PLOS ONE

GSCMap - A Gene-Set-based version of Connectivity Map and an application using Local Hierarchical Clustering to characterize Bioactive Compounds in terms of Biological Functional Groups

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Article
Full Title:	GSCMap - A Gene-Set-based version of Connectivity Map and an application using Local Hierarchical Clustering to characterize Bioactive Compounds in terms of Biological Functional Groups
Short Title:	GSCMap for the therapeutic characterization of bioactive compounds
Corresponding Author:	Hoong-Chien Lee, Ph.D. Chung Yuan Christian University Zhongli City, TAIWAN
Keywords:	Genome-wide gene expression; functional gene-sets; gene-set-based analysis; individual gene analysis; functional genomic profile; Connectivity Map; Gene-Set Connectivity Map; Gene-Set Local Hierarchical Clustering; drug repurposing; functional drug classification; histone deacetylase inhibitor; cyclin-dependent kinase inhibitor; anti-cancer; antibiotic; anesthetic agent; anti-inflammatory agent
Abstract:	Gene-set-based analysis (GSA), which uses the relative importance of functional gene-sets as units for analysis of genome-wide gene expression data, has exhibited major advantages with respect to greater accuracy, robustness, and biological relevance, over individual gene analysis (IGA), which uses log-ratios of individual genes for analysis. Yet IGA remains the dominant mode of analysis of gene expression data. The Connectivity Map (CMap), an extensive database on genomic profiles of effects of drugs and small molecules and widely used for studies related to repurposed drug discovery, has been mostly employed in IGA mode. Here, we constructed a GSA-based version of CMap, Gene-Set Connectivity Map (GSCMap), in which all the genomic profiles in CMap are converted, using gene-sets from the Molecular Signatures Database, to functional profiles. We showed that GSCMap essentially eliminated cell-type dependence, a weakness of CMap in IGA mode, and yielded significantly better performance on sample clustering and drug-target association. As a first application of GSCMap we constructed the platform Gene-Set Local Hierarchical Clustering (GSLHC) for discovering insights on coordinated actions of biological functions and facilitating classification of heterogeneous subtypes on drug-driven responses. GSLHC was shown to tightly clustered drugs of known similar properties. We used GSLHC to identify the therapeutic properties and putative targets of 18 compounds of previously unknown characteristics listed in CMap, eight of which suggest anti-cancer activities. We expect GSCMap and GSLHC to be widely useful in providing new insights in the biological effect of bioactive compounds, in drug repurposing, and in function-based classification of complex diseases.
Order of Authors:	Feng-Hsiang Chung
	Zhen-Hua Jin
	Tzu-Ting Hsu
	Chueh-Lin Hsu
	Hoong-Chien Lee, Ph.D.
Suggested Reviewers:	Ying Xu, PhD Professor, University of Georgia xyn@bmb.uga.edu Leading bioinformatics and systems biology scholar, cancer specialist
	Andrew Torda, PhD Professor, Hamburg University torda@zbh.uni-hamburg.de

	Leading bioinformatics and computational biology scholar
	Frank Emmert-Streib Professor, Queen's University Belfast f.emmert-streib@qub.ac.uk Expert on computational biology, gene expression studies, pathological pathways of complex diseases
	Yu Xue, PhD Professor, Huazhong University of Science and Technology xueyu@mail.hust.edu.cn Expert on computational studies of post-translational modifications, systems biology, bioinformatics
	Remy K Aziz, PhD Assistant professor, Cairo University ramy.aziz@salmonella.org Expert on genomics, phamarcology
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. A complete funding statement should do the following: Include **grant numbers and the URLs** of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who received the funding. **Describe the role** of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If they had no role in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript*." If the study was **unfunded**, provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work*." * typeset	Grants NSC-102-2911-I-008-001, NSC-100-2911-I-008-001, NSC-100-2627-M-008-004, Ministry of Science and Technology (Republic of China) http://www.most.gov.tw/ Grants 100CGH-NCU-A5, 101CGH-NCU-A5, Ministry of Education Republic of China) http://english.moe.gov.tw/ Grants 10110021-5, 10210061-5, National Central University http://www.ncu.edu.tw/?hl=en; and Cathay General Hospital http://www.cgh.org.tw/en/index.html The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work,	The authors have declared that no competing interests exist.

GSCMap – A Gene-Set-based version of Connectivity Map and an application using Local Hierarchical Clustering to characterize Bioactive Compounds in terms of Biological Functional Groups

Feng-Hsiang Chung[+1,2], Zhen-Hua Jin[+1], Tzu-Ting Hsu[1], Chueh-Lin Hsu[1] and Hoong-Chien Lee*[1-4]

[1]Institute of Systems Biology and Bioinformatics, National Central University, Zhongli, Taiwan 32001
[2]Center for Dynamical Biomarkers and Translational Medicine, National Central University, Zhongli, Taiwan 32001
[3]Department of Physics, Chung Yuan Christian University, Zhongli, Taiwan 32023
[4]Physics Division, National Center for Theoretical Sciences, Hsinchu, Taiwan 30043

[+]FHC (email: fabian415@gmail.com) and ZHJ made equal contributions.
*Corresponding author, HCL, email: hclee12345@gmail.com.

ABSTRACT

Gene-set-based analysis (GSA), which uses the relative importance of functional gene-sets as units for analysis of genome-wide gene expression data, has exhibited major advantages with respect to greater accuracy, robustness, and biological relevance, over individual gene analysis (IGA), which uses log-ratios of individual genes for analysis. Yet IGA remains the dominant mode of analysis of gene expression data. The Connectivity Map (CMap), an extensive database on genomic profiles of effects of drugs and small molecules and widely used for studies related to repurposed drug discovery, has been mostly employed in IGA mode. Here, we constructed a GSA-based version of CMap, Gene-Set Connectivity Map (GSCMap), in which all the genomic profiles in CMap are converted, using gene-sets from the Molecular Signatures Database, to functional profiles. We showed that GSCMap essentially eliminated cell-type dependence, a weakness of CMap in IGA mode, and yielded significantly better performance on sample clustering and drug-target association. As a first application of GSCMap we constructed the platform Gene-Set Local Hierarchical Clustering (GSLHC) for discovering insights on coordinated actions of biological functions and facilitating classification of heterogeneous subtypes on drug-driven responses. GSLHC was shown to tightly clustered drugs of known similar properties. We used GSLHC to identify the therapeutic properties and putative targets of 18 compounds of previously unknown characteristics listed in CMap, eight of which suggest anti-cancer activities. We expect GSCMap and GSLHC to be widely useful in providing new insights in the biological effect of bioactive compounds, in drug repurposing, and in function-based classification of complex diseases.

Keywords: Genome-wide gene expression; functional gene-sets; gene-set-based analysis; individual gene analysis; functional genomic profile; Connectivity Map; Gene-Set Connectivity Map; Gene-Set Local Hierarchical Clustering; drug repurposing; functional drug classification; histone deacetylase inhibitor; cyclin-dependent kinase inhibitor; anti-cancer; antibiotic; anesthetic agent; anti-inflammatory agents

INTRODUCTION

Microarray technique has been a powerful tool for profiling gene expression on a genome-wide scale and to study associations between gene expression and the pathology of common diseases, including various cancers and Alzheimer's disease [1, 2]. A common practice, the Individual Gene Analysis (IGA) of microarrays, focuses on statistics-based identification of differentially expressed genes (DEGs) between two phenotypes. Standard and popular methods of this type include student *t*-test, *z*-test, SAM, Limma, and ANOVA [3-7]. While most biological processes, including metabolic process, signal transduction, and regulation of transcription, typically involve the collaborative activation of large sets of genes, IGA methods emphasize the independence of individual genes and neglect the expected correlations in gene expression.

An improvement on IGA is to explore whether, among IGA-selected DEGs, functionally related gene sets, such as those given by Gene Ontology [8] and KEGG [9], are significantly expressed. An example of this approach is Fisher's exact test [10]. A drawback in this approach is that genes not among DEGs, namely the vast majority of genes, are excluded from the consideration. In the event when the DEG set is large, the correspondingly long list of functional gene-sets makes it cumbersome to compare results between studies. Most importantly, this approach tends to be dominated by large gene-sets, such as those of immune response and metabolic pathways, and results in the neglect of possibly important functions represented by smaller gene-sets.

The Connectivity map (CMap) was first developed as a generic solution for identifying the functional associations between diseases, genes, and drugs [11]. This approach provides a common analytical platform using genomic profiles as a shared language to connect diseases, gene functions, and drug activities. Many studies have employed disease-defined gene-sets to query CMap for the discovery of repurposed drug activities against common diseases, including diabetes [12] and Alzheimer's disease [13, 14], and solid tumours such as colon cancer [15], breast cancer [16], and lung adenocarcinoma [17]. The standard application of CMap has been IGA based [18]. However, results of IGA-based application of CMap on human samples tend to be dominated by cell types (Supporting Information in [11]).

Gene-Set Analysis (GSA) was developed to address the shortcomings of IGA [19,20]. GSA uses sets of genes connected by biological functions, instead of individual genes, as units of analysis. In Gene Set Enrichment Analysis (GSEA) [21], the first GSA method, the relative importance of a functional gene-set is represented by an enrichment score (ES). GSEA was employed to generate a map that links genomic profiles of diseases to corresponding drug responses in CMap [11].

More recent variants of GSEA, including GSA [22], SAFE [23], Catmap [24], ErmineJ [25], and SAM-GS [26], employ variations in matrix ranking, definition for enrichment scores, or scheme for significance estimation. Other methods including FunNet [27], PARADIGM [28], and COFECO [29] are network-based and more sophisticated, but their application may also be limited by the availability of gene-gene interactions. GSA methods have been employed to explore functional relationships in large-scale compendiums of clinical cancer cohort samples and to elucidate associations in drug-driven signatures for therapeutic purposes [30].

In GSA, a genomic profile may be expressed as the set of ESs for a comprehensive list of gene-sets computed from that genomic profile; we shall call that set a functional genomic profile (hereafter, functional profile). Because a functional profile neither relies on an arbitrary threshold for gene selection, as does IGA, nor by definition is it dominated by a few functionalities involving large gene-sets, it is expected to be more accurate and sensitive in reflecting the global as well as detailed properties of a genome-wide gene expression than IGA.

Here, we built Gene-Set Connectivity Map (GSCMap), an enhanced version of CMap where the genomic profiles of drugs in the CMap database are converted to functional profiles. Like CMap, GSCMap may be used for repurposed drug discovery, except that in GSCMap the functional signature of a phenotype is matched to functional profiles of drugs. The goal is to construct a database that one may expect to yield a more robust drug-phenotype association. We conducted tests to establish the internal consistency of GSCMap. We showed that grouping of drugs with similar biological activities is much more robust with GSCMap than with CMap in IGA mode. For an application of GSCMap we developed Gene-set-based Local Hierarchical Clustering (GSLHC), which utilizes an agglomerative hierarchical method for clustering a subset of functional gene-sets associated with "local" drugs responses. The idea is that, given a very large matrix of gene-set enrichment scores, a clear pattern of coordinated expression in sets of functionalities are usually confined to a subgroup of samples, a pattern that may not be easily detected by global measurements [31, 32]. Through GSLHC we identified the therapeutic properties and putative targets of 18 compounds of previously unknown characteristics listed in CMap, placing each in a subclass of drugs grouped by the similarity of the functional response they induce. Eight of the 18 subclasses contain putative anti-cancer activities. Our results revealed novel links in terms of gene-sets, and drug-versus-functions.

Our results showed GSCMap to be a robust and biologically more reliable version of CMap, and GSLHC, in combination with GSCMap, to be useful in discovering linkages among bioactive compounds characterized by their functional properties.

MATERIALS AND METHODS

External database

The CMap database. Four types of human cancer cell lines (MCF7, PC3, HL60, SKMEL5) were treated with 1,309 distinct small-molecules including U.S. Food and Drug Administration (FDA) approved drugs and uncharacterized bioactive compounds (call perturbagens by the authors of CMap, here simplicity referred as drugs), for a total of 6,097 treatments [11]. Gene (total RNA) expressions from the 6,097 "instances" (an instance is a cell line treated with a drug at a dosage, and its non-treated control) were recorded in two batches of microarrays: 671 HG-U133A (Affymetrix) chips (on 407 drugs) and 5,426 HT-HG-U133A chips (for a total of 6,097 chips on 1,309 drugs). These were downloaded from Gene Expression Omnibus (GEO http://www.ncbi.nlm.nih.gov/geo/; accession no. GSE5258). The same data are available from the CMap website (http://www.broadinstitute.org/cmap/).

Gene-set database. We downloaded the annotated 4,884 gene-sets from the Molecular Signatures Database (MSigDB: http://www.broadinstitute.org/gsea/msigdb/index.jsp) [21]. These gene-sets include four types: C2: curated gene-sets from known pathways, online databases, and knowledge of domain experts; C3: motif gene-sets based on conservative cis-regulatory motifs from human, mouse, rat, and dog genomes; C4: computational gene-sets determined by co-expression neighbourhoods centered on 380 cancer-related genes; C5: gene-ontology gene-sets collected from the same GO annotations of genes. C1(positional gene-sets on each human chromosome) was not included in this study for saving the time on big size of gene-sets. For convenience, gene symbols in each gene-set were combined and transformed in HG-U133A Affymetrix ID according to the updated annotation file from Affymetrix website (http://www.affymetrix.com/estore/).

Chemical structure database. In order to cluster compounds based on 3D structure similarity, we queried 1,309 drug names on NCBI PubChem database (http://pubchem.ncbi.nlm.nih.gov/). Next, the retrieved 1,267 compounds (97% of CMap databsets) were hierarchically clustered by *Chemical Structure Clustering* tool based on the 3D structure (fingerprint) similarity using the single linkage algorithm on PubChem website [33]. Finally, we partitioned the tree into K clusters with K ranging from 10 to 200, and evaluated the clustering performance using F-score [34].

Pharmacological classification system. We retrieved class information of 798 compounds (61% of CMap databsets) from the Anatomical Therapeutic Chemical (ATC) classification system in the World Health Organization (WHO) website (http://www.whocc.no/) for information on similar therapeutic classes. In this system, drugs are classified into groups at 5 different levels: the first level of code indicates the anatomical main group; the second level of code indicates the therapeutic main group; the third level of code indicates the therapeutic/pharmacological subgroup; the fourth level of code indicates the chemical/therapeutic/pharmacological subgroup; the fifth level of code indicates the chemical substance. We used the first four levels of ATC to evaluate the gene and gene-set clusters performance using F-score. The fifth level of the code was not included in our analysis because at this level CMap was too fragmented – almost one drug to a class – for the code to be useful.

Molecular target database. We extracted information on known therapeutic protein targets, relevant diseases or cancers, and corresponding drugs (787 drugs; 60% of CMap datasets) from the Therapeutic Target Database (TTD: http://bidd.nus.edu.sg/group/ttd/) [35]. The working types on specific targets by the corresponding drugs (including activator, adduct, agonist, antagonist, antibody, binder, blocker, breaker, cofactor, inducer, inhibitor, intercalator, modulator, multitarget, opener, regulator, stimulator, and suppressor) were simply divided into two major groups: inhibition or activation. Because drugs and targets do not have one-to-one correspondence, we did not calculate F-score based on the small class size. Instead, we computed drug-drug correlations by target group in IGA and GSA. The drug-pair is assumed to have correlation value of 1 if they have similar effects on the same protein target.

Local database

CMap mirror database. Following the original methods described in CMap*,* the raw image of CEL files for the 6,097 instances from the CMap database were converted to average log-ratios and confidence calls using the algorithms MAS 5.0 (Affymetrix) and linear-fit-on-Pcall [11]. For each instance the log-ratios for the 22,283 HG-U133A probesets were ranked and the ranked data for all instances were saved in matrix form locally.

Local CMap program. The web version of CMap cannot be queried in batch mode. Furthermore, in each individual query the number of "tags", or the size of the gene-set, is limited to 1000. To overcome these limitations, we used C++ language to build a local program encoding the same algorithms and datasets used by CMap. This program allows CMap-type queries to be made locally in single or batch mode, and permits GSEA (Gene Set Enrichment Analysis [21]) parameters be varied. The program was tested for reliability and speed before applied to the current study (see Results).

The log-ratio matrix CMap and the enrichment-score matrix GSCMap and their sub-matrices

Cmap is a 22,283x6,097 probe-set versus instance matrix; elements of matrix are log-ratios of expression intensities. From this a number of extend maps/matrices were constructed:
Cmap1 – The 22,283x671 sub-matrix of CMap involving the 671 instances in CMap v1.0.
tCMap1 – A 300x671 sub-matrix of CMap1 involving the 300 highest variance probe-sets.
CMd – A 22,283x1,309 probe-set versus drug matrix reduced from CMap by averaging over same-drug instances.
IGCMd – A 4,884x1,309 sub-matrix of CMd involving the 4,884 highest variance probe-sets.
GSCMap – A 4,884x6,097 gene-set versus instance matrix; elements of the matrix are enrichment scores (ESs). For each of the 4,884 gene-sets (called tags) from MSigDB (collections C2-C5), we queried the 6,097 instances in CMap (version 2.0) to yield a 6,097-component vector (called Vd) of Kolmogorov-Smirnov statistic [11, 36, 37] based ESs, as defined in [21]. GSCMap is the set of 4,884 Vd's.
GSCMap1 – The 4,884x671 sub-matrix of GSCMap involving the 671 instances in CMap v1.0.

tGSCMap1 – A 300x671 sub-matrix of GSCMap1 involving the 300 largest ES variances gene-sets.
GSCMd – A 4,884x1,309 gene-set versus drug matrix. In CMap each drug were treated a variably multiple (averaging 6,097/1,309=4.66) times. For each gene-set and each drug the matrix element is the Kolmogorov-Smirnov statistic score (as in GSEA [21]) obtained by ranking the vector Vd corresponding to the gene-set and querying it using the multiple treatments for that drug.

Significance by permutation and normalized enrichment score (NES)

We tested the significance of the ES of a gene-set-drug pair by random permutation. Given a gene-set and a drug, and suppose the drug had t treatments in CMap and an (gene-set versus drug) enrichment score ES_0. We generated a distribution of randomized ESs by running r trials, in each trial recalculating the Kolmogorov-Smirnov ES by replacing the t treatments for the drug by t randomly selected treatments among the 6,097 treatments. A randomization (two-sided) p-value for the ES was computed from ES_0 and the distribution. The normalized enrichment score (NES) was taken to be ES_0 divided by the mean of the distribution [21]. In this work we set r = 10,000.

Gene-set based Local Hierarchical Clustering (GSLHC)

GSLHC is an application of GSCMap for discovering links among drugs through gene-sets strongly acted on by the drugs. Its implementation involves the steps: (i) Select a query drug set, which may be a single drug or a group of drugs with known shared property, or a drug of unknown property. (ii) For the query drug set, cull from GSCMap the functional profiles of drugs a subset gene-sets, each of which significantly enriched against every drug in the query drug set, where significant enrichment is determined by a threshold randomization p-value below an upper bound (we used $p < 0.005$). In the randomization test we generate a distribution of ESs by computing the ES for a gene-set-drug pair many times, each time replacing the genes in the gene-set by randomly selected genes from the entire gene pool [19]. (iii) Do a two-way hierarchical clustering of the culled gene-sets with the entire set of 1,309 drugs, and cut out from the resulting heatmap the clade of drugs that includes the query drug set with correlation above a threshold value (we used 0.9).

Cluster evaluation

We used the F-score, a harmonic mean of precision and recall [34], to evaluate a cluster as a classifier of a known classification. Let TP, FP, and FN be true positive, false positive, and false negative, respectively. The precision rate P and recall R rate of the cluster are respectively given by P = TP/(FP + TP) and R = TP/(TP+ FN). Suppose several nodes in a cluster are meant to represent a classification, then, for class i, the F-score F_i for that class is the maximum nodal value for 2PR/(P+R), and the F-score for the classification is the weighted average of F_i summed over the nodes. The higher the F-score, the better the classification by cluster. The F-score ranges from 0 to 1.

Ethic information None.

RESULTS

The local program reproduced results from CMap server with better efficiency

We used a gene-set called BRUINS_UVC_RESPONSE_LATE, which contains 1,137 genes differentially expressed only 12 h after UV-C irradiation of MEF cells, from MSigDB to compare the local program with the remote CMap server on the 100 drugs with the smallest *p*-value. The two programs yielded practically identical ESs (Figure S1A, dashed lines), and almost identical permutation *p*-values (Figure S1A, solid lines). Identical *p*-value were not expected; proportionally large differences in *p*-value occurred only when $p < 10^{-3}$. We used the 772 gene-sets in the C2 collection of MSigDB (number of genes in gene-sets ranged from 50 to 1000) to compare the speed of the local program and the CMap server and found that the computation times were comparable, but the local program was slower when the gene number in the gene-set exceeded 600 (Figure S1B). The slower speed of the local program was more than compensated by allowing querying in batch mode; if needed it could speeded up by modifying its sorting algorithm.

DEGs have low reproducibility in CMap genomic profiles

In CMap each of the 1,309 perturbagens has an average of 4.7 genomic profiles (from different treatments) resulting from the total of 6,097 treatments. We computed the fractional overlaps of top-1,000 DEGs between pairs of genomic profiles. The average reproducibility (common DEGs/1000) between different-perturbagen pairs has a sharp peak at 0.05, with few cases exceeding 0.1. That of the same-perturbagen pairs also peaks strongly at 0.06, but has a long weak tail (Figure S2), with 10,771 of cases having a reproducibility greater than 0.2.

The CMap and GSCMap matrices and their sub-matrices were constructed

Here, by CMap we mean the 22,283 (probes) x 6,097 (instances) matrix of log-ratios from CMap database. Using CMap we constructed the 4,884 (gene-sets) x 6,097 GSCMap matrix of ESs using the 4,884 gene-sets in MSigDB. Then we constructed sub-matrices of CMap, CMap1 (22,283x671), tCMap1 (300x671), CMd (22,283x1,309), and IGCMd (4,884x1,309), and sub-matrices for GSCMap, GSCMap1 (4,884x671), tSGCMap1 (300x671) and SGCMd (4,884x1,309), where 1,309 refers to the number of drugs/small molecules in CMap, 671 refers to the number of instances in CMap v1.0, 4,884 refers to the number of gene-sets in MSigDB or the 4,884 highest variance probe-sets (for IGCMd), and 300 refers to 300 highest variance probe-sets (CMap) or ESs (GSCMap) (detail in Methods).

Cell-type dependence of CMap data was strong in IGA but weak in GSA

As a first comparison between the IGA and GSA, we separately hierarchically clustered the two (300x671) matrices tCMap1 (representing IGA) and tGSCMap1 (representing GSA) using a Pearson distance metric and average-linkage and examined the properties of the two resulting 671-branch dendograms as cell-type classifiers. Under visual inspection the tCMap1 dendrogram was overwhelmingly dominated by cell type (Figure 2A) whereas the tGSCMap dendogram was not (Figure 2B). Quantitatively, F-scores (Materials and methods) for the tCMap1 dendogram indicated

that it provided a close to perfect classification for the four cell types (Table 1, permutation *p*-value < 0.01). In contrast, the tGSCMap dendogram was a poor (but fair for HL60) classifier for cell types. A similar result was found in a Principle Component Analysis on the full CMap dataset (Figure S3). These results implied GSA results had a significantly better chance than IGA of not being masked by cell-type dependence.

Testing drug responses in IGA and GSA

GSA had clearer and more varied drug response than IGA. We separately two-way hierarchically clustered the two (4,884x1,309) matrices IGCMd (for IGA) and GSCMd (for GSA) using Pearson distance metric and average-linkage (Figure 3). All computations were carried out over two days on a personal computer with an Intel(R) dual core Quad CPU, 2.40 GHz processor with a 8GB RAM. While the vast majority of gene-sets responded to the drugs as being either positively or negatively enriched (Figure 3A), the vast majority of high-variance genes were neither up-regulated nor down-regulated with respect to the drugs (Figure 3B).

GSA gave a better drug classifier than IGA. In CMap a drug typically is represented by several instances. For example, the pairs of drugs, trichostatin and LY-294002, respectively occur in 15 and 9 instances, each instance represented by a vector of 4,884 ESs (in GSCMap) or 22,283 intensity log-ratios (in CMap). We separately hierarchically clustered the two sets of combined 24 instances. Viewed as classifiers of the two drugs, the GSA cluster had a F-score of 0.98, and the IGA cluster, 0.72 (Figure 4a). The superiority of GSA over IGA in its ability to tell one drug from another happened to be a general feature. We repeated the above comparison for all the 20,736 drug-pairs with multiple instances in CMap1 and in GSCMap1 and found that the (drug classification) F-score for GSA was about 0.036 higher then IGA over an average of 0.75 (Figure 4b, two-sample Kolmogorov-Smirnov test: *p*-value < 2.2e-16).

GSA and IGA responded similarly to chemical properties of CMap drugs. The F-scores of clusters, constructed through GSA (using ESs from GSCMd) and IGA (using gene expression log-ratios from IGCMd), of drugs classified according to their anatomical, chemical, therapeutic, pharmacological (Anatomical Therapeutic Chemical (ATC) classification system, World Health Organization, http://www.whocc.no/) and structural (PubChem Structure Database [37]) properties (Material and Methods) were indistinguishable (Figure S4).

Genomic signatures of same-target drug pairs had higher correlation in GSA than in IGA. We expect the genomic signatures of drugs sharing a target to be more similar than drugs that do not. Information on drug targets were obtained from the Therapeutic Target Database (TTD) [35] (Material and Methods). The same-target drug-pairs correlated much better under GSA (ESs from GSMCd) than IGA (gene expression log-ratios from IGMCd) (Figure 5). An outstanding case was the triplet vorinostat, valproic acid, and trichostatin A that targets the histone deacetylase (HDAC) protein. The three pair-wise correlations for the triplet ranged from 0.8 to 1.0 in GSA and from 0.05 to 0.15 in IGA. Averaged over all 5,034 pairs involving 639 drugs, the mean of GSA correlation was 0.35 (S.D. =

0.27) and the mean of IGA correlation was 0.18 (S.D. = 0.15) (two-sample *t*-test, *p*-value < 2.2e-16).

Validation of GSLHC and novel HDAC inhibitors

There are 106 active compounds in the CMap database that are poorly studied, and GSLHC was developed as an application on GSCMap to discover drug partners of known therapeutic properties for the compounds. We tested the GSLHC by giving it a set of gene-sets common to and significantly enriched in the functional profiles of three histone deacetylase (HDAC) inhibitors – vorinostat (also known as suberoylanilide hydroxamic acid or SAHA), valproic acid, and trichostatin A – and see if it can recover them from GSCMap. The three HDAC inhibitors were chosen because they have been fully studied [38-40]. A set of 597 gene-sets significantly enriched with permutation $p< 0.005$ were selected for the test (Material and Methods). The selected gene-sets had functions related to HDAC inhibitor activities. For example, among the down-regulated functions were histone acetylating, histone and chromatin modification, and maintenance of chromatin structures (Figure 6c). The test was successful; the triplet was among the six recovered drugs (Figures 6a and 6b). The three extras are not known as HDAC inhibitors but two of the three, scriptaid and HC toxin, have been reported to have HDAC inhibition activities [41, 42].

Application of GSLHC to characterize active compounds of unknown therapeutic properties

A novel cyclin-dependent kinase inhibitor (CDKi) The compound 0175029-0000 is among molecules in CMap known to be active in certain biological roles [11] but poorly studied in literature. Its ES profile had 1,080 significantly enriched gene-sets with permutation $p< 0.005$. Our GSLHC search showed it to be closely associated with three CDKi's with correlation coefficient (CE) > 0.97 and five DNA topoisomerases with CE > 0.92 (Figure 7a and 7b). Biological functions negatively regulated by these drugs included those related to cell cycle and checkpoint on cell cycle (Figure 7c).

A novel antibiotic, anesthetic, and anti-inflammatory agent The ES profile of compound CP-863187 had 36 significantly enriched gene-sets with permutation $p< 0.005$. Our GSLHC search showed it to be closely associated with an antibiotic (piperacilin; CE > 0.98), an anesthetic (benzocaine), an anti-inflammatory agents (betunlinic acid; CE > 0.97), as well as with another anti-inflammatory agent (CE > 0.96) and five other antibiotics (CE > 0.90) (Figure 8a and 8b). Biological functions affected by drugs associated with CP-863187 included negative regulation of integrin signalling pathway and hydrolases (Figure 8c).

Summary of drug discovery by GSLHC (Table 2)

Eighteen previously uncharacterized compounds in CMap, including 0175029-0000 and CP-863187, were discovered by GSLHC to have closely associated drug partners (in CMap), putative targets, and therapeutic indications (Table 2; detail in Figures S5 - S20). Among the discoveries, eight compounds: tyrphostin AG-825, 5248896, 0175029-0000, H-7, U0125, STOCK1N-35215, 0297417-0002B, and F0447-0125, were identified as having potential anti-tumor activities. Depending on their closest putative drug partners, their molecular mechanisms differ. Camptothecin, irinotecan, and

betulinic acid, with closest partners tyrphostin AG-825, U0125, and CP-944629, respectively, were predicted to block DNA transcription by inhibiting DNA topoisomerase activities. The compounds 0175029-0000 and H-7, with closest partner GW-8510, were predicted to be cyclin-dependent kinase inhibitors. Compounds predicted to have therapeutic activities on non-cancer diseases include 5186324 (closest partner neostigmine bromide and therapeutic activity on myasthenia gravis) and Prestwick-692 (closest partner isoflupredone and therapeutic activity on rheumatoid arthritis).

DISCUSSION and SUMMARY

We used CMap as a vehicle for the demonstration that GSA is a better way than IGA in utilizing genome-wide gene expression. Because this would involve repeated and massive application of CMap, we constructed a local extended version of CMap. The local CMap was stored and computation using it were conducted on a personal computer equipped with Intel(R) dual core Quad CPU, 2.40 GHz processor with a 8GB RAM. Advantages of the local program over the remote CMap include: (i) No reliance on the Internet and the ensuing network connection time saved; (ii) Length of the list of querying gene not limited to 1000; (iii) Capability for batch mode operation. Extensive tests conducted on the local version confirmed its accuracy, and verified that in single mode its running speed is comparable to the remote CMap (Figure S1).

We implemented a GSA-based application of CMap by constructing GSCMap, an analog of CMap where gene-based genomic profiles of instances in CMap are replaced by gene-set-based functional profiles.

Hierarchical clustering based on gene expression has been an important tool in genomic technology. We showed that IGA-based hierarchical clustering of the CMap (the matrix) was dominated by cell-types, a dominance absent in the GSA-based GSCMap (Figure 2). This notion was strengthened by our quantitative measure, using F-scores, of the clusters as classifies of cell types. We confirmed a previous report that CMap was an excellent classifier of cell types, a result that imposes strong constraints of it being a good classifier of drug effects. In contrast, our F-score analysis showed GSCMap to be a poor classifier of cell types (Table 2).

Having demonstrated that GSCMap has far weaker cell-type dependence than CMap, we conducted three tests to show the former had more discriminating responses to drug properties than the latter. The first test (using the 4,884x1,309 matrices GSCMd and IGCMd) showed gene-set response to drugs in GSCMap exhibited a much wider range then gene expression response to drugs in CMap (Figure 3). A second test showed that GSCMap clustered same-drug instances consistently better than CMap (Figure 4). A third test showed that the genomic profiles of a pair of drugs having the same target had higher correlation in GSCMap than in CMap (Figure 5). In contrast, the same analysis applied to drug-pairs having structural similarities at the chemical level or therapeutic indications at the clinical level did not exhibit any different between GSCMap and CMap (Figure S4).

GSLHC was designed to discover, through GSCMap, functional links among drugs in CMap. The principle of the method, local hierarchical clustering, is generally applicable to any large list that may or may not represent drug effects. We validated GSLHC by using three known HDAC inhibitors as bait and saw that they were recovered as part of a tight cluster (correlation > 0.92) returned by GSLHC. The cluster also included three drugs, scriptaid, HC toxin, and rufabutin, not previously known as HDAC inhibitors. GSLHC showed all three as having significant correlation with biological functions relating to switching histone modification and destroying chromatin maintenance (Figure 6);

scriptaid and HC toxin have been reported to inhibit HDAC proteins [41, 42], and rifabutin is primarily used in the treatment of tuberculosis. We regard all three as potential novel HDAC inhibitors.

Of the 106 uncharacterized compounds in the CMap dataset, GSLHC found drug partners of known indications for 18 (Table 2), 8 of which, tyrphostin AG-825, 0175029-0000, H-7, U0125, STOCK1N-35215, 0297417-0002B, F0447-0125, and CP-944629 were inferred to have anti-tumor activities. In each case we found significantly correlations between the compound with newly inferred indication and biological functions related to that indication (Figures 7-8, and S5-20).

The compound 0175029-0000 was shown to be closely associated with three CDKi's – GW-8510 [46-53], alsterpaullone [46-53], H-7 [46-53] – and five DNA topoisomerases – doxorubicin [46-53], camptothecin [46-53], azacitidine [46-53], mitoxantrone [46-53], and ellipticine [46-53] (Figure 7), and was inferred as a putative CDKi/DNA topoisomerases, all of which have been reported to have anti-tumour activities [46-53] and significantly expressed biological functions that negatively regulate cell cycle and checkpoint on cell cycle (Figure 7c).

The compound CP-863187 was shown to be closely associated with an antibiotic (piperacilin), an anesthetic (benzocaine), and an anti-inflammatory agent (betunlinic) (Figure 8), and to significantly express negative regulation of integrin signaling and hydrolases (Figure 8c). There are studies suggesting that antibiotics may have inflammatory and anesthetic properties [54, 55]. The source of the shared properties may be that as a signal transductors, integrins are involved in activities on cell membranes and cell-cell interactions. Hydrolases are ubiquitous and play important roles among bacteria including digesting the murein of bacteria [43], acting as a pacemaker for cell wall growth [44], and splitting the septum during cell division [45].

Despite its apparent success, the GSLHC approach has its own limitations. Statistical concerns regarding the neutrality of GSEA has been raised [56] (and replied [57]). There is not a perfect method for extracting hypothesis-free information from something as rich as a modern set of genome-wide gene expression data. The several tests shown in this work does show that for practical purposes, GSA, including GSEA and two algorithms derived from it, PAGE and GAGE, is superior to IGA. Of the 106 unknown compounds in CMap (version 1.0), we only found drug partners for 18. That we failed to do the same for the other 88 compounds have many possible reasons: a weakness of GSLHC; the gene-sets in MSigDB is not sufficiently comprehensive; the sets of compounds presently included in MCap is too restrictive. Improvements on all three fronts are possible, even expected. Already in its current form, we expect the GSLHC approach to be more widely applicable to many areas other than what was demonstrated here. To name a few: repurposed drug discovery based on functional-profile characterization of phenotypes, function-based diagnosis and classification of complex diseases, and prognosis on advance-stage patients after chemotherapy treatment.

REFERENCES

1. Perez-Diez, A., A. Morgun, and N. Shulzhenko, *Microarrays for cancer diagnosis and classification.* Adv Exp Med Biol, 2007. **593**: p. 74-85.
2. Miller, J.A., M.C. Oldham, and D.H. Geschwind, *A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging.* J Neurosci, 2008. **28**(6): p. 1410-20.
3. Cui, X. and G.A. Churchill, *Statistical tests for differential expression in cDNA microarray experiments.* Genome Biol, 2003. **4**(4): p. 210.
4. Zaravinos, A., et al., *Identification of common differentially expressed genes in urinary bladder cancer.* PLoS One, 2011. **6**(4): p. e18135.
5. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
6. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
7. Pavlidis, P., *Using ANOVA for gene selection from microarray studies of the nervous system.* Methods, 2003. **31**(4): p. 282-9.
8. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
9. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
10. Hosack, D.A., et al., *Identifying biological themes within lists of genes with EASE.* Genome Biol, 2003. **4**(10): p. R70.
11. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.
12. Aramadhaka, L.R., et al., *Connectivity maps for biosimilar drug discovery in venoms: the case of Gila monster venom and the anti-diabetes drug Byetta(R).* Toxicon, 2013. **69**: p. 160-7.
13. Meng, F., et al., *Constructing and characterizing a bioactive small molecule and microRNA association network for Alzheimer's disease.* J R Soc Interface, 2014. **11**(92): p. 20131057.
14. Chen, F., et al., *Gene expression profile and functional analysis of Alzheimer's disease.* Am J Alzheimers Dis Other Demen, 2013. **28**(7): p. 693-701.
15. Garman, K.S., et al., *A genomic approach to colon cancer risk stratification yields biologic insights into therapeutic opportunities.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19432-7.
16. Huang, L., et al., *An integrated bioinformatics approach identifies elevated cyclin E2 expression and E2F activity as distinct features of tamoxifen resistant breast tumors.* PLoS One, 2011. **6**(7): p. e22274.
17. Wang, G., et al., *Expression-based in silico screening of candidate therapeutic compounds for lung adenocarcinoma.* PLoS One, 2011. **6**(1): p. e14573.
18. Qu, X.A. and Rajpal, D.K., *Application of connectivity map in drug discovery and development.* Drug Discov Today, 2012. **17**(23-23): p. 1289-98.
19. Nam, D. and S.Y. Kim, *Gene-set approach for expression pattern analysis.* Brief Bioinform, 2008. **9**(3): p. 189-97.
20. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nat Genet, 2003. **34**(3): p. 267-73.
21. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
22. Efron, B. and R. Tibshirani, *On Testing the Significance of Sets of Genes.* Annals of Applied Statistics, 2007. **1**(1): p. 107-129.

23. Barry, W.T., A.B. Nobel, and F.A. Wright, *Significance analysis of functional categories in gene expression studies: a structured permutation approach.* Bioinformatics, 2005. **21**(9): p. 2043-9.
24. Breslin, T., P. Eden, and M. Krogh, *Comparing functional annotation analyses with Catmap.* BMC Bioinformatics, 2004. **5**: p. 193.
25. Lee, H.K., et al., *ErmineJ: tool for functional analysis of gene expression data sets.* BMC Bioinformatics, 2005. **6**: p. 279.
26. Dinu, I., et al., *Improving gene set analysis of microarray data by SAM-GS.* BMC Bioinformatics, 2007. **8**: p. 242.
26. Prifti, E., et al., *FunNet: an integrative tool for exploring transcriptional interactions.* Bioinformatics, 2008. **24**(22): p. 2636-8.
28. Vaske, C.J., et al., *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.* Bioinformatics, 2010. **26**(12): p. i237-45.
29. Sun, C.H., et al., *COFECO: composite function annotation enriched by protein complex data.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W350-5.
30. Wong, D.J., et al., *Revealing targeted therapy for human cancer by gene module maps.* Cancer Res, 2008. **68**(2): p. 369-78.
31. Ben-Dor, A., et al., *Discovering local structure in gene expression data: the order-preserving submatrix problem.* J Comput Biol, 2003. **10**(3-4): p. 373-84.
32. Tanay, A., R. Sharan, and R. Shamir, *Discovering statistically significant biclusters in gene expression data.* Bioinformatics, 2002. **18 Suppl 1**: p. S136-44.
33. Li, Q.L., et al., *PubChem as a public resource for drug discovery.* Drug Discovery Today, 2010. **15**(23-24): p. 1052-1057.
34. Manning, C.D., P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press.
35. Zhu, F., et al., *Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1128-36.
36. Hollander, M. and D. Wolfe, *Nonparametric Statistical Methods* 2ed. 1999, New York: Wiley.
37. Lamb, J., et al., *A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer.* Cell, 2003. **114**(3): p. 323-34.
38. Richon, V.M., *Cancer biology: mechanism of antitumour action of vorinostat (suberoylanilide hydroxamic acid), a novel histone deacetylase inhibitor.* Br J Cancer, 2006. **95**(S1): p. S2-S6.
39. Gottlicher, M., et al., *Valproic acid defines a novel class of HDAC inhibitors inducing differentiation of transformed cells.* EMBO J, 2001. **20**(24): p. 6969-78.
40. Kemp, M.G., et al., *The histone deacetylase inhibitor trichostatin A alters the pattern of DNA replication origin activity in human cells.* Nucleic Acids Res, 2005. **33**(1): p. 325-37.
41. Keen, J.C., et al., *A novel histone deacetylase inhibitor, scriptaid, enhances expression of functional estrogen receptor alpha (ER) in ER negative human breast cancer cells in combination with 5-aza 2'-deoxycytidine.* Breast Cancer Res Treat, 2003. **81**(3): p. 177-86.
42. Balakin, K.V., et al., *Histone deacetylase inhibitors in cancer therapy: latest developments, trends and medicinal chemistry perspective.* Anticancer Agents Med Chem, 2007. **7**(5): p. 576-92.
43. Smith, T.J., S.A. Blackman, and S.J. Foster, *Autolysins of Bacillus subtilis: multiple enzymes with multiple functions.* Microbiology, 2000. **146 (Pt 2)**: p. 249-62.
44. Holtje, J.V., *From growth to autolysis: the murein hydrolases in Escherichia coli.* Arch Microbiol, 1995. **164**(4): p. 243-54.
45. Garcia, P., et al., *LytB, a novel pneumococcal murein hydrolase essential for cell separation.* Mol Microbiol, 1999. **31**(4): p. 1275-81.

46. Wood, E.R., et al., *Discovery and in vitro evaluation of potent TrkA kinase inhibitors: oxindole and aza-oxindoles.* Bioorg Med Chem Lett, 2004. **14**(4): p. 953-7.
47. Lahusen, T., et al., *Alsterpaullone, a novel cyclin-dependent kinase inhibitor, induces apoptosis by activation of caspase-9 due to perturbation in mitochondrial membrane potential.* Mol Carcinog, 2003. **36**(4): p. 183-94.
48. Keller, H.U., A. Zimmermann, and V. Niggli, *Diacylglycerols and the protein kinase inhibitor H-7 suppress cell polarity and locomotion of Walker 256 carcinosarcoma cells.* Int J Cancer, 1989. **44**(5): p. 934-9.
49. Tan, C., et al., *Daunomycin, an antitumor antibiotic, in the treatment of neoplastic disease. Clinical evaluation with special reference to childhood leukemia.* Cancer, 1967. **20**(3): p. 333-53.
50. Rose, M.G., *Hematology: Azacitidine improves survival in myelodysplastic syndromes.* Nat Rev Clin Oncol, 2009. **6**(9): p. 502-3.
51. Ko, M.W., et al., *Acute promyelocytic leukemic involvement of the optic nerves following mitoxantrone treatment for multiple sclerosis.* J Neurol Sci, 2008. **273**(1-2): p. 144-7.
52. Kim, J.Y., et al., *Ellipticine induces apoptosis in human endometrial cancer cells: the potential involvement of reactive oxygen species and mitogen-activated protein kinases.* Toxicology, 2011. **289**(2-3): p. 91-102.
53. Ulukan, H. and P.W. Swaan, *Camptothecins: a review of their chemotherapeutic potential.* Drugs, 2002. **62**(14): p. 2039-57.
54. Rubin, B.K. and J. Tamaoki, *Antibiotics as anti-inflammatory and immunomodulatory agents*. Pir. 2005, Basel ; Boston: Birkhäuser. xiii, 273 p.
55. Sanders, W.E., Jr., *Antibiotics during anesthesia and surgery.* Int Anesthesiol Clin, 1968. **6**(1): p. 211-8.
56. Damian, D. and M. Gorfine, *Statistical concerns about the GSEA procedure.* Nature Genetics, 2004. **36**(7): p. 663-663.
57. Mootha, V.K., et al., *Statistical concerns about the GSEA procedure - Reply.* Nature Genetics, 2004. **36**(7): p. 663-663.

FIGURE LEGENDS

Figure 1. The GSLHC workflow

Figure 2. Hierarchical clustering of CMap instances less dominated by cell-type when clustering is based on multiple gene-set enrichment scores

Dendrograms are hierarchical clustering of CMap instances based on gene expression (A) and gene-set enrichment score (B). Colors in color bar below dendogram respectively represent the cell lines SKMEL5 (red), PC3 (green), MCF7 (blue), and HL60 (purple). For each instance top-300 genes or gene-sets with the top-300 expression log-ratios or ES scores were selected for clustering based on Pearson distance metric and average linkage.

Figure 3. Two-way hierarchical clustering heatmep of 1309 CMap perturbagens shows higher contrast when clustering is based on multiple gene-set enrichment scores

Two-way hierarchical clustering heatmaps were generated based on Pearson distance metric and average linkage using, for each CMap perturbagen: (A) normalized enrichment scores (NESs) of 4,884 gene-sets from MSigDB, and (B) log-ratios for expression levels of the top-4884 high-variance genes. Color code: on red, positive NES or log-ratio; green, black, NES or log ratio ~0; green, negative NES or log-ratio.

Figure 4. Separation of two drugs among all instances involving the drug pair is done better using multiple multiple gene-set enrichment scoress

Quality of separation is determined by F-scores for hierarchical clusters, constructed using gene-set enrichment scores and log-ratios for gene expressions, respectively, of all instances involving the drug pair. (A) Two clusters for the drug-pair valproic acid and trichostatin A; cluster based on gene expression, cluster on left, and on gene-set, cluster on right. Two-color bar indicates drug classification. (B) Ranking by F-scores of ~20,000 drug-pairs from the CMap development batch on HG-U133A platform involving 407 drugs and 674 chips; black, gene expression and red, gene-set.

Figure 5. Same-target drug-pairs correlate better when evaluated by multiple gene-set enrichment scores

Figure plots correlation of same-target drug pair evaluated by gene-set enrichment score (ES) versus that evaluated by gene expression. Drug targets were those given by TTD database. In the gene-set approach, each drug, or CMap perturbagen, was represented by the ESs of 4884 MSigDB gene-sets. In the gene expression case, each drug was represented by the set of top-4884 high variance genes. The three red dots are from the three pairs formed by the three drugs, vorinostat, valproic acid, and trichostatin A, all targeting the histone deacetylase (HDAC) protein.

Figure 6. GSLHC finds novel HDAC inhibitors

The three know HDAC inhibitors valproic acid, trichostatin A, and vorinostat, are all significantly enriched by 597 gene-sets with permutation $p < 0.005$; these 597 gene-sets were used in a new

heatmap in the GSLHC protocol. (A) A sub-heatmap including the three HDAC inhibitors and all neighbors with correlation > 0.9. (B) Detail of the drug cluster associated with the sub-heatmap. The two drugs rifabutin and scriptaid in the cluster, not previously known as HDAC inhibitors, has literature support as having inhibition functions on HDAC proteins. (C) Detail of the gene-set cluster with the sub-heatmap shows several functions known to be related to HDAC inhibitor activities.

Figure 7. GSLHC identifies 0175029-0000 as a novel cyclin-dependent kinase inhibitor (CDKi)
(A) A correlation > 0.9 sub-heatmap including the compound 0175029-0000 of unknown function from a GSLHC-generated heatmap based on the ES of 1080 gene-sets significantly enriched in 0175029-0000 with permutation p< 0.005. (B) Detail of the drug cluster associated with the sub-heatmap. According to the TTD database, GW-8510, alsterpaullone, and H-7 (red asterisk) CDK inhibitors, and doxorubicin, camptothecin, azacitidine, mitoxantrone, and ellipticine (blue asterisk) are DNA topoisomerase inhibitors. All have anti-tumor activities. (C) Detail of the gene-set cluster with the sub-heatmap shows functions known to be related to the inhibition activities of cell cycle.

Figure 8. GSLHC identifies CP-863187 as a potential antibiotic
(A) A correlation > 0.9 sub-heatmap including the compound CP-863187 of unknown function from a GSLHC-generated heatmap based on the ES of 36 gene-sets significantly enriched in CP-863187 with permutation p< 0.005. (B) Detail of the drug cluster associated with the sub-heatmap. According to TTD database, piperacillin, dapsone, tocainide, ampicillin, sulfadimethoxine, metronidazole (red asterisk) are antibiotics, betulinic acid and isoflupredone (blue asterisk) are anti-inflammatory agents, and benzocaine (green asterisk) is an anesthetic. (C) Detail of the gene-set cluster with the sub-heatmap shows functions known to block the formation of bacteria cell wall by inhibition of integrin signaling pathway.

SUPPORTING INFORMATION LEGENDS

Figure S1. The local program reproduces results of CMap server
(A) The local program (blue) tracks results given by CMap for permutation p-value (solid lines), with small deviations when drug list is less than 30, and enrichment score (dash lines). (B) Run times for the local program and CMap are comparable, with the former slightly faster when size of probe set is less than 700, and slight slower otherwise.

Figure S2. Most of replicates treating with the same perturbagen show low reproducibility on the top-1000 differentially expressed genes (DEGs) across all CMAP datasets
The reproducibility between two treatments (blue: the same perturbagen; red: two different perturbagens) is defined by the frequency of number of the overlapping genes verse the number of 1000 DEGs.

Figure S3. Principle component analysis for full C-MAP dataset
The first two components, together accounting for 21.7% of the total weight, show a clear separation of data from the HC60 (black circle) and PC3 (green cross) cell lines.

Figure S4. Performance test (F-score) showed that no difference between gene and gene-set clusters by Anatomical Therapeutic Chemical (ATC) classification system and PubChem structure database
(A) In PubChem database, we use chemical structure clustering tool to cluster compounds based on the structure (fingerprint) similarity using the Single Linkage algorithm; number of cluster decreases with cluster size. Both results indicated that F-score increases with decreasing class size. (B) In ATC system, drugs are classified into groups at 4 different levels – from general anatomical groups to detail chemical/therapeutic/pharmacological subgroups.

Figure S5. GSLHC identified the compound 5186324 as a novel acetylcholinesterase inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound 5186324 of unknown function from a GSLHC-generated heatmap based on gene-sets in 5186324 significantly enriched with permutation $p< 0.005$. (B) Detail of the dendrogram showing 5186324 (marked by black asterisk) with its partner drugs.

Figure S6. GSLHC identified the compound DL-PPMP as a novel cyclooxygenase-1 inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound DL-PPMP of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in DL-PPMP with permutation $p< 0.005$. (B) Detail of the dendrogram showing DL-PPMP (marked by black asterisk) with its partner drugs.

Figure S7. GSLHC identified the compound Prestwick-692 as a novel glucocorticoid receptor agonist
(A) A correlation > 0.9 sub-heatmap including the compound Prestwick-692 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in Prestwick-692 with permutation $p< 0.005$. (B) Detail of the dendrogram showing Prestwick-692 (marked by black asterisk) with its partner drugs.

Figure S8. GSLHC identified the compound tyrphostin AG-825 as a novel DNA topoisomerase I inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound tyrphostin AG-825 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in tyrphostin AG-825 with permutation $p< 0.005$. (B) Detail of the dendrogram showing tyrphostin AG-825 (marked by black asterisk) with its partner drugs.

Figure S9. GSLHC identified the compound 5248896 as a novel human epidermal growth factor receptor (HER)-2/neu inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound 5248896 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in 5248896 with permutation p< 0.005. (B) Detail of the dendrogram showing 5248896 (marked by black asterisk) with its partner drugs.

Figure S10. GSLHC identified the compound H-7 as a novel Cyclin-dependent kinase 2 Inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound H-7 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in H-7 with permutation p< 0.005. (B) Detail of the dendrogram showing H-7 (marked by black asterisk) with its partner drugs.

Figure S11. GSLHC identified the compound Prestwick-1103 as a novel Tumor necrosis factor antibody
(A) A correlation > 0.9 sub-heatmap including the compound Prestwick-1103 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in Prestwick-1103 with permutation p< 0.005. (B) Detail of the dendrogram showing Prestwick-1103 (marked by black asterisk) with its partner drugs.

Figure S12. GSLHC identified the compound U0125 as a novel DNA topoisomerase I inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound U0125 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in U0125 with permutation p< 0.005. (B) Detail of the dendrogram showing U0125 (marked by black asterisk) with its partner drugs.

Figure S13. GSLHC identified the compound 5109870 as a novel Alpha adrenergic receptor antagonist
(A) A correlation > 0.9 sub-heatmap including the compound 5109870 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in 5109870 with permutation p< 0.005. (B) Detail of the dendrogram showing 5109870 (marked by black asterisk) with its partner drugs.

Figure S14. GSLHC identified the compound MG-132 as a novel Proteasome Inhibitor (A) A correlation > 0.9 sub-heatmap including the compound MG-132 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in MG-132 with permutation p< 0.005. (B) Detail of the dendrogram showing MG-132 (marked by black asterisk) with its partner drugs.

Figure S15. GSLHC identified the compound PHA-00851261E as a novel CGMP-inhibited 3',5'-cyclic phosphodiesterase
(A) A correlation > 0.9 sub-heatmap including the compound PHA-00851261E of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in PHA-00851261E with permutation p< 0.005. (B) Detail of the dendrogram showing PHA-00851261E (marked by black asterisk) with its partner drugs.

Figure S16. GSLHC identified the compound STOCK1N-35215 as a novel Histone deacetylase inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound STOCK1N-35215 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in STOCK1N-35215 with permutation p< 0.005. (B) Detail of the dendrogram showing STOCK1N-35215 (marked by black asterisk) with its partner drugs.

Figure S17. GSLHC identified the compound 0297417-0002B as a novel Purine nucleoside phosphorylase Inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound 0297417-0002B of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in 0297417-0002B with permutation p< 0.005. (B) Detail of the dendrogram showing 0297417-0002B (marked by black asterisk) with its partner drugs.

Figure S18. GSLHC identified the compound F0447-0125 as a novel DNA Inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound F0447-0125 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in F0447-0125 with permutation p< 0.005. (B) Detail of the dendrogram showing F0447-0125 (marked by black asterisk) with its partner drugs.

Figure S19. GSLHC identified the compound W-13 as a novel Mineralocorticoid receptor agonist
(A) A correlation > 0.9 sub-heatmap including the compound W-13 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in W-13 with permutation p< 0.005. (B) Detail of the dendrogram showing W-13 (marked by black asterisk) with its partner drugs.

Figure S20. GSLHC identified the compound CP-944629 as a novel DNA polymerase beta inhibitor
(A) A correlation > 0.9 sub-heatmap including the compound CP-944629 of unknown function from a GSLHC-generated heatmap based on gene-sets significantly enriched in CP-944629 with permutation p< 0.005. (B) Detail of the dendrogram showing CP-944629 (marked by black asterisk) with its partner drugs.

TABLES

Table 1. Cell-type effects are eliminated in hierarchical clustering based gene-set enrichment
Cluster evaluation by F score was computed for cell-type classification of hierarchical clustering based on individual genes and on gene-based enrichment (see materials and methods). Permutation *p*-value was calculated for the F scores by 100 random permutations of cell-type labels.

Cell type	F		Permutation *p*-value	
	Gene	Gene-set	Gene	Gene-set
MCF7	0.92	0.33	< 0.01	0.83
HL60	0.99	0.59	< 0.01	0.01
PC3	0.97	0.31	< 0.01	0.48
SKMEL5	1.00	0.30	< 0.01	0.09

Table 2. Putative molecular target and pharmacology application deduced from GSLHC of CMap perturbagens without known indication

Partner drug of a perturbagen (test drug in first column) was given by GSLHC. Putative target and indication were those associated with partner drug as given by the TTD database. Perturbagens found to be anti-tumouric are marked by the $^{+}$ symbol.

Test drug (+ anti-tumor)	Cor.	Partner drug	Putative targets	Indications
5186324	0.99	Neostigmine bromide	Acetylcholinesterase inhibitor	Myasthenia gravis
DL-PPMP	0.99	Indoprofen	Cyclooxygenase-1 inhibitor	Non-steroidal anti-inflammatory drug
Prestwick-692	0.99	Isoflupredone	Glucocorticoid receptor agonist	Rheumatoid arthritis
tyrphostin AG-825^{+}	0.995	Camptothecin	DNA topoisomerase I inhibitor	Cancer
	0.990	GW-8510	Cyclin-dependent kinase 2 inhibitor	Cancer
	0.990	Doxorubicin	DNA topoisomerase II inhibitor	Cancer
	0.975	duanorubicin	DNA topoisomerase II inhibitor	Leukemia, cancer
	0.970	Irinotecan	DNA topoisomerase I inhibitor	Colorectal Cancer
	0.96	Mitoxantrone	Human epidermal growth factor receptor-2/neu inhibitor	Acute myeloid leukemia, metastatic breast cancer
	0.96	alsterpaullone	Glycogen synthase kinase 3 inibitor	Cancer, type II diabetes
5248896^{+}	0.98	tyrphostin AG-825	Human epidermal growth factor receptor-2/neu inhibitor	Myeloid leukemia
0175029-0000^{+}	0.98	GW-8510	Cyclin-dependent kinase 2 inhibitor	Cancer
CP-863187	0.98	Piperacillin	Sodium channel blocker	Anesthetic
H-7^{+}	0.98	GW-8510	Cyclin-dependent kinase 2 inhibitor	Cancer
	0.98	Doxorubicin	anthracycline antibiotic	Cancer
Prestwick-1103	0.98	Pentoxifylline	Tumor necrosis factor antibody	Intermittent claudication, vascular dementia
U0125$^{'}$	0.98	Irinotecan	DNA topoisomerase I inhibitor	Colorectal Cancer
5109870	0.97	Phenoxybenzamine	Alpha adrenergic receptor antagonist	Hypertension, hypoplastic left heart syndrome
MG-132	0.97	MG-262	Proteasome Inhibitor	---
PHA-00851261E	0.97	Amrinone (inamrinone)	CGMP-inhibited 3',5'-cyclic phosphodiesterase	Congestive heart failure
STOCK1N-35215^{+}	0.97	MS-275	Histone deacetylase inhibitor	Hodgkin's lymphoma (phase II trial)
0297417-0002B^{+}	0.95	8-azaguanine	Purine nucleoside phosphorylase Inhibitor	Acute leukemia
F0447-0125^{+}	0.95	Lomustine	DNA Inhibitor	Brain tumours, Hodgkin's lymphoma
W-13	0.95	Fludrocortisones	Mineralocorticoid receptor agonist	Addison's disease, cerebral saltwasting syndrome
CP-944629	0.92	Betulinic acid	DNA polymerase beta inhibitor	Melanoma (in development)

Figure 1
Click here to download high resolution image

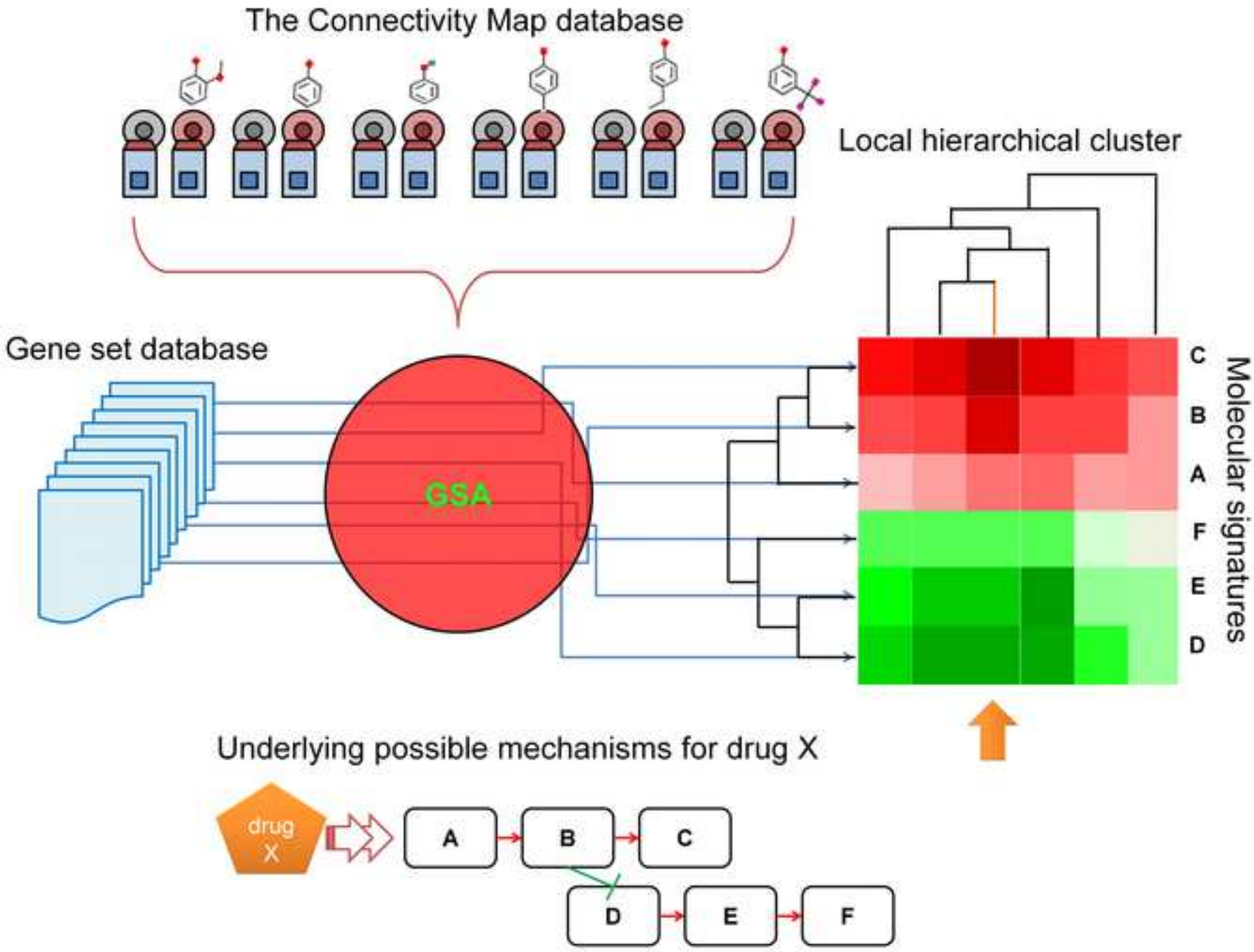


Figure 2
Click here to download high resolution image

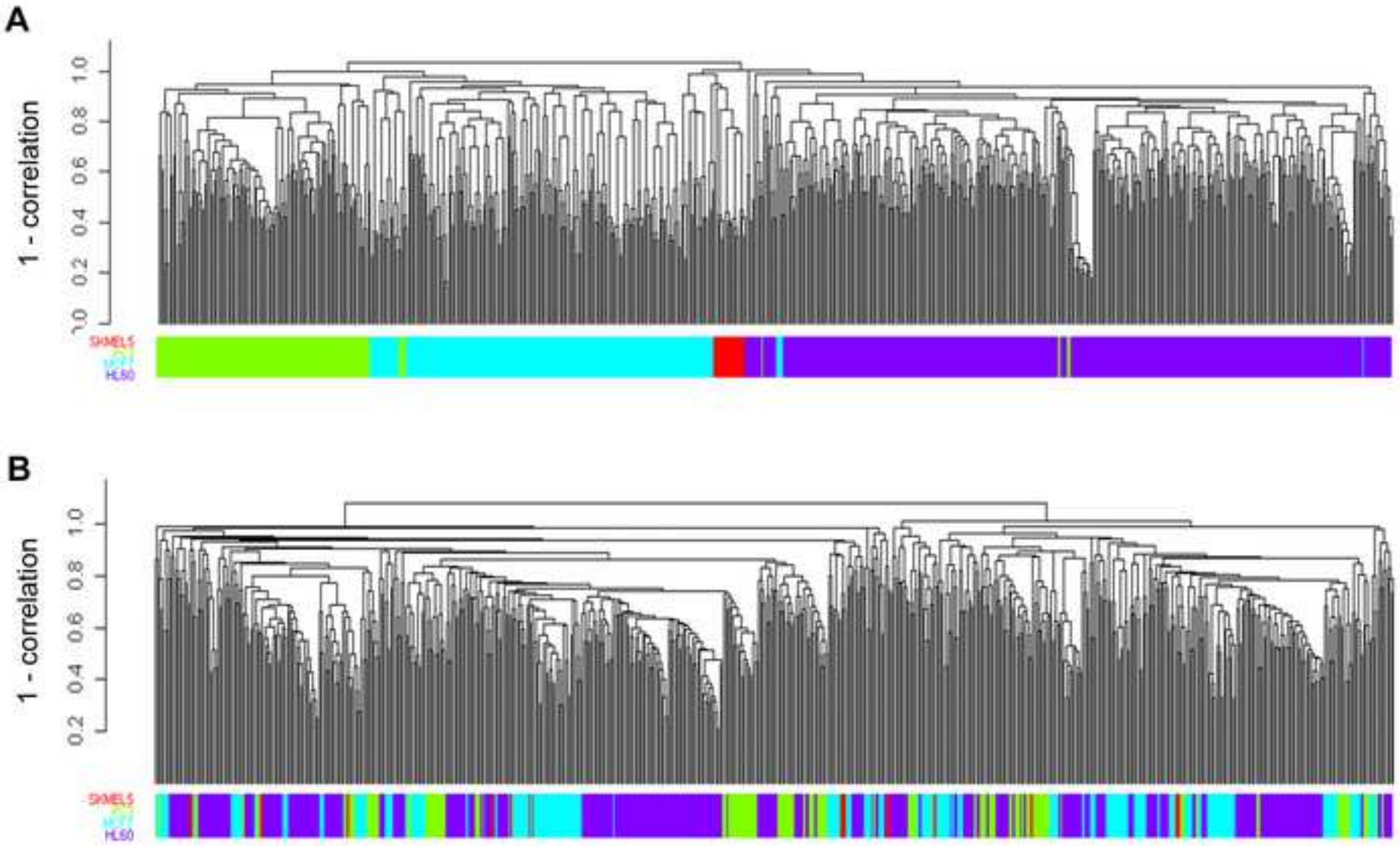


Figure 3
Click here to download high resolution image

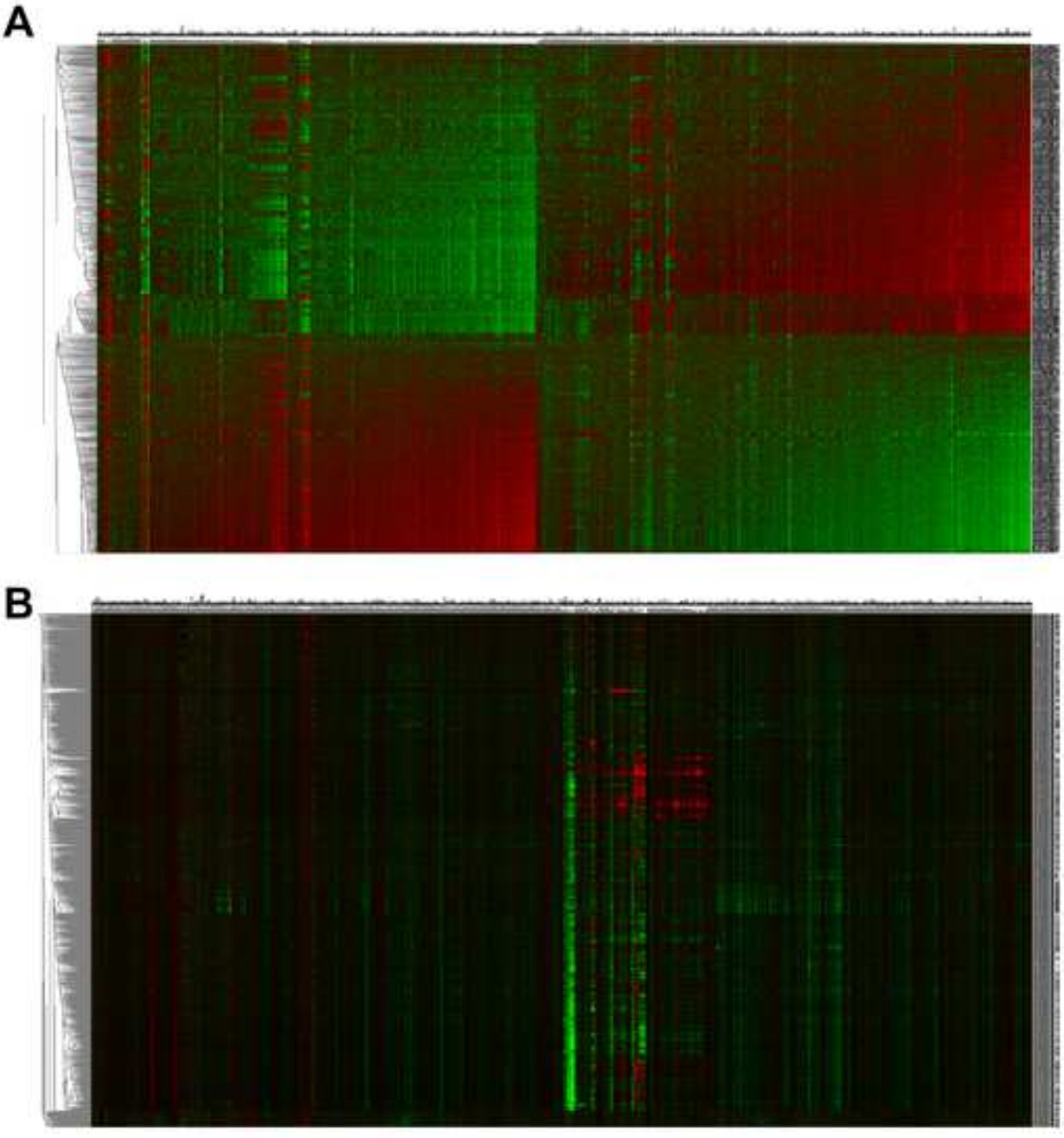


Figure 4
Click here to download high resolution image

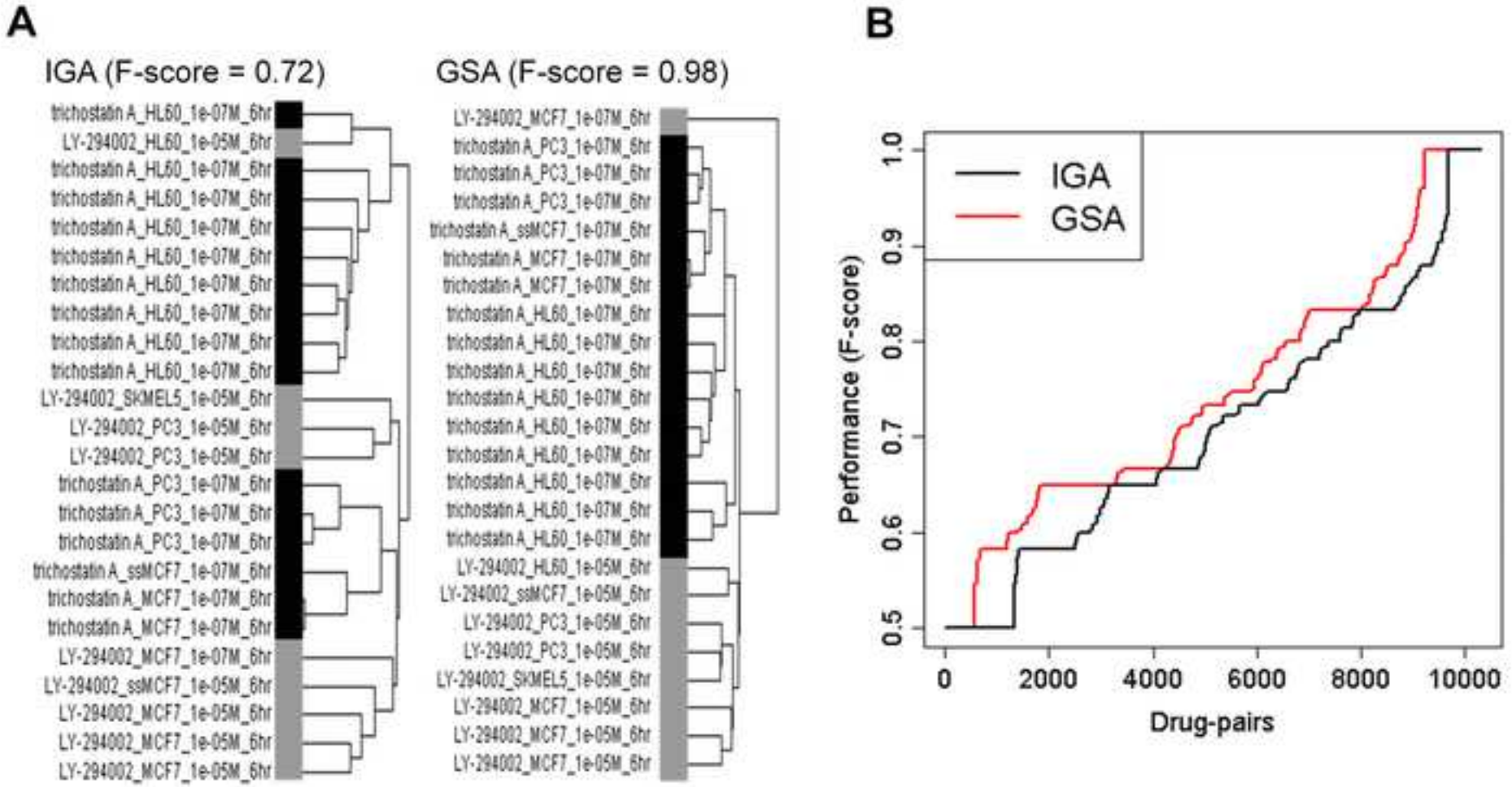


Figure 5
Click here to download high resolution image

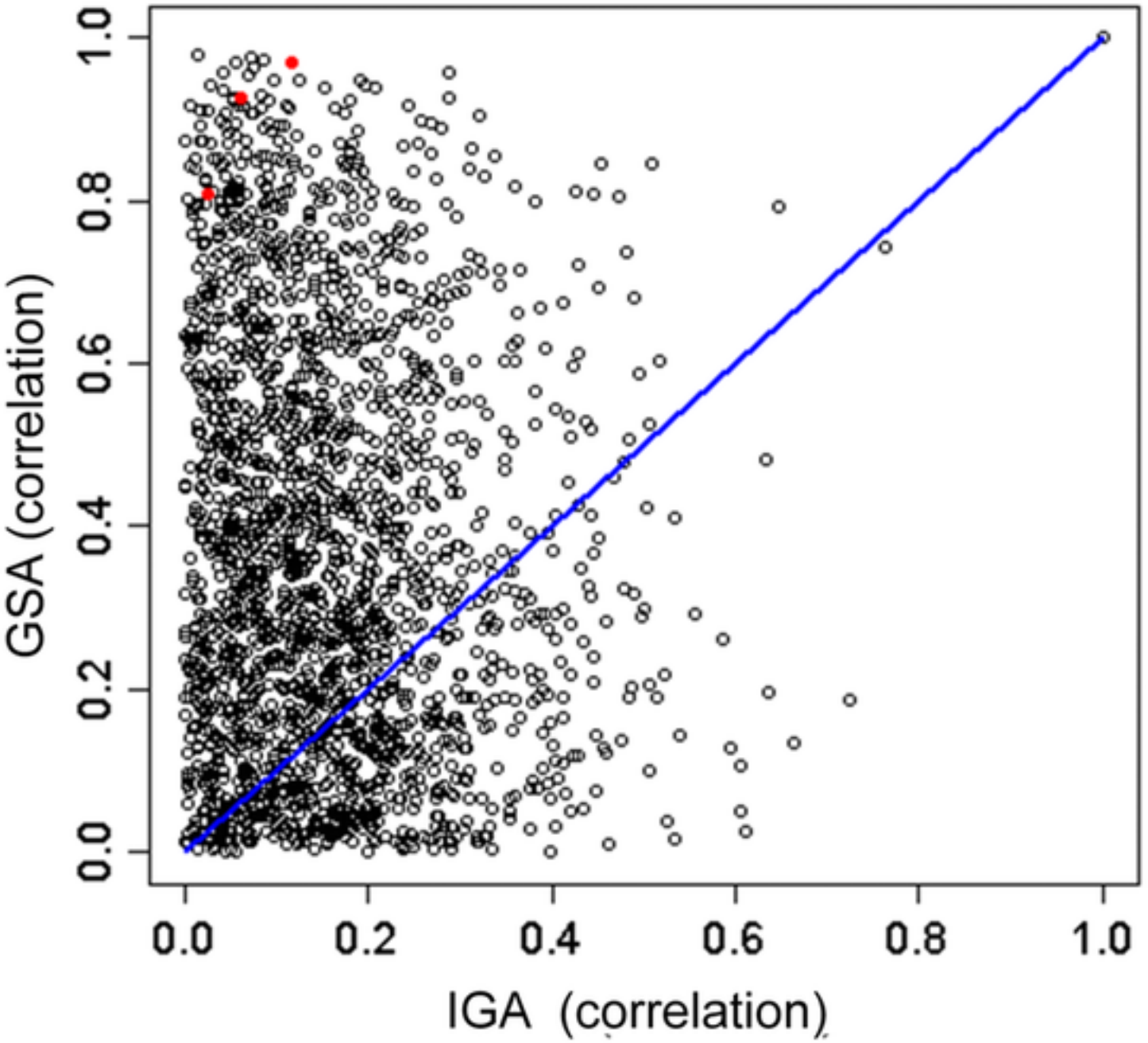


Figure 6
Click here to download high resolution image

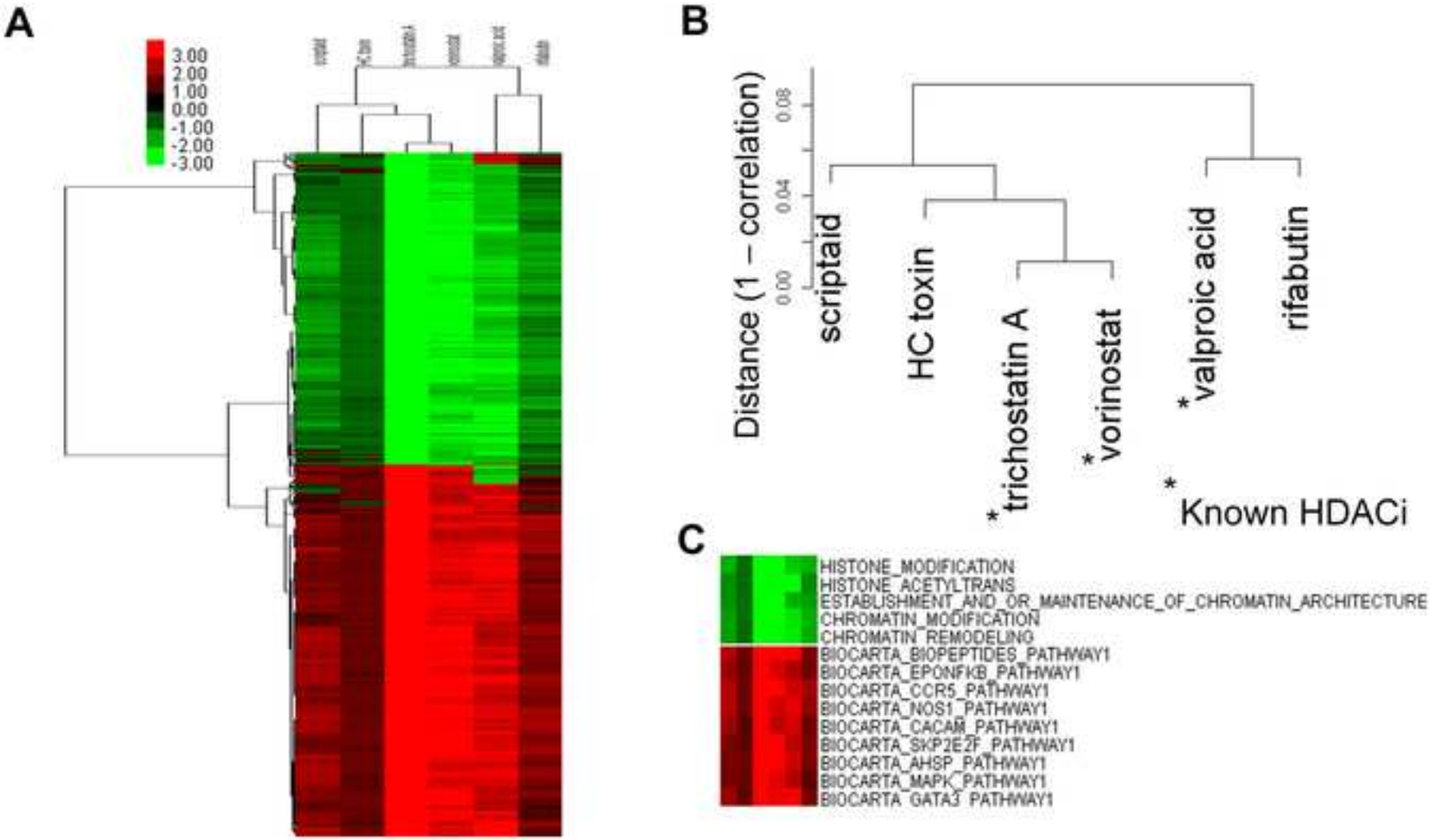


Figure 7
Click here to download high resolution image

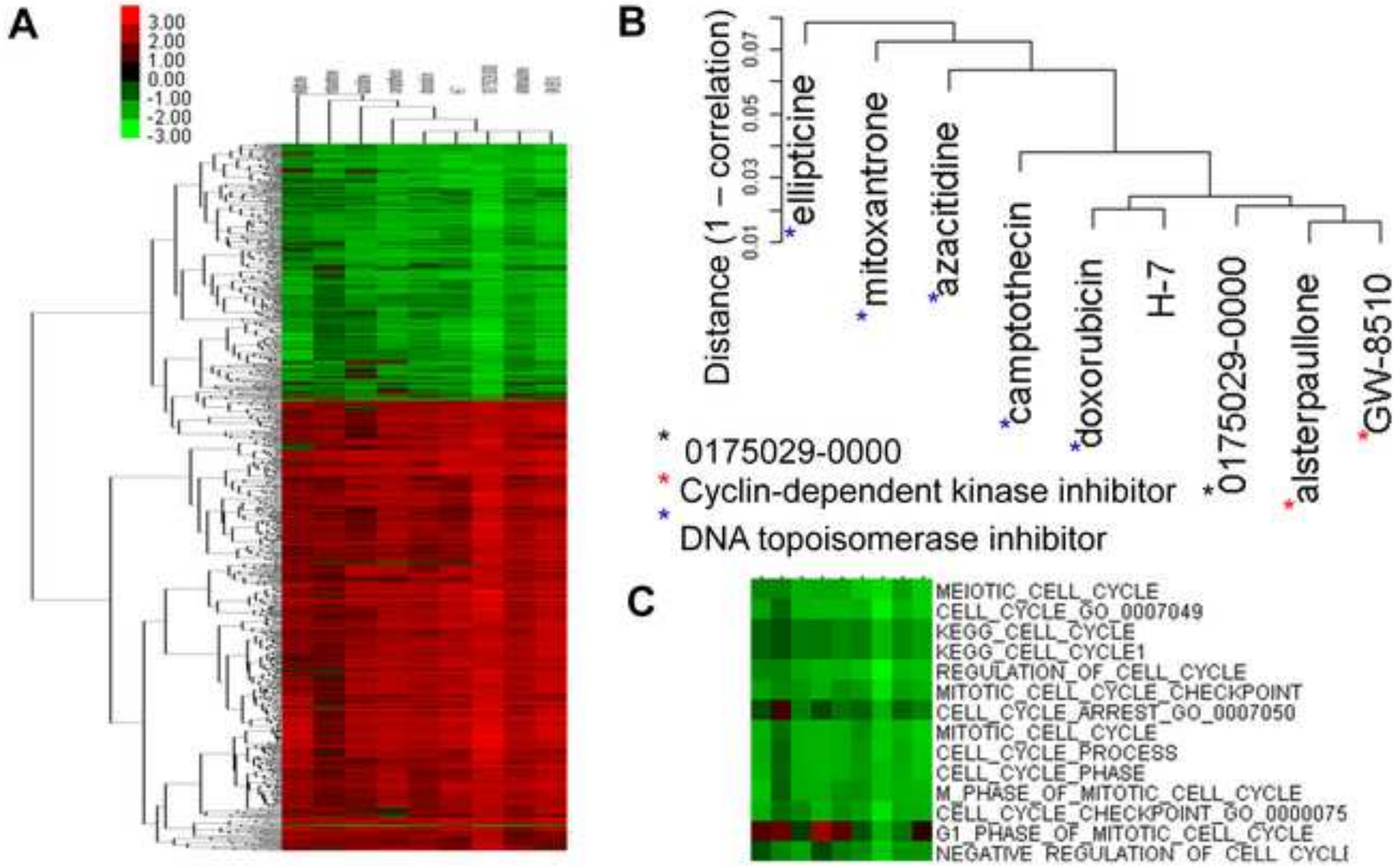
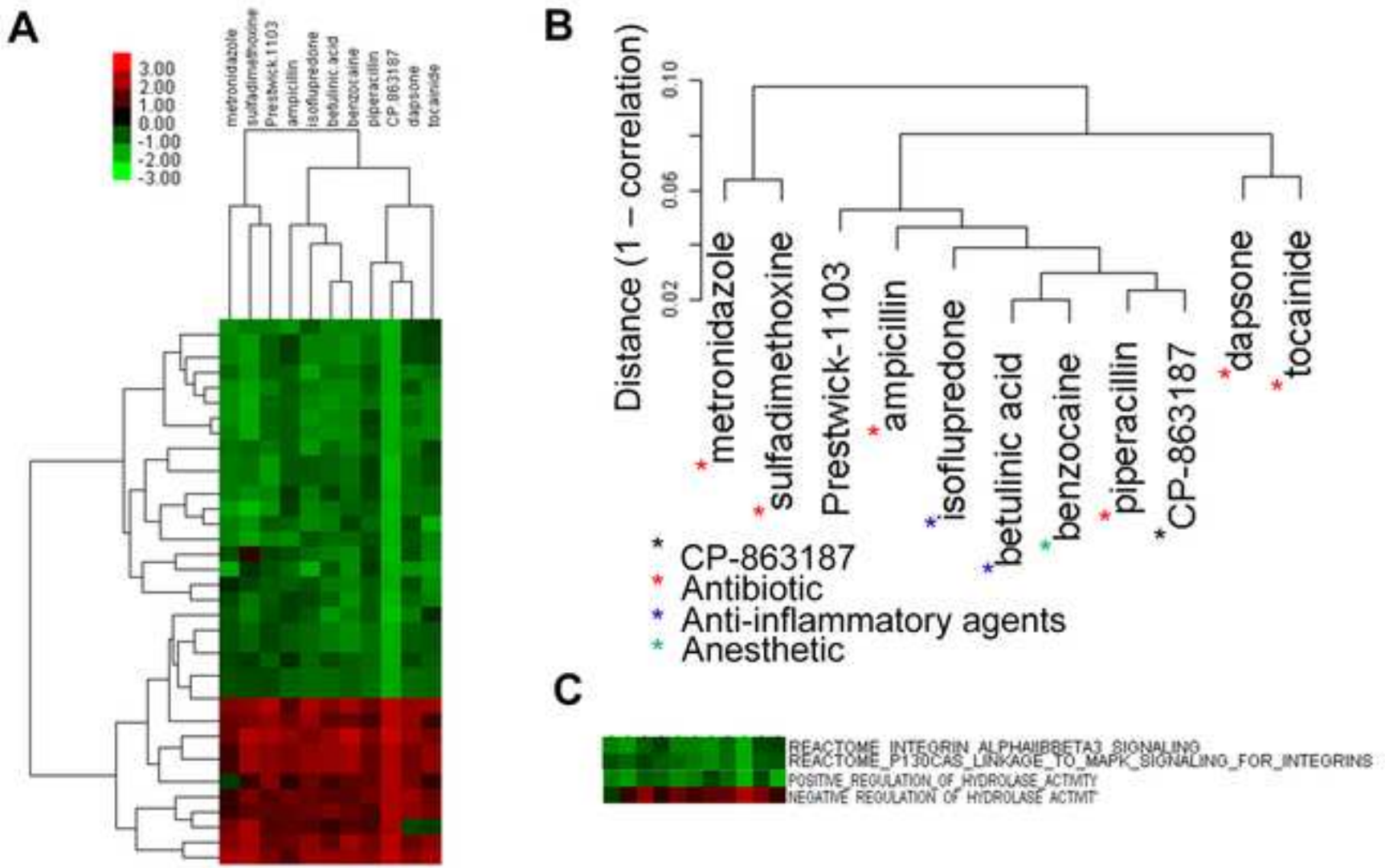


Figure 8
Click here to download high resolution image



Supporting Information: Figure S1.tif

Supporting Information: Figure S2.tif

Supporting Information: Figure S3.tif

Supporting Information: Figure S4.tif

Supporting Information: Figure S5.tif

Supporting Information: Figure S6.tif

Supporting Information: Figure S7.tif

Supporting Information: Figure S8.tif

Supporting Information: Figure S9.tif

Supporting Information: Figure S10.tif

Supporting Information: Figure S11.tif

Supporting Information: Figure S12.tif

Supporting Information: Figure S13.tif

Supporting Information: Figure S14.tif

Supporting Information: Figure S15.tif

Supporting Information: Figure S16.tif

Supporting Information: Figure S17.tif

Supporting Information: Figure S18.tif

Supporting Information: Figure S19.tif

Supporting Information: Figure S20.tif