DrugMatrix®

Calculations White Paper

March 2011

NIEHS/NTP

111 TW Alexander Drive

Durham, NC 27709

http://ntp.niehs.nih.gov/

http://www.niehs.nih.gov/

Table of Contents

Introduction **5**

Chapter 1. Compound Curation **6**
1.1. Structure Activity Class 6
1.2. Activity Class 6
1.3. Indication 6
1.4. Therapeutic Class 6
1.5. Product Class 7
1.6. Mechanism 7
1.7. Mode Class 7
1.8. Development Status 7
1.9. Adverse Drug Reaction Classification 7
1.10. Known Toxicity 8
1.11. Tissue Toxicity 8

Chapter 2. DrugMatrixScreen™ Pharmacology Assay Panel **9**
2.1. Molecular Pharmacology Assay Panel 9
2.2. Assay Selection Criteria 10
2.3. Assay Protocol 10
2.4. Overview of Assay Results 11

Chapter 3. In-life Data Domain of DrugMatrix **15**
3.1. Introduction 15
3.2. Blood Chemistry Data 15
3.2.1. Introduction 15
3.2.2. Log Ratio and Significance Calculation 15
3.2.3. Normal Range of Historical Controls 15
3.2.4. Denominator Calculations 15
3.3. Histopathology Data 15
3.3.1. Ridit Scores for Histopathology Data 16
3.3.2. Calculating Ridit Scores 16
References 17

Chapter 4. Gene Expression Domain **18**
4.1. Introduction 18
4.2. Probe or Probe-set, Expresson and Gene Definitions 18
4.3. Gene Annotation 19
4.3.1. Gene Names: 20
4.3.2. GO Annotation: 22
4.3.3. Synonyms and Details: 22
4.3.4. Update: 23
4.3.5. Sources of Information: 23
4.4. Expression Data Processing 23
4.4.1. Data Pre-processing 23
4.4.2. CodeLink RU1 Array 23
4.4.2.1. Bad Probe Selection 23

4.4.2.2. Data Exclusion 25
4.4.2.3. Handling of No Data Spots 25
4.4.3. Affymetrix Array Quality Control Assessment 26
4.4.4. Normalization 27
4.5. Ratio Calculations and Significance Determination 29
4.5.1. Formalizing the Problem 29
4.5.2. Control Set Selection 29
4.5.3. Log Ratios 35
4.5.4. Significance of Results 35
4.5.5. Empirical Bayes Variance Estimates 36
4.6. Normalizing Log Ratios Across Platforms 38
4.7. Filtering Log Ratios 38

Chapter 5. Signature Technology 39
5.1. Sparse Linear Programming (SPLP) Algorithm 39
5.3. Signature Mining 39
5.1.1. Class Size Requirements: 40
5.1.2. Performance Requirements: 40
5.1.3. Scalar Product Score Calculations: 41
5.1.4. Posterior Probability Score Calculation: 41
5.3. T-Rank Signatures 42
5.4. Signature Projection Score (SPS) 43
5.5. Cross Platform Signatures 46
5.6. Motif 46

Chapter 6. DrugMatrix Analytical Tools 48
6.1. Compound/Gene Similarity Comparison 48
6.1.1. Similar Genes and Compounds Calculation 48
6.1.2. Chemical Structure Comparisons 48
6.2. Significant Gene Finder 51
6.3. Most Variable Probe Set 51
6.4. Pathway Analysis 52
6.4.1. Pathway Curation 52
6.4.2. Pathway Visualization 52
6.4.3. Pathway Impact Analysis 53
6.5. GO Data Query 54
6.6. Pattern Analysis 55

Figure Legend 55

Table of Figures

Figure 1: Frequency histogram of 107,478 percent inhibition measurements from 861 compounds tested in up to 130 assays 11
Figure 2: Frequency histogram of 3,845 IC_{50} (log-μM) measurements from 506 compounds tested in up to 130 assays 12
Figure 3: Frequency histogram showing the fraction of 123 assays inhibited by 156 randomly chosen compounds. 13
Figure 4: Frequency histogram showing the fraction of 156 compounds that inhibit each of 123 assays 14
Figure 5: Model for genomics data in DrugMatrix 19
Figure 6: Strategy for derivation of gene names. (DBSS: Determined by sequence similarity) 21
Figure 7: Procedure for GO annotation 22
Figure 8: Examples of bad probes 24
Figure 9: The density distribution of data for a CodeLink™ array. 28
Figure 10: Pre- and post-curve fitting of array data to a reference template. 28
Figure 11: Principal components analysis of Log gene expression signal for control and drug-treated arrays. 31
Figure 12: Process drift when controls are not matched to experiments 32
Figure 13: Protocol version specific denominators perform well on proper window, but poorly on wrong window 33
Figure 14: The signature projection score 45
Figure 15: Distribution of Tanimoto scores 50
Figure 16: Distribution of 3DMI scores 51
Figure 17: Example of pathway visualization 53

Table of Tables

Table 1: Example assays from the DrugMatrixScreen molecular pharmacology panel. 9
Table 2: Different QC metrics used for quality assessment 26

Introduction

The contextual DrugMatrix® database is a unique toxicogenomic reference database that contains data on over 600 drugs, chemicals, toxicants, and biochemical reagents. These compounds have been extensively curated and profiled using large-scale gene expression analysis as well as pharmacologic activity (MDS Pharma assays) and in-life evaluations, including hematology, clinical chemistry and histopathology.

DrugMatrix database contains various types of data that require different modeling and handling, including normalization, ratio calculation, and denominator groupings. This document specifically addresses the key computations currently used by DrugMatrix for different types of data modeling. A detailed description of raw data handling, such as removal of no data spots, exclusion of various data points and data normalization is given. The gene expression changes between treated and control samples and the supporting statistics to determine their significance in the current version of DrugMatrix is summarized in detail. The rationale for using the different denominators for ratio calculation of gene expression, blood chemistry and histopathology data is described. The algorithms and computations that form the basis of Drug Signature technology are elaborated in detail, as well as the various analytical tools that are utilized to assist data visualization and interpretation in DrugMatrix database. In addition, compound curation, pathway curation and gene annotation processes are described in detail in this document.

For ease of reference, this document is divided into six chapters each addressing a specific domain of the DrugMatrix database: compound curation, pharmacology assay panel, in-life data domain, gene expression domain, Drug Signature technology and DrugMatrix analytical tools.

Chapter 1. Compound Curation

1.1. Structure Activity Class

Each compound studied in DrugMatrix™ is assigned a structure activity class based on the molecular target associated with its approved clinical use, as reported in the literature. When the molecular target is not clearly defined, the compound is classified based on the accepted or most relevant mechanism of action and/or clinical indications. When appropriate or necessary, a compound is classified with both its mechanism of action and molecular target in order to bin compounds into a more general group. When diverse chemical structure types are active against the same protein target, the compounds are sub-grouped into pharmacophore types, and a structure activity class is assigned to each sub-group based on the molecular target and chemical structure type.

1.2. Activity Class

Each compound studied in DrugMatrix™ is grouped into an activity class, which represents a more generic compound annotation than structure activity class. Compounds are grouped together based on having structure activity class annotations that are related by a common therapeutic activity (i.e. anti-inflammatory) or toxicological activity (DNA damager). Structures are not considered when grouping compounds, such that unrelated structures that act through a common molecular target are grouped together. Likewise, compounds with distinct, but pharmacologically related targets, are also grouped together under a single activity class term. Compounds with a structure activity annotation unrelated by therapeutic or toxicological activity are simply annotated with their structure activity until such time that related compounds are added to the database.

1.3. Indication

Each compound studied in DrugMatrix™ is associated with its approved clinical indications. A comprehensive vocabulary list of 1822 indications was created for this purpose. While curating clinical indications for a given drug, terms from this list of 1822 matching those found in the Physician's Desk Reference (PDR), drug package insert, subscribed online databases, or primary literature for the drug are selected and curated.

1.4. Therapeutic Class

The purpose of the "therapeutic class" category is to classify a compound based on its therapeutic uses with respect to its indications. A comprehensive list of 120 therapeutic classes, such as "Antidiabetic Agents", "General Anesthetics, Intravenous" or "Antibacterials, Systemic", was established for this purpose. During the process of curating drug indications, each indication is associated with an appropriate therapeutic class term from this list of 120. The same therapeutic class may be associated with several indications of a given compound. For example, the therapeutic class "Antibacterials, Systemic" may be associated with

the indications “Mycobacterium tuberculosis”, “Erythema Nodosum Leprosum (ENL)”, “Leprosy”, and “Atypical Mycobacterial Diseases” of the same compound.

1.5. Product Class

Product class, such as “Hormones, Endocrine and Metabolic”, “Central Nervous System (CNS)”, or “Anti-infectives” is a general industry classification of the drug. A vocabulary list of 24 terms was established for product class. Each therapeutic class is pre-associated with a product class in the curation database so that when a therapeutic class is selected, the product class is determined automatically.

1.6. Mechanism

Mechanism displays one or more descriptive phrases established in DrugMatrix™ to classify a compound based on its phenotypic or physiological effect such as “Inhibit platelet aggregation”, “Increase lipid catabolism” or “Block neural transmission.” A list of 75 Mechanism terms was established for compound curation.

1.7. Mode Class

One or more descriptive phrases were established in DrugMatrix™ to describe a compound based on how it affects its molecular target. A list of 75 Modes have been created for the assignment of the Mode Class, such as “Enzyme Inhibitor”, “Receptor Agonist, Selective” or “Channel Blocker.”

1.8. Development Status

The development status category tracks the status of a compound in a typical drug development pipeline. A vocabulary list of 16 was developed for curation. A compound can be “FDA Approved”, “Marketed”, or “Withdrawn” at various development phases. Compounds which are not drugs may be “Toxicants” or “Biochemical” standards.

1.9. Adverse Drug Reaction Classification

A comprehensive vocabulary list of 1,600 adverse effects was created for the purpose of curating adverse drug effects. Each of the adverse effects was associated with a tissue/organ ID and given an “Adverse Effect Group Name”. Each of the adverse effects is also assigned a severity index based on the seriousness of the adverse effect or clinical event. Genotoxicity, mutagenicity, carcinogenicity, and life threatening or potential life threatening events are given a score of “SSS”, serious but manageable events are given a score of “SS”, and measurable but not serious events are give a score of “S”.

During the curation of adverse drug effects, terms from the list of 1,600 matching those in the PDR, drug package insert, or primary literature for a particular drug are selected and given a frequency score (1- most frequent, 2-less common, 3-rare) according to the information in the documents. The “Adverse Effect”

associated with each drug in DrugMatrix™ is thus the curated adverse effect terms for each drug prefixed with the associated tissue/organ ID and the frequency score (e.g., LIV_1_Increased Liver Enzymes; CVS_3_Bradycardia). If an “SSS” severity label is associated with an adverse effect, then the curated adverse effect appears in “Tissue Toxicity” of the compound, while the “Adverse Effect Group Name” appears in “Known Toxicity” of the compound.

1.10. Known Toxicity

This category displays one or more common terms used to highlight serious adverse effects associated with a compound, such as “Carcinogenicity” or “Electrolyte Disturbance.” All compounds associated with an adverse drug reaction classified with a severity score of “SSS”, as well as compounds associated with any cardiovascular toxicity, hepatotoxicity, or nephrotoxicity with a frequency score of 1, are highlighted by the respective “Adverse Effect Group Name”.

1.11. Tissue Toxicity

The adverse effect term associated with known toxicity as defined above is displayed.

Chapter 2. DrugMatrixScreen™ Pharmacology Assay Panel

2.1. Molecular Pharmacology Assay Panel

An important component of DrugMatrix™ is the molecular pharmacology assay panel. These assays provide additional biological depth and annotation to the compounds and to important drug targets (gene products) in DrugMatrix. The panel consists of 125 biochemical assays of receptors, ion channels and enzymes. These include the targets of important drug classes, targets of important side effects, and recombinant P450's important for human drug metabolism and drug-drug interactions. Table 1 provides a brief, incomplete list of some of the assays as an example.

Receptors and Ion Channels	Panel Selection reason(s)	Enzymatic Activities	Panel Selection reason(s)	P-450's	Panel Selection reason(s)
Adrenergic a_{2B}	Blood pressure changes	**Cyclooxygenase COX-1**	Important drug Target	**CYP450, 3A4 Human**	Drug-Drug Interaction, and Drug Metabolism
Cannabinoid CB_1	Cannabinoids are an area of active research in Pharma	**Peptidase, Matrix Metalloproteinase-1 (MMP-1)**	MMPs are an area of active research in Pharma	**CYP450, 2D6 Human**	"
Glycine, Strychnine-Sensitive	Site of important toxicity	**Monoamine Oxidase MAO-A**	Important Drug site of concern in psychiatric settings	**CYP450, 2C9 Human**	"
Insulin	Important drug target	**Protein Serine/Threonine Kinase, p38a**	P38s are an area of active research in Pharma and site of concern for liver and other toxicities	**CYP450, 2C19 Human**	"
Histamine H_1	Important drug target, of concern in relation to drowsiness	**Protein Tyrosine Kinase, Lck**	Tyr-Kinases are an area of active research in Pharma	**CYP450, 1A2 Human**	"

Table 1: Example assays from the DrugMatrixScreen molecular pharmacology panel.

A complete list of the assays included in the molecular pharmacology panel may be assessed on the web at:

http://discovery.mdsps.com/Catalog/Discovery/AssayPackage.aspx?id=89

2.2. Assay Selection Criteria

The assays included in the panel were selected from the MDS-Pharma Services catalog of more the 800 assays. These assays were selected to include assays of drug targets with known side effects potentially undesirable for new drug candidates, assays of targets for very important drug classes, assays for targets popular within the drug industry and sites of human drug-drug interaction and human drug metabolism. Sixty-nine percent (69%) of the assays are based on expression of the human recombinant form of the receptor or enzyme. Rat-based assays comprise 14% of the assays, and the remainder comprises mouse, bovine, bacteria, guinea pig, and rabbit-based assays.

2.3. Assay Protocol

The assays were performed using receptor and enzyme preparations at the MDS-Pharma Services (MDSPS) high throughput facility using highly automated robotic method. Overall characteristics of every assay were assessed by MDSPS on a daily basis and compared to historical values for the assay; assays not meeting quality criteria were repeated. The performance of key standards against every assay and the exact assay conditions may be viewed at the MDSPS website cited above. Each compound was dissolved in DMSO and applied to duplicate assays at 10 μM final concentration. Compounds were selected for more extensive testing if they achieved or exceeded 70% inhibition at the 10 μM concentration. The follow-up test assayed the response at eight concentrations at ½ -log intervals in duplicate. Concentrations tested began at 10^{-5}, 10^{-6}, 10^{-7} or 3 x 10^{-9} M, as appropriate. The IC50 was estimated using non-linear curve fitting. The K_I is an equilibrium dissociation constant equal to the concentration of the test compound that would bind to half the binding sites at equilibrium in the absence of radioligand or other competitors. The Ki was calculated using the Cheng-Prusoff relationship and the Kd of the radioligand.

$$K_i = \frac{IC_{50}}{1 + \frac{[\text{ligand}]}{K_d}}$$

2.4. Overview of Assay Results

Assaying the DrugMatrix™ compound set with the DrugMatrixScreen panel of assays provide findings never before reported. These kinds of large scale screens provide insight into the assay technology, the characteristics of individual compounds, and individual assays. The statistics and figures provided below are shown to provide the reader some background so that facts drawn from DrugMatrix might be viewed in context. None of these figures is comprehensive or current; however, given that each figure is drawn from a large data set, in an approximately random manner, one could use these figures as a guide since new data will likely continue to exhibit these types of performance characteristics

Figure 1 illustrates the distribution of %Inhibition at 10 μM across 861 drugs for about 130 assays tested over a several year period. These data show that for most compounds, and most assays, the compound is not an effective inhibitor, since the mode of the nearly normal curve is zero percent inhibition. However there are a small minority of all compound-assay pairs (<1%) for which the compound has very high affinity for the target site, causing >90% inhibition at the 10 μM concentration.

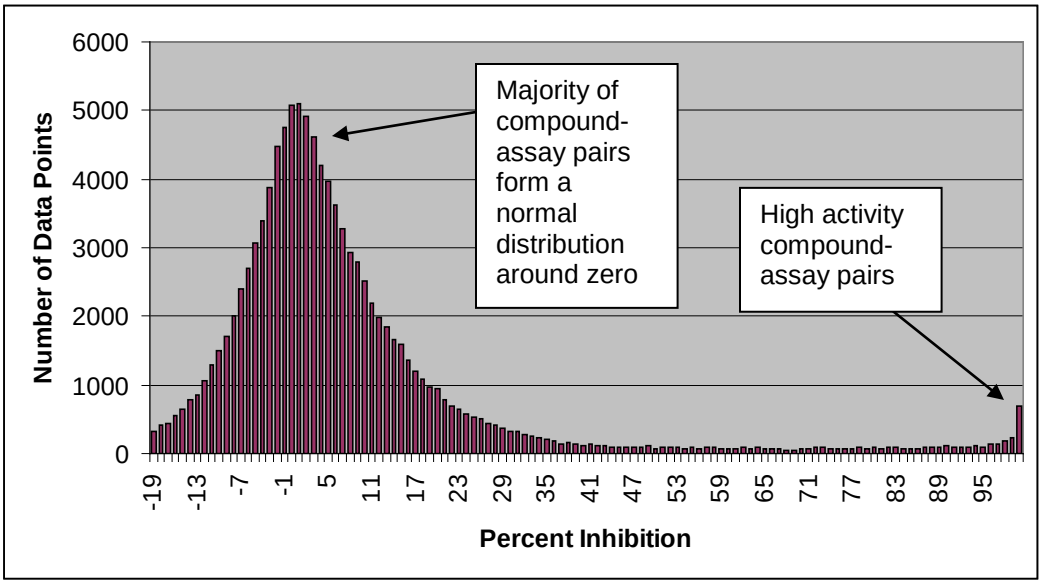


Figure 1: Frequency histogram of 107,478 percent inhibition measurements from 861 compounds tested in up to 130 assays

Figure 2 illustrates the distribution of IC_{50} values observed for a panel of 506 compounds tested in up to 130 assays over the course of several years. This curve indicates, as one might expect based on the screening protocol, that the mode of this histogram is 2.8 μM, indicating that most of the compound affinities for various receptor or enzyme binding sites are weak. However, significant minorities of compounds have IC_{50}'s less then 1 μM; these higher affinity

interactions might be the cause of drug-specific off-target effects and should be considered when examine the biological and transcriptional effects of those drugs and compounds.

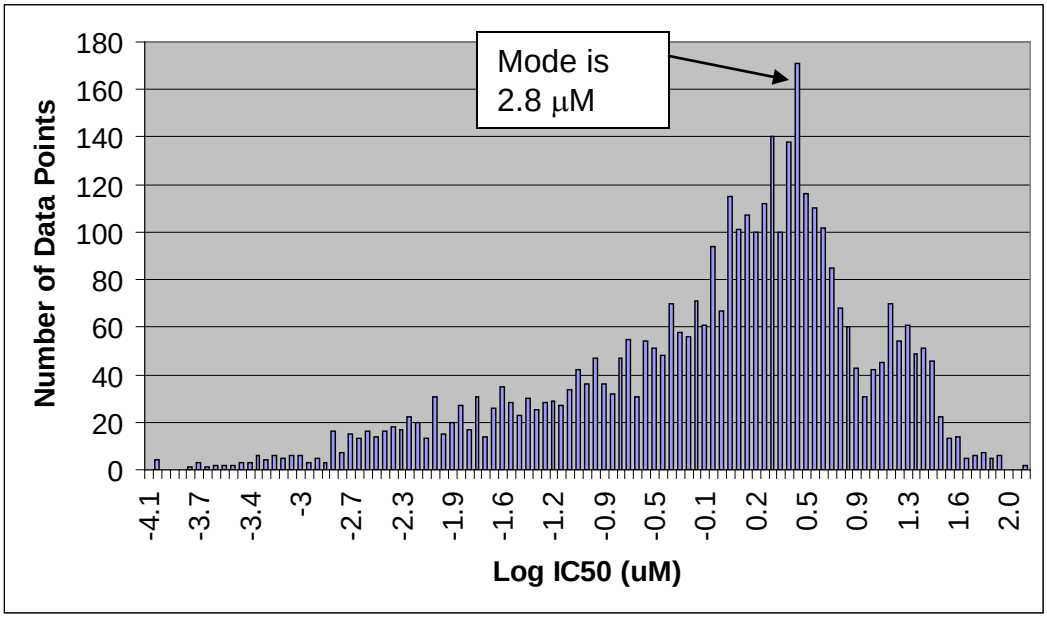


Figure 2: Frequency histogram of 3,845 IC_{50} (log-μM) measurements from 506 compounds tested in up to 130 assays

Note: the apparent bimodal nature of this curve is a result of an early testing process. During early work, the first 100 compounds of the 506 represented here, were screened at 30 μM and retested in an eight concentration point ½ log dilution series if their percent inhibition exceeded 50%. The protocol as outlined above was adopted after the first 100 compounds were tested.

Figure 3 gives a more in-depth look at the characteristics of individual assays. This figure illustrates the frequency of any of 156 randomly chosen compounds inhibiting 123 of the assays. The data show that some assays, such as the 5-HT2b receptor assay, are quite promiscuous; 16% of the compounds tested bind to this receptor. In contrast, other assays such as the insulin receptor binding assay (not shown) are highly specific since none of the 871 compounds tested, except for insulin itself, bind to this receptor. These data are provided to indicate that not all assays are equal in fidelity, and care should be exercised when interpreting potential off-target effects on assays that have a high degree of promiscuity. In contrast, for other highly specific assays, any compound binding to the receptor or channel or inhibiting the enzyme should be considered as highly significant.

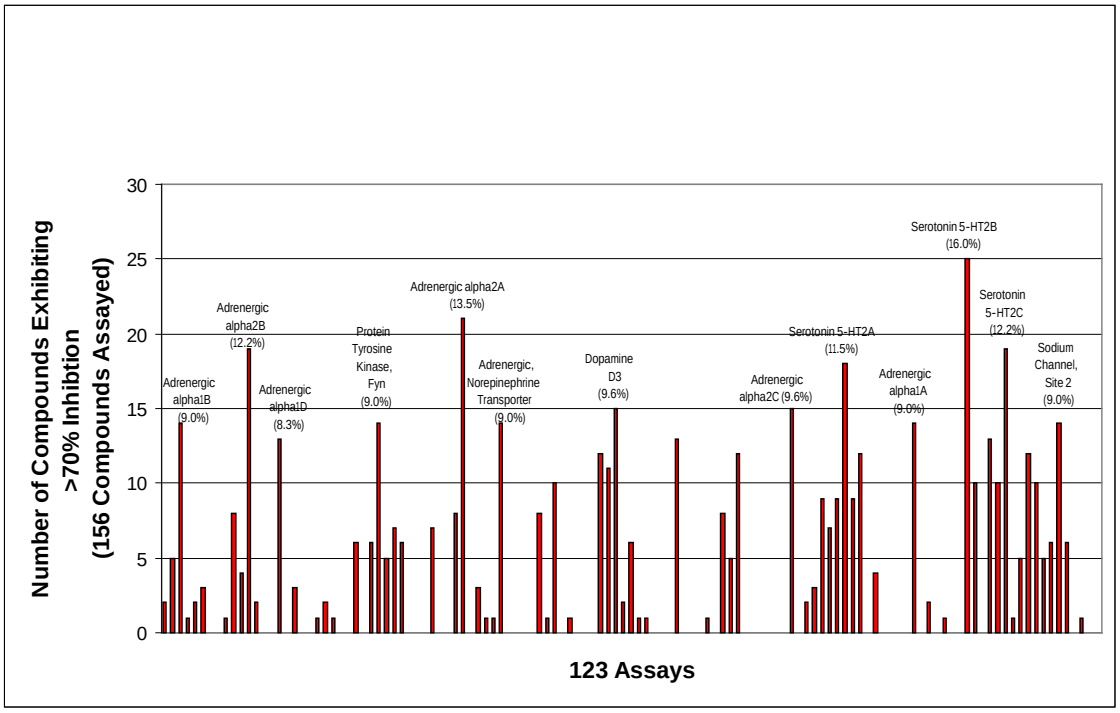


Figure 3: Frequency histogram showing the fraction of 123 assays inhibited by 156 randomly chosen compounds.

Figure 4 examines the characteristics of the 156 compounds shown in Figure 3. Here the specificity of each compound is examined against 123 of the assays. Some compounds such as amitriptyline or sulconazole inhibit a high proportion (24% and 16% respectively) of the 123 assays tested. In contrast, certain compounds such as enalapril (not shown) inhibit only the enzyme that is the designed target of that drug (angiotensin converting enzyme). These data are provided to indicate that not all compounds should be viewed as equally selective; some compounds are quite promiscuous and bind to numerous sites while other are highly selective and bind only to its intended target. Thus a compound's "native selectivity" should be considered when interpreting biological effects; promiscuous compounds might be expected to have a wide variety of effects, while unusual effects of a compound that is otherwise very selective should be considered as highly significant and might indicate that there is another, perhaps novel receptor. Taken together, figures 3 and 4 indicate that the effect of any given compound on any given assay should be interpreted in the context of both the compound and the assay specificity. The data from the DrugMatrixScreen panel of *in vitro* molecular pharmacology assays provide the necessary context.

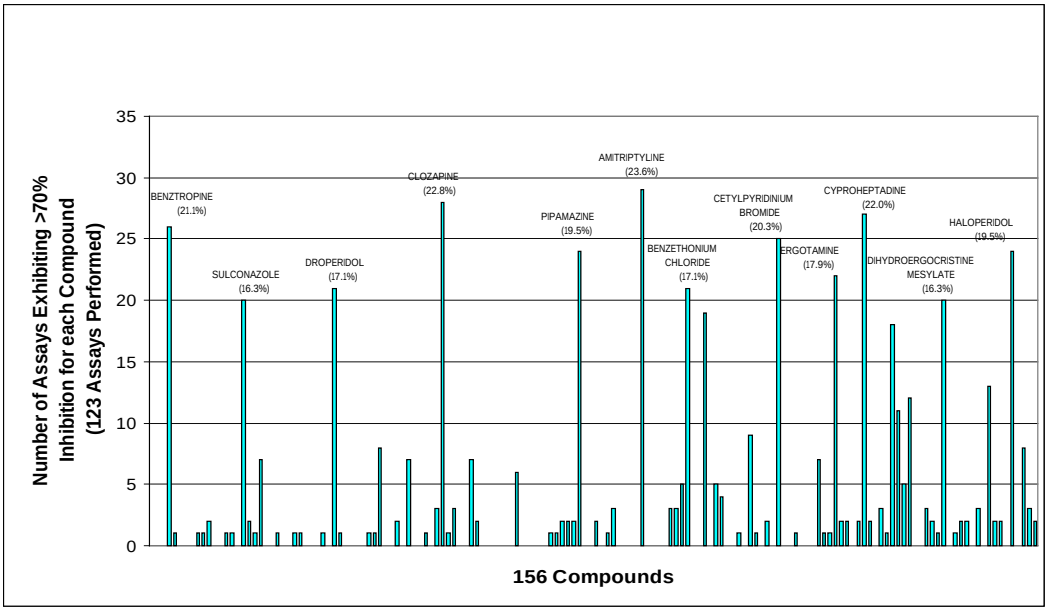


Figure 4: Frequency histogram showing the fraction of 156 compounds that inhibit each of 123 assays

Chapter 3. In-life Data Domain of DrugMatrix

3.1. Introduction

The in-life domain of DrugMatrix is composed of *in vivo* body and organ weight data, hematology and clinical chemistry data and histopathology observations. Please refer to *in vivo* experimental protocols for details on compound and dose selection, animal treatment and data collection for various in-life parameters.

3.2. Blood Chemistry Data

3.2.1. Introduction

Blood chemistry data are collected for control and drug-treated animals for all *in vivo* studies. 19 different clinical chemistry assays are performed on the Hitachi 911, and 21 different clinical hematology assays are run on the Baker 9000 at QCL labs. For details on each individual clinical chemistry and hematology assay please refer to the clinical pathology description document in DrugMatrix.

3.2.2. Log Ratio and Significance Calculation

The log ratio for blood chemistry data is calculated in the same manner as the gene expression data in DrugMatrix using denominator specified in the following section. For significance determination, one-way ANOVA was performed followed by Dunnett's test to compare the mean of each dose group to the control.

As in the case of the gene expression data, all valid data is therefore used. Assay data for an experimental animal tissue is used regardless of whether an expression array has been run for the sample.

3.2.3. Normal Range of Historical Controls

The normal range of historical controls is calculated as the inter-percentile value between the 5th percentile and the 95th percentile.

3.2.4. Denominator Calculations

There are four major variables, route, vehicle, timepoint, and study date that can potentially influence the baseline control values for each of the assays measured. The denominator was selected for each assay taking into consideration of the relative importance of each of these variables for calculating the mean log ratios for treated versus control animals. The expression levels show a dramatic shift after the date of 3/1/2002. As a result, the decision was made to build reference sets using the vehicle and time matched controls separated into two groups based on a study date before or after 3/1/2002 for all clinical data.

3.3. Histopathology Data

Histopathology evaluations were conducted by board-certified veterinary pathologists sub-contracted by the original developers of DrugMatrix, Iconix

Pharmaceuticals. The histopathology data is mapped from the CRO source files through the experiment conditions (study date, drug, dose, route, vehicle, and time point). The incidence of each lesion is listed and coded with distribution information and the severity of each lesion is scored on a severity scale of 0-4 (normal, minimal, mild, moderate, and marked). For the replicate sets, histopathology scores are averaged across the animals in the set. Statistical significance was determined using the ridit score calculation (see below). Histopathology data is presented in the DrugMatrix data warehouse and also through the PathLab reports available through the DrugMatrix software.

3.3.1. Ridit Scores for Histopathology Data

Ridit analysis is a statistical technique that works with ordered categorical data by evaluating categories in terms of their frequency of occurrence relative to a control or reference group. The term ridit is derived from the initials of “relative to an identified distribution”, though it is written in lower case, not as an acronym. It has been applied in a wide variety of fields since it was invented in the 1950s.

The advantages of the ridit approach are category scores are not assigned an arbitrary numerical value and the scoring adapts to the distribution of scores in the reference population.

3.3.2. Calculating Ridit Scores

Let the histopathology scores take on values of increasing severity from 0 to K (K is always 4 in our scoring scheme). For each histopathology measure and each tissue construct and save the ridit table for that measure and tissue by the following:

Count the frequency of occurrence of each score level across all control animals in the tissue as n_k, $k = 0, \ldots, K$

so that no frequency is zero, add $\frac{1}{2}$ to each n_k, $k = 0, \ldots, K$

compute the total (modified) frequency $N = \sum_{0}^{K} n_k$

calculate the ridit score table as $r_0 = \frac{n_0}{2N}$ and $r_k = \left(\frac{n_k}{2} + \sum_{0}^{k-1} n_j \right) \Big/ N$, $k = 1, \ldots, K$

Apply the ridit score table to all individual histopathology outcomes for that measure and tissue, storing the value of r_k. This will be a new entry in the tables where histopathology data is stored. The individual ridit scores are a measure of the probability of having an outcome score of k or less on that particular measure in that particular tissue.

For each treatment condition, summarize the mean ridit score $\bar{r}$ across that treatment condition and record it. The mean ridit score can be interpreted as the estimated probability that a randomly selected individual from the treated group

has a more severe outcome than a randomly selected individual from the control group.

Compute and save a p-value for testing whether the treatment group has a statistically significantly higher average ridit score than the control group. If there are n animals with measured histopathology score in the treatment group then compute $z = \frac{\bar{r} - 0.5}{se(\bar{r})}$, where $se(\bar{r}) = \frac{1}{2\sqrt{3n}}$ is the approximate standard error. Then the p-value is $1 - \Phi(z)$, where Φ is the unit normal distribution function. This p-value should be stored along with the ridit score.

The ridit tables in step 1 were constructed once based on all available control data and stored for reference. If significant additional data becomes available for a particular tissue, the tables for that tissue will be recomputed and all ridit calculations for that tissue recalculated.

References

Fleiss, J., Levin, B. and Paik, M. *Statistical Methods for Rates and Proportions, Third Edition*. Wiley-Interscience, 2003, pp198-205.

Chapter 4. Gene Expression Domain

4.1. Introduction

This chapter addresses gene annotation and key computations for determining gene expression changes between treated and control samples and the supporting statistics to determine their significance in the current version of DrugMatrix™. Details such as data point removal, normalization methods and cross microarray platforms analysis are explained.

4.2. Probe or Probe-set, Expresson and Gene Definitions

The DrugMatrix database uses a data model that organizes genomic data into three layers (see Figure 5) with different functions. They are:

Probes or Probe-set: the basic units for measuring gene expression as implemented on oligo-based microarrays. Probes are built from specific short nucleotide sequence runs, typically 25 to 30 nucleotides, chosen from longer sequences pulled from sequence databases. Probe-set is a summarization of a group of approximately 16 probes, consisting of perfect-match and mis-match probes, as calculated by a summarization algorithm on Affymetrix arrays. Probe and Probe-set are usually annotated with a Genbank accession number that represents the DNA sequence record from which they were derived.

Expressons: a conceptual layer that represents aggregations of probes that are known based on empirical data, to have similar expression profiles across a large number of compound treatments. Expressons can be thought of as “expression units” that remove some of the redundancy in the probe sets found on commercial microarrays. By aggregating probes we reduce the complexity of interpreting expression data from a number of redundant measurements. In addition, we use the expresson layer to map probes from different chips and chip platforms into an entity that represents a platform-independent expression unit. In this way, the expresson layer captures the cross-platform mapping between probe-sets on the Affymetrix and probes on Codelink arrays. There can be one or more probes mapping to a single expresson. All the basic DrugMatrix™ computations are completed at the expresson level.

Genes: a layer that represents a functionally-annotated unit of gene expression. The gene information is used to aggregate the expresson-level data and simple averages are computed. Aggregates of expresson are formed from probes with accession numbers accorded to the same gene by Unigene. While there may be objections to simple averaging of statistical computations, e.g. standard errors, in practice this is not considered an egregious distortion of the results.

Genes are annotated based on public domain information in Unigene and Entrez Gene at the NCBI. And are also populated with the Gene Ontology Consortium terms for function, localization and process, when known. Many rat gene

annotations are based on suspected homology to characterized genes in the human and mouse genomes. There can be one or more expressons mapping to a single gene. When two different expressons map to a single gene, the resulting gene expression measurements are labeled with the single gene name, but the expression measurements may not be well correlated, so they are displayed separately as two expressons.

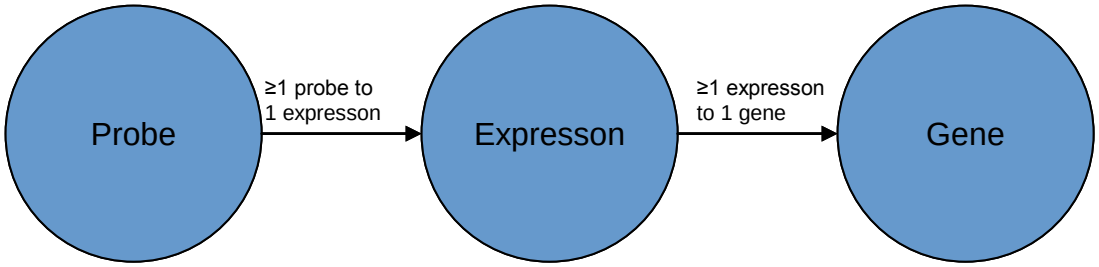


Figure 5: Model for genomics data in DrugMatrix

In DrugMatrix expression changes are assigned to a genomic entity that has a name derived from a combination of gene and probe names. The expresson itself has no name, just an internal numeric database identifier in the data warehouse, but is represented by the group of probes it aggregates. The following example shows the names of two expressons that correspond to the same TAA1 gene.

tumor-associated antigen 1 (L12025_PROBE1)

tumor-associated antigen 1 (L12025_at,NM_017076_PROBE1)

These two expressons both “belong” to the “tumor-associated antigen 1” gene, but the expression of this gene is differentially measured by different sets of probes, some on Affymetrix chips and some on Codelink. When viewing expression data in the DrugMatrix user interface, one must read the “Gene Name” column with this concept of expressons in mind.

4.3. Gene Annotation

Microarray technology measures transcript levels for hundreds to thousands of genes. To understand the biological significance of gene expression changes, a detailed gene annotation for each probe is the essential early step in interpretation. This includes not only the accurate assignment of genes to the specific DNA /probe sequences, but also the detailed gene function, cellular location and their biological contexts.

DrugMatrix™ currently provides the following information from its Gene Domain: gene names, accession numbers (cross-reference IDs for public gene databases), official gene symbol, synonyms, Gene Ontology (GO) assignments, detailed description of gene function, pathway associations, pharmacology assay assignments as well as comparable probes across multiple microarray platforms.

DNA and protein level annotations for microarray elements include not only information automatically extracted and mined from public databases, but also information hand-curated from the literature. The gene information is updated periodically to be consistent with the public domain genomic databases.

The annotations derived from public databases include descriptive and functional annotations from current NCBI releases of the UniGene, Entrez Gene, Homologene, as well SwissProt and Rat Genome Database (RGD). Each probe, where possible, is matched with one of the UniGene and/or Entrez Gene entries. The UniGene and Entrez Gene identifiers are determined according to the probe's representative sequence provided by the manufacturer. The UniGene title, gene symbol, expression tissue, and protein similarities (by BLASTX) are also extracted from the Unigene Database. Similarly, gene name, synonym, EC number, function summary, domain and Gene Ontology information are extracted from Entrez Gene and assigned to the probes via Entrez Gene identifiers. Additionally, Homologene provides some ortholog/homolog relationships for genes from different species.

4.3.1. Gene Names:

The logic flow for choosing a most descriptive gene name for each probe is illustrated in Figure 6. Gene names are chosen from the following sources, in order of preference:

Entrez Gene title, UniGene Title, modified name from mammalian homologs (homolog name + Species + (DBSS_LevelOfSimilarity)) (DBSS: stands for "Determined by sequence similarity"), Swiss-Prot title, modified Genbank description provided by the manufacturers.

First, an automated program searches the Entrez Gene and UniGene databases with NCBI accession numbers provided by the manufacturer. If the accession number for the probe is associated with a Entrez Gene record, then the Entrez Gene official gene name is chosen. If the accession number is associated with an informative UniGene record , then its UniGene Title will be formatted (e.g. removing non informative descriptions such as: *Transcribed sequence, complete CDS, mRNA, partial CDS, Rattus norvegicus* etc.) and assigned to the probes. However, if the UniGene title is not very informative (e.g., *Rn.8666, Similar to RIKEN cDNA 2010106G01 (LOC311401), mRNA*), then its homologs from other mammals will be examined. If they are more descriptive, then the annotation will be modified according to its more descriptive names rather than the one with the highest sequence similarity (e.g. the above name will be changed to "*putative intramembrane cleaving protease (Hs.) (DBSS_moderate)*"). If the accession is not associated with either a Entrez Gene or UniGene record, the probe is checked for a Swiss-Prot-Trembl record, and the subsequent Swiss-Prot name will be assigned to the specific probes. The rest of the probe names will be assigned from BLASTN searches if informative hits are found; otherwise they will be assigned the Genbank descriptions provided by the manufacturer.

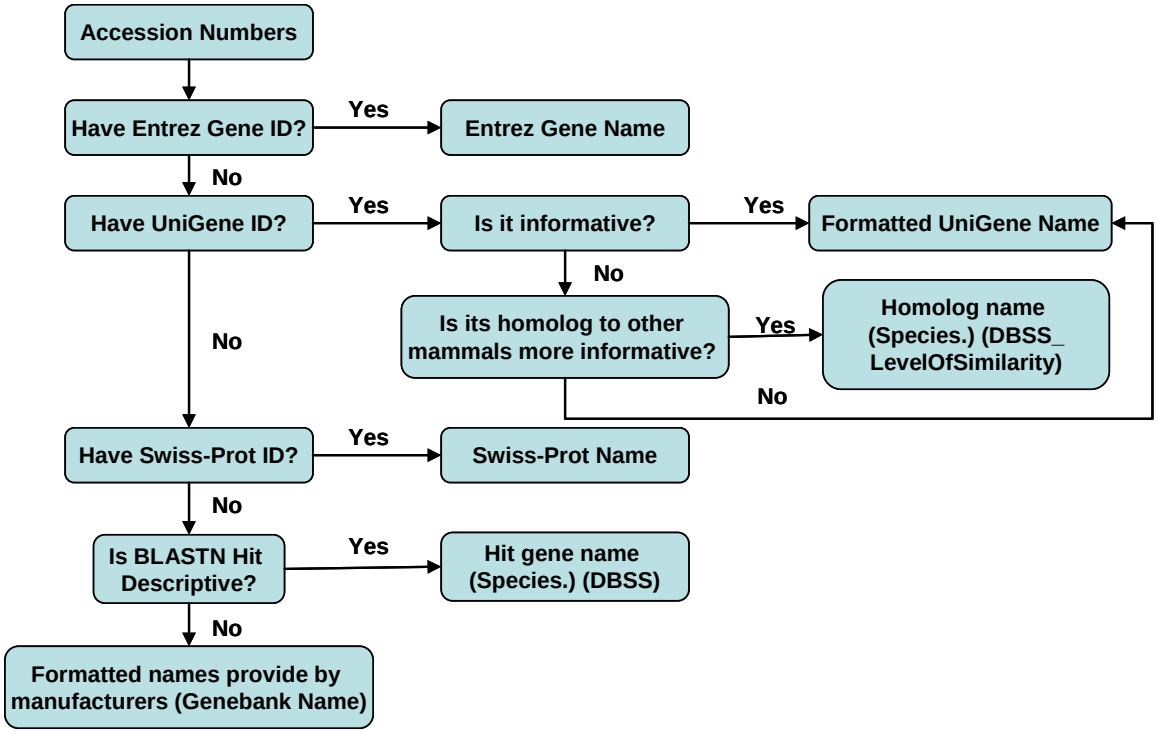


Figure 6: Strategy for derivation of gene names. (DBSS: Determined by sequence similarity)

4.3.2. GO Annotation:

The GO annotations associated with rat arrays are derived by direct extraction from Entrez Gene, comparative trans-annotation of GO terms from rat-mouse or rat-human homologs, and manual curation (Figure 7).

To create GO annotations for rat and human arrays, the accession numbers are used to search for GO annotations contained in Entrez Gene first. Due to the relative scarcity of GO terms provided by Entrez Gene for rat genes, the accession numbers are also used to search for Swiss-Prot-Trembl database, and the GO terms were matched to specific probes by their Swiss-Prot IDs. The GO annotations were further expanded using comparative trans-annotation: the human and mouse homologs of the rat genes are derived from UniGene and/or Homologene. If the homologs have GO annotations in Entrez Gene or Swiss-Prot, the GO annotation assigned to mouse/human gene records were transferred to their respective rat homologs. Manual annotation is an ongoing process to derive GO annotations for the unannotated probes based on textual function summary, publications, domain/motif information etc.

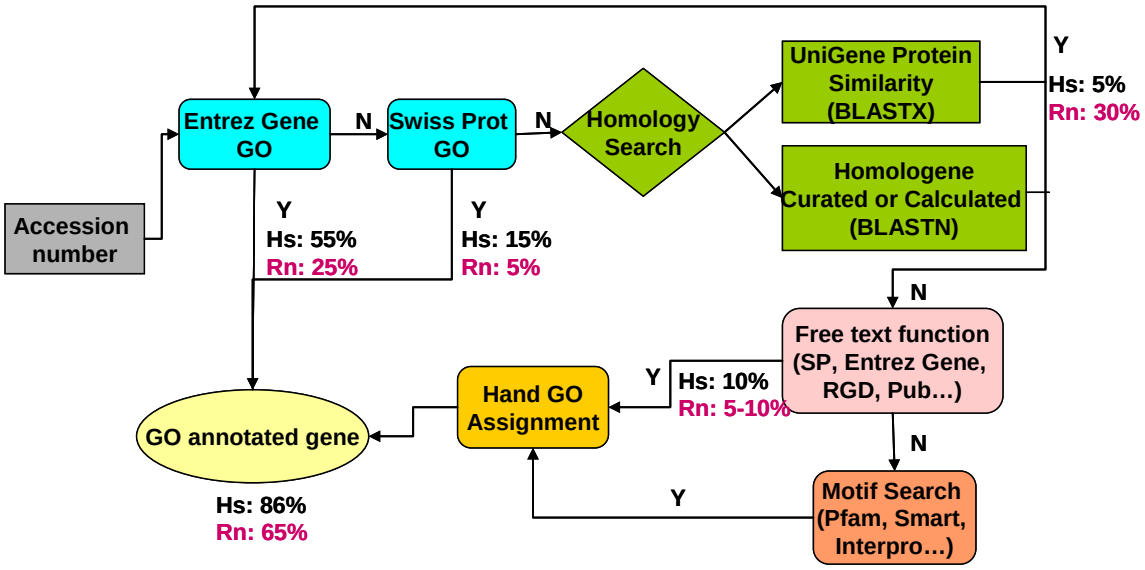


Figure 7: Procedure for GO annotation

4.3.3. Synonyms and Details:

Gene synonyms are a combined list of synonyms mined from Entrez Gene and SwissProt, while gene details are a short paragraph describing the detailed function of the gene, protein family, splice variants, pathway involvement, domain information and expression profiles. They are collected from the same highly curated sources and also from their mouse/human homologs as well as literature information curated by hand.

4.3.4. Update:

Periodically, typically once per year, the most recent UniGene and Entrez Gene reports are parsed and checked against the existing version, and the new UniGene, Entrez Gene ID, updated names and/or other functional annotations are incorporated into the databases. Updates replace the old annotation with a current version in effected DrugMatrix releases, and this fact is noted in the Release Notes included on disk and online.

4.3.5. Sources of Information:

Unigene (http://www.ncbi.nlm.nih.gov/UniGene/)

SwissProt (http://www.expasy.ch/cgi-bin/sprot-search-ful)

Gene Ontology Consortium (http://www.geneontology.org/)

Homologene (http://www.ncbi.nlm.nih.gov/HomoloGene/)

Resourcer (http://compbio.dfci.harvard.edu/tgi/cgi-bin/magic/r1.pl)

RGD (http://rgd.mcw.edu/)

Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene)

Pubmed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed)

4.4. Expression Data Processing

4.4.1. Data Pre-processing

Prior to the computational step, the raw data must be purged of aberrant readings due to manufacturing defects and array processing errors. All arrays pass though a stringent internal QC step before being accepted as valid for inclusion in the final DrugMatrix™ data set. Subsequent to the QC assessment procedure, individual spot readings may be excluded for a number of reasons, impacting the computation of the statistics. These processes are explained in the next two sections.

4.4.2. CodeLink RU1 Array

4.4.2.1. Bad Probe Selection

Currently 1346 of the 9911 probes on the Amersham CodeLink™ RU1 array have been “masked” because of performance issues. Generally, these probes do not behave consistently at the signal level from array to array. The variation is not due to biological noise, but rather to protocol differences and process drift that occur over time.

Bad probes were initially identified when ratios were calculated using denominators derived from control arrays processed before a particular lab protocol change and numerators derived from experimental arrays processed after that same lab protocol change. Lab protocol changes can change base line

signals for a number of probes by as much as 20-fold, resulting in ratios that indicate up to 150-fold difference in regulation of a particular probe between experiment and control, when in fact there is not a difference. Ratios of similar magnitude can be calculated by using as denominators any control arrays processed before the protocol change and as numerators any control arrays processed after the protocol change. Examination of signals of various probes from approximately 200 control arrays showed distinct patterns associated with protocol differences (Figure 8).

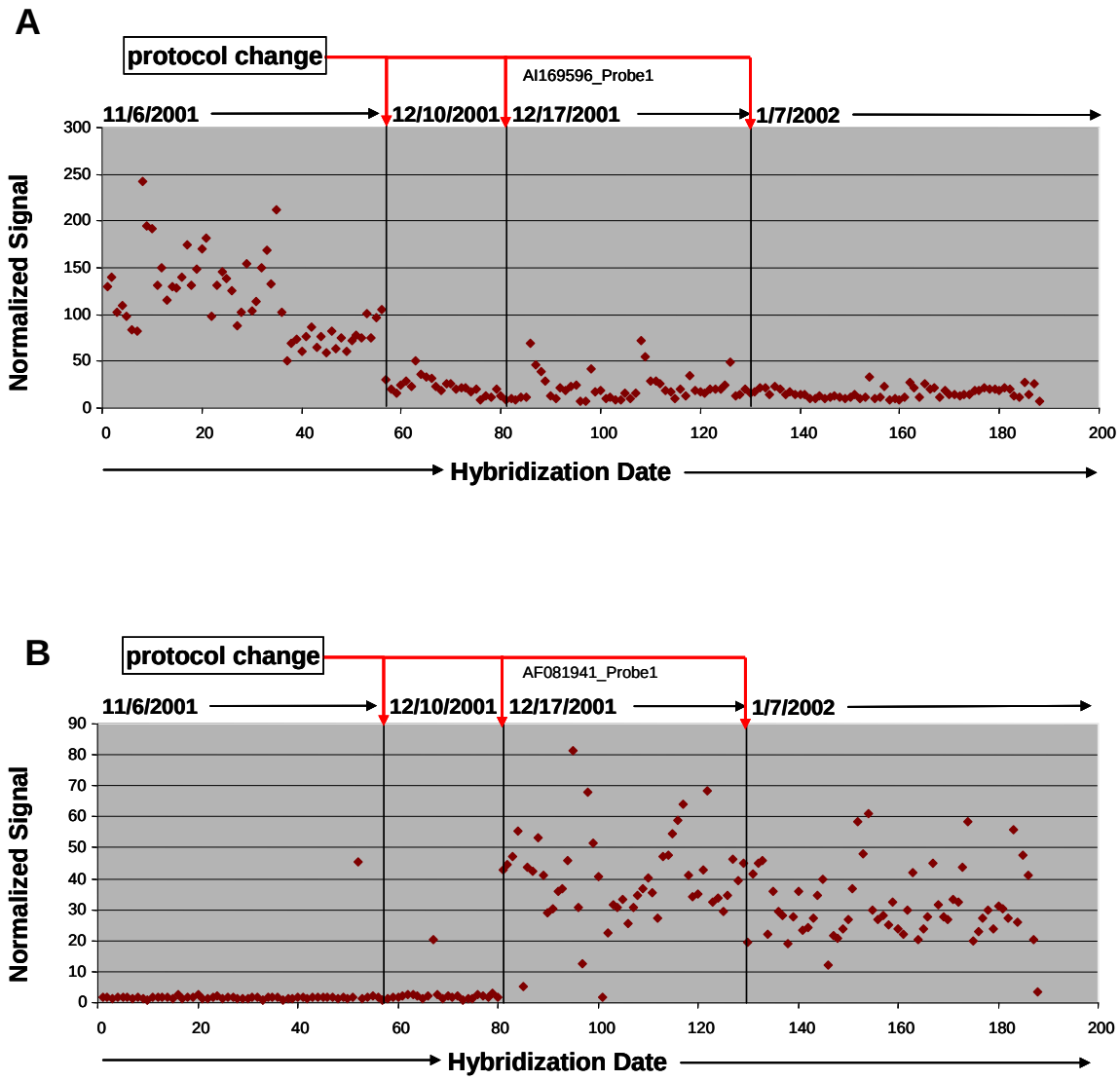


Figure 8: Examples of bad probes

In addition, signal variation was found to occur over time as part of process drift. Clustering of control arrays across the normalized $\log_{10}$ signal of 500 randomly selected probes shows that controls cluster by hybridization date rather than other variables such as time point, vehicle, or route of administration.

Probes that exhibited this unusual behavior were eliminated from standard data analyses so as not to bias the biological interpretation. Probes were selected for elimination first by calculating for each probe the percentage of significant (p-value < 0.05) ratios relative to the number of experiments within a particular time period that used a single control set. Those probes that exceeded 20% were visually inspected individually by graphing experimental ratios for each of the almost 1400 liver experiments arranged by experiment (RepID) hybridization date. Experimental ratios that were not significant at p-value < 0.05 were set to zero to visualize more easily unusual patterns. Probes that showed unusual behavior in liver, kidney, and heart were removed for a total of 1346 probes on the CodeLink™ RU1 array.

4.4.2.2. Data Exclusion

Amersham CodeLink™ RU1 array has a small subset of probes that are declared defective due to manufacturing process errors such as blocked probe deposition, and a larger subset of probes with a “no call” or “unreliable call” reading due to the signal being below some threshold value such as local background. Therefore, some sets of arrays will have fewer readings for each gene than there are arrays in the set.

Manufacturing defects on CodeLink™ biochips are declared in the QC report for each chip. The defective spots are flagged and identified in the output files by a negative number code. Any values associated with these spot readings are discarded.

A subtler problem is probes that have signal readings above a local background threshold, but below some statistical global threshold range. CodeLink™ flags these spot readings; however, since DrugMatrix™ uses multiple signals, any valid local background adjusted spot reading that is greater than zero is considered valid. The measurement noise of these low signal spots is taken into account with the statistical computations on the replicate samples. Thus, any spot values flagged as zero or less are input into the database as null values. All spot readings greater than zero in arrays that have passed QC and not flagged as defective by manufacturing QC are considered valid for further computations.

4.4.2.3. Handling of No Data Spots

DrugMatrix™ computations require at least two valid spot readings (as defined above) for each condition, treated and control. This is the natural lower limit to complete statistical computations, such as standard errors. When this rule is not met, the ratio computations are not completed and the resulting computed values, including the simple mean log ratio, are set to null in the database. When the rule is met, the valid normalized spot readings are used to compute ratios and other statistics. Null readings are stored because some other computations, outside the scope of this paper, require information on the probes present on an array, regardless of whether a valid signal is generated or not.

4.4.3. Affymetrix Array Quality Control Assessment

All probe sets on Affymetrix arrays are closely examined during the development phase. Any probe sets not behaving as expected are either masked out or removed from the array before commercial release. Therefore, quality control assessment at probe-set level is not deemed necessary for Affymetrix arrays.

Affymetrix array quality control assessment at array level is performed using GCOS software as described in the following protocol:

Array scanning and gridding

1. After array processing, the arrays are scanned using the 7G scanner with autoloader
2. Arrays are placed in the carousel immediately after washing (maximum 16 arrays per batch).
3. Following scanning, the DAT raw images file is automatically gridded by the GCOS program to produce the CEL file.

Array QC assessment procedure – as described in the Array Quality Control Assessment Procedure.

1. QC information is loaded to our database and is available for review.
2. First round of QC is based on non-normalized array data using the 5 metrics listed in the table below.

	Metric	Fail if
QC Eval. #1	BG Ave	> 100
	% P	< 30%
	RawQ	> 10
	Noise Ave	> 10
	SFAC	> 12
QC Eval. #2	CC QC	< 0.65
		if correlation to different tissue reference is higher than to annotated tissue

Table 2: Different QC metrics used for quality assessment

It is necessary for an array to meet ALL of the above criteria to be considered a passed array and to be qualified for subsequent rounds of evaluations. None of the other QC metrics suggested by Affymetrix have been shown to be useful in this process. These include 3'/5' ratios of housekeeping genes (Actin, Hexokinase, and GAPDH) and Poly-A+ and Hyb Control spikes. It seems there is little correlation between the ratios of the housekeeping genes to the quality of the array. This is also true to the Poly-A+ Control spikes. For the Hyb Control spikes, any array with a Hyb control spike value less than 100% were already failed due to one or more of the five different metrics listed above.

3. Second round of QC is based on normalized array data, using Pearson correlation coefficient analysis. The Pearson correlation analysis

compares the correlation of any test array (experimental or control array) to each of the tissue reference standards (mean of the average normalized signal across all probes for a set of control arrays for each tissue). In this process the test array is compared to the tissue reference using the Pearson correlation analysis, which results in the identification of mislabeled (correlation to different tissue reference standard higher than to annotated tissue reference standard) or poorly processed arrays. An array is considered to be mislabeled, if the correlation between the tissue annotated test array (experimental or control array) and tissue reference is < 0.8 AND the correlation of the test array to any other tissue reference is > 0.8. An array is considered a poorly processed array, if the correlation between the test array to any tissue reference (except bone marrow and spleen) is below 0.65. Spleen and bone marrow tissues are highly correlated to each other, which is why they cannot be used for filtering at this stage of the process.

4.4.4. Normalization

Prior to statistical computation, spot reading data is normalized. Drugmatrix data normalization uses a proprietary algorithm briefly described in the following paragraphs.

The normalization of the data assumes that in general, for an array with many probes, the majority of the signal represents genes that have unchanged expressions compared to controls and that the outlier values represent the true biology of the process and not some artifact due to measurement noise. The algorithm does not make the assumption that the true mRNA abundance being measured is linearly proportional to the spot reading signal. Rather, the assumption of unchanged signals representing the bulk of the signal measurements is used to center a nonlinear curve fit to a reference template (Figure 9).

This reference template is constructed for a set of tissue, time, and vehicle-matched control arrays, computing a median log signal for each probe. The curve fit can correct for some array processing problems such as partial signal saturation and improves the overall quality of the data compared to simple linear normalization methods. Essentially the curve fits the mode of the signal distribution for sets of genes against the expected value for that set. The effect of this curve fit is shown in Figure 10.

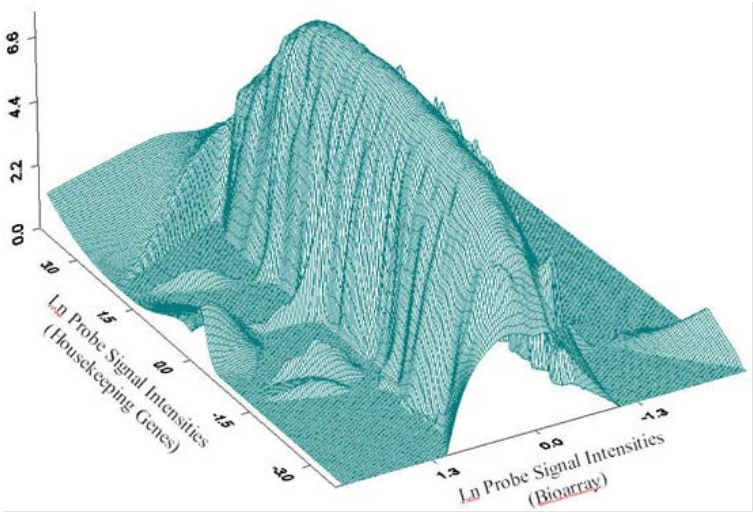


Figure 9: The density distribution of data for a CodeLink™ array.

Plot is for all genes in signal size window against the window signal value.

A. **B.**

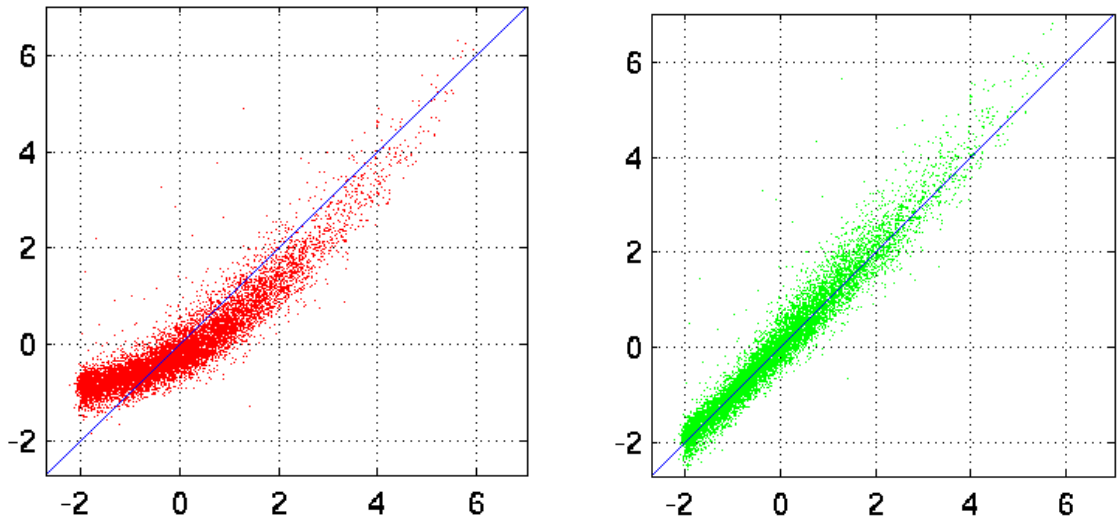


Figure 10: Pre- and post-curve fitting of array data to a reference template.

Reference template log signals on x-axis, log signals of array to be normalized on y-axis. A. Pre curve fit, B. Post curve fit.

4.5. Ratio Calculations and Significance Determination

The primary goal of all calculations is to determine which genes are significantly up or down regulated.

4.5.1. Formalizing the Problem

For $g = 1,\ldots,G$ genes, $k = 1,\ldots K$ (drug/dose/time) treatment conditions, and

$i = 1,\ldots N_X$ replicate measurements with $X_{gki} = log\ (Signal)$ for each gene in each condition (assume *Signal* values are already array-wise normalized). These data are taken with respect to an overall measurement situation defined by microarray type and model, animal strain, tissue, treatment time and administration mode (vehicle/route). For each measurement situation there is a set of log (signal) values for vehicle-treated control measurements, C_{gj}, $j = 1,\ldots N_c$. These control measurements reflect the reference gene-expression levels that form the basis for comparison of each treatment condition.

Statistically, the observed measurements are modeled as having a Gaussian (i.e. normal) distribution, $X_{gki} \sim N(\mu_{Xgki}, \sigma^2_{Xgk})$, and the control measurements as also having a distribution, $C_{gki} \sim N(\mu_{Cgki}, \sigma^2_{Cgk})$.

All of the discussions here are done separately for each gene, so the subscript *g* is suppressed in what follows, for clarity.

4.5.2. Control Set Selection

The effects of each experimental treatment on each gene are measured by comparing the gene expression level in treated rat samples versus an assigned set of control samples. Using proper controls removes the effect of extraneous factors, leaving only the treatment effects. This section documents how sets of controls are assigned for our *in vivo* data.

In early versions of DrugMatrix, study-matched controls were used, where each treatment was matched with the specific control samples from the same animal study with the same treatment time, but this was found not to be the best solution, for several reasons:

Studies contain several vehicle/route combinations, so study controls are not always matched on vehicle or route of administration.

Not all control tissue samples are hybridized

Study controls are not always hybridized near in time to the study experimental samples

Larger control sets reduce standard error

Alternative methods of assigning control sets were studied to find the most effective methods, given the set of available control hybridizations. Overall,

control samples comprise 11% of the DrugMatrix samples. The effective use of these controls allows better analysis of significant expression changes.

Several factors were studied for possible use in forming control sets:

Tissue

Treatment Time

Vehicle & Route of Administration

Study Date (Beginning Date of Treatment)

Laboratory Protocol Version

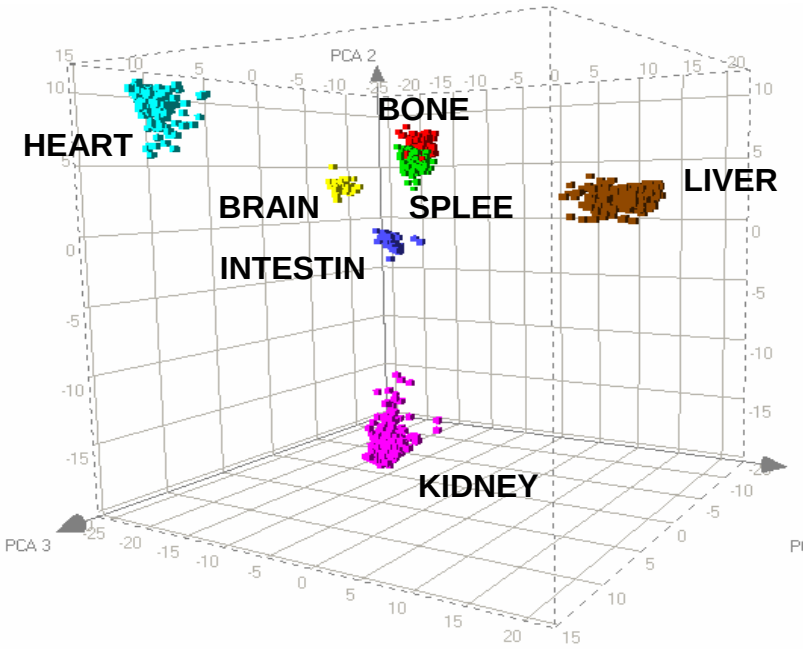
RNA Preparation Date

Hybridization Date

Figure 11 below shows Principal Components Analysis plots of log gene expression signal for each sample, including control samples and drug-treated samples. The PCA plot indicates clear separation among tissues and that the variation due to drug treatment is much less than due to tissue. Therefore, tissue-specific controls must be used.

Upon examination, it also became clear that Laboratory Protocol Version corresponded almost exactly to hybridization date, and that in addition hybridization date also separated some events such as a building move that were not specific protocol changes, but did show impacts in the processing results. RNA preparation date also correspond to hybridization date almost exactly. So of these three variables, only hybridization date was considered.

1,697 control arrays



9,300 drug treated arrays

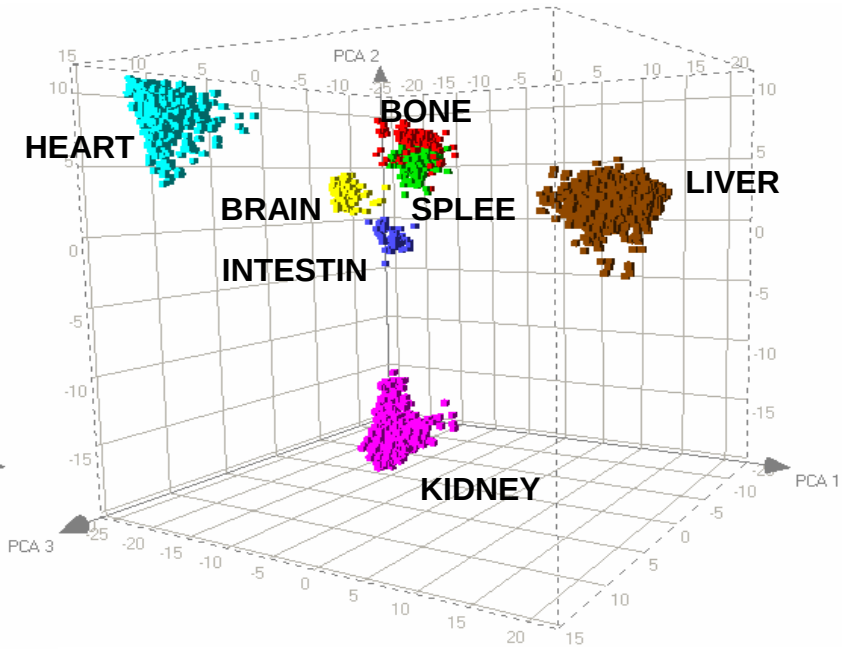


Figure 11: Principal components analysis of Log gene expression signal for control and drug-treated arrays.

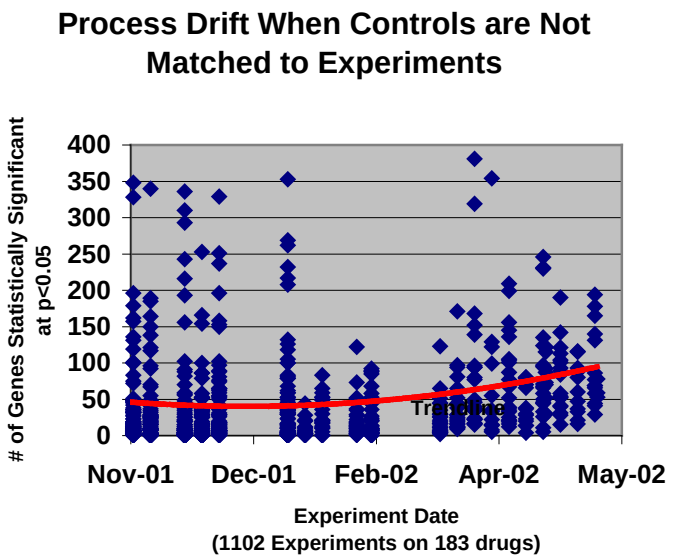


Figure 12: Process drift when controls are not matched to experiments

In considering control set effectiveness, we used a metric of the number of significantly differentially expressed genes to examine different variable combinations. If the control sets are not well-matched to the treatment samples, the apparent number of differentially expressed genes is artificially inflated. So this can be used as a metric to examine different control strategies. Figure 12 illustrates the effect of using a fixed control sample from early in the sampling, and the increasing number of apparently differentially-expressed genes that were seen in later experimental situations.

Another metric of relationship that was used was the across-gene correlation between experimental and control samples. Higher correlations indicated better control matching

Numerous alternatives were considered for control set assignment, including

Static set of controls that don't change that are time and vehicle-matched

All controls (all times, vehicles)

All controls time- and vehicle-matched

Rolling controls from a certain time period (all times, vehicles) matched with experiments from same period (experimental arrays hybridized on different days)

Rolling controls time- and vehicle-matched and from a certain time period matched with experiments from same period

Strict time- and vehicle-matched controls that are processed on the same days as the experiments

Composite of experimental and control arrays (all times, vehicles)

Composite of experimental and control arrays time- and vehicle-matched

Figure 13 illustrates the effect of using control sets from each laboratory protocol period (as noted, equivalent to a period based on hybridization date) as control sets for other periods. Again the best results match controls by period.

Control sets matched on vehicle and route of administration were studied for a number of vehicle/route combinations and found to only minimally improve the outcome of experiments. Similarly matching control sets based on treatment timepoint only showed minimal improvement.

There was improvement found using control sets containing 20 control samples versus 10 control samples. Also, it was found that hybridization of individual arrays of a replicate set on different days led to a more accurate estimate of the within-set replicate variability, which led to fewer false positive gene expression results.

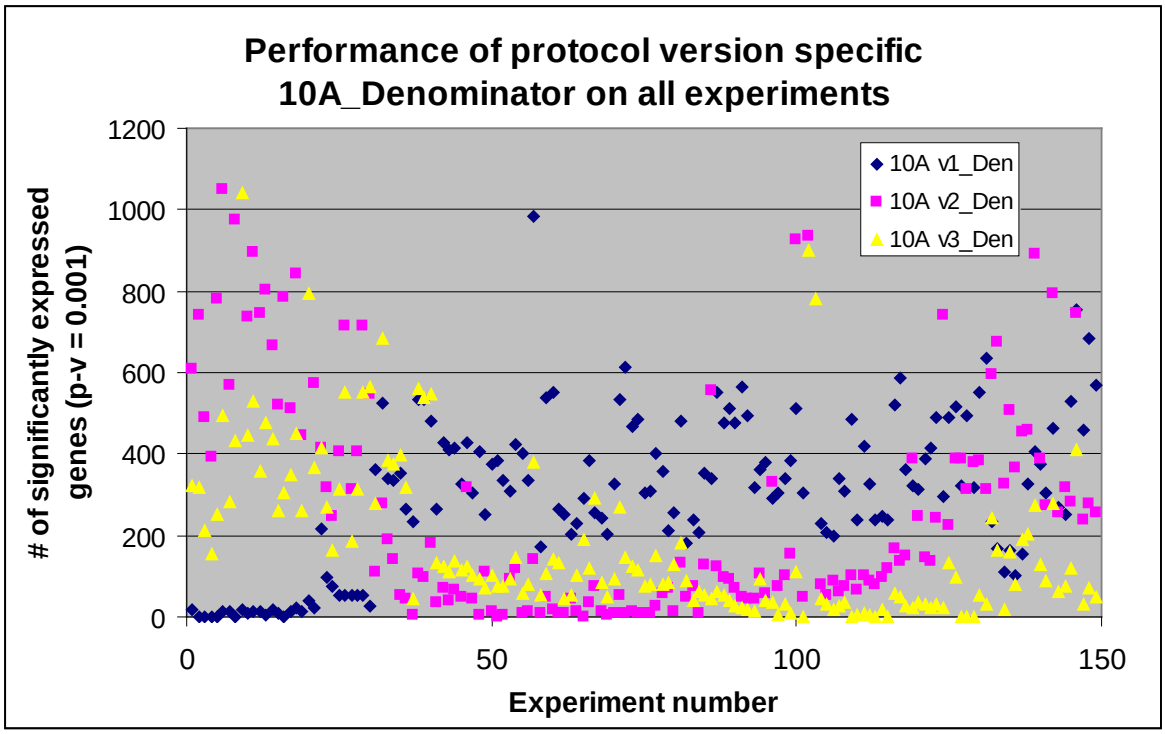


Figure 13: Protocol version specific denominators perform well on proper window, but poorly on wrong window

The conclusions of our study were that the most important factors in assigning effective control sets were:

Match controls by tissue

Match controls using time periods based on Hybridization Date

Samples from treatment replicate sets are each hybridized on separate days, to avoid specific day effects

Up to 20 controls (minimum of 10) are used to form a control set.

Improved precision due to larger control sets is more valuable than effects on a few genes

Hybridization Date acts as a proxy for Laboratory Protocol, RNA Processing Date, and other array processing details.

Matching on vehicle/route and treatment time show only marginal improvement

Where choice is available, statistically balance selection against vehicle/route and treatment time for that period

Our current procedure for in vivo data in DrugMatrix™ is to establish, for each tissue, a series of time periods defined in terms of hybridization date. As far as possible, period changes are made to coincide with laboratory process changes. Separate period definitions are made for each tissue, depending on hybridized control availability. For each period, a set of controls are selected, by the following process:

First, control sets are identified for use in nonlinear array normalization. These are based on the same control periods, but are not limited in size. All control arrays for a given tissue run in a specified period that pass QC are examined. Pairwise Pearson correlations of log signal across genes are computed between each of the samples of a control set. Samples with a median correlation versus all other samples of greater than 0.85 are selected for the normalization control sets.

Then out of these sets, control sets are selected for ratio calculation in DrugMatrix™. Among passing control arrays as given above, up to 20 controls (or all available) are selected. If too many controls are available, selection is made based on a statistical survey balancing method against other factors (e.g. timepoint, route, or vehicle). This control set is then used for calculating log ratios for all treated samples for that given tissue and time period.

Where experimental treatment replicates include experimental samples corresponding to multiple control periods, the control period containing the largest number of experimental samples is used. If this number is the same, the latest such period is used.

In vitro and client data

Control sets for *in vitro* and client gene-expression data are the control arrays from the corresponding study, passing QC and inter-hybridization correlation checks, matched by treatment time.

4.5.3. Log Ratios

For comparative purposes, expression changes in the treated group are compared to a control group and the relative expression values are computed. The differential expression is calculated as a log ratio for ease of statistical manipulation.

The estimated mean log ratio (for a particular gene) in condition *k* is:

$$D_k = \overline{X}_k - \overline{C}.$$

This is the same as the log of the ratio of the geometric means of the treated and control measurements.

4.5.4. Significance of Results

It is important to distinguish between the standard deviation of a population and the standard deviation of an estimated quantity around its true value, called the standard error of the estimate. The classic relationship for simple replication is that the standard error of a mean of *n* measurements with individual standard deviation, σ, is $\sigma / \sqrt{n}$. In our situation, there are two populations, the treated and control groups, and the standard error of D_{gk} around its true value would be

$$SE(D_k) = \sqrt{\frac{\sigma_{X\,k}^2}{N_X} + \frac{\sigma_C^2}{N_C}}$$

Of course, the true values of the σs are not known so it is necessary to substitute an estimate of them. If it is assumed that $\sigma^2_{Xgk} = \sigma^2_{Cg}$, then it is optimal to use a pooled estimate of variance to combine the individual variance estimates for each group, resulting in the denominator of the classic Student's two-sample t with degrees of freedom $df = N_X + N_C - 2$.

However, the assumption of equal variances for controls and experimental animals seems unwarranted, and a safer version of the t-test is the Welch t-test, which estimates the variances separately for each group, using a denominator of

$$SE(D_k) = \sqrt{\frac{(N_X - 1)S_{X\,k}^2 + (N_C - 1)S_C^2}{N_X + N_C - 2}}$$

$$SE(D_k) = \sqrt{\frac{S_{X\,k}^2}{N_X} + \frac{S_C^2}{N_C}}$$

with estimated (and possibly non-integer) degrees of freedom

$$df = \frac{(S^2_{X_g}/N_X + S^2_C/N_C)^2}{\frac{(S^2_{X_g}/N_X)^2}{N_X - 1} + \frac{(S^2_C/N_C)^2}{N_C - 1}}$$

In either case, the t-statistic is computed as $T = \frac{D_k}{SE(D_k)}$, and compared with a standard t-table based on *df* degrees of freedom in order to obtain the corresponding p-values, confidence intervals, etc. (Doing this across many genes at once leads to multiple testing issues, which are beyond the scope of this discussion).

Estimates of SE based only on the data for each situation are very specific, but don't have enough information to get precise estimates of error. In the most common case where $N_X = 3$, there are only 2 degrees of freedom for S^2_X, which leads to imprecise estimates, and in particular, sometimes the estimated sigma can be very small, leading to false positives. Additionally, if only one observation is available, there is no estimate of sigma.

A global estimate of σ could be used, assuming that for each gene, the σs are constant across conditions. This may not be reasonable as different treatments may affect the biological variability of the gene, so it would be better to have a method that allows for this possibility.

The Empirical Bayes approach suggested below uses a weighted combination of these two approaches, where the weights reflect the empirically fitted prior distribution of standard deviations across replicate treatment conditions for each gene. Having many replicated situations, the process can be modeled gene by gene.

For each expresson and gene, DrugMatrix™ displays the log ratio, standard error and p value for the expression changes between the treated and control arrays.

4.5.5. Empirical Bayes Variance Estimates

Log ratio estimates of the differential effect of treatment on each gene are estimated as the difference between the average log signals over two independent set of measurements:

$$\log ratio = \overline{X}_{Treated} - \overline{X}_{Control}$$

The precision of this quantity as an estimate of the true log ratio due to the treatment is represented as the standard error of log ratio, calculated as:

$$stderr(\log ratio) = \sqrt{\frac{\sigma^2_{Treated}}{n_{Treated}} + \frac{\sigma^2_{Control}}{n_{Control}}}$$

where the sigmas are the variance of the log signal in each group and the *n's* are the number of measurements in each group. Of course practice the true variance in each group is not known, and variance terms are unknown and must be estimated from the data. The simplest estimate of variance uses only the data for the specific gene and specific condition, yielding the classical estimate:

$$\hat{\sigma}^2 = S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

However, in gene expression data only a very limited amount of data is available in each group (e.g. for treated conditions, usually n=3). This leads to very unstable and imprecise estimate of variance (only 2 degrees of freedom if n=3). This can lead to the inability to reliably identify significantly differentially expressed genes, since the corresponding confidence intervals must be made large to account for the small number of degrees of freedom. The estimate of variance could be substantially smaller or larger by chance than the true value.

In order to improve on this and to take advantage of the common information in the many similar measurement situations, we have adopted a Bayesian model for variance. In this model, we assume that for each gene, the variance from different treatments is similar (though not necessarily exactly the same), under each treatment condition. In particular, we model σ^2 as having a Bayesian prior, following a Scaled Inverse Gamma distribution with two hyperparameters

$$\nu = \text{degrees of freedom}$$
$$\sigma_0^2 = \text{global variance estimate}$$

Because of our large database of gene expression data, we can estimate these hyperparameters separately for each gene based on observed variances over large collection of treatments. Then for each situation, the Empirical Bayes variance estimates for each situation are calculated by a Bayesian point estimate of

$$\hat{\sigma}^2_{EB} = \frac{\nu\sigma_0^2 + (n-1)S^2}{\nu + n - 1}$$

a weighted combination of global variance and specific variance estimate. The benefits of this process are a more stable variance estimate, larger degrees of freedom, and avoiding getting variance estimates too small by chance.

This method is similar to method used in the Cyber-T software presented in Baldi, P. and Long, A (2001) *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes* Bioinformatics, 17, 6, 509-519. In that application they estimate the hyperparameters for variance in pools of genes grouped by signal intensity, whereas we have enough data to estimate parameters for each gene separately.

4.6. Normalizing Log Ratios Across Platforms

The distribution density of log ratio responses is not equal for all platforms. This impacts the results of computations that rely on the values of log ratios, such as signature scores. To ensure that platforms other than the core CodeLink™ platform respond well, the ratios of the target platform are normalized to the CodeLink™ platform. The method of normalization uses the correlation of signal changes for a set of experiments where the same mRNA samples are used between the target platform and CodeLink™.

4.7. Filtering Log Ratios

The usage of the term “filtering” differs from the traditional definition. Usually, to filter is to select records that fulfill a set of criteria. Internally, filtering is being used to describe the process of setting non-significant data to zero, a procedure also known as squelching.

The purpose of squelching or filtering is to set non-significant expression ratios to zero to repress the noise for additional data mining. The importance of filtering is due to its usage in additional analyses, e.g. Pathway Impact, GO query, etc. Filtering occurs after the calculations of expression log ratios and significance tests as described in Section 4.5.4. Essentially, non-significant ratios, as determined by p-value, are set to zero. P-values of 0.05 or 0.01 are common settings. For clinical and hematology ratios, the calculations for expression ratios and significance tests are the same as those for expression data, and the method of filtering is also the same.

Chapter 5. Signature Technology

5.1. Sparse Linear Programming (SPLP) Algorithm

In order to predict pathological and pharmacological effects using gene expression data, we utilize several previously described classification algorithms (El Ghaoui et al. 2003). Our studies show that several types of linear classifiers based on Linear Programming, Logistic Regression and Minimax Probability Machines can be used to derive readily interpretable drug signatures with high classification performance (Natsoulis et al., 2005). Our methods differ from the well known Support Vector Machines (SVMs) in that they explicitly use the information in the standard error of the $\log_{10}$ ratio in that they maximize a margin based on a “1-norm” or “Manhattan distance”, rather that a “2-norm” or “Euclidean distance”. This tends to produce sparse (short) classifiers. Most of the signatures in DrugMatrix use SPLP (Sparse Linear Programming) and ASPLP (Adjusted Sparse Linear Programming) algorithm. The ASPLP algorithm differs from SPLP in the penalty for a misclassified experiment. SPLP penalized false positive and false negatives equally. ASPLP penalizes false negatives more, compensating for the unbalanced training data (with many more negatives than positives). ASPLP is performed as a grid search across these penalty parameters, and the best-performing signature is selected. Both methods can be tuned to produce classifiers in the form of short, weighted gene lists, an important step towards the development of useful diagnostic biomarkers and for interpretation of the biological basis of classification. Upon analysis it can be seen that some of the signature genes act as positive impact or “rewards” for the class of interest, while others act as negative impact or “penalties”. The combination of reward and penalty genes enhances performance by keeping the number of false positives low. As discussed in Natsoulis et al (2005), the results of these algorithms can be combined with feature selection techniques in order to further reduce the length of the drug signatures.

5.3. Signature Mining

The goal of signature mining is to classify experiments in DrugMatrix by biologically relevant criteria, either manually or systematically, in order to create suitable training sets that can be used to derive a Drug Signature that will predict a phenotype of interest. Quality is ensured by enforcing minimum class size and diversity requirements, as well as performance characteristics estimated from cross-validation. Class size and diversity requirements ensure that a signature is derived from a training set that is broadly representative of that phenotype, rather than representative of any one molecule or sub-structural class that may be inherent in a smaller training set. Minimum performance requirements ensure that the signature has a reliable level of sensitivity and specificity for practical use.

5.1.1. Class Size Requirements:

Signature training sets meet the following minimum conditions in order to be deemed valid:

Number of experiments in positive (+1) class: >=6

Number of compounds in positive (+1) class: >=3

Number of experiments in negative (-1) class: >=44

Number of Activity_Class terms represented in positive (+1) class: >=3 [this requirement only applies to pathology signatures so that the pathology being predicted is not confounded by a structure activity class that may be linked to the pathology]

Experiments must all be on the same platform

Experiments must be from the same sex and strain of rat

Number of missing values: Probes are removed from consideration if they have any missing values (>= 0%) in the +1 class OR >= 5% in the -1 class. Any remaining missing values are replaced with null values.

5.1.2. Performance Requirements:

To estimate the performance of the signature on test data, a split sample cross validation procedure is used. The training set is partitioned such that 60 % of each of the positive experiments and negative experiments are randomly selected and used as a training set to derive a unique signature, which is subsequently used to classify the remaining positive and negative test experiments of known label. An experiment is defined as the average $\log_{10}$ ratio calculated across the three biological replicate hybridizations within a single drug-dose-time combination. This process is repeated at least 20 times, and the overall performance of the signature is measured as the percent true positive rate (sensitivity) and true negative rate (specificity) averaged over the 20 partitions of the data set. The average log odds ratio (LOR) may be used to summarize the performance, and is defined as the natural log of the ratio of the odds of predicting a subject to be positive when it is positive, versus the odds of predicting a subject to be positive when it is negative. It is estimated from a set of training/test cross-validation partitions by the following equation,

$$LOR = \ln \frac{(\sum_{i=1}^{c} TP_i + 0.5) * (\sum_{i=1}^{c} TN_i + 0.5)}{(\sum_{i=1}^{c} FP_i + 0.5) * (\sum_{i=1}^{c} FN_i + 0.5)}$$

where c (= 20) equals the number of partitions and TP_i, TN_i, FP_i, and FN_i represent the true positive, true negative, false positive, and false negative counts on the test cases of the ith partition, respectively. The final signature that

is used on new test data in DrugMatrix is derived from 100% of the training set. This is termed the complete signature.

Validity requirements for signatures depend on their intended use. Signatures designed for assessing toxicity must have a sensitivity (true positive rate) of at least 40%, and a specificity (true negative rate) of at least 98%. Signatures designed to assess pharmacology must have sensitivity of at least 50%, and specificity of at least 95%.

5.1.3. Scalar Product Score Calculations:

A "signature score" is a method of calculating, for any given signature and any given set of measurements on an experimental condition, the class to which the experiment belongs. The signature score of a particular experiment is the linear combination of the form

$$\sum_{i=1}^{n} a_i x_i - b$$

where the x_i are the $\log_{10}$ ratio gene expression values for gene i, a_i are the weights of the signature, and b is the signature bias. If the signature score is positive, then the experiment is classified as positive for the characteristic of that signature. Likewise, if the signature score is negative, then the experiment is classified as negative for the characteristic of that signature.

As can be seen, the weights and bias of a signature can be multiplied by an arbitrary constant without changing the decision rule. For SPLP signatures, as described above, in cases where the fitted signature completely separates the training data (the vast majority of cases, in practice), it is always possible to scale the signature coefficients such that the smallest observed signature score in the positive training set takes on the value +1, and the largest observed signature score in the negative training set takes on the value –1. We refer to signature scores, calculated using signature coefficients that are adjusted in this fashion, as "scalar product" scores. A scalar product greater than or equal to 1 indicates that the effect of the experiment is as great, or greater, than the positive experiments used in the training set, while a scalar product score less than 1 indicates that the effect of the experiment is less than or equal to the negative experiments used in the training set. Scalar products close to the zero threshold should be interpreted as potentially having the characteristic of the signature.

5.1.4. Posterior Probability Score Calculation:

The posterior probability score (PPS) is created to supplement the scalar product score to estimate the probability for an experiment to have the property a signature is designed to detect, allowing evaluation of the confidence in the signature call and providing a fair means of comparison of the strength of signature matches between signatures. The Bayesian posterior probability is calculated from an assumed prior probability of membership in the positive class,

which is determined as the ratio of positive to negative class members in the training data, and the expected distributions of scalar product scores for positive and negative experiments not included in the training set. These distributions are estimated from the scores of positive and negative experiments in the cross-validation partitions. If this probability is very near 1, then we have high confidence that the experiment has the property, if the probability is near 0 we have high confidence that the experiment does not have the property, and values near 0.5 would mean that there is no clear evidence that the experiment does or does not have the property.

Since the posterior probability score ranges only from 0 to 1, it will not be very sensitive to variations at the outer end of the range for purposes of correlations. The probability score can be converted by the logit transform to:

$$PLS = \ln\left(\frac{PPS}{1 - PPS}\right)$$

The PLS be thought of as the log-odds that the experiment belongs to the positive class. A PPS value of 0.5 will translate to a PLS value of 0.

5.3. T-Rank Signatures

T-rank signatures are not binary classifiers, but rather a means to rank a test experiment relative to positive experiments used in the training set. T-rank signatures are created for each SPLP-based signature in DrugMatrix using the same positive experiments in the training set. The signature is created by ranking all genes according to a distinction score that evaluates which genes individually best distinguish the experiments in the positive class of the training set from a reference group of treatments, which can be defined by the negative class of the training set, or defined as zero. The genes that are most differentially expressed, either up or down, are chosen based on a modified version of the distinction metric proposed in T.R. Golub et al., Science 286(5439), 531-37 (1999), calculated as

$$D = \frac{mean_1 - mean_2}{stdev_1 + stdev_2 + k}$$

where $mean_1$ and $stdev_1$ are the mean and standard deviation of the $\log_{10}$ expression ratio for the gene (or other classification variable) across the treatment group, $mean_2$ and $stdev_2$ are the corresponding values across the reference group, and k is a small value, usually 0.1, that is added to the denominator to avoid errors in cases where the standard deviation (stdev) terms in the denominator are zero or close to zero.

The number of genes desired in a signature and their direction are specified as part of signature creation. A number of genes can be specified for the T-rank signature. Four modes of selection of genes are commonly used:

Upregulated genes only, ranked by the largest values of *D*

Downregulated genes only, ranked by the largest values of *-D*

Genes from both directions, ranked by the largest values of |*D*|

A balanced selection of genes, where half of the specified number of genes are selected according to the largest values of *D* and half by the largest values of –*D*.

For the T-rank signatures in DrugMatrix, the forty genes with the highest |*D*| are used, in which $mean_2$ is set to zero. If the number of genes found to have the particular quality or direction of distinction in the given situation is less than the specified number, all genes of the given type are included in the signature, and the actual number of genes in the signature may be less than the originally specified number.

When applied as a classifier for new data, the signature score of this signature applied to a new experimental conditions c is calculated as

$$Score(c) = \sum_{g \in G_S} D_g X_g^c \quad ,$$

where X_g is the experimental measurement (e.g. log ratio expression) for gene *g*, D_g is the distinction value for gene *g* as calculated above, and G_s is the set of genes in the signature. Often these weights are translated into the equivalent form of a Signature Projection Score as discussed in the next section. If this score is to be used for classification, a classification cutoff is established by ranking the scores as calculated above across all of the experimental conditions in the treatment and reference groups, and taking the cutoff to be the value of the tth highest score, where t is the number of conditions in the treatment group, and then classifying new experimental conditions as having the signature attribute if the signature score equals or exceeds this cutoff.

5.4. Signature Projection Score (SPS)

The Signature Projection Score is a way of scaling and interpreting the result of the T-rank signature and to standardize their interpretation across multiple experiments and signatures. It is based on the concept that the score for a signature against a particular compound can be thought of in terms of the relative length of the projection of the gene expression vector (usually in log differential gene expression space) for a particular experimental condition onto a reference vector representing the average differential expression of the genes in the signature between the treatment and reference groups.

The signature projection score for experimental condition c with (log ratio) expression values X_g^c , for a signature with genes g in G_S is calculated as

$$SPS(c) = \sum_{g \in G_S} \frac{\left(X_g^c - R_g\right)\left(T_g - R_g\right)}{S_g}$$

where

T_g is average measurement of the across the signature dimensions among the treated group of interest.

R_g represents the reference, or null, state of the signature on gene g. For signatures this can be taken either a zero log ratio, i.e. no change, or the average gene expression across the reference group (which is often close to 0 for a large negative class). Alternatively, in cases where the Signature Projection Score is to be calculated for a linear classifier such as calculated from support vector machines, R_g can be chosen to correspond to the boundary of the classification rule, such that the linear classification decision would be correspond to the sign of the Signature Projection Score.

S_g represents scaling factors for the contribution of each measurement dimension to the signature score. In the simplest case, the value of S_g is the same for each dimension, and is chosen in such a way that the SPS score for a particular condition or maximum across a particular set of conditions (e.g. the set of conditions used to derive the signature) achieves a specified outcome value, e.g. 100. Alternatively the values of S_g can be adjusted in order to differentially weight the signature dimensions for the precision and accuracy of that measurement dimension. For any signature method such as support vector machines that yields a linear discriminant classification rule, the gene weights of the linear classifier can be easily translated into corresponding values of S_g such that the linear classification rule can be expressed as a Signature Projection Score. In all cases a final consistent scale factor is applied to the S_g in order that the specified outcome value is achieved on the specified condition.

If several conditions are to be evaluated conjointly over a set of treatment conditions C, the score for the entire set can be evaluated by summarizing the SPS score over the individual conditions in various ways, for instance the maximum can be taken over all conditions. Alternatively the score with the largest absolute value can be chosen. Other options are also possible.

One way to visualize this is to take the two vectors of gene expression, experimental and reference, and take the length of the perpendicular projection of the experimental vector onto the reference vector, divided by the length of the reference vector, as a percent. So if the experimental and reference vectors are closely aligned, you get a large signature score, if they are at right angles you get a small signature score. This is of course very much like a correlation between these two vectors, but here the normalization is only with respect to the reference vector, so that we can evaluate the magnitude of the experimental effect in the

direction of interest, not just its direction (if we used correlation, a very small effect that happened to be co-aligned with the signature vector would get a correlation of 1, as would a very large effect in the same direction).

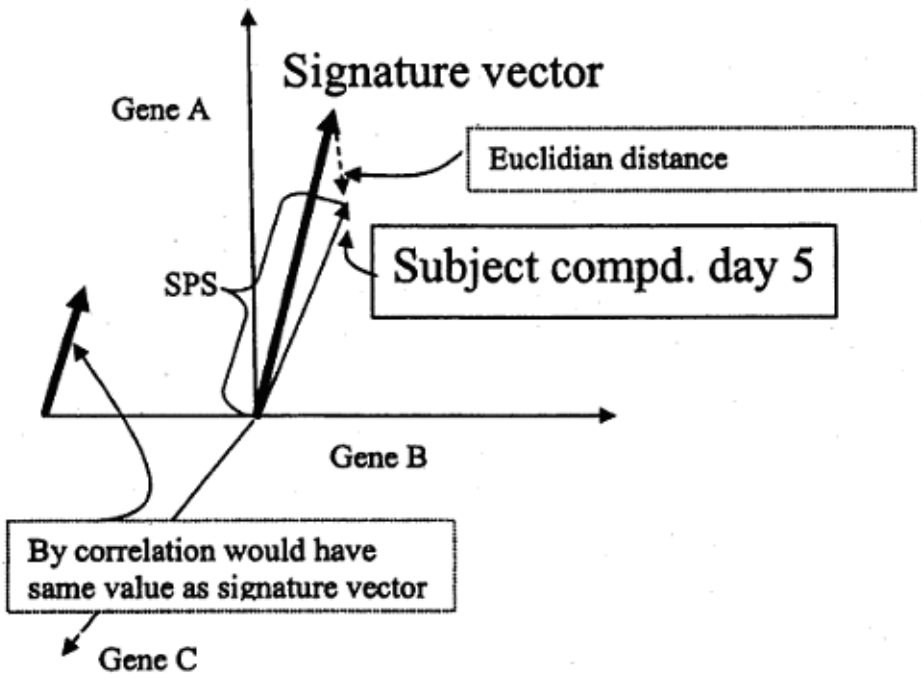


Figure 14: The signature projection score

This method is advantageous over the correlation coefficient method because it takes into account the amplitude of the expression changes as well as their direction; it is superior to Euclidian distance methods because it measures direction and magnitude relative to the untreated or control state and not just the distance between a test compound and some standard compound treatment.

Advantages of the Signature Projection Score (SPS) over other measures include:

It provides a number that relates a compound's gene signature to other compounds and compound families.

It reflects both the intensity of the effect and the closeness of match.

Hypermatches as well as anti-matches to the signature property can be evaluated.

It has an intuitive interpretation feel.

It can be rapidly and easily calculated.

Standard error calculations easily can be performed to represent the uncertainty of the score.

The statistical significance of the finding (p value) can be easily computed.

These properties would not hold for other traditional measures of similarity, such as Pearson correlation or Euclidean distance. Pearson correlation derives only from the similarity of the direction of change (the angle between the two vectors), and does not evaluate the magnitude of the change. Hence it would be impossible to distinguish a small effect in the direction of interest from a large one. Euclidean distance evaluates only the distance between the compound and signature vectors, so a) it does not focus on in the dimensions relevant to the signature, and b) it would score increasing distance (i.e. decreasing similarity) if the compound showed more activity in the direction of the signature than the cases used to build the signature. Other measures of similarity currently in use share similar deficiencies relative to the Signature Projection Score.

5.5. Cross Platform Signatures

Drug signatures are derived from Codelink RU1 data or Genechip Rat230 v.2 data. Signatures derived from one microarray platform, but intended for application to data from a different platform, are called cross-platform signatures. For example, signatures intended for application to data from the RGU34A, and RAE230A were based on Codelink RU1 data. For this purpose, the training sets were identical to those used for the native RU1 signatures, but the variables (probesets) were restricted to those that could be mapped to the RU1 platform. Probesets measuring the same gene as a Codelink probe may still have a different quantitative response to the same experimental sample. To account for these differences, Genechip log ratio data applied to Codelink-based signatures are multiplied by a scaling factor. These factors are specific to each probeset-probe pair, and were determined empirically based on extensive cross platform studies described elsewhere.

5.6. Motif

Motif is a novel application of SPLP signature technology. It utilizes the sets of pathway genes as classifiers to discriminate between different classes of compounds, which are defined in the same way as the training set for the drug signatures. Motif analysis provides a contextual link between drug signature match, which is predictive of certain physiological or toxicological end points, and pathway gene profiling, which can provide insight into the possible mechanism associated with those end points. Similar to the drug signature, a match to a motif by test compound is determined by using scalar product score: a positive scalar product indicates that the test experiment has the property defined by the signature and significantly regulates pathway genes in a way that provide possible explanation of the property predicted by the signature. A negative scalar

product indicates that the test experiment is either not in class as defined by the signature or it has the property predicted by the signature, but can not be explained by the effects of test experiment on pathway genes.

Chapter 6. DrugMatrix Analytical Tools

6.1. Compound/Gene Similarity Comparison

6.1.1. Similar Genes and Compounds Calculation

a. Global Similarity

Two of the data domains available in DrugMatrix™, Genes and Experiments, provide details of nearest neighbor genes or experiments for each domain subject based on hierarchical clustering. Because of the enormous size of the data set, this is not carried out on demand, but rather pre-computed. The similarity score is the Pearson's correlation coefficient between the relevant profiles on log ratio gene expression values. Pearson's correlation measures the degree of linear relationship between two variables. Ranges are from +1 to -1. A correlation of +1 indicates a perfect positive linear relationship. A correlation of -1 indicates a perfect negative linear relationship. A value of R = 0 is random. For experiments, the profile vector is the set of expressons on the array. Missing values are ignored. Where the vector is not the same, for example between different platforms, only the intersecting expressons are used for the calculation of the correlation coefficient. For genes, the vector is the total set of experiments. This means that, unlike experiments, the vector size increases as the database grows.

The 25 highest scoring experiments to the subject experiment are stored. The experiment dendrogram that is displayable in the DrugMatrix™ user interface is the nearest cluster of 25 or fewer experiments based on a global clustering of all experiments across all genes. The cluster is selected by a linear search of the experiment profiles for the closest match. The cluster is expanded by walking up the cluster hierarchy until a cluster exceeds the maximum size of 25 experiments. The cluster of 25 or fewer experiments is returned. For experiments uploaded by the user, the experiment cluster set is re-clustered to include the new experiment.

b. Local Similarity

Expression experiment matrix in the DrugMatrix™ toolbox uses the same clustering method as the global clustering. The data sets are the intersections of the data, and the clustering is done on user demand by the application. This functionality is also available for the bio-activity matrix and blood chemistry matrix calculation in the toolbox.

6.1.2. Chemical Structure Comparisons

DrugMatrix™ contains precomputed structural comparisons from three commercially available and/or published chemical structure comparison algorithms that are readily available.

Thr*ee Dimensional Molecular Imprint* (3DMI) is a proprietary algorithm developed by Ajay Jain at UCSF. It computes a 3-D surface around each molecule and then computes a molecular distance based on comparing all possible surfaces for all possible conformations of two molecules. This extremely compute-intensive process is very slow and cannot be used to compute all pair wise distances between all 8000 structures in DrugMatrix™ (64 million comparisons). Instead, the 3DMI method uses a set of 20 reference molecules that have been chosen to represent a broad chemical space. By computing the molecular distance between the 8000 structures and the 20 reference molecules, the computation is reduced to only 160,000 comparisons. The resulting vectors can be used to compute the distance between any two structures using a simple Euclidean metric in 20-dimensional space. The 3DMI method yields molecular distances, ranging from 0 to ~4, which can uncover interesting 3-D structural relationships between drugs. A distance of 0 represents an exact match.

The *Tanimoto Method* is a well-know metric for measuring similarities of structural fingerprints generated from small molecules. DrugMatrix uses the Tanimoto method as implemented in the Fingerprint Toolkit produced by Daylight Chemical Information Systems, Inc. The Daylight method constructs a structural fingerprint from each molecule, where a fingerprint represents a pattern of bits that describe the molecule's structure. For more information on the theory behind creating structural fingerprints, please see the Daylight Theory Manual (http://www.daylight.com/dayhtml/doc/theory/theory.finger.html). The Tanimoto coefficient is one of many methods to compute similarity between bitmaps, and it is simply a measure of the fraction of bits that are in common between two fingerprints (number of bits in common / total number of bits). The Tanimoto method yields molecular similarity measures in the range from 0 to 1, where typically 0.7 represents a good match. A value of 1 represents a perfect match.

MOLSIM is a molecular similarity method implemented in MDL Information Systems' ISIS Relational Chemical Server product. The MOLSIM method yields a molecular similarity metric in the range from 0 to 100. A value of 100 represents a perfect match. A value of 50 has a p-value of approximately 0.05.

While each of these algorithms provides a metric on compound similarity, metrics from the different methods are not comparable. To create a more comparable number that would allow the user to compare results from multiple similarity algorithms and to rank-order these mixed results, a simple statistical metric, a p-value, is computed for each type of algorithm.

The method to compute the p-value associated with a compound-similarity score is quite straightforward and empirical. To do this, for each algorithm the pair-wise similarity or distance between at least 1000 randomly selected structures from a much larger database of structures is computed. These 1 million pair-wise comparisons are assumed to represent the population of all structural comparisons in chemical structure space. The distribution of similarity or distance scores is analyzed. For the three algorithms described so far, these distributions

are normal or near-normal. Since the distributions are near-normal, we simply identify a cutoff that represents the similarity or distance score of the top one thousandth match in the distribution and assign this score as the P=0.001 value for that algorithm.

The distribution of Tanimoto scores, a similarity metric where 1 represents a perfect match and 0 is the minimal score, is shown in Figure 15. Since the higher-scoring matches represent the most similar compounds, a score of 0.8 or greater was identified as having a P-value of 0.001 or better. The rule of thumb, according to the Daylight documentation, is that a score of 0.7 or greater represents a significant similarity.

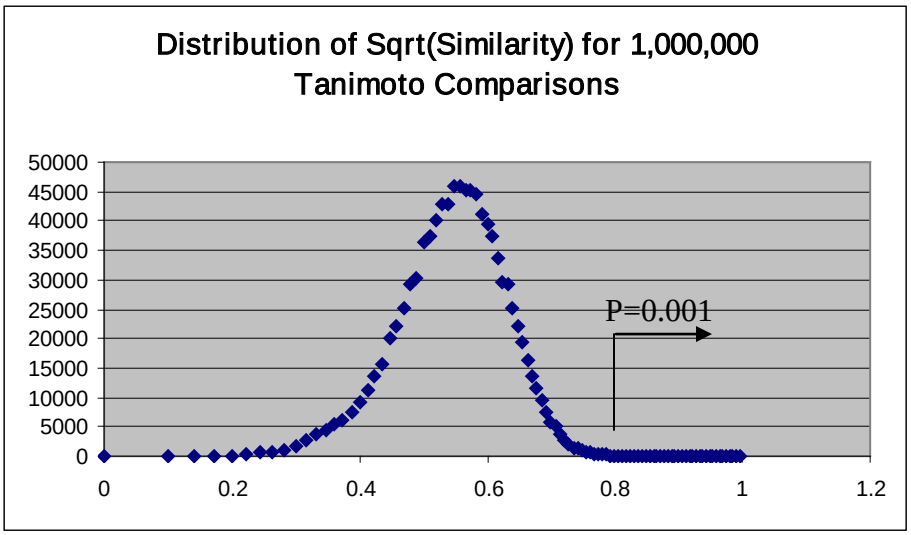


Figure 15: Distribution of Tanimoto scores

The distribution of 3DMI scores, a distance metric where 0 represents identity and scores may go up as high as 3 to 5, is shown in Figure 16. To achieve normality in the distribution, we needed to use the square-root of the 3DMI distance; so in fact, it is that distribution that is shown. Lower scores represent the best matches, so a square-rooted score of 0.3 is identified as the threshold for a P-value of 0.001.

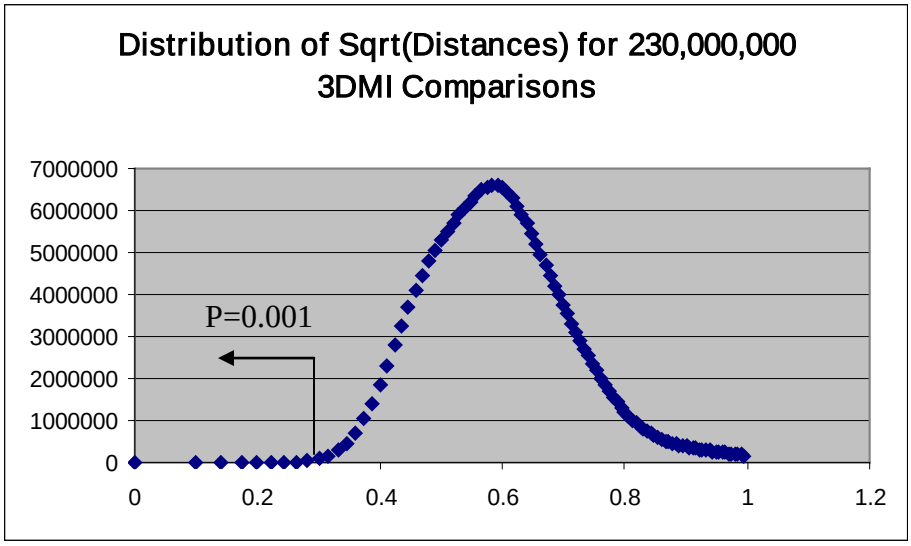


Figure 16: Distribution of 3DMI scores

6.2. Significant Gene Finder

This tool extracts the most significant set of genes for a list of experiments relative to vehicle controls. Alternatively, another list of experiments can be used as an optional control set. The analysis extracts genes that are most consistently perturbed within the set of experiments. The optional control set can be used when determining differences between experiment sets of related compounds.

The statistical significance is determined using student t-test based on customer defined p value cutoff. The resulting gene list will be displayed with a score that indicates how consistently the expression differs from the control value (or zero if no control) and the direction of the expression change.

6.3. Most Variable Probe Set

This probe set is limited to those that are most variable in mean log ratios across all treatments in the appropriate tissue. From a complete gene expression file that lists mean log ratios for all probes and experiments, the standard deviation of mean log ratios for each probe across all experiments with valid ratios is computed. This results in a standard deviation value for each probe. These standard deviations are sorted in descending order, and the gene expression data across all treatments for the top user-defined number (usually 1000) of

probes with the highest standard deviations is retained in a separate file for import and analysis into the clustering software application.

This functionality is available in the Expression Experiments -> Genes Conversion tool in DrugMatrix by choosing the standard deviation option.

6.4. Pathway Analysis

6.4.1. Pathway Curation

The suite of DrugMatrix pathways, by design, is a companion to the gene expression data that makes up the bulk of the DrugMatrix database. The purpose of the DrugMatrix pathways is to help us better understand, at the molecular level, the action and the consequences of the action of an individual or a group of compounds in living mammals. Thus, pathways associated with mammalian metabolic functions known to be affected by drug treatments including: 1) well defined signal transduction cascades, 2) biosynthesis or catabolism of key signaling or building block molecules, and 3) regulation of specific biological processes or responses, were selected for curation. Each DrugMatrix pathway is a functionally connected metabolic response or process. Details of concurrent or alternative processes of a complex physiological response are often curated into separate pathways. Where ever relevant, and whenever possible, physiological responses as well as known drug actions are included in the pathway drawings.

After selecting a pathway theme for curation, an in-depth literature search is conducted and both review articles of the field, and immediately relevant primary literature were carefully reviewed. The initial focus is on key genes well documented to be associated with the metabolic function of interest, and simple connection maps are constructed to include how the key genes expression affect the function and/or expression of each other, and the physiological/pathological consequences of such interactions. In the following step, genes or gene products that interact directly with the above key molecules, for the metabolic function in focus, are added to the map. In the final step, genes, gene products and signaling cascades involved in the regulation of the key molecules for the above metabolic function are added. Additional literature search are usually necessary for completing a pathway after constructing the initial simple connection map.

Existing pathways are revised as new or missing information was gathered during the curation of new pathways. Current articles in Science, Nature, Nature Review Drug Discovery, Nature Medicine, Cell, and Cell Metabolism are monitored routinely for new pathway themes, and to keep the existing pathways up to date.

6.4.2. Pathway Visualization

There are currently 135 pathways that have been internally curated and integrated into the DrugMatrix toolbox. Genes on each pathway are mapped,

where possible, to corresponding probes in the microarray platforms. Users can visualize the pathways with gene expression data of the treatment of interest overlaid onto the visualization (see Figure 17). Each colored circle or diamond represents a unique Unigene or LocusLink ID, and the color is determined by the average log ratio of all probes that map to this Unigene or LocusLink cluster, where shades of red represent gene up-regulated and shades of green represent gene down-regulated. Significantly induced or repressed genes, with a significance level of p<0.05, are indicated with a diamond.

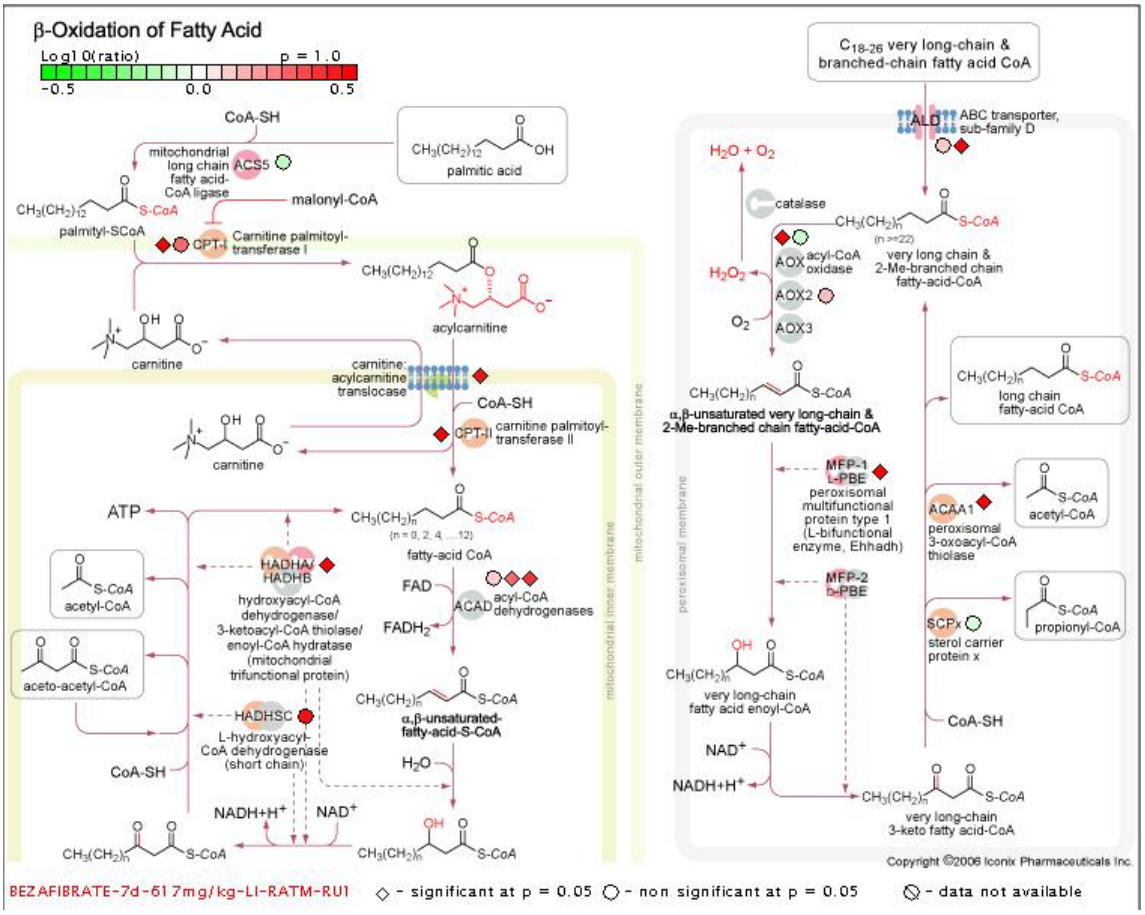


Figure 17: Example of pathway visualization

6.4.3. Pathway Impact Analysis

In addition to pathway visualization, DrugMatrix™ offers an analysis tool that calculates the impact a particular treatment has on a given pathway using significantly induced, repressed or perturbed genes at a given p value. This tool allows the user to determine quickly whether the treatment of interest has an effect on a pathway with regard to gene expression changes.

In the pathway impact matrix, statistical enrichment of significantly impacted genes at a given p value in a pathway is determined using a hypergeometric distribution metric as described below. The universe is defined as the genes on

the relevant platform with pathway association. No multiple testing correction is applied. The pathway impact is expressed as a –log (p-value), the impact score, which reflects the likelihood a treatment of interest perturbing that pathway. An impact score between 1.3 and 2, corresponding to a p-value of 0.05 to 0.01, represents a weak impact on the pathway. An impact score greater than 3, corresponding to a p-value of 0.001, indicates a strong impact, while an impact score between 2 and 3, represents moderate impact on a pathway.

Hypergeometric enrichment analysis

In probability theory and statistics, the hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of n draws from a finite population without replacement.

For example, on a microarray that contains N genes, D are involved in a particular user defined category, a pathway, or a GO term, or a structure activity class. A treatment of interest significantly induced n genes on this microarray, of which k are involved in the category described above. The hypergeometric distribution describes the probability that in a number of n genes induced by the treatment exactly *k* genes are within the defined category by random chance:

$$f(k; N, D, n) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$

The smaller the probability value, the more likely it is that the relative enrichment of n genes in the defined category is significant.

6.5. GO Data Query

DrugMatrix™ offers an analysis tool that determines whether a subset of genes perturbed by a given experiment is enriched in any GO (gene ontology) terms in the three main GO categories: biological process, molecular function and cellular component. This tool allows the user to determine whether a particular treatment affects a particular gene ontology term through transcriptional gene expression changes.

The GO analysis is conceptually very similar to the pathway impact analysis. The probability of the coincidence of the genes within the selected subset and the genes in each GO term is determined by the hypergeometric distribution. The input parameters include the universe set of genes, the query subset of genes, and the subset of genes in the GO term and the intersection of the two gene subsets. The universe is defined as the genes on the relevant platform (e.g. CodeLink™ RU1 rat array) that have gene ontology annotation.

Because the analysis typically makes many tests, a multiple test correction must be done. The Bonferroni multiple test correction is applied to each GO term

probability value. The parameter for this correction is not constant across all GO terms, but is adjusted depending on the number of nodes at and above the node being tested. The score is the adjusted probability value that the genes that intersect could be found by chance.

6.6. Pattern Analysis

Patterns are lists of weighted Genes (at the expresson, or cross-platform mapping level) that can be used to search across the Expression Experiment domain for experiments that have similar expression pattern across a specified list of genes. Patterns can be created in one of two ways, either by using the Pattern Creator tool, or by performing a series of steps using the Export and Import Pattern tools. The Pattern Creator tool takes as input an experiment list and a gene list. It takes either a mean or median log ratio value across those experiments specified for each gene in the gene list, and uses those as the weighted values of the pattern. The log ratio values can be unfiltered, or filtered by p-value. The Export Pattern tool takes as input either a gene list or a pre-existing pattern and converts that into a tab delimited file that can be easily edited using Excel. The Import Pattern tool takes a tab delimited file and uses it to create a pattern list in the user's workspace. Once created, a pattern can be used in conjunction with the Pattern Search tool to search across the expression experiment domain. Pattern Search tool takes as input a single pattern list, and uses that pattern list to search across the entire DrugMatrix Expression Experiment domain. The results returned by this search is a list of experiments and a corresponding correlation coefficient score which describes the correlation between the expressed genes in the experiment and the weighted genes of the pattern. The search may be performed in Trinary mode. This sets both weighted values and expression data into a trinary state represented by 1, 0, or -1.

Figure Legend

Figure 1 Histogram showing the frequency distribution of percent inhibition data on 107,478 measurements from 861 compounds tested in 130 molecular pharmacology assays. This data was collected over the course of several years.

Figure 2 Histogram showing the frequency distribution of IC50 data in LogμM on 3,845 measurements from 506 compounds tested in 130 molecular pharmacology assays. This data was collected over the course of several years.

Figure 3 Histogram showing the frequency distribution of the molecular pharmacology target inhibition by 156 compounds tested in 123 assays. Shown on Y-axis is the number of compounds that caused higher than 70% inhibition of the molecular pharmacology target in each of the 123 assays. A total of 637 compound-assay combinations showed more than 70% inhibition. The 156 compounds were randomly chosen.

Figure 4 Histogram showing the frequency distribution of the molecular pharmacology target inhibition by 156 compounds tested in 123 assays. Shown on Y-axis is the fraction of the 123 assays that exhibited higher than 70% inhibition for each of the 156 compounds. A total of 637 compound-assay combinations showed more than 70% inhibition. The 156 compounds were randomly chosen.

Figure 5 Model for genomics data in DrugMatrix explaining the relationship among probe, expresson and gene.

Figure 6 Diagram depicting the strategies for gene name derivation.

Figure 7 Diagram depicting the strategies for Gene Ontology term annotation.

Figure 8 Examples of bad probes on CodeLink RU1 array showing the impact of laboratory protocol change on probe response. The normalized signal for the probe averaged across 200 control arrays (Y-axis) were plotted against the hybridization date of the target to the array (X-axis). Dates on the top of the graphs show the initial date on which arrays were first hybridized, followed by three different dates on which distinct laboratory protocols were implemented. A) Probe shows a 4 to 6-fold difference in normalized signal between the first protocol used before 12/10/2001 and the other protocols implemented after 10/12/2001. B) Probe shows a 10-20-fold difference in normalized signal after 12/17/2001, when the second protocol change was implemented.

Figure 9 Density plot in three-dimension showing the distribution of signal intensities for house-keep genes verses probes on the CodeLink RU1 array.

Figure 10 Scatter plot showing the effect of normalization on array signal intensity relative to the reference template. Shown is the log signal for the test array on Y-axis and the reference template on X-axis. The reference template is constructed by computing the median log signal for each probe for a set of tissue, vehicle and time matched control arrays. A) pre-curve fit. B) post curve fit.

Figure 11 Principle component analysis of log gene expression signal for control and drug-treated arrays. Left is the plot for 1,697 control arrays and right is the set of 9,300 drug-treated arrays.

Figure 12 Scatter plot of number of significantly changed genes across experimental date, illustrating the process drift. Number of significantly changed genes was determined using the same control set as denominator against 1102 experiments corresponding to 183 compounds, for which gene expression data were collected across a large span of experimental date.

Figure 13 Scatter plot of number of significantly changed genes computed using control sets in different laboratory protocol periods, illustrating the effect of protocol change on gene expression data ratio calculation. Number of significantly changed genes is determined by P value cutoff of 0.001. Control sets grouped by different laboratory protocol period were colored differently.

Figure 14 Diagram demonstrating the derivation of signature projection score.

Figure 15 Histogram showing the distribution of Tanimoto scores across the 1,000,000 compound comparisons. Y-axis represents the number of compound comparisons having a specific Tanimoto score and X-axis represents the Tanimoto score.

Figure 16 Histogram showing the distribution of 3DMI scores across the 230,000,000 compound comparisons. Y-axis represents the number of compound comparisons having a specific 3DMI score and X-axis represents the square-root of 3DMI distance.

Figure 17 Example of pathway visualization showing the gene expression changes induced by bezafibrate treatment at 617 mg/kg at 7 days in the liver as visualized on the beta oxidation of fatty acids pathway. Gene responses are represented by the colored spots, where red indicates up-regulation and green indicates down-regulation. Significantly changed (p value <0.05) genes are represented by diamonds.