Structured Application of Biological Ontologies to Annotate High-Throughput Screening Assays and their Targets of Activity

Jimmy Phuong

A technical report submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Masters of Science in Public Health in the Department of Environmental Sciences and Engineering.

Chapel Hill

2014

Approved by:

Ivan Rusyn, M.D., Ph.D.

Rebecca Fry, Ph.D.

Matthew Martin, Ph.D.

Abstract

Jimmy Phuong
Structured Application of Biological Ontologies to Annotate
High-Throughput Screening Assays and their Targets of Activity
(Under the direction of Matthew Martin)

High-throughput screening (HTS) assays have changed the pace of chemical data collection, enabling assessments at various levels of biological relevance. EPA's ToxCast Program has 328 assays (experiments) generating 541 assay components (readouts), which produces 795 assay component endpoints (analyses), with intentions to increase the number of assays and the number of substances tested. As new assays are developed, it becomes a challenge to communicate what kind of data and features are associated with each assay. This report uses the BioAssay Ontology and other publicly available ontologies to produce the ToxCast Assay Annotation, a structured resource for descriptive information that uses controlled vocabulary to aid in the communication and use of ToxCast HTS assay data. Organized by 34 annotations including 'assay design type' and 'detection technology type', this structure allows for a concise reference to the pertinent attributes of an assay. Additionally, the perspective differences between the technological and intended target are separately captured. This structured annotation also allows for the identification of comparable ToxCast assay endpoints, and offers the potential to link with other HTS data repositories.

Acknowledgements and Dedication

I would like to give thanks to my mentor, Dr. Matt Martin, for his advice and insight which has allowed me to develop as a scientist. Thank you sincerely for giving me the opportunity and creative freedom to explore an area of science that is under-developed and challenging. I'd also like to thank Drs. Ivan Rusyn and Rebecca Fry for their support while planning and executing my Masters work-plan.

A special thank you goes out to all the past and present project team members at US EPA ORD-National Center for Computational Toxicology that I had the pleasure of interacting with. Without their diverse expertise, time, and patience, this highly integrative work could not have been completed.

I dedicate this work to my family and friends, whose warmth and support, numerous health concerns, and out-of-the-box thinking have driven me to view science from different angles.

Table of Contents

Abstract iii

Acknowledgements and Dedication iv

List of Figures vi

List of Tables vii

List of Abbreviations viii

Chapter 1: Literature review 1

Chapter 2: Introduction 8

Chapter 3: Methods 11

Chapter 4: Results 21

Chapter 5: Discussion, Conclusions, Limitations, and Future Directions 40

Chapter 6: Indoor environmental health sampling, Science talk panels,
and mammary gland tumor bioinformatics investigation:
the Practicum experience with Silent Spring Institute 48

Appendix 53

References 63

List of Figures

Figure 1: The annotation workflow 13

Figure 2: PCA workflow of the chemical screening data 19

Figure 3: The annotation structure 22

Figure 4: Intended target family annotation terms 34

Figure 5: Heat map of the PCA loadings clustered by the annotation terms for the first five principle components 38

List of Tables

Table 1: The 37 annotations used and the short description of the concepts they capture 15

Table 2: Assay information and content readout types.....24

Table 3: Assay design types annotated to ToxCast Assay Components Endpoints 25

Table 4: Detection technology types annotated to ToxCast Assay Components Endpoints 27

Table 5: A comparison of the assay design subtypes by the detection technology subtypes 27

Table 6: Reagent and components information for the APR_HepG2_1hr assay 29

Table 7: Organism/tissue types and cell format types 30

Table 8: Comparison of technological and intended target types 31

List of Abbreviations

AC50.....50% of the maximal activity concentration

ACEA.....Acea Biosciences

ACToR.....Aggregate Computational Toxicology Resource Database

APR.....Apredica

AR.....Androgen receptor

ATG .....Attagene

BAO .....BioAssay Ontology

BAO-GPCR .....G-Protein coupled Receptor Ontology

BARD .....BioAssay Research Database

BSK.....Bioseek

CL .....Cell Ontoloy

CLO.....Cell Line Ontology

CTD.....Comparative Toxicogenomics Database

DSSTox.....Distributed Structure-Searchable Toxicity Database

ELISA .....Enzyme-linked immunosorbent assay

EPA.....United States Environmental Protection Agency

ER .....Estrogen receptor

ESR1 .....Estrogen Receptor Alpha

ExpoCastDB .....ExpoCast Database

GPCR.....G-Protein coupled Receptor

GSEA.....Gene Set Enrichment Analysis

HCS.....High Content Screening

HTS.....High-Throughput Screening

IRB.....Institutional review board

LEC.....Lowest effect concentration

MeSH .....Medical Subject Headings

MLP .....Molecular Libraries Program

MLSCN.....Molecular Libraries Screening Center Network

NAS.....National Academy of Sciences

NCATS .....National Center for Advancing Translational Science

NCBI Taxon.....NCBI Organismal Classifications

NCCT.....National Center for Computational Toxicology

NCIT .....National Cancer Institute Thesaurus

NGO.....non-government organization

NIH .....National Institute for Health

NVS.....NovaScreen

NRC .....National Research Council

OT .....Odyssey Thera

OWL .....web ontology language

PC.....Principle component

PCA.....Principle Components Analysis

PPARA.....human peroxisome proliferator-activated receptor alpha

PPARD.....human peroxisome proliferator-activated receptor delta

PPARG.....human peroxisome proliferator-activated receptor gamma

PPRE.....Peroxisome Proliferator-activated Response Element

RTU.....reporter transcription unit

SOP .....Standard operating procedures

SSI.....Silent Spring Institute

Tox21 .....Toxicity Testing in the 21st century

ToxCastDB .....ToxCast Database

ToxRefDB.....ToxRef Database

TSCA .....Toxic Substances Control Act

Chapter 1

Literature Review

Chemical Testing Demand and High-Throughput Screening Assays

The foremost concern about the chemicals in the environment is that most are insufficiently evaluated for their bioactivity and potential hazards. Since 1976, the Toxic Substances Control Act (TSCA) inventory has registered over 66,000 chemicals manufactured in or imported into the United States (Congress 1976). TSCA mandates the United States Environmental Protection Agency (EPA) to protect the public from adverse human health or environmental outcomes downstream of these chemicals. For the past 40 years, these risk assessments used any bioactivity and adverse effect information available, which largely relied on expensive, time-consuming animal model experiments. Due to this slow pace and unevenness of chemical testing, a small fraction of chemicals becomes data-rich while the vast majority remains with little or no available data.

The National Academy of Sciences' (NAS) National Research Council (NRC) has addressed the current chemical testing paradigm with the desire to move in a direction that reduces the number of animals used, reduces the cost and testing time, and increases the mechanistic understanding of the chemical effects (NRC 2007). While encouraging the recycled use of existing *in vivo* data, these desires have turned the scientific and regulatory communities towards *in vitro* assays, particularly high-throughput screening (HTS) and high content screening (HCS) assays. Compared with *in vivo* studies, HTS and HCS assays require smaller amounts of testing space and volumes of testing material. Integrated with new methodologies such as

toxicogenomics, bioinformatics and computational toxicology, the data collected through HTS and HCS assays can more easily enable mechanistic assessments. The assays may probe human genes, cells, or tissues to reflect on how chemicals may elicit perturbations at the molecular level or cumulatively as pathway responses (Dix et al. 2007; Morisseau et al. 2009; Judson et al. 2010; Kavlock et al. 2012).

At its foundation, an assay is a manufactured test to detect perturbations away from the normal biological activity. The activity tested will be dependent on how the assay is conducted and what it measures. HTS assays are assays that have been optimized to allow simultaneous testing while reducing the cost and time expenditures. Consider a HTS assay conducted in a 384 well-plate as the optimized form of the same assay that was previously conducted in single test tubes—the data yield is in orders of magnitude faster. HTS methodologies are predominantly drug discovery approaches; however, reapplying these same approaches towards environmental chemicals can help address the number of data gaps existing for the large portion of environmental chemicals (Dix et al. 2007; Judson et al. 2009).

In response to the NRC report (2007), EPA has chosen approximately 10,000 chemicals to be considered for the ToxCast screening and prioritization program (Dix et al. 2007; Judson et al. 2009). Out of the TSCA inventory, these chemicals were selected due to medium- and high-production volumes (exceeding 10,000 lb/year), known industrial functions as pesticides actives, presence in the environment as drinking-water contaminants, or known inert chemicals (Judson et al. 2009). Some of these chemicals are data-rich, enabling a way to compare the assay results with precedent knowledge of the chemicals' activities (Martin et al. 2011). In ToxCast, testing would occur in phases; each ToxCast phase is a separate group of nominated chemical that will

be tested through a large, diverse number of HTS and HCS assays. In addition to ToxCast, EPA is an active participant in Toxicity Testing in the 21st Century consortium (Tox21), an Inter-Agency collaboration that takes a different strategy towards executing chemical tests. Using the latest in automated HTS technologies, Tox21 tests all 10,000 chemicals through small groups of assays; each Tox21 phase is a different set of assays (Huang et al. 2011). From these two programs, at each phase, ToxCast would provide a broad view of chemical activity across diverse biological endpoints while Tox21 would provide a means to rank the 10,000 chemicals using assays that target endpoints of high concern.

Data Storage

With the large number of chemicals to be tested and even larger nest of HTS data expected, data storage becomes a big factor. To list them explicitly, there would be the chemical or substance identifiers, the structural features for each chemical or substance, the plate maps for each tested chemical plate, the assay identifiers, the readout data and the analyzed data. To cover these different needs, separate databases were devoted to capture the information. Judson et al. (2012) mentions that ToxCast plans to disseminate the data storage to a number of different databases of specific function. The EPA Distributed Structure-Searchable Toxicity (DSSTox) program is dedicated to the chemical structure and linkage between chemical structures to external data sources (Williams-DeVane et al. 2009; Judson et al. 2012). The EPA ToxCast Database (ToxCastDB) would serve as the data repository for both the ToxCast assay data as well as descriptive information about each assay. As a key component for biological modeling, the EPA ToxRef Database (ToxRefDB), a database devoted to systematic curation of *in vivo* experimental outcomes, would be the anchorage point between the ToxCast chemicals and the

historic toxicity endpoint information that may be available for them (Martin et al. 2009). Similarly, the EPA ExpoCast Database (ExpoCastDB) stores data pertinent for exposure and environmental presence modeling. Separately, the EPA Aggregate Computational Toxicology Resource (ACToR) database combines the information from each of these EPA databases and to other publicly available data sources. In general, these databases can communicate or be queried via chemical structure and identity. This is a chemical-centric, test substance oriented point of view, which is not developed for assay-centric or target-centric options.

Assay Terminology

With the push for advancing chemical testing, more questions and challenges about HTS and HCS assays arise. The language for different aspects of *in vitro* assays was not formally established to enable assay-centric search options. Perhaps more pertinent to ACToR and ToxCastDB, this area of assay terminologies gets revisited when trying to communicate similarities and differences between assays. For instance, protein assays were previously synonymous with binding assays. Now, with new assay technologies developed to probe different facets of protein function, a protein assay seems vague. Within ToxCast, a protein assay could now mean enzyme-substrate reactions, receptor-ligand binding, protein expressions by enzyme-linked immunosorbent assays (ELISA), changes in protein-protein interaction, or even a marker protein for cytotoxicity or a pathway response (Kavlock et al. 2012). This area could continue to propagate as new, abstract ways to consider a protein's biological processes and systems biological impact are developed.

Beyond the assay technology, the content readout has shifted towards more multiplexed and multiparametric approaches. A single assay could now be equipped to interrogate a battery

of targets (Romanov et al. 2008; Houck et al. 2009; Giuliano et al. 2010; Martin et al. 2010; Rotroff et al. 2010). This is the real challenge: communicating to the general public what the readouts of increasingly complex assays are with respect to the data already available from the previous generation of single readout assays. This inherently demands a definition for the minimum amount of information for an assay, for which there currently is not an agreed upon standard across different technologies (Visser et al. 2011).

Moreover, different HTS campaigns have their own approaches towards describing and categorizing their *in vitro* assay libraries. The ToxCast program purchases testing data for the ToxCast chemical sets from various contract vendors, who have the technology and expertise to perform patented assay protocols. Some of the past ToxCast publications emphasized the biomedical innovation from assays purchased from different contract vendors while focusing on utility of the data for modeling chemical and biological endpoints (Judson et al. 2010; Martin et al. 2011; Sipes et al. 2011; Kleinstreuer et al. 2013). Hence, the mechanics, biological innovations, and utility behind each HTS assay were separately explained with variations in the terminology. A similar scenario can be seen of the Molecular Libraries Program (MLP), an NIH funded HTS campaign that began in 2003. MLP uses different testing centers within the Molecular Libraries Screening Center Network (MLSCN) to focus on different assay protocols then deposits the chemical testing data into PubChem repositories (Wang et al. 2009; Chen and Wild 2010). As such, the use of varying vocabulary has preset difficulties in understanding the assays and in applying cross-analysis methods (Schürer et al. 2011).

The ToxCast HTS assays were previously annotated in a number of ways using unstructured text. Containing a breadth of initial annotations, these have served as the foundation

for chemical-to-pathway modeling and for anchoring *in vitro* to *in vivo* outcomes for predictive models (Judson et al. 2010; Knudsen and Kleinstreuer 2011; Kleinstreuer et al. 2013). However, with exception to the gene target annotations, the use of unstructured text without quality controls gave way to mistakes in representation that were later remedied. Visser et al. (2011) has described the utility of maintaining quality control checkpoints, a workflow step to inspect for annotation mistakes in previous annotations and for annotating new assays moving forward. Visser et al. also promotes the use of controlled, ontology vocabulary to help unify synonymous concepts, where applicable. In doing so, the use of better annotation terms gets highlighted or, in the lack thereof, the usage highlights the need for new concepts and terms.

Aside from publishing the articles with the HTS and HCS assay data, there are communication challenges that sit between access to the data and knowledgeable use of the data. There is currently no recognized, uniform guidance for the minimum amount of information (e.g. metadata) needed to be supplied with the assay data across technologies (Vempati et al. 2012). Alternatively, ontologies provide controlled vocabulary that may address relationships between different assay concepts. There are a large number of database schemas and biological ontologies currently available within BioPortal (bioportal.bioontology.org) that could provide controlled vocabulary for annotation purposes. Among them, the BioAssay Ontology (BAO), an ontology created from the University of Miami, has proposed a guidance framework that incorporates vocabulary from different ontologies for objectively annotating HTS assay. This includes provisions for the assay design, assay formats, detection technologies, perturbagen (further referred as the tested chemical), and endpoints (Vempati and Schürer 2004; Visser et al. 2011; Vempati et al. 2012). Within each of these annotations, subclasses may branch further, where each term has descriptive information provided to explain its contextual usage.

While ontologies may change, the ontology's framework can be reused to guide annotations. BAO version 1.6 has been used to annotate the assays deposited into PubChem for relevant assay descriptive data (Schürer et al. 2011; Vempati et al. 2012). Schürer et al. (2011) has applied these PubChem annotations as the basis for assay promiscuity evaluations, a calculation method similar to a principle components analysis (PCA) to determine what annotations of the assay are major contributors to the amount of noise and non-specific chemical hits.

Because of this precedence, the use of BAO version 1.6 to annotate assays moving forward may enable a uniformed language through which assay comparisons may be made. However, several limitations in BAO were noted; the amount of terms incorporated into BAO from other ontologies is not representative of each of those full ontologies. This suggests that BAO keeps only the fraction of those ontology terms that have been encountered with each update; therefore, newer and diverse assays may have concepts and terms outside of BAO's capacities. BAO is currently at version 2.0. Between versions, BAO has incorporated more terms from Gene Ontology (GO) biological processes and Cell Line Ontology (CLO) into their respective branches. The same creators of BAO have also created the G-Protein Coupled Receptor Ontology (BAO-GPCR), enabling BAO annotation terms to link with BAO-GPCR terms for concepts relating to G-protein coupled receptors (GPCR).

Chapter 2

Introduction

Seen as versatile, cost-effective, and a way to gain mechanistic insight, testing chemicals in concentration-response using HTS and HCS assays with diverse biological endpoints overcomes a number of disadvantages in whole animal toxicity testing (Judson et al. 2009; Morisseau et al. 2009). The main advantage in using HTS assays is that novel biological targets can be investigate through a wide range of *in vitro* assay technologies (e.g., receptor binding, transcriptional activation, protein fragment complementation). The multiplexed and multiparametric approaches make it possible to use a single assay for interrogating a battery of targets, allowing for more pathway-based analyses and mechanistic learning all while reducing testing cost and time (Romanov et al. 2008; Houck et al. 2009; Giuliano et al. 2010; Martin et al. 2010; Rotroff et al. 2010). With increased throughput, *in vitro* assay technologies are key strategic tools to generate data on chemical-biological activity and step away from the heavy reliance upon preliminary *in vivo* whole animal testing.

As HTS technologies continue to improve, questions are raised about how to capture the increasing complexity while retaining the ability to relate between assays. Often described non-uniformly, some HTS assays may have higher minimum information standards than what is required for other technologies. Without adopting a consensus structure, the approaches to describe assays may vary between different HTS initiatives. These inconsistencies in uses of vocabulary can presets communication and cross-analysis difficulties.

In the early stages of Tox21, the National Institute for Health (NIH) National Center for Advancing Translational Sciences (NCATS) came up with a list of reporting parameters for HTS and HCS assays (Inglese et al. 2007). This was a short list meant to minimally capture the necessary components for the HTS screening and post-analysis. While the concepts were clear, how to address the parameters required more clarity and instruction. Annotations would still have used unstructured free-text.

Since then, BAO has proposed a guidance framework based on the screening information produced from MLP. This framework nominates concepts that need to be annotated and annotation terms for each annotation to use. What's more, it captures more than the NCATS reporting parameters while clarifying certain parameters that are better separated. Featuring clear semantics and hierarchical relationships, BAO makes use of several ontologies of biomedical and pharmacological focus and has applied their framework to annotate assays in the public domain from PubChem and other data repositories (Vempati and Schürer 2004; Visser et al. 2011; Vempati et al. 2012). This highlights the BAO framework as a broad and integrative foundation for capturing similarities and differences between assays within the realm of toxicity testing.

The ToxCast Assay Annotation was developed to describe features represented in ToxCast assay endpoints. Led by the EPA National Center for Computational Toxicology (NCCT), ToxCast is tasked to test chemicals in the environment with the purpose of increasing the biological-toxicological knowledge and informing chemical-testing decision-making. Contracting with various laboratories and platform vendors, the ToxCast Phase I and II chemical sets (n=1060) were completely tested through innovative set of assay technologies (Kunkel et al.

2004; MacDonald et al. 2006; Hartig et al. 2008; Romanov et al. 2008; Giuliano et al. 2010; Huang et al. 2011; Sakamuru et al. 2012; Sipes et al. 2013). Assays vary from simple interactions (e.g., biochemical binding or enzyme inhibition/activation) to complex biological reporters (e.g. multiple targets, inferred targets, or cellular processes). The discussion on minimum information standards has become increasingly technology-oriented (Goetz et al. 2011; Patlewicz et al. 2013). Adhering to minimum information standards makes a global assay annotation framework difficult. Therefore, in annotating the ToxCast assays which have more technological and biological diversity than in any screening program before it, it requires a uniform list of annotations. Since BAO is the most comprehensive assay-oriented ontology currently available among the 370 ontologies in BioPortal (http://bioportal.bioontology.org/), it was applied in a structured approach that would allow for expansions to the annotations and annotation terms.

The primary objective of this study is to demonstrate the role of the ToxCast Assay Annotation in understanding and analyzing HTS data. First, we describe the structured approach used in the global annotation. Next, we show how these annotations can provide resolution to understand the assay identification, design, target, or analysis information. Finally, we report on an initial cross-analysis of the chemical-biological activity using the annotation terms. Example scenarios are provided to illustrate how design and target annotations can be used to represent assay biological and technological space.

Chapter 3

Methods

Data Sources

The inventory list of ToxCast assays was obtained from ToxCastDB (Judson et al. 2012). The assays used to build the first version of ToxCast Assay Annotation was limited to only those that have completed testing and analysis with ToxCast Phase I and II chemical sets. Assays that met this criteria were found to belong to the following assay sources: 1) ACEA biosciences (ACEA, www.aceabio.com) (Rotroff et al. 2013), 2) Apredica (APR, www.cyprotex.com) (Shah et al., in-progress), 3) Attagene (ATG, www.attagene.com) (Romanov et al. 2008; Martin et al. 2010), 4) Bioseek (BSK, www.bioseekinc.com) (Kunkel et al. 2004; Houck et al. 2009), 5) Novascreen (NVS, www.perkinelmer.com) (Knudsen et al. 2011; Sipes et al. 2013), 6) Odyssey Thera (OT, www.odysseythera.com) (Yu et al. 2003; MacDonald et al. 2006), and 7) Tox21 (Huang et al. 2011; Sakamuru et al. 2012). EPA purchases the assay data generated from the Assay Sources, while Tox21 provides assay data as part of an interagency collaboration. The methodology for each assay was obtained from their respective ToxCast platform manuscripts, vendor/program publications, or standard operating procedures (SOPs). For data analysis with chemical testing data, the *High-Throughput Chemical Screening Data from ToxCast & Tox21* as part of the *December 2013* ToxCast Phase II Data Release was used (ToxCast 2013).

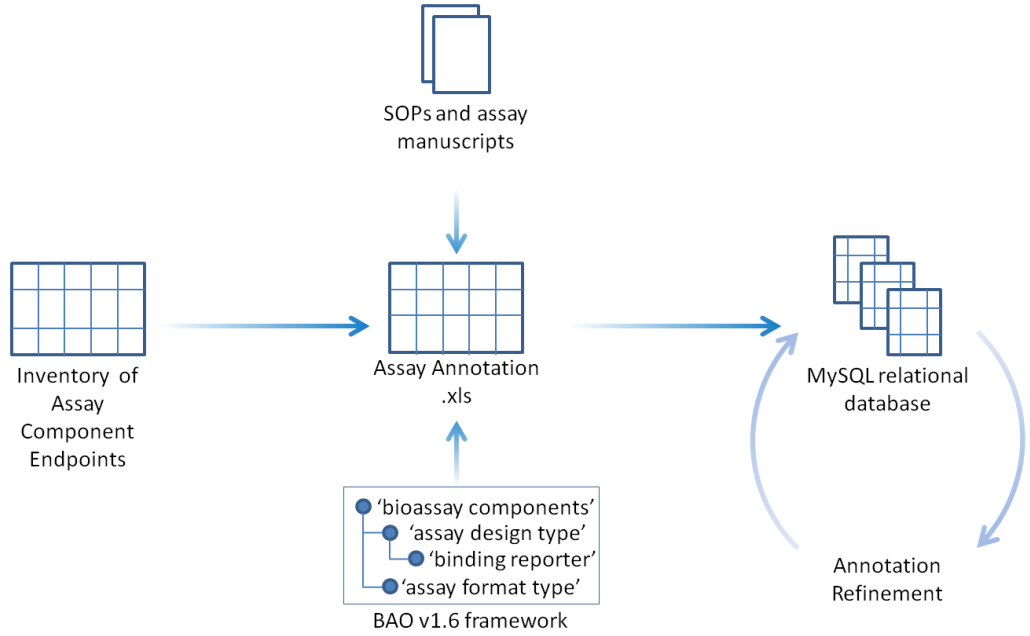
Annotation Framework

The BAO version 1.6 (www.bioassayontology.org) was developed around six concepts of biological screening (Schürer et al. 2011; Visser et al. 2011; Vempati et al. 2012). To summarize,

(1) *perturbagen* (perturbing agents that are screened), (2) *assay design* (the underlying methodology and strategy used for detecting a perturbation), (3) *assay format* (the chemical- and biological-features common to the test condition), (4) *detection technology* (the physical method used to detect and record perturbation signals), (5) *meta-target* (the molecular entity, biological process, or event interrogated by the assay), and (6) *endpoint* (the analyzed measurements, parameters and values). For the ease of maintaining the chemical inventory, the ToxCast chemical library, representing the perturbagen component, is stored and routinely updated within the DSSTox database (Williams-DeVane et al. 2009; Judson et al. 2012); hence, the ToxCast Assay Annotation is meant to describe the other five concepts of biological screening.

The annotations can be separated into four sets of information. These four sets include assay identification information (identifiers for each level, the assay source, and peripheral catalog information), design information (format, design, and technology aspects that decompress the assay's innovations), target information (various perspectives about the assay's target), and analysis information (how the data were processed and analyzed). Each set of information can be further separated into smaller concepts—the individual annotations. Each annotation is rationally assigned to one of three levels, which represent the stage of the assay as they undergo processing. Each annotation is annotated with an annotation term (the controlled vocabulary) with respect to its level. These levels includes: the assay level–the experiment or test event, the assay component level—the individual raw readouts within the experiments, and the assay component endpoint level—the analyzed readouts which have been data fitted, such as to a four-parameter Hill curve. In this way, the assignment puts an annotation as a feature of a certain level, a communication option for focusing the amount of information.

Figure 1: The annotation workflow



Data Input

The initial, data entry steps follow the left-and-middle portions of the workflow shown in Figure 1. Initially, from the table of ToxCast *assay component endpoint*s, each assay was manually annotated to an Excel spreadsheet, following the BAO version 1.6 annotation template as a model (Vempati and Schürer 2004). In this format, each column was an *assay component endpoint* obtained from ToxCastDB and each row was an annotation. Thereafter, annotation terms were selected in reference to the SOPs or the assay manuscripts.

After the initial steps the annotations were transitioned to a MySQL relational database for better data structure and ease of quality control. The spreadsheet format permitted rigid one-to-one assignments. This poses problems for the annotations that have one-to-many relationships, such as reagents, which were kept semicolon delimited within each cell of the spreadsheet format until they could be transitioned. This also allows for the linkages between

each annotation to Assay, Assay Component, or *assay component endpoint*s to be defined and represented together as a table.

The ToxCast Assay Annotation MySQL database is mainly comprised of seven tables. The *Assay Source* table provides some description for the contract vendor that performed the assays. The *Assay*, *Assay Component*, and *assay component endpoint* tables are analogous to their levels, and contain the annotations assigned to the respective level. Reagents, technological targets, and intended targets contain one-to-many relationships so they were separated as their own tables. Respectively, *Assay Reagents*, *Assay Component Target*, and *assay component endpoint Target* were mapped as dependents of the *Assay*, *Assay Component*, and *assay component endpoint* tables.

Quality Control

Quality control checkpoints to inspect the manual curations occurred at two steps: (1) transitioning the spreadsheet to the MySQL database and (2) for refining the annotations kept in the MySQL database. At the first quality control checkpoint, the spreadsheet was transitioned into a MySQL database and inspected for mismatched, mistyped, or erroneous entries. At the second quality control checkpoint, the annotation terms are extensive reviewed for appropriate coverage and representation. We also reviewed the annotations to inspect for appropriate transmission of information. Wherever necessary, additional ontologies were incorporated to supplement the annotation terms from BAO. Table 1 displays the 37 annotations selected for further use, where six annotations have subset annotations.

Table 1: The 37 annotations used and the short description of the concepts they capture

Annotation name	Short description of the annotation
assay source name	a short name for the entity that conducted the assay
assay source long name	the long name for the entity that conducted the assay
assay name	a short name for the assay
assay component name	a short name containing the assay and its component readout
assay component endpoint name	a short name containing the assay, the component readout, and the analysis applied
timepoint hr	the duration length to conduct the test portion of the assay
organism id	the NCBI taxonomy id for the organism or cellular derivative used for the assay
organism*	the organism related to the target of the assay
tissue	the organ-level, anatomical entity of the protein or cell used in the assay
cell format	the cellular or subcellular format of the assay
cell free component source	the cellular or sample tissue source of the assayed gene protein
cell short name	the name of the cell line or primary cell used
cell growth mode*	the growth mode of a cell line
assay footprint	the physical format, such as plate density, in which an assay is performed
assay format type* ¥	the conceptual biological and/or chemical features of the assay system
content readout type*	the throughput and information content generated
assay design type* ¥	the method that a biological or physical process is translated into a detectable signal
detection technology type ¥	the type of detection signal
detection technology	the name of the detection technology method
key positive control	the designated positive control
dilution solvent	the solvent used as the negative control and to make the test chemical soluble
dilution solvent percent max	the maximal amount of the dilution solvent that could be present during an assay
key assay reagent type	the type of key determinant substance of the assay
key assay reagent	the name of the key determinant substance of the assay
assay function type	the purpose of the analyzed readout in relation to others from the same assay
biological process target	the biological process or processes investigated by the assay
normalized data type	the fold induction or percent activity scale in which the assay data is displayed
signal direction type	the expected direction of the detected signal in relation to the negative control
analysis direction	the analyzed positive (upward) or negative (downward) direction
signal direction	the direction observed of the detected signal in relation to what was expected of it
technological target type ¥	the measured chemical, molecular, cellular, or anatomical entity
technological target gene id	the Entrez gene ID for the molecular target measured by the assay
technological target gene symbol	the Entrez gene symbol for the molecular target measured by the assay
intended target type ¥	the objective chemical, molecular, cellular, pathway or anatomical entity
intended target gene id	the Entrez gene ID for the molecular target that is the objective of the assay
intended target gene symbol	the Entrez gene symbol for the molecular target that is the objective of the assay
intended target family ¥	the target family of the objective target for the assay
Culture or Assay	the culture or assay condition for reagent annotations
Reagent Name Value type	the type of substance or function served by the reagent
Reagent Name Value	the name of the reagent

* The descriptions are borrowed in part from the BAO definitions; ¥ Has an annotation to describe a subset

Standardized Vocabulary for Annotation

BAO is the principle source of annotation terms. Additional ontologies were used to standardize and expand certain annotations for broader representation of annotation terms. The

NCBI Organismal Classifications (NCBI Taxon), an ontology for species taxonomy, was used for the *organism* and *organism id* annotations. To standardize the *cell short name* annotation terms, the Cell Line Ontology (CLO), an ontology for cell line information, was used for 'cell line' *cell format* (e.g. 'CHO-K1' is the more commonly used derivative of Chinese Hamster Ovaries) (Sarntivijai et al. 2011), while Cell Ontology (CL), an ontology for *in vivo* cell types, was used for 'primary cell' or 'primary cell co-culture' *cell format* (e.g. 'umbilical vein endothelium' is the annotation term for what is more commonly referred to HUVEC cells) (Meehan et al. 2011). In this way, we can identify higher level anatomical entities that the annotation terms may belong to (e.g. 'brain' *tissue* includes 'Rat forebrain', 'Rat cortical membranes', 'KAN-TS cells' and 'Bovine hippocampal membranes').

While this is not an ontology, the NCBI Entrez Gene annotation files for human, rat, mouse, and bovine were used to annotate gene symbols and gene ids for the *technological_target_gene_id, technological_target_gene_symbol, intended_target_gene_id,* and *intended_target_gene_symbol* annotations (Maglott et al. 2011).

The *organism* and *organism id* annotations are generally annotated according to the *cell short name*. In an assay performed in a cell-based format, the general understanding is that the host cell will have cellular machinery that influences the gene and outcome of the assay. The exception is given to assays performed in a cell-free format, where genes are transfected into expression vehicles, extracted, and used in assays without additional cellular component. In that situation, the *organism* is annotated with reference to the gene's species of origin.

Modifications to annotations and annotation terms

From the BAO framework, the *assay format* concept was reorganized and separated into two annotations, *assay format type* and *cell format. Assay format type* identifies the overall chemical-feature or biological-response being investigated as 'biochemical', 'physicochemical', 'cell-based', or 'whole organism' (Vempati and Schürer 2004). The *cell format type* adds resolution to the *assay format type* by represents the cells as immortalized cell lines or primary cells, and whether they are treated as homogenous cultures, heterogeneous co-cultures, or cell-free extracts during the course of the assay.

Two reporter types were added to the *assay design type* annotation terms. The 'background reporter' was introduced as a reporter type for baseline noise, while the 'growth reporter' characterizes assays that measure cellular development without intracellular or extracellular morphological endpoints.

To make the assay protocol information transparent, a few annotations are devoted to capturing these details for reporting purposes. At present, due to the lack of formal annotation terms for reagent materials, the *reagent name value* uses unstructured text. However, the *reagent name value type* contains 32 possible annotation terms to annotate the base medium, serum variety, culture or assay duration, additional reagent types (e.g. antibody, extractor, fixing agent, or stain), environmental factors (e.g. pH level and temperature), and the expected number of cells in the well by the beginning of the assay. Since the assay is influenced by the preparatory protocols during cell culturing and conditioning, the annotation separates the reagents used for the preparatory, cell 'culturing' conditions from the experimental, 'assay' conditions. If it is applicable to the readout, key reagents are highlighted as a separate annotation.

Within the BAO version 1.6 framework, the 'molecular targets' was a branch of the *meta targets* concept. Here, the annotation terms in the 'molecular targets' was used as the basis for the target type annotations—*technological target type* and *intended target type*—and their subtype annotations. The difference between these two annotations is displayed in Table 1.

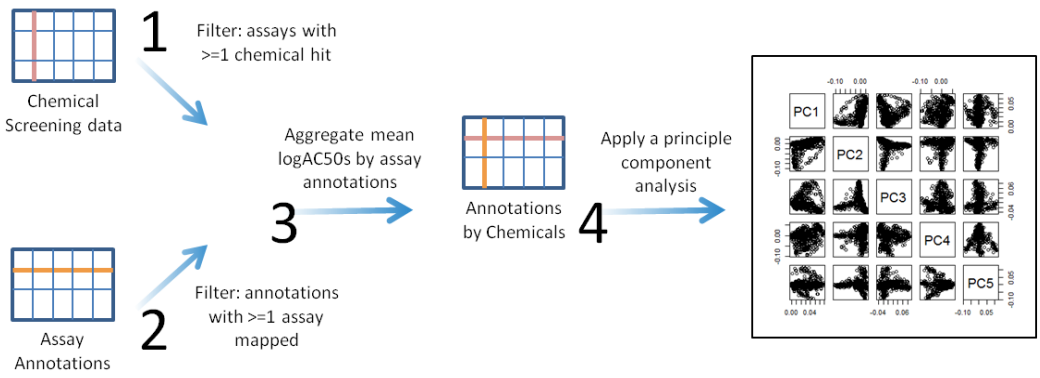
'Chemical', 'cellular', and 'pathway' annotation terms were added to the target type annotation. Specifically, the 'chemical'-type targets were given the subtypes 'physical feature', 'ATP', or a hormone chemical name (e.g. 'Cortisol', 'Corticosterone', or 'Estrone'). "Cellular"-type targets can be given to scenarios where the focus is a morphology or function, so the subtypes may include the "cellular" or a subcellular object (e.g. mitochondria, nucleus, or lysosome). For 'protein'-type targets, the subtypes 'protein-specified' was included for targets where the gene protein is known but not pursued for a certain function (e.g. not functioning for 'enzyme'-substrate or 'receptor'-ligand reactions). Furthermore, 'protein-unspecified' was included for targets that non-specifically tagged proteins. Similarly, 'pathway'-type targets and the 'pathway-specified' subtype was introduced as an annotation term for assays that probed known gene-mediated biological pathway, such as assays screening for estrogen receptor-alpha agonists.

Visualization Software and Data Analysis

NCBO BioPortal is a website used as the main source for viewing different ontologies and how they map to other ontologies, and to download archived ontology files. In addition, Protégé, an open-source software, was used to view web ontology language (owl) format files and for conducting SPARQL queries. Cytoscape, an open-source software platform for visualizing and integrating complex networks, was used to website were used. R statistical

software was used to read-in and rearrange data files, and to perform the principle components analysis using the prcomp function from the R Statistical library.

Figure 2: PCA workflow of the chemical screening data



A principle components analysis (PCA) was conducted to investigate the variances observed in the *High-Throughput Chemical Screening Data from ToxCast & Tox21* (December 2013) with regards to the ToxCast Assay Annotations, shown in Figure 2. 50% of Maximal Activity Concentration (AC50) values were obtained for each chemical and *assay component endpoint* pair. Using R statistical software, these values were filtered for assays with at least one chemical hit (i.e. values not equal to '100000' or 'NA'), then they were divided by 100000 and negative log transformed. Moreover, we removed the APR_1hr assays (n=20 *assay component endpoints*), which were discontinued after Phase I testing, and the BSK_SM3C assay (n=28 *assay component endpoints*), which had undergone a name change to BSK_CASM3C prefix between phases. Next, an assay annotation binary table was generated to show mapping between *assay component endpoints* and annotation terms. Reagent information and assay identification annotations aside from the assay component endpoint were excluded. The log-AC50 values then were merged with this Assay Annotation binary table, and filtered to retain only the annotation terms mapped with at least one *assay component endpoint* in use (i.e. the column sum of each

annotation term is at least 1). The log-AC50 values were then aggregated by each annotation term, and NA or NaN data on each row (annotation term) were normalized to the mean of the numeric values. The data were then processed using the prcomp function and the loadings were visualized using the heatmap.2 function from the gplot library. The R script is displayed as Appendix 3.

Chapter 4

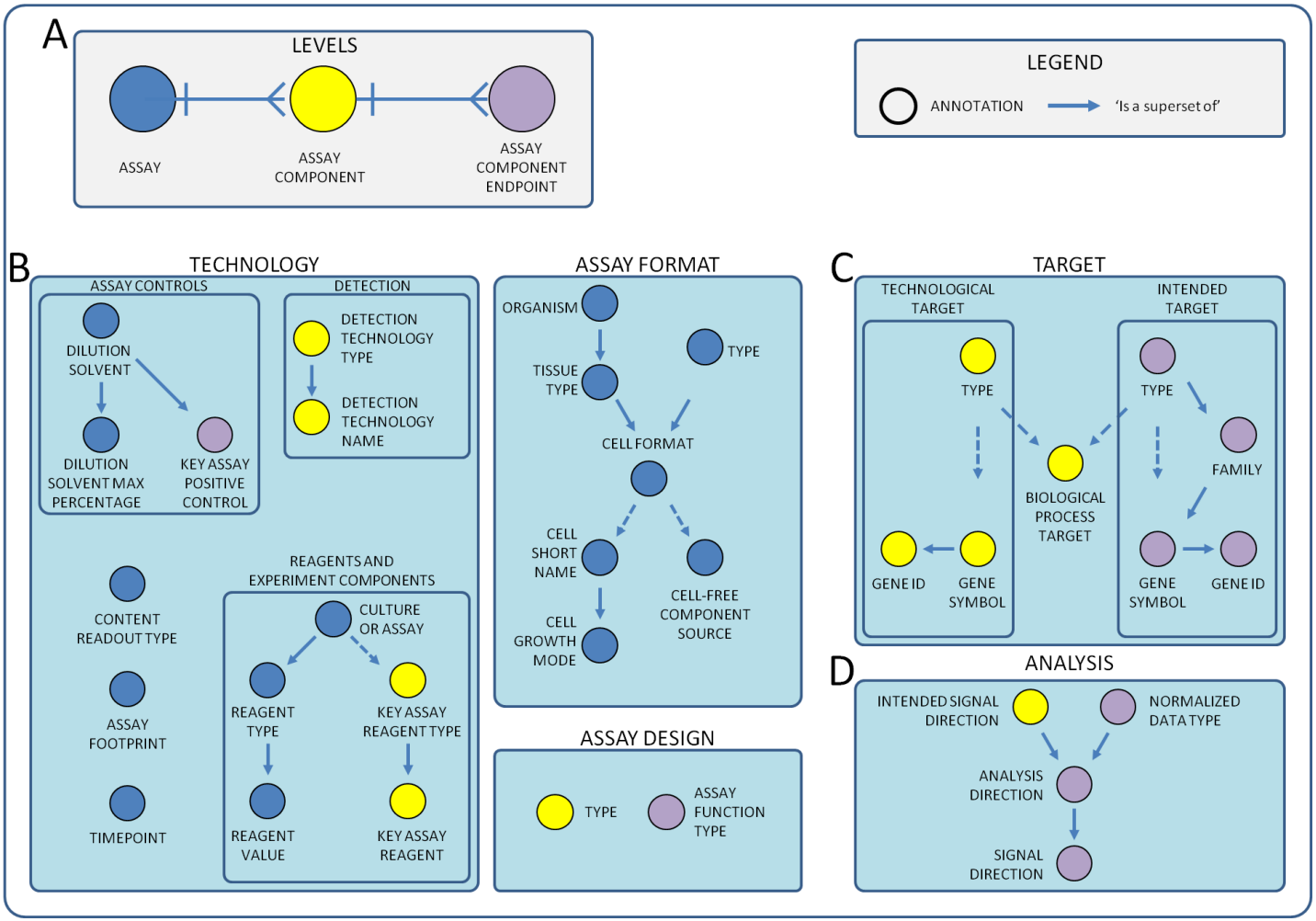
Results

Assay Annotation Structure

The ToxCast Assay Annotation structured the 37 annotations adopted from the BAO framework and used them to capture annotations for 795 ToxCast *assay component endpoint*s. The annotation structure follows the four sets of information and progresses from the 'assay' level to the 'assay component endpoint' level (Figure 1). Used to describe HTS assays, the primary goal of this thesis is to establish a structured annotation scheme that uses ontology-based annotation terms as controlled vocabulary, where applicable. In addition, these annotations can be used to understand general trends observable among the annotated HTS assays as well as explain variances observed among the screening data.

The annotation structure displays dependencies between annotations that follow the same concepts. Solid arrows are depicted for relationships between annotations where one annotation term influences the next. As seen between the *intended target type* and *intended target family*, 'protein' or 'pathway' *intended target types* would merit the *intended target family* to be annotated with a gene family; when the *intended target type* is 'cellular', the *intended target family* may be a 'cell cycle' or 'cell morphology' annotation term to follow suit. Alternatively, annotations linked by dashed arrows suggest a conditional relationship. If given certain annotation terms, the subsequent annotations may or may not be annotated. For instance, for an

Figure 3: The annotation structure. The annotations can be grouped into A) assay identification information, B) design information, C) target information, and D) analysis information. Relationships between annotations are either one-to-many (solid arrow) or conditional (dashed arrows), where certain dependencies may not be applicable.



assay that uses ‘primary cell co-culture’ *cell format*, it is unnecessary to annotate the *cell-free component source*. In that situation, the *cell-free component source* get default values equal to NA or 0, if the annotation is numeric. In another example, for a ‘cellular’ *intended target type,* it may be unnecessary to annotate target gene symbol or gene id, so both are defaulted. The use of dashed arrows is a reflection that different assay technologies may have minimum information standards that may be seen as inapplicable with each other.

Assay Identification Annotation

Kavlock et al. (2012) reports the general study designs, technologies applied, and unique features from each *Assay Source*. Here, after controlling vocabulary used across Assay Sources, Table 2 reflects that total unique features per Assay source with reference to each level. Using the 795 ToxCast *assay component endpoints* library found to have complete ToxCast Phase I and II chemical screening data, the analyzed data were linked back to 541 unique assay components, which were generated from 328 assays (Table 2). Over 23,000 annotation terms have been annotated across the 26 design and analysis information annotations for these 795 *assay component endpoints*. For their respective assays, there are roughly 2,800 records for reagents and testing protocol information (approximately 8,400 annotation terms as both structured and unstructured text), and about 1,400 records for target information (approximately 7,000 annotation terms).

Table 2: Assay information and content readout types

Assay Source	Assay	Assay Component	Assay Component Endpoint	Content readout type	
				Single	Multiplexed
ACEA	1	1	2	•	
Apredica (APR)	2	20	40		•
Attagene (ATG)	2	82	82		•
Bioseek (BSK)	8	87	174		•
NovaScreen (NVS)	276	276	422	•	
Odyssey Thera (OT)	20	20	20	•	
Tox21	19	55	55	•	
	328	541	795		

The *assay component* is meant to normalize single versus multiplexed/multiparametric assays according to individual readouts. Depending on the *content readout type,* the number of targets that a single experiment can probe is analogous to the number of *assay components* deriving from the same *assay*. At present, the ToxCast assays seem to trend as single or multiplexed. Shown in Table 2, NVS and OT assays are found to be characteristic single-readout assays displaying equal assay to assay component counts. In contrast, ATG, APR, and BSK are multiplexed-readout assays that measure a battery of individual targets including some that serve as background detection or as a measure of viability. NVS accounts for the highest number of *assays*, *assay components*, and *assay component endpoints*, while ATG assays account for the highest number of *assay component endpoints* per *assay* conducted.

Some assays do not follow strictly to the conventions of *content readout type* but provide interesting variants of the single-readout type. The ACEA assay only generates a single readout; however, the upward and downward curve-fit analysis can yield findings for two different intended targets. The cell line used by the ACEA_T47D *assay* is sensitive to estrogen-receptor (ER) agonists, and can be used to detect ER-pathway-mediated cell proliferation, when analyzed

in the 'gain' of *signal direction*, while serving a cell viability purpose in the 'loss' of *signal direction*. Similarly, Tox21 assays are meant to generate a ratio to represent the pattern of activity. This is seen first as background and raw readout (i.e. Channel 1 and 2 wavelength measurements), and a viability readouts. The ratio can be calculated as a ratio of Channel 2/Channel 1 readouts, and the viability readouts could then be used to inspect for possible artifacts or excessive cytotoxicity affecting the quality of the readouts.

Assay Design Annotation

The *assay design type* and *detection technology type* annotates the objective of the measurement and the method of collecting quantified data. A majority of ToxCast assays were found to be 'binding reporter', 'enzyme reporter' or 'inducible reporter' *assay design types* (Table 3). These reporter types assess different facets of how chemicals may affect genes of concern.

Table 3: Assay design types annotated to ToxCast Assay Components Endpoints

		Totals	Assay Sources						
			ACEA	APR	ATG	BSK	NVS	OT	Tox21
Assay design type	**subtype**								
Binding reporter	ELISA	149	0	0	0	**148**	**1**	0	0
	Fluorescent polarization	1	0	0	0	0	**1**	0	0
	Protein fragment complementation	14	0	0	0	0	0	**14**	0
	Radioligand binding	120	0	0	0	0	**120**	0	0
	FRET	8	0	0	0	0	**8**	0	0
Conformation reporter	Protein conformation	4	0	**4**	0	0	0	0	0
Enzyme reporter	Enzyme activity	296	0	**4**	0	0	**292**	0	0
Inducible reporter	Beta lactamase induction	7	0	0	0	0	0	0	**7**
	Luciferase induction	13	0	0	0	0	0	**4**	**9**
	mRNA induction	84	0	0	**82**	0	0	0	**2**
	Fluorescent protein induction	2	0	0	0	0	0	**2**	0
Growth reporter	Real-time cell-growth kinetics	2	**2**	0	0	0	0	0	0
Morphology reporter	Cell phenotype	18	0	**16**	0	**2**	0	0	0
Membrane potential reporter	Dye binding	5	0	**4**	0	0	0	0	**1**
Viability reporter	Cell number	4	0	**4**	0	0	0	0	0
	DNA content	8	0	**8**	0	0	0	0	0
	ATP content	10	0	0	0	0	0	0	**10**
	Protein content	24	0	0	0	**24**	0	0	0
Background reporter	Artifact detection	26	0	0	0	0	0	0	**26**

Some *assay component endpoints* such as BSK_3C_IL8_down are binding reporters by way of ELISA immunoassay systems, which assess target protein expression levels (i.e. decreases in IL8). In contrast, NVS_ADME_hCYP1A1_Activator considers how the gene protein's normal enzyme-substrate functions get affected by chemical competitive or inhibitory action; in this case, it assesses the level by which the enzyme-substrate functions increases. While inducible reporters may vary, some like OT_AR_ARE_LUC_Agonist_1440 use transfected firefly luciferase to probe the level of transcriptional induction.

In addition, APR and Tox21 assays were found to have made use of 'conformation reporters', 'enzyme reporters', 'morphology reporters', 'membrane potential reporters', 'viability reporters', and 'background reporters'. This identifies that certain *assay design type* may be specific to certain assay technologies or methodologies.

Most ToxCast assays use 'fluorescence' or 'radiometry' *detection technology types* (Table 4). 'Fluorescence intensity' is often the method of quantification for 'fluorescence'-type assays, which are observed in assays from all assay sources except ACEA. For assay component endpoints associable to 'radiometry'-type detection technology, 'scintillation counting' is often the method of choice for radioligand binding assays, found here to be specific to NVS assays. Though in low presence, 'label-free technologies', 'luminescence', 'microscopy', and 'spectrophotometry' detection technology type annotation terms were annotated for at least one ToxCast *assay component endpoints*.

Table 4: Detection technology types annotated to ToxCast Assay Components Endpoints

Detection technology type	subtype	Totals	Assay Sources ACEA	APR	ATG	BSK	NVS	OT	Tox21
Fluorescence	Fluorescence intensity	582	0	**40**	**82**	**148**	**260**	**14**	**38**
	Fluorescence other	1	0	0	0	0	**1**	0	0
	FRET: TR-FRET	8	0	0	0	0	**8**	0	0
Label Free Technology	Electrical Sensor: Impedance	2	**2**	0	0	0	0	0	0
Luminescence	Bioluminescence	21	0	0	0	0	0	**4**	**17**
	Chemiluminescence	1	0	0	0	0	**1**	0	0
Microscopy	Optical microscopy: Fluorescence microscopy	4	0	0	0	**2**	0	**2**	0
Radiometry	Scintillation counting	136	0	0	0	0	**136**	0	0
Spectrophotometry	Absorbance	40	0	0	0	**24**	**16**	0	0

Table 5: A comparison of the assay design subtypes by the detection technology subtypes

Assay design subtypes	No of *assay component endpoints*	Detection technology subtypes: Fluorescence intensity	Scintillation counting	Absorbance	Bioluminescence	FRET: TR-FRET	Fluorescence microscopy	Electrical Sensor: Impedance	Fluorescence polarization	Chemiluminescence
		594	136	40	21	8	4	2	1	1
enzyme activity	296	264	16	16	0	0	0	0	0	0
immunoassay: elisa	149	148	0	0	0	0	0	0	0	1
radioligand binding	120	0	120	0	0	0	0	0	0	0
mRNA induction	84	84	0	0	0	0	0	0	0	0
cell phenotype	26	24	0	0	0	0	2	0	0	0
protein content	24	0	0	24	0	0	0	0	0	0
artifact detection	20	20	0	0	0	0	0	0	0	0
protein fragment complementation assay	14	14	0	0	0	0	0	0	0	0
luciferase induction	13	0	0	0	13	0	0	0	0	0
DNA content	12	12	0	0	0	0	0	0	0	0
ATP content	10	2	0	0	8	0	0	0	0	0
FRET	8	0	0	0	0	8	0	0	0	0
beta lactamase induction	7	7	0	0	0	0	0	0	0	0
dye binding	7	7	0	0	0	0	0	0	0	0
cell number	6	6	0	0	0	0	0	0	0	0
protein	6	6	0	0	0	0	0	0	0	0
fluorescent protein induction	2	0	0	0	0	0	2	0	0	0
real-time cell-growth kinetics	2	0	0	0	0	0	0	2	0	0
fluorescent polarization	1	0	0	0	0	0	0	0	1	0

A comparison between the two annotations suggests that the same *detection technology type* may assess different *assay design types*; conversely, the same *assay design type* may be assessed by different *detection technology types* (Table 5). It is noteworthy to mention 'fluorescence' *detection technology type*, which have been applied to all *assay design types* except 'growth reporters', reflects that fluorescent protein and probe technologies have developed in greater extents for HTS targeted measurements compared with other technologies.

Related to the detection technology, it was found that the reagent and experimental components annotations come secondary to the format annotations. We attempted to capture the conditions of the test environment, but in doing so found that separate protocols are used for preparations prior to the assay (the culture conditions) compared to the actual assay. Shown in Table 6 is an example of the reagent information for the APR_HepG2_1hr *assay*. Under the Culture and Assay conditions, we display the types of reagent of condition used (left) and the name or value to that reagent or condition (right). Table 6 is somewhat representative of the cell-based high-content screening assays, as opposed to cell-free biochemical assays which are only annotated with assay conditions.

Table 6: Reagent and components information for the APR_HepG2_1hr assay

Culture		Assay	
media_base	Eagle's minimum essential media/ Earle's balanced salt solution	media_base	Eagle's minimum essential media/ Earle's balanced salt solution
media_serum	10% FBS	media_serum	10% FBS
cofactor	non-essential amino acids	extracellular matrix	rat tail collagen I
cofactor	glutamine	buffer	Hank's balanced salt solution
inhibitor	penicillin	fixing agent	3.7% formaldehyde
inhibitor	streptomycin	antibody	primary antibodies
media_temp_celcius	37	stain	Hoechst-33342 dye
media_time_hr_min	18	antibody	anti-phospho-histone-H2AX antibody
media_time_hr_max	24	antibody	anti-a-tubulin antibody
media_cell_aliquot	4200	antibody	anti-p53 antibody
		stain	MitoTracker Red
		antibody	anti-phospho-c-jun antibody
		antibody	anti-phospho-histone-H3 antibody
		media_temp_celcius_min	25
		media_temp_celcius_max	37
		media_time_hr	1

In addition, depending on the assay, if a reagent is the key factor(s) towards quantifying signal changes, then the reagent(s) would also be annotated as the *key assay reagent*, an annotation of the *assay component* level. Generally, this means that the reagent is captured as a factor of the 'assay' conditions. Taking the APR_HepG2_1hr reagents, the MitoTracker Red (stain) is necessary for identifying the mitochondria within each cell, and so it is highlighted as the key assay reagent for APR_MitoMembPot_1hr and APR_MitoMass_1hr *assay components*. However, like for ATG CIS and TRANS assays, specific reporter transcription unit (RTU) used during the cell-culturing preparatory protocols are central towards the reporting of the respective readouts. Moreover, it is worthy to mention that label-free technologies, such as the cell electrical sensing used for the ACEA_T47D *assay*, would not have a *key assay reagent*.

While the connections could not be displayed in Figure 3, the *assay format type* and the *cell format*, and in some instances the *cell short name*, have a predominant influence upon the reagent use. Certain reagents are necessary for the culture of cell-based versus biochemical assay

format types. The same is observed for assays that use 'primary cells' or 'primary cell co-cultures' versus 'cell lines'. Certain cell lines will have more specific media and serum specifications than others. Table 7 provides a short summary of the three most commonly used organism and tissue types by each assay source, and how many unique *cell short names* or *cell free component source* used by each assay source were annotated to one of five possible *cell format types*.

Table 7: Organism/tissue types and cell format types

	ACEA	APR	ATG	BSK	NVS	OT	Tox21
Unique Organism & Tissue-types	**1**	**1**	**1**	**3**	**30**	**5**	**6**
Most frequently used	**human** (breast)	**human** (liver)	**human** (liver)	**human** (vascular)	**human** (recombinant)¥	**human** (kidney)	**human** (kidney)
2nd most frequent				**human** (skin)	**rat** (brain)	**Chinese hamster** (ovary)	**human** (breast)
3rd most frequent				**human** (lung)	**rat** (recombinant)¥	**human** (cervix)	**human** (liver)
Cell format types							
cell line	1	3	2	0	0	20	19
primary cell	0	0	0	4	0	0	0
primary cell co-culture	0	0	0	4	0	0	0
cell-free	0	0	0	0	188	0	0
tissue-based cell-free	0	0	0	0	88	0	0

¥ ***Target gene proteins were extracted from expression vehicles (e.g. insect cells, bacterial, or cell lines)***

Target information

The names of the ToxCast *assay component endpoints* may not immediately focus on the intended target. In fact, with just the assay component endpoints alone, it will be a challenge to determine what the targets are at all. The intended target is the objective probe of the chemical bioassay, and it can often be said to be the center of communication in regards to the chemical's activity. This can discount the value of the technological target. After all, the intended target is

measured either directly or through an interpretation that uses the technological target. The current annotation uses *technological target type*, *intended target type*, *biological process target*, *intended target family type* and gene annotations to distinguish these different motifs.

Across the 795 ToxCast *assay component endpoints*, by gene ID and target type combinations, there are 383 unique technological targets and 387 unique intended targets. The technological target types range from 'DNA', 'RNA', 'protein', 'cellular' and 'chemical' types, while the intended targets include 'protein', 'cellular', 'pathway', and 'chemical' types. Though it may be simplest to annotate one target per *assay*, this approach overlooks the value in multiplexed assay readouts. The technological and intended target annotations were created to dissociate assays that use different means to measure the same intended target.

A comparison of the technological and intended targets shows that some assays make direct measurements, while others use the technological target as a quantifiable surrogate for the intended target. Table 8 summarizes the occurrence of these measurement relationships with regards to the target types from each assay source. Assays that make direct measurements have the same annotations for technological and intended target types and gene ids. Alternatively, assays may make use of technological targets as quantifiable surrogates or close substitutes to approximate the intended target, shown boxed in Table 8.

Table 8: Comparison of technological and intended target types

	ACEA	APR	BSK	Tox21	ATG	BSK	NVS	Tox21	ACEA	APR	OT	Tox21
Intended	**cellular**				**protein**				**pathway**			**chemical**
Technological												
cellular (25)	1	8	12	3	0	0	0	0	1	0	0	0
protein (662)	0	4	8	0	0	154	422	24	0	16	34	0
RNA (138)	0	0	0	0	138	0	0	0	0	0	0	0
DNA (12)	0	8	0	0	0	0	0	0	0	4	0	0
chemical (22)	0	0	0	10	0	0	0	0	0	0	0	12

Most of the ToxCast HTS assays are straight-forward technologies that make direct measurements. Displayed in Table 8 as the unboxed values, there are 24 ‘cellular’, 600 ‘protein’, and 12 ‘chemical’ targets directly measured by an *assay component endpoint* from each *assay source*. For example, the BSK ELISA-based assays (e.g. BSK_3C_MIG_dn) use protein-specific antibodies to bind to specific target genes. The change in fluorescence would be directly relative to the protein expression level at each concentration tested. In a similar way, the Tox21 autofluorescence assays (e.g. Tox21_Autofluor_HEPG2_Cell_green) aim to detect inherent fluorescent properties from the test substance. These assays probe different color wavelengths to observe baseline changes that could be concentration-dependent artifact fluorescence from the chemical.

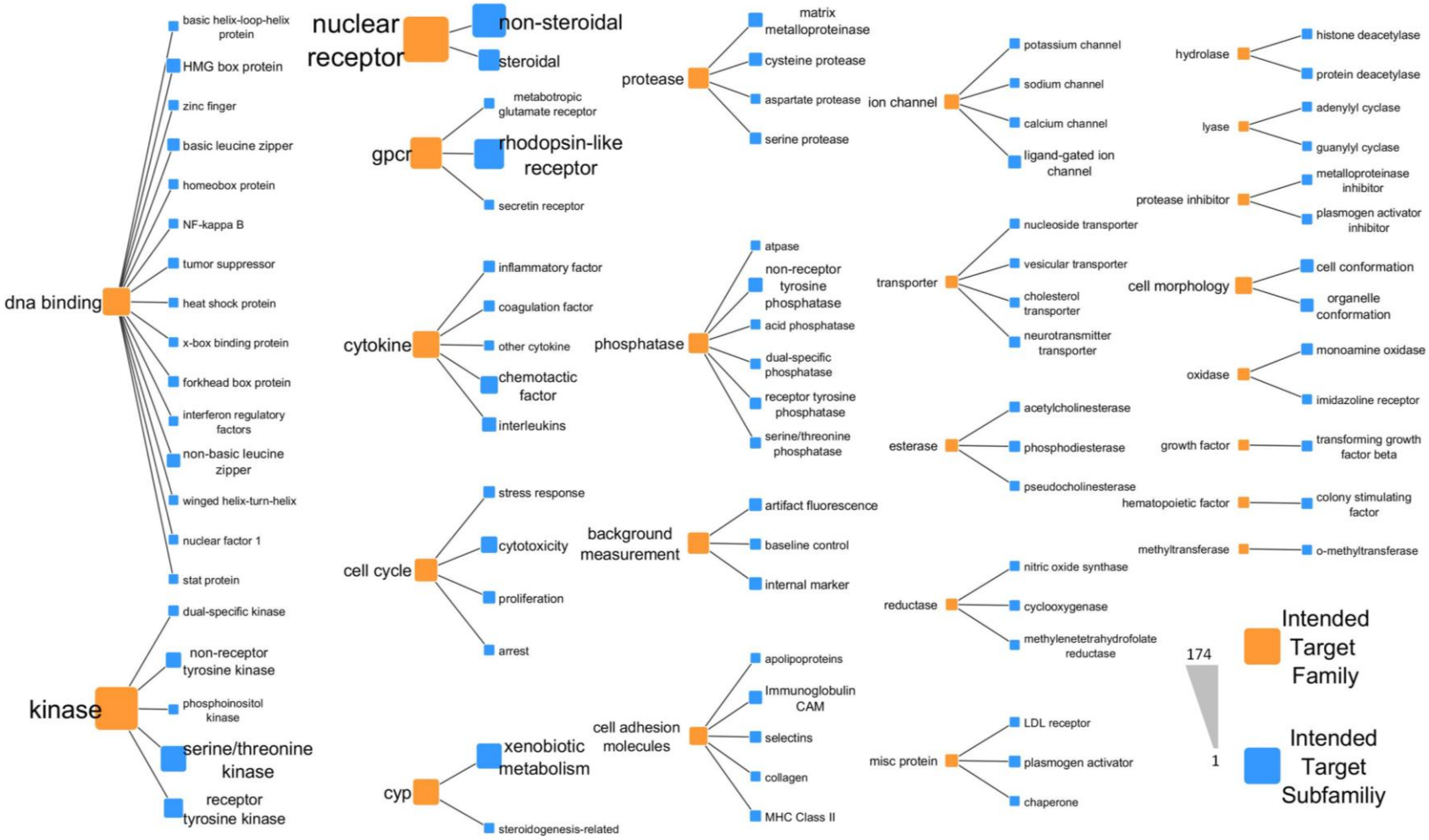
Alternatively, assays may target an abstract component of the intended target’s biology as a function of the technological target. Shown in Table 8 as boxed values, there are 30 ‘cellular’, 138 ‘protein’, and 55 ‘pathway’ targets assessed by various methods and technological target types. For instance, OT_PPARg_PPARgSRC1_0480 measures the fluorescence generated from the complementary binding of human peroxisome proliferator-activated receptor gamma (gene symbol: PPARG, gene ID: 5468) with the v-src kinase (gene symbol: SRC, gene ID: 6714). Changes in the measured level of fluorescence and relative localization within the cell are indicative of changes along the PPARG signaling pathway. For assays where the concept becomes too complex to represent by target type and gene annotations alone, the *biological process target* would be annotated. Take ATG_PPRE_CIS for example. human Peroxisome Proliferator-Activated Receptor Alpha, Delta and Gamma—PPARA (gene ID: 5465), PPARD (gene ID: 5467), and PPARG (gene ID: 5468), respectively—are the technological targets

measured together as a unit as downstream products of transcription factors binding to the Peroxisome Proliferator-activated Response Element (PPRE). Changes to the mRNA levels of these target genes reflect chemical effect to the upstream transcriptional events, so the *biological process target* is the 'regulation of the transcription factor activity'.

Furthermore, to represent a complex signaling or regulatory pathway as the assay target, the *biological process target* is used in conjunction with reference gene focal to that pathway. For instance, the ACEA assay monitors cellular growth kinetics as an indication of cytotoxicity or estrogen receptor alpha (ESR1) signaling for cell proliferation. Respectively, the cytotoxicity intention has a 'cellular' *intended target type* and 'cell cycle'- 'cytotoxicity' *intended target family*. The cell proliferation intention has 'pathway' *intended target type*, ESR1 as the *intended target gene symbol*, 'nuclear receptor' and 'steroidal' as the *intended target family*, and 'regulation of cell proliferation' as the *biological process target.*

To group the 387 unique intended targets, we developed 24 *intended target families* (84 subfamilies). Figure 4 is a simple connection map displaying the *intended target families* to their *intended target subfamilies,* sized by the number of times an intended target gene is mapped to those annotation terms. Three main categories currently exist in the intended target families: 2 cellular aspects (i.e. 'cell cycle' or 'cell morphology'), 21 gene families (e.g. 'GPCR,' 'kinase,' or 'protease'), and one for quality control aspects (i.e. background measurements), shown in Appendix 1. The *intended target families* have one-to-many relationships with the intended target genes, providing a means to filter down to the targets of interest or to query for assay identifiers associated within the *intended target family* groupings.

Figure 4: Intended target family annotation terms



The structure to the *intended target families* is a work-in-progress that attempts to pull in groupings from various ontology sources. First, we realize that gene protein vocabulary bear intricate connections in terms of the active/inactive sequence domains, functions, and relationships between superfamily, family, subfamily and many more categories. These categories are often unevenly distributed between categories making it difficult to communicate as some are more developed than others. For the majority of gene proteins, groupings are formed based on function and similarity in protein sequence. Some groupings are under debate as more gene proteins are characterized and the functions annotated. The *intended target families and intended target subfamilies* are over-simplifications; they are an attempt to cross-sectionally group the target genes within reason by their first and second order associations. Often, this means the gene-oriented *intended target families* are the class of proteins, and the *intended target subfamilies* are actually the regarded superfamilies under that class.

From same creators as BAO, the G-Protein Coupled Receptors Ontology (BAO-GPCR) was used to define the high-leveled subfamilies within the domain of G-protein coupled receptors (GPCRs) (Przydzial et al. 2013). Unfortunately, for most of the other gene-oriented *intended target families,* there is not a single well established and publicly ontology to represent each topic area. This can be seen with the kinase family, which continues to have new gene proteins discovered and new subfamilies introduced; thereby, older classifications are antiquated and new introductions are not well-characterized (Manning et al. 2002). The five *intended target subfamilies* for kinases displayed are a product between the schemes used in the KinaseDB, the WikiKinome, National Cancer Institute Thesaurus (NCIT) and Medical Subject Headings (MeSH). The same can be said for the ‘Protease’, ‘Phosphatase’, and many other gene-oriented *intended target families*.

When mapped out, we found that 62 *intended target gene symbols* have homologs genes investigated within ToxCast. These gene symbols were mapped back to 145 *assay component endpoints* that could be used for comparison of orthogonal assays (see Appendix 2), excluding those genes that were mapped to only one *assay component*. These genes belong to the 'nuclear receptor', 'dna binding', 'growth factor', 'protease', 'cyp', 'esterase', 'gpcr', 'ion channel', and transporter *intended target families.* Furthermore, among these identified *assay component endpoints*, 24 non-human genes investigated by NVS cell-free protein-binding assays have a human homolog investigated among the rest of the identified 145 *assay component endpoints*. These *assay component endpoints* open the possibilities to compare orthogonal assays for understanding different chemical interaction patterns and for comparisons across different species.

Assay Analysis Information

Assay component endpoints distinguish the data processing decisions applied unto the raw *assay component* data, represented in Figure 4 by four annotations. Differences in the concepts covered in these annotations are represented in Table 1. The *normalized data type* prominently observed in ToxCast *assay component endpoints* is 'percent change'. Subsets of ATG (n=80) and BSK (n=174) *assay component endpoints* are annotated as 'fold induction' type. 'Fold induction' normalization uses the performance of the negative control as the baseline reference, while 'percent change' uses the performance of negative control as the baseline and positive control for normalizing the maximal activity. The *analysis direction* details whether the *assay component* data were fitted in a 'positive' or 'negative' activity direction, important

concepts when using values generated, such as the lowest effective concentration (LEC) and the activity scores (e.g. AC50).

The data analysis was used to also identify findings that are both expected and unexpected of the *assay component*. With respect to increasing test chemical concentrations, the *signal direction type* annotates whether the raw readout is expected to have a 'gain' or 'loss' in signal activity relative to that of the negative control, may go 'both' ways, or where the expectation is no change ('none'). The actual analysis applied is described by the *analysis direction*, annotated as either 'positive' or 'negative'.

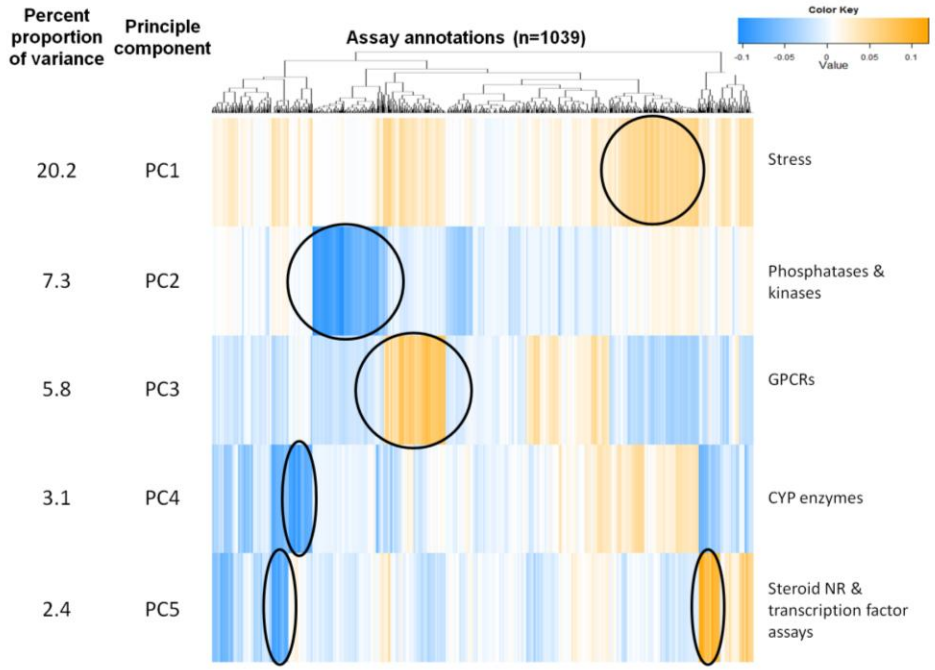
For *assay component endpoints* with theoretical intents (i.e. *signal direction type*: 'gain' and 'loss'), the *signal direction* is annotated as 'gain' if the *analysis direction* annotation corresponds (i.e., 'positive') or as 'loss' if they do not correspond (i.e. 'negative'). For example, the NVS_ADME_hCYP1A1_Activator is expected to have a 'loss' of signal (*signal direction type*) as increasing chemical concentrations impedes the cell-free reaction. However, it is fitted in the 'negative' *analysis direction*, indicating that a decrease in the fluorescent substrate presence (the reaction was promoted to generate the product) relative to the negative control was detected with increasing test concentrations. Chemicals active in this *assay component endpoint* would, therefore, have caused a 'gain' in the protein's activity (*signal direction*).

For *signal direction types* annotated as 'both' or 'none', there was not a theoretically intended direction. In these cases, the *signal direction* only corresponds to the *analysis direction*, where 'positive' yields 'gain' and 'negative' yields 'loss', respectively.

Principle Component Analysis

With the December 2013 ToxCast Phase II Data Release, there are 795 *assay component endpoints* with complete chemical screening data. After generating the binary file to map each *assay component endpoint* to their annotation terms (data not shown), we used the global PCA using 1,039 of the annotation terms. These terms come from all annotations, except the assay identification annotations (removed as non-descriptive annotation terms) and the reagent and experimental components annotations (had not undergone full quality control review). The PCA produced 1,039 principle components with loading values for each annotation term at each principle component (correlation values to each principle component). Figure 5 displays a heat map of PC1 through PC5 with the clustered annotation terms, where the loading values are presented as heat between dodger blue (negative value) to orange (positive value).

Figure 5: Heat map of the PCA loadings clustered by the annotation terms for the first five principle components



The percent proportion of variance for the first three principle components is about 33%, where PC1 alone can explain about 20.2% of the variance observed across the screening data. This is not unexpected since the percent proportion for each principle component continues to decrease as it approaches PC1039, where the summation of all principle component percent proportions approaches 100%. Our findings suggest that this global PCA cannot explain the variability very well. However, it does help identify the conceptual clusters of annotations terms contributing to each of the principle components (Figure 5, at right). For instance, for PC1, we observed extreme heat or correlation values coming from a number annotation terms related to stress on the test environment and contributors to promiscuous activity. This includes but is not limited to factors of the technology, assay and cell format, and where the target of interest is cytotoxicity or cellular stress. These peripheral factors to an assay are potentially the main contributors to noise in the data and that this may interfere with the cell-free as well as cell-based assays.

Continuing along the principle components, there seems to be certain cluster of annotations contributing to more variance than others. Along the right side of Figure 5, phosphatases and kinases were identified as next most predominant factor contributing to variability in the screening data (PC2), followed by GPCRs, ion channel proteins, and transporters (PC3), then CYP enzymes (PC4), and steroidal nuclear receptors and a number of transcription factors (PC5). Assays that have these features or combinations of these features may be more prone to promiscuous hits that others. What this calls for is new analytical approaches to adjust or further analyze these hit variance issues.

Chapter 5

Discussion, Conclusions, Limitations, and Future Directions

Discussion

The diversity of *in vitro* assays has continually broadened, bringing new options for investigative toxicology to light while increasing the complexity and information demand. There was no structure to meet this information demand. In 2006, at the beginnings of ToxCast, the only information sources available to understand an assay and its readouts would have come from manuscripts, summary protocols, and other printed sources. Even then, each source may use different vocabulary or report different amounts of detail, making it a tedious effort to fully grasp one assay let along comparing between multiple, diverse assays. The situation became truly challenging when multiple assays of different technologies provided non-concordant results for the same intended target. Finally, in recent years, frameworks and minimum information standards have arisen. This annotation collates these recent developments and proposes a structured approach based on ToxCast assays to better address information capture about an assay and to organize them for focused communication.

The ToxCast assay library has been annotated using a customized, systematic method for assay data integration and aggregation. Reflective of one of the visions from MLP, ToxCast Phase I and II chemical sets are tested across all 328 ToxCast assays to comprise a complete data matrix of comprehensive chemical-biological data (Vempati et al. 2012). We successfully applied this annotation approach across 328 ToxCast *assays* with minimal terminology additions to allow for rapid data processing and quality control.

Numerous consortia have discussed the minimum information standards appropriate for quality reporting. These discussions have become technology-oriented, developing specific standards for subsets of assay technologies (Goetz et al. 2011; Patlewicz et al. 2013). The BAO framework was constructed with a wide breadth of coverage, capturing the NCATS reporting parameters and parameters from other technologies (Sittampalam et al. 2012; Vempati et al. 2012). Our approach applies structure to the BAO framework, where different annotations (e.g. *content readout type, detection technology name, intended target family*) may flag the use of certain guidelines. This unifies the annotation standards though it requires critical review of the annotation databases and their limitations. In the future, it may be useful to have these guidelines incorporated alongside the *detection technology* and design annotations.

Two novel concepts were developed as a result of this study. The first concept is the hierarchical assignment of assay annotations as attributes of a particular level. Using the order displayed in Figure 3A, the lineage between an *assay* to the *assay component endpoints* can be defined for conventional usage. Following this example, we can communicate about 'Bioseek' (name of assay source--BSK) with an interest in the BSK_3C *assay* (name of the experiment). Among its readouts, the *assay component* BSK_3C_IL8 (name of the raw readout) was analyzed to produce the *assay component endpoint* BSK_3C_IL8_down (name of the analyzed readouts). Conversely, the assay component endpoint can be tracked back through its lineage to identify other assay components and assay components that may be have generated from the same experiment, such as BSK_3C_IL8_up.

A similar approach is taken by the BioAssay Research Database (BARD), a collaborative project by MLP that also uses BAO as their foundational framework. BARD applies a higher

level of separation where HTS projects are broken down into experiments and further into individual assays. The BARD approach can be adopted into our approach. It essentially stops short of the *assay component* and *assay component endpoint* levels, but it is in agreement that annotations can be separated perspectively by level for better organization and communication.

Secondly, the assay target annotations were developed to conceptually capture the technological target(s) for measurement as well as the intended objective target for assessment. Under the BAO framework, a subjective target needed to be selected. The vocabulary did not lend to separate that target as the reagent fluorescent probe or a gene of interest. Here, we use the *key assay reagent* annotation to set aside signal probes, and by taking this conceptual separation of technological versus intended targets a structure is established to capture targets that are more complex and abstract.

The design information annotations provide context for how the target was measured. By simply using the target annotations to filter assays, features such as the objective and experimental conditions may be ignored. The study design information is the key component for reproducibility issues and often questioned when discussing the influencing factors. For instance, fluorescence screening assays are prone to artifact interference from test substances that may refract light. Nanomaterials, particulate matter, coloring agents, and volatile test chemicals may suffer from the artifact fluorescence interference as well as agglomeration and solubility issues, a reason for new assays to use the liquid-air interface culture technologies (Ghio et al. 2013). Knowing these limitations, one might say certain test substances are inapplicable to most HTS technologies, resonating the need for new, better assay technologies for activity testing.

Conclusion

As an organization tool, we've shown that the ToxCast Assay Annotations follows a logical structure that can filter down and simplify differences between assays. Combination use of the annotations, as seen with *assay design type* and *detection technology type*, can enhance these aggregations and differentiations. Similarly, combination use of the technological and intended target annotations can reveal the different modes of assessing the same intended target using different technological means. Furthermore, to support the filtering feature, the *intended target family* annotations were introduced to guide groups of target concepts.

A key idea generated from these results was that the annotations are indeed interconnected, and that these connections can help with reducing redundant annotations which can be grouped at a higher level. The flattened table approach for annotation, while useful for initial annotations, was too flat and made it difficult to convey the annotations. The annotations assigned to the *assay component endpoint* level are features that can only be separated at the most terminal level, whereas the annotations for the raw readouts (e.g. assay design, detection technology, and technological target) can be grouped up at the *assay component* level, and again for the *assay* level. These level groupings provide the organization necessary to concentrate the communication, keeping details focused on the experiment, the raw data, or the analyses separately.

Ultimately, the annotations are a support tool to understand the results of the chemical screening data. The combined display of multiple annotations can highlight the similarities and differences observable in the chemical hits. The principle component analysis showed that there are certain annotations terms, certain features belonging to the assays, which may be associable

to promiscuous hits and high variance in detected hits. The primary component was isolated to be differences in cell sensitivity and technology factors that may add stress on the testing environment. Therefore, while the annotations may support informed use of the chemical screening data, they can be applied to identify noise that may be truly or artificially stimulated.

Limitations and Future Directions

We recognize that there needs to be more definition for assays that target gene or pathway 'agonist' and 'antagonist' activities as well as for those that don't probe gene targets. Beyond the 328 assays, we also annotated over 100 pilot assays, discontinued assays, and in-progress assays. In particular, several annotations are in-development to capture 'antagonist' or 'agonist' target status, and to capture the *agonist stimulators* necessary for antagonist assays. Similar to the positive controls and reference chemicals used in each assay, the *agonist stimulators* were chosen from published literature and do not have clear justifications for the selection. As more chemical probes are discovered, this area and annotation should be further developed for better clarity. This also extends to metabolic and pharmacokinetics assays, which target certain chemical derivatives as the measure of gene-mediated biological processes.

Furthermore, new HTS assays that target developmental endpoint currently meet certain limitations within our annotation approach. Many of the zebrafish assays in the public domain look for time-dependent developmental effects, many of which are not gene targets but are malformations or disease-state targets (Sipes et al. 2011; Padilla et al. 2012; Truong et al. 2014). One solution is to incorporate more Gene Ontology biological process and cellular component terms or another source for formal vocabulary on developmental effects. However, these assays tend to use multiparametric approaches, which there currently is no formal vocabulary set for

certain analytical techniques and calculations. The same can be said of stem cell and pluripotent cell differentiation assays. To a lesser degree, multi-cell organotypic cultures are another promoted area of new assay development, which can be captured using our annotation approach with a few minor adjustments to the *assay format type* and *cell format* annotations.

As a byproduct of the assays that were used for annotation here, the target annotations have a bias towards gene-oriented assays. For cell-based assays, the targets can be generalized using the 'cell cycle' and 'cell morphology' *intended target families*. Annotation terms such as 'mitochondria' and 'nucleus' were incorporated into the target subtypes annotations. To separate the assay target as an indicator for localization versus organelle disruption would require the use of the *biological process target*. This might force the use of Gene Ontology *biological process target* annotations, so new annotations might be necessary to better capture cellular targets. Visual mapping of cellular event pathways or adverse outcome pathways will also aid in the representation of 'pathway' target types.

Descriptive elements of HTS assays have previously been used to aid the validation of assay results. Patlewicz et al. (2013) had focused on the analytical validation of the ToxCast Androgen Receptor (AR) assays, where differences in the human and rat *cell free component source* displayed variant binding capacities. ToxCast now includes a Chimpanzee AR binding assay, NVS_NR_cAR, which was compared against different AR homolog performance including a wild-type human AR recombinant expression in COS monkey kidney cells (Hartig et al. 2008). The ToxCast human AR assay uses a different AR protein extracted from LNCaP human Leydig cells (Knudsen et al. 2011). The utility in adding the chimpanzee AR assay is that it expands the detection sensitivity of perturbing the gene target across mammalian homologs

(Knudsen et al. 2011; Sipes et al. 2013). At the same time, it places more importance on annotating features of the experimental protocol to enable similar analyses to that of the AR assays. New annotations should be considered to capture species 'wild-type', 'polymorphic or mutant' gene variants, whether obtained through 'recombinant' expression or from 'endogenous' biological sources, or if there are other protein modifications such as 'ligand binding domain only' recombinant expression products.

Presently, there is not a minimal way to query for supplemental parameter readouts. As seen in the Tox21 AR beta-lactamase assay, we observe all assay components as inducible reporters but some as 'background control' or 'reporter gene' *assay function types.* This assay took multiple reads through which the Tox21_AR_BLA_Agonist_ch1 readout measures the baseline ('background control') while Tox21_AR_BLA_Agonist_ch2 and Tox21_AR_BLA_Agonist_ratio take the differential, 'reporter gene' comparisons (Huang et al. 2011). Similarly, ATG assays take reporter gene readouts (e.g. ATG_AR_TRANS), while a number of internal markers are used as 'background control 'parameters (e.g. ATG_GAL4_TRANS, ATG_M_06_TRANS), a common aspect of multiplexed assays (Romanov et al. 2008). A possible solution is to remind the user that 'signaling', 'reporter gene', or 'binding' *assay component endpoints* may need to be analyzed with relation to parameter *assay component endpoints* that are derived from the same *assay* and have 'viability' or 'background control' *assay function type*s.

The ToxCast Assay Annotation is presently used to support software developments with assay descriptions, like for the ToxCast Dashboard. This software tool is meant to promote informed use of the HTS data and consistent data representation in regulatory decision making.

Currently, the BioPortal interface can map between ontologies by their mutual annotation terms. This feature opens the possibility of ontology-based data integration, linking out with external databases, and normalization of data format heterogeneity and semantics. Judson et al. (2012) has described a knowledgebase essentially as a database supported with an ontology. The ontology makes the database better built to conduct search options. As new technologies become available, incorporating new vocabulary into the existing ontologies enables a consistent means to update the knowledgebase. Applying these functions to the ToxCast Dashboard can greatly improve its usability to accommodate the clients' work needs.

The 'assay component map' is a separate table meant to map the raw data files names to the appropriate assay component. This is a recent development that attempts to foster the data analysis pipeline with the annotation resource.

As mentioned, additional annotations are in-progress to highlight more features of an assay. Moving forward, the ToxCast Assay Annotation may incorporate parallel annotations to support a main annotation with the respective vocabulary and schema used by other ontologies in the same domain. For instance, the Brenda Tissue and Enzyme Source Ontology contains interesting annotation terms for endogenous enzyme source and anatomical entity hierarchy, which could be used alongside CL and CLO for *cell short name* annotations (Gremse et al. 2011). Similarly, the reagent and experimental components information, presently BAO annotation terms are used, but the Experimental Factor Ontology carries similar annotation terms which could also extend to the *assay format type* and *cell format* annotations.

Chapter 6

Indoor environmental health sampling, Science talk panels, and mammary gland tumor bioinformatics investigation: the Practicum experience with Silent Spring Institute

Joining Silent Spring Institute (SSI) for a two week practicum, I had the great pleasure and opportunity to listen in on several of these developments, learn from their experts, and contribute to them a segment of database and pathway informatics. After some conversations with SSI Director of Research and my practicum preceptor, Ms. Ruthann Rudel, the practicum was initiated with three intents, each having environmental science competencies. The first was to shadow field scientists and observe their exposure sampling protocol, which exercise diversity and cultural competencies. The second was to participate in Silent Spring Institute's scientific meetings and get a feel for how business and science operates in a non-government organization (NGO), which involve program planning discussions. The third was to work with SSI experts to learn about the Mammary Carcinogen Review Database, one of SSI's product resources, and to use it with a bioinformatics and cheminformatics approach, an application of communication and informatics skills.

Diversity and Cultural Competency

The goal of the first intent was to demonstrate awareness of and sensitivity to the varied perspective, norms, and values of others based on individual and ethnic differences. This was done by shadowing two specialists, Mr. Oscar Zarate from SSI and Ms. Meryl Corton from Harvard School of Public health. The two specialists were conducting an indoor exposure sampling study in a newly-occupied housing-community in the Boston area. On one sampling

event, a sampling module was set up in the home of a native Spanish-speaking participant, so Mr. Zarate administered the intake survey entirely in Spanish. The survey was very extensive and it is important to mention the necessity of interpretive clarifications; for instance, "Number of chairs" versus "Number of fabric-covered chairs", details which may be overlooked are separate survey questions within the intake form.

To efficiently get through the visit in 1 hr, I assisted Ms. Corton in setting up the module, sample for floor dust, and conducted the walk-through inspection. The engineered module was equipped with different passive air badges and particulate matter collectors, and it was set up to sample for 1 full week at a single location within the household. Like some of the intake questions, the walk-through inspection was a very eye-opening exercise, requiring awareness of direct and indirect exposure factors that may influence the indoor air quality in each room (e.g. the number of air purifiers, use of aromatic candles or air fragrances, mold spots, and chipped or cracked paint along the infrastructure), a careful reminder of what good housekeeping would include and what it could prevent.

Program Planning

The second intent was to participate in SSI's scientific meetings. These discussions identified the needs of the scientific and public community, and how to better meet these needs as a research entity. Some big concepts out of these meetings include consideration for who the proverbial stakeholders are, what it means to outreach and maintain community involvement, and what are the scientific products. For example, SSI has been a long-time researching entity for the Cape Cod, Massachusetts community. Having looked into the environmental contaminants, pesticides used, and the unusually high incidence of cancer and health concerns

from the region, what more could be done? Is the amount of research conducted sufficient? This resonates whether an NGO may get too niche or may have departed from specialization. This has implications for the scientific direction of the institute, the kind of skill sets and human resource to develop, and the kind of contribution and amount of people-time to spend towards a collaborative product.

The scientific meetings also demonstrate collective information sharing, a communication competency. It was similar to an academic lab meeting, where decisions made or results gathered are weighed as "useful or not," "may contribute towards policy or other publication products," or whether it may be good practice to compare with a respected data source or review summary. Through this, scientific results may get digested when reporting out. Particularly when reporting to local, non-scientific communities, phrasing and semantics need to be considerately and sensitively applied.

Communication and Informatics

Finally, the third intent involves collective information sharing, problem-solving, data interpretation, and considering genetic factors for adverse health outcomes. Though it was not able to be completed within the 2 week timeframe and follow-up developments have not been successful, this has been useful as a learning exercise in understanding different data resources. During the 2 week practicum, I was asked to present on a bioinformatics approach performed for a Human Leukemogen Project (Thomas et al. 2012) with the intent to do a similar analysis using a chemical set from the Mammary Carcinogen Review Database. This was a hypothesis generating exercise and an effort to identify chemical clusters based on enrichment of KEGG pathways. A few issues immediately surface: 1) unlike the Human leukemogens, the mammary

gland carcinogen chemicals are largely obtained from mouse and rat models, and it was not clear whether they lead to tumors of specific neoplastic or non-neoplastic types; 2) while the original plan was to use Comparative Toxicogenomics Database (CTD) files, the data files had missing and confusing data, and initial statistical approaches using Gene Set Enrichment Analysis (GSEA) failed to re-capitulated CTD displayed results; and 3) because the chemicals are mouse and rat actives, it would be important to consider enrichments of mouse and rat pathways.

To address these bioinformatics issues, with support from Drs. Lisa Truong and Richard Judson from EPA NCCT and Dr. Reuben Thomas from UC Berkeley, Ms. Janet Ackerman from SSI and I data-mined through different chemical-biological databases. For the first issue, we looked into the mouse and rat pathology data in Chemical Effects in Biological Systems database. However, while the data is rich, it required significant computing power and re-analysis of raw data; therefore, the tumor specificity categorizations could not be determined. Next, we identified that due to licensing limitations the CTD gene-pathway file was incomplete of certain KEGG pathways of interest and that an alternative source would be needed to consider the respective mouse or rat pathways. We then investigated the KEGG REST web services to extract pathway information, and we found that this can directly provide species-specific pathways, rather than conduct a gene-ortholog translation effort between species. Still, recapitulating the pathway enrichments failed; we discovered that CTD takes a human-health oriented approach so only human pathway enrichments would be displayed. Therefore, there is a need for more inter-species genetic and physiologic research to better understand the mechanisms of toxicity specificity and differences in susceptibility for translational science.

In addition, within the CTD chemical-gene interaction data set, it was discovered that transgenic studies make up a large proportion of the available mouse and rat data subsets (e.g. while the tested organism was a mouse species, the genes assessed are inserted from the human genome). These data sets have been corrected for using NCBI files for the respective rat, mouse and human genomes and will be re-run for enrichment patterns in the near future. This did, however, identify that rat and mouse chemical-gene data is actually very meager for the 273 chemicals of interest (800-1200 interactions in rat; 4000-6000 interactions in mouse) compared with the human data (50000-160,000 interactions in humans for the same set of 273 chemicals), whereas it was previously thought of as comparable numbers.

Through the bioinformatics exercise, it brings to light the limitations and challenges present in cross-species modeling (e.g. data gaps in testing, transgenic studies, and even coverage of species-specific pathways). Hopefully, after the enrichments have been re-run, we may see some clusters form out of the human data subset that may be similarly detected from the mouse or rat data subsets. Thereafter, we can analyze for a primary set of gene targets useful for characterization, and determine whether or not a certain gene-form (e.g. DNA, RNA, or protein) of the gene targets contributes most to the characterization. To the larger picture, these outputs may help in understanding the mammary gland carcinogens and their similarities and differences in physiologic responses between human, mouse, or rat models.

APPENDIX

Appendix 1: Orthologous Gene Targets

- Intended target gene symbols that are non-human genes are highlighted in blue.
- The rank column is used to order the groups of *assay component endpoints* in terms of their number of orthologous assays

Assay component endpoint	Intended target type	Intended target subtype	Intended target family	Intended target subfamily	Intended target gene ID	Intended target gene symbol	Rank
ACEA_T47D_80hr_Positive	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
ATG_ERa_TRANS	protein	transcription factor	nuclear receptor	steroidal	2099	ESR1	1
ATG_ERE_CIS	protein	transcription factor	nuclear receptor	steroidal	2099	ESR1	1
NVS_NR_bER	protein	receptor	nuclear receptor	steroidal	407238	ESR1	1
NVS_NR_hER	protein	receptor	nuclear receptor	steroidal	2099	ESR1	1
NVS_NR_mERa	protein	receptor	nuclear receptor	steroidal	13982	Esr1	1
OT_ER_ERaERa_0480	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
OT_ER_ERaERa_1440	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
OT_ER_ERaERb_0480	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
OT_ER_ERaERb_1440	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
OT_ERa_EREGFP_0120	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
OT_ERa_EREGFP_0480	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
OT_ERa_ERELUC_AG_1440	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
OT_ERa_ERELUC_ANT_1440	pathway	pathway-specified	nuclear receptor	steroidal	2099	ESR1	1
Tox21_ERa_BLA_Agonist_ch2	protein	receptor	nuclear receptor	steroidal	2099	ESR1	1
Tox21_ERa_BLA_Agonist_ratio	protein	receptor	nuclear receptor	steroidal	2099	ESR1	1
Tox21_ERa_BLA_Antagonist_ratio	protein	receptor	nuclear receptor	steroidal	2099	ESR1	1
Tox21_ERa_LUC_BG1_Agonist	protein	receptor	nuclear receptor	steroidal	2099	ESR1	1
Tox21_ERa_LUC_BG1_Antagonist	protein	receptor	nuclear receptor	steroidal	2099	ESR1	1
ATG_AR_TRANS	protein	transcription factor	nuclear receptor	steroidal	367	AR	2
NVS_NR_cAR	protein	receptor	nuclear receptor	steroidal	747460	AR	2
NVS_NR_hAR	protein	receptor	nuclear receptor	steroidal	367	AR	2
NVS_NR_rAR	protein	receptor	nuclear receptor	steroidal	24208	Ar	2
OT_AR_ARELUC_AG_1440	pathway	pathway-specified	nuclear receptor	steroidal	367	AR	2
OT_AR_ARSRC1_0480	pathway	pathway-specified	nuclear receptor	steroidal	367	AR	2
OT_AR_ARSRC1_0960	pathway	pathway-specified	nuclear receptor	steroidal	367	AR	2
Tox21_AR_BLA_Agonist_ch2	protein	receptor	nuclear receptor	steroidal	367	AR	2
Tox21_AR_BLA_Agonist_ratio	protein	receptor	nuclear receptor	steroidal	367	AR	2
Tox21_AR_BLA_Antagonist_ratio	protein	receptor	nuclear receptor	steroidal	367	AR	2
Tox21_AR_LUC_MDAKB2_Agonist	protein	receptor	nuclear receptor	steroidal	367	AR	2
Tox21_AR_LUC_MDAKB2_Antagonist	protein	receptor	nuclear receptor	steroidal	367	AR	2

ATG_PPARg_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	5468	PPARG	3
ATG_PPRE_CIS	protein	transcription factor	nuclear receptor	non-steroidal	5468	PPARG	3
NVS_NR_hPPARg	protein	receptor	nuclear receptor	non-steroidal	5468	PPARG	3
OT_PPARg_PPARgSRC1_0480	pathway	pathway-specified	nuclear receptor	non-steroidal	5468	PPARG	3
OT_PPARg_PPARgSRC1_1440	pathway	pathway-specified	nuclear receptor	non-steroidal	5468	PPARG	3
Tox21_PPARg_BLA_Agonist_ch2	protein	receptor	nuclear receptor	non-steroidal	5468	PPARG	3
Tox21_PPARg_BLA_Agonist_ratio	protein	receptor	nuclear receptor	non-steroidal	5468	PPARG	3
APR_p53Act_1h_dn	pathway	pathway-specified	dna binding	tumor suppressor	7157	TP53	4
APR_p53Act_1h_up	pathway	pathway-specified	dna binding	tumor suppressor	7157	TP53	4
APR_p53Act_24h_dn	pathway	pathway-specified	dna binding	tumor suppressor	7157	TP53	4
APR_p53Act_24h_up	pathway	pathway-specified	dna binding	tumor suppressor	7157	TP53	4
APR_p53Act_72h_dn	pathway	pathway-specified	dna binding	tumor suppressor	7157	TP53	4
APR_p53Act_72h_up	pathway	pathway-specified	dna binding	tumor suppressor	7157	TP53	4
ATG_p53_CIS	protein	transcription factor	dna binding	tumor suppressor	7157	TP53	4
ATG_Ahr_CIS	protein	transcription factor	dna binding	basic helix-loop-helix protein	196	AHR	5
Tox21_AhR	protein	receptor	dna binding	basic helix-loop-helix protein	196	AHR	5
ATG_CAR_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	9970	NR1I3	6
ATG_PBREM_CIS	protein	transcription factor	nuclear receptor	non-steroidal	9970	NR1I3	6
NVS_NR_hCAR_Agonist	protein	receptor	nuclear receptor	non-steroidal	9970	NR1I3	6
NVS_NR_hCAR_Antagonist	protein	receptor	nuclear receptor	non-steroidal	9970	NR1I3	6
ATG_DR5_CIS	protein	transcription factor	nuclear receptor	non-steroidal	5914	RARA	7
ATG_RARa_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	5914	RARA	7
NVS_NR_hRAR_Antagonist	protein	receptor	nuclear receptor	non-steroidal	5914	RARA	7
NVS_NR_hRARa_Agonist	protein	receptor	nuclear receptor	non-steroidal	5914	RARA	7
ATG_FXR_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	9971	NR1H4	8
ATG_IR1_CIS	protein	transcription factor	nuclear receptor	non-steroidal	9971	NR1H4	8
NVS_NR_hFXR_Agonist	protein	receptor	nuclear receptor	non-steroidal	9971	NR1H4	8
NVS_NR_hFXR_Antagonist	protein	receptor	nuclear receptor	non-steroidal	9971	NR1H4	8
ATG_GR_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	2908	NR3C1	9
ATG_GRE_CIS	protein	transcription factor	nuclear receptor	non-steroidal	2908	NR3C1	9
NVS_NR_hGR	protein	receptor	nuclear receptor	non-steroidal	2908	NR3C1	9
Tox21_GR_BLA_Agonist_ch2	protein	receptor	nuclear receptor	non-steroidal	2908	NR3C1	9
Tox21_GR_BLA_Agonist_ratio	protein	receptor	nuclear receptor	non-steroidal	2908	NR3C1	9
Tox21_GR_BLA_Antagonist_ratio	protein	receptor	nuclear receptor	non-steroidal	2908	NR3C1	9
ATG_PPARa_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	5465	PPARA	10
ATG_PPRE_CIS	protein	transcription factor	nuclear receptor	non-steroidal	5465	PPARA	10
NVS_NR_hPPARa	protein	receptor	nuclear receptor	non-steroidal	5465	PPARA	10
ATG_PXR_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	8856	NR1I2	11
ATG_PXRE_CIS	protein	transcription factor	nuclear receptor	non-steroidal	8856	NR1I2	11
NVS_NR_hPXR	protein	receptor	nuclear receptor	non-steroidal	8856	NR1I2	11

ATG_RXRa_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	6256	RXRA	12
OT_NURR1_NURR1RXRa_0480	pathway	pathway-specified	nuclear receptor	non-steroidal	6256	RXRA	12
OT_NURR1_NURR1RXRa_1440	pathway	pathway-specified	nuclear receptor	non-steroidal	6256	RXRA	12
ATG_TGFb_CIS	protein	transcription factor	growth factor	transforming growth factor beta	7040	TGFB1	13
BSK_BE3C_TGFb1_down	protein	protein-specified	growth factor	transforming growth factor beta	7040	TGFB1	13
BSK_BE3C_TGFb1_up	protein	protein-specified	growth factor	transforming growth factor beta	7040	TGFB1	13
BSK_KF3CT_TGFb1_down	protein	protein-specified	growth factor	transforming growth factor beta	7040	TGFB1	13
BSK_KF3CT_TGFb1_up	protein	protein-specified	growth factor	transforming growth factor beta	7040	TGFB1	13
ATG_THRa1_TRANS	protein	transcription factor	nuclear receptor	non-steroidal	7067	THRA	14
NVS_NR_hTRa	protein	receptor	nuclear receptor	non-steroidal	7067	THRA	14
Tox21_TR_LUC_GH3_Agonist	protein	receptor	nuclear receptor	non-steroidal	7067	THRA	14
Tox21_TR_LUC_GH3_Antagonist	protein	receptor	nuclear receptor	non-steroidal	7067	THRA	14
BSK_BE3C_MMP1_down	protein	protein-specified	protease	matrix metalloproteinase	4312	MMP1	15
BSK_BE3C_MMP1_up	protein	protein-specified	protease	matrix metalloproteinase	4312	MMP1	15
BSK_hDFCGF_MMP1_down	protein	protein-specified	protease	matrix metalloproteinase	4312	MMP1	15
BSK_hDFCGF_MMP1_up	protein	protein-specified	protease	matrix metalloproteinase	4312	MMP1	15
NVS_ENZ_hMMP1	protein	enzyme	protease	matrix metalloproteinase	4312	MMP1	15
NVS_ENZ_hMMP1_Activator	protein	enzyme	protease	matrix metalloproteinase	4312	MMP1	15
BSK_KF3CT_MMP9_down	protein	protein-specified	protease	matrix metalloproteinase	4318	MMP9	16
BSK_KF3CT_MMP9_up	protein	protein-specified	protease	matrix metalloproteinase	4318	MMP9	16
NVS_ENZ_hMMP9	protein	enzyme	protease	matrix metalloproteinase	4318	MMP9	16
NVS_ENZ_hMMP9_Activator	protein	enzyme	protease	matrix metalloproteinase	4318	MMP9	16
NVS_ADME_hCYP1A1	protein	enzyme	cyp	xenobiotic metabolism	1543	CYP1A1	17
NVS_ADME_hCYP1A1_Activator	protein	enzyme	cyp	xenobiotic metabolism	1543	CYP1A1	17
NVS_ADME_rCYP1A1	protein	enzyme	cyp	xenobiotic metabolism	24296	Cyp1a1	17
NVS_ADME_rCYP1A1_Activator	protein	enzyme	cyp	xenobiotic metabolism	24296	Cyp1a1	17
NVS_ADME_hCYP1A2	protein	enzyme	cyp	xenobiotic metabolism	1544	CYP1A2	18
NVS_ADME_hCYP1A2_Activator	protein	enzyme	cyp	xenobiotic metabolism	1544	CYP1A2	18
NVS_ADME_rCYP1A2	protein	enzyme	cyp	xenobiotic metabolism	24297	Cyp1a2	18
NVS_ADME_rCYP1A2_Activator	protein	enzyme	cyp	xenobiotic metabolism	24297	Cyp1a2	18
NVS_ADME_hCYP2E1	protein	enzyme	cyp	xenobiotic metabolism	1571	CYP2E1	19
NVS_ADME_hCYP2E1_Activator	protein	enzyme	cyp	xenobiotic metabolism	1571	CYP2E1	19
NVS_ADME_rCYP2E1	protein	enzyme	cyp	xenobiotic metabolism	25086	Cyp2e1	19
NVS_ADME_rCYP2E1_Activator	protein	enzyme	cyp	xenobiotic metabolism	25086	Cyp2e1	19
NVS_ENZ_hAChE	protein	enzyme	esterase	acetylcholinesterase	43	ACHE	20
NVS_ENZ_hAChE_Activator	protein	enzyme	esterase	acetylcholinesterase	43	ACHE	20
NVS_ENZ_rAChE	protein	enzyme	esterase	acetylcholinesterase	83817	Ache	20
NVS_ENZ_rAChE_Activator	protein	enzyme	esterase	acetylcholinesterase	83817	Ache	20
NVS_GPCR_bAdoR_NonSelective	protein	receptor	gpcr	rhodopsin-like receptor	282133	ADORA1	21
NVS_GPCR_hAdoRA1	protein	receptor	gpcr	rhodopsin-like receptor	134	ADORA1	21

NVS_GPCR_bDR_NonSelective	protein	receptor	gpcr	rhodopsin-like receptor	281125	DRD1	22
NVS_GPCR_hDRD1	protein	receptor	gpcr	rhodopsin-like receptor	1812	DRD1	22
NVS_GPCR_bH1	protein	receptor	gpcr	rhodopsin-like receptor	281231	HRH1	23
NVS_GPCR_hH1	protein	receptor	gpcr	rhodopsin-like receptor	3269	HRH1	23
NVS_GPCR_gLTB4	protein	receptor	gpcr	rhodopsin-like receptor	100379538	Ltb4r	24
NVS_GPCR_hLTB4_BLT1	protein	receptor	gpcr	rhodopsin-like receptor	1241	LTB4R	24
NVS_GPCR_gMPeripheral_NonSelective	protein	receptor	gpcr	rhodopsin-like receptor	100379235	Chrm3	25
NVS_GPCR_hM3	protein	receptor	gpcr	rhodopsin-like receptor	1131	CHRM3	25
NVS_GPCR_hAdra2A	protein	receptor	gpcr	rhodopsin-like receptor	150	ADRA2A	26
NVS_GPCR_rAdra2_NonSelective	protein	receptor	gpcr	rhodopsin-like receptor	25083	Adra2a	26
NVS_GPCR_hAdrb1	protein	receptor	gpcr	rhodopsin-like receptor	153	ADRB1	27
NVS_GPCR_rAdrb_NonSelective	protein	receptor	gpcr	rhodopsin-like receptor	24925	Adrb1	27
NVS_GPCR_hNTS	protein	receptor	gpcr	rhodopsin-like receptor	4923	NTSR1	28
NVS_GPCR_rNTS	protein	receptor	gpcr	rhodopsin-like receptor	366274	Ntsr1	28
NVS_GPCR_hOpiate_mu	protein	receptor	gpcr	rhodopsin-like receptor	4988	OPRM1	29
NVS_GPCR_rOpiate_NonSelective	protein	receptor	gpcr	rhodopsin-like receptor	25601	Oprm1	29
NVS_GPCR_rOpiate_NonSelectiveNa	protein	receptor	gpcr	rhodopsin-like receptor	25601	Oprm1	29
NVS_GPCR_hV1A	protein	receptor	gpcr	rhodopsin-like receptor	552	AVPR1A	30
NVS_GPCR_rV1	protein	receptor	gpcr	rhodopsin-like receptor	25107	Avpr1a	30
NVS_LGIC_bGABAR_Agonist	protein	receptor	ion channel	ligand-gated ion channel	282235	GABRA1	31
NVS_LGIC_bGABARa1	protein	receptor	ion channel	ligand-gated ion channel	282235	GABRA1	31
NVS_LGIC_rGABAR_NonSelective	protein	receptor	ion channel	ligand-gated ion channel	29705	Gabra1	31
NVS_MP_hPBR	protein	transporter	transporter	cholesterol transporter	706	TSPO	32
NVS_MP_rPBR	protein	transporter	transporter	cholesterol transporter	24230	Tspo	32
NVS_NR_bPR	protein	receptor	nuclear receptor	non-steroidal	280895	PGR	33
NVS_NR_hPR	protein	receptor	nuclear receptor	non-steroidal	5241	PGR	33
NVS_TR_gDAT	protein	transporter	transporter	neurotransmitter transporter	100714898	Slc6a3	34
NVS_TR_hDAT	protein	transporter	transporter	neurotransmitter transporter	6531	SLC6A3	34
NVS_TR_hAdoT	protein	transporter	transporter	nucleoside transporter	2030	SLC29A1	35
NVS_TR_rAdoT	protein	transporter	transporter	nucleoside transporter	63997	Slc29a1	35
NVS_TR_hNET	protein	transporter	transporter	neurotransmitter transporter	6530	SLC6A2	36
NVS_TR_rNET	protein	transporter	transporter	neurotransmitter transporter	83511	Slc6a2	36
NVS_TR_hSERT	protein	transporter	transporter	neurotransmitter transporter	6532	SLC6A4	37
NVS_TR_rSERT	protein	transporter	transporter	neurotransmitter transporter	25553	Slc6a4	37

Appendix 2: R Script for a Global Principle Components Analysis of ToxCast Assay Data

Prior to running this code, the "Ontology_realm" E-drive was established on an external
hard drive. It contains the ToxCast Phase II Data release screening data, design info data
file, and target info data files were pre-processed to calculate $-Log_{10}$(value/100000)
transformed values.

```
# Progression of objects:
#
# Design_info -> Design_melt -> Design_melted merges with Target_melted -> AxM
# Target_info -> Target_melt -> Target_melted merges with Design_melted -> AxM
# AxM merges into MxC_prep -> MxC_prep -> MxC (end)
# DATA_LEVEL6 -> DL6
# DL6 + AxM -> MxC_prep

# install.packages("Rtools")
# install.packages("devtools")
# install.packages("reshape2")
# install.packages("Rcpp")
library(Rcpp)
library(devtools)

# set directory #
annotation.dir <- "E:/Ontology realm"
setwd(annotation.dir)


# Read in Data release and Annotations#
list.files()
DATA_LEVEL6 <- read.csv ("ToxCast_E1k_1858_LEVEL6_AC50_COMPLETE_2014MAR17.csv",
header=TRUE)
head(DATA_LEVEL6)[1:10]
dim(DATA_LEVEL6)

Design_info <- read.csv ("toxcast_assay_annotation_study_design_Mar2014.csv", header=TRUE)
Design_info <- Design_info[order(Design_info$assay_component_endpoint_name),]
Target_info <- read.csv ("toxcast_assay_annotation_target_info_Mar2014.csv", header=TRUE)
Target_info <- Target_info[order(Target_info$assay_component_endpoint_name),]


# Merging the BSK CASM3C gaps with the SM3C data #
BSK_mod <- colnames(DATA_LEVEL6)[grep("BSK_CASM3C|BSK_SM3C",colnames(DATA_LEVEL6))]
#BSK_mod
sum(!is.na(DATA_LEVEL6[,BSK_mod[1]])) #775 chemicals with CASM3C data
sum(!is.na(DATA_LEVEL6[,BSK_mod[29]])) #292 chemicals with SM3C data

for (i in 1:28){ #for each line of CASM3C that is NA, replace with the matching SM3C data
  j <- i+28
  DATA_LEVEL6[is.na(DATA_LEVEL6[,BSK_mod[1]]),BSK_mod[i]] <-
DATA_LEVEL6[is.na(DATA_LEVEL6[,BSK_mod[1]]),BSK_mod[j]]
}
sum(!is.na(DATA_LEVEL6[,BSK_mod[1]])) #1058 chemicals with CASM3C data
```

```
# id and remove APR 1hr (Phase 1 only) data #
APR_mod <- colnames(DATA_LEVEL6)[grep("APR_",colnames(DATA_LEVEL6))]
APR_mod <- APR_mod[grep("_1h_",APR_mod)]
sum(!is.na(DATA_LEVEL6[,APR_mod[15]])) #292 chemicals with each APR 1hr assay

# remove the SM3C and APR_1hr data from further usage #
DATA_LEVEL6 <- DATA_LEVEL6[,!colnames(DATA_LEVEL6) %in% BSK_mod[29:56]]
DATA_LEVEL6 <- DATA_LEVEL6[,!colnames(DATA_LEVEL6) %in% APR_mod]
rm(BSK_mod, APR_mod, i, j)


# create annotation:term pairs #
desc_term_pair <- function (x) {
  y <- matrix(NA,nrow=nrow(x),ncol=ncol(x))
  for (i in 1: nrow(x)) {
    for (j in 1:length(x)) {
      #y[i,j] <- toupper(paste(names(x)[j],":",x[i,j],sep=""))
      y[i,j] <- paste(names(x)[j],":",x[i,j],sep="")
    }
  }
  colnames(y) <- colnames(x)
  y <- cbind(as.character(x$assay_component_endpoint_name),y)
  colnames(y)[1] <- "assay_component_endpoint"
  y <- y[, colnames(y)!="assay_component_endpoint_name"]
}
Design_melt <- desc_term_pair(Design_info)
head(Design_melt)

Target_melt <- desc_term_pair(Target_info)
head(Target_melt)


# Separate out the description variables, annotation IDs, and other variables to be removed from PCA#
remove <- c("_desc","assay_source_long_name","assay_name","assay_component_name"
            ,"organism_id","target_gene_id"
            ,"wavelength_","key_assay_","key_positive_control","dilution_solvent_","timepoint_"
            )

Design_melt <- as.matrix(Design_melt[,grep(gsub(", ","\\|",toString(remove)),colnames(Design_melt),
invert=TRUE)])
Target_melt <- as.matrix(Target_melt[,grep(gsub(", ","\\|",toString(remove)),colnames(Target_melt),
invert=TRUE)])

colnames(Design_melt)
colnames(Target_melt)


# remove duplicate columns except for "the "assay_component_endpoint" #
colnames(Design_melt)[colnames(Design_melt) %in% colnames(Target_melt)]
remove <- colnames(Design_melt)[colnames(Design_melt) %in% colnames(Target_melt)][-1]
Target_melt <- Target_melt[,!colnames(Target_melt) %in% remove]
Design_melt <- as.data.frame(Design_melt)
Target_melt <- as.data.frame(Target_melt)
```

```
# reshape the matrices #
library(reshape2)
Melted <- function(x) {
  y <- melt(x, id="assay_component_endpoint", na.rm=FALSE)
  y <- dcast(y, assay_component_endpoint~value, function(x) 1, fill=0)
}

Design_melted <- Melted(Design_melt)
head(Design_melted,30)

Target_melted <- Melted(Target_melt)
head(Target_melted,30)


# merge the two matrices for uniform representation of Design and Target #
AxM <- merge(Design_melted, Target_melted, by="assay_component_endpoint", all=TRUE)
rownames(AxM) <- AxM$assay_component_endpoint
colnames(AxM)
dim(AxM)
head(AxM,10)[1:10]

plot(colSums(AxM[,-1]), xlim=c(0,300), #ylim=c(0,100),
    ylab="No. of Assays", xlab="AxM term id", main="No. of Assays across the AxM terms")
summary(colSums(AxM[,-1]))


# checking the format of the AxC table #
tail(DATA_LEVEL6)
retired <- colnames(DATA_LEVEL6)[!colnames(DATA_LEVEL6) %in% rownames(AxM)][-(1:3)]
retired
#### Note: ATG perc series, Tox21 ch1+ch2, & Tox21 mito fitc+rhodamine
#### have data released, but not the annotations for them


# Generate the MxC matrix #
names(AxM)
AxM[1:5,1:5] #still has the assay_component_endpoint column
AxM[,"assay_component_endpoint"]
dim(AxM)
dim(DATA_LEVEL6)


# Generate the chemical data as.numeric and as AxC #
head(DATA_LEVEL6)[1:6]
DL6 <- matrix(0, ncol=nrow(DATA_LEVEL6), nrow=ncol(DATA_LEVEL6)-3)
for (i in 1:nrow(DATA_LEVEL6)){
  x <- lapply(DATA_LEVEL6[i,4:ncol(DATA_LEVEL6)], as.character)
  x <- unlist(lapply(x, as.numeric))
  DL6[,i] <- x
}
class(DL6[,1845]) #check to see if it is numeric
tail(DL6)
DL6 <- as.data.frame(DL6)
rownames(DL6) <- colnames(DATA_LEVEL6)[-(1:3)]
```

```
# Excluding the retired assays and selecting the chemical identifiers to keep #
colnames(DL6) <- DATA_LEVEL6[,"TS_ChemName"] #assigning the chemical column names
DL6 <- cbind(rownames(DL6),DL6)
head(DL6)[1:5]
colnames(DL6)[1] <- "assay_component_endpoint"
DL6 <- DL6[!rownames(DL6) %in% retired,] #removing retired assays
DL6[1:10,1:10]
dim(DL6)


# Remove the chemicals with too many NAs #
plot(colSums(is.na(DL6)),
    xlab = "Chemical ID", ylab = "No of NA values for Assays", main = "Distribution of NA chemical average
values")
summary(colSums(is.na(DL6)))
  #graph and summary table suggests that the magic cutoff numbers are 100, 150, 300, and 600)
maybeE1K <- colnames(DL6)[colSums(is.na(DL6))>500]
length(maybeE1K)
DL6 <- as.data.frame(DL6[,!colnames(DL6) %in% maybeE1K])
length(colnames(DL6)) #1059 + assay_component_endpoint_names


# Generating the wide file #
colnames(DL6)[1:10]
dim(AxM)
AxM <- AxM[rownames(AxM) %in% rownames(DL6),]

filtered <- colnames(AxM[,-1])[colSums(AxM[,-1])<1]
length(filtered)
filtered[1:100]
AxM <- as.data.frame(AxM[,!colnames(AxM) %in% filtered])
AxM[,1]

AxM[1:2]
MxC_prep <- merge(DL6, AxM, by="assay_component_endpoint")
head(MxC_prep)[c(1:10,1060:1069)]
dim(DL6) # 783 by 1060


# Getting rid of the 'assay_component_endpoint' columns #
row.names(MxC_prep) <- MxC_prep[,"assay_component_endpoint"]
MxC_prep <- MxC_prep[,!names(MxC_prep) %in% "assay_component_endpoint"]
DL6 <- DL6[,!names(DL6) %in% "assay_component_endpoint"]
AxM <- AxM[,!names(AxM) %in% "assay_component_endpoint"]
head(MxC_prep)[1:5]

MxC <- matrix(NA, nrow=ncol(AxM), ncol=ncol(DL6))
dim(MxC) # Metadata Annotations by Chemical data; AxM by dataset

dim(DL6) # Assays(+retired) by Chemical data
dim(AxM) # Assays by Metadata Annotations
dim(MxC_prep) # assays incommon by [Chemical data + Metadata Annotation]
MxC_prep[,1060]
```

```
colnames(MxC) <- colnames(DL6)
rownames(MxC) <- colnames(AxM)

# Calculate the average(-log10(AC50)) for each chemical to annotation term #
x <- MxC_prep[,1:1059]
ptm <- proc.time()
  for (j in 1000:ncol(AxM)) {
    k <- j+ncol(DL6)
    y = MxC_prep[,k]
    MxC_p <- data.frame(x,y)
    MxC_p <- MxC_p[which(MxC_p$y=="1"),]
    z <- aggregate(MxC_p, by=list(MxC_p[,"y"]), FUN=mean, na.rm=TRUE)
    MxC[j,] <- unlist(z[1,2:1060])
  }
proc.time()-ptm
MxC[1000:1050,1:5]

rownames(MxC)[is.na(rowSums(MxC))]

rm(i,j,k,x,y,z,MxC_p,ptm)


# writing the MxC table #
write.csv(MxC,"MetaxChem_averagescores.csv")
# PCA prep #1
install.packages("gplots")

# Check for the number of NAs per chemical #
plot(colSums(is.na(MxC)))

# There are still too many NA and NaN values. Trying Matt's solution: Avg()=>NaN) #
MxC_1 <- MxC
for (i in 1:nrow(MxC)) {
  if (sum( is.na(MxC[i,]) )>0) {
    MxC_avg <- mean(MxC[i,],na.rm=TRUE)
    MxC_1[i,] <- as.numeric(gsub(NaN,MxC_avg, unlist(MxC_1[i,])))
  }
}
rm(i)
MxC_1[1:10,1:10]
MxC_1 <- t(MxC_1)
  rownames(MxC_1)=NULL #remove the chemical names
  remove <- colnames(MxC_1)[colSums(MxC_1)==0] #remove features that have 0 hits
MxC_2 <- MxC_1[,!colnames(MxC_1) %in% remove]
MxC_2
write.csv(MxC_2, "MxC_data.csv")

try1 <- matrix(NA,ncol=ncol(MxC_2),nrow=2)
for (i in 1:ncol(MxC_2)){
  try1[1,i] <- mean(MxC_2[,i],)
  try1[2,i] <- sd(MxC_2[,i])
}

plot(try1[1,], col="blue")
```

```
par(new=T)
plot(try1[2,], col="red")
par(new=F)

MxC_pca <- prcomp(MxC_2,center=TRUE, scale=TRUE)
tail(MxC_pca)
names(MxC_pca)

PCA_loading <- MxC_pca$rotation #is.matrix
length(PCA_loading[,1])
write.csv(PCA_loading, "MxC_loading.csv")

Vimportance <- summary(MxC_pca) #generates the importance column
Vimportance <- Vimportance$importance #is.matrix
rownames(Vimportance)
write.csv(Vimportance, "MxC_importance.csv")


# examining the data #
dimnames(PCA_loading)
pairs(PCA_loading[,1:5]) #scatterplot of PC1 thru PC10

library(gplots)
my_palette <- colorRampPalette(c("dodger blue", "white", "orange"))(n = 299)
PCA_heat <- heatmap.2(PCA_loading[,1:5],dendrogram=c("row"),Rowv=TRUE,Colv=FALSE,trace="none",
col=my_palette,
                density.info="none", denscol=FALSE, key=TRUE)
```

References

Chen, B. and D. J. Wild (2010). "PubChem BioAssays as a data source for predictive models." Journal of Molecular Graphics and Modelling **28**(5): 420-426.

Congress, U. S. (1976). "Toxic substances control act." Public Law **99**: 469.

Dix, D. J., K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer and R. J. Kavlock (2007). "The ToxCast program for prioritizing toxicity testing of environmental chemicals." Toxicological Sciences **95**(1): 5-12.

Ghio, A. J., L. A. Dailey, J. M. Soukup, J. Stonehuerner, J. H. Richards and R. B. Devlin (2013). "Growth of human bronchial epithelial cells at an air-liquid interface alters the response to particle exposure." Particle and fibre toxicology **10**(1): 25.

Giuliano, K. A., A. H. Gough, D. L. Taylor, L. A. Vernetti and P. A. Johnston (2010). "Early safety assessment using cellular systems biology yields insights into mechanisms of action." Journal of biomolecular screening **15**(7): 783-797.

Goetz, A. K., B. P. Singh, M. Battalora, J. M. Breier, J. P. Bailey, A. C. Chukwudebe and E. R. Janus (2011). "Current and future use of genomics data in toxicology: Opportunities and challenges for regulatory applications." Regulatory Toxicology and Pharmacology **61**(2): 141-153.

Gremse, M., A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling and D. Schomburg (2011). "The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources." Nucleic acids research **39**(suppl 1): D507-D513.

Hartig, P., M. Cardon, C. Blystone, L. Gray and V. Wilson (2008). "High throughput adjustable 96-well plate assay for androgen receptor binding: a practical approach for EDC screening using the chimpanzee AR." Toxicology letters **181**(2): 126-131.

Houck, K. A., D. J. Dix, R. S. Judson, R. J. Kavlock, J. Yang and E. L. Berg (2009). "Profiling bioactivity of the ToxCast chemical library using BioMAP primary human cell systems." Journal of biomolecular screening **14**(9): 1054-1066.

Huang, R., M. Xia, M.-H. Cho, S. Sakamuru, P. Shinn, K. A. Houck, D. J. Dix, R. S. Judson, K. L. Witt and R. J. Kavlock (2011). "Chemical genomics profiling of environmental chemical modulation of human nuclear receptors." Environmental health perspectives **119**(8): 1142.

Inglese, J., C. E. Shamu and R. K. Guy (2007). "Reporting data from high-throughput screening of small-molecule libraries." Nature chemical biology **3**(8): 438-441.

Judson, R., A. Richard, D. J. Dix, K. Houck, M. Martin, R. Kavlock, V. Dellarco, T. Henry, T. Holderman and P. Sayre (2009). "The toxicity data landscape for environmental chemicals." Environmental Health Perspectives **117**(5): 685.

Judson, R. S., K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, H. M. Mortensen, D. M. Reif, D. M. Rotroff, I. Shah and A. M. Richard (2010). "In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project." Environmental health perspectives **118**(4): 485.

Judson, R. S., M. T. Martin, P. Egeghy, S. Gangwal, D. M. Reif, P. Kothiya, M. Wolf, T. Cathey, T. Transue and D. Smith (2012). "Aggregating data for computational toxicology applications: the US environmental protection agency (ePA) Aggregated Computational toxicology Resource (ACtoR) system." International journal of molecular sciences **13**(2): 1805-1831.

Kavlock, R., K. Chandler, K. Houck, S. Hunter, R. Judson, N. Kleinstreuer, T. Knudsen, M. Martin, S. Padilla and D. Reif (2012). "Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management." Chemical Research in Toxicology **25**(7): 1287-1302.

Kleinstreuer, N. C., D. J. Dix, K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, K. B. Paul, D. M. Reif, K. M. Crofton and K. Hamilton (2013). "In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis." Toxicological Sciences **131**(1): 40-55.

Knudsen, T. B., K. A. Houck, N. S. Sipes, A. V. Singh, R. S. Judson, M. T. Martin, A. Weissman, N. C. Kleinstreuer, H. M. Mortensen and D. M. Reif (2011). "Activity profiles of 309 ToxCast™ chemicals evaluated across 292 biochemical targets." Toxicology **282**(1): 1-15.

Knudsen, T. B. and N. C. Kleinstreuer (2011). "Disruption of embryonic vascular development in predictive toxicology." Birth Defects Research Part C: Embryo Today: Reviews **93**(4): 312-323.

Kunkel, E. J., I. Plavec, D. Nguyen, J. Melrose, E. S. Rosler, L. T. Kao, Y. Wang, E. Hytopoulos, A. C. Bishop and R. Bateman (2004). "Rapid structure-activity and selectivity analysis of kinase inhibitors by BioMAP analysis in complex human primary cell-based models." Assay and drug development technologies **2**(4): 431-442.

MacDonald, M. L., J. Lamerdin, S. Owens, B. H. Keon, G. K. Bilter, Z. Shang, Z. Huang, H. Yu, J. Dias and T. Minami (2006). "Identifying off-target effects and hidden phenotypes of drugs in human cells." Nature Chemical Biology **2**(6): 329-337.

Maglott, D., J. Ostell, K. D. Pruitt and T. Tatusova (2011). "Entrez Gene: gene-centered information at NCBI." Nucleic acids research **39**(suppl 1): D52-D57.

Manning, G., D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam (2002). "The protein kinase complement of the human genome." Science **298**(5600): 1912-1934.

Martin, M. T., D. J. Dix, R. S. Judson, R. J. Kavlock, D. M. Reif, A. M. Richard, D. M. Rotroff, S. Romanov, A. Medvedev and N. Poltoratskaya (2010). "Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program." Chemical research in toxicology **23**(3): 578-590.

Martin, M. T., R. S. Judson, D. M. Reif, R. J. Kavlock and D. J. Dix (2009). "Profiling chemicals based on chronic toxicity results from the US EPA ToxRef Database." Environmental health perspectives **117**(3): 392.

Martin, M. T., T. B. Knudsen, D. M. Reif, K. A. Houck, R. S. Judson, R. J. Kavlock and D. J. Dix (2011). "Predictive model of rat reproductive toxicity from ToxCast high throughput screening." Biology of reproduction **85**(2): 327-339.

Meehan, T. F., A. M. Masci, A. Abdulla, L. G. Cowell, J. A. Blake, C. J. Mungall and A. D. Diehl (2011). "Logical development of the cell ontology." BMC bioinformatics **12**(1): 6.

Morisseau, C., O. Merzlikin, A. Lin, G. He, W. Feng, I. Padilla, M. S. Denison, I. N. Pessah and B. D. Hammock (2009). "Toxicology in the fast lane: application of high-throughput bioassays to detect modulation of key enzymes and receptors." Environmental health perspectives **117**(12): 1867.

NRC, N. R. C. C. o. T. T. A. o. E. A. (2007). Toxicity testing in the 21st century: a vision and a strategy, National Academies Press.

Padilla, S., D. Corum, B. Padnos, D. Hunter, A. Beam, K. Houck, N. Sipes, N. Kleinstreuer, T. Knudsen and D. Dix (2012). "Zebrafish developmental screening of the ToxCast™ Phase I chemical library." Reproductive Toxicology **33**(2): 174-187.

Patlewicz, G., T. Simon, K. Goyak, R. D. Phillips, J. Craig Rowlands, S. Seidel and R. A. Becker (2013). "Use and Validation of HT/HC Assays to Support 21st Century Toxicity Evaluations." Regulatory Toxicology and Pharmacology.

Przydzial, M. J., B. Bhhatarai, A. Koleti, U. Vempati and S. C. Schürer (2013). "GPCR ontology: development and application of a G protein-coupled receptor pharmacology knowledge framework." Bioinformatics **29**(24): 3211-3219.

Romanov, S., A. Medvedev, M. Gambarian, N. Poltoratskaya, M. Moeser, L. Medvedeva, M. Gambarian, L. Diatchenko and S. Makarov (2008). "Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors." Nature methods **5**(3): 253-260.

Rotroff, D. M., A. L. Beam, D. J. Dix, A. Farmer, K. M. Freeman, K. A. Houck, R. S. Judson, E. L. LeCluyse, M. T. Martin and D. M. Reif (2010). "Xenobiotic-metabolizing enzyme and transporter gene expression in primary cultures of human hepatocytes modulated by ToxCast chemicals." Journal of Toxicology and Environmental Health, Part B **13**(2-4): 329-346.

Rotroff, D. M., D. J. Dix, K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, D. M. Reif, A. M. Richard, N. S. Sipes and Y. A. Abassi (2013). "Real-Time Growth Kinetics Measuring Hormone Mimicry for ToxCast Chemicals in T-47D Human Ductal Carcinoma Cells." Chemical research in toxicology **26**(7): 1097-1107.

Sakamuru, S., X. Li, M. S. Attene-Ramos, R. Huang, J. Lu, L. Shou, M. Shen, R. R. Tice, C. P. Austin and M. Xia (2012). "Application of a homogenous membrane potential assay to assess mitochondrial function." Physiological genomics **44**(9): 495-503.

Sarntivijai, S., Z. Xiang, T. F. Meehan, A. D. Diehl, U. Vempati, S. C. Schürer, C. Pang, J. Malone, H. E. Parkinson and B. D. Athey (2011). "Cell Line Ontology: Redesigning the Cell Line Knowledgebase to Aid Integrative Translational Informatics." ICBO **833**.

Schürer, S. C., U. Vempati, R. Smith, M. Southern and V. Lemmon (2011). "BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets." Journal of biomolecular screening **16**(4): 415-426.

Sipes, N. S., M. T. Martin, P. Kothiya, D. M. Reif, R. Judson, A. M. Richard, K. Houck, D. Dix, R. Kavlock and T. Knudsen (2013). "Profiling 976 ToxCast chemicals across 331 enzymatic and receptor signaling assays." Chemical research in toxicology.

Sipes, N. S., M. T. Martin, D. M. Reif, N. C. Kleinstreuer, R. S. Judson, A. V. Singh, K. J. Chandler, D. J. Dix, R. J. Kavlock and T. B. Knudsen (2011). "Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data." Toxicological Sciences **124**(1): 109-127.

Sipes, N. S., S. Padilla and T. B. Knudsen (2011). "Zebrafish—As an integrative model for twenty-first century toxicity testing." Birth Defects Research Part C: Embryo Today: Reviews **93**(3): 256-267.

Sittampalam, G. S., N. Gal-Edd, M. Arkin, D. Auld, C. Austin, B. Bejcek, M. Glicksman, J. Inglese, V. Lemmon and Z. Li (2012). "Development and Applications of the Bioassay Ontology (BAO) to Describe and Categorize High-Throughput Assays."

Thomas, R., J. Phuong, C. M. McHale and L. Zhang (2012). "Using bioinformatic approaches to identify pathways targeted by human leukemogens." International journal of environmental research and public health **9**(7): 2479-2503.

ToxCast (2013). High-Throughput Chemical Screening Data from ToxCast & Tox21. O. o. R. a. D. Environmental Protection Agency, National Center for Computational Toxicology.

Truong, L., D. M. Reif, L. St Mary, M. C. Geier, H. D. Truong and R. L. Tanguay (2014). "Multidimensional In Vivo Hazard Assessment Using Zebrafish." Toxicological Sciences **137**(1): 212-233.

Vempati, U. D., M. J. Przydzial, C. Chung, S. Abeyruwan, A. Mir, K. Sakurai, U. Visser, V. P. Lemmon and S. C. Schürer (2012). "Formalization, Annotation and Analysis of Diverse Drug and Probe Screening Assay Datasets Using the BioAssay Ontology (BAO)." PloS one **7**(11): e49198.

Vempati, U. D. and S. C. Schürer (2004). "Development and Applications of the Bioassay Ontology (BAO) to Describe and Categorize High-Throughput Assays."

Visser, U., S. Abeyruwan, U. Vempati, R. P. Smith, V. Lemmon and S. C. Schürer (2011). "BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results." BMC bioinformatics **12**(1): 257.

Wang, Y., J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant (2009). "PubChem: a public information system for analyzing bioactivities of small molecules." Nucleic acids research **37**(suppl 2): W623-W633.

Williams-DeVane, C. R., M. A. Wolf and A. M. Richard (2009). "DSSTox chemical-index files for exposure-related experiments in ArrayExpress and Gene Expression Omnibus: enabling toxico-chemogenomics data linkages." Bioinformatics **25**(5): 692-694.

Yu, H., M. West, B. H. Keon, G. K. Bilter, S. Owens, J. Lamerdin and J. K. Westwick (2003). "Measuring drug action in the cellular context using protein-fragment complementation assays." Assay and drug development technologies **1**(6): 811-822.