BMC Bioinformatics

BioMed Central
The Open Access Publisher

This Provisional PDF corresponds to the article as it appeared upon acceptance. Fully formatted PDF and full text (HTML) versions will be made available soon.

Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs

BMC Bioinformatics 2012, **13**:112 doi:10.1186/1471-2105-13-112

Suleiman A Khan (suleiman.khan@aalto.fi)
Ali Faisal (ali.faisal@aalto.fi)
John P Mpindi (john.mpindi@helsinki.fi)
Juuso A Parkkinen (juuso.parkkinen@aalto.fi)
Tuomo Kalliokoski (tuomo.kalliokoski@novartis.com)
Antti Poso (antti.poso@uef.fi)
Olli P Kallioniemi (olli.kallioniemi@helsinki.fi)
Krister Wennerberg (krister.wennerberg@helsinki.fi)
Samuel Kaski (samuel.kaski@hiit.fi)

ISSN 1471-2105

Article type Research article

Submission date 31 August 2011

Acceptance date 9 April 2012

Publication date 30 May 2012

Article URL http://www.biomedcentral.com/1471-2105/13/112

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

http://www.biomedcentral.com/info/authors/

© 2012 Khan *et al.* ; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs

Suleiman A Khan[1*]
[*] Corresponding author
Email: suleiman.khan@aalto.fi

Ali Faisal[1]
Email: ali.faisal@aalto.fi

John Patrick Mpindi[2]
Email: john.mpindi@helsinki.fi

Juuso A Parkkinen[1]
Email: juuso.parkkinen@aalto.fi

Tuomo Kalliokoski[3]
Email: tuomo.kalliokoski@novartis.com

Antti Poso[4]
Email: antti.poso@uef.fi

Olli P Kallioniemi[2]
Email: olli.kallioniemi@helsinki.fi

Krister Wennerberg[2]
Email: krister.wennerberg@helsinki.fi

Samuel Kaski[1,5*]
[*] Corresponding author
Email: samuel.kaski@hiit.fi

[1] Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, PO Box 15400, Espoo 00076, Finland

[2] Institute for Molecular Medicine Finland FIMM, University of Helsinki, PO Box 20, Helsinki 00014, Finland

[3] CADD, Global Discovery Chemistry, Novartis Institute for Biomedical Research, Basel CH4002, Switzerland

[4] School of Pharmacy, Faculty of Health Sciences, University of Eastern Finland, PO Box 1627, Kuopio 70211, Finland

[5] Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, PO Box 68, Helsinki 00014, Finland

Abstract

Background

Detailed and systematic understanding of the biological effects of millions of available compounds on living cells is a significant challenge. As most compounds impact multiple targets and pathways, traditional methods for analyzing structure-function relationships are not comprehensive enough. Therefore more advanced integrative models are needed for predicting biological effects elicited by specific chemical features. As a step towards creating such computational links we developed a data-driven chemical systems biology approach to comprehensively study the relationship of 76 structural 3D-descriptors (VolSurf, chemical space) of 1159 drugs with the microarray gene expression responses (biological space) they elicited in three cancer cell lines. The analysis covering 11350 genes was based on data from the Connectivity Map. We decomposed the biological response profiles into components, each linked to a characteristic chemical descriptor profile.

Results

Integrated analysis of both the chemical and biological space was more informative than either dataset alone in predicting drug similarity as measured by shared protein targets. We identified ten major components that link distinct VolSurf chemical features across multiple compounds to specific cellular responses. For example, component 2 (hydrophobic properties) strongly linked to DNA damage response, while component 3 (hydrogen bonding) was associated with metabolic stress. Individual structural and biological features were often linked to one cell line only, such as leukemia cells (HL-60) specifically responding to cardiac glycosides.

Conclusions

In summary, our approach identified several novel links between specific chemical structure properties and distinct biological responses in cells incubated with these drugs. Importantly, the analysis focused on chemical-biological properties that emerge across multiple drugs. The decoding of such systematic relationships is necessary to build better models of drug effects, including unanticipated types of molecular properties having strong biological effects.

Background

The mechanism of action of drugs at the biochemical level has typically been studied by investigating specific chemical properties of the drug and the biological properties of its specific target [1,2]. This is the standard paradigm in Quantitative Structure Activity Relationship (QSAR) studies, where multivariate mathematical models are used for modeling the relationships between a set of physiochemical or structural properties and biological activity. In previous QSAR studies, such as in the classical 3D-QSAR work by Cramer et al. [3], values of a single biological activity measure are predicted.

However, biological responses at the cellular level are diverse and each drug typically binds to a multitude of targets in the cells and elicits a number of other off-target effects. Systems-level approaches are thus needed to get a more comprehensive view of drug effects in living cells. Genome-wide massively multivariate description of the cellular responses caused by the drugs, such as in the Connectivity Map program (CMap; Lamb et al., [4]), requires new kinds of tools for analysis and interpretation.

Chemical systems biology has emerged at the interface of systems biology and chemical biology with the goal of constructing a systems-level understanding of drug actions. Systematic analysis of a network of drug effects, *i.e.* network pharmacology, offers great opportunities for drug design in the future [5]. Chemical systems biology has also been used to predict drug side effects [6] as well as in other types of toxicological analysis [7].

Here, we undertook a complementary approach, by studying the impact of a host of chemical descriptors across a large panel of drugs on the biological response profiles measured at a genome-wide scale. We linked key structural components of the drug molecules, as defined by 3D VolSurf descriptors, with the consistent biological properties, as measured by microarray gene expression profiles. We developed a data-driven approach to analyze relationships between patterns of chemical descriptors of the drugs on one hand, and matching patterns in the cellular responses measured by genome-wide expression profiles, as shown in Figure 1.

Figure 1 Data-driven search for statistical relationships between Chemical space (formed of VolSurf features) and Drug response space (gene expression)

As biological response data we used the Connectivity Map (CMap, [4,8]), which consists of gene expression measurements from three cancer cell lines (MCF7-breast, PC3-prostate and HL60-blood) treated with over a thousand different drug molecules (Figure 1C). These data offer a unique view to the genome-wide responses of the cells to drug treatments and has been used to find new biological links *e.g.* between heat shock protein inhibitors, proteasome inhibitors, and topoisomerase inhibitors [8].

Our key assumption is that the chemical structure as encoded in the 3D descriptors of drugs impacts on the drug response resulting in specific patterns of gene expression. Furthermore, if there is any statistical relationship between the occurrence of patterns in the chemical space and the patterns in biological response space, those patterns are informative in forming hypotheses on the mechanisms of drug action. Given proper controls, the statistical responses can be attributed to the specific features of the chemicals tested out of a diverse drug library. In this paper we used comprehensive but readily interpretable models for finding the

statistical dependencies. We searched for distinct components that correlate the patterns in the chemical space with the biological response space. Assuming linear relationships, the task reduces to Canonical Correlation Analysis (CCA [9]) for searching for correlated components from the two data spaces (Figure 1D). We visualized the components in a comprehensive way to facilitate interpretation (Figure 1E and 1F) and validate them both qualitatively and quantitatively.

Canonical Correlation Analysis was recently used for drug side effect prediction and drug discovery by Atias and Sharan [10]. They applied CCA to combine known side effect associations of drugs with (i) 2D structure fingerprints and (ii) bioactivity profiles of the chemicals. The CCA results from both combinations were then successfully used to predict side effects for the drugs, suggesting that CCA is effective in finding relevant components from heterogeneous data sources.

Drugs generally act on a multitude of direct and intended targets as well as on a number of non-specific off-targets. All these targets and effects together connect to a phenotypic response. As most of these effects are still poorly understood, modelling of the structure-target-response profiles across a large drug library is an important, but challenging goal. In this study we modelled the structure-response relationships of 1159 drug molecules directly, with CCA components playing the role of unknown mechanistic processes.

The lack of information on all of the possible targets prompted us to select a particular set of chemical descriptors that allows capturing of generic response patterns. Many kinds of chemical descriptors are available for characterizing chemical structures in a quantitative way. Simple classical 2D fingerprints can be used to detect close analogs, but they would miss most if not all scaffold-hopping situations, where the different chemical entities give rise to similar pharmacophoric properties. Fingerprints and pure pharmacophoric descriptors require clearly defined individual targets, which are not known in many cases. In the present study, we aimed to bridge the chemical and biological space by using a set of VolSurf descriptors of the drugs ([11]; Figure 1A) that are ideal for capturing both structural similarities and general chemical features, such as solubility and permeation properties (ADMET: Absorption, Distribution, Metabolism, Excretion and Toxicology properties). Although VolSurf descriptors are not thought to explain detailed structure-activity relationships, such as the binding to a single target, they offer a good overall interpretation of the molecular shape, hydrogen bonding, lipophilicity, and related properties, which are more conservative than individual binding cavities. It has also been shown that shape is a major factor when trying to find compounds with similar biological activity but dissimilar 2D structures [12].

The idea of correlating chemical structures with biological expression was introduced by Blower et al. in [13]. By combining 2D fingerprint data with biological activity profiles for the chemicals over 60 cancer cell lines (NCI60), and with steady-state gene expression measurements from those cell lines before drug treatments, they obtained indirect relationships between chemical substructures and the gene targets. In a more recent work, Cheng et al., [14] investigated correlations between the chemical structures, bioactivity profiles, and molecular targets for a set of 37 chemicals. This small-scale study demonstrated that combinations of biological activity and chemical structure information can provide insights into drug action mechanisms on a molecular level.

By using the direct gene expression responses to a large set of drug treatments from the Connectivity Map, along with comprehensive component-level decomposition of response profiles, we are able to make more direct observations on how compounds impact on cells and which features of the chemical molecules and the biological responses are correlated.

Results and discussion

We analysed the 1159 drug treatment gene expression responses of three cancer cell lines of the Connectivity Map, with the methods summarized in Figure 1 and detailed in Methods. The analysis decomposed the relationship between the "chemical space" and the "biological space" into components. The chemical space consists of the selected 76 chemical descriptors of each drug and the biological space contains gene expression responses of corresponding drugs. Each component relates a characteristic statistical gene expression pattern with a pattern of the drug properties. We will call the components "CCA components" as the core method is Canonical Correlation Analysis (CCA). In this section we analyse further the identified components and the statistical relationships they discovered.

Quantitative validation of functional similarity of drug components

We evaluate the biological relevance of the extracted CCA components by studying the functional similarity of drugs associated with each component. In particular, we measure the performance of the component model in retrieving similar drugs, as indicated by external data about their function, and compare it to retrieval based on either the biological or chemical data separately. Details of the validation procedure are described in Methods. The mean average precision obtained for the retrieval task on the four data sets (CCA components, chemical space, biological space as represented by GSEA and Gene expression) are plotted in Figure 2.

Figure 2 Quantitative validation of functional similarity of drug components. The figure shows the mean average precision for retrieving functionally similar chemicals as a function of the number of top chemicals considered. Results are shown for three representations: CCA (red), Chemical space (green), and Biological space (GSEA: blue, Gene expression: grey). Error bars show one standard error of the mean precision

The results show that retrieval based on the chemical space, *i.e.* VolSurf descriptors, performs clearly better than retrieval based on the biological space (activities of gene sets and genes), indicating that the chemical information is more relevant for evaluating the functional similarity of the chemicals. The biological space encoded by gene sets performs similarly to the original gene expression, indicating that the gene sets are a sensible encoding of the biological space; information lost due to dimensionality reduction is balanced by introduction of prior biological knowledge in the form of the sets. Finally, the combined space formed by the CCA components shows significantly better retrieval performance than either of the data spaces separately. The results are consistent over the range of drugs considered in the retrieval task. These results show that CCA is able to extract and combine relevant information about the chemical structure and biological responses of the drugs, while filtering out biologically irrelevant structural information and also biological responses unrelated to the chemical characteristics.

Response components and their interpretations

We next analyze the top ten CCA components having the highest significant correlations between the spaces. Figure 3 summarizes the relationships between the VolSurf descriptors and the gene sets as captured by the components. Each component is divided into two subcomponents 'A' and 'B', where in the first, the compounds have positive canonical score and in the second negative (the characteristic response patterns are otherwise the same, details in Methods). For each CCA-subcomponent the 20 highest-scoring compounds are listed in the Additional file 1: Top_Compounds.xls.

Figure 3 Relationships decomposed into components. "Eye diagram" showing the top 10 significant CCA components ordered by correlation from top to bottom (middle), VolSurf descriptors (left), and top gene sets (right). The CCA components are shown as circles, with numbers indicating the decreasing order of canonical correlation and letters A and B indicating subcomponents (A: positive canonical score, B: negative canonical score). The widths of the curves from the components to VolSurf descriptors and gene sets indicate the strength of the corresponding associations. For VolSurf descriptors the subcomponent-specific activity is shown, whereas for the gene sets we show the overall activity in the component. For an example compound, VolSurf fields are illustrated in the top-left corner while three gene sets are listed along with their five most significant genes in the top right corner

VolSurf descriptors, unlike more typically used 2D or 3D fingerprints and pharmacophores, do not have clear structural counterparts such as fragments or functional groups. Therefore, these descriptors are able to group together compounds with very dissimilar chemical structures and yet having the same type of chemical properties. This is especially important in our study since the same biological activity may not have been created because of the same biological target protein but because of another target protein in the same pathway. Within a pathway, binding cavity properties may change, but general descriptors describing the size, lipophilicity, and shape are more conservative. Therefore VolSurf descriptors are better suited for classifying these aspects of the compounds.

In the case of the 1st and 2nd components, the selected VolSurf descriptors are dominated by hydrophobicity related descriptors. The 3rd component is connected to hydrogen bonding, polar interactions, and dispersion-related descriptors. Component 4 is also strongly connected with hydrophobicity related descriptors, besides the component is influenced by the molecular volume and surface area. Components 5–7 are mainly connected to "pharmacophoric" descriptors that are describing distribution of strong interaction points over the molecular space. Components 8 and 9 are both strongly influenced by integy-moments, describing imbalance of either hydrophobic or hydrophilic areas over the whole molecular volume. Component 10 is mainly affected by shape and size-related parameters, and also lipophilic integy moments.

On the biological space we observe that the enriched gene sets in component 1 indicate a mitogenic signaling response. Component 8 appears similar but has an additional link to cell adhesion signaling. Component 4 in turn very directly connects to cytoskeletal regulation and cell adhesion. While there appears to be a considerable overlap between compounds in components 2, 3 and 10, the enriched gene sets in component 2 show a strong link to DNA damage response, 10 is associated with common cancer signals, and component 3 is associated with an anabolic response. Components 5 and 6 are connected to different

differentiation events. Component 7 links to gene expression changes seen in GPCR signaling. Component 9 links to amino acid and nitrogen metabolism.

We further extracted the significant genes in each component and performed GO enrichment analysis on them. Additional file 2: Top_GOTerms.xls shows the top 10 significant GO terms for each component while Additional file 3: Top_Genes.xls shows the top 30 significantly differentially expressed genes for each.

Based on the Eye diagram (Figure 3) and lists of significant genes, gene sets, GO terms, and the top 20 compounds, we summarize the biological and chemical patterns in Table 1.

Table 1 Summarized interpretation of top 10 components. Group A and B are the subcomponents of Figure 3

Comp.	Biological Interpretation	Compounds in Group A	Compounds in Group B	VolSurf Interpretation
1	Classic growth factor signaling: (MAP and protein kinase signaling)	Sulfonamides, antibiotics, carbonic anhydrase inhibitors	Antipsychotic and antihistaminic compounds	High lipophilicity
2	DNA damage	Contrast agents, antibiotics,	DNA damaging agents, antimetabolites	Strong lipophilic areas emphasized
3	Stress response, mitochondrial and anabolic metabolism	DNA damaging agents	GPCR antagonists, ion channel blockers	Polar interactions enriched
4	Cytoskeleton, cell adhesion and migration	GPCR liganda, macrocyclic cmpds and contrast agents	Beta adrenergic agonists, other GPCR ligands	N/A
5	Differentiation, EMT, stemness	NSAIDS, cAMP signaling promoting compounds	HDAC Inhibitors, HDAC-like	Significantly enriched with pharmacophoric features*
6	Inflammatory and differentiation signaling	N/A	Protein synthesis inhibitors, anti-diabetics, cardiac glycosides	Pharmacophoric features*
7	GPCR and cytokine signaling	N/A	Cardiac glycosides, cephalosporins	Pharmacophoric features*
8	Growth factor and cell adhesion signaling	Cardiac glycosides	β-adrenergic agonists, Ca^{2+} channel blockers	Integy-moment and significant pharmacophoric enriched*
9	Amino acid and nitrogen metabolism	Protein synthesis inhibitors	Anti-diabetics	Integy-moment and significant pharmacophoric enriched*
10	Cancer signaling	DNA damaging agents	Corticosteroids, ionophores	Size shape type descriptors

The pharmacophoric enrichment analysis (marked with "*") was carried out over VolSurf features (Additional file 5: VolSurf_Classification.xls) considered as a gold standard, and measuring enrichment of the list in a component by a hypergeometric test

Enrichment of known targets

Existing drug response research relies heavily on target and pathway-based analysis. Our novel approach attempts to go beyond known targets and pathways to find drug response groups in an entirely data driven way linking pharmacophoric descriptors to specific gene expression response patterns. Some of the found patterns may rationally be explainable by known targets and next we analyze the discovered biological process classes (components), for enriched protein targets.

We collected a set of drug-protein target pairs from several drug databases (see Methods for details); that contains targets for almost half of the CMap drugs. We then did enrichment analysis of the compounds sharing common targets in each component (see Methods for details). The common protein targets and component p-values are listed in Additional file 6: Target_Breakdown.xls.

As expected, the analysis shows that only few subcomponents have compounds that significantly share protein targets ($p < 0.05$). These are 1B, 2B, 10A and 10B. Subcomponent 1B appears to be strongly driven by endogenous amine GPCR interactions. Subcomponent 2B and 10A, like 3A, are driven by compounds inducing DNA damage and are discussed below in more detail. Subcomponent 10B is strongly influenced by glucocorticoids primarily targeting the glucocorticoid receptor, but the fact that several other ion flux modulating compounds (erastin, clofilium tosylate, colforsin, monensin and lasalocid) also link to this subcomponent argues that the corticosteroid effects are through the mineralocorticoid receptor and that the subcomponent response is driven by a selected ion flux.

The fact that other subcomponents do not significantly link to target classes in our target analysis is a strong indicator that the associations between the compounds in most components are not restricted to known primary target mechanisms and that our method allows for discovering novel, but still undefined, mechanism of action and target linkages between compounds. The evaluation of mechanisms of action that cannot be explained via known protein targets is a challenging research direction, which requires uncovering the vast hidden mechanisms that might make two seemingly non-similar compounds similar. We feel that our approach provides a step forward towards the goal of understanding drug associations extracted from the actual measurement data that could potentially provide hypotheses for unexplored polypharmacology and both target and off-target drug mechanisms.

Components 3/3A – A cell stress component

We observed that in component 3, the top genes and gene sets indicate mostly mitochondrial and metabolic stress-related processes. Top gene sets associating with this component include many gene sets connecting to mitochondrial function (Figure 3). Similarly, on the gene level several known cell stress genes such as PGK1, PGD, and PRMT1 [15-17] are upregulated. A deeper look into the 3D structures of the top compounds in this component (Additional file 1: Top_Compounds.xls) reveals a possibility of 4–12 hydrogen bonds in all of the top compounds of set 3A. Thus, these compounds may be able to affect proton transportation

processes, which is in agreement with the biological interpretation that mitochondria act like proton pumps. As an example, Figure 4 shows the hydrogen bond donor and acceptor regions of azacitidine and idarubicin.

Figure 4 Compounds high in hydrogen Bonding. Azacitidine (left) and Idarubicin (right) showing H-bonding areas with blue (hydrogen-bond donor) and red (hydrogen-bond acceptor)

To help understand how the biological variability caused by the drugs and the chemical features in component 3A compares with the intrinsic variability from one cell type to another, we visualize drug response transitions on MCF7 cells. Thirty different breast cancer cell types are used as references in their unperturbed condition (as described in Methods). The MCF7 treatments from subcomponent 3A and the thirty independent breast cancer cells are plotted in Figure 5 using a recent multidimensional scaling method called NeRV [18,19]. NeRV shows these different cell instances mapped onto the 2-dimensional display such that similarities are preserved as faithfully as possible. Subcomponent 3A contains many DNA-damaging agents such as the DNA intercalating and topoisomerase inhibitory camptothecin, daunorubicin, and mitoxantrone, the CDK inhibitors alsterpaullone, GW-8510 and 0175029–0000, the cardiac glycoside lanatoside C, which at high concentrations is likely to inhibit topoisomerases [20] the antimetabolite methotrexate, as well as rescinnamine, which has been suggested to induce a DNA damage response without itself inducing DNA damage [21] and the aromatase inhibitor letrozole. The NeRV plot based on the top changed genes in treated and untreated MCF7 cells as well as a panel of other breast cancer cell lines, shows that after treatment with these drugs, the gene expression of the luminal, ER-positive MCF7 cells starts to resemble the basal, ER-negative breast cancer types. Interestingly, while MCF7 cells are relatively chromosomally stable, the drug-treatments make them appear like chromosomally unstable and intrinsic high DNA damage cell lines such as HCC1937 or MDA-MB-231. Cell stress is an emerging cancer therapy target and it is interesting that a) this subcomponent including topoisomerase inhibitors, antimetabolites and CDK inhibitors induces stress-related metabolic responses in MCF7 cells similar to what is seen in a basal level in other, more malignant and undifferentiated breast cancer cells, and b) it raises the hypothesis that the compounds in this subcomponent could be combined with cell stress targeting compounds. This finding is strengthened by the fact that many of the top upregulated genes in the 3A subcomponent; ACHY, CDC37, GPI, ME2, PMRT1, P4HB, PGD, and PGK1 are also overexpressed in breast cancers as compared to normal tissue (Additional file 4: HeatMaps.pdf, Figure A).

Figure 5 3A drug transitions. NeRV visualization showing Drug Treatment Transitions. Lines indicate the transition from Pretreated MCF7 to treated MCF7 cells

Components 2B & 10A – functionally similar but gene-wise different responses

We observe that component groups 2B, 3A, and 10A share several compounds such as the DNA-intercalating topoisomerase inhibitors mitoxantrone and irinotecan, the cyclin-dependent kinase (CDK) inhibitors alsterpaullone, GW-8510 and 0175029–0000, and the antimetabolites methotrexate and azacitidine, 5 of the top 20 between each paired group. Most of the non-overlapping compounds in each component group are not linked functionally or structurally in any obvious way, on the other hand. To verify that the components capture different phenomena despite sharing several compounds, we compute chemical composition

and biological similarity matrices over all component pairs. We use the Tanimoto similarity measure to compute overlap between the top 30 genes of each subcomponent pair; as shown in Additional file 4: HeatMaps.pdf, Figure D. The analysis of biological similarity between these subcomponents with compound overlap (out of top 20 compounds for each component) indicates that there is minimal biological and chemical sharing between any two components. Almost all component pairs that are highly biologically similar have a non-significant and low chemical composition similarity, and vice versa. This is a strong indication that we have identified sets of VolSurf descriptors that link to different biological responses. In some cases, several of these features can be identified in a single molecule like the etidronic acid, which is linked to both components 3 and 6. The chemical properties of component 6 are connected with pharmacophoric features and component 3 with hydrogen bonding, while biologically the components are related to differentiation and stress response, respectively.

To get a deeper view of the underlying biological response mechanisms we explore the response patterns of the components using heatmaps. In the first heatmap, we consider the most active genes in a subcomponent and plot their expression levels across the top compounds of every subcomponent (Figure 6). In the figure we search for the subcomponents that have a unique expression pattern across other subcomponents in a column. Components 2B and 10A show a unique structure. These seem to represent two separate aspects of DNA damage response, which are connected to two separate molecular features; hydrophobicity in component 2B and shape-type VolSurf descriptors in component 10A. The gene expression changes in both subcomponents are strongly linked to a DNA damage and mitotic arrest response with many proto-oncogenic cell division and mitogenic signaling genes being down regulated (Figure 6; Additional file 3: Top_Genes.xls). The same genes are commonly seen upregulated in cancers (Additional file 4: HeatMaps.pdf, Figure B and C) and many of them have been and are pursued as drug targets. Therefore both the components are similar on a larger biological scale, but do in fact have little gene-wise overlap.

Figure 6 Finding interesting components. Heatmap across the 10 highest scoring significant CCA components: X-axis lists the top 30 significant genes in each subcomponent, while y-axis represents the top 20 scoring compounds in each. Two unique components 2B and 10A are zoomed in to show the detailed expression pattern along with 3D VolSurf descriptors (green areas are the lipophilic fields and the purple water fields). Only a subset (5 compounds and 10 genes) is shown in the zoomed version due to space constraints

To validate these hypotheses, we checked for reported growth inhibition for the top 20 chemicals in these two subcomponents in the NCI/DTP *in vitro* cell line testing database (NCI60 testing program, http://dtp.nci.nih.gov/docs/cancer/cancer_data.html). Four compounds from 2B and 10 from 10A were represented in the NCI60 datasets (Table 2). For almost all of the compounds for which NCI60 data are available, in CMap the compounds were used at doses that very effectively will stop the cells from growing or kill them.

Table 2 Growth Inhibition verification of 2B/10A Compounds

Chemical	CMap (μM)	GI50 (μM)	Subcomponent	Cell line
berberine	**10**	**25.1**	2B	MCF7
irinotecan	**100**	**6.3**	2B, 10A	MCF7
mitoxantrone	**7.9**	**0.004**	2B	MCF7
amiodarone	**6.3**	**4.0***	2B	PC3
8-azaguanine	**25.1**	**0.32**	10A	HL60

apigenin	**15.9**	**25.1**	10A	HL60
azacitidine	**15.9**	**0.79**	10A	PC3
camptothecin	**12.6**	**<0.01****	10A	MCF7
chrysin	**15.9**	**15.8*****	10A	MCF7
methotrexate	**7.9**	**0.03**	10A	MCF7
thioguanosine	**12.6**	**0.32**	10A	MCF7
esculetin	25.1	>100**	10A	HL60
fulvestrant	1.0	>100**	10A	PC3

GI50 values (drug concentration causing a 50% growth inhibition) from NCI/DTP are shown along with the corresponding concentrations used in the Connectivity Map (CMap) data. By comparing the GI50 and CMap values we can get an idea of expected cell killing effect of the drug in the CMap data. Drugs that are expected to eventually kill the cells are shown in bold. GI50 and CMap concentration values are in μM scales.
* GI50 value at the end of the tested range.
** Mean of GI50 values from HL60 and MCF7 cell lines.
*** Value from HL60 cell line.

Component 7B – A leukemia-specific subcomponent

Based on studying the heatmaps, 7B is another interesting subcomponent: It has a dominant effect on HL60 as compared to MCF7 and PC3, indicating that this subcomponent and the link between structure and gene expression may be specific for leukemic cells and leukemic-specific molecular targets.

Figure 7 shows the activity of most significant genes (columns) corresponding to the top compounds (rows) across the three cell lines. The top drugs are mainly cardiac glycosides and these drugs are known to have a strong toxic effect on leukemic cells at the concentrations used. It is worth noting that FLT3LG (FLT3 ligand) is one of the most significantly up-regulated genes. The FLT3 receptor, to which FLT3LG binds, is an emerging target in leukemia [22].

Figure 7 Heatmap for subcomponent 7B. Y-axis lists the top 10 active compounds in the component, replicated over the three cell lines, while the X-axis lists the most significantly active genes in the component. The genes are clearly activated systematically and exclusively in the HL60 cell line, hence indicating an HL60 specific response

Conclusions

We have introduced a chemical systems biology approach for analyzing the complex relationship patterns between chemical structures of drug molecules and their genome-wide responses in cells. With Canonical Correlation Analysis, we are able to find statistical dependencies between the two data spaces in the form of correlated components. We have demonstrated quantitatively that these components are more informative about drug similarity than either chemical or biological data separately.

Our approach finds the relationships in an entirely data driven way without being constrained to known target information. Uncovering the detailed mechanisms of actions of a diverse

library of drugs, including those not having known target classes is a major research goal. Our approach provides the first step, by generating hypotheses for unexplored polypharmacology and both target and off-target drug mechanisms.

In our study, we used gene sets to introduce biological knowledge into the analysis. Iorio et al. [8] have recently got promising results with an alternative method of analyzing gene expression responses. It is an interesting and non-straightforward research question whether that approach can be generalized to searching for structure-response relationships.

We have also demonstrated the use of advanced visualization methods to facilitate detailed interpretation of the chemical and biological characteristics of the components. Our findings show connections between the biological responses of many known drug groups to their general chemical properties (Table 1). As an example of the ability of the model to discover detailed drug response mechanisms we were able to separate different DNA damage responses that appear to be driven by different chemical features in compound sets having considerable overlap. Subcomponents 2B, 3A, and 10A all contain classic chemotherapeutic and DNA damaging agents as described in the results section. Strikingly, subcomponents 2B and 10A are driven primarily by hydrophobic/lipophilic descriptors and are more similar in their biological output. They both connect to the downregulation of many proto-oncogenic and mitotic genes but notably still through almost entirely non-overlapping gene sets and genes. Subcomponent 3A, on the other hand, which is connected to hydrogen bonding and hydrophilic features, connects to a very different cellular response: the turning on of many stress-induced “defense” genes. In other words, we document how within the same compound or related compounds, hydrophobic and size features drive a mitotic arrest response while hydrogen bonding and hydrophilic features drive a reparative response. This knowledge, in combination with gene expression data in the solid tumors may allow us to design and utilize the chemotherapeutic agents with the appropriate balance of hydrophilic, size and hydrogen bonding for each cancer patient to hit the correct balance between anti-growths to damage response induction for best possible efficacy.

Methods

Gene expression data

We used the Connectivity Map gene expression profile data set as biological response profiles to drug treatments, forming the *biological space.* Instead of the rank-based procedure of the original Connectivity Map paper, we used a different preprocessing method since ranking amplifies noise. Even small differences in low intensities, which contain mostly noise are ranked, and this has a significant impact on the identification of differentially expressed genes. Hence, we downloaded the raw data files in original CEL-format, from http://www.broadinstitute.org/cmap/, which we RMA-normalized [23] before computing differential expression. We used expression profiles from the most abundant microarray platform (HT-HG-U133A) in the data collection. Differential expression was then taken with respect to the control measurements in each measurement batch. In the case of multiple controls per batch, we formed a more robust control by removing as an outlier the control with the highest Euclidean distance to the other controls, and then used the mean of the rest as the controls. To further reduce the noise in the expression data, we discarded 5% of the genes having the highest variance in the control measurements, that is, genes having high level of variation unrelated to chemical responses. As simple means of balancing between the

varying sample sizes for different chemicals in the CMap data, we chose for each chemical the one instance with the strongest effect, measured with the highest norm of response, on the cell line for further analysis. The resulting gene expression data consisted of profiles for 1159 compounds over 11,350 genes.

To bring in prior knowledge of biological responses, and to reduce the dimensionality of the gene expression data, we performed Gene Set Enrichment Analysis (GSEA). GSEA gives as output for each gene set the direction (positive/negative) and strength of the activity, as measured by the false discovery rate (FDR) q-values, ranging from 0 (indicating highest activity) to 1. We transformed the q-values for the CCA by first inverting the q-values such that 1 indicates the highest activity, and then further mirroring q-values for the negatively activated gene sets with respect to zero to take the sign of activity into account. This results in a reasonably unimodal distribution of the data around zero, with higher positive and negative values indicating higher positive and negative activation of the gene sets, respectively. In the resulting data we have biological activation profiles over 1321 gene sets for 1159 distinct chemicals (see Figure 1B).

As the gene sets, we used the C2 collection (curated gene sets v2.5) from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb/). The extensive collection of gene sets covered 90% genes in our data and is thus a sensible representation of the biological space. GSEA was computed with the Java software package version 2–2.05 (http://www.broadinstitute.org/gsea).

Chemical descriptors

The *chemical space* was formed by representing each chemical with a set of descriptors of its structure and function. In the analysis, the chemical similarity will be dependent on the selected descriptors and thus the selection is of utmost importance. This is especially true when the aim is to find small molecules that share targets and biological functions regardless of structural similarity. We use the VolSurf descriptors, calculated using MOE version 2009.10 (http://www.chemcomp.com/software.htm). Original sdf-files were translated into 3D using Maestro/LigPrep (Maestro version 9.0) since VolSurf descriptors are based on 3D molecular fields. The resulting data contains 76 descriptors for each chemical. Additional file 5: VolSurf_Classification.xls lists these descriptors.

Canonical correlation analysis

Drug action mechanisms are indirectly visible in relationships between the chemical properties of the drug molecules and the biological response profiles. We carry out a data-driven search for such relationships with a method that searches for correlated components in the two spaces, as shown in Figure 1.

Canonical Correlation Analysis (CCA [9]) is a multivariate statistical model for studying the interrelationships between two sets of variables. CCA explores correlations between the two spaces whose role in the analysis is strictly symmetric, whereas classical regression approaches like Partial Least Squares [24] typically explain one or possibly several response variables in one space by a set of independent variables in the other one. The result of the CCA analysis is an underlying component subspace relating chemical descriptors with gene sets.

Let us consider two matrices X and Y, of the size of n x p and n x q, representing the chemical and biological spaces. The rows represent the samples and the columns are the features (gene set activation values or chemical descriptors). In the following we describe the CCA learning algorithm as a stepwise process.

First, two projection vectors w_1 and v_1 are sought such that they maximize the correlation P_1 between components of data formed by,

$$P_1 = \max \mathrm{cor}_{w_1,v_1}\left(Xw_1, Yv_1\right),$$

subject to the constraint that the variance of the components is normalized, *i.e.*, $\mathrm{var}(Xw_1) = \mathrm{var}(Yv_1) = 1$.

The resulting linear combinations Xw_1 and Yv_1 are called the first canonical variates or components, and P_1 is referred to as the first canonical correlation. The first canonical variates explain the maximum possible shared variance of the two spaces along a single linear pair of projections: w_1 and v_1.

The next canonical variates and correlations can be found as follows. For each successive step $s = 2,3,\ldots\min(p,q)$, the projection vectors (w_s,v_s), can be found by maximizing

$$P_s = \max \mathrm{cor}_{w_s,v_s}\left(Xw_s, Yv_s\right),$$

subject to the constraint $\mathrm{var}(Xw_s) = \mathrm{var}(Yv_s) = 1$, and with a further constraint of uncorrelatedness between different components.

Let $C_{xx} = XX^T$ and $C_{yy} = YY^T$ denote the scaled sample covariance matrices for the two input spaces, and $C_{xy} = XY^T$ the sample cross-covariance. Then the first canonical correlation is

$$P_1 = w_1^T C_{xy} v_1 \Big/ \sqrt{w_1^T C_{xx} w_1}\sqrt{v_1^T C_{yy} v_1}\,.$$

If C_{xx} and C_{yy} are invertible the vectors w_1 and v_1 maximizing the above equation can be found. Generally, in omics data and also in our study, the number of genes/gene sets is large compared to the number of experiments. In such cases the classical CCA solution may not exist or it can be very sensitive to collinearities among the variables. This issue can be addressed by introducing regularization [25-27], *i.e.*, by penalizing the norms of the associated vectors. Hence, we seek projection vectors that maximize the penalized correlation

$$P_1 = w_1^T C_{xy} v_1 \Big/ \sqrt{w_1^T C_{xx} w_1 + L_1\|w_1\|}\sqrt{v_1^T C_{yy} v_1 + L_2\|v_1\|}\,.$$

The regularization coefficients L_1 and L_2 were estimated with 20-fold cross validation over a grid of values, while maximizing the retrieval performance on known drug properties. The retrieval procedure and performance measure are described in the drug similarity validation section below. In each fold, the model was first applied to a training data set, and the test data were then projected to the obtained components. Estimated regularization parameter values were $L_1 = 100$ and $L_2 = 0.001$. We used R-package “CCA” [26].

Drug similarity validation procedure

To quantitatively validate the performance of the component model in extracting functionally similar drugs, we carried out the following analysis. For a given data set, we first computed pairwise similarities of drugs. In practice, we used each chemical in turn as a query, and ranked the other chemicals based on their similarity to the query. For the similarity measure, we had three alternatives, similarity in the CCA component space, in the biological space, and in the chemical space. Finally, we computed the average precision of retrieving chemicals that are functionally similar to the query, *i.e.* share at least one known property in an external validation set. We report the mean of the average precisions for all chemicals. We repeat the results as a function of the number of the top ranked chemicals used to compute the average precision (from 5 to 100).

We constructed the external validation set about the functional similarity of the drugs from their known protein targets and ATC (Anatomical Therapeutic Chemical, http://www.whocc.no/atc_ddd_index/) codes. Drug target information was obtained from ChEMBL (https://www.ebi.ac.uk/chembl/), DrugBank (http://www.drugbank.ca/), DUD (http://dud.docking.org/), and ZINC (http://zinc.docking.org/). We additionally extracted targets and ATC codes for the CMap chemicals from the supplementary material provided in [8]. From the ATC codes we used the fourth level information, indicating the chemical/therapeutic/pharmacological subgroup and hence high similarity of drugs. In total we have 4427 associations between 821 CMap chemicals and 796 targets or ATC codes.

Visualizing through an “eye diagram”: relationship between gene sets, extracted components, and VolSurf descriptors

The CCA components summarize statistical relationships between the chemical and biological spaces. The relationships can be visualized as in Figure 1 and 3: The components in the middle are connected to the chemical descriptors that activate them (left) and to the gene sets that are differentially expressed when the component is active. We selected the top 10 significant components from the CCA model for the visualizations. The significances of the components were estimated by a permutation test, using the observed correlations as a test statistic. The samples in one of the spaces were randomly rearranged removing the relationship with the other space. One thousand such random permutations were formed and their canonical correlations computed. Component significances were then determined as the proportion of random correlations that are greater than the observed correlation.

The components were further split into two subcomponents labeled ‘A’ and ‘B’, in A the canonical scores are positive and in B negative. Compounds in the two subcomponents behave in the opposite fashion on the gene sets and VolSurf features, such that when one of the subcomponents activates some biological processes, the other either has no effect or deactivates them. For visual clarity the eye diagram shows only the top 10 correlated gene sets for each component, out of the 1321 gene sets used. All 76 VolSurf features are shown. The eye diagram was originally introduced in [28] for visualizing component models.

Differentially expressed genes and GO enrichment

To get a comprehensive view of the biology in each component we extracted the genes and Gene Ontology classes active in each one of them. For each component, we took the top 20 positively and top 20 negatively correlated gene sets and listed the genes in them. The differential expression of these genes was tested for activation/repression of the gene in the top 10 active compounds in the component using a regularized *t*-test [29]. The genes having p-values < 0.05 were considered to be significantly activated by the compounds in the component. This procedure ensures that the extracted genes are most representative of the top correlated gene sets in the component, hence contributing the most to the canonical correlation.

The component-specific list of significantly differentially expressed genes was used to compute the corresponding Gene Ontology Enrichment for each component. Enrichment was computed for Biological Process classes using GOstats R-package (www.bioconductor.org/help/bioc-views/release/bioc/html/GOstats.html). Additional file 2: Top_GOTerms.xls lists the top 10 significant GO terms for each component while Additional file 3: Top_Genes.xls lists the top 30 significantly differentially expressed genes.

Target enrichment procedure

The target enrichment analysis of each subcomponent was done based on sharing of known targets. Specifically, the shared targets of top 20 compounds of each subcomponent was compared to target sharing in 1000 random draws of the same number of compounds. P-value is given by the proportion of enriched targets in the random samples. Additional file 6: Target_Breakdown.xls lists the common targets and p-values for each subcomponent.

Drug Target data was obtained from ChEMBL (https://www.ebi.ac.uk/chembl/), DrugBank (http://www.drugbank.ca/), DUD (http://dud.docking.org/), and ZINC (http://zinc.docking.org/). We additionally extracted targets from the supplementary material provided in [8]. In total 716 CMap chemicals had target information.

Characterizing drug response on breast cancer cells

We investigated if the components reveal interesting patterns in the responses to drugs, by plotting the transitions caused by each drug in the gene subspace defined by the component. This was done by extracting the 100 most significant genes as an effective representative of changes caused by treatments in the genome (using the procedure described in the above sub-section). The profiles of 30 independent cell lines in a steady-state, unperturbed conditions, were included to act as references for calibrating the display. These independent breast cancer cell lines were obtained from ArrayExpress experiment ID E-MTAB-37 (www.ebi.ac.uk/arrayexpress) with replicates merged to make a single representation for each of the cell types. All cell lines were annotated as BasalA, BasalB, Luminal, or progenitor using classifications by Kuemmerle et al.,[30]. Only MCF7 (breast cancer) treatments were used from CMap data.

The breast cancer cell line and CMap data come from different Affymetrix platforms, HG-U133plus_2.0 and HT-HG-U133A, respectively. We therefore normalized them separately by computing differential expression as the expression value divided by the mean of each gene within the platform. These normalized data were scaled using log2.

Both the CMap-selected instances and breast cancer cell data were collected into a single matrix. To visualize the transitions, pairwise correlation similarities were computed over this matrix and plotted using the state-of-the-art non-linear dimensionality reduction and visualization tool NeRV [18,19]. The result is a mapping of the high-dimensional expression profiles to a two-dimensional display for easier visualization, such that if two points are similar in the visualization, they can be trusted to have been similar before the projections also. NeRV visualization of component 3A, which is analyzed in the Results, is shown in Figure 5.

Authors' contributions

SuK carried out the GSEA, CCA, Gene/GO Enrichment while participating in preprocessing the data, retrieval and NeRV studies and plotting heatmaps. AF worked on model selection and statistical significance tests for the CCA, developed the initial versions for the heatmaps, the eye diagram and coordinated the first writing of the manuscript. J-PM participated in pre-processing the Connectivity Map data and NeRV study. JP plotted the Eye Diagram and participated in NeRV and retrieval studies. TK calculated the VolSurf descriptors used in the study. The PI's AP, OK, KW and SK supervised the study. All authors participated in planning, writing, and read and approved the final manuscript.

Acknowledgements

SuK, AF, JP, and SaK belong to the Adaptive Informatics Research Center at Aalto University School of Science. This work was supported by the Academy of Finland [Center of Excellence funding no. 213502 (Translational Genome-scale Biology) and funding no. 140057 (Computational modeling of the biological effects of chemicals)], the PASCAL2 Network of Excellence, ICT [216886], Sigrid Juselius Foundation, Cancer Society of Finland, the Jane and Aatos Erkko Foundation, the HBGS graduate school (J-PM), the FICS graduate school (SuK and AF), and the HECSE graduate school (JP).

References

1. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL: **Predicting new molecular targets for known drugs.** *Nature* 2009, **462:**175–191.

2. Iskar M, Campillos M, Kuhn M, Jensen LJ, Noort VV, Bork P: **Drug-Induced Regulation of Target Expression.** *PLoS Computational Biology* 2010, **6(9)**.

3. Cramer RD III, Patterson DE, Bunce JD: **Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins.** *J. Am. Chem. Soc.* 1988, **110(18)**.

4. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR: **The Connectivity Map: Using Gene-**

Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* 2006, **313(5795):**1929–1935.

5. Hopkins AL: **Network pharmacology: the next paradigm in drug discovery.** *Nature Chemical Biology* 2008, **4(11):**682–690.

6. Tatonetti NP: **Predicting drug side-effects by chemical systems biology.** *Genome Biology* 2009, **10(238)**.

7. Spicker JS, Brunak S, Frederiksen KS, Toft H: **Integration of Clinical Chemistry, Expression, and Metabolite Data Leads to Better Toxicological Class Separation.** *Toxicological Sciences* 2008, **102(2):**444–454.

8. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D: **Discovery of drug mode of action and drug repositioning from transcriptional responses.** *PNAS* 2010, **107(33):**14621–14626.

9. Hotelling H: **Relations between two sets of variants.** *Biometrika* 1936, **28:**321–327.

10. Atias N, Sharan R: **An Algorithmic Framework for Predicting Side-Effects of Drugs.** *Journal of Computational Biology* 2011, **18(3):**207–218.

11. Cruciani G, Pastora M, Gubab W: **VolSurf, a new tool for the pharmacokinetic optimization of lead compounds.** *Eur. J. Pharm. Sci* 2000, **11(2):**29–39.

12. Naylor E, Arredouani A, Vasudevan SR, Lewis AM, Parkesh R, Mizote A, Rosen D, Thomas JM, Izumi M, Ganesan A, Galione A, Churchill GC: **Identification of a chemical probe for NAADP by virtual screening.** *Nature Chemical Biology* 2009, **5(4):**220–226.

13. Blower PE, Yang C, Fligner MA, Verducci JS, Yu L, Richman S, Weinstein J: **Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data.** *The Pharmacogenomics Journal* 2002, **2(4):**259–271.

14. Cheng T, Wang Y, Bryant SH: **Investigating the correlations among the chemical structures, bioactivity profiles and molecular targets of small molecules.** *Bioinformatics* 2010, **26(22):**2881–2888.

15. Wang J, Wang J, Dai J, Jung Y, Wei CL, Wang Y, Havens AM, Hogg PJ, Keller ET, Pienta KJ, Nor JE, Wang CY, Taichman RS: **A glycolytic mechanism regulating an angiogenic switch in prostate cancer.** *Cancer Res.* 2007, **67(1):**149–59.

16. Gessner T, Vaughan LA, Beehler BC, Bartels CJ, Baker RM: **Elevated pentose cycle and glucuronyltransferase in daunorubicin-resistant P388 cells.** *Cancer Res* 1999, **50(13):**3921–7.

17. Scoumanne A, Chen X: **Protein methylation: a new mechanism of p53 tumor suppressor regulation.** *Histol Histopathol* 2008, **23(9):**1143–9.

18. Venna J, Peltonen J, Nybo K, Aidos H, Kaski S: **Information retrieval perspective to nonlinear dimensionality reduction for data visualization.** *Journal Machine Learning Research* 2010, **11:**451–490.

19. Venna J, Kaski S: **Nonlinear dimensionality reduction as information retrieval.** In Marina Meila and Xiaotong Shen, editors, Proceedings of *AISTATS* 2007, the 11th International Conference on Artificial Intelligence and Statistics. Omnipress 2007. JMLR Workshop and Conference Proceedings, Vol. 2: AISTATS 2007.

20. Prassas I, Diamandis EP: **Novel therapeutic applications of cardiac glycosides.** *Nat Rev Drug Discov* 2008, **7(11):**926–35.

21. Vasilyeva A, Clodfelter JE, Gorczynski MJ, Gerardi AR, King SB, Salsbury F, Scarpinato KD: **Parameters of Reserpine Analogs That Induce MSH2/MSH6-Dependent Cytotoxic Response.** *J Nucleic Acids* 2010, **2010**(2010):162018.

22. Sanz M, Burnett A, Lo-Coco F, Löwenberg B: **FLT3 inhibition as a targeted therapy for acute myeloid leukemia.** *Curr Opin Oncol* 2009, **21(6):**594–600.

23. Irizarry RA: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Research* 2003, **31(4)**.

24. Wold H: *Estimation of principal components and related models by iterative least squares*. Analysis: Multivariate; 1966:391–420.

25. Leurgans SE, Moyeed RA, Silverman BW: **Canonical Correlation Analysis when the Data are Curves.** *Journal of the Royal Statistical Society B* 1993, **55(3):**725–740.

26. Gonzalez I, Dejean S, CCA: **An R Package to extend Canonical Correlation Analysis.** *Journal of Statistical Software* 2008, **23(12)**.

27. Bickel P, Li B: **Regularization in Statistics.** *TEST* 2006, **15(2):**271–344.

28. Caldas J, Gehlenborg N, Faisal A, Brazma A, Kaski S: **Probabilistic retrieval and visualization of biologically relevant microarray experiments.** *Bioinformatics* 2009, **25(12):**145–153. ISMB/ECCB.

29. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17:**509–519.

30. Kuemmerle NB, Rysman E, Lombardo PS, Flanagan AJ, Lipe BC, Wells WA, Pettus JR, Froehlich HM, Memoli VA, Morganelli PM, Swinnen JV, Timmerman LA, Chaychi L, Fricano CJ, Eisenberg BL, Coleman WB, Kinlaw WB: **Lipoprotein lipase links dietary fat to solid tumor cell proliferation.** *Mol Cancer Ther* 2011, **10**(3): 427–436. doi:10.1158/1535-7163.MCT-10-0802.

Additional files

Additional_file_1 as XLS

Additional file 1: Table S1. Top_Compounds.xls. Contains lists of Top 20 active compounds in each subcomponent

Additional_file_2 as XLS

Additional file 2: Table S2. Top_GOTerms.xls. Contains lists of Top 10 significant GO Terms in each component

Additional_file_3 as XLS

Additional file 3: Table S3. Top_Genes.xls. Contains lists of Top 30 significant Genes in each subcomponent

Additional_file_4 as PDF

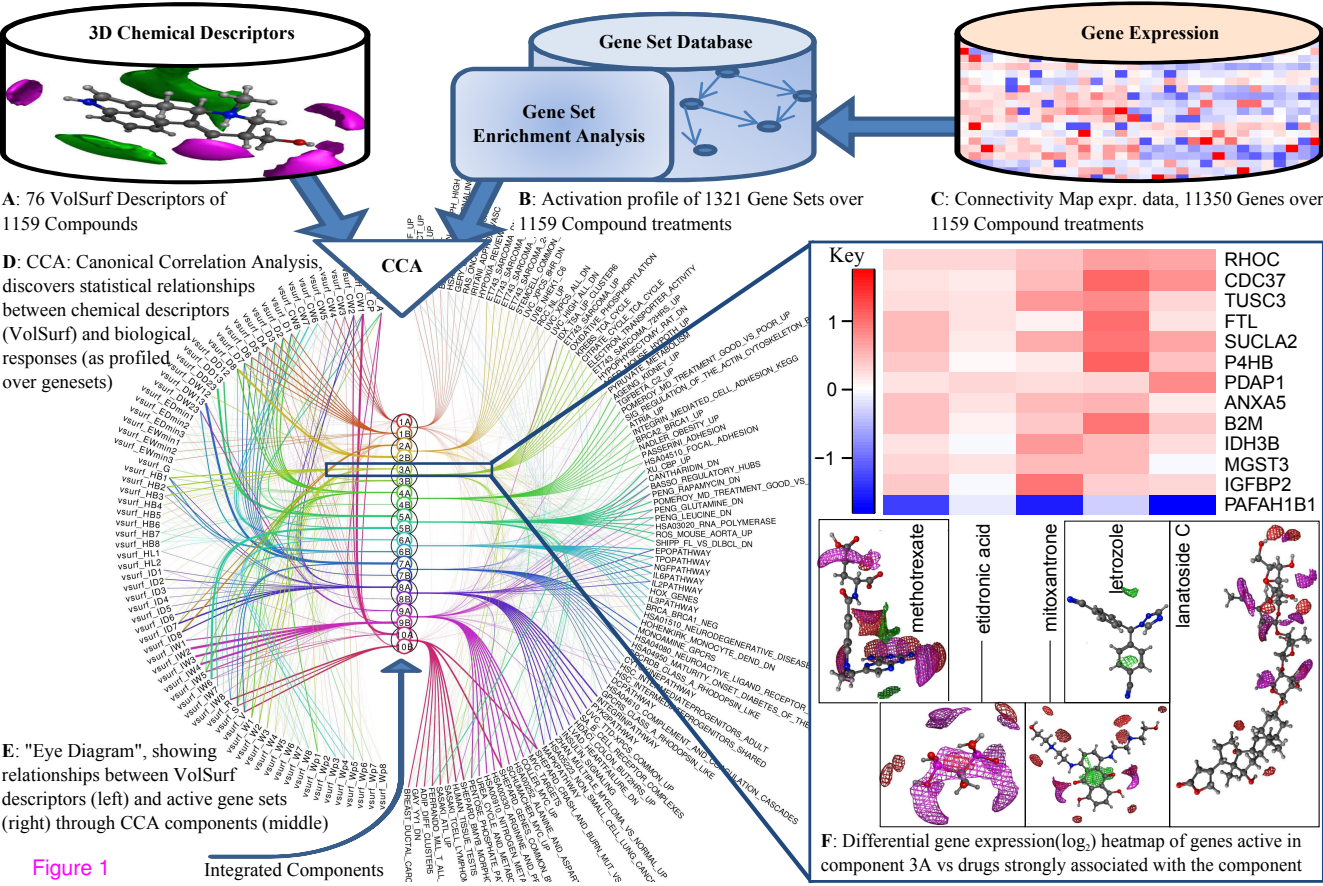
Additional file 4: Figure S1. HeatMaps.pdf. Contains breast cancer patient heat maps and chemical/biological composition plot

Additional_file_5 as XLS

Additional file 5: Table S4. VolSurf_Classification.xls. Contains lists of all VolSurf features and their pharmacophoric classification

Additional_file_6 as XLS

Additional file 6: Table S5. Target_Breakdown.xls. Contains list of commonly occurring protein targets in each subcomponent



A: 76 VolSurf Descriptors of 1159 Compounds

B: Activation profile of 1321 Gene Sets over 1159 Compound treatments

C: Connectivity Map expr. data, 11350 Genes over 1159 Compound treatments

D: CCA: Canonical Correlation Analysis, discovers statistical relationships between chemical descriptors (VolSurf) and biological responses (as profiled over genesets)

E: "Eye Diagram", showing relationships between VolSurf descriptors (left) and active gene sets (right) through CCA components (middle)

F: Differential gene expression($\log_2$) heatmap of genes active in component 3A vs drugs strongly associated with the component

Figure 1

Integrated Components

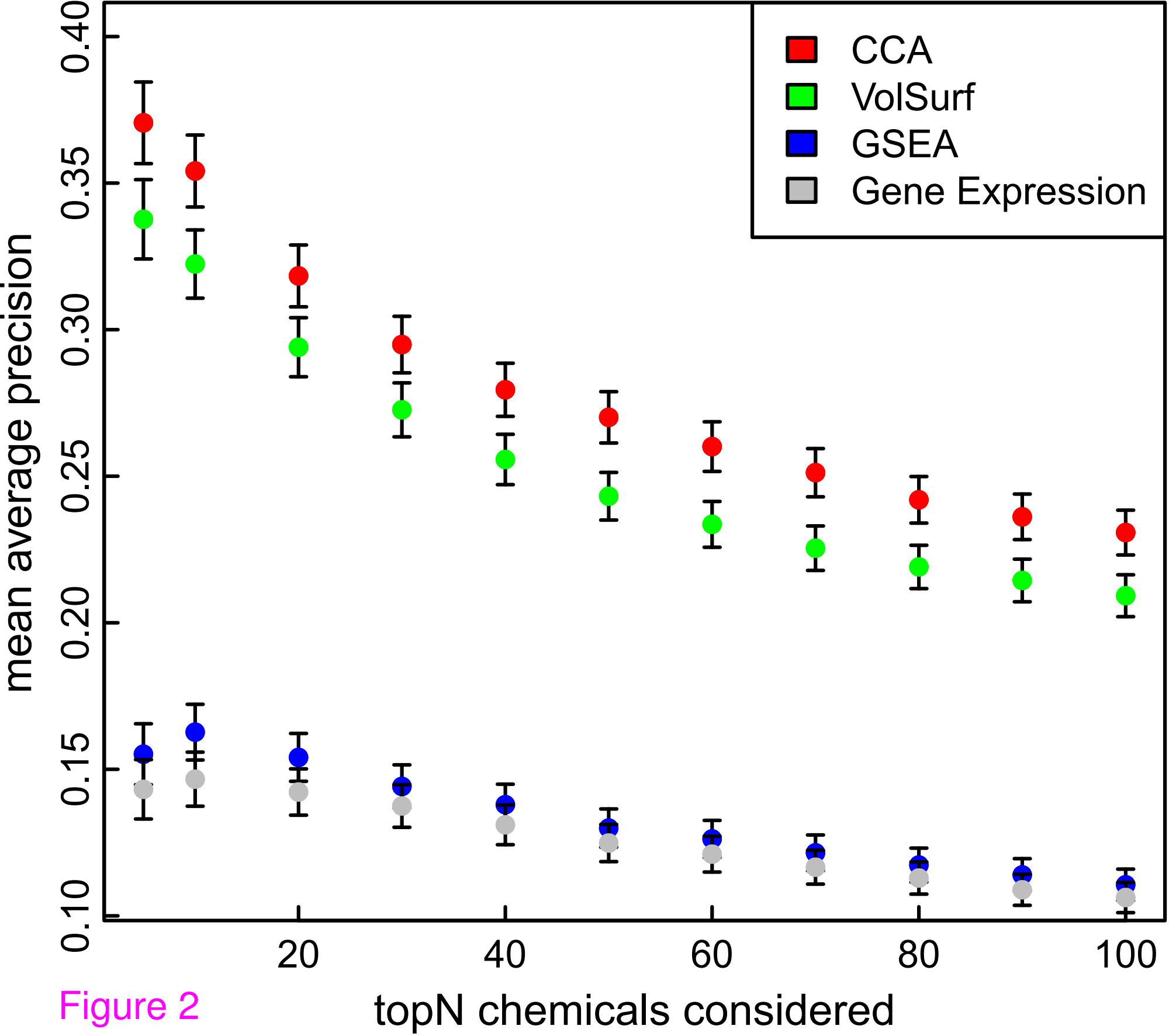
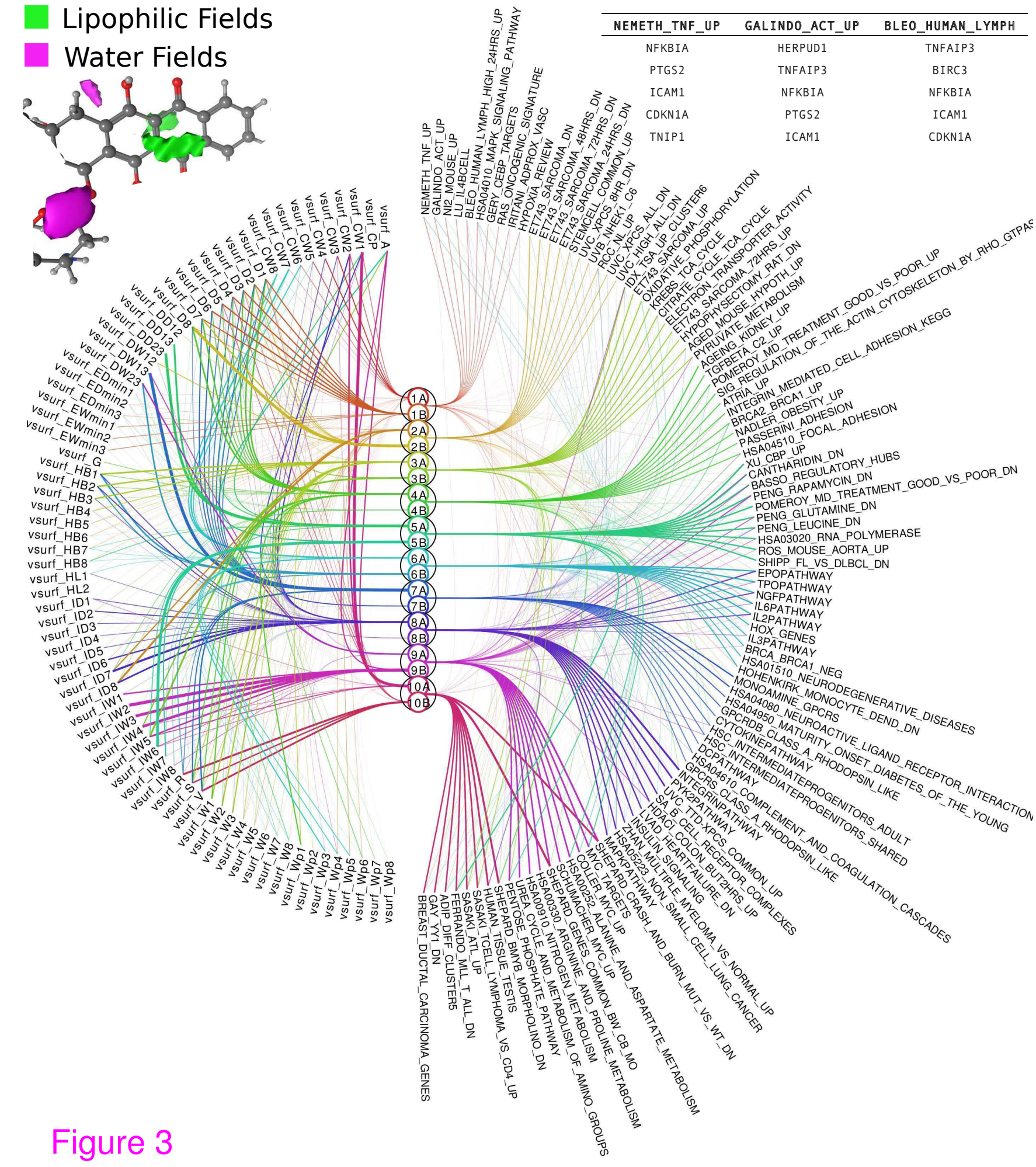

CCA
VolSurf
GSEA
Gene Expression
mean average precision
0.10
0.15
0.20
0.25
0.30
0.35
0.40
20
40
60
80
100
topN chemicals considered

Figure 2



NEMETH_TNF_UP	GALINDO_ACT_UP	BLEO_HUMAN_LYMPH
NFKBIA	HERPUD1	TNFAIP3
PTGS2	TNFAIP3	BIRC3
ICAM1	NFKBIA	NFKBIA
CDKN1A	PTGS2	ICAM1
TNIP1	ICAM1	CDKN1A

Figure 3

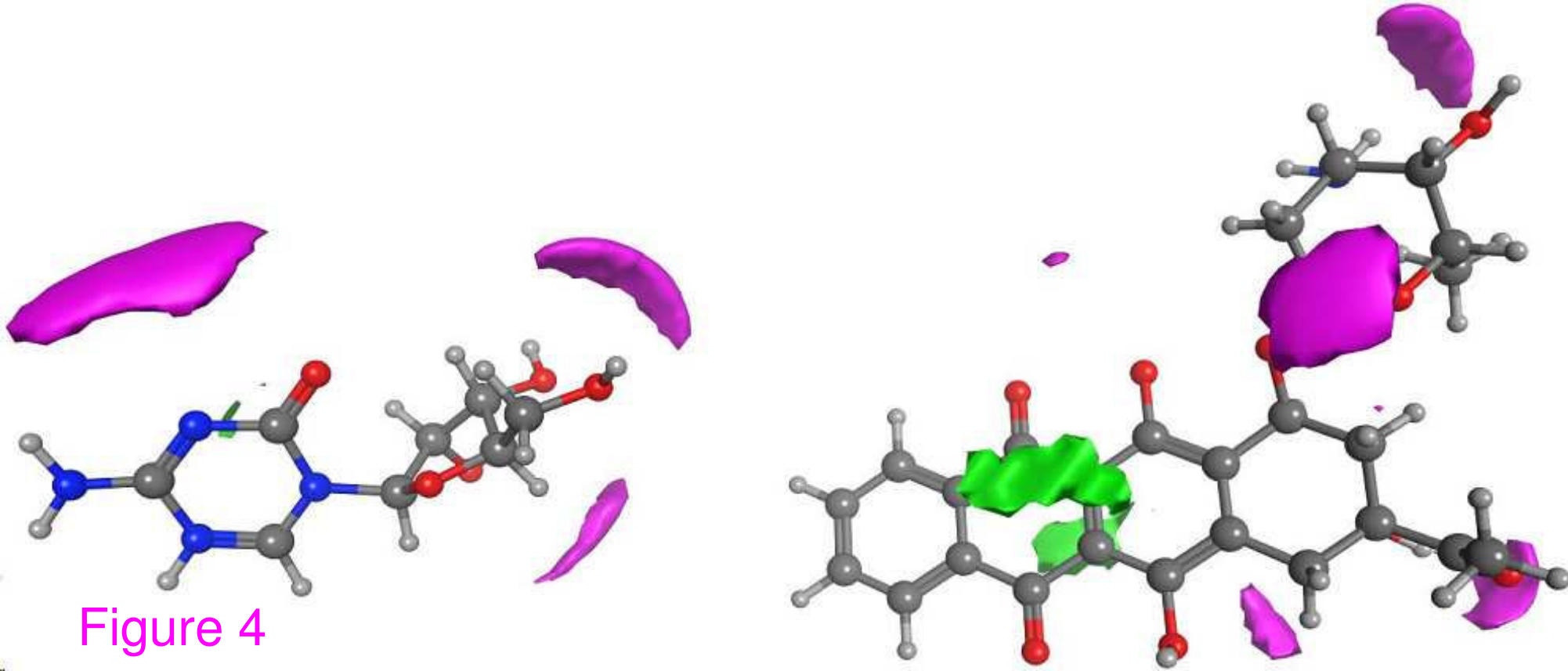


Figure 4

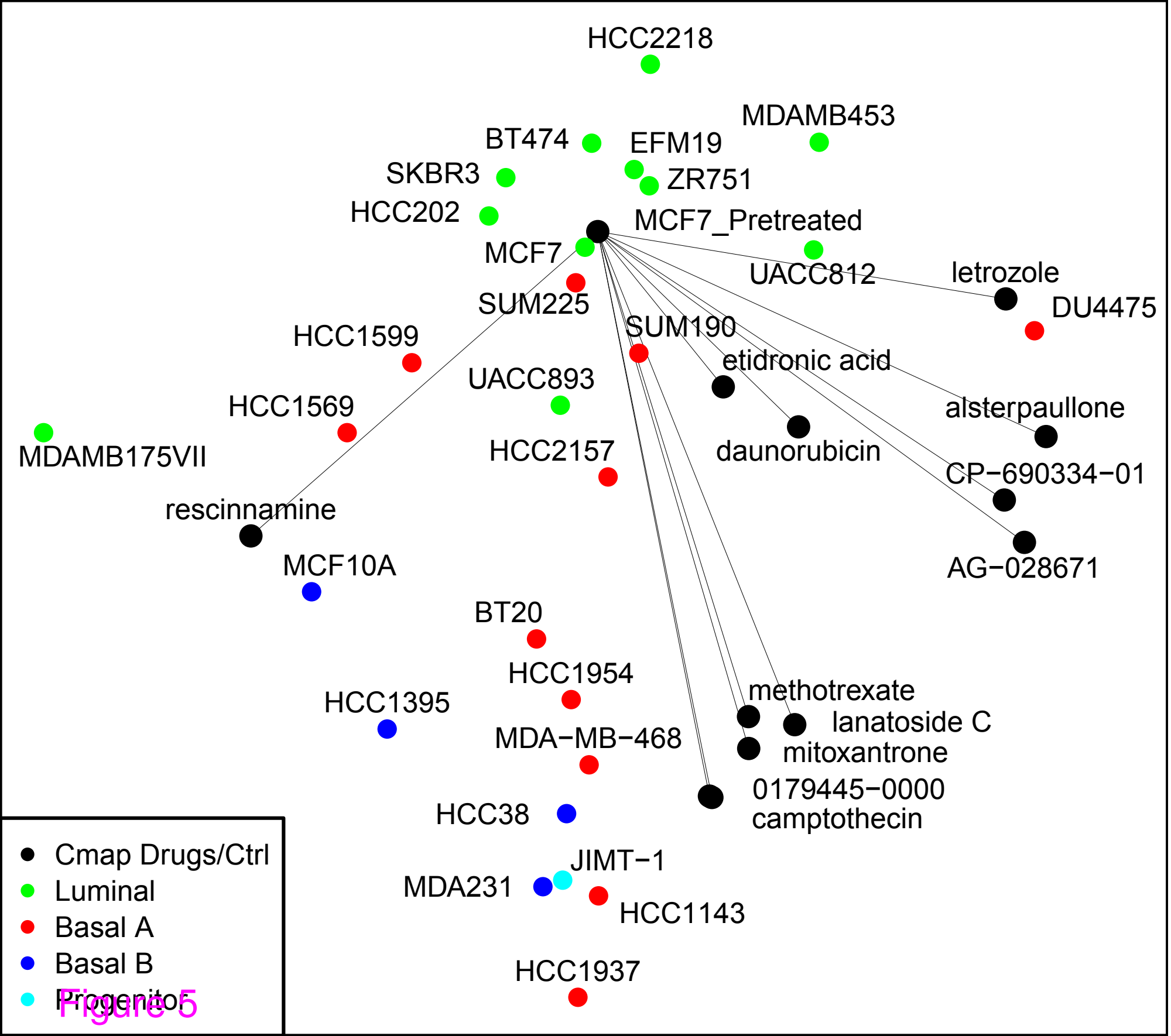
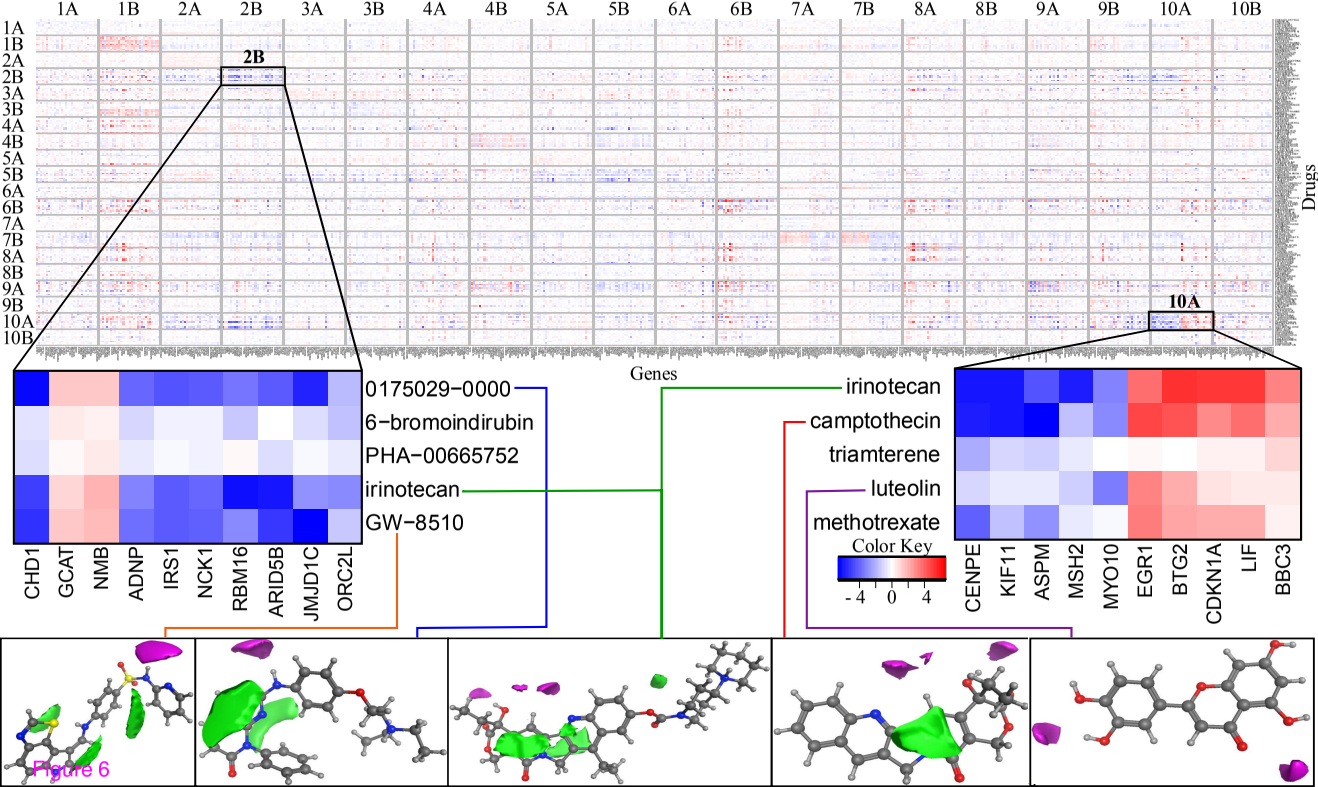

HCC2218
MDAMB453
BT474
EFM19
SKBR3
ZR751
HCC202
MCF7_Pretreated
MCF7
UACC812
letrozole
SUM225
DU4475
HCC1599
SUM190
etidronic acid
UACC893
HCC1569
alsterpaullone
MDAMB175VII
HCC2157
daunorubicin
CP−690334−01
rescinnamine
AG−028671
MCF10A
BT20
HCC1954
methotrexate
HCC1395
lanatoside C
MDA−MB−468
mitoxantrone
0179445−0000
HCC38
camptothecin
JIMT−1
MDA231
HCC1143
HCC1937
Cmap Drugs/Ctrl
Luminal
Basal A
Basal B
Progenitor

Figure 5


1A
1B
2A
2B
3A
3B
4A
4B
5A
5B
6A
6B
7A
7B
8A
8B
9A
9B
10A
10B
Drugs
Genes
0175029-0000
6-bromoindirubin
PHA-00665752
irinotecan
GW-8510
CHD1
GCAT
NMB
ADNP
IRS1
NCK1
RBM16
ARID5B
JMJD1C
ORC2L
irinotecan
camptothecin
triamterene
luteolin
methotrexate
Color Key
-4
0
4
CENPE
KIF11
ASPM
MSH2
MYO10
EGR1
BTG2
CDKN1A
LIF
BBC3
Figure 6

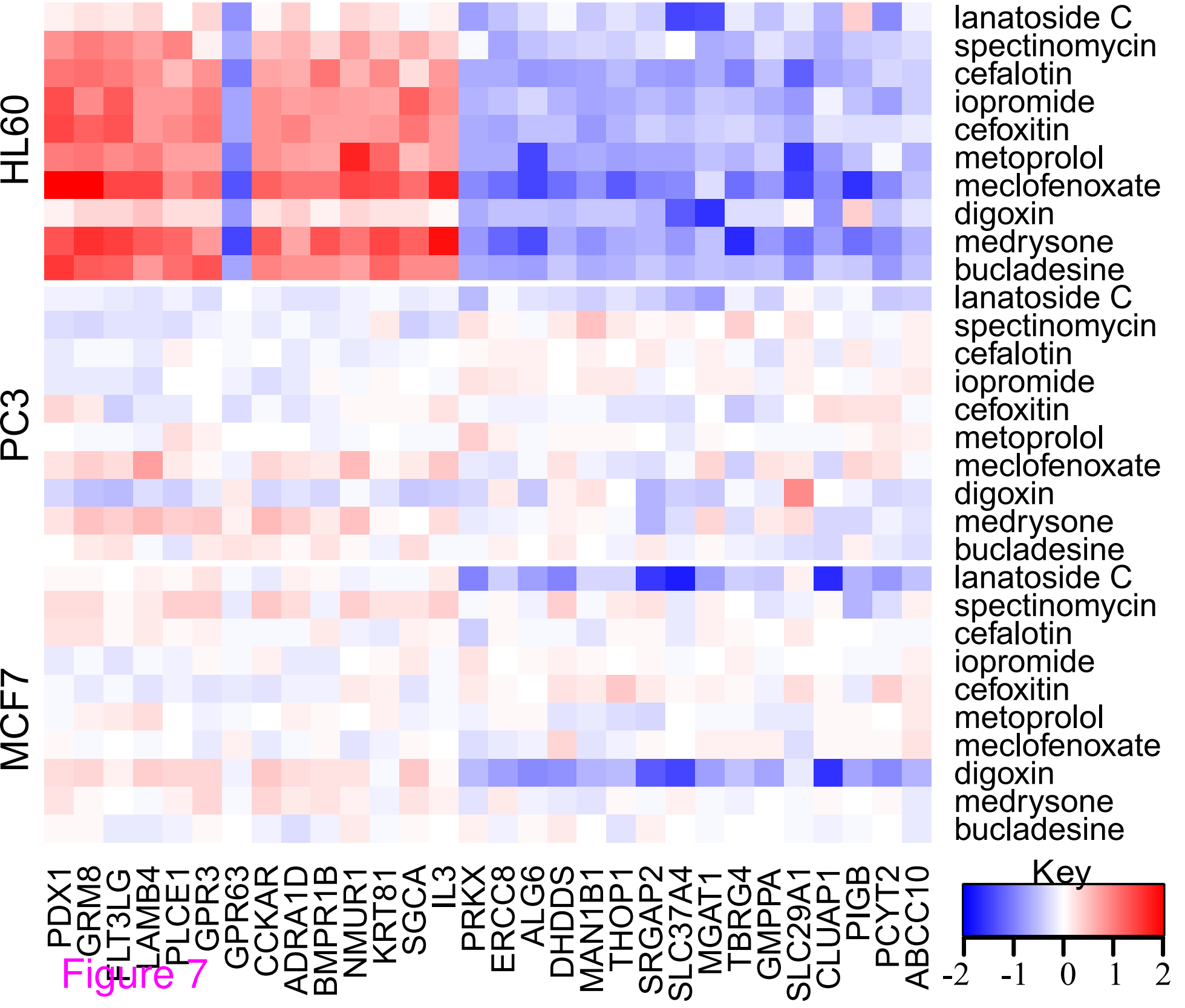

HL60
PC3
MCF7
lanatoside C
spectinomycin
cefalotin
iopromide
cefoxitin
metoprolol
meclofenoxate
digoxin
medrysone
bucladesine
PDX1
GRM8
LT3LG
LAMB4
PLCE1
GPR3
GPR63
CCKAR
ADRA1D
BMPR1B
NMUR1
KRT81
SGCA
IL3
PRKX
ERCC8
ALG6
DHDDS
MAN1B1
THOP1
SRGAP2
SLC37A4
MGAT1
TBRG4
GMPPA
SLC29A1
CLUAP1
PIGB
PCYT2
ABCC10
Key
-2
-1
0
1
2

Figure 7

Additional files provided with this submission:

Additional file 1: Top_Compounds.xls, 38K
http://www.biomedcentral.com/imedia/1754931398591280/supp1.xls
Additional file 2: Top_GOTerms.xls, 36K
http://www.biomedcentral.com/imedia/2083859901591280/supp2.xls
Additional file 3: Top_Genes.xls, 50K
http://www.biomedcentral.com/imedia/3315415835912808/supp3.xls
Additional file 4: HeatMaps.pdf, 721K
http://www.biomedcentral.com/imedia/1276199475912808/supp4.pdf
Additional file 5: VolSurf_Classification.xls, 31K
http://www.biomedcentral.com/imedia/3207988305912808/supp5.xls
Additional file 6: Additional file 6_Target_Breakdown.xls, 35K
http://www.biomedcentral.com/imedia/1918131168678035/supp6.xls