BIOINFORMATICS

Vol. 30 ECCB 2014, pages i579–i586
doi:10.1093/bioinformatics/btu468

Unveiling new biological relationships using shared hits of chemical screening assay pairs

Xueping Liu[1,2] and Monica Campillos[1,2,*]

[1]Institute of Bioinformatics and Systems Biology and [2]German Center for Diabetes Research, Helmholtz Center Munich, 85764 Neuherberg, Germany

ABSTRACT

Motivation: Although the integration and analysis of the activity of small molecules across multiple chemical screens is a common approach to determine the specificity and toxicity of hits, the suitability of these approaches to reveal novel biological information is less explored. Here, we test the hypothesis that assays sharing selective hits are biologically related.

Results: We annotated the biological activities (i.e. biological processes or molecular activities) measured in assays and constructed chemical hit profiles with sets of compounds differing on their selectivity level for 1640 assays of ChemBank repository. We compared the similarity of chemical hit profiles of pairs of assays with their biological relationships and observed that assay pairs sharing non-promiscuous chemical hits tend to be biologically related. A detailed analysis of a network containing assay pairs with the highest hit similarity confirmed biological meaningful relationships. Furthermore, the biological roles of predicted molecular targets of the shared hits reinforced the biological associations between assay pairs.

Contact: monica.campillos@helmholtz-muenchen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The screening of a library of compounds in a biological assay is a common first step in drug discovery to find chemical hits for the drug leads. A single chemical screening experiment provides information about the activity of compounds on a target or biological process. However, to determine the suitability of the chemical hit as chemical probe or drug lead, it is important to know additional properties of the compound such as its specificity and toxicity. An inexpensive and efficient manner to obtain information about these properties is to learn about the activity of this compound across multiple chemical screens. This approach is followed routinely in chemical screening programs such as the NCI60 project run by 'US National Cancer Institute (NCI)' where the activity of a compound across 60 different cancer cell lines is measured to detect selective chemical hits for a particular cancer and avoid general toxicity (Shoemaker, 2006).

In the past decade several initiatives including the NIH Molecular Libraries Program (Austin *et al.*, 2004) and ChemBank (Seiler *et al.*, 2008) have compiled chemical biology experiments performed by different laboratories using diverse experimental set-ups ranging from cell-free to cell-based and even whole organism-based assays. The analysis of these heterogeneous datasets is challenging yet offers the possibility to obtain a global view of the chemical and biological activities of chemicals. In this regard, the integration and analysis of the collection of assays stored in the PubChem BioAssay (Wang *et al.*, 2010) repository has proven to be useful to predict adverse drug reactions (Pouliot *et al.*, 2011) and to determine chemical properties of promiscuous compounds, that is, those that appear as frequent hitters in many high-throughput assays (Canny *et al.*, 2012; Chen *et al.*, 2009; Schürer *et al.*, 2011).

The results of these studies suggest that a plethora of hidden molecular and biological information in these repositories can be uncovered using integrative computational methods. This is particularly relevant for the hits of phenotypic assays, for which the underlying molecular targets responsible for their activity is unknown. To determine the protein targets of the chemical hits of these assays, *in silico* target prediction methods (Keiser *et al.*, 2007; Liu *et al.*, 2013; Wang *et al.*, 2012) are arising as an efficient approach to obtain insights into the compound mode of action. For instance, Young *et al.* have shown recently that the predicted molecular targets of hits are able to explain complex readouts of high-content screening assays (Young *et al.*, 2008).

Here, we exploited the vast amount of publicly available chemical screening assays present in the ChemBank database to evaluate in a systematic manner if a pair of biological processes or molecular activities (hereafter named 'biological activities') modulated by common chemicals in phenotype- or target-based assays, respectively, is related. We tested and confirmed this hypothesis by the systematic analysis of the biological activities measured in pairs of assays sharing non-promiscuous compounds in this repository. Subsequently, to understand the molecular mechanism linking pairs of phenotypic assays sharing chemical hits, we annotated the molecular targets of the shared hits. To that aim, we used HitPick (Liu *et al.*, 2013), a recently developed *in silico* target prediction method to predict the molecular targets of compounds. We found that the known biological role of the predicted targets of common chemical hits confirms the biological processes relationships between the phenotypic assay pairs and provides mechanistic understanding of the relationships. This approach allows us to find relationships between biological activities and to understand better the molecular basis of the shared biological activities.

2 MATERIALS AND METHODS

2.1 ChemBank assay data structure

The ChemBank (Seiler *et al.*, 2008) data were downloaded in May 2011 and comprised 193 projects with loaded screening plates, including 3852 assays and 228 887 tested compounds. We also extracted information

*To whom correspondence should be addressed.

about the assays and projects including 'assay names', 'assay description', 'project names', 'project description' and 'project motivation'. Three projects containing 18 assays were discarded because they lacked information about compound IDs. If a project comprises assays containing in the 'assay name' an annotation of 'raw' and 'user', such as the project of 'Pseudomonas Cell Wall Synthesis', we only kept the assay annotated as 'user', as we observed that it often reports the specific activity of the compounds. This step retained 3617 assays. Then, we combined the assays performed with the same experimental protocol indicated by identical 'assay name' and 'assay description', such as assay ID 1133.0005, ID 1133.0006 and ID 1133.0007 of the project 'Glioblastoma Modulators', into the same 'assay type'. In total, 3617 assays were grouped into 1640 assay types. The analysis presented here was based on the assay type, which for simplicity we named 'assay'. We assigned the activity of a compound both on an assay level and a project level. A compound is active in a project when it is active in at least one of their assays.

We classified the assays into 'cell-free', 'cell-based' or 'microorganism' assays according to the assay description provided by ChemBank. If the assay was performed in a cell line (e.g. all the assays in the 'Glioblastoma Modulators' project were done in U251 human glioma cells), this assay was classified as 'cell-based'; if the assay was performed in a microorganism (e.g. the 'SigB Inhibition' project that identified small-molecule inhibitors of Listeria SigB transcription factor was performed in Vibrio sp. S1063), this assay was classified as 'microorganism'; the remaining biochemical or biophysical assays were classified as 'cell-free'.

2.2 Chemical hit identification methods

To identify chemical hits in the ChemBank data set, we applied three published methods, namely, the ChemBank (Seiler *et al.*, 2008), the B-Score (Malo *et al.*, 2006) and the Well-Correction (Makarenkov *et al.*, 2007) methods and five modifications of them to adapt the methods to the ChemBank data structure. These methods are summarized briefly as follows:

The ChemBank method aims to normalize the activity in the assay based on mock signals. B-Score method uses the median polish procedure to remove the row/column biases in a plate. First, a residual activity of a compound is calculated and then all data are normalized for plate- and row/column-specific effects. Chemical hits were determined using median absolute deviation (MAD) or *P*-value statistics, i.e. (i) compounds with a residual larger than 2*MAD ('2MAD'), (ii) $P<0.01$, (iii) $P<0.05$, were defined as hits. Well-Correction method rectifies the distribution of assay measurements by normalizing data within each considered well across all assay plates. In the end, both $P<0.01$ and 0.05 were applied to capture the hits.

As the B-Score method requires ideally the controls to be located randomly among the wells of each plate, or at most localized in the first and last columns, we created a modification of the method called B-Score_A adapted it to the ChemBank dataset structure where some plates only contain positive-control wells (e.g. plate ID 1031.0004.Pos.A and B). For this, positive controls were not considered in the median polish procedure and their residual activity was computed by subtracting the mean median effects of non-positive controls from their raw values. The next steps, including hit detection thresholds, were identical to those of the B-Score method.

The Well-Correction method requires the compounds measured across all assay plates to be randomly distributed. In the ChemBank dataset, many wells across different plates contain high number of positive controls (e.g. well A24 of assay ID 1017.0030) and therefore, the Well-Correction method cannot be applied directly. To correct for this, we discarded wells with higher number of positive controls (i.e. number of positive controls $\geq$ number of non-positive controls). To keep all the methods comparable, we applied this modification for the above four methods (marked as * in Fig. 1).

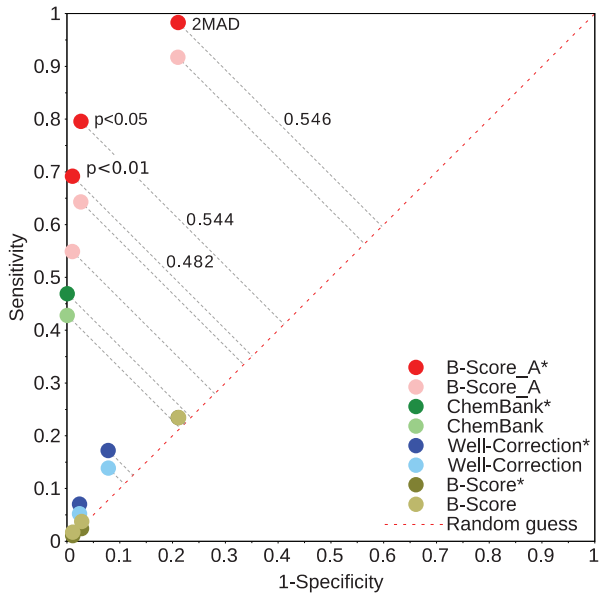

Fig. 1. ROC space showing the performance of the eight hit identification methods for the ChemBank assay dataset. To assess the performance of the eight methods, we calculated the distance of the coordinate (1-Specificity, Sensitivity) to a random guess line. The greater the distance to the random line, the better the method is. Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP). TP: true positive, TN: true negative, FN: false negative, FP: false positive. Asterisks denote modifications of the corresponding methods

If the assay contains replicates of compounds, we required all replicates to be identified as hits to consider them as chemical hits (also named actives, Fig. 2a). We determined the performance of the eight hit identification methods using the receiver operating characteristic (ROC) graph (Fawcett, 2006) and the positive and negative controls (including mock treatments) of the assays were used as a benchmark set. In all 3852 assays, the total number of positive controls is 96 and the number of negative controls is 7 590 042 and 7 620 521 for non-modified and modified versions of methods, respectively. The modification of the B-Score_A with two different thresholds, namely, '2MAD' and '$P<0.05$', showed the best performances. We selected the latter one due to its higher specificity (97.4%) with 79.6% of sensitivity.

2.3 Promiscuity filters

To increase computational efficiency, we applied filter F1 to keep compounds from the initial ChemBank dataset showing activity in more than one project. The removal of the compounds active in only one project or inactive in all the projects does not have an effect on the hit similarity (calculated by continuous Tanimoto coefficient, Tc (Willett *et al.*, 1998) (see Supplementary Methods) between assays (see Supplementary Fig. S1). Then, we applied two additional filters to keep selective compounds at project level (F2) and assay level (F3), respectively. F3 was applied to projects with at least nine assays, which was determined by averaging the number of assays per project in the ChemBank screening repository.

3 RESULTS

3.1 Assay structure and chemical hit identification

We chose the ChemBank repository of chemical screens to test the hypothesis of whether a pair of biological activities

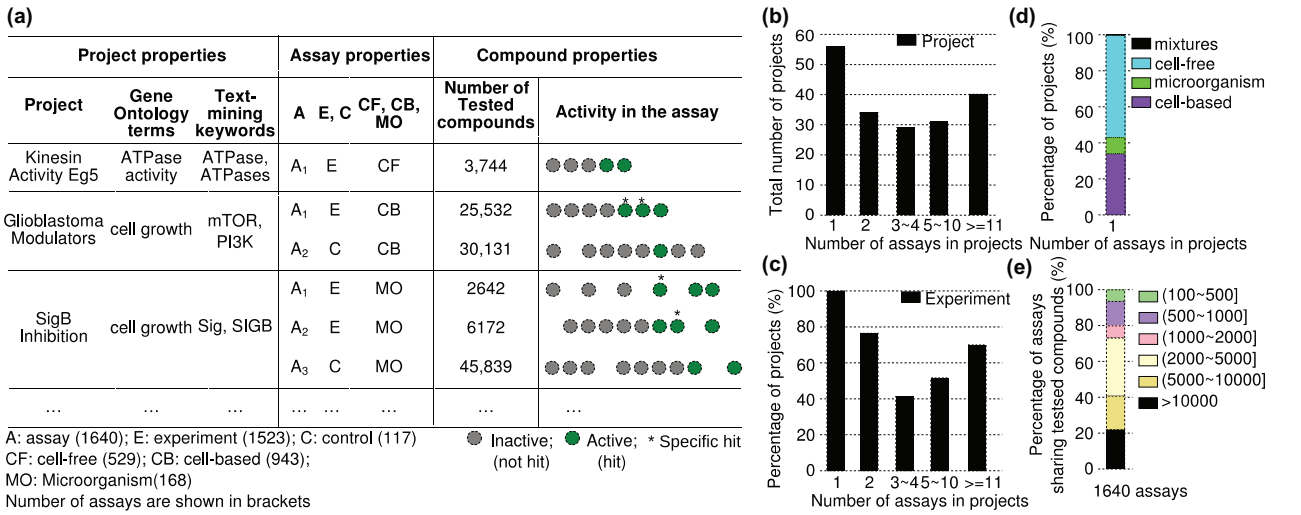

Fig. 2. Data structure of the ChemBank repository. (**a**) Classification of the different projects. Grey dots represent inactive compounds, while green dots represent active hits in the assay. Asterisks indicate that the hit is specific to the 'experiment' assay. (**b**) Distribution of the number of assays in projects. (**c**) Distribution of experimental assays in projects. (**d**) Distribution of cell-free, cell-based and microorganism assays in projects. (**e**) Percentage of assays sharing tested compounds

(i.e. biological processes or molecular activities) modulated by the same chemicals is related. In the ChemBank repository, the raw activity of a total number of 228 887 compounds in 3852 assays (representing experimental batches) of 190 diverse projects is available.

In a first step, we identified the chemical hits of the assays. As several approaches have been proposed to identify chemical hits in chemical screenings (Makarenkov *et al*., 2007; Malo *et al*., 2006; Seiler *et al*., 2008), we decided to test a collection of eight different methods (see Section 2) to select the best-performing hit identification method for the ChemBank dataset repository. To that aim, we determined the method that best discriminated between the compounds representing positive and negative controls within the assays. The B-Score_A method, a modification of the well-known B-Score method (Malo *et al*., 2006) achieved the best performance with a sensitivity of 79.6% and a specificity of 97.4% (Fig. 1). We thus selected this method to determine the chemical hits of ChemBank assays. Then, we grouped chemical screen batches performed using identical experimental protocols into 'assay types' (hereafter named 'assays') reducing the number of assays to 1640 (see Section 2).

Next, we annotated and classified the assays part of ChemBank projects to be able to compare them in terms of their biological relatedness. We first classified the assays into 'experiment' and 'control', according to whether the activity measured in the assay was the intended biological activity of the project or unspecific activities, respectively (Fig. 2a). In the second place, we classified the assays into cell-free, cell-based and microorganism based on the biological object of the experiments (Fig. 2a) (see Section 2). Lastly, we annotated the molecular activities and biological processes measured in the projects by assigning manually specific Gene Ontology (GO) (Ashburner *et al*., 2000) terms (biological process for phenotypic assays or molecular function for cell-free assays) to the projects Fig. 2a). As an additional description of the activity tested in projects, we manually assigned suitable keywords representing protein/gene names or biological processes to the projects (Fig. 2a). The distribution of the number of manual GO and keywords assigned to projects is listed in Supplementary Table S1. We then propagated the GO terms and keywords of each project to its 'experiment' assays.

We observed that the projects differ both in the number of assays (ranging from 1 to 113, Fig. 2b) and the percentage of 'experiment' assays (Fig. 2c) they include. This observation underlines the heterogeneity of the composition of ChemBank dataset. The distribution of cell-free, cell-based and microorganism assays is also heterogeneous. More than 40% of the projects are composed of phenotypic assays (cell-based and microorganism), and the majority of them are cell-based assays (Fig. 2d, also see Supplementary Fig. S2). Interestingly, despite the inhomogeneity of the ChemBank dataset, we found that ~80% of the assays have >1000 tested compounds (Fig. 2e) in common, indicating that the different assays can be compared based on the activity of a large number of compounds.

3.2 Promiscuity filters and similarity in biological activity

Next, we tested the hypothesis of whether chemical screening assays belonging to different projects with a similar chemical hit profile are biologically related. To that aim, we applied the Lin measurement (Lin, 1998) that quantifies the semantic similarity between GO terms assigned to the assays. Additionally, we applied the biomedical text-mining tool 'EXtraction of Classified Entities and Relations from Biomedical Texts (EXCERBT)' (Mewes *et al*., 2011) that detects terms co-mentioned in abstracts of scientific literature to evaluate whether the keywords linked to the assays of the pair are related.

Afterwards, for every assay and with the set of compounds that show activity in at least two projects (Filter 1, F1) (Fig. 3, F1), we constructed a binary fingerprint vector representing the activity of the set of compounds in the assays (1 active chemical

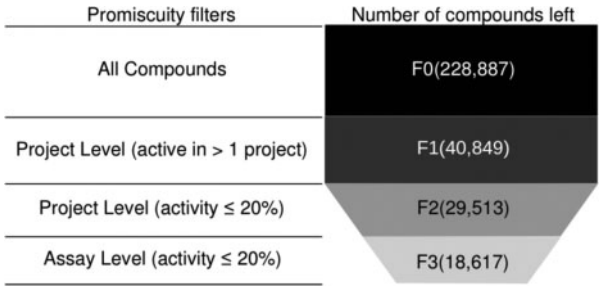

Fig. 3. Promiscuity filters. F0 contains all the compounds of the dataset. F1 keeps the compounds active in at least one project, and F2 retrieves the compounds active in ≤20% of the projects. F3 retains compounds active in ≤20% of the assays for the projects with higher than average number of assays (average number of assays per project is 9 for ChemBank). The number of remaining compounds after filtering is given in brackets

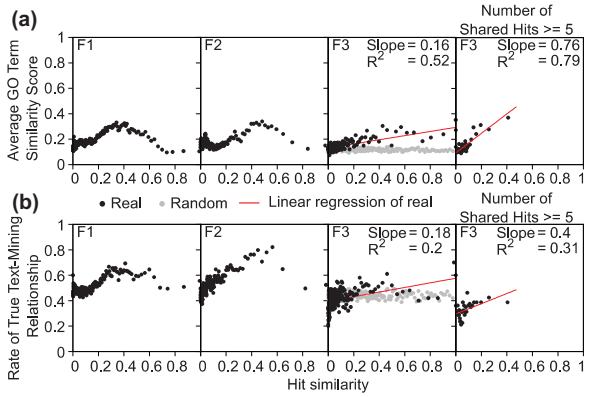

Fig. 4. Correlation between hit similarity and known relationships of ChemBank assay pairs. Hit similarity was calculated by continuous Tc. (**a**) Relationships indicated by GO terms and (**b**) relationships indicated by text mining. Each point in the plot represents a bin of assay pairs according to the sorted Tc values. In F1, each bin contains 1000 assay pairs. Bins in F2 and F3 contain 500 and 100 pairs, respectively. Separately, the performance of assay pairs in F3 sharing five or more hits is shown for (a) and (b)

hit, 0 inactive). Next, for all possible pairs of 'experiment' type assays belonging to different projects, we calculated the chemical hit similarity using a continuous Tc. Under these conditions, chemical hit similarity appeared not to be related to similar biological activities of assay pairs (Fig. 4a and b, F1). We reasoned that promiscuous compounds might be responsible for the high chemical hit similarity in unrelated assays. The prevalence of non-specific or promiscuous compounds is a well-known problem in high-throughput screening (HTS) assays commonly explained by their ability to form aggregates and act on unrelated targets (Feng *et al.*, 2005). Thus, their presence might be especially disturbing for the detection of biological connections between assay pairs.

Based on this assumption, we tested if the removal of promiscuous compounds increases the biological relatedness for assays sharing hits. To that aim, we applied two promiscuity filters. The first filter retained compounds with activity observed in <20% of the projects (Fig. 3, F2) and the second filter (F3) kept compounds that are active in <20% of the assays within a project. To avoid discarding specific chemical hits in projects with low number of assays where 'experiment' assays represent >20% of all assays, the filter F3 was applied only to projects with at least nine assays (Fig. 3, F3) (see Section 2). For example, the latter filter would discard all specific chemical hits in projects composed of one experiment and one control assay like the project 'Glioblastoma Modulators' (Fig. 2a) that searched for PI3K and mTOR modifiers in glioblastoma cells. If applied to this project, this filter would remove all specific hits, that is, those compounds that are active in cells treated with rapamycin ('experiment') and inactive in cells not treated with the mTOR inhibitor ('control'), as they are active on 50% (>20%) of the assays in this project.

As can be observed in Figure 4a and b, only after the application of the most stringent promiscuity filter F3, a linear relationship between hit similarity and known biological relationships was observed. Interestingly, such relationship disappeared when we compared assays with random hits, reinforcing the reliability of the relationships between biological activities captured by this approach. Furthermore, this trend became stronger when we discarded combinations of assays sharing low number of hits (Fig. 4a and b, number of shared hits ≥5, also see Supplementary Fig. S3a and b), indicating that the larger the number of common chemical hits is, the more likely it is to capture biological relationships between assays. An example of a known relationship between assay pair captured with our approach is the 'Bacterial Viability'–'Antibacterial' assay pair. This pair has a hit similarity of 0.54 (it shares 25 hits of the 51 and 25 tested compounds in each assay, respectively) and a biological similarity of 1 (the same 'GO:0016049 cell growth' term was annotated to both assays).

3.3 Assay interaction network

Next, we visualized and inspected manually the assay pairs showing high chemical hit similarity. For that, we constructed an assay interaction network with the assay pairs showing the highest hit similarity (Tc > 0.4) and sharing five or more chemical hits. This network contains 32 nodes and 26 edges (Fig. 5).

Interestingly, 92% of the edges in the network connect assays of the same experimental type. That is, phenotypic assays share hits with other phenotypic assays and cell-free assays tend to share hits with other assays of the same type. We found, for instance, a group of four interconnected assay pairs of the 'microorganism' type (i.e. 'Bacterial Viability', 'SigB Inhibition', 'Worm Anti-Infective' and 'Anti-Bacterial' assays) where the same biological activity, that is, the antibacterial activity, was sought in all of them. An example of a connection of two clearly related cell-free assays is the link between 'Kinesin Activity Eg5' and 'Kinesin Activity MKLP1' comprised by two assays aiming to find inhibitors of proteins of the Kinesin family. These instances provide evidence that relationships between the biological activities measured in the assays can be captured by our approach.

Intriguingly, we found a high number of edges (11, representing 42% of the edges) connecting 'control' assays to 'experiment' assays, the majority of them (9) linking two cell-based assays.

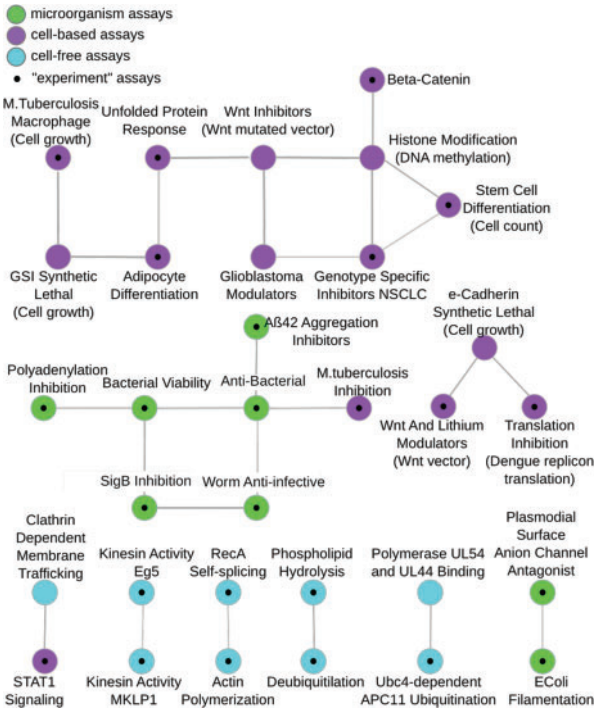

Fig. 5. Network of assay pairs from ChemBank repository sharing selective hits

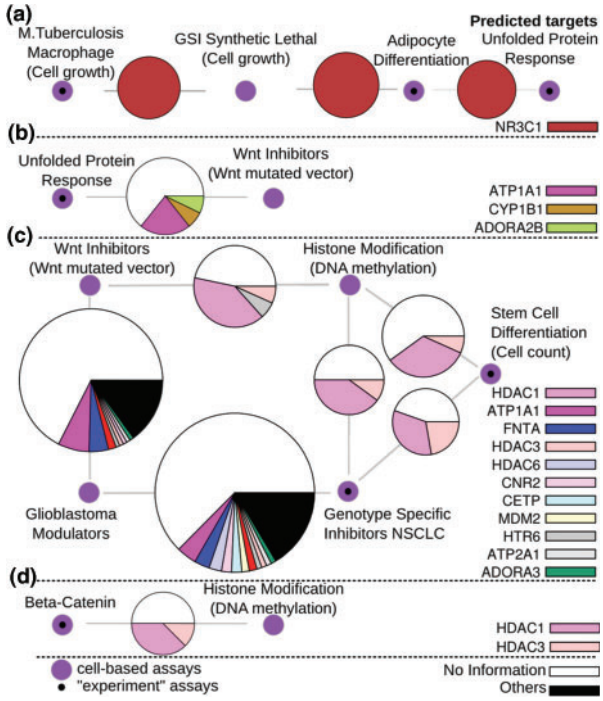

Fig. 6. Enriched targets between assay pairs. (**a–d**) are examples of assay connections (shown by assay name). The size of each pie chart is proportional to the logarithm of the number of shared hits. For simplicity, in the pie charts we show the most frequently predicted targets (with a precision >50%) of the shared chemical hits (see Supplementary Table S2 for the full target list of each assay pair in Fig. 6). The fraction of the pie charts representing hits with no predicted targets is shown in white as 'No Information'. In Figure 6c, only those representative targets common to three hits for assays pairs in the group are shown, and the remaining targets common to ≤2 hits are shown in black as 'Others'

A closer inspection of the activities measured in these assays indicates that cell growth-related processes, such as differentiation or growth inhibition, were often measured in the assays as the sought activity, for example, in assays seeking for chemicals with anticancer activity or in assays controlling the cytotoxicity of compounds. To gain deeper insights into the molecular basis of these assay combinations, we extracted molecular information of the chemical hits shared by these pairs by annotating predicted human drug targets of the compounds. For that, we applied the HitPick target prediction method (Liu *et al*., 2013) to predict the molecular targets of hits with high confidence (precision > 50%). Interestingly, we found the same predicted drug targets related to several assay pairs. For example, compounds specifically targeting the glucocorticoid receptor (NR3C1) are active in four consecutive assays in the network, namely '*Mycobacterium tuberculosis* (*M.tuberculosis*) Macrophage', 'Gamma Secretase Inhibitor (GSI) Synthetic Lethal (Cell growth)', 'Adipocyte Differentiation' and 'Unfolded Protein Response (UPR)' (Fig. 6a). The role of NR3C1 in macrophages as the target of anti-inflammatory agents (Barnes, 1998) and its anticancer activity (Cook *et al*., 1988) provides an explanation for the molecular basis of the relationship between the '*M.tuberculosis* Macrophage' that screened for inhibitors of *M.tuberculosis* growth in macrophages and 'GSI Synthetic Lethal (Cell growth)', a 'control' assay that tested the growth inhibitory activity of molecules in T-cells. Moreover, the known ability of NR3C1 to induce adipocyte differentiation (Xu *et al*., 1990) explains the common link between the cell growth and differentiation activities measured in 'GSI Synthetic Lethal (Cell growth)' and 'Adipocyte Differentiation' assays, respectively. Interestingly, although the link between UPR and differentiation processes has been proposed in the literature (Hetz, 2012), the molecular basis of this connection is not fully understood. Here, our results suggest the function of NR3C1 as intermediary between UPR induction and differentiation. However, this proposal should be taken with caution, as the specificity of the chemical hits on UPR process cannot be assessed owing to the lack of control assays in the project. In this context, the UPR assay is linked to a control assay of the 'Wnt Inhibitors (Wnt mutated vector)' project, which measures the promoter activity of a mutated version of Wnt responsive construct (Fig. 6b). A closer look at this relationship reveals that ATP1A1 (ATPase, Na^+/K^+ transporting, alpha 1 polypeptide), CYP1B1 (cytochrome P450, family 1, subfamily B, polypeptide 1) and ADORA2B (adenosine A2b receptor) are the predicted targets of the chemical hits of this pair. The role in cancer of ATP1A1 (Newman *et al*., 2008), CYP1B1 (Gajjar *et al*., 2012) and ADORA2B (Ma *et al*., 2010) indicate that the activity of compounds in the 'Wnt Inhibitors (Wnt mutated vector)' assay is likely due to their cytotoxicity. Although the known role of UPR to induce cell cycle arrest (Brewer and Diehl, 2000) and the recently reported role of ouabain, specific inhibitor of ATP1A1, on the modulation of UPR (Ozdemir *et al*., 2012), would suggest that the relationship between this assay pair is

due to the UPR-dependent growth inhibitory activity, further research is needed to assess the specificity of the shared hits on the UPR assay.

The growth inhibition measured in the 'Wnt Inhibitors (Wnt mutated vector)' assay is further confirmed by the association of this assay with the anticancer 'Gliobastoma Modulators' and 'Genotype-Specific Inhibitors in Non-Small Cell Lung Cancer' assays (Fig. 6c). Our target prediction approach revealed that, within this group of growth inhibitory assays, the cytotoxic activity is partly mediated through well-known anticancer targets, such as histone deacetylases (HDACs) (Wagner *et al*., 2010), ATP1A1 (Newman *et al*., 2008), farnesyltransferase, CAAX box, alpha (FNTA) (Rowinsky *et al*., 1999) and mouse double minute 2 homolog (MDM2) (Shangary and Wang, 2008). Furthermore, the modulation of these targets also explains the link between the chemical screens measuring stem cell differentiation ['Stem Cell Differentiation (Cell count)' assay] and DNA methylation [by 4,6-diamidino-2-phenylidole staining in 'Histone Modification (DNA methylation)' assay]. Intriguingly, other predicted targets behind the growth inhibition activity in this group of cancer-related assays include adenosine receptor A3 (ADORA3), cannabinoid receptor 2 (CNR2), cholesteryl ester transfer protein, plasma (CETP), 5-hydroxytryptamine receptor 6 (HTR6) and ATPase, Ca^{2+} transporting cardiac muscle, fast twitch 1 (ATP2A1). The modulation of these targets in anticancer screens suggests the possible role of these proteins in growth inhibition. In fact, the activity of ADORA3 as a potential target for tumor growth inhibition has been proposed before (Madi *et al*., 2004).

Another well-known biological connection is represented by the link between 'Beta-Catenin' assay that measured the nuclear translocation of beta-catenin and 'Histone Modification (DNA methylation)' assay (Fig. 6d). HDAC, the predicted target of the common hits, has been shown to inhibit Wnt signalling through disruption of the interaction between beta-catenin and T cell factor (Ye *et al*., 2009). Thus, the biological relationship between these two assays is explained by the known relationship of HDACs.

In summary, after retrieving the chemical hits from the ChemBank assays, we observed that the biological activities measured in two assays sharing selective hits are related. The close inspection of the assay pairs sharing specific hits in the network is able to confirm the biological and molecular associations of assay pairs and reveal molecular information underling the shared activity.

4 DISCUSSION

In this work, we have integrated and analysed the information stored in ChemBank and demonstrated that the biological activities of assay pairs sharing selective chemical hits are often related. The relationships between the biological processes of phenotypic assays are furthermore supported by the role of protein targets predicted for the shared hits.

Fingerprint-based approaches, where profiles of a collection of predefined features of an object such as a compound or protein is compared, have often been exploited in Chemistry and Biology fields to infer properties of compounds (Willett *et al*., 1998) (Willett, 2000) and genes (Liu *et* al., 2013). These approaches are based on the observation that similar fingerprint profiles correlate with similar properties (Fan *et al*., 2006). For example, compounds with similar chemical fingerprint profiles tend to have similar biological activities (Petrone *et al*., 2012). Likewise, compounds with similar modes of action have also been observed to exhibit similar behavior across multiple assays (Dancík *et al*., 2014). In contrast, in this study we use chemical hit-based fingerprints constructed with selective compounds to infer biological relationships between assays. Interestingly, we show that the relationships between assays can only be captured when a stringent selectivity filter is applied to discard promiscuous compounds from the chemical hit profile. Currently, there is no consensus for the definition of compound promiscuity, and different promiscuity filters have been proposed in the literature. Schürer *et al*. (2011) and Jacob *et al*. (2012) defined promiscuous compounds as those showing activity in >50 or 30% of the assays, respectively, while Gamo *et al*. (2010) calculated an 'inhibition frequency index' for each compound and applied a variable threshold, ranging from 5 to 20% of screens, depending on the number of HTS screens a given compound had been through. Although these studies have revealed interesting chemical moieties associated to unspecific signals in chemicals screens, the question of what level of selectivity is necessary to capture hits carrying information about specific biological signals has not been addressed yet. In this study, we have shown that a stringent promiscuity filter that first selects hits active in <20% of the projects (filter F2) and subsequently retains compounds with activity in <20% of the assays within a project (filter F3) is necessary to enrich for hits with specific biological activities. We reason that the low number of projects performed in the same experimental backgrounds generating the same unspecific signals might be the cause for the lack of correlation between hit and similarity of biological activities of two assays after the application of filter F2. Although this is partially overcome by discarding compounds active in several assays of the same project and consequently, performed in similar experimental backgrounds (filter F3), our approach also detects connections between cell-free assays that are apparently unrelated. For example, the 'Phospholypid Hydrolysis' assay is associated to the 'Deubiquitilation' assay (Fig. 5). A closer look at this connection reveals artefactual yet non-promiscuous hits, as the shared hits of the two connections appear active in the control assays of the project (termed 'unspecific' chemical hits, see Fig. 2a). This indicates that the stringent promiscuity filters applied here might, for some experimental conditions, be insufficient to discard unspecific hits, and additional control assays might be necessary to remove non-selective chemical hits.

The presence of unspecific hits is also evidenced by the occurrence of edges that connect 'control' and 'experiment' assays. For example, the 'e-Cadherin Synthetic Lethal (Cell growth)' 'control' assay that controlled for the cytotoxicity of compounds in the human mammary epithelial HMLE cell line is connected to the 'Wnt And Lithum Modulators (Wnt vector)' 'experiment' assay (Fig. 5), suggesting that the shared hits of the pair are not specific of the Wnt signalling process. This hypothesis is further corroborated by the known or suspected anticancer activity of the predicted targets [HDAC1 (Wagner *et al*., 2010), FNTA (Rowinsky *et al*., 1999) and sigma non-opioid

intracellular receptor 1 (SGIMAR1) (Aydar *et al*., 2006), Supplementary Table S2] of the shared hits and the modulation of these targets in a control assay of 'Wnt Inhibitors (Wnt mutated vector)' (Fig. 6c, also see Supplementary Table S2). Similarly, the link between the cytotoxic 'control' assay of the 'e-Cadherin synthetic lethal (Cell growth)' project and the 'Translation Inhibition (Dengue replicon translation)' assay that detected inhibitors of the translation of Dengue virus replicon (Fig. 5) points to the unspecificity of the chemical hits in the 'Translation Inhibition' assay. These examples illustrate the need of additional control assays in these screening projects to assess the specificity of the compounds. Nonetheless, we show that this approach was able to capture meaningful biological connections even between different types of assays, such as the link between a microorganism assay with a cellular assay. For example, the microorganism 'Anti-Bacterial' assay is connected with cellular '*M.tuberculosis* Inhibition' assay performed in BG1 ovarian cancer cells.

We observe that many relationships between different phenotypic assays are established based on the shared cytotoxicity of compounds in cell- or whole organism-based assays. Cytotoxicity appears thus as underlying biological effect common to phenotypic assays that account for the activity of many hits in these assays. Interestingly, the target prediction for those 'non-promiscuous' but 'cytotoxic' compounds reveals targets of drugs used as anticancer therapies, such as the HDACs (Wagner *et al*., 2010) and ATP1A1 (Newman *et al*., 2008), or targets that have been proposed for cancer treatment such as FNTA (Rowinsky *et al*., 1999) and MDM2 (Shangary and Wang, 2008). Hence, other predicted targets connecting these assays might represent potential targets for the treatment of cancers, such as CNR2, CETP, HTR6, ATP2A1 and ADORA3. Indeed, ADORA3 has been proposed as a potential therapeutic cancer target (Madi *et al*., 2004).

In summary, this work shows the potential of integrative approaches dealing with high-throughput chemical screening data to reveal novel connections between the biological processes and molecular activities measured in chemical screens. In the future, with the expected increase in HTS assay data available in public repositories, it is envisioned that many more biological relationships will be discovered with the application of this or similar computational approaches.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Dr Benedikt Wachinger and Dr Volker Stümpflen from Clueda AG for the support of providing the EXCERBT text-mining tool. We also acknowledge members of SBSM group at the Helmholtz Center Munich for helpful discussions and the support of the TUM Graduate School's Faculty Graduate Center Weihenstephan at the Technische Universität München, Germany.

Funding: This study was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes research (DZD e.V).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
Austin,C.P. *et al.* (2004) NIH molecular libraries initiative. *Science*, **306**, 1138–1139.
Aydar,E. *et al.* (2006) The expression and functional characterization of sigma (sigma) 1 receptors in breast cancer cell lines. *Cancer Lett.*, **242**, 245–257.
Barnes,P.J. (1998) Anti-inflammatory actions of glucocorticoids: molecular mechanisms. *Clin. Sci. (Lond)*, **94**, 557–572.
Brewer,J.W. and Diehl,J.A. (2000) PERK mediates cell-cycle exit during the mammalian unfolded protein response. *Proc. Natl Acad. Sci. USA*, **97**, 12625–12630.
Canny,S.A. *et al.* (2012) PubChem promiscuity: a web resource for gathering compound promiscuity data from PubChem. *Bioinformatics*, **28**, 140–141.
Chen,B. *et al.* (2009) PubChem as a source of polypharmacology. *J. Chem. Inf. Model.*, **49**, 2044–2055.
Cook,P.W. *et al.* (1988) Glucocorticoid receptor-dependent inhibition of cellular proliferation in dexamethasone-resistant and hypersensitive rat hepatoma cell variants. *Mol. Cell. Biol.*, **8**, 1449–1459.
Dancík,V. *et al.* (2014) Connecting small molecules with similar assay performance profiles leads to new biological hypotheses. *J. Biomol. Screen.*, **19**, 771–781.
Fan,X.-H. *et al.* (2006) Multiple chromatographic fingerprinting and its application to the quality control of herbal medicines. *Anal. Chim. Acta*, **555**, 217–224.
Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**, 861–874.
Feng,B.Y. *et al.* (2005) High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.*, **1**, 146–148.
Gajjar,K. *et al.* (2012) CYP1B1 and hormone-induced cancer. *Cancer Lett.*, **324**, 13–30.
Gamo,F.-J. *et al.* (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature*, **465**, 305–310.
Hetz,C. (2012) The unfolded protein response: controlling cell fate decisions under ER stress and beyond. *Nat. Rev. Mol. Cell Biol.*, **13**, 89–102.
Jacob,R.T. *et al.* (2012) MScreen: an integrated compound management and high-throughput screening data storage and analysis system. *J. Biomol. Screen.*, **17**, 1080–1087.
Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
Lin,D. (1998) An information-theoretic definition of similarity. In: *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, pp. 296–304.
Liu,X. *et al.* (2013) HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics*, **29**, 1910–1912.
Ma,D.-F. *et al.* (2010) Hypoxia-inducible adenosine A2B receptor modulates proliferation of colon carcinoma cells. *Hum. Pathol.*, **41**, 1550–1557.
Madi,L. *et al.* (2004) The A3 Adenosine receptor is highly expressed in tumor versus normal cells potential target for tumor growth inhibition. *Clin. Cancer Res.*, **10**, 4472–4479.
Makarenkov,V. *et al.* (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics*, **23**, 1648–1657.
Malo,N. *et al.* (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.*, **24**, 167–175.
Mewes,H.W. *et al.* (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.*, **39**, D220–224.
Newman,R.A. *et al.* (2008) Cardiac glycosides as novel cancer therapeutic agents. *Mol. Interv.*, **8**, 36–49.
Ozdemir,T. *et al.* (2012) Ouabain targets the unfolded protein response for selective killing of HepG2 cells during glucose deprivation. *Cancer Biother. Radiopharm.*, **27**, 457–463.
Petrone,P.M. *et al.* (2012) Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem. Biol.*, **7**, 1399–1409.
Pouliot,Y. *et al.* (2011) Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clin. Pharmacol. Ther.*, **90**, 90–99.
Rowinsky,E.K. *et al.* (1999) Ras protein farnesyltransferase: a strategic target for anticancer therapeutic development. *J. Clin. Oncol.*, **17**, 3631–3652.
Schürer,S.C. *et al.* (2011) BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *J. Biomol. Screen.*, **16**, 415–426.
Seiler,K.P. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–359.
Shangary,S. and Wang,S. (2008) Targeting the MDM2-p53 interaction for cancer therapy. *Clin. Cancer Res.*, **14**, 5318–5324.
Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.

Wagner,J.M. *et al.* (2010) Histone deacetylase (HDAC) inhibitors in recent clinical trials for cancer therapy. *Clin. Epigenetics*, **1**, 117–136.

Wang,J.-C. *et al.* (2012) idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. *Nucleic Acids Res.*, **40**, W393–399.

Wang,Y. *et al.* (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.*, **38**, D255–266.

Willett,P. (2000) Chemoinformatics – similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.*, **11**, 85–88.

Willett,P. *et al.* (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **38**, 983–996.

Xu,X.F. *et al.* (1990) Progestin binds to the glucocorticoid receptor and mediates antiglucocorticoid effect in rat adipose precursor cells. *J. Steroid Biochem.*, **36**, 465–471.

Ye,F. *et al.* (2009) HDAC1 and HDAC2 regulate oligodendrocyte differentiation by disrupting the beta-catenin-TCF interaction. *Nat. Neurosci.*, **12**, 829–838.

Young,D.W. *et al.* (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, **4**, 59–68.