The Tox21 Phase III "1500 Genes" High Throughput Transcriptomics Project

Project Leader: Richard Paules, PhD, NIEHS/DNTP/Biomolecular Screening Branch

A major challenge for understanding the effects of environmental exposures on human health is the lack of sufficient safety or toxicological information for the tens of thousands of chemical compounds that are currently in use, even though significant numbers of people are exposed to them in measurable levels. Both health care providers and regulatory authorities need toxicological data on these compounds in order to characterize the potential health risk to exposed populations.

Traditional toxicological methods rely largely on *in vivo* animal testing in order to assess the risk to human health associated with chemical exposures. However, this approach is both expensive and time-consuming and the findings are often difficult to extrapolate to effects on human health. In 2004, through its *Vision and Roadmap for the 21st Century* document, the NTP proposed using high throughput biochemical- and cell-based *in vitro* assays to rapidly, and at reduced cost, gain valuable toxicological information about thousands of compounds. Support for this approach was exemplified in the 2007 National Research Council (NRC) report *Toxicity Testing in the 21st Century: A Vision and a Strategy*. In 2008, to accelerate the implementation of NTP's and NRC's vision and strategy, the NIEHS/NTP, the U.S. EPA's National Center for Computational Toxicology (NCCT), and NHGRI's NIH Chemical Genomics Center (NCGC, now within the National Center for Advancing Translational Sciences or NCATS) entered into a formal Memorandum of Understanding partnership on "High-Throughput Screening, Toxicity Pathway Profiling and Biological Interpretation of Findings," with the U.S. FDA joining this collaborative effort in 2010.

Informally referred to as Tox21 (for Toxicology in the 21st Century), the goals of this collaboration are to 1) identify patterns of compound-induced biological responses in order to characterize toxicity/disease pathways, to facilitate cross-species extrapolations, and to model low-dose effects; 2) prioritize compounds for more extensive toxicological evaluation; and 3) develop predictive models for biological responses in humans. The initial phase of Tox21 involved proof-of-principal experiments to demonstrate the feasibility of using a high throughput screening (HTS) approach for toxicity testing and involved the screening of approximately 2500 compounds in 140 quantitative (q)HTS assays representing 77 predominantly cell-based reporter gene endpoints. Phase II of Tox21, currently in progress, involves the screening of over 10,000 compounds at 15 concentrations in 30 assays, with each assay run three times. The assays were prioritized to target known stress response pathways (including pathways such as oxidative stress, genotoxic stress, endoplasmic reticulum stress, mitochondrial stress, inflammation, etc.) and nuclear receptor signaling responses

(including signaling by such receptors as the estrogen receptor α, androgen receptor, thyroid hormone receptor β, peroxisome proliferator-activated receptor γ, glucocorticoid receptor). The Tox21 Phase I and Phase II data collected to date are publicly available on PubChem. While these data undoubtedly provide valuable information on potential adverse effects associated with exposures to compounds for which little or no toxicological information exists, these qHTS assays are limited in terms of biology.

Proposed activities in the recently initiated Tox21 Phase III include a focus on high-content imaging and mid- to high-throughput gene expression screens using a variety of normal human cells and cell lines, including metabolically-competent human cells. In particular, a project was initiated to capture information from the whole transcriptome (i.e., the entirety of all expressed RNA molecules in a cell or biological sample) using a targeted subset of genes in a HTS or semi-HTS assay to gain insight into how biological systems respond to chemical exposures. This project is being referred to as the “1500 Genes High Throughput Transcriptomics Project,” although neither the actual number of genes to be utilized nor the specific transcriptomics platform to be used has yet to be determined.

The transcriptome of a biological system is dynamic, changing in composition in response to endogenous and exogenous factors. As such, the transcriptome can be considered an integrated product of all factors that might impinge on the genome to alter gene expression and thus a read-out of the physiological or pathophysiological status of a biological system. These factors include the function of gene activators, repressors, silencers, DNA and histone modifications, epistatic interactions among genes, feedback mechanisms, and the entire genetic background of a particular biological system, tissue, organ, or individual. Since the transcriptome reflects the current condition of a biological system, it has the potential to reveal responses to chemical exposures such as molecular initiating events associated with acute toxicity or adverse outcome pathways, adaptive changes, or irreversible “point-of departure” alterations involved in toxicities and disease development.

One aspect of the transcriptome is that, even though the human genome has over 20,000 genes that potentially can be transcribed as unique messenger RNA (mRNA) molecules (not counting splice variants or alternately processed mRNAs), these mRNAs are not all independently expressed but, in fact, many are co-regulated or coordinately expressed. These coordinately regulated mRNAs often encode proteins that function in biological modules generally referred to as networks or pathways. The ultimate physiological or pathophysiological status of a biological system is largely the consequence of interactions between these modules or networks. For example, two major networks center around the p53 and the NF-κB proteins and it is the interaction

and balance between the signals generated by these two networks that can drive a cell toward either cell cycle arrest or programmed cell death. The degree of perturbation, that is the direction and magnitude of modulation, of members of networks in response to chemical exposures is critical information that can determine the biological and toxicological outcome of an exposure.

Based on this coordinate regulation of gene expression, Todd Golub and Justin Lamb at the Massachusetts Institute of Technology and their colleagues hypothesized that a connection could be made between biological perturbations and treatments such as exposures to particular chemicals or "perturbagens" (i.e., a chemical, drug, or genetic manipulation that would perturb the normal homeostasis of a biological system), using the transcriptome as a means to link those two endpoints. The result of their initial work, utilizing publically available human gene expression data as well as their own data, was published as the "Connectivity Map" (Lamb, *et al.* 2006). Furthermore, Golub and colleagues hypothesized that for most networks mRNA expression levels of key representative genes (i.e., "landmark" genes) in a network could be used as surrogates for all members of that network. If correct, measuring the expression levels of an appropriate subset of landmark genes could be used to impute whole transcriptome information. Through a series of subsequent bioinformatic analyses, Golub, Lamb and colleagues were able to determine that a subset of approximately 1000 landmark genes used in a Luminex bead-based assay of RNA transcripts (referred to as the "L1000") was able to capture approximately 80% of the linkages between chemical and genetic perturbations and biological outcomes.

Golub, Lamb and colleagues have extended their studies by utilizing the L1000 platform in a number of studies that are part of the NIH Common Fund's *Library of Integrated Network-based Cellular Signatures (LINCS)* program and currently have collected over 1.4 million gene expression profiles. The goal of the LINCS program is to develop a "library" of molecular signatures that describe how different types of cells respond to a variety of perturbagens. One attractive aspect of the L1000 technology is its relatively low cost (i.e., the L1000 platform provides imputed whole-genome expression coverage of the transcriptome for approximately 1/10th of the current cost of a whole genome microarray chip analysis and approximately 1/20th or less of the current cost of Next Generation RNA-Sequencing analysis of the transcriptome). The development of transcriptomic technologies that exploit a surrogate subset of the transcriptome and are amenable to HTS approaches opens up the possibility of obtaining whole transcriptome information on multiple cells and tissues from multiple species exposed in a concentration- and time-dependent manner to thousands of perturbagens. This is the impetus for the initiation of the Tox21 "1500 Genes" Project.

The Tox21 "1500 Genes" Project

The selection of an appropriate subset of key representative or "sentinel" genes is critical to the success of the "1500 Genes" Project. Desired characteristics of the Tox21 sentinel "1500 Genes" set (i.e., the "S1500") would be to provide information useful to understanding mechanisms of adverse effects from exposures to perturbagens, as well as to provide health scientists and regulators with critical information to assist in hazard identification and human risk assessment. Therefore, the NTP solicited input from the scientific community through a **Federal Register** (FR) notice published on July 29th, 2013 requesting the "Nomination and Prioritization of Environmentally Responsive Genes for Use in Screening Large Numbers of Substances Using Toxicogenomic Technologies." The FR notice elicited 17 responses from a variety of sectors. This was followed by a workshop held at NIEHS in September 2013 with invited speakers and public participation in order to present and discuss ideas about the gene prioritization criteria to be use during the gene selection process. Subsequent to the workshop, an interagency working group composed of members of the Tox21 community was established to consider the input provided toward the selection criteria to be used in the gene selection process and to move forward with the gene selection process. The working group has arrived at the following consensus strategy.

"1500 Genes" Selection Strategy: The goal of this effort is to select and evaluate a gene set that will be used to query gene expression as a "reduced representation" of the whole-genome transcriptome. This sentinel gene set will be implemented using a high throughput, cost effective transcriptomics technology (i.e., the "S1500 Platform"). The S1500 gene set will be used to measure whole-genome effects in a comprehensive variety of biological contexts, such as treatment of a particular cell type with a variety of chemicals or treatment of a variety of different cell types with a set of chemicals, etc. A primary use will be to screen chemicals for toxicological research prioritization. The actual number of genes to be used in the S1500 will depend on the technical limitations of the selected transcriptomics platform and will likely increase as technologies improve.

The S1500 gene set should have the following attributes:

1. It captures maximal expression variability and dynamics: This will ensure that perturbagens will trigger changes in at least some of the S1500 genes.
2. Extrapolatability: This property refers to the ability to extrapolate or predict with some accuracy the expression changes in all genes from those observed in this reduced set of sentinel genes.

3. Maximal coverage of pathways based on inclusion of specifically selected genes to ensure maximal biological pathway coverage. Much of biological knowledge is described or understood in terms of pathways, so information on pathway genes can help scientists interpret and make use of the data generated using the S1500.
4. Selective inclusion of toxicity and disease related genes: Specific genes will be selected for their reported roles in toxicity-related and disease-related processes, with the expectation that they will prove to be useful in toxicity and risk assessment evaluations.

Genes that satisfy attributes 1 and 2 will be derived by data-driven, unsupervised, bioinformatics approaches while genes that satisfy attributes 3 and 4 will be derived by expert-driven, supervised approaches. An additional desirable criterion is that the S1500 incorporate all, or a large fraction of, the LINCS L1000 gene set in order to facilitate comparison and possible integration of the S1500- and L1000-derived data sets.

S1500 selection efforts will focus initially on human genes and will be used to test perturbations in human cells and tissues. Subsequently, similar gene sets will be selected for mouse, rat, zebrafish, and *Caenorhabditis elegans*, with weighted consideration for orthologous genes.

Proposed gene sets will be evaluated using an objective, informatic approach that produces a weighted sum of terms relating to criteria 1-4. The L1000 gene set will serve as a baseline gene set against which other gene sets will be compared.

Gene sets will be built and tested using existing whole-genome microarray data. Specific data sets using the same technology (e.g., the Affymetrix human chip) will be utilized, splitting them into training and test sets. Evaluations will be performed that will include extrapolatability, maximal coverage of pathways, connectivity with responses to related perturbagens, etc.

The “1500 Genes” Selection Project Workgroup Participants:

Scott Auerbach, Biomolecular Screening Branch, DNTP, NIEHS
Pierre Bushel, Biostatistics Branch, DIR, NIEHS
Jennifer Collins, Exposure, Response & Technology Branch, DERT, NIEHS
Agnes Forgacs, National Center for Computational Toxicology, US EPA
David Gerhold, Genomic Toxicology Group, National Center for Advancing Translational Sciences (NCATS)

Richard Judson, National Center for Computational Toxicology, US EPA
Elizabeth Maull, Biomolecular Screening Branch, DNTP, NIEHS
Alex Merrick, Biomolecular Screening Branch, DNTP, NIEHS
Rick Paules, Biomolecular Screening Branch, DNTP, NIEHS
Ruchir Shah, Social & Scientific Systems, Inc.
Deepak Mav, Social & Scientific Systems, Inc.
Dan Svoboda, Social & Scientific Systems, Inc.

References

J. Lamb, *et al.* 2006. *Science*, **313**, 1929-1935.
Federal Register. **Vol. 78, No. 145**, 45542-45543. July 29th, 2013.