largest module and removing its genes for the next iteration, until the modules had fewer than 5 genes or were not statistically significant (p-value > 0.05). To recall strong drug-gene correlations "proximal" to the drug modules (missed, most likely, by the incomplete coverage of Reactome), we used the DIAMOnD module-expansion algorithm [29]. We considered only genes that: (i) were correlated to the drug response; (ii) were not present in any of the Reactome pathways; and (iii) were in the top 200 closest genes to the module, according to DIAMOnD (this cutoff was proposed by the authors of DIAMOnD based on orthogonal functional analyses). Hence, we obtained at least one *positively* correlated module for 175 of the drugs (48 genes [median; Q1: 23, Q3: 83]) and one *negatively* correlated module for 154 of the drugs (40 genes [median; Q1: 21, Q3: 78]). Robustness analysis of this procedure is found in Additional file 1: Fig. S1D. A GMT list of the drug modules can be found in Additional file 2. The correlation values of the genes in the drug modules is available in Additional file 3.

Distances between drug targets and modules. DIAMOnD [29] provides a list of genes sorted by their network-based proximity to the module. Accordingly, we retrieved from the STRING interactome the top closest 1,450 genes (~10% of the largest connected component of the network) for every drug module. We then checked the ranking of drug targets in the resulting DIAMOnD lists, (conservatively) taking the median value when more than one target was available. To assess the proximity of drug targets to the modules, we measured distances to three different sets of random proteins. The first random set corresponded to the STRING proteome.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For the second, we collected all genes defined as *Tclin* or *Tchem* in the Target Central Resource Database [33] (i.e. "druggable proteins"). Finally, the third random set included all pharmacologically active drug targets reported in DrugBank (https://www.drugbank.ca/).

Distances between modules. We calculated distances between *positively* and *negatively* correlated modules separately using the network distance proposed by Menche et al. [34]. This distance measure is sensitive to the number of genes (size) included in the modules. To normalize this measure, we devised the following procedure. First, we grouped drug modules on the basis of their size. Then, for each module, we calculated the distribution of shortest distances from each gene to the most central one [35]. We used this distribution to sample random modules from the network. When the distribution constraint could not be fully met, we used the DIAMOnD algorithm [29] to retrieve the remaining genes (50% of the genes at maximum). We repeated this process to obtain 10 random modules of each size. Next, we distributed the random modules into ranges (intervals) of 5 (i.e. from 10 to 14 genes, from 15 to 19, etc.; 50 random modules per interval). Then, for each pair size, we randomly retrieved 100 pairs of modules and calculated the network-based distance between them. The mean and standard deviation of the distances at each pair size were used to normalize the observed distances, correspondingly (z-score normalization) (we checked that 100 random pairs were sufficient to approximate the mean and standard deviation of the population). The more negative the network distance (d_{net}), the more proximal the modules are. We provide the network distances as an Additional file 4.

Drug response prediction using drug modules. We performed drug response predictions in the GDSC dataset by using drug modules (only first PCMs and NCMs, to make results comparable between drugs). We devised a simple GSEA-like predictor in which CCLs were evaluated for their up/down-regulation of the modules, correspondingly. To this end, we first normalized the expression of each gene across the CCL panel (z-score). Then, for each drug, we ranked CCLs based on the GSEA enrichment scores (ES), taking drug modules as gene sets. To evaluate the ranking, we chose the top 25, 50 and 100 CCLs based on the *known* drug sensitivity profile. Performance was evaluated using the AUROC metric. Results were compared to those obtained with positively and negatively correlated genes (PCG, NCG) from the full signatures (z_{cor} beyond +/- 3.2).

To check whether modules derived from GDSC generalize to other CCL panels, we applied the same procedure to the Cancer Therapeutics Response Portal (CTRP) (https://ocg.cancer.gov/programs/ctd2/data-portal). As done with the GDSC panel, we removed all CCLs derived from neuroblastomas, hematopoietic, bone and small cell lung cancer tissues,

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

leaving a total of 636 CCLs, 397 in common with our GDSC panel (67 drugs in common). Drug response predictions for CTRP were performed as detailed above. We used the best ES among all modules associated with the drug. In addition, we did the analysis using CCLs exclusive to CTRP (i.e. not shared with the GDSC panel).

Module 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). We evaluated each gene set independently using a hypergeometric (Fisher's exact) test (first and second modules were merged, when applicable; the gene universe was that of GDSC). Enrichments can be found in the Additional file 5.

Drug module enrichments in the TCGA cohort. We downloaded gene expression data (median z-scores) for 9,788 patients and 31 cancer tissues from the PanCancer Atlas available in the cBioPortal resource (http://www.cbioportal.org). "Presence" or "expression" of the module in each patient was evaluated using GSEA (P-value < 0.001), ensuring that the direction (up/down) of the enrichment score corresponded to the "direction" of the module (PCM/NCM). For a complete list of enrichment results, please see Additional file 6 (results are organized by tumor type). Further, to identify associations between drug modules and cancer driver genes, we checked whether patients "expressing module of drug X" (P-value < 0.001) were "harboring a mutation in driver gene Y" (Fisher's exact test). We considered 113 driver genes (obtained as described in [36], using the "known" flag) (Additional file 7).

Characterization of drug modules. In order to characterize drug modules from different perspectives, we designed 21 features belonging to the following categories: (i) *General features* derived directly from the pharmacogenomics panel; (ii) *Network features* related to network measures such as topological properties; and (iii) *Biological features* encompassing a series of orthogonal analyses related to drug biology. For more information, please see Additional file 8 and its corresponding legend.

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, GI) for 265 drugs screened against 1,001 cell lines derived from 29 tissues, together with *basal* transcriptional profiles of the cells (among other omics data). Aware of the work by Rees et al. [9], we first looked

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for the dominant effect of certain tissues in determining associations between drug response and gene expression. We found that CCLs derived from neuroblastoma, hematopoietic, bone and small cell lung cancers may confound global studies of drug-gene correlations due to their unspecific sensitivity to drugs (Additional file 1: Fig. S2A). These tissues were excluded 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 217 drugs, 15,944 genes and 671 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 each drug (area over the growth-inhibition curve; 1-AUC) across the CCL panel. We used a z-transformed version of Pearson's ρ, as recommended elsewhere [25]. Figure 1A shows the pair-wise distribution of the z-correlation (z_{cor}) measures between the 15,944 genes and the 217 drugs. We validated the correlations identified in the GDSC panel on an independent set by applying the same protocol to the Cancer Therapeutic Response Portal (CTRP) panel [9] (Additional file 1: Fig. S4B). To identify the strongest drug-gene associations, we set a cutoff of +/- 3.2 z_{cor}, based on an empirical null distribution obtained from randomized data (see Additional file 1: Fig. S1C and Methods). Please note that this is a widely adopted procedure that is not designed to detect *single* drug-gene associations (which would require multiple testing correction) [37]. Instead, and similar to signature identification in differential gene expression analysis, the goal is to identify sets of genes that are (mildly) correlated with drug response. For each drug, we obtained a median (Med) of 249 *positively* correlated genes [first quartile (Q1): 120, third quartile (Q3): 584], and Med of 173 *negatively* correlated genes [Q1: 59, Q3: 484] (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, had hardly a dozen. We observed that the number of genes that correlate with drug response strongly depends on the drug class (Figure 1C), EGFR and ERK-MAPK signaling inhibitors being the classes with the largest number of associated genes, and JNK/p38 signaling and chromatin histone acetylation inhibitors being those with the fewest correlations. This variation may be partially explained by the range of drug potency across the CCL panel, as it is "easier" to detect drug-gene correlations when the drug has a wide sensitivity spectrum (Additional file 1: Fig. S5).

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Similarly, analysis of independent drug-gene correlations suggests that some genes are *positively* correlated to many drugs. For instance, we found 5% of the genes to be associated with more than 10% of the drugs (Figures 1B and Additional file 1: Fig. S3). The transcripts of these "frequent positively correlated genes" are enriched in membrane processes, specifically focal adhesion (P-value < $5.2 \cdot 10^{-12}$) and extracellular matrix (ECM) organization (P-value < $5 \cdot 10^{-16}$), including subunits of integrin, caveolin and platelet-derived growth factors (PDGFs). These genes determine, among others, the activation of Src kinases [38-41]. Overall, ECM proteins are known to play an important role in tumor proliferation, invasion and angiogenesis [42, 43] and are often involved in the upstream regulation of cancer pathways [44] such as PI3K/mTOR [38-40], MAPK [39] and Wnt signaling [45], and in cell cycle and cytoskeleton regulation [46]. It is thus not surprising that ECM genes determine drug response in a rather unspecific manner.

On the other hand, "frequent *negatively* correlated genes" are associated with small molecule metabolism (xenobiotic metabolic processes, P-value < 3.2×10^{-3}). In this group, 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 [47-50]. 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 27 different solute carriers (SLCs) to be *negatively* correlated to the potency of many drugs. Of note, we also found the ABCA1 transporter and other 8 SLCs to be among the frequent *positively* correlated genes, thus emphasizing the key role of transporters and carriers in determining drug potency.

All of the above suggests that 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). However, while these determinants are recognized to play a crucial role, they do not inform targeted therapies, as they are usually unrelated to the mechanism of action of the drug. Thus, we assessed whether measuring drug-gene correlations would also be sufficient to elucidate drug targets, i.e. we tested whether the expression level of the target correlates with the potency of the drug. Since most drugs had more than one annotated target, to measure significance we randomly sampled 1,000 times an equal number of genes and derived an empirical Z-score (Methods). Figure 1D shows that the expression level of most drug targets did *not* correlate with drug response. In fact, only ~10% of the drugs had "positively correlated targets" (Z-score > 1.9, P-value ~ 0.05). Remarkably, the 6 EGF pathway inhibitors in our dataset were among these drugs, as were 3 of the 4 IGF pathway and 3 of the 21 RTK

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

pathway inhibitors. We noticed that the molecular targets for these pathways were usually cell surface receptors, e.g. EGFR, IGFR, ALK, ERBB2, MET and PDGFRA. Overall, of the 20 drugs with positively correlated targets, 13 bind to cell surface receptors, showing a propensity of drug-gene correlations to capture membrane targets (odds ratio = 15.13, P-value = 1.9×10^{-7}). In Additional file 1: Fig. S6 we show how this trend is driven mostly by the over-expression of the target on the cell surface.

The relatively small number of positively correlated targets illustrates how the analysis of expression levels alone is insufficient to reveal MoAs, especially when the drug target is located downstream of the cell surface receptors in a signaling pathway. Some authors have suggested that the tissue of origin of the cells might play a confounding role in defining drug response signatures. To address this notion, we repeated the calculation of Pearson's z_{cor} correlations separately for each of the 13 tissues in our dataset. In general, the trends observed at the tissue 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) (Additional file 1: Fig. S4A, right panel). Accordingly, we confirmed that none of the tissues had a globally dominant effect on the measures of drug-gene correlations (Additional file 1: Fig. S2B) and verified that certain tissue-specific associations were still captured by the analysis. For instance, going back to the targeting of EGFR (which was *positively* correlated with Afatinib and Gefinitib), we show in Figure 1E that the "global" correlation can be partly attributed to non-small cell lung cancer (NSCLC) cells ($z_{cor} > 1.96$, P-value < 0.05). Indeed, Afatinib and Gefitinib have an approved indication for NSCLC. Both drugs correlate with EGFR also in the aerodigestive tract, an observation reported in an independent study dedicated to the discovery of drug-tissue/mutation associations (ACME) [7]. Moreover, and consistent with recent findings [51-54], Gefitinib has a significant correlation to EGFR in breast cancers, whereas Afatinib correlates with this target in pancreatic CCLs. Afatinib, in turn, is associated with ERBB2 in breast CCLs, as also confirmed by ACME analysis (Additional file 1: Fig. S4C).

From drug-gene correlations to drug modules

The previous analysis demonstrates that conventional drug-gene correlations do *not* directly identify drug targets, and suggests that standard transcriptional 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 to make them more functionally homogeneous, and (ii) allow for the measurement of network distances so that genes

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

proximal to the target can be captured and connected to it, even if the expression of the target itself is not statistically associated with 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). In the Methods section, we explain in detail the module detection procedure. In brief, starting from drug-gene correlations (Figure 2A), we first filtered out those genes whose expression was frequently (and unspecifically) correlated to the potency of many drugs (Additional file 1: Fig. S3). This reduced the number of associations to 182 [median; Q1: 84, Q3: 372] *positively* and 122 [median; Q1: 41, Q3: 337] *negatively* correlated genes per drug, respectively. Next, in order to identify genes acting in coordination (i.e. participating in enriched Reactome pathways [29, 30]), we adapted the gene set enrichment analysis (GSEA) algorithm [28] to handle drug-gene correlations (instead of gene expression fold-changes) (Figure 2B). The resulting GSEA-filtered list of genes kept 100 [median; Q1: 49, Q3: 277] *positive* and 77 [median; Q1: 30, Q3: 221] *negative* correlations per drug. After this filtering, we submitted this list to HotNet2 [31], a module detection algorithm that was originally developed for the identification of recurrently mutated subnetworks in cancer patients (Figure 2C; Additional file 1: Fig. S7 shows the importance of the Reactome-based filtering previous to HotNet2). As a reference network (interactome) for HotNet2, we chose a high-confidence version of STRING [32], composed of 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 the DIAMOnD module expansion algorithm [29] to recover strong drug-gene correlations that had been discarded along the process. Although this step made a relatively minor contribution to the composition of the modules (less than 5% of the genes; Additional file 1: Fig. S8), we did not want to lose any strong association cause by the limited coverage of the Reactome database (Figure 2D).

Our pipeline yielded at least one "*positively* correlated module" (PCM) for 175 of the 217 drugs (48 genes [median; Q1: 23, Q3: 83]). Similarly, we obtained "*negatively* correlated modules" (NCMs) for 154 of the drugs (40 genes [median; Q1:21, Q3:78]). Thus, compared to the original signatures, drug modules are considerably smaller (80% reduction) (Figure 3A) and are commensurate with manually annotated pathways in popular databases (Additional file 1: Fig. S9). For roughly two thirds of the drugs, we obtained only one PCM and one NCM. For the remaining drugs, a second (usually smaller) module was also identified (Additional file 1: Fig. S10A). The complete list of drug modules can be found in Additional file 2. Pair-wise drug-gene

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

correlations of the modules are listed as Additional file 3. Additionally, the code of the module-detection pipeline is available at: https://github.com/sbnb-lab-irb-barcelona/GDSC-drug-modules.

Drug modules are tightly related to mechanisms of action

To assess the mechanistic relevance of drug modules, we measured their distance to the corresponding drug targets, i.e. we formulated the hypothesis that drug targets should be "proximal" to dysregulated network regions. To this end, we used the DIAMOnD algorithm again [29], this time to retrieve, for each drug, a list of genes ranked by their proximity to the corresponding drug module(s) (Methods). Figure 3B shows that drug targets are remarkably up-ranked in these lists, making them closer to the drug modules than any other set of random proteins, including druggable genes and pharmacological receptors [33], 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 proximal proteins (top decile), which means a dramatic increase in mechanistic interpretability compared to the 12.25% of drugs that could be linked to their targets via conventional analysis of drug-gene correlations.

A unique feature of drug modules is that network-based distances can be natively measured between them [34]. We computed the distance between drug modules pair-wise (Additional file 4) and grouped them by drug class (Figure 3C) (see Methods and alternative statistical treatments in Additional file 1: Fig. S11). The diagonal of Figure 3C clearly indicates that drugs belonging to the same category tend to have "proximal" modules (some of them in a highly specific manner, like in the case of ERK-MAPK signaling cascade inhibitors). Most interestingly, we could observe proximities between modules belonging to different drug classes. For instance, modules of drugs targeting RTK signaling were "located" near to those of drugs affecting genome integrity, in good agreement with recently reported cross-talk between these two processes [55, 56]. Likewise, and as proposed by some studies [57-59], IGFR-related drugs were "proximal" to drugs affecting cell replication events such as mitosis, cell cycle and DNA replication.

Drug modules retain the ability to predict drug response

We have shown that drug modules are related to the MoA of the drug, but the question remains as to the extent to which they retain the predictive capabilities of the full transcriptional profiles/signatures. In the CCL setting, gene expression profiles are valuable predictors of drug response [5, 11, 60] and crucially contribute to state-of-the-art pharmacogenomic models. To test whether our (much smaller) drug modules retained predictive power, we devised a simple drug sensitivity predictor based on the GSEA score (see Methods). In brief, given a drug, we tested

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

whether cell lines *sensitive* to a certain drug were enriched in the corresponding drug modules. We expect genes in PCMs to be *over*-expressed in sensitive cell lines and those in NCMs to be *under*-expressed. Analogously, we took the *positively* and *negatively* correlated genes from the full drug-gene associations (signatures), and also performed a GSEA-based prediction. 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, based on the distribution of sensitive cell lines provided by the authors of the GDSC (Additional file 1: Fig. S12A). This simple binarization is, in practice, proportional to more sophisticated "sensitive/resistant" categorizations such as the waterfall analysis [14], and it yields prediction performance metrics comparable between drugs.

Additional file 1: Fig. S13 suggests that, when applied to the GDSC, drug module enrichment analysis can classify sensitive cell lines with high accuracy, especially for the top 25 sensitive cell lines (area under the ROC curve (AUROC): 0.77), which is a notable achievement considering that drug modules are 80% smaller than the original signatures. To assess the applicability of our modules outside the GDSC dataset, we performed an external validation with the CTRP panel of cell lines. About 37% of our drugs were also tested in this panel. In CTRP, drug sensitivity is measured independently of GDSC, which poses an additional challenge for prediction as a result of experimental inconsistencies [14]. Of the CCLs, 397 are shared between GDSC and CTRP, and gene expression data are also measured independently. We performed the GSEA-based sensitivity prediction for *all* CTRP CCLs. Figures 3D and E show the distribution of prediction performances for the 70 drugs, and illustrative ROC curves corresponding to four drugs (namely Daporinad, Vorinostat, I-BET-762 and Docetaxel), respectively. We found that, when focusing on the top 25 sensitive CCLs, over a quarter of the drugs had AUROC > 0.7, including Daporinad. Acceptable (AUROC > 0.6) predictions were achieved for half the cases (e.g. Vorinostat and I-BET-762), which is a comparable result to recent attempts to translate sensitivity predictors between different CCL panels [61]. For the remaining drugs, predictive performance did not differ to random expectation (AUROC < 0.6) (e.g. Docetaxel). Notably, performance declined only slightly when considering CCLs that were *exclusive* to the CTRP panel (i.e. not part of the GDSC dataset) (Figure 3D, blue boxes). The figure was comparable, if not better, to that obtained using full signatures (PCGs and NCGs) (Additional file 1: Fig. S13, gray boxes). These observations support previous recommendations to pre-filter pharmacogenomic data based on prior knowledge [62] (Additional file 1: Fig. S14).

Module-based characterization of drugs

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Since drug modules are highly connected in biological networks, they are expected to be (at least to some extent) functionally coherent and easier to interpret. Accordingly, we tested the enrichment of drug modules in a collection of high-order biological processes (the Hallmark gene sets) available from the Molecular Signatures Database (MSigDB) [63]. Additional file 1: Fig. S15A shows that the number of enriched Hallmark gene sets depends upon the MoA of the drug. The results of the enrichment analysis are given in Additional file 5 and as an interactive exploration tool based on the CLEAN methodology (Additional file 9; https://figshare.com/s/932dd94520d4a60f076d) [64]. We chose three drug classes to illustrate how to read these results, namely drugs targeting mitosis, RTK signaling inhibitors, and ERK-MAPK signaling inhibitors (Figure 4A).

Drugs targeting mitosis have modules enriched in cell cycle and replication processes (Figure 4A, top). Specifically, genes related to the Myc transcription factor are over-represented in three of the drug modules (NPK76-II-72-1, GSK1070916 and MPS-1-IN-1). The modules of these drugs have a rather distinct composition, NPK76-II-72-1 having the largest coverage of Myc-related genes and being, together with MPS-1-IN-1, related to both Myc1 and Myc2 processes. In Additional file 1: Fig. S15B we show how, for these two drugs, cell line sensitivity is dependent on Myc expression levels.

In contrast to mitosis inhibitors, drugs targeting the RTK pathway are enriched in biological processes outside the nucleus (Figure 4A, middle), among these hypoxia and the epithelial-mesenchymal transition (EMT). Both mechanisms are known to be associated with tyrosine kinases [65, 66]. Interestingly, a subgroup of RTK inhibitors (namely ACC220, CEP-701, NVP-BHG712 and MP470) is characteristically associated with the PI3K-AKT-mTOR signaling cascade. With the exception of NVP-BHG712, these inhibitors have the tyrosine kinase FLT3 as a common target [67, 68]. Deeper inspection of FLT3 inhibitors reveals module proximities to certain PI3K inhibitors (e.g. GDC0941), and the PI3K-AKT-mTOR pathway is enriched in ERBB2 inhibitors as well (Additional file 4 and 5).

As for ERK-MAPK pathway inhibitors, we observed a total of 17 enriched Hallmarks, making this class of drugs the one with most variability in terms of enrichment signal of the modules (Figure 4A, bottom; Additional file 1: Fig. S15A). However, while some processes like apoptosis are detected in most of the drugs in this category, others are target-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 [69, 70].

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [71-74], and Additional file 1: Fig. S15C demonstrates that the variability observed between drugs in this class is driven mostly by differences in the sensitivity profiles of the drugs.

Overall, the enrichment signal (i.e. the functional coherence) of drug modules is substantially higher than that of full signatures (PCGs and NCGs) (Figure 4B,C). This facilitates, in principle, the mechanistic interpretation of drug-gene correlation results (Additional file 1: Fig. S15D). We show an illustrative module (CEP-701) in Figure 4D.

We next examined whether our results could be extended beyond CCL panels. We found that drug modules are indeed identified (GSEA P-value < 0.001) in the majority of patients in the TCGA clinical cohort (Additional file 1: Fig. S15E; see Methods for details). Closer inspection by TCGA tumor type further supports the clinical relevance of our results (Additional file 6). For example, drugs affecting MAPK signaling (specifically, BRAF inhibitors, e.g. Dabrafenib) have a tendency to “occur” in skin cutaneous melanomas (SKCM), as expected (Figure 4E, blue). Of note, one PPAR inhibitor (FH535) was also found enriched in a high number of SKCM patients, in good agreement with work by others proposing the use PPAR inhibitors to treat skin cancer [75, 76]. Similarly, we observed an abundance of EGFR inhibitor modules among pancreatic cancers (PAAD) (Figure 4E, green), in line with the known crucial role of EGFR in pancreatic tumorigenesis [77, 78]. As for gliobastomas (GBMs) (Figure 4E, purple), we found two GSK3 inhibitors (CHIR-99021 and SB216763) and one TNKS inhibitor (XAV939), all of them targeting WNT signaling, which is potential mechanism against this tumor type [79]. We also found one EGFR inhibitor (Gefitinib) and the PLK inhibitor NPK76-II-72-1 mentioned above in the context of Myc enrichment analysis. Both mechanisms have shown promise in EGFR- and Myc-activated gliomas, respectively [80, 81]. Finally, we encountered a more heterogeneous pattern in breast cancer patients (BRCA) (Figure 4E, orange), including mechanisms supported by the literature, such as AKT, IRAK1 and PLK3 inhibition [82-84].

Beyond the tumor-type level, we looked for modules that were significantly enriched (odds ratio > 2, P-value < 0.001) in patients harboring specific driver mutations (Methods). A full account of this enrichment analysis is given in Additional file 7. We found, for instance, that modules of drugs targeting ERK/MAPK signaling are related to patients with mutations in HRAS and BRAF [85, 86], and that, in turn, BRAF is (together with KRAS) frequently mutated in patients “expressing”

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

modules of EGFR signaling inhibitors [87]. Taken together, and although TCGA treatment response data is too scarce to allow for prediction assessment [88], these results indicate that the drug modules identified in CCLs hold promise for translation to clinical practice.

Conclusions

Two limitations of large-scale pharmacogenomic studies are the difficulty to reproduce results across screening platforms and the eventual translation to clinics, as it remains unclear whether immortalized cells are able to model patient samples [89]. Another important limitation is the overwhelming number of drug-gene correlations that can be derived from these experiments, yielding signatures of drug sensitivity that are almost impossible to interpret. We have shown, for example, that (i) the number of correlated genes is highly variable across drugs; (ii) some genes are unspecifically correlated to many drugs; and (iii) not all drug-gene pairs are equally correlated in every tissue. We propose that converting transcriptional signatures to network modules may simplify the analysis, since network modules are smaller, more robust and functionally coherent. We have validated this strategy by proving that drug response modules, which are enriched in biological processes of pharmacological relevance and exhibit comparable predictive power to the full signatures, are tightly related to the MoA. Further, we have characterized the modules extensively (Additional file 8 and e.g. Additional file 1: Fig. S16) and confirmed their occurrence in the TCGA clinical cohort (Additional file 6).

However, our approach does have some of the limitations of ordinary transcriptomic analyses. Expression levels of mRNA do not perfectly match protein abundance, nor are they able to capture post-translational modifications such as phosphorylation events, which are key to some of the pathways studied here. Moreover, wide dynamic ranges in gene expression and drug sensitivity data are necessary for drug-gene correlations to be captured, thus 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 transcriptional traits that may be irrelevant to drug response. Our module-based approach partially corrects for this confounding factor, although the integration of other CCL omics data (such as mutations, copy number variants and chromatin modifications) could further ameliorate these issues and also provide new mechanistic insights. In this context, systems biology tools that learn the relationships between different layers of biology are needed. Along this line, the release of CCL screens with readouts other than growth inhibition or proliferation rate [90, 91] will help unveil the connections between the genetic background of the cells and the phenotypic outcome of drug treatment.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

All in all, 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 [92]. The L1000 Next-generation Connectivity Map, for instance, contains about one million post-treatment gene expression signatures for twenty thousand molecules [90]. These signatures await to be interpreted and annotated and, more importantly, they have to be associated with pre-treatment signatures in order to identify therapeutic opportunities. We believe that network biology strategies like the one presented here will enable this connection, encircling relevant "regions" of the signatures and measuring the distances between them.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Gene expression, drug response and drug target datasets are available in the cancerxxgene resource (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html). External validation data for drug response predictions are available in the ctd2 repository (https://ocg.cancer.gov/programs/ctd2/data-portal). Hallmark gene sets were obtained from the Molecular Signature Database (MSigDB) (http://software.broadinstitute.org/gsea/index.jsp). TCGA gene expression data were downloaded from cbioportal (http://www.cbioportal.org; "PanCancer Atlas" flag). TCGA cancer drivers are available in the OncoGenomic Landscapes resource (https://oglandscapes.irbbarcelona.org/). All data generated or analyzed during this study are included in this published article and its additional files. Code for the module detection pipeline is available at https://github.com/sbnb-lab-irb-barcelona/GDSC-drug-modules.

Competing interests

The authors declare that they have no competing interests.

Funding

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A.F-T. is a recipient of an FPI fellowship. P.A. acknowledges the support of the Spanish Ministerio de Economía y Competitividad (BIO2016-77038-R) and the European Research Council (SysPharmAD: 614944).

Author contributions

A.F-T., M.D.F. and P.A. conceived the study. A.F-T. and M.D-F. performed the analyses. A.F-T., M.D.F. and P.A. interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Prof. Ben Raphael and Dr. Matthew Reyna (Princeton University) for guidance in the use of HotNet2.

Additional files

Additional file 1. Contains Supplementary Figures 1-16 (PDF).

Additional file 2. Collection of drug modules in GMT format. The first column indicates the name of the drug while the second column indicates whether the module is a secondary module ("second_module") or not ("na"). From the third column onwards, there are the genes composing the module (gene names). Excel file (XLSX).

Additional file 3. Drug module – gene correlations across tissues. Excel file (XLSX).

Additional file 4. Pair-wise distances between drug modules. Network distances (d_{net}) are normalized (z-scores): negative values denote proximity. Secondary modules receive with the suffix "_md2". See Methods for a detailed explanation of the network distance measurement. Excel file (XLSX).

Additional file 5. Enrichment scores and p-values between drug modules (rows) and Hallmark gene sets (columns). For simplicity, secondary modules were merged with the main ones. Excel file (XLSX).

Additional file 6. Enriched (p-value < 0.001) drug module count across 31 TCGA cancer types, i.e. number of patients where each module is "expressed". Excel file (XLSX).

Additional file 7. Cancer driver and drug module associations (OR > 2, p-value < 0.001), based on patients "expressing/not-expressing" a module and "having/not-having" a driver mutation in the TCGA cohort. Excel file (XLSX).

Additional file 8. We have chosen 21 features from network-based measures and other functional data: (i) **General features** *(columns 2-9)*. They illustrate basic and general features derived from the omics panel.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We provide, for instance, the number of genes in each module, the average correlation among module genes and a measure of how "unique" are those genes with respect to other modules. Besides, we annotate drug classes and the AUROC predictions in both the GDSC and CTRP panels. (ii) **Network features** *(columns 10-12)*. These include distances between module genes and drug targets, "connectivity" within module genes (i.e. distance between them), and proximity to genes from other modules. (iii) **Biological features** *(columns 12-21)*. A series of biological features related to drug biology. We give most of them as simple proportions of genes/proteins. Among others, we provide the cellular compartmentalization of the genes, transcription factor specificity and the proportion of disease-related and "druggable" genes inside the module. Annotated drug targets are listed as well. Excel file (XLSX).

Additional file 9. CLEAN cluster results using drug module genes and Hallmark gene sets. We provide an additional table with the significant associations between gene clusters and hallmark gene sets. Compressed folder (ZIP). The file can be found at https://figshare.com/s/932dd94520d4a60f076d

Additional file 10. Peer Review file (PDF).

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Nevins JR, Potti A: **Mining gene expression profiles: expression signatures as cancer phenotypes**. *Nat Rev Genet* 2007, **8**(8):601.
2. Pammolli F, Magazzini L, Riccaboni M: **The productivity crisis in pharmaceutical R&D**. *Nat Rev Drug Discov* 2011, **10**(6):428-438.
3. Swinney DC, Anthony J: **How were new medicines discovered?** *Nat Rev Drug Discov* 2011, **10**(7):507-519.
4. Monks A, Scudiero D, Skehan P, Shoemaker R, Paull K, Vistica D, Hose C, Langley J, Cronise P, Vaigro-Wolff A: **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):757-766.
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D *et al*: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity**. *Nature* 2012, **483**(7391):603-607.
6. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR *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-D961.
7. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, Jones V, Bodycombe NE, Soule CK, Gould J *et al*: **Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset**. *Cancer Discov* 2015, **5**(11):1210-1223.
8. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H *et al*: **A Landscape of Pharmacogenomic Interactions in Cancer**. *Cell* 2016, **166**(3):740-754.
9. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, Javaid S, Coletti ME, Jones VL, Bodycombe NE *et al*: **Correlating chemical sensitivity and basal gene expression reveals mechanism of action**. *Nat Chem Biol* 2016, **12**(2):109-116.
10. Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, Bhutra S, Grossman RL, Stephanie Huang R: **Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies**. *Genome Res* 2017, **27**(10):1743-1751.
11. Lee S-I, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J *et al*: **A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia**. *Nat Commun* 2018, **9**(1):42.
12. Liu X, Yang J, Zhang Y, Fang Y, Wang F, Wang J, Zheng X, Yang J: **A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia**. *Sci Rep* 2016, **6**:22811.
13. Niepel M, Hafner M, Duan Q, Wang Z, Paull EO, Chung M, Lu X, Stuart JM, Golub TR, Subramanian A *et al*: **Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling**. *Nat Commun* 2017, **8**(1):1186.
14. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, Quackenbush J: **Inconsistency in large pharmacogenomic studies**. *Nature* 2013, **504**(7480):389-393.
15. Stransky NaG, M. and Kryukov, G. V. and Garraway, L. A. and Lehar, J. and Liu, M. and Sonkin, D. and Kauffmann, A. and Venkatesan, K. and Edelman, E. J. and Riester, M. and Barretina, J. and Caponigro, G. and Schlegel, R. and Sellers, W. R. and Stegmeier, F. and Morrissey, M. and Amzallag, A. and Pruteanu-Malinici, I. and Haber, D. A. and Ramaswamy, S. and Benes, C. H. and Menden, M. P. and Iorio, F. and Stratton, M. R. and McDermott, U. and Garnett, M. J. and Saez-Rodriguez, J.: **Pharmacogenomic agreement between two cancer cell line data sets**. *Nature* 2015, **528**(7580):84--87.
16. Geeleher PaG, E. R. and Seoighe, C. and Cox, N. J. and Huang, R. S.: **Consistency in large pharmacogenomic studies**. *Nature* 2016, **540**(7631):E1-E2.
17. Garnett MJaE, E. J. and Heidorn, S. J. and Greenman, C. D. and Dastur, A. and Lau, K. W. and Greninger, P. and Thompson, I. R. and Luo, X. and Soares, J. and Liu, Q. and Iorio, F. and Surdez, D. and Chen, L. and Milano, R. J. and Bignell, G. R. and Tam, A. T. and Davies, H. and Stevenson, J. A. and Barthorpe, S. and Lutz, S. R. and Kogera, F. and Lawrence, K. and McLaren-Douglas, A. and Mitropoulos, X. and Mironenko, T. and Thi, H. and Richardson, L. and Zhou, W. and Jewitt, F. and Zhang, T. and O'Brien, P. and Boisvert, J. L. and Price, S. and Hur, W. and Yang, W. and Deng, X. and Butler, A. and Choi, H. G. and Chang, J. W. and Baselga, J. and Stamenkovic, I. and Engelman, J. A. and Sharma, S. V. and Delattre, O. and Saez-Rodriguez, J. and Gray, N. S. and Settleman, J. and Futreal, P. A. and Haber, D. A. and Stratton, M. R. and Ramaswamy, S. and McDermott, U. and Benes, C. H. : **Systematic identification of genomic markers of drug sensitivity in cancer cells**. *Nature* 2012, **483**:570--575.
18. Schadt EE: **Molecular networks as sensors and drivers of common human diseases**. *Nature* 2009, **461**(7261):218-223.
19. Kim Y-A, Przytycka TM: **Bridging the Gap between Genotype and Phenotype via Network Approaches**. *Front Genet* 2012, **3**:227.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

20. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**(6761supp):C47-C52.
21. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer**. *J Comput Biol* 2011, **18**(3):507-522.
22. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A: **Functional characterization of somatic mutations in cancer using network-based inference of protein activity**. *Nat Genet* 2016, **48**(8):838-847.
23. Barabási A-L, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease**. *Nat Rev Genet* 2011, **12**(1):56-68.
24. 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.
25. Dančík V, Carrel H, Bodycombe NE, Seiler KP, Fomina-Yadlin D, Kubicek ST, Hartwell K, Shamji AF, Wagner BK, Clemons PA: **Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses**. *J Biomol Screen* 2014, **19**(5):771-781.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
27. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration**. *Nucleic Acids Res* 2006, **34**(Database issue):D668-672.
28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *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):15545-15550.
29. 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.
30. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B *et al*: **The Reactome Pathway Knowledgebase**. *Nucleic Acids Res* 2018, **46**(D1):D649-D655.
31. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M *et al*: **Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes**. *Nat Genet* 2015, **47**(2):106-114.
32. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P *et al*: **The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible**. *Nucleic Acids Res* 2016, **45**(D1):D362-D368.
33. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J *et al*: **Unexplored therapeutic opportunities in the human genome**. *Nat Rev Drug Discov* 2018.
34. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L: **Disease networks. Uncovering disease-disease relationships through the incomplete interactome**. *Science* 2015, **347**(6224):1257601.
35. Freeman LC: **Centrality in social networks conceptual clarification**. *Soc Networks* 1978, **1**(3):215-239.
36. Mateo L, Guitart-Pla O, Duran-Frigola M, Aloy P: **Exploring the OncoGenomic Landscape of cancer**. *Genome Medicine* 2018, **10**(1):61.
37. Smirnov PaK, V. and Maru, A. and Freeman, M. and Ho, C. and El-Hachem, N. and Adam, G. A. and Ba-Alawi, W. and Safikhani, Z. and Haibe-Kains, B.: **PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies**. *Nucleic Acids Res* 2018, **46**(D1):D994-D1002.
38. Gilcrease MZ: **Integrin signaling in epithelial cells**. *Cancer Lett* 2007, **247**(1):1-25.
39. Givant-Horwitz V, Davidson B, Reich R: **Laminin-induced signaling in tumor cells**. *Cancer Lett* 2005, **223**(1):1-10.
40. 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):632-636.
41. 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):5908-5915.
42. Patarroyo M, Tryggvason K, Virtanen I: **Laminin isoforms in tumor invasion, angiogenesis and metastasis**. *Semin Cancer Biol* 2002, **12**(3):197-207.
43. Keely P, Parise L, Juliano R: **Integrins and GTPases in tumour cell growth, motility and invasion**. *Trends Cell Biol* 1998, **8**(3):101-106.
44. 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):139-151.
45. Du J, Zu Y, Li J, Du S, Xu Y, Zhang L, Jiang L, Wang Z, Chien S, Yang C: **Extracellular matrix stiffness dictates Wnt expression through integrin pathway**. *Sci Rep* 2016, **6**:20395.
46. Moreno-Layseca P, Streuli CH: **Signalling pathways linking integrins with cell cycle progression**. *Matrix Biol* 2014, **34**:144-153.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

47. Lee CA, Neul D, Clouser-Roche A, Dalvie D, Wester MR, Jiang Y, Jones JP, 3rd, Freiwald S, Zientek M, Totah RA: **Identification of novel substrates for human cytochrome P450 2J2**. *Drug Metab Dispos* 2010, **38**(2):347-356.
48. 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):103-141.
49. 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.
50. Hayes JD, Flanagan JU, Jowsey IR: **Glutathione transferases**. *Annu Rev Pharmacol Toxicol* 2005, **45**:51-88.
51. Green MD, Francis PA, Gebski V, Harvey V, Karapetis C, Chan A, Snyder R, Fong A, Basser R, Forbes JF *et al*: **Gefitinib treatment in hormone-resistant and hormone receptor-negative advanced breast cancer**. *Ann Oncol* 2009, **20**(11):1813-1817.
52. Zhang X, Zhang B, Liu J, Liu J, Li C, Dong W, Fang S, Li M, Song B, Tang B *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.
53. 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):371-381.
54. Ioannou N, Dalgleish AG, Seddon AM, Mackintosh D, Guertler U, Solca F, Modjtahedi H: **Anti-tumour activity of afatinib, an irreversible ErbB family blocker, in human pancreatic tumour cells**. *Br J Cancer* 2011, **105**(10):1554-1562.
55. Mahajan K, Mahajan NP: **Cross talk of tyrosine kinases with the DNA damage signaling pathways**. *Nucleic Acids Res* 2015, **43**(22):10588-10601.
56. Chen M-K, Hung M-C: **Regulation of therapeutic resistance in cancers by receptor tyrosine kinases**. *Am J Cancer Res* 2016, **6**(4):827-842.
57. Ish-Shalom D, Christoffersen CT, Vorwerk P, Sacerdoti-Sierra N, Shymko RM, Naor D, DeMeyts P: **{{M}itogenic properties of insulin and insulin analogues mediated by the insulin receptor}**. *Diabetologia* 1997, **40 Suppl 2**:25-31.
58. Borisov N, Aksamitiene E, Kiyatkin A, Legewie S, Berkhout J, Maiwald T, Kaimachnikov NP, Timmer J, Hoek JB, Kholodenko BN: **Systems-level interactions between insulin–EGF networks amplify mitogenic signaling**. *Molecular Systems Biology* 2009, **5**(1):256.
59. Mairet-Coello G, Tury A, DiCicco-Bloom E: **Insulin-Like Growth Factor-1 Promotes G₁/S Cell Cycle Progression through Bidirectional Regulation of Cyclins and Cyclin-Dependent Kinase Inhibitors via the Phosphatidylinositol 3-Kinase/Akt Pathway in Developing Rat Cerebral Cortex**. *The Journal of Neuroscience* 2009, **29**(3):775.
60. 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.
61. Juan-Blanco T, Duran-Frigola M, Aloy P: **Rationalizing Drug Response in Cancer Cell Lines**. *J Mol Biol* 2018.
62. Ferranti D, Krane D, Craft D: **The value of prior knowledge in machine learning of complex network systems**. *Bioinformatics* 2017, **33**(22):3610-3618.
63. 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):417-425.
64. Freudenberg JM, Joshi VK, Hu Z, Medvedovic M: **CLEAN: CLustering Enrichment ANalysis**. *BMC Bioinformatics* 2009, **10**(1):234.
65. 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**:101-114.
66. Thiery JP: **Epithelial–mesenchymal transitions in development and pathologies**. *Curr Opin Cell Biol* 2003, **15**(6):740-746.
67. Lindblad O, Cordero E, Puissant A, Macaulay L, Ramos A, Kabir NN, Sun J, Vallon-Christersson J, Haraldsson K, Hemann MT *et al*: **Aberrant activation of the PI3K/mTOR pathway promotes resistance to sorafenib in AML**. *Oncogene* 2016, **35**(39):5119-5131.
68. Nogami A, Oshikawa G, Okada K, Fukutake S, Umezawa Y, Nagao T, Kurosu T, Miura O: **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):9189-9205.
69. Haq R, Shoag J, Andreu-Perez P, Yokoyama S, Edelman H, Rowe GC, Frederick DT, Hurley AD, Nellore A, Kung AL *et al*: **Oncogenic BRAF regulates oxidative metabolism via PGC1α and MITF**. *Cancer Cell* 2013, **23**(3):302-315.
70. Haq R, Fisher DE, Widlund HR: **Molecular pathways: BRAF induces bioenergetic adaptation by attenuating oxidative phosphorylation**. *Clin Cancer Res* 2014, **20**(9):2257-2263.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

71. 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):1965-1975.
72. 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.
73. Moens U, Kostenko S, Sveinbjørnsson B: **The Role of Mitogen-Activated Protein Kinase-Activated Protein Kinases (MAPKAPKs) in Inflammation**. *Genes* 2013, **4**(2):101-133.
74. Cho Y-Y, He Z, Zhang Y, Choi HS, Zhu F, Choi BY, Kang BS, Ma W-Y, Bode AM, Dong Z: **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):3596-3603.
75. Schadendorf D: **Peroxisome Proliferator-Activating Receptors: A New Way to Treat Melanoma?** *Journal of Investigative Dermatology* 2009, **129**(5):1061-1063.
76. Borland MG, Kehres EM, Lee C, Wagner AL, Shannon BE, Albrecht PP, Zhu B, Gonzalez FJ, Peters JM: **Inhibition of tumorigenesis by peroxisome proliferator-activated receptor (PPAR)-dependent cell cycle blocks in human skin carcinoma cells**. *Toxicology* 2018, **404-405**:25-32.
77. Ardito Christine M, Grüner Barbara M, Takeuchi Kenneth K, Lubeseder-Martellato C, Teichmann N, Mazur Pawel K, DelGiorno Kathleen E, Carpenter Eileen S, Halbrook Christopher J, Hall Jason C *et al*: **EGF Receptor Is Required for KRAS-Induced Pancreatic Tumorigenesis**. *Cancer Cell* 2012, **22**(3):304-317.
78. Tzeng C-WD, Frolov A, Frolova N, Jhala NC, Howard JH, Buchsbaum DJ, Vickers SM, Heslin MJ, Arnoletti JP: **Epidermal growth factor receptor (EGFR) is highly conserved in pancreatic cancer**. *Surgery* 2007, **141**(4):464-469.
79. Lee Y, Lee J-K, Ahn SH, Lee J, Nam D-H: **WNT signaling in glioblastoma and therapeutic opportunities**. *Laboratory Investigation* 2015, **96**:137.
80. Rao SK, Edwards J, Joshi AD, Siu IM, Riggins GJ: **A survey of glioblastoma genomic amplifications and deletions**. *Journal of Neuro-Oncology* 2010, **96**(2):169-179.
81. Higuchi F, Fink AL, Kiyokawa J, Miller JJ, Koerner MVA, Cahill DP, Wakimoto H: **PLK1 Inhibition Targets Myc-Activated Malignant Glioma Cells Irrespective of Mismatch Repair Deficiency–Mediated Acquired Resistance to Temozolomide**. *Molecular Cancer Therapeutics* 2018, **17**(12):2551.
82. Paplomata E, O'Regan R: **The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers**. *Therapeutic Advances in Medical Oncology* 2014, **6**(4):154-166.
83. Wee ZN, Yatim SMJM, Kohlbauer VK, Feng M, Goh JY, Bao Y, Lee PL, Zhang S, Wang PP, Lim E *et al*: **IRAK1 is a therapeutic target that drives breast cancer metastasis and resistance to paclitaxel**. *Nature Communications* 2015, **6**:8746.
84. Fallah Y, Brundage J, Allegakoen P, Shajahan-Haq AN: **MYC-Driven Pathways in Breast Cancer Subtypes**. *Biomolecules* 2017, **7**(3).
85. Van Aelst L, Barr M, Marcus S, Polverino A, Wigler M: **Complex formation between RAS and RAF and other protein kinases**. *Proceedings of the National Academy of Sciences* 1993, **90**(13):6213.
86. Moodie SA, Willumsen BM, Weber MJ, Wolfman A: **Complexes of Ras.GTP with Raf-1 and mitogen-activated protein kinase kinase**. *Science* 1993, **260**(5114):1658.
87. Fitzgerald TL, Lertpiriyapong K, Cocco L, Martelli AM, Libra M, Candido S, Montalto G, Cervello M, Steelman L, Abrams SL *et al*: **Roles of EGFR and KRAS and their downstream signaling pathways in pancreatic cancer and pancreatic cancer stem cells**. *Advances in Biological Regulation* 2015, **59**:65-81.
88. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV *et al*: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics**. *Cell* 2018, **173**(2):400-416.e411.
89. Jaeger S, Duran-Frigola M, Aloy P: **Drug sensitivity in cancer cell lines is not tissue-specific**. *Molecular Cancer* 2015, **14**(1):40.
90. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK *et al*: **A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles**. *Cell* 2017, **171**(6):1437-1452.e1417.
91. Wawer MJ, Li K, Gustafsdottir SM, Ljosa V, Bodycombe NE, Marton MA, Sokolnicki KL, Bray M-A, Kemp MM, Winchester E *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):10911-10916.
92. Frumkin I, Schirman D, Rotman A, Li F, Zahavi L, Mordret E, Asraf O, Wu S, Levy SF, Pilpel Y: **Gene Architectures that Minimize Cost of Gene Expression**. *Mol Cell* 2017, **65**(1):142-153.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

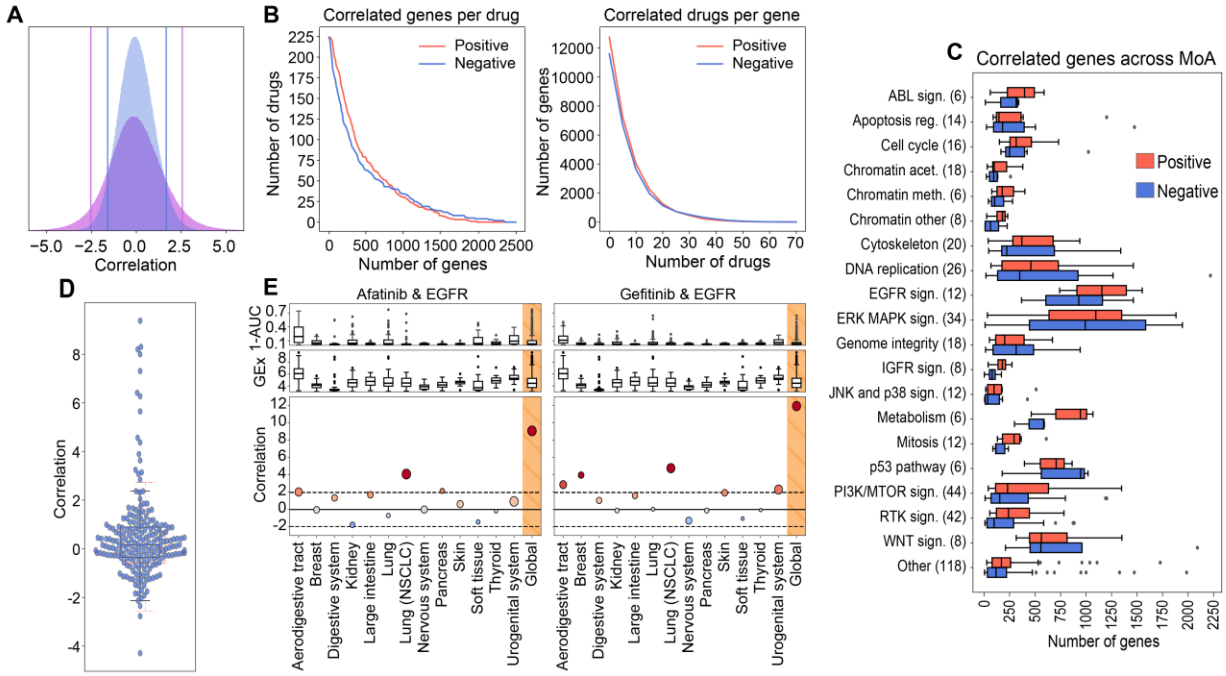

Figure 1. Analysis of drug-gene correlations. (A) Observed drug-gene correlation distribution (purple) and randomized drug-gene correlation distribution (blue) (random permutation of expression values). Vertical lines denote the percentiles 5 and 95 of each distribution. (B) The left panel shows the "number of correlated genes per drug", while the right panel shows the "number of correlated drugs per gene". In the left panel one can read, for example, that there are about 25 drugs (y-axis) with at least 1,250 correlated genes (x-axis). Likewise, in the right panel one can read that about 4,000 genes (y-axis) are correlated to at least 10 drugs (x-axis). (C) Number of positively (red) and negatively (blue) correlated genes across drug classes. (D) Positively-correlated targets (see Methods for details on the z-score normalization procedure of this correlation measure). Each dot represents one drug-target correlation. A full account of drug-target annotations is provided in Additional file 8. The red boxplot shows the background (random) distribution. (E) Drug-gene correlations (z_{cor}) between Afatinib/Gefitinib and the Epidermal Growth Factor Receptor (EGFR) across tissues. In the upper plots, we show the drug sensitivity (1-AUC) across tissues. In the middle plots, we show basal gene expression of EGFR across tissues. Bottom plots show the Afatinib/Gefitinib-EGFR correlation. The rightmost values refer to the correlation when all tissues are considered (Global). Size of the bubbles is proportional to the number of CCLs in each tissue.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

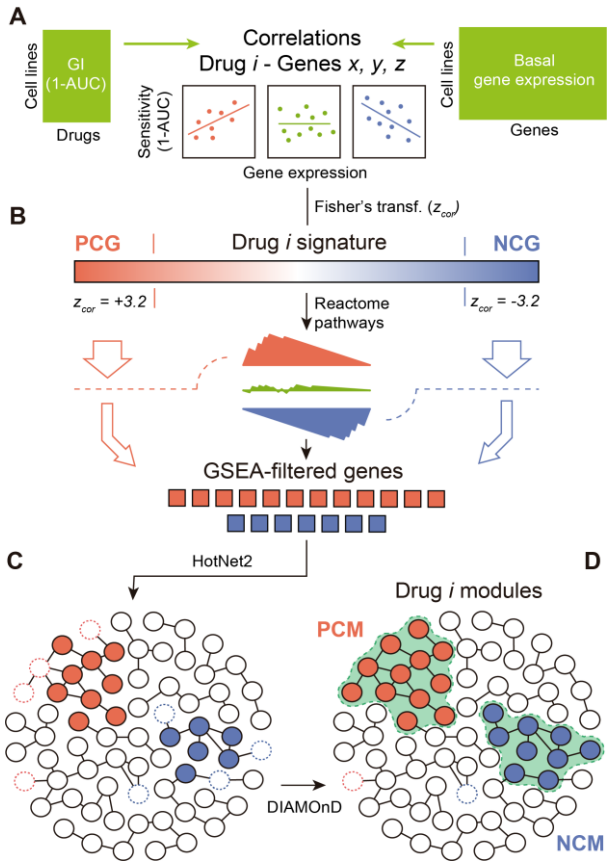

Figure 2. Methodological pipeline to identify drug modules. (A) The process of obtaining modules starts with the calculation of z-transformed Pearson's correlation (z_{cor}) between gene expression and drug sensitivity data for each drug-gene pair. Correlations (z_{cor}) beyond +/- 3.2 are considered to be significant. (B) We then run a gene-set enrichment analysis (GSEA) for each drug in order to identify genes that participate in enriched Reactome pathways. (C) GSEA-filtered genes are submitted to HotNet2 on the STRING interactome in order to identify drug modules. (D) Finally, modules are expanded (when possible) using the DIAMOnD algorithm to recall the few correlated genes that might have been excluded in step C as a result of the limited coverage of the Reactome database. This final step has a minor impact on the composition of the module.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

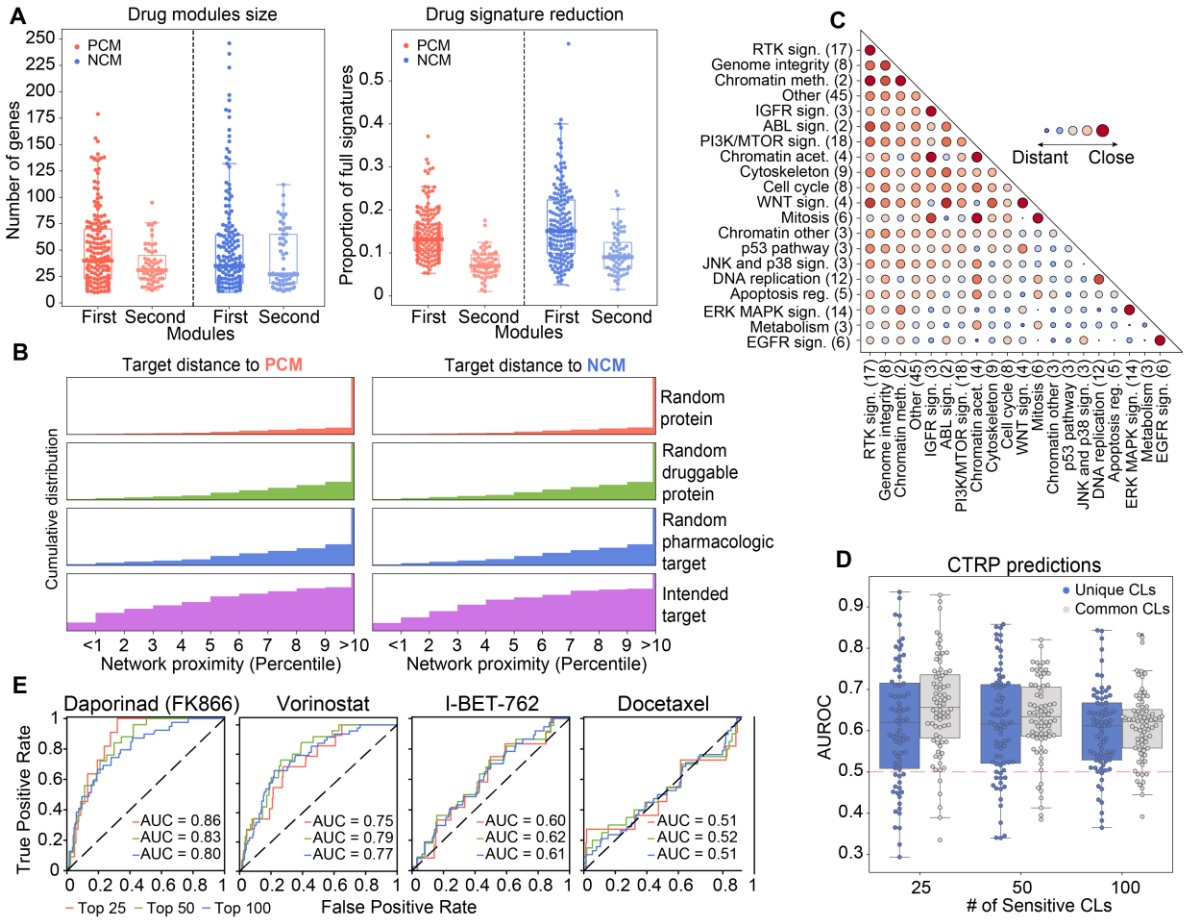

Figure 3. Global drug module analysis. (A) Number of genes in positively and negatively correlated modules (PCMs and NCMs) (left). Proportion of genes in the modules with respect to PCGs and NCGs (i.e. full signature). (B) Distance between drug targets and PCMs/NCMs (purple cumulative distribution). Results are compared to random proteins from the STRING interactome (red), proteins sampled from the "druggable proteome" (Target Central Resource Database) (green) and proteins sampled from the pharmacological targets in DrugBank (blue). (C) Network-based distances between drug classes. The bigger the bubble, the closer the distance between drug classes. Drug classes are sorted by specificity in their proximity measures. Please note "distant" values in the diagonal are possible due differences in drug modules belonging to the same class. The network-based distance calculation is detailed in Methods. (D) Predictive performances (AUROC) of the drug modules evaluated in the CTRP panel (top 25, 50 and 100 sensitive CCLs). Blue distributions correspond to results using unique CCLs (i.e. not shared with the GDSC panel). (E) Illustrative ROC curves for Daporinad (FMK866), Vorinostat, I-BET-762 and Docetaxel.

bioRxiv preprint doi: https://doi.org/10.1101/383588; this version posted March 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

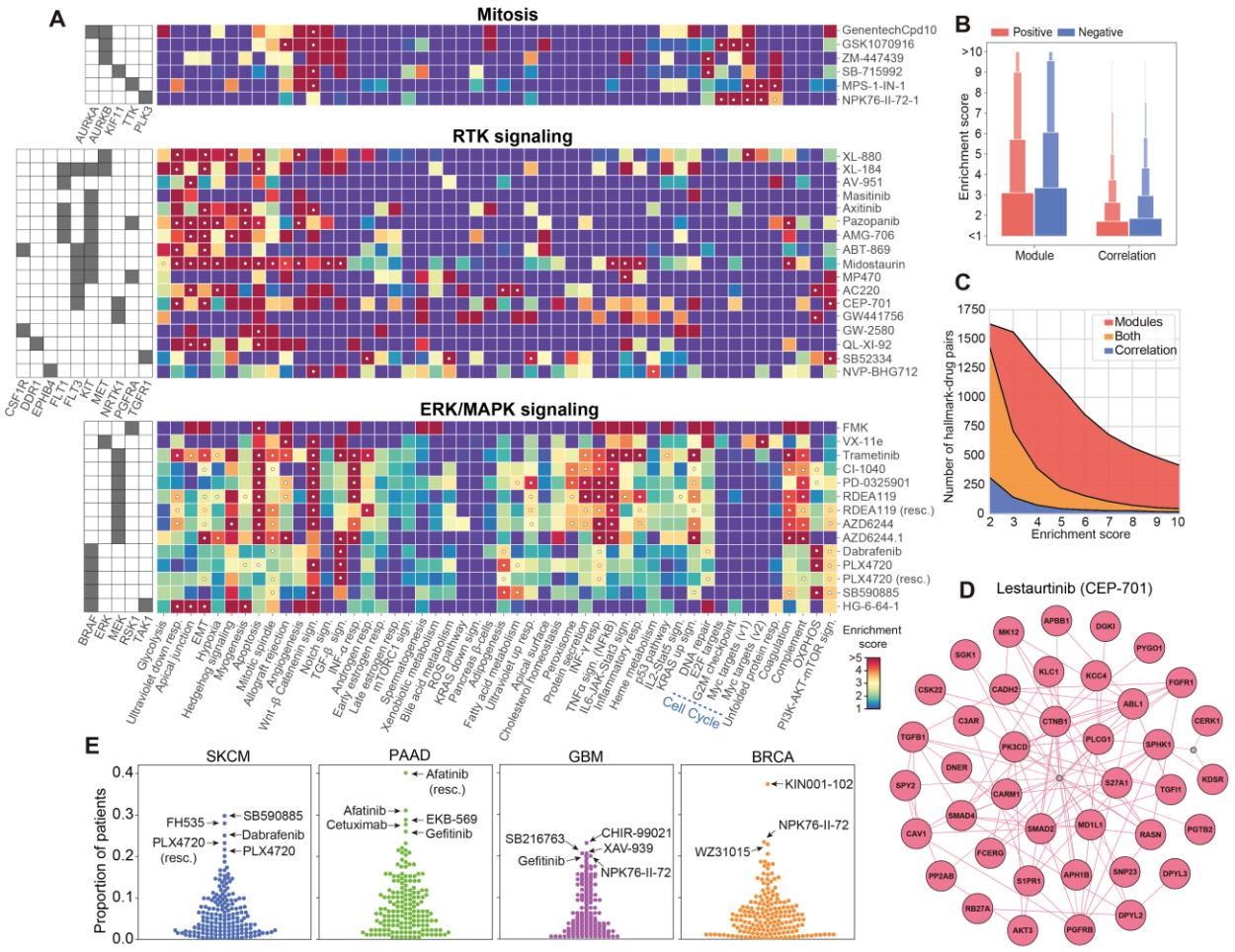

Figure 4. Drug module characterization. (A) Drug module enrichment analysis based on the Hallmark gene set (odds ratios in color, p-values < 0.05 are marked with a white dot). For simplicity, three drug classes are shown: drugs affecting mitosis, RTK signaling, and ERK/MAPK signaling. (B) Distribution of the enrichment scores in the Hallmark collection gene sets. Overall, higher enrichment scores are obtained using modules than using full signatures (PCGs and NCGs) (the gene universe used here is that of Reactome). (C) Similarly, number of Hallmark-drug pairs at different enrichment scores. We show the pairs found only with the modules (PCMs and NCMs, red), only with correlations (PCGs and NCGs, blue), or in both (orange). (D) A view of Lestaurtinib (CEP-701) module. For illustrative purposes, two out-of-the-module (non-correlated) proteins are shown (gray), one being very central and one being peripheral but acting as a "bridging" node. (E) TCGA enrichment analysis of PCMs in four types of cancer: SKCM (Cutaneous Melanoma), PAAD (Pancreatic Adenocarcinoma), GBM (Glioblastoma Multiforme) and BRCA (Breast Carcinoma).