# Picking Cell Lines for High-Throughput Transcriptomic Toxicity Screening

R. Woodrow Setzer, Russell S. Thomas, National Center for Computational Toxicology, US Environmental Protection Agency, Research Triangle Park, NC

High throughput, whole genome transcriptomic profiling is a promising approach to comprehensively evaluate chemicals for potential biological effects. To be useful for *in vitro* toxicity screening, gene expression must be quantified in a set of representative cell types that captures the diversity of potential responses across chemicals. The ideal dataset to select these cell types would consist of hundreds of cell types treated with thousands of chemicals, but does not yet exist. However, basal gene expression data may be useful as a surrogate for representing the relevant biological space necessary for cell type selection. The goal of this study was to identify a small (< 20) number of cell types that capture a large, quantifiable fraction of basal gene expression diversity. Three publicly available collections of Affymetrix U133+2.0 cellular gene expression data were used: 1) 59 cell lines from the NCI60 set; 2) 303 primary cell types from the Mabbott et al (2013) expression atlas; and 3) 1036 cell lines from the Cancer Cell Line Encyclopedia. The data were RMA normalized, log-transformed, and the probe sets mapped to HUGO gene identifiers. The results showed that <20 cell lines capture only a small fraction of the total diversity in basal gene expression when evaluated using either the entire set of 20960 HUGO genes or a subset of druggable genes likely to be chemical targets. The fraction of the total gene expression variation explained was consistent when evaluated using either linear combinations of genes and cell lines or by centroids of clusters of cell lines. Alternatively, only ten cell lines are required to capture the diversity in the fraction of genes that are highly expressed in at least one of the cell lines. This analysis demonstrates the challenges in profiling the entire genome based on a limited number of cells. Future efforts are needed to explore alternative approaches to selecting representative cell lines that adequately cover relevant biological space for high-throughput transcriptomic toxicity screening. *This abstract does not necessarily reflect U.S. EPA policy.*