Published in final edited form as:
Stat Anal Data Min. 2009 September 1; 2(3): 175–185. doi:10.1002/sam.10049.

Gene expression associations with the growth inhibitory effects of small molecules on live cells: specificity of effects and uniformity of mechanisms

Kerby Shedden[1], **Yang Yang**[1], and **Gus Rosania**[2]

[1] Department of Statistics, University of Michigan, Ann Arbor MI USA

[2] Department of Pharmaceutical Sciences, College of Pharmacy, University of Michigan, Ann Arbor MI USA

Abstract

The NCI60 human tumor cell line screen is a public resource for studying selective and non-selective growth inhibition of small molecules against cancer cells. By coupling growth inhibition screening data with biological characterizations of the different cell lines, it becomes possible to infer mechanisms of action underlying some of the observable patterns of selective activity. Using these data, mechanistic relationships have been identified including specific associations between single genes and small families of closely related compounds, and less specific relationships between biological processes involving several cooperating genes and broader families of compounds. Here we aim to characterize the degree to which such specific and general relationships are present in these data. A related question is whether genes tend to act with a uniform mechanism for all associated compounds, or whether multiple mechanisms are commonly involved. We address these two issues in a statistical framework placing special emphasis on the effects of measurement error in the gene expression and chemical screening data. We find that as measurement accuracy increases, the pattern of apparent associations shifts from one dominated by isolated gene/compound pairs, to one in which families consisting of an average of 25 compounds are associated to the same gene. At the same time, the number of genes that appear to play a role in influencing compound activities decreases. For less than half of the genes, the presence of both positive and negative correlations indicates pleiotropic associations with molecules via different mechanisms of action.

Keywords

High throughput screen; gene expression; chemical biology; measurement error; false discovery rate; toxicity

Introduction

The Developmental Therapeutics Program (DTP) of the National Cancer Institute has utilized a panel of 60 human tumor-derived cell lines to assess the cytotoxic activity of more than 75,000 compounds, in what is commonly known as the “NCI60 cell line screen” [1,2]. These cell lines are derived from various tissue types, and represent cancers such as leukemias and melanomas, and solid tumors of the lung, colon, brain, ovary, breast, prostate and kidney. The expression of a large number of genes has been measured in these cell lines using various transcriptional profiling techniques. Since the gene expression is measured before exposure to any of the chemical agents, it can be used to identify biological factors predisposing a cell line to sensitivity or resistance to a particular compound or class of

compounds. In particular, the NCI60 cell lines have been used as a platform for analyzing relationships between the selective activity of potential anticancer agents, and the expression of genes encoding drug targets, drug transporters, and proteins involved in drug activation, cellular stress response, and detoxification [3,4,5,6,7,8,9].

A natural first step toward understanding this dataset is to identify pairwise associations between the expression of individual genes and the activity of individual compounds. Subsequently, higher order relationships involving multiple genes and/or compounds can be pursued [10,11,12]. In this paper we focus on the identification of pairwise associations. Based on our statistical findings, we can consider a sequence of questions of scientific interest:

- *Do numerous distinct genes show associations with the activity of at least one compound, and do numerous distinct compounds show associations with the expression of at least one gene?* The answer to this question tells us in broad terms whether gene expression is an important determinant of chemosensitivity, and whether the influence of gene expression on chemosensitivity is restricted to a few key genes.
- *Are associations between gene expression and compound activity highly specific, in that small groups of genes are uniquely associated with small groups of compounds; or are they more general, in that a single compound is often associated with many genes and a single gene is often associated with many compounds?* The answer to this question is informative about whether the mechanisms by which genes influence compound activity are highly specific, for example through binding or molecular recognition events, or alternatively are broad, as in a particular gene marking a cell state that results in sensitivity or resistance to a diverse class of compounds.
- *Do genes act in a uniform manner, by always increasing or always decreasing chemosensitivity, or do they act differently for different compounds?* The answer to this question tells us whether the consequences of an individual gene's expression are mechanistically simple, in that the effect of the gene is the same for all associated compounds, or are mechanistically complex, in that different compounds are affected in qualitatively different ways.

Our ability to accurately identify associations between gene expression and compound activity, and consequently to address the three questions posed above, is largely limited by the diversity of the cell lines in the screen, and by the measurement accuracy of the experiments. Cell line diversity is important since we will be unable to detect associations with biological states that are not represented in the assayed panel. Thus the associations we identify here represent a subset of all true associations, and in particular some classes of compounds that show no associations in these data may show strong and interesting associations in a more biologically diverse collection of cell lines. Measurement error limits the identification of associations since it diminishes the apparent strength of true associations. In addition, as our goal will be to focus on associations that occur when gene expression and toxicity vary beyond some fixed level, the presence of measurement error makes it more difficult to accurately identify which genes and compounds have the required level of variability in their true, as opposed to measured, expression and activity levels. In this paper we aim to identify associations between gene expression and compound activity, and subsequently to answer as fully as possible the three questions posed above, while accounting for and where possible compensating for the limited measurement accuracy in the data.

Methods

Experimental data

Compound activity data are available from the NCI Developmental Therapeutics Program (DTP) Human Tumor Cell Line Screen (http://dtp.nci.nih.gov/ docs/cancer/cancer data.html) [1]. We used the July 2007 data release. We used data for the 59 cell lines that had more than 1000 GI50 values and that also had microarray gene expression data. Compound activity results are reported as the negative of the log10 GI50, where GI50 is the compound concentration required to inhibit cell growth by 50%. Compounds are tested in a series of dilutions with maximum concentration typically at 10^{-4}M. Microarray gene expression data on the NCI cell line panel from multiple sources are available through the DTP web site (http://dtp.nci.nih.gov/mtargets/download. html). We used 3 gene expression datasets provided by GeneLogic and Novartis. The two GeneLogic data sets are single replications per cell line, one using the Affymetrix U133 microarrays and one using the Affymetrix U95A/B microarray. The Novartis data is the average of three replicate assays per cell line using the Affymetrix U95A microarray. The probe sets in the Novartis data are a subset of those in the GeneLogic U95A/B data, with only the A chip being used. A fraction of the GI50 data points are missing for quality control reasons. There are no missing values in the microarray data.

Filtering genes and compounds

All analysis was done on log scale data (log10 for GI50, log2 for gene expression). Compounds were excluded if they had fewer than 50 experimental GI50 measurements for the 59 cell lines (the threshold was changed from 50 to 45 when the leukemia cell lines were omitted, see below for details). Compounds were also excluded if their log10 GI50 standard deviation was below 0.2. When the standard deviation is 0.2, the probability that a GI50 will deviate by more than two-fold (higher or lower) from the mean value is approximately 1/8. Thus under this selection criterion, at least 1/8 of the (at most 59) GI50 values for a given compound will vary at least two-fold from the mean. We chose the fraction 1/8 since it approximately represents the fraction of the total data from one tissue type. GI50 standard deviations were estimated using the robust biweight midvariance estimator [13]. Expression data for a given probeset was excluded if the standard deviation for log2 data was below 0.65. The interpretation of this threshold is analogous to the threshold 0.2 used for the GI50 data; the numerical threshold differs only due to the change in scale. To enable direct comparison, the standard deviation thresholds 0.2 and 0.65 were also used when the leukemia cell lines were omitted (see below).

Identifying associations

All probe set/compound pairs that passed the filtering step were evaluated for association using the biweight midcovariance [13] to robustly estimate correlation coefficients without excessive loss of statistical efficiency. All variances and covariances were calculated using a robust procedure since in earlier work with these data we found that many of the strongest Pearson correlations between gene expression and compound activity were due to a single cell line's data. While these “outlying” measurements may reflect true values, we wanted to focus here on associations that are reflected in multiple cell lines.

All available data for each compound/gene pair was used to calculate the correlation estimates (data for between 50 and 59 cell lines if all tissue types are included, or for between 45 and 51 cell lines if the leukemia panel is omitted). Estimated correlation values greater in magnitude than thresholds of either 0.5 or 0.55 were retained for subsequent analysis as described below. Due to the coding of GI50 as -log(GI50), a positive association

means that higher gene expression is associated with chemosensitivity, and a negative association means that higher gene expression is associated with chemoresistance.

False discovery rate analysis

The null expected number of associations and false discovery rate (FDR) were obtained by simulating independent standard normal value pairs (corresponding to the GI50 and expression data) and estimating their correlation coefficient using the biweight procedure discussed above. The sample sizes for the simulation analysis (the number of value pairs, corresponding to cell lines with observed data) were frequency-matched to the distribution of sample sizes in the actual data. We simulated 10^7 sets of value pairs to estimate the null expected number of associations. False discovery rates were then estimated using the simple expression N_n/N_o, where N_o is the number of associations identified in the data, and the expected number of false positives N_n is estimated as $N_t p_0$, where N_t is the number of tests and p_0 is the null probability of an association. Data from other distributions, and permuted experimental data were also used, to explore the sensitivity of p_0 to the data distribution. We elected to use this simple approach to FDR analysis, similar to the approach of Benjamini and Hochberg [14] rather than more sophisticated approaches [15,16] since the number of tests is very large and the proportion of positives is very small ($\ll 1\%$).

Tissue type specificity

Many genes are strongly differentially expressed across the tissue types, and many compounds show strong patterns of tissue-specific activity. We aimed to identify associations between gene expression and GI50 that are primarily due to tissue type, as well as those that are not. Therefore, the filtering and association identification steps were carried out three times, as follows. One set of results, denoted A below, was obtained using all cell lines and compounds that pass the filtering step. A second set of results, denoted C below, was obtained after mean-centering gene expression and GI50 values within each tissue type. Since a major component of the tissue-specific effects are due to leukemia cells (and the underlying difference between cells grown on a solid surface and cells grown in suspension), a third set of results, denoted S below, was generated omitting the leukemia cell lines. For centering, the tissue types were assigned according to the updated designation of MDA-MB-435 and NCI/ADR-RES as melanoma and ovarian cancer in origin, respectively [17,18].

Results

Overall statistical significance

We began by assessing the statistical strength of apparent associations in the data. Table 1 summarizes our association findings in each data set, and for each of the inclusion and processing rules A, C, and S defined in the methods. For example, for the A data from the GeneLogic U133 platform with threshold 0.5 (first row of Table 1), the null probability of an association was $p_0 = 4.4 \times 10^{-5}$, so the expected number of false associations is $26391 \times 11196 \times 4.4 \times 10^{-5} = 13001$. Since we observed 308165 associations, the FDR is estimated as $13001/308165 \approx 0.04$.

To assess the sensitivity of these results to the choice of a normal distribution for determining the null probability of an association p_0, the analysis was repeated with other forms of null data. First, several parametric distributions were used, including Bernoulli trials with equal probabilities (0.5/0.5) and unequal probabilities (0.75/0.25), various finite mixtures of normals, a continuous uniform distribution, and standard exponential and Cauchy distributions. Most of these distributions gave fewer positives calls than in the normal case, suggesting that the use of a normal reference distribution is conservative if it is

biased at all. Only a strongly skewed distribution (exponential) and a heavy-tailed distribution (Cauchy) gave more positive calls than in the normal case, by around 25% and 50%, respectively. However consideration of skew coefficients and tail indices in the data do not suggest that either the gene expression or GI50 data are often as skewed as an exponential distribution or as heavy-tailed as a Cauchy distribution. We also used a simple permutation approach, in which genes and compounds were randomly selected, and the GI50 values for the compound were randomly permuted prior to calculating the correlation coefficient. These results were within 10% of the value for the normal case (on the low side), which is not a statistically significant difference with respect to Monte-Carlo error, based on 10^7 permutations as we performed.

The number of positives for the A data tends to be an order of magnitude or more greater than the number for the C data, whereas the number of positives drops by around half when comparing the C to the S data. This suggests that a large fraction of associations are at least partially related to tissue-specific cell growth inhibition, and that not all of these effects are explained by the differences between solid tumor-derived and leukemic cell lines. For all three datasets, the number of positive calls as a function of the correlation threshold is steeply decreasing in the range of thresholds considered – there is around an order of magnitude drop in the number of positives as the threshold is raised from 0.5 to 0.55, even though scatterplots in this range of correlation values tend to be visually indistinguishable.

We did a limited analysis of the structures of the compounds involved in associations to assess whether a restricted class of chemical structures gives rise to a large fraction of the associations. Out of the 11196 compounds passing the filtering step, 8605 had CACTVS fingerprints available from PubChem. Using these fingerprints, we calculated Tanimoto similarity measures between all compound pairs within the 5879 compounds involved in associations, and between all compound pairs within the 2726 compounds not involved in associations. The mean and standard deviation of Tanimoto coefficients in these two groups, respectively, are 0.56(0.24) and 0.42(0.15). Thus the compounds involved in associations are somewhat less diverse than the set of compounds that are not involved in associations, suggesting that some common structural families are contained within the set of compounds whose activity is influenced by gene expression. For example, P-glycoprotein substrates are a structurally restricted class of compounds that are selective against cell types lacking P-glycoprotein mediated drug efflux activity, which is reflected in expression of the ABCB1 gene.

The results in Table 1 indicate that for the A data, associations between GI50 and gene expression can be identified with high accuracy, whereas for the C data there is insufficient power to identify any true associations that might exist. For the A data, associations appear to be reliably identifiable at a correlation threshold of 0.5. In the S data, there is strong evidence that numerous associations exist, but at the 0.5 correlation threshold the number of false positives will be high. For the remainder of the analysis we will use the 0.5 threshold for the A data and the 0.55 threshold for the S data. The Novartis data has the lowest number of genes passing the filtering step among the three platforms, and also consistently has the lowest FDR value. Since the Novartis data is the average of three experiments, it presumably has the least measurement error. This suggests that a number of spurious associations in the GeneLogic datasets are due to measurement error. Nevertheless, even the GeneLogic datasets have favorable FDR values, so we will continue to use all three datasets in the analysis.

To further assess the significance of these results, we hypothesized a mean/variance relationship for the null distribution of the number of gene/compound pairs associated at a given threshold level. As a working relationship we took variance = $f \cdot$mean, for some $f > 1$.

If there were no dependence within the genes or within the compounds, and if each gene and each compound were tested only once (rather than in all gene/compound pairs), the null distribution would be Poisson and we would have $f = 1$. Since the C data showed no evidence of associations, we used these data to estimate f. Specifically, we used simple linear regression to regress $(Y - X)^2$ against X, where Y is the observed number of associations and X is the null mean. The regression fit over the 6 points for the C data (as given in rows 7-12 of Table 1) had an adjusted R^2 of 0.94, and an intercept that was not significantly different from zero ($p > 0.1$), consistent with the hypothesized mean/variance relationship. The estimated slope was highly significant ($p < 10^{-3}$), and gave an estimate $\sqrt{f}=90$ which we used as an inflation factor for standard deviations (i.e. estimating the standard deviation as $90 \cdot \sqrt{\text{mean}}$). In this way we calculated the 98$^{\text{th}}$ percentile of the FDR distribution for the A and S datasets with each threshold as the mean plus two standard deviations. For the A data with correlation threshold 0.5, these values are below 0.15, and for the S data with correlation threshold 0.55 these values are below 0.41. From this we conclude that associations identified from the A and S data are unlikely to be dominated by false positives.

Positive controls

As a positive control we considered the relationship between triciribine (TCN, NSC154020) or triciribine phosphate (TCN-P, NSC280594) and adenosine kinase (ADK). Expression of ADK is necessary for the growth inhibitory activity of TCN and TCN-P, as it participates in the phosphorylation/ dephosphorylation of TCN that converts it between its toxic and cell-permeable forms [19,20]. The U95 array and U133 array each contain two probesets for ADK. On the U133 array, one TCN/ADK pairing (NSC154020, 204119 s at) had a correlation coefficient of 0.60, and another (NSC280594, 204119_s_at) had a correlation of 0.50. On the GeneLogic U95 array, the pairing (NSC154020, 168_at) had a correlation of 0.54. These signed correlations are in the expected direction. The correlations given are for the A data, and are equal or slightly stronger in the S data. In the Novartis U95 data set, the two probesets for ADK did not meet our filtering criteria.

Proportion of measured genes associated with chemosensitivity or chemoresistance

The percentages of probe sets that passed the filtering steps and that are associated with at least one compound are shown in Table 2. Given the low false discovery rates for the A and S data, this suggests that a substantial fraction of all genes that vary in expression in the NCI cell line panel are truly associated with the activities of one or more compounds. Indeed, if the false positives are distributed independently of the genes, the proportion of genes associated with at least one compound drops by only 1-2% for the A data and by 4-5% for the S data. These percentage are obtained by considering the number of compounds associated to a particular gene, then calculating the probability that all of the associations are false positives based on the FDR values given in Table 1. Specifically, if the FDR is f and a given gene is associated to m compounds, the probability that the gene has at least one true association is $1 - f^m$. The values 1-2% and 4-5% cited above are expected values calculated by summing the probabilities $1 - f^m$ over all genes. The reduction is smaller than the FDR since most genes are associated with multiple compounds.

For both the A and S data, the Novartis dataset with its presumed lower level of measurement error had the fewest probe sets passing the filtering, but the greatest proportion of non-filtered probe sets involved in associations. The observation that fewer probesets pass the filtering in the Novartis set is likely due to the unreplicated GeneLogic datasets overcalling genes with expression variation. If true associations are less common among the less variable genes, the proportion of genes with at least one association will be lower in the

unreplicated datasets. The observation that the Novartis data has the greatest proportion of non-filtered probe-sets involved in associations can also be explained in terms of measurement error. If the false positives are distributed randomly across the genes and compounds, and given that associations are rare given the number of tests performed, false positives will tend to introduce genes with associations to only one compound. This is discussed further below. Notably, when using the Novartis data, the fraction of substantially varying genes that are associated with GI50 for at least one compound exceeds 1/2.

Specificity of associations

Figure 1a shows the distribution of the number of compounds associated with each non-filtered probe set (restricted to probe sets associated with at least one compound). The results shown in the figure are for the A data, from each of the three array platforms. Probe sets associated with many compounds are likely to represent one or more mechanisms of selective activity or resistance. All three datasets have a small number of probe sets associated with more than 100 compounds, with a few genes being associated with more than 500 compounds. The Novartis dataset has the greatest fraction of probe sets associated with large numbers of compounds. Conversely, the two unreplicated GeneLogic sets have a greater fractions of genes associated with only one or two compounds.

As noted above, measurement error in the gene expression and GI50 measurements is likely to lead to isolated associations between genes and compounds, rather than producing genes with associations to numerous compounds. Consistent with this, the Novartis data has the lowest proportion of genes with small numbers of associated compounds.

Although the Novartis data presumably has the lowest measurement error among the expression datasets considered here (since it is an average of replicates), its measurement error is not zero, and there is also measurement error in the GI50 data. Therefore we sought to understand what the distribution of the number of associated compounds per gene would look like in the complete absence of experimental measurement error in both the gene expression and GI50 data. Using the original data from which the Novartis averages were obtained, the measurement variance in the averaged Novartis data can be estimated as 0.08=0.24/3, where 0.24 is the measurement variance within replicates averaged over the genes. Unlike the gene expression data, the GI50 data are not systematically replicated. However there are around 10^5 compound/cell line pairs for which at least two experimental points are available. From these, we estimated the GI50 measurement error variance as 0.45, averaged over the compounds.

Given nominal levels for the gene expression and GI50 measurement variance, it becomes possible to employ a procedure analogous to the SIMEX procedure [21], which is used to adjust coefficient estimates in regression models for the effects of measurement errors. We added simulated centered normal errors with variances 0.08 and 0.45 to the Novartis gene expression and the GI50 data, respectively (following the A processing). Then we recalculated all associations from these data. If f is the estimated density for the experimental data, and g is the estimated density after adding additional simulated errors of a similar magnitude to the actual measurement errors, then $(2f - g)$ (x02228)0 can be used to estimate the density that would be found if no measurement errors were present. This estimate is justified under an assumption that for the magnitudes of measurement errors under consideration, the effect of measurement error on the density is approximately linear in a pointwise sense.

Figure 1b shows the results of this analysis. Only the Novartis dataset is analyzed in this way since it is the only dataset where we have replication allowing measurement error variances to be directly estimated. The estimated distribution for the number of compounds

associated to a gene when no measurement error is present is markedly different from the direct estimate shown in Figure 1a. The direct estimate has a mode at zero, whereas the adjusted estimate has a mode at around 1.4 (on the log10 scale), and has very little mass at zero. Thus while an analysis ignoring measurement error suggests that the associations are dominated by isolated gene/compound pairs, consideration of measurement error suggests that very few such isolated pairs actually exist, and that a typical gene that is associated to at least one compound is associated to around $10^{1.4} \approx 25$ compounds.

Balance between genes conferring chemosensitivity and chemoresistance

Positive and negative directions of association between gene expression and -log(GI50) are indicative of different mechanisms of action. Genes associated with upregulation of detoxification mechanisms such as drug efflux and catalytic deactivation should lead to negative associations, but in general both directions are possible. For the A, C, and S datasets, the genes associated with -log(GI50) at either a 0.5 or 0.55 threshold are nearly perfectly balanced between positive and negative associations. For genes that are associated with multiple compounds, the directions of associations can provide us with an indication of whether the gene acts with a uniform mechanism for all the compounds. If the gene acts via the same mechanism for all compounds, we expect the direction of association to be the same (either positive or negative) for all associated compounds.

We assessed whether there is evidence that individual genes exhibit both positive and negative associations with different sets of compounds. To examine this, we identified the probe sets that are associated with at least 20 compounds in a single direction in the A dataset at threshold 0.5. There are 2466, 1134, and 2397 probe sets satisfying this condition from the GeneLogic U133, Novartis U95A, and GeneLogic U95A/B arrays, respectively. The overlap between the Novaratis and GeneLogic U95A arrays is 541 probe sets, or roughly half the list from the Novartis data. The average proportion of associations within each of these probe sets that were oriented in the less frequent direction for that probeset was 0.12, 0.12, and 0.11 in the three data sets. In other words, almost 90% of the associations are in the same direction, on average.

We next aimed to determine whether the observed level of balance between positive associations and negative associations within a gene was consistent with the true associations for the gene being either entirely positive or entirely negative. We did this by correlating the actual GI50 data to simulated iid standard normal data as a proxy for gene expression when no association is present. This provided us with estimates of the mean and variance of the number of false positives when testing a single gene against the 11196 compounds in set A. The variance in the number of positives (based on threshold 0.5) was 22 times the mean number of positives. This substantial overdispersion relative to the Poisson distribution is due to the clustered nature of the compound set, which largely consists of small sets of related compounds with similar GI50 profiles (note that the overdispersion here is much less than found above when testing all genes against all compounds). From the simulation described above, the probability of a false association with either a positive or negative sign for a particular gene and compound is $p_0 = 4.4 \cdot 10^{-5}$, so the expected number of false associations with a particular sign for a given gene in the A data tested against all compounds is $11196 \cdot 2.2 \cdot 10^{-5} = 0.25$. Thus the upper 99.9^{th} percentile for the number of false associations with a particular sign is approximately $0.25 + 3 \cdot \sqrt{0.25 \cdot 22} = 7.3$.

Based on this analysis, genes with 7 or fewer compounds associated in a negative direction can be presumed to have only positive true associations, and genes with 7 or fewer compounds associated in a positive direction can be presumed to have only negative true

associations. Table 3 shows the results of this analysis. Among genes in the A data having 20 or more associations in at least one direction, the fraction having 7 or fewer associations in the less common direction are 0.64, 0.60, 0.68, in the GeneLogic U133, Novartis U95A, and GeneLogic U95AB datasets, respectively (calculated as $(N_{++} + N_{--})/N$ in Table 3). We conclude that the experimental data for these genes are consistent with a uniform mechanism of action for all associated compounds, although uniformity in the direction of association is not sufficient to conclude this. Conversely, we can infer that the 30%-40% of genes that are associated with multiple compounds in differing directions are unlikely to do so with a uniform mechanism of action.

Genes associated with many compounds

Some of the genes associated with GI50 for large number of compounds are involved in cellular processes identified as being related to small molecule toxicity in previous studies using the NCI60 cell line screening data. For example, in the S dataset only three probesets are associated with the GI50 of more than 300 compounds. Two of these three probesets are for the cysteine/glutamate transporter SLC7A11, previously identified as being involved in chemoresistance [22,3]. In the A dataset, five separate probesets for ABCC3 are each positively associated with chemosensitivity for more than 250 compounds. ABCC3 is well known as playing a major role in multidrug resistance [23].

We analyzed the set of genes having 100 or more associations in at least one of the datasets using the DAVID/EASE functional classification tool [24]. For genes whose expression was associated with chemosensitivity, eight functional clusters were identified: these include clusters of various actins, integrins, and collagens. The clearest theme in the eight clusters is that they contain numerous genes related to maintenance of the extracellular membrane, and more generally to a differentiated cell state. For genes whose expression was associated with chemoresistance, six clusters were identified. These included clusters centered on growth factors, transcription factors, and genes associated with cell signaling and proliferation.

For detailed manual analysis, we focused on the genes whose expression is associated with 200 or more compounds in the S data. A total of 146 probesets on the three arrays satisfy this condition. Among the genes represented by these probesets there is a consistent trend of signal transduction genes being upregulated in cells with a resistant phenotype, and extracellular matrix and cytoskeletal genes being upregulated in cells with a sensitive phenotype. The subset of these genes associated with 200 or more compounds in at least two of the three S data sets is shown in Table 4.

Signal transduction genes whose expression is associated with chemoresistance include the epidermal growth factor receptor (EGFR), erythroblastic leukemia viral oncogene homolog 3 (ERBB3), cysteine-rich transmembrane regulator (CRIM1), fasciculation and elongation protein zeta 2 (FEZ2), tumor necrosis family receptor superfamily member 12A (TNFRSF12A), ubiquitin-associated SH3 domain containing protein A (CLIP4), and hypoxia inducible factor alpha (HIF1A). Interestingly, CLIP4 is a regulator of EGFR signaling, that acts by inhibiting internalization of EGFR into endocytic vesicles and therefore blocking its degradation. Also, the FEZ2 gene is a regulator of protein kinase C, which is downstream of the EGFR. CRIM1 has been implicated in growth factor secretion and HIF1A is an important signal transduction molecule that triggers an angiogenic response and induces drug resistance when it is upregulated in cancer cells in response to hypoxic conditions. Therefore, these genes are potentially linked to a common pathway involved in autocrine or paracrine growth factor signaling mechanisms associated with angiogenic phenotypes, which are upregulated in aggressive cancer cells.

As noted above, expression of extracellular matrix and cell adhesion genes tend to be associated with drug sensitivity. These include collagen 7A1 (COL7A1), one of the main extracellular matrix components of the basement membrane; dentin matrix protein (FAM20C) a component of the tooth extracellular matrix; versican (VCAN), which is chondroitin sulfate proteoglycan that binds hyaluronic acid and is a component of cartilage; fibrillin 1 (FBN1) which is the major component of extracellular matrix microfibrils; myosin light chain kinase (MYLK), which is associated with regulating contractility and cell shape; and palladin (PALLD), which is a component of actin stress fibers and has a role in cell adhesion. In addition to these genes, the integrins ITGAV and ITGA5 which are the vitronectin and fibronectin receptors are associated with drug sensitivity, although overexpression of another integrin ITGA3 is more associated with resistance. Extracellular matrix and adhesion genes are likely to be associated with more differentiated cancer cells with anchorage-dependent growth regulation.

Discussion

We have identified several attributes of the NCI/DTP screening data that are not apparent from simple tabulation of pairwise associations. First, and least surprisingly, we found that if no attempt is made to remove associations that may be due to tissue type differences, a large number of associations can be identified with only a modest rate of false positives. Since many associations that depend on tissue type variation may reflect confounding effects of tissue-specific gene expression, it is also desirable to identify associations that are not completely due to differences between the tissues. Our results suggest that attempts to minimize tissue-specific effects should be considered in the context of statistical power. Mean centering each cell type in the panel is the most direct way to completely remove tissue specific effects, but it appears to remove so much of the association signal that little power remains. It is likely that mean-centering the tissues over-corrects for tissue-specific expression, and moreover, not all tissue specific expression is uninformative about mechanisms of activity. As a less drastic strategy, focusing on solid tumors by removing leukemic cell lines enriches for relationships that are less tissue-specific, at a modest cost in power.

We found that numerous (thousands) of genes are involved in associations with one or more compounds, and that that these genes comprise a substantial fraction (more than 1/2) of all genes varying beyond a fixed threshold. One reason for focusing on the more variable genes is that associations involving such genes are presumed to be more likely to be causal rather than responsive, although this is certainly not universally true. A second reason for focusing on more variable genes is that there is greater statistical power for identifying the associations in which they are involved.

The effects of measurement error should be considered when assessing the number of genes involved in associations. Comparing our findings with replicated and unreplicated data, it appears that the unreplicated datasets substantially overcall the set of genes with a given threshold level of variability. Moreover, the genes introduced by this inflation are likely to introduce a greater fraction of false associations compared to the genes that are truly variable. The fact that our positive control gene ADK did not vary sufficiently in the Novartis data to be tested for associations suggests that the variance thresholds we chose were too high to capture all mechanistic relationships. However the false positive rate presumably increases as the threshold is lowered. Thus the situation with ADK may best be viewed as reflecting the limits of power of this dataset. Also of note is that previous work involving the association between ADK and TCN [7] began with a subset of the NCI screening compounds consisting of well-characterized compounds. Thus the multiple testing

issue in that context was of a much lower order, and the success of that analysis does not imply that such associations should be detectable in the entire dataset.

A number of the previously published findings from this dataset have focused on associations between single genes and single compounds, which are presumably indicative of chemically specific relationships, perhaps mediated through molecular recognition or binding. Other reported associations have been more general, including the identification of efflux and detoxification systems for which numerous compounds can serve as substrates. Our findings suggest that genes with the capacity to be associated with small molecule toxicity are likely to be associated with multiple compounds, whereas genes associated with a single compound are rare. The DTP screening set is diverse, but it does contain clusters of related structures, particularly around approved drugs. However, we estimated the mean number of associated compounds per gene as 30, and clusters of 30 or more closely related compounds are uncommon. Therefore it is unlikely that the gene/drug association clusters we have shown to exist are exclusively due to clustering of structural analogues in the compound set. To further refine our understanding of the generality or specificity of gene/compound associations in this dataset, subsequent analyses should move beyond the number of compounds associated with a gene, and explore the the actual diversity of chemical structures in more detail.

The balance between associations in the positive and negative directions within a gene can be informative about the uniformity of the mechanism underlying the associations. If substantial numbers of associations occur in both directions, multiple mechanisms are likely present, whereas predominance of one direction is consistent with a uniform mechanism. Our results suggest that more than half of genes may act with a uniform mechanism.

Finally, we performed systematic and manual analyses of genes associated with large numbers of compounds. These genes are presumed to be involved with general, less specific mechanisms of action. In contrast, genes associated with only a few compounds are more likely to engage in chemically specific interactions [9,7]. The genes associated with large numbers of compounds include transporters already identified as involved in drug efflux and cell detoxification [3,22]. In addition we observed that expression of a set of genes relating to EGFR signaling is related to chemoresistance, whereas genes relating to the extracellular membrane and cell adhesion are associated with chemosensitivity. The latter association may be reflect a general tendency for the less differentiated cell lines, presumably derived from more advanced cancers, to be broadly less sensitive to cytotoxic agents.

In terms of general methodology for high throughput screening and biomolecular assay data, we have demonstrated that the use of false discovery rates together with straightforward techniques to account for the effects of measurement error can yield insights about how associations among variables are distributed in a large and complex dataset. Our approach utilizes a simple formulation of FDR that was calibrated using simulations due to the lack of an analytic variance formula for the robust association estimators we used. To accommodate the clustered nature of the gene expression and compound activity data, over-dispersion factors were calculated from the data. We note that this part of the analysis could alternatively have been carried out with a permutation approach. For measurement error analysis, we used direct estimates of the measurement error variance based on replication in the gene expression and compound activity data sets. We then followed a simulation procedure much like the SIMEX approach to determine the distribution of associations that is expected in a data set that is free of measurement error. We found this to be a straightforward and effective means to clarify the overall pattern of gene/compound associations, and propose that it could be meaningfully applied in a variety of other data analytic settings involving large sets of data measured with error.

Our findings indicate that the NCI60 cell line screen is most effective at identifying associations involving multiple genes or compounds, since that provides a form of internal replication. Accordingly, our results point to the types of associations that are most likely to be detectable in this dataset. Since it appears that genes associated with compound activity are often associated with multiple, distinct compounds, power for future studies will be maximized by focusing on that subset of the data.

Acknowledgments

The authors gratefully acknowledge NIH support to Shedden and Rosania (grant 3P20HG003890).

References

1. Shoemaker, RobertH. The nci60 human tumour cell line anticancer drug screen. Nat Rev Cancer Oct;2006 6(10):813–823. [PubMed: 16990858]
2. Alley MC, Scudiero DA, Monks A, Hursey ML, Czerwinski MJ, Fine DL, Abbott BJ, Mayo JG, Shoemaker RH, Boyd MR. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. Cancer Res Feb;1988 48(3):589–601. [PubMed: 3335022]
3. Dai, Zunyan; Huang, Ying; Sadee, Wolfgang; Blower, Paul. Chemoinformatics analysis identifies cytotoxic compounds susceptible to chemoresistance mediated by glutathione and cystine/glutamate transport system xc- J Med Chem Apr;2007 50(8):1896–1906. [PubMed: 17367118]
4. Marx, KennethA; O'Neil, Philip; Hoffman, Patrick; Ujwal, ML. Data mining the nci cancer cell line compound gi(50) values: identifying quinone subtypes effective against melanoma and leukemia cell classes. J Chem Inf Comput Sci 2003;43(5):1652–1667. [PubMed: 14502500]
5. Ring, BrianZ; Chang, Stella; Ring, LWinston; Seitz, RobertS; Ross, DouglasT. Gene expression patterns within cell lines are predictive of chemosensitivity. BMC Genomics 2008;9(74)
6. Covell, DavidG; Wallqvist, Anders; Huang, Ruili; Thanki, Narmada; Rabow, AlfredA; Lu, XiangJun. Linking tumor cell cytotoxicity to mechanism of drug action: an integrated analysis of gene expression, small-molecule screening and structural databases. Proteins May;2005 59(3):403–433. [PubMed: 15778971]
7. Shedden, Kerby; Townsend, LeroyB; Drach, JohnC; Rosania, GustavoR. A rational approach to personalized anticancer therapy: chemoinformatic analysis reveals mechanistic gene-drug associations. Pharm Res Jun;2003 20(6):843–847. [PubMed: 12817886]
8. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR. Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A Sep;2001 98(19):10787–10792. [PubMed: 11553813]
9. Wei, Guo; Twomey, David; Lamb, Justin; Schlis, Krysta; Agarwal, Jyoti; Stam, RonaldW; Opferman, JosephT; Sallan, StephenE; den Boer, MoniqueL; Pieters, Rob; Golub, ToddR; Armstrong, ScottA. Gene expression-based chemical genomics identifies rapamycin as a modulator of mcl1 and glucocorticoid resistance. Cancer Cell Oct;2006 10(4):331–342. [PubMed: 17010674]
10. Li KC, Yuan S. A functional genomic study on nci's anticancer drug screen. Pharmacogenomics J 2004;4(2):127–135. [PubMed: 14993929]
11. Huang R, Wallqvist A, Thanki N, Covell DG. Linking pathway gene expressions to the growth inhibition response from the national cancer institute's anticancer screen and drug mechanism of action. Pharmacogenomics J 2005;5(6):381–399. [PubMed: 16103895]
12. Lee, AdamC; Shedden, Kerby; Rosania, GustavoR; Crippen, GordonM. Data mining the nci60 to predict generalized cytotoxicity. J Chem Inf Model Jul;2008 48(7):1379–1388. [PubMed: 18588283]
13. Hoaglin, DC.; Mosteller, F.; Tukey, JW. Understanding robust and exploratory data analysis. New York: Wiley; 1983.
14. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society series B-Methodological 1995;57:289–300.

15. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical bayes analysis of a microarray experiment. Journal of the American Statistical Association 2001;96:1151–1160.

16. Storey J. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B 2002;64:479–498.

17. Ellison G, Klinowska T, Westwood RFR, Docter E, French T, Fox JC. Further evidence to support the melanocytic origin of mda-mb-435. Mol Pathol Oct;2002 55(5):294–299. [PubMed: 12354931]

18. Garraway, LeviA; Widlund, HansR; Rubin, MarkA; Getz, Gad; Berger, AaronJ; Ramaswamy, Sridhar; Beroukhim, Rameen; Milner, DannyA; Granter, ScottR; Du, Jinyan; Lee, Charles; Wagner, StephanN; Li, Cheng; Golub, ToddR; Rimm, DavidL; Meyerson, MatthewL; Fisher, DavidE; Sellers, WilliamR. Integrative genomic analyses identify mitf as a lineage survival oncogene amplified in malignant melanoma. Nature Jul;2005 436(7047):117–122. [PubMed: 16001072]

19. Porcari AR, Ptak RG, Borysko KZ, Breitenbach JM, Vittori S, Wotring LL, Drach JC, Townsend LB. Deoxy sugar analogues of triciribine: correlation of antiviral and antiproliferative activity with intracellular phosphorylation. J Med Chem Jun;2000 43(12):2438–2448. [PubMed: 10882371]

20. Ptak RG, Borysko KZ, Porcari AR, Buthod JL, Holland LE, Shipman C, Townsend LB, Drach JC. Phosphorylation of triciribine is necessary for activity against hiv type 1. AIDS Res Hum Retroviruses Oct;1998 14(15):1315–1322. [PubMed: 9788672]

21. Cook J, Stefanski LA. A simulation extrapolation method for parametric measurement error models. Journal of the American Statistical Association 1995;89:1314–1328.

22. Liu, Ruqing; Blower, PaulE; Pham, AnhNhan; Fang, Jialong; Dai, Zunyan; Wise, Carolyn; Green, Bridgette; Teitel, CandeeH; Ning, Baitang; Ling, Wenhua; Lyn-Cook, BeverlyD; Kadlubar, FredF; Sade, Wolfgang; Huang, Ying. Cystine-glutamate transporter slc7a11 mediates resistance to geldanamycin but not to 17-(allylamino)-17-demethoxygeldanamycin. Mol Pharmacol Dec;2007 72(6):1637–1646. [PubMed: 17875604]

23. Borst P, Evers R, Kool M, Wijnholds J. A family of drug transporters: the multidrug resistance-associated proteins. J Natl Cancer Inst Aug;2000 92(16):1295–1302. [PubMed: 10944550]

24. Dennis, Glynn; Sherman, BradT; Hosack, DouglasA; Yang, Jun; Gao, Wei; Lane, HClifford; Lempicki, RichardA. David: Database for annotation, visualization, and integrated discovery. Genome Biol 2003;4(5):3.

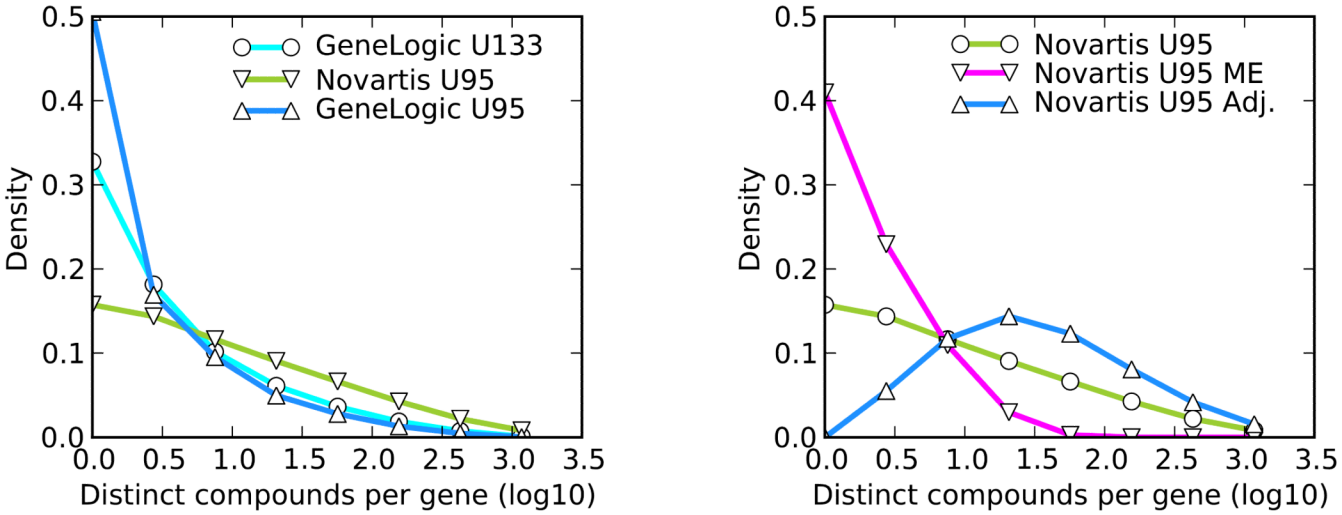

Figure 1.

a: Estimated distribution of the number of associated compounds per gene for each of the three experimental data sets. b: The experimental distribution of associated genes in the Novartis set, along with the number of associations observed when extra measurement error is added (Novartis ME), and the adjusted distribution estimating what would be found in the absence of measurement error (Novartis Adj.).

Table 1

Summary of association results compared to null expected findings. The primary data sets are coded in column 1 as 1: GeneLogic U133, 2: Novartis U95A, 3: GeneLogic U95A/B. Three data filtering and processing schemes are indicated in column 3: all data were used without centering (A), all data were used following centering within tissue types (C), or only data from solid tumor derived cell lines were used without tissue type centering (S).

Data set	Threshold	Processing	#Probe Sets	#Compounds	#Assoc.	Null exp. #Assoc.	FDR(98%)
1	0.5	A	26391	11196	308165	13001	0.04(0.11)
2	0.5	A	5144	11196	145365	2534	0.02(0.08)
3	0.5	A	35196	11196	296205	17338	0.06(0.14)
1	0.55	A	26391	11196	102098	1861	0.02(0.09)
2	0.55	A	5144	11196	50018	363	0.01(0.08)
3	0.55	A	35196	11196	92515	2483	0.03(0.12)
1	0.5	C	4601	11117	11601	13810	0.84
2	0.5	C	988	11117	3692	2966	0.80
3	0.5	C	20488	11117	43353	61497	1.00
1	0.55	C	4601	11117	2490	3375	1.00
2	0.55	C	988	11117	850	725	0.85
3	0.55	C	20488	11117	9128	15032	1.00
1	0.5	S	25657	9484	212117	58399	0.28(0.48)
2	0.5	S	4833	9484	82899	11001	0.13(0.36)
3	0.5	S	34883	9484	232279	86016	0.37(0.60)
1	0.55	S	25657	9484	66538	7300	0.11(0.34)
2	0.55	S	4833	9484	28239	1375	0.05(0.29)
3	0.55	S	34883	9484	68754	9924	0.14(0.41)

Table 2

Percentages of genes associated with chemosensitivity or resistance for at least one compound. The data sets (1, 2, 3) and data filtering/processing schemes (A, C, S) are coded as in table 1. Results are shown for correlation thresholds 0.5 and 0.55.

	0.5			0.55		
	1	**2**	**3**	**1**	**2**	**3**
A	64	79	60	33	53	28
C	66	69	63	26	30	24
S	80	89	79	46	62	43

Table 3

Summary of results on uniformity of associations. N_+: genes with at least 20 positive associations; N_-: genes with at least 20 negative associations; N: genes with at least 20 positive or at least 20 negative associations; N_{++}: genes with at least 20 positive and at most 7 negative associations; N_{--}: genes with at least 20 negative and at most 7 positive associations.

Data	N_+	N_-	N	N_{++}	N_{--}
1	1195	1776	2466	535	1033
2	608	777	1134	265	416
3	1220	1600	2397	657	970

Table 4

Genes associated with at least 200 compounds in at least two of the three data sets. Data sets 1, 2, 3, are denoted along the top margin as in Tables 1-3. For each dataset, three numbers are given: the number of distinct probesets showing an association, the percentage of associations that are positive, and the total number of associations. The percentage and total are calculated over all compounds associated to any probe set for the gene.

	1			2			3		
CRIM1	2	20	474	1	4	212	1	19	284
MYLK	1	76	268	1	86	239	2	67	558
FHL1	1	65	232	1	74	200	1	61	224
MIPEP	1	73	276	1	67	248	1	37	281
LOXL2	1	59	240	1	76	285	1	58	208
TRPC1	1	86	260	2	80	472	1	82	232
CYR61	2	49	545	1	67	231	1	51	269
NMT2	1	31	220	1	11	202	1	29	217
ANLN	1	32	205				1	37	258
GLRB				1	85	259	1	81	265
PDGFC	2	78	482				1	69	251
CLIP4	1	1	208				1	0	219
COL4A1	2	74	497				1	63	252
C6ORF192	1	26	208				1	31	233
PEA15				1	63	233	1	30	210
THBS1	3	63	755				2	75	492
PALLD	2	82	457				1	87	230
BCAT1	1	60	281				2	51	479
JUB	1	17	232				1	20	290
MYH9	1	19	278				1	12	312
IL1R1				1	65	224	1	68	219
POLR1D	1	90	205				1	85	249
EGFR	2	5	445				2	8	461
TNFRSF12A	1	19	277				1	17	200
NUAK1	1	72	264				1	87	210