Novel Informatic Approaches to Analyze Gene Expression Data with the ToxCast™ 320 Chemical Library in Cultures of Primary Human Hepatocytes

Andrew Beam[1], Daniel Rotroff[3], Kimberly Freeman[1], Adam Farmer[1], Howard Bondell[2], Keith Houck[3], Richard Judson[3], David Dix[3], Edward LeCluyse[1], Stephen Ferguson[1]

[1] CellzDirect/Life Technologies Corporation

[2] N.C. State University – Department of Statistics

[3] U.S. EPA National Center for Computational Toxicology (NCCT)

Abstract #: 1093

Poster Board #: 311

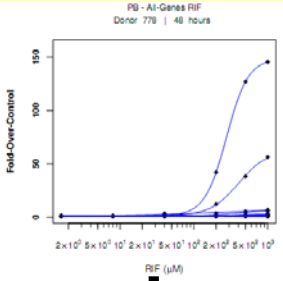
Introduction

The U.S. Environmental Protection Agency (EPA) is conducting a large scale research program called ToxCast™ to assess the potential toxicity to humans *in vivo* by collecting a large number of predictive assays *in vitro*. Cultures of primary human hepatocytes have been shown to be a viable *in vitro* model for human liver function. Traditional approaches such as microarrays allow researchers to assess a compound's effect on hepatocytes by measuring changes in protein expression levels. However, these approaches often neglect to incorporate full concentration and time related responses due to limitations in throughput and sensitivity.

We have developed a high throughput assay suite using primary human hepatocyte cultures as i*n vitro* models that retain liver-like functionality to generate more comprehensive data across time and concentration. Using the ToxCast™ 320 chemical library, mRNA expression was determined using a quantitative nuclease protection assay using the Omix™ Imaging System (High Throughput Genomics, Inc., Tucson, AZ). Fourteen gene targets representing Phase I/II metabolism and transport were monitored based on their role in liver function and sensitivity to receptor pathways (AhR, CAR, PXR, PPARα, FXR).

Techniques from machine learning were used to cluster compounds by gene response profiles. Dose-responses were mathematically abstracted as vectors in multidimensional space (rather than classical scalar representations traditionally associated with standard microarray analyses) and used in algorithms such as K-means and algometric clustering to create representative chemical phylogenies. Unique to this approach is assessment of concentration-response changes over time (6, 24, 48 hr in culture) as well as correlation of gene targets with one another. From these analyses, inclusion of data from all time points resulted in more accurate clustering of the replicate ToxCast™ 320 and reference chemicals that reduced donor-dependent variability. This approach has significant implications in standardizing primary hepatocyte data analysis across donors and profiling chemical response with *in vivo* endpoints.

Figure 1: Vector Abstraction



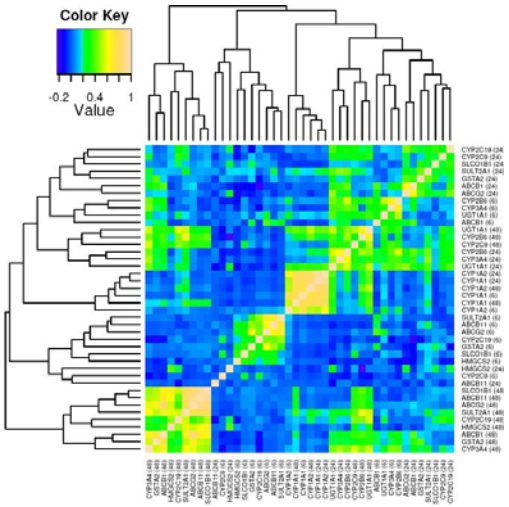
$$\langle ABCB1_{x1}, ABCB1_{x2}, ABCB1_{x3}, ABCB1_{x4}, \ldots UGT1A1_{x3}, UGT1A1_{x4}, UGT1A1_{x5} \rangle$$

Each compound's entire dose response for all gene targets across 3 time points was abstracted as vector in 70 dimensional space (14 genes x 5 concentrations x 3 time points = 70 dimensions). By using this approach two compounds' relative "dissimilarity" can be calculated by computing the Euclidean distance between their vector representations. Each of the 70 dimensions may also be Z-scored (a.k.a. normal score) to reduce the effects of large responses (e.g. induction of CYP1A1 vs. SUL2A1) and different donors may be Z-scored independently to reduce donor variability. This approach also lends itself to adding new information later as new, independent dimensions may be concatenated to the original vector.

Results and Conclusions

- A compound's dose-response can be abstracted mathematically as a vector
- Gene to gene correlations can provide insights into previously unknown regulation as well as validating an assay by confirming correlations known in current literature.
- Utilizing all of the information for a compound's response, as opposed to a scalar statistic such E_{max} or EC_{50} , gives a more robust and accurate representation of a compound's performance.
- The most accurate clustering, measured by the proximity of replicate compounds to each other, was achieved by using the vector representation and incorporating all three time points.

Figure 2: Gene to Gene Correlations

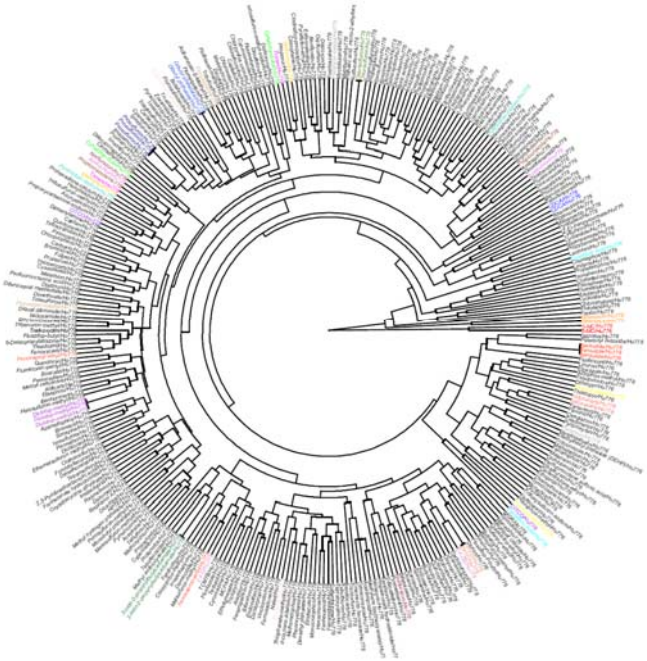


A heatmap representing the Pearson correlation between all 14 gene targets across 3 time points. The statistic used in calculating the correlation was the dynamic range (E_{max} – E_{min}) of a gene's response. The table below contains the ten most correlated gene targets.

Ten Most Correlated Gene Targets

Gene 1 (hr)	r	Gene 2 (hr)
CYP1A1 (24)	.9445	CYP1A2 (24)
ABCB11 (48)	.9053	SLCO1B1 (48)
ABCG11 (48)	.8859	ABCG2 (48)
ABCG2 (48)	.8645	SLCO1B1 (48)
CYP1A1 (24)	.8610	CYP1A2 (48)
ABCB1 (48)	.8515	ABCG2 (48)
CYP1A1 (6)	.8312	CYP1A2 (24)
CYP1A2 (24)	.8220	CYP1A2 (48)
CYP1A1 (24)	.8159	CYP1A1 (48)
CYP1A1 (6)	.8009	CYP1A1 (48)

Figure 3: Clustering and Dendrograms



A dendrogram of the ToxCast™ 320 chemical library created using the vector abstraction of each compound's dose-response for all 14 gene targets across all three time points. Different donors were Z-scored independently, to reduce both inter-donor variability and bias towards more efficacious genes. Incorporating all three time points into a 210 dimensional space (70 dimensions per time point x 3 time points = 210) resulted in more accurate clustering of EPA replicates that were blinded during the study as well as the positive controls. Z-scoring donors independently also drastically reduced donor dependence; without it the clustering results were greatly affected by the donor in which the compound was tested. Corresponding replicates and positive controls are given the same color.

Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy.