A Method for the Detection of Meaningful and Reproducible Group Signatures from Gene Expression Profiles

Louis Licamele[1,2] and Lise Getoor[1]

[1] University of Maryland, College Park MD 20740, USA,
licamele@cs.umd.edu
[2] Vanda Pharmaceuticals, Inc., Rockville MD 20850, USA

Abstract. Gene expression microarrays are commonly used to detect the biological signature of a disease or to gain a better understanding of the underlying mechanism of how a group of drugs treat a specific disease. The outcome of such experiments, e.g., the signature, is a list of differentially expressed genes. Reproducibility across independent experiments remains a challenge. We are interested in creating a method that can detect the shared signature of a group of expression profiles, e.g., a group of samples from individuals with the same disease or a group of drugs that treat the same therapeutic indication. We have developed a novel Weighted Influence - Rank of Ranks (WIMRR) method, and we demonstrate its ability to produce both meaningful and reproducible group signatures.

Keywords: gene expression analysis, gene expression profiles, drug discovery, bioinformatics, data mining.

1 Background

Microarray technology is often credited with leading the advancement in the field of modern biological research and was coined as an Array of Hope shortly after its introduction [1]. As microarrays have become commonplace in the laboratory, the amount of gene expression data available in the public domain continues to grow at a rapid pace. Microarray experiments, whether they set out to discover biomarkers for a particular disease or to characterize a group of similar tissue samples, tend to have the same outcome: a list of differentially expressed genes (DEGs). In recent years, a growing debate has developed surrounding the scientific validity of microarrays in respect to their reliability [2-3]. Low reproducibility of DEGs across independent experiments testing the same hypothesis has become the norm [4]. Novel methods to detect robust group signatures from gene expression experiments are needed.

Gene expression profiling has traditionally been used to detect genetic differences between various types of groups including detecting gender differences [5], predicting cancer prognoses [6], segmenting and explaining diseases and their

J. Suzuki and T. Nakano (Eds.): BIONETICS 2010, LNICST 87, pp. 387–401, 2012.

subtypes [7], and understanding the underlying mechanism of biological processes and pathways [8]. Gene expression data are good sources for investigating and predicting the potential therapeutic effects of a drug because they characterize the response of the cell to external stimuli. A method that generates more reliable and reproducible results (e.g., lists of DEGs) from gene expression data is well positioned to become the core predictive model of a drug discovery system.

It is important to note, however, that many factors complicate analysis of gene expression experiments, including assumptions about the biological processing of mRNA and confounding factors inherent in mRNA expression data. Furthermore, reproducibility has remained low among these types of experiments, calling into doubt the validity of the detected signatures. For example, using an identical set of RNA samples across several different commercial platforms, Tan et al. [9] found only four common DEGs. Both Ramalho-Santos [10] and Ivanova [11] independently found only six DEGs in common among roughly 200 that had been identified in each study (even though they had a similar study design using the same platform). In another study by Miller et al. [12], who compared the effect of varying platforms on the same samples, there were only 11 DEGs in common of 425 DEGs that were found by CodeLink and 138 DEGs found by the Affymetrix platform. These are all examples of studies that exhibit how current methods are producing irreproducible signatures. This lack of reproducible findings indicates that false positives are being detected, and that these methods may be overfitting the data. Furthermore, many methods are complex and only explain a group in a piecewise fashion (e.g., a decision tree-type model). We believe that the ideal method does not require such strict filtering and instead dynamically weights the influence of each probe based on the relative rank of that probe within each member of the group.

We propose the creation of a group profile that will serve as the representative profile for a given group of interest. A gene expression profile is the representation of the activity of thousands of genes at once for a given sample. A group profile represents the shared activity of these thousands of genes across all of the member samples belonging to the group. For example, we can create a group profile consisting of all available antipsychotic drugs; we refer to this as an antipsychotic profile. Traditionally, researchers attempt to find probes or genes that form the signature for a group by evaluating probes above a certain fold-change threshold. These methods will detect the signature common to the group in the rare case that the shared effect is incredibly strong (and there are no large experimental biases between the expression profiles). However, the majority of the time, the true signal is missed because it is not significantly up- or down-expressed in every one of the instances that make up a group (we refer to this as the full group). These methods preferentially detect very big changes within a subgroup of samples and then merge all of these differentially expressed genes with a combination function. Unfortunately, this approach does not find true signatures common to the full group and allows the method to overfit the data. Our method differs from most previous methods by focusing on detecting

signatures common to the full group, signatures that are normally overlooked by other methods, e.g., decision trees and support vector machines, which can explain a group as a combination of rules defining unknown subgroups.

The representation of a group profile is a ranked list of all probesets on the microarray. A benefit of our approach is that this is the same representation as a single profile. This representation allows any current and future methods for non-parametric gene expression data to be used with our group profiles. We can focus on the most up- and down-expressed probesets from the profile, which we refer to as the signature of the group (separately they are the up and down signatures respectively). For example, we can make use of methods developed by others (e.g., Connectivity Map (CMAP) [13]) to use this antipsychotic group profile to search a database for drugs sharing the same signature. Alternatively, we can use still other methods (e.g., the L2L Microarray Analysis Tool [14]) to evaluate if any particular biological process is overrepresented within this signature, an approach that would provide additional insight into the common mechanism of antipsychotic therapies.

In this paper, we introduce and describe our rank of ranks method for group profile creation. We evaluate the utility of this group analysis method using a pilot study in which we focus on the antipsychotic group from the original CMAP build 01 dataset. Our evaluation consists of both understanding the group profiles biologically and demonstrating the ability to use a signature from these profiles as a predictive model of therapeutic use. We conclude with a full analysis of the newer, and larger CMAP build 02 dataset, including a sensitivity evaluation of each group as well as the validation of the most robust profiles within an independent dataset. All the results are available at GEPedia.org.

2 Problem Definition

Given a database D of treatments (i.e., drugs or other compounds), D = t_1,..., t_n, we are interested in creating a set of group profiles. A group can be defined as a set of instances (e.g., cells treated with a particular drug) that share something of interest in common (e.g., the same therapeutic use, mechanism of action, side effect, chemical structure). We are interested in understanding what is biologically common for a given group profile as well as evaluating the ability to query the database with the group profile to predict new members of the group. Our goal is to discover other drugs or treatments, perhaps originally developed for a different therapeutic purpose, which are likely to also share the same therapeutic properties as the query group. These therapeutic agents are thus good candidates for which new uses can then be evaluated.

For each treatment instance t in the database, there is both general information about the experimental conditions of the sample as well as the actual experiment data from the microarray itself. The gene expression profile is represented as a ranked list (amplitude of the treatment as compared to the control). Information specific to the treatment (i.e., the name of the drug, the therapeutic class [class] and subclass [subclass] as defined by the chemicals Anatomical

Therapeutic Chemical [ATC] code) is represented. There is also information that describes the experimental conditions of the sample, specifically the molar amount of substance (mol), the vehicle used for delivery of the drug (e.g., water, EtOH, MeOH, DMSO), and the batch or round in which the sample was run. A group, and therefore a group profile, can be created from any of these meta-labels associated with the samples.

3 Group Profile Creation (Weighted Influence Model - Rank of Ranks Method)

Previous methods have demonstrated that weighted distribution-based statistics can be more robust in detecting similarity in the pairwise comparison of gene expression data [13]; therefore, we propose a method for determining what is common among a group by also using a weighted method. This dynamic weighting of probes allows us to avoid strictly filtering any probes as is done with a fold-change threshold approach. We calculate the average rank of each probe across the members of the group and then re-rank the probes based on this average rank. We refer to this as the Weighted Influence Model, Rank of Ranks (WIMRR) method. The rank of each probe within each treatment t is known: rank(p, probes(t)). Let us assume we have a binary membership function, member(t, g), that returns 1 if treatment instance t is a member of group g and returns 0 otherwise. The size of the group is equal to the number of treatment instances that are members of the group. The average rank for each probe is then calculated. Given this set of average ranks across the members of a particular group, the probes are now re-ranked according to how consistently they are up- or down-expressed across the group. We define Profile(g) as the probes in probes(g) sorted by their average rank across all members of the group.

4 Group Profile Evaluation - A Pilot Study

We make use of the original CMAP dataset (build 01) from the Broad Institute to evaluate our group profile method as part of a pilot study. We refer to this as the CMAP 1.0 dataset. We use this smaller, simpler dataset to characterize our method. Later, we analyze the newer CMAP build 02 dataset (CMAP 2.0), which contains many more treatments. For each treatment instance in the CMAP dataset, probe sets are first ranked based on their level of expression relative to the vehicle control in a fashion similar to the method described by Lamb et al. [13]. A group profile is then created for each therapeutic use according to the ChemBank annotation for the instances using our novel WIMRR method. The signature of each group profile is created by selecting the top and bottom k probes. For this evaluation, we set $k = 50$.

4.1 Antipsychotics from Pilot Study

We focus on the antipsychotic profile from the CMAP 1.0 dataset as an example by which to analyze the WIMMR group profile creation method. The

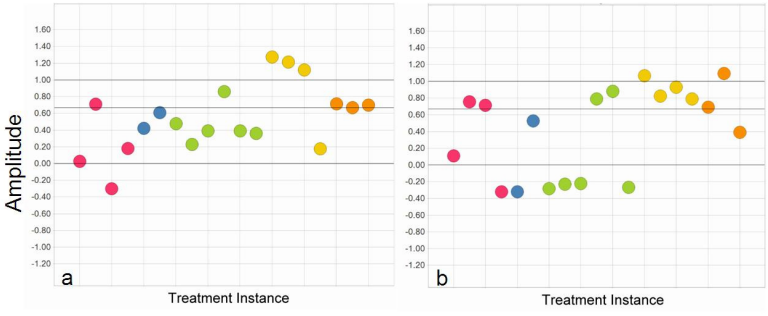
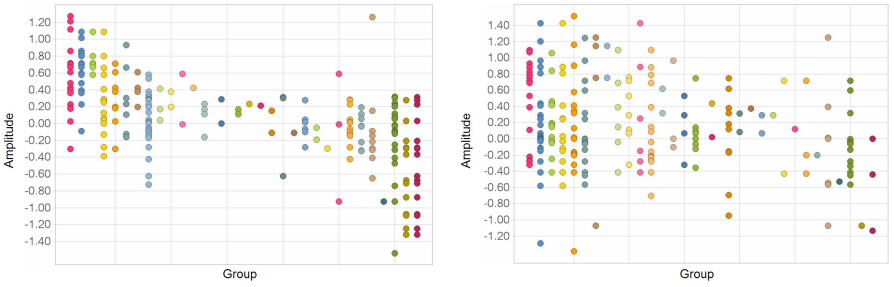


Fig. 1. The amplitude values for a) the top probe found by the group profile method is from the BHLHB2 gene and b) the top probe by the fold-change method that is greater than 2. The lines correspond to a fold-change of 2 and 3, respectively.

antipsychotic group is selected as the example because it includes a large number of unique drugs. The instances from the CMAP 1.0 dataset that are labeled as antipsychotic agents according to ChemBank are used to create this group. The antipsychotics profiled in this dataset include chlorpromazine, clozapine, haloperidol, thioridazine, and trifluoperazine. There are 19 profiles total for this group, consisting of replicates across different concentrations. The group profile is created and the top and bottom 50 probes are selected to serve as the signature for this group.

The top and bottom probes can both provide valuable insight. We focus on the top 50 probes, but the same analysis can be performed with the bottom 50 probes in an analogous way. The amplitude value for the top probe (Affymetrix probe id 201170 s at) is shown in Fig. 1A. This probe, which corresponds to the basic helix-loop-helix domain containing, class B, 2 (BHLHB2) gene, is almost exclusively up-expressed in all of the antipsychotic instances. We evaluate the specificity of this probe by determining how this probe behaves across the whole database (Fig. 2A). All but one of the antipsychotic instances (pink dots in first column) show a clear increase in expression levels. The next set of groups all contain drugs that are known to also act as antipsychotics; this is expected if this probe is predictive of antipsychotic activity. The second group is the tranquilizers (includes prochlorperazine, fluphenazine, and trifluoperazine), the third group is antiemetics (includes prochlorpromazine and trifluoperazine), and the fourth group is the antineoplastics (includes prochlorpromazine). There is a clear pattern of antipsychotic activity related to the up-expression of this probe across the database.

We now compare what we have seen with the top probe from our method with a probe selected using more conventional methods. A potential alternative method for selecting probes (and genes) of interest that has been used extensively in the field has been to select probes that are commonly up-, or down-, expressed above a particular threshold. The most common thresholds used in the literature are fold-changes greater than or equal to either 2 or 3, which correspond to amplitude values of 0.67 and 1.0, respectively. We select the best probe from this



(a) (b)

Fig. 2. (a) Specificity of top probe, BHLHB2, from the group profile method. From left to right, the first group is the antipsychotics, the second is the tranquilizers (includes prochlorperazine, fluphenazine, and trifluoperazine), the third group is antiemetics (includes prochlorpromazine and trifluoperazine), and the fourth group is the antineoplastics (includes prochlorpromazine). (b) The top probe from the fold-change greater than 2 method is not specific to antipsychotics. The first group is the antipsychotics, the second is anti-inflammatory, the third is antineoplastics and the fourth is analgesics.

alternative method, determining the probe that exhibits a fold-change greater than 2 in the most antipsychotic instances. The best probe found by this method was for the SEMA3B gene. The amplitude values across all of the antipsychotics for this probe are shown in Fig. 1B. Note that even though some of the individual instances have a very high amplitude value, roughly one-third of the instances have the opposite effect. Again, we determine the specificity of this probe to the antipsychotics by evaluating how it behaves across the rest of the database (Fig. 2B). Visually, we can see that this probe is not specific to the antipsychotics at all.

As validation of our group profile method, we examine BDNF. BDNF (Brain-Derived Neurotrophic Factor) has long been a candidate gene for both schizophrenia and bipolar disorder [15-17]. This additional information demonstrates how this method can give insight into the etiology of the disease that these drugs treat. It also demonstrates how the method extends beyond solely learning about the mechanism of action of drugs. Turning back to the best result from the alternative (fold-change threshold) method, there is no known link between SEMA3B and antipsychotics, schizophrenia, bipolar disorder or other topics expected to be related to antipsychotic agents.

5 Understanding Group Signatures

As mentioned earlier, one of the major benefits of our group profile method is that we can easily plug our group profile results into many algorithms and tools developed to analyze (individual) gene expression data. The probe sets in the group profile signatures can be evaluated for significant overrepresentation of gene ontology (GO) terms, e.g., GO Biological Processes, using the L2L analysis

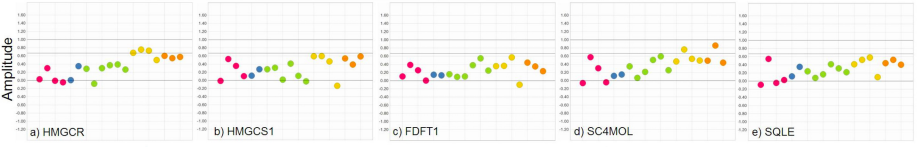


Fig. 3. The amplitude values for the probes in the most significantly up-expressed GO term for the antipsychotic group: sterol biosynthetic process. The probes correspond to the a) HMGCR, b) HMGCS1, c) FDFT1, d) SC4MOL, and e) SQLE genes.

tool [14]. Given a list of probe sets, e.g., DEGS, and a list to match them to, e.g. GO:BiolProc, L2L calculates the expected number of matches given the probes found on the microarray. From the actual and expected matches, an enrichment score and the corresponding P value for each GO term is then calculated [18]. Additional lists of published probe sets are also evaluated, including GO Cellular Component, GO Molecular Function, reactome protein-protein interactions [19], predicted human MicroRNA targets [20], and cancer gene expression modules [21].

We use the L2L method to evaluate the example group profile of the antipsychotics. The top 50 probes are evaluated for significant overrepresentation of GO Biological Process terms. The most significant terms are all related to lipid homeostasis (Table 1). There are five genes involved in the sterol biosynthetic process (GO:0016126) within the top 50 probes. Out of over 22,000 probes, only 41 are annotated as belonging to this GO term, so 0.11 probes for this term are expected by chance. This GO term, along with the next three in Table 1, pass Bonferroni correction for multiple testing ($p \leq 1.11E\text{-}05$ after correction for all four GO terms). The amplitude values for the five genes that are involved in this pathway are shown in Fig. 3. There is an obvious trend that the expression of these probes is increased in almost every antipsychotic instance in our database. However, even though they are always up-expressed, the amplitude value is normally below the common threshold used by other researchers (fold-change of 2 or 3). This is a good example of how the group profile method is able to detect consistent, and therefore more robust, signals in gene expression data; signals that are normally overlooked by current methods.

Table 1. The most significantly overrepresented GO Biological Process terms from the up-expressed antipsychotic signature

GO Term	GO ID	Probes	Expected	Actual	Enrichment	P Value
sterol biosynthetic process	GO:0016126	41	0.11	5	44.73	1.04E-07*
steroid biosynthetic process	GO:0006694	88	0.24	5	20.84	4.89E-06*
alcohol metabolic process	GO:0006066	371	1.01	8	7.91	1.05E-05*
sterol metabolic process	GO:0016125	104	0.28	5	17.63	1.11E-05*
steroid metabolic process	GO:0008202	211	0.58	6	10.43	2.91E-05
cholesterol biosynthetic process	GO:0006695	31	0.08	3	35.50	8.60E-05
lipid biosynthetic process	GO:0008610	281	0.77	6	7.83	1.40E-04

Support for these GO Biological Process findings comes from the work of other researchers aimed at understanding the molecular origin of the known metabolic side effects of antipsychotics that include increased weight gain and propensity to adiposity and insulin resistance [22]. Our observation is consistent with literature reports of an antipsychotic drug effect on the same or overlapping sets of genes involved in lipid homeostasis. Interestingly, a genome-wide screen of Saccharomyces cerevisiae heterozygotes had previously revealed that the antipsychotics haloperidol, chlorpromazine, and trifluoperazine had a strong effect on genes involved in yeast fatty acid biosynthesis (OLE1, the ortholog of the human SCD), sterol biosynthesis or phospholipid transport [23].

6 Querying with Group Signatures

The WIMRR method is able to create a specific representative profile for a group of gene expression profiles. We have demonstrated the ability to gain insight into the mechanism of action of a drug class (as well as the disease that it is used to treat) using WIMRR group profiles. Now we utilize the strength of a group profile to detect and predict the therapeutic use of a drug based on an individual gene expression profile.

We use the truncated KS statistics described previously for pairwise (instance-to-instance) similarity calculations [13] to detect instances that are similar to a group profile of interest (instance-to-group). Using the same antipsychotic group profile, we query the database of instances using k = 50 (i.e., the signature discussed previously). The instances most similar to this group profile are shown in Table 2, along with their KS score. The last column in Table 2 represents membership in the group of interest, i.e., if a given treatment is a member of the antipsychotic group used in creating the profile. Scanning the list, we see that prochlorperazine (Instance ID = 995) is the most similar non-antipsychotic drug. It turns out that prochlorperazine is in fact a phenothiazine antipsychotic; however, it is more commonly used for the treatment of nausea and vertigo. Prochlorperazine is a highly potent neuroleptic, which is considered a typical antipsychotic. The next non-antipsychotic is fluphenazine, for which two replicates show up as extremely similar to the antipsychotic profile. Fluphenazine is a typical antipsychotic drug used for the treatment of psychosis, e.g., schizophrenia and bipolar disorder. Fluphenazine is also an extremely potent phenothiazine. The next novel compound is calmidazolium, which is a calmodulin inhibitor. Though it is not used as an antipsychotic, it is validated because many of the antipsychotic drugs are potent inhibitors of calmodulin [24].

In fact, it turns out that many of the most significant results are already used as an antipsychotic agent even though they are not labeled in ChemBank as such. These examples are a validation of our method and increase the confidence in the other results that are not already supported by the literature, as these are potentially the important and still unknown alternative uses for these therapeutic agents.

Table 2. The database was queried with the antipsychotic signature (up and down together) and the most similar

Rank	Instance ID	Name	KS Score	Antipsychotic Member
1	1010	thioridazine[INN]	1.58	X
2	1068	thioridazine[INN]	1.483	X
3	1004	trifluoperazine[INN]	1.469	X
4	995	prochlorperazine[INN]	1.435	
5	910	trifluoperazine[INN]	1.408	X
6	417	thioridazine[INN]	1.387	X
7	983	haloperidol[INN]	1.352	X
8	1024	haloperidol[INN]	1.346	X
9	1017	fluphenazine[INN]	1.317	
10	1075	fluphenazine[INN]	1.293	
11	421	trifluoperazine[INN]	1.256	X
12	906	calmidazolium	1.223	
13	870	pyrvinium	1.209	
14	1053	prochlorperazine[INN]	1.201	
15	418	haloperidol[INN]	1.167	X
16	1009	clozapine[INN]	1.162	X
17	419	chlorpromazine[INN]	1.138	X
18	1003	nordihydroguaiareticacid	1.1	
19	416	clozapine[INN]	1.09	X
20	1105	monensin[INN]	1.077	
21	978	pyrvinium	1.065	
22	893	pararosaniline	1.051	
23	882	ionomycin	1.027	
24	941	rottlerin	1.023	
25	1012	troglitazone[INN]	1.018	
26	1082	haloperidol[INN]	1.009	X
27	1055	chlorpromazine[INN]	0.997	X
28	1041	haloperidol[INN]	0.992	X
29	997	chlorpromazine[INN]	0.99	X

7 Analysis of CMAP V2.0

We have introduced our method for creating group profiles from gene expression data. For this, we have used the original version of the CMAP dataset as our motivating example. We have seen how we can gain biological insight from these profiles as well as how to predict new members by querying the group signature. Here we present our analysis of the newly released CMAP 2.0 dataset with our method and describe the results. Groups are defined according to the compounds ATC code. We have analyzed all the groups at ATC level 3 and level 4. ATC level 3 defines the therapeutic/pharmacological subgroup, e.g., N05A = Antipsychotics. ATC level 4 further defines a subgroup based on chemical properties, e.g., N05AE = Indole Derivative Antipsychotics. We focus on groups with three or more compounds, resulting in 117 ATC level 3 groups and 148 level 4 groups.

7.1 GEPedia.org

We have compiled all of the results from our analysis of CMAP 2.0 and have made them available online at GEPedia.org. In this manuscript, we focus on evaluating our group profile method and only highlight a few interesting results from this analysis. We assume that there are many undiscovered biological insights within this dataset. We are releasing all of the data allowing researchers to examine the results for further discoveries and to compare with their own datasets.

Currently, the organization of GEPedia.org is based around the analysis presented in this paper. We include the output of the complete analysis of all groups. For every group, i.e., for all ATC groups, we have made available a) the profile itself, including the up- and down-expressed signatures, b) the analysis of the profile according to the L2L tool, c) the sensitivity analysis of the profile, and d) the results of searching across the database with the signature. In the future, we plan to modify the website to allow more interactive analysis of the data in addition to allowing scientists to upload, analyze, and share their own gene expression data.

7.2 Sensitivity Analysis and Independent Validation

A sensitivity analysis is performed in order to prioritize the evaluation of the most promising group profiles. To do this, we randomly divide the group into two equal-sized subgroups: a training group that contains half of the treatment instances from the group and a test group composed of the remainder of the group. A group profile is created for both subgroups, and the top (up-tags) and bottom (down-tags) 100 probes are selected. The number of probes in common between the two subgroups is calculated for both the up- and down-tags respectively. The treatment instances are re-randomized and this process is repeated for a total of 10 iterations. The average number of probes in common across the 10 iterations is calculated for the up- and down-tags. The higher the average number of probes in common (for the up-tags, down-tags, or both up- and down-tags), the more robust we consider the group profile. From this value, i.e., the average number of probes in common, we estimate the probability assuming a binomial distribution.

The most robust ATC level 3 (therapeutic/pharmacological) group profiles are shown in Table 3 for both the up and down signatures together (full results in Supplemental Table 1 and Supplemental Table 2 for the up and down signatures, respectively). The full results for the level 4 ATC (chemical/therapeutic/pharmacological) group profiles for the up and down signatures are shown in Supplemental Table 3 and Supplemental Table 4, respectively. The associated probability for each of these profiles is also listed. The observed probabilities indicate that some of these profiles are not random. Corrections for multiple testing are performed, and the Bonferroni-corrected P values are also included in each of the tables.

At the onset of this paper, we mention that we are interested in creating a gene expression profile for groups sharing a therapeutic use, and so we focus our analysis on the ATC level 3 groups. There are 36 groups with significant (Bonferroni-corrected $P^* < 0.05$) up-expressed signatures and 28 for the down-expressed signatures. Out of these groups, 25 groups are robust for both up- and down-expressed signatures. While a robust up- or down-expressed signature can independently give novel insight into the underlying shared biological function of a group, we focus on groups that are significant for both because we also want to use these profiles to help predict novel uses of the drugs in our database. The similarity metric that we have adopted requires both the up and down signatures

Table 3. The most robust group profiles across the whole database are presented here

Group	Drugs	Up	P Up*	Down	P Down*	Label
N05A	28	70.6	3.09E-139	49.4	9.59E-86	Antipsychotics
R06A	27	23.7	1.07E-31	12	5.70E-12	Antihistamines for Systemic Use
N06A	25	29.6	5.70E-43	12.1	4.04E-12	Antidepressants
D07A	19	49.8	1.11E-86	19.7	1.64E-24	Corticosteroids, Plain
G01A	18	12.1	4.04E-12	6.5	1.98E-04	Antiinfectives and Antiseptics
D01A	16	10.2	2.42E-09	11.7	1.59E-11	Antifungals for Topical Use
S01B	16	7.9	3.36E-06	8.9	1.56E-07	Antiinflammatory Agents
N03A	11	13.1	1.22E-13	17.8	3.01E-21	Antiepileptics
H02A	11	18.5	1.94E-22	5.2	6.72E-03	Corticosteroids for Systemic Use
R03B	10	15.6	1.35E-17	4.6	3.09E-02	Drugs for Obstructive Airway Diseases, Inhalents
D10A	9	18.7	8.80E-23	6.5	1.98E-04	Anti-Acne Preparations (Topical)
L04A	8	39.9	2.46E-64	31.4	1.47E-46	Immunosuppressants
D07X	8	25.3	1.12E-34	6.7	1.13E-04	Corticosteroids, (Dermatologicals)
G03D	8	11.1	1.22E-10	4.7	2.41E-02	Progestogens
L01X	7	19.9	7.31E-25	11.5	3.16E-11	Other Antineoplastic Agents
L02B	6	19.3	8.12E-24	13.5	2.93E-14	Hormone Antagonists (and related)
R03A	6	9.8	8.89E-09	5.3	5.17E-03	Adrenergics, Inhalents
C08C	6	5.6	2.34E-03	4.5	3.95E-02	Selective Calcium Channel Blockers
G03C	5	30.5	9.35E-45	10.1	3.36E-09	Estrogens
S01C	5	10.3	1.74E-09	5.3	5.17E-03	Anti-inflammatory -infective (Combo)
C08E	4	11.7	1.59E-11	7.3	1.99E-05	Non-selective Calcium Channel Blockers
C01A	3	61.1	2.94E-114	62.2	4.60E-117	Cardiac Glycosides
L01D	3	6.4	2.62E-04	22.9	3.16E-30	Cytotoxic Antibiotics (and related
L01B	3	7.9	3.36E-06	19.8	1.09E-24	Antimetabolites

to be used together. We now present a deeper analysis of the most robust profiles. The larger the set of unique drugs that compose a group, the more evidence we have that the therapeutic mechanism is what is being detected in the profile. For this reason, we focus on the significant groups with the largest number of unique drugs. We compare our results to those from an independent dataset using the same method (Table 4).

7.3 Antipsychotic Group (N05A)

We start our analysis with the largest group that meets our significance threshold: the antipsychotic group with 28 unique drugs. The ATC level 3 code for this group is N05A. The antipsychotic profile is the most robust result from the ATC level 3 groups when evaluating the up-expressed signature (Bonferroni-corrected P value: P*=3.10E-139). This corresponds to an average of 70.6 probes that are shared between the top 100 probes of two random subgroups. Interestingly, this same group is the second most significant when evaluating the robustness of the down-expressed signature (P*=9.59E-86; Average probes in common = 49.4). In an attempt to discover what the underlying shared biological process is within these antipsychotic agents, we turn to the L2L analysis. The most overrepresented GO Biological Process term is Sterol Biosynthetic Process (GO:0016126; P*=6.45E-20). This is the same term that was found over-expressed within the smaller pilot study and demonstrates that our group profile method can detect the true signature with a small set of samples.

We have the ability to compare this profile with the antipsychotic profile recently published by Polymeropoulos et al. [25]. It is important to note that

these two profiles were created by two independent laboratories, with different cell lines and with a different, but overlapping, set of antipsychotics. These two profiles are very similar, and they share 34 probes in common among their top 100 probes (P=6.42E-54). The most significant GO Biological Process term from the Polymeropoulos et al. antipsychotic group profile is Lipid Biosynthesis. Given the significant overlap of the profiles, it is not surprising that this term is actually a grandparent of Sterol Biosynthetic Process (connected through the GO term Steroid Biosynthetic Process). The GO term Lipid Biosynthesis is also highly significant within the CMAP v2.0 antipsychotic group (P*=2.70E-13).

The down-expressed signatures also share several probes in common (Probes=6; P=6.79E-06). The GO Biological Process analysis points to a significant down-regulation of the DNA regulation process (GO:0006260; P*=3.61E-07). Barochovsky et al. have demonstrated in vivo that compounds acting on the central nervous system, specifically those that affect noradrenergic, dopaminergic, and serotoninergic neurotransmitters, reduce brain cell replication [26]. This observation of compounds acting on the CNS was a dose-dependent effect and was seen for both agonists and antagonists. This down-expressed signature, like the up-expressed signature, is well supported by the literature. The antipsychotic profile that we have discovered is robust, both in and across datasets. Furthermore, we have demonstrated the ability of our group profile method to give biological insights into the potentially unknown shared biological process exhibited by a group of drugs.

Table 4. The most robust profiles were evaluated against an independent dataset (Polymeropoulos et al)

Group	Vanda PDR Group	Probes In Common	P
N05A	CNS:Antipsychotics	34	6.42E-54
R06A	Resipiratory Agent:Histamine Antagonist	4	1.13E-03
N06A	CNS:Antidepressants	15	1.07E-18
D07A	Dermatological:Corticosteroids	30	7.88E-46

7.4 Antihistamine Group (R06A)

The second-largest group that meets our significance criteria is the antihistamines (full annotation: Antihistamines for Systemic Use; ATC Code: R06A). This group contains 27 unique drugs. The sensitivity analysis reveals 23.7 probes on average shared within the up-expressed signature and 12 for the down-expressed (P*=1.07E-31 and P*=5.70E-12, respectively). The up-expressed signature exhibits a common underlying theme related to negative regulation of I-kappaB kinase / NF-kappaB cascade (GO:0043124; P=6.08E-05). This GO signature is not as strong as some of the other profiles and is not significant when corrected for multiple testing. However, it is interesting to note that this signature is consistent with the known effect of antihistamines on NF-kappaB. Roumestan et al. have shown that antihistamines inhibit NF-kappaB through both H1 receptor-dependent and independent mechanisms [27]. This profile does

not replicate when compared to the equivalent group (Respiratory Agent: Histamine Antagonist) from the dataset presented by Polymeropoulos et al., though a similar trend is seen. The average number of probes in common is four and one respectively, for the up- and down-expressed signatures (P = 1.13E-03 and P = 3.60E-01).

7.5 Antidepressant Group (N06A)

Next, we discuss the third-largest group: the antidepressants (ATC Code: N06A). There are 25 unique drugs within this group. The sensitivity analysis results in an average of 29.6 and 12.1 probes in common for the up- and down-expressed signatures (P*=5.70E-43 and P*=4.04E-12, respectively). Evaluating the up-expressed signature, the most overrepresented GO Biological Process term is Sterol Biosynthetic Process (GO:0016126; P*=1.19E-09). This is the same core mechanism seen within the antipsychotic group, but this signature is seen on a smaller scale. Polymeropoulos et al. demonstrated the same relationship between the expression profile of antipsychotic and antidepressant drugs [25]. When we compare our antidepressant profile to the antidepressant profile from the dataset from Polymeropoulos et al., we find 15 probes in common (P=1.07E-18). The down-expressed signature does not reproduce within the Polymeropoulos et. al. dataset, sharing only one probe in common.

7.6 Corticosteroid Group (D07A)

The last group that we evaluate in depth is the corticosteroids (N=19; ATC Code: D07A). This profile is also robust according to the sensitivity analysis. The average number of probes in common for the up-expressed signature is 49.8 (P*=1.11E-86). The down-expressed signature has an average of 19.7 probes in common (P*=1.64E-24). Individually, the up- and down-expressed signatures do not exhibit a significant result for any GO Biological Process, but evaluated together they demonstrate an effect on the regulation of the interleukin-6 biosynthetic process (P*=1.38E-02). Corticosteroids are involved in a wide range of physiological systems such as stress response, immune response and regulation of inflammation. Interleukin-6 acts as both a pro-inflammatory and anti-inflammatory cytokine that can be secreted to stimulate response to trauma [28]. There is a significant overlap between this profile and the corresponding profile (Dermatological: Corticosteroids) from Polymeropoulos et al. The up-expressed signatures share 30 probes in common while the down-expressed share nine probes, corresponding to probabilities of P=7.88E-46 and 9.72E-10, respectively.

8 Conclusions

We have introduced and evaluated our method for creating group profiles from gene expression data. The ability to have reproducible sets of differentially expressed genes from microarray experiments has been a big challenge, and we have demonstrated how our method is able to overcome this obstacle. Furthermore, we

have illustrated how to gain biological insight from such group profiles as well as the ability to use them as a signature to query a database. In our example domain of a drug discovery system, this biological insight allows researchers to potentially learn about the etiology of the disease that these compounds are being used to treat and gives them a predictive tool to find novel uses for other drugs.

Though a major focus of this work has been to introduce our method and validate it across independent datasets, we are also releasing all group profiles from the full CMAP 2.0. This includes all corresponding meta-analysis that has been performed: L2L analysis, similarity searching results, etc. We this resource contains of hidden biological insight into many groups of drugs and their target diseases, and for further in-depth research.

There are many possible avenues of further improvements and research. Thus far, we have assumed that explicit groups are given a priori. Our sensitivity analysis validates how coherent a group is; however, it does not dictate what to do if the outcome is not positive. For example, a leave-one-out analysis can be done to exclude members that do not fit well within a group. Lastly, it is important to note that our method is focused on determining a reproducible genetic profile for a group of samples; in this case, drugs of a particular class. We provide no guarantee as to the uniqueness of such profiles and instead claim that these profiles can be used to compare groups. We have kept the full ranked list as the profile, and so it is straightforward for extensions to this method to be developed to further refine and learn what genetic components make up a more unique signature if that was the end goal. In keeping the full profile, i.e., the re-ranked list of probesets, we allow further research methods, which are developed for individual expression profiles, e.g., the L2L method, to also be applicable to our group profiles.

References

1. Lander, E.S.: Array of hope. Nat. Genet. 21, 3–4 (1999)
2. Frantz, S.: An array of problems. Nature Reviews Drug Discovery 4, 362–363 (2005)
3. Marshall, E.: Getting the noise out of gene arrays. Science 306, 630–631 (2004)
4. Ein-Dor, L., Zuk, O., Domany, E.: Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc. Natl. Acad. Sci. USA 103, 5923–5928 (2006)
5. Zhang, W., Huang, R.S., Duan, S., Dolan, M.E.: Gene set enrichment analyses revealed differences in gene expression patterns between males and females. Silico. Biol. 9, 55–63 (2009)
6. van t Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536 (2002)
7. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
8. DeRisi, J.L., Iyer, V.R., Brown, P.O.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686 (1997)
9. Tan, P.K., Downey, T.J., Spitznagel, E.L., Xu, P., et al.: Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. 31, 5676–5684 (2003)

10. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C., et al.: Stemness: transcriptional profiling of embryonic and adult stem cells. Science 298, 597–600 (2002)
11. Ivanova, N.B., Dimos, J.T., Schaniel, C., Hackney, J.A., et al.: A stem cell molecular signature. Science 298, 601–604 (2002)
12. Miller, R.M., Callahan, L.M., Casaceli, C., Chen, L., et al.: Dysregulation of gene expression in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-lesioned mouse substantia nigra. J. Neurosci. 24, 7445–7454 (2004)
13. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., et al.: The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313, 1929–1935 (2006)
14. Newman, J.C., Weiner, A.M.: L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol. 6, R81 (2005)
15. Gupta, M., Chauhan, C., Bhatnagar, P., Gupta, S., et al.: Genetic susceptibility to schizophrenia: role of dopaminergic pathway gene polymorphisms. Pharmacogenomics 10, 277–291 (2009)
16. Gerard, S.: Reviewing medications for bipolar disorder: understanding the mechanisms of action. The Journal of Clinical Psychiatry 70, e02 (2009)
17. Xiu, M., Hui, L., Dang, Y., Hou, T.D., et al.: Decreased serum BDNF levels in chronic institutionalized schizophrenia on long-term treatment with typical and atypical antipsychotics. Prog. Neuropsychopharmacol. Biol. Psychiatry 33, 1508–1512 (2009)
18. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000)
19. Vastrik, I., DEustachio, P., Schmidt, E., Joshi-Tope, G., et al.: Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 8, R39 (2007)
20. John, B., Enright, A.J., Aravin, A., Tuschl, T., et al.: Human MicroRNA targets. PLoS Biol. 2, e363 (2004)
21. Segal, E., Friedman, N., Koller, D., Regev, A.: A module map showing conditional activity of expression modules in cancer. Nat. Genet. 36, 1090–1098 (2004)
22. Newcomer, J.W., Sernyak, M.J.: Identifying metabolic risks with antipsychotics and monitoring and management strategies. J. Clin. Psychiatry 68, e17 (2007)
23. Lum, P.Y., Armour, C.D., Stepaniants, S.B., Cavet, G., et al.: Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. Cell 116, 121–137 (2004)
24. Donohoe, D.R., Aamodt, E.J., Osborn, E., Dwyer, D.S.: Antipsychotic drugs disrupt normal development in Caenorhabditis elegans via additional mechanisms besides dopamine and serotonin receptors. Pharmacol. Res. 54, 361–372 (2006)
25. Polymeropoulos, M.H., Licamele, L., Volpi, S., Mack, K., Mitkus, S.N., et al.: Common effect of antipsychotics on the biosynthesis and regulation of fatty acids and cholesterol supports a key role of lipid homeostasis in schizophrenia. Schizophr Res. 108, 134–142 (2009)
26. Barochovsky, O., Patel, A.J.: Effect of central nervous system acting drugs on brain cell replication in vitro. Neurochem. Res. 7, 1059–1074 (1982)
27. Roumestan, C., Henriquet, C., Gougat, C., Michel, A., et al.: Histamine H1-receptor antagonists inhibit nuclear factor-kappaB and activator protein-1 activities via H1-receptor-dependent and -independent mechanisms. Clin. Exp. Allergy 38, 947–956 (2008)
28. Heinrich, P.C., Behrmann, I., Haan, S., Hermanns, H.M., et al.: Principles of interleukin (IL)-6-type cytokine signalling and its regulation. Biochem. J. 374, 1–20 (2003)