INTEROPERABILITY IN TOXICOLOGY: CONNECTING CHEMICAL, BIOLOGICAL, AND COMPLEX DISEASE DATA

Sean Mackey Watford

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Gillings School of Global Public Health (Environmental Sciences and Engineering).

Chapel Hill
2019

Approved by:

Rebecca Fry

Matt Martin

Avram Gold

David Reif

Ivan Rusyn

© 2019
Sean Mackey Watford
ALL RIGHTS RESERVED

ABSTRACT

Sean Mackey Watford: Interoperability in Toxicology: Connecting Chemical, Biological, and Complex Disease Data
(Under the direction of Rebecca Fry)

The current regulatory framework in toxicology is expanding beyond traditional animal toxicity testing to include new approach methodologies (NAMs) like computational models built using rapidly generated dose-response information like US Environmental Protection Agency's Toxicity Forecaster (ToxCast) and the interagency collaborative Tox21 initiative. These programs have provided new opportunities for research but also introduced challenges in application of this information to current regulatory needs. One such challenge is linking *in vitro* chemical bioactivity to adverse outcomes like cancer or other complex diseases. To utilize NAMs in prediction of complex disease, information from traditional and new sources must be interoperable for easy integration. The work presented here describes the development of a bioinformatic tool, a database of traditional toxicity information with improved interoperability, and efforts to use these new tools together to inform prediction of cancer and complex disease. First, a bioinformatic tool was developed to provide a ranked list of Medical Subject Heading (MeSH) to gene associations based on literature support, enabling connection of complex diseases to genes potentially involved. Second, a seminal resource of traditional toxicity information, Toxicity Reference Database (ToxRefDB), was redeveloped, including a controlled vocabulary for adverse events used to map identifiers in the Unified Medical Language System (UMLS), thus enabling a connection to MeSH terms. Finally, gene to MeSH associations were used to evaluate the biological coverage of ToxCast for cancer to understand the capacity to use ToxCast to identify chemical hazard potential. ToxCast covers many gene targets putatively linked to cancer; however, more information on pathways in cancer progression is needed to identify robust associations between chemical exposure and risk of complex disease. The findings herein demonstrate that increased interoperability between data resources is necessary to leverage the large amount of data currently available to understand the role environmental exposures play in etiologies of complex diseases.

ACKNOWLEDGEMENTS

I must recognize and thank Katie Paul-Friedman for her mentorship and guidance through this program. Katie has been my supervisor and mentor at USEPA’s National Center for Computational Toxicology for the past two years. She has been an invaluable advocate supporting my interests as I develop expertise and skills to apply within public health and toxicology.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF ABBREVIATIONS x

CHAPTER 1: INTRODUCTION 1

Current issues in data interoperability to support computational toxicology and chemical safety evaluation 1

Current state of the data landscape in toxicology 3

Example Use Cases 9

Conclusion 13

Scope of dissertation 13

CHAPTER 2: NOVEL APPLICATION OF NORMALIZED POINTWISE MUTUAL INFORMATION (NPMI) TO MINE BIOMEDICAL LITERATURE FOR GENE SETS ASSOCIATED WITH DISEASE 15

Introduction 15

Methods 19

Results 28

Discussion 36

Summary 42

CHAPTER 3:TOXREFDB VERSION 2.0: IMPROVED UTILITY FOR PREDICTIVE AND RETROSPECTIVE TOXICOLOGY ANALYSES 44

Introduction 44

Meeting Challenges in ToxRefDB 2.0 47

Conclusions 72

Summary 73

CHAPTER 4: EVALUATING THE CANCER- RELATED BIOLOGICAL COVERAGE OF TOXICITY FORECASTER (TOXCAST): IDENTIFICATION OF KNOWLEDGE GAPS AND IMPLICATIONS FOR CHEMICAL SCREENING 74

Introduction 74

Methods 78

Results 86

Discussion 91

Summary 95

CHAPTER 5: CONCLUSIONS, PERSPECTIVES, AND FUTURE DIRECTIONS 96

APPENDIX 1: USEPA PRODUCTS SUPPORTING COMPUTATIONAL TOXICOLOGY 103

APPENDIX 2: METADATA FOR PUBMED ARTICLES CITING EITHER MARTIN ET AL. (2009) (28) OR MARTIN ET AL. (2009) (29) AS OF AUGUST 8, 2018 104

APPENDIX 3: SUBSET OF TOXREFDB ENDPOINTS AND EFFECTS CONSIDERED CANCER-RELATED FOR RESEARCH PURPOSES 109

APPENDIX 4: CANCER -RELATED ENDPOINTS FROM TOXREFDB AND THE CORRESPONDING TOXCAST INTENDED GENE TARGETS AND ASSAY ENDPOINTS 117

APPENDIX 5: RELEVANCE DECISIONS FOR TOP TEN TERMS FROM EACH REFERENCE LIBRARY 118

REFERENCES 134

LIST OF TABLES

Table 2.1: Equations for ranking GeneID-MeSH co-occurrences 24

Table 2.2: Manually curated resources used to construct EMCON 29

Table 2.3: Manual Precision for 17 selected MeSH terms 34

Table 2.4: Top MeSH terms for genes BRCA1, BRCA2, ESR1, ESR2, and PGR from EMCON 35

Table 3.1: Extraction progress as of ToxRefDB v2 release 57

Table 3.2: OCSPP 870 series health effects guidelines in ToxRefDB 61

Table 3.3: Observations for guideline profiles 62

Table 3.4: Guideline adherence scoring added to ToxRTool 64

Table 3.5: Effect profiles in ToxRefDB v2 for POD computation 66

Table 3.6: Number of datasets in ToxRefDB v2 for BMDS modeling 69

Table 4.1: Carcinogenesis concepts and the corresponding MeSH term(s) 83

Table 4.2: The number of MeSH terms used to query EMCON and the total number of genes retrieved from EMCON for the cancer concept 87

Table 4.3: Precision of enriched topics retrieved from each reference library for each CC 88

Table 4.4: Summary statistics for reference libraries 90

Table 4.5: Enriched topics with the highest number of ToxCast genes present in gene set 91

Table 5.1 Example use cases for using EMCON to address ongoing efforts in toxicology 101

LIST OF FIGURES

Figure 1.1: Evolving infrastructure to support modern chemical safety evaluation 6

Figure 2.1: MeSH tree branches for “Carcinoma, Ductal, Breast” 18

Figure 2.2: Workflow for building Entity MeSH Co-occurrence Network (EMCON) 21

Figure 2.3: Example of the naive GeneID-MeSH network 22

Figure 2.4: Example of MeSH term frequency normalization 23

Figure 2.5: EMCON NPMI distribution. 30

Figure 2.6: MeSH frequency normalization for two branches 31

Figure 2.7: Comparison of mean rank of ELR, breast cancer-specific gene set to random gene sets within EMCON search results 33

Figure 2.8: The rank comparison of the ELR gene set from EMCON and Génie 36

Figure 2.9: Number of genes retrieved for the 17 selected MeSH terms 41

Figure 3.1: Number of studies by study type and species in ToxRefDB v2 48

Figure 3.2: Three generation MGR example 49

Figure 3.3: ToxRefDB general schema with changes made from ToxRefDB v1 to ToxRefDB v2 50

Figure 3.4: Example of the controlled effect terminology in ToxRefDB v2 52

Figure 3.5: Terminology sources and membership of UMLS concept codes cross-referenced to ToxRefDB endpoints and effects 54

Figure 3.6: Data extraction and review workflow 55

Figure 3.7: Inference workflow to determine negative effects 62

Figure 3.8: Proportion of effects in ToxRefDB that can be modeled 70

Figure 3.9: The number and type of models for each dataset 71

Figure 4.1: Overall approach for identifying cancer-related gene sets to estimate biological coverage of ToxCast for cancer 80

Figure 4.2: Building the MeSH co-occurrence network 82

Figure 4.3: Filtering gene sets for cancer concepts “Immunomodulation” and “Angiogenesis” 87

Figure 4.4: The ToxCast coverage for cancer-related gene sets and enriched topics from four gene set reference libraries 90

Figure 5.1 Example IARC workflow for using ToxCast to identify carcinogenic potential of chemicals 102

LIST OF ABBREVIATIONS

ACToR	Aggregated Computational Toxicology Resource
AO	Adverse Outcome
AOP	Adverse Outcome Pathway
BMD	Benchmark dose
BMR	Benchmark Response
BRIDG	Biomedical Research Integrated Domain Group
CARC	Cancer Assessment Review Committee
CC	Cancer Concept
CDISC	Clinical Data Interchange Standards Consortium
CE	Critical Effect
CEBS	Chemical Effects in Biological Systems
CMAP	Connectivity Map
CPDB	Carcinogenic Potency Database
CTD	Comparative Toxicogenomics Database
CV	Controlled Vocabulary
ECHA	European Chemicals Agency
ELR	Expert Literature Review
EMCON	Entity MeSH Co-occurrence Network
FAIR	Findable, Accessible, Interoperable, Reusable
GEO	Gene Expression Omnibus
GWAS	Genome Wide Association Study
HAWC	Health Assessment Workspace Collaborative
HL7	Health Level-7
HTTK	High Throughput Toxicokinetics
HTTr	High Throughput Transcriptomics
IARC	International Agency for Research on Cancer

ICE	Integrated Chemical Environment
INHAND	International Harmonization of Nomenclature and Diagnostic Criteria
IRIS	Integrated Risk Information System
IUCLID	International Uniform Chemical Information Database
KE	Key Event
KER	Key Event Relationship
LEL	Lowest Effect Level
LOAEL	Lowest Observed Adverse Effect Level
MeSH	Medical Subject Heading
MGI	Mouse Genome Informatics
MIE	Molecular Initiating Event
MOA	Mode of Action
NAM	New Approach Methologies
NCCT	National Center for Computational Toxicology
NEL	No Effect Level
NER	Named Entity Recognition
NICEATM	NTP Interagency Center for the Evaluation of Alternative Toxicological Methods
NIH	National Institutes of Health
NLM	National Library of Medicine
NOAEL	No Observed Adverse Effect Level
NPMI	Normalized Pointwise Mutual Information
NTP	National Toxicology Program
OCSPP	Office of Chemical Safety and Pollution Prevention
OECD	Organisation for Economic Co-operation and Development
OHT	OECD Harmonized Templates
OMIM	Online Mendelian Inheritance in Man®
OPP	Office of Pesticide Programs
PMI	Pointwise Mutual Information

PMID	PubMed Identifier
PPRTV	Provisional Peer-Reviewed Toxicity Value
RESTful	Representational State Transfer
RGD	Rat Genome Database
SEND	Standards for Exchange of Nonclinical Data
TKCC	Ten Key Characteristics of Carcinogens
Tox21	Toxicity Testing in the 21st Century
ToxCast	Toxicity Forecaster
ToxPi	Toxicological Priority Index
ToxRefDB	Toxicity Reference Database
TR	Treatment related
TSCA	Toxic Substances Control Act
U.S. EPA	United States Environmental Protection Agency
UMLS	Unified Medical Language System

CHAPTER 1: INTRODUCTION

Current issues in data interoperability to support computational toxicology and chemical safety evaluation

Toxicology has been undergoing a period of rapid change and growth to meet the challenge of safety assessment for tens of thousands of chemicals with both potential environmental exposure and a lack of a complete dataset for hazard identification (1–4). After over a decade since the publication of the seminal National Research Council report, *Toxicity Testing in the 21st Century: A Vision and a Strategy* (5) calling for advancements in the field of toxicology using new approach methodologies (NAMs) (6,7), substantial progress has been made initially driven by the interagency collaboration for Toxicity Testing in the 21st Century (Tox21) (8,9) and the US Environmental Protection Agency (EPA) Toxicity Forecaster (ToxCast) program (10,11). These massive data generation efforts have produced dose-response information for chemical interactions with biological targets (2,9,12), and further motivated development of aggregated digital resources of legacy toxicity information (13,14) and software to access and analyze this information (15–20). Many fit-for-purpose applications have been developed to understand how to use the generated information. As a result, information is siloed, which prevents easy integration and exchange of data (i.e. interoperability) creating problems like inconsistent versioning, lack of provenance, and unnecessary duplication. Ultimately the consequence of the lack of data interoperability is that progress in understanding biological and toxicological effects of chemical exposures is hampered despite an abundance of information. To fully leverage the resulting information from NAMs for toxicology and public health questions, efforts must be applied to enabling connections between data sources (6). Overcoming the high opportunity cost of enabling interoperability of various data streams will help realize the following goals: rapidly and reproducibly associate NAM-based information with phenotypes and outcomes of interest for development of computational models (21–29), hypothesis generation aimed at increased mechanistic knowledge (30–35), and systematic literature reviews (36), all of which can support efficient and state-of-the-art screening level chemical safety evaluation (37).

Interoperability refers to “the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort” (38). Data interoperability can be accomplished through numerous means like development and adherence to controlled vocabularies (CVs) and compliance with formatting standards for exchange of data. Computational efforts in toxicology to generate and analyze massive amounts of data are relatively new, so CVs and formatting standards are not widely used and accepted. Of course, data interoperability challenges are not unique to toxicology and, in fact, are one of the key challenges facing each industry from finance to social media to public health and biomedicine (38–44). As an example of how interoperability promotes greater consumption of data for biological learning, platforms from companies like Affymetrix were developed to rapidly and affordably capture and analyze transcriptomic data. The application of Affymetrix platforms and other microarray technologies in a clinical setting was aided by standardization efforts (i.e. to support interoperability) for mass distribution of kits as well as standard reporting of results, which subsequently led to development of tool suites that could consume and analyze the information (45). The adherence to data formatting standards allowed for aggregation into a single resource called Gene Expression Omnibus (GEO), which allows access to the research community (46,47). For toxicology, the lack of consensus on how the vast amount of concentration-response data collected from a myriad of *in vitro* platforms can be applied to regulatory toxicology applications has clarified the need for implementation of data management strategies that maximize interoperability. For instance, “big data” is being generated via whole-genome sequencing (48), high-content imaging (49,50), and high-throughput screening (9,11), and how they are formatted, processed, analyzed, stored, and accessed are dissimilar, between data types and data generators, which creates an additional obstacle for data integration to answer applied questions. Building consensus on reporting standards, both for assay design principles and observed effects, would contribute to progress in the use of these data for regulatory applications.

Data interoperability is a salient and critical need to address if computational toxicology is to succeed in supporting modern chemical safety evaluation and research in public health and toxicology. Indeed, in alignment with the amended Toxic Substances Control Act (TSCA) (4) the EPA is required to develop a risk-based method for chemical prioritization and in doing so to use NAMs or the equivalent in lieu of traditional methods. A state has been reached in which volumes of data can be generated, but full utilization of these information to find creative scientific solutions absolutely necessitates taking the time to adopt improved data management practices in order to connect the appropriate data to a biological target and to understand the methodology employed. Good data management practices are embodied by the FAIR Data Principles or Findability, Accessibility, Interoperability, and Reusability (38). These principles were defined to guide existing and future endeavors in scientific research as technology advances and data is generated to support knowledge discovery. Without proper data interoperability, progress in other areas will remain limited. In support of public health research goals, these principles are echoed in the National Institutes of Health (NIH) Strategic Plan for Data Science, with emphasis on infrastructure development and support for good data management practices as a crucial effort for continued success (51). One of the first steps outlined in NIH's approach is to update the current NIH infrastructure by connecting related systems for increased data interoperability. The answer to the overall challenge of achieving interoperability is simple to describe but difficult to implement, not only due to mountains of legacy data trapped in antiquated or difficult to process formats, but also due to rapid data generation efforts with lack of standardization creating data "silos". Clearly the NIH has identified data interoperability as a key measure needed to achieve near and long-term goals for health-related research, but the field of toxicology needs more examination and consideration of why data interoperability is needed and how it could be achieved (52). The objective of this chapter is to provide the needed introduction to the current data landscape in toxicology, including specific use case examples that demonstrate a need for increased data interoperability for computational toxicology. As part of this introduction, research needs and key questions relevant to data interoperability in the public health and toxicology fields are highlighted.

Current state of the data landscape in toxicology

Toxicology is a diverse and applied field where health-related information from models of animal and/or human toxicity are translated into actionable items for chemical safety assessment. Decisions made based on toxicity data can not only dramatically affect human health and the environment but also have major economic implications. Thus, any changes to the existing paradigms for data collection, evaluation, and analysis as technology and science advances come under intense scrutiny. However, NAM-based data generation is proceeding, and myriad ongoing efforts continue to demonstrate how this information could be used to answer regulatory toxicology questions (21–23,53–58). In this section, the apparent lack of data interoperability for both traditional and NAM-based toxicity information are reviewed.

Much of the available traditional toxicology data for human health safety evaluations has been collected through animal experimentation to identify doses that do not cause adverse health effects and to identify hazards. This information is captured in physical and digital text documents. Many of these documents are used for regulatory purposes and not computationally accessible, e.g. the data are available for capture in text or PDF or in database formats that are not easily integrated. The existing information can be found in various formats scattered across different digital systems like Integrated Risk Information System (IRIS) (59), PubMed (60), https://www.regulations.gov, Chemical Effects of Biological Systems (CEBS) (61), eChemPortal (62), Provisional Peer-Reviewed Toxicity Value (PPRTV) (63), Carcinogenic Potency Database (CPDB) (64), and Toxicity Reference Database (ToxRefDB) (13). This exemplifies lack of interoperability that promotes duplication of information and challenges in data provenance that culminate in a lack of data interoperability. A specific example of these issues is the inability to identify identical National Toxicology Program (NTP) reports in databases that collect this information: CEBS (61), ToxRefDB (13), and CPDB (64). These resources are databases that have extracted data from animal toxicity studies conducted by NTP; however, the source documents are available as either full reports from various online locations or broken up as separate publications that can be found across different scientific journals. Because of source document management that was initiated without understanding of the future database needs (i.e. lack of versioning and unique identifiers) due to the age of some of the studies, as well as differences in how entities like a “study” are defined, it’s extremely difficult to identify the overlap between the two resources. These issues primarily encompass the legacy or historical data problems the field faces, but extensive efforts are under way to increase data interoperability to mitigate such issues. Addressing these challenges is critical as the field is rapidly changing because the success of new approaches often depends on the use of legacy information as a reference.

One of the first implementations of a public repository that integrated information on a by-chemical basis from *in vitro* bioassays and *in vivo* toxicology data from myriad public sources was the Aggregated Computational Toxicology Resource, or ACToR (14) (Figure 1.1A). ACToR initially provided the legacy information and access to two prominent projects within NCCT, ToxCast and ToxRefDB through a single web application and subsequently began developing Representational State Transfer (RESTful) web services (65) for increased availability of the resources. ToxCast and ToxRefDB, among others, moved the field forward because of the massive amount of information made available to explore computational modeling approaches to examine chemical hazard (13,21–23,25,66,67). With the progress made through ToxCast and Tox21, other projects grew into defined research areas or domains like "Exposure" (3) and "Use" (68) that spawned the development of different databases, applications, and software packages to meet specific research needs (Figure 1.1B). Many of these efforts have been siloed endeavors that have led to duplication of information across databases and difficulty in managing this information with time and resources spent on "data cleaning", version control, and quality assurance measures. Other centralized user interfaces for accessing both NAMs and traditional data, including the Comptox Chemicals Dashboard (17), NTP BioPlanet (18), National Library of Medicine (NLM) PubChem (69), Comparative Toxicogenomics Database (35), and NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) Integrated Chemical Environment (ICE) (20), have all led to increased access of data. Based on the rapid growth of these user interfaces, and their capabilities, it is clear that tools are needed to help data stakeholders integrate and organize this information, either by chemical or by biological process.

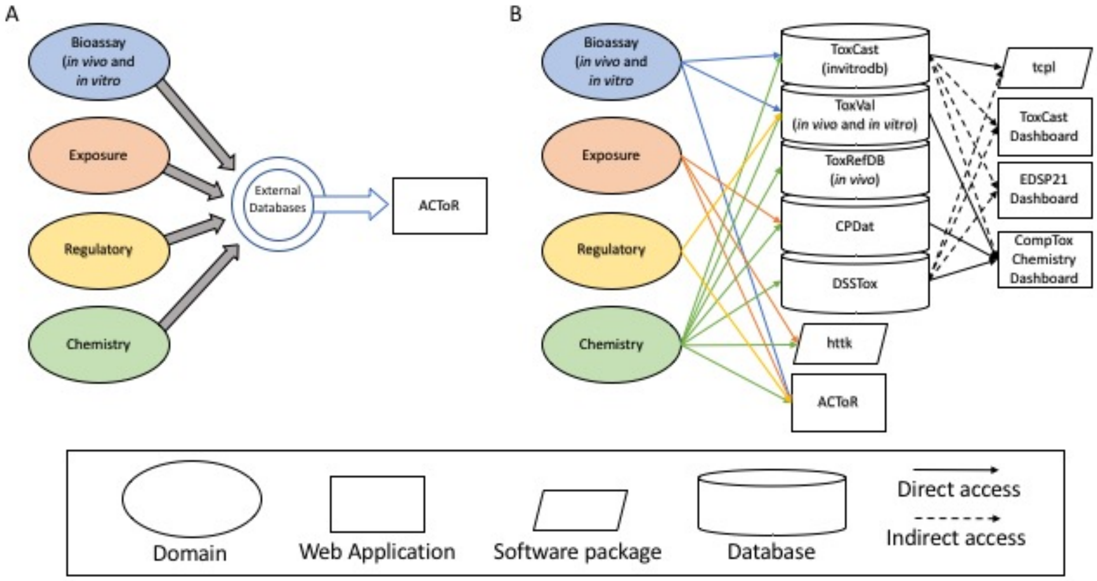


Figure 1.1: Evolving infrastructure to support modern chemical safety evaluation
Pictured is an abstract representation of the changing infrastructure that supports USEPA's computational toxicology efforts. (A) Initially, information across relevant domains in toxicology were aggregated from external databases to a single database accessed through a single web application called ACToR. (B) With continued success in data generation projects like ToxCast, multiple products were developed. The dashed arrows represent indirect access to the needed information. Indirect access means that the underlying information was duplicated because each web application is supported by a separate database, which is consistent with siloing and reinforcing data inconsistency. Appendix 1 further describes each product shown in this figure.

Modern chemical safety evaluation requires a framework to link relevant information

To enable modern chemical safety evaluation, several knowledge domains must be considered and integrated: (1) information on chemicals or substances; (2) information on chemical bioactivity, phenotypes, and toxicity; and, (3) information by testing methodology, assay principle, and intended target. Linking chemical exposure to disease or toxicity remains a challenge in part because chemical risk assessment is chemical centric, and information about the development and progression of disease or toxicity is thus not easily examined to allow inference of which chemical exposures are linked to these biological outcomes. Thus, efforts to increase data interoperability for bioactivity, toxicity, or phenotype and testing methodology, assay principle, or intended target would enable more inferences from disease or toxicity back to chemical exposure. Already, there is a framework to link biological observations: the adverse outcome pathway (AOP) (70–76). AOP networks are excellent for organizing information related to complex diseases and phenotypes; this cataloguing of interrelated biological processes and actions is advancing efforts in toxicology to use NAMs that may be used to predict adverse outcomes. However, novel, hypothetical associations between chemical exposure(s) and diseases may require bioinformatic tools and unsupervised approaches to putatively link chemical exposures with adverse outcomes. Though AOPs provide organization, in consideration of the AOP framework, it is clear that computationally accessible databases and additional bioinformatic tools to link information to AOPs rapidly are increasingly needed in modern chemical safety evaluation.

The Adverse Outcome Pathway (AOP) framework has been proposed as a method to aggregate relevant information together and define a set of processes contributing an adverse outcome (AO), or an event relevant to regulatory concern. An AOP is defined as "an analytical construct that describes a sequential chain of causally linked events at different levels of biological organization that lead to an adverse health or ecotoxicological effect" (77). An AOP is mapped as a linear progression of a series of key events (KEs) linked together by qualitative or quantitative key event relationships (KERs) across biological levels of organization, beginning with a specialized KE known as a molecular initiating event (MIE) and culminating with another specialized KE known as an adverse outcome (AO). As high-throughput screening, transcriptomics, epidemiology, etc. have generated large datasets for evaluating the effects of chemical exposure, the field of toxicology has been evolving to create new strategies for linking macromolecular and cellular changes with adverse outcomes to leverage all of this new data for safety evaluation. Key strategies for doing this include systematic literature review and predictive modeling (21–23,36). Either of these can inform a network of AOPs to describe biology pertinent to mechanisms of disease. However, problems persist in development of AOPs: (1) how can more AOPs be developed via rapid linkage of MIE or KE related information from NAM-based and traditional toxicology screening methods? (2) How can hypotheses be generated to suggest more potential MIE to AO associations? To address these persistent questions, additional tools need to be developed to make the NAM and traditional information more accessible, and further, existing information from disparate resources need to be leveraged to support discovery of novel MIE to AO relationships.

Relevant to the questions above, knowledge discovery and chemical safety evaluation for toxicology necessitate improvements to the FAIR data landscape as mentioned previously, thus enabling integration of data that were previously siloed, and these efforts are ongoing. To use NAMs, i.e. high-throughput transcriptomics, high-throughput screening, and high-content imaging data, more effectively in screening level assessment, better tools to link these data to MIEs, KEs, and AOs of interest for regulatory toxicology are needed. However, even the traditional toxicity information could benefit from more standardization and computationally accessibility for linkages to be established to AOs within AOPs.

A primary obstacle to interoperability of NAM and legacy-based information in siloed resources is balancing the need for domain-specific details with the need to reduce complexity to enable use of the data. There are several opposing drivers in balancing data complexity and simplicity, including the differences among data stakeholders. For instance, the needs of data scientists in terms of available tools and datasets may differ from the needs of the general public for data transparency and availability and the needs of regulatory toxicologists charged with making public health decisions. Currently, domain specific complexity has led to a number of standalone resources with separate databases and programming utilities (such as R packages to manage ToxCast or high-throughput toxicokinetic data) that require the user to develop a deeper understanding of how to integrate the resources.

Increased data interoperability of traditional toxicity and NAM-based toxicity information is discussed further in the facets of specific use cases. One example for hypothesis generation for AOP development is to relate complex disease phenotypes, such as cancer, with gene information that may be informative. The need to associate NAM information with outcomes at higher levels of biological organization necessitates high quality curation and structuring of legacy toxicology information, from animal and human studies. The archiving and curation of legacy data supports computational model development to predict adverse outcomes, but this process of data management is complex. An additional use case for computational toxicology is the integration of both NAM and traditional data resources to better understand how to predict complex outcomes and prioritize adverse outcome pathway development.

Example Use Cases

Standardization and mapping of vocabularies

As previously stated, development and adherence to formatting standards and CVs increases interoperability, especially for legacy information systems or new data streams without existing standards. The updates to Toxicity Reference Database (ToxRefDB) further described in Chapeter 3 (13) are an example of how legacy information can be modernized for easier integration and use to advance research. ToxRefDB is the largest publicly available digital resource aggregating results from animal toxicity studies that was initially created for retrospective analysis and as a reference to validate both ToxCast bioactivities and computational models (28,29). The impetus for the recent update was to collect dose-response information that was not originally extracted from the studies. However, endeavors to increase interoperability were also undertaken. The studies in ToxRefDB span decades where the language for reporting adverse events is inconsistent from either subjective expert preferences or updates as knowledge about pathology has advanced. The terminology in ToxRefDB was standardized for a ToxRefDB-specific CV. The CV was mapped to concepts in Unified Medical Language System (UMLS), which is a resource managed by National Library of Medicine integrating over 150 biomedical vocabularies into a semantic network (78). By mapping to a standard that is already integrated, interoperability is achieved with any other resource that is also mapped to the same standard.

A goal of this type of work will be to more easily pass information across different resources that could benefit from the information. For example, CEBS also captures information from animal toxicity studies and the adverse event reporting for histopathology results adheres to a CV called International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) and accounted for in NTP's Nonneoplastic Lesion Atlas (NLA) (79). Despite adherence to a CV, INHAND is not mapped to any other resources, which makes interoperability difficult. Since INHAND is not mapped to any of the UMLS vocabularies, interoperability between the reported adverse events in ToxRefDB and CEBS is difficult. However, a primary user of INHAND is the eTox consortium (80), which is a group of pharmaceutical companies that have compiled animal studies into a single resource. Continuing efforts of eTox include increased interoperability though ontology development and mapping (81). Another resource that collects animal toxicity information is the Health Assessment Workspace Collaborative (HAWC) (82). Although a

CV is available, it does not capture the granularity needed for interoperability. Finally, another resource that collects information on animal toxicity studies is International Uniform Chemical Information Database (IUCLID) (83). Like HAWC, IUCLID has a limited CV available as a “picklist” (84) and still lacks granularity for adverse events that is captured in ToxRefDB. IUCLID is the primary tool used by European Chemicals Agency (ECHA) to collect and evaluate chemicals for regulatory applications. IUCLID is separate from the previously mentioned applications because it adheres to data formatting standards developed in conjunction with Organisation for Economic Co-operation and Development (OECD) called OECD Harmonised Templates (OHTs). IUCLID can consume any data formatted according to OHTs. Both HAWC and IUCLID have been developed for chemical-centric regulatory applications, therefore aggregation of information has also been primarily chemical-centric. However, moving forward with research endeavors investigating NAMs and to answer questions about reproducibility in animal toxicity studies, the adverse event reporting also must adhere to CVs and formatting standards and fully support interoperability.

A massive amount of information is readily available from each of the information systems above, yet interoperability is still lacking primarily due to lack of CVs and data format standards. The progress made in ToxRefDB with CV development and mapping was a manual effort; however, automatic mapping is possible. Several tools like National Center for Biomedical Ontologies (NCBO) Bioportal Annotator and UMLS MetaMap are available map text to respective CVs using Natural Language Processing (NLP) techniques. Without definitions or full text input, these methods are limited to string comparisons, which are not always very accurate. For example, the ToxRefDB term “pathology microscopic” was manually mapped to the UMLS term “Histopathology Result”. When using the BioPortal Annotator, the UMLS terms that are mapped to “pathology microscopic” are “Pathology” and “Microscopic”, which, even together, do not represent “pathology microscopic” as well as “Histopathology Result”. In many cases, manually mapping terms may be the best option because of the accuracy, but automatic mapping pipelines can still be utilized and should always be investigated as an option.

Putative gene-outcome relationships for complex phenotypes

As previously stated, one of the most prominent challenges for adopting NAMs for chemical risk assessment is understanding how results can be applied to current public health issues like cancer. One approach from Kleinsteuer et al. (2013) (25) attempted to use odds ratios between *in vitro* bioactivity in ToxCast assay and cancer-related phenotypes in rodents, as documented in ToxRefDB, to develop chemical cancer hazard scores. Subsequently, the biological plausibility of links between ToxCast assays and ToxRefDB cancer outcomes was manually assessed by a literature review. The limitations of this approach were made clear in Cox et al. (2016) (85) stating that small changes to the dataset dramatically changed the results. This model instability could be the result of false positives i.e. the chemical bioactivity observed in ToxCast is not related to the cancer outcome observed in ToxRefDB. The approach could benefit if each ToxCast assay, which are linked to gene target(s), can be established in cancer AOPs. Indeed, this type of approach was taken by the International Agency for Research on Cancer (IARC) (86,87) where each ToxCast assay was reviewed and binned into the ten key characteristics of carcinogens (TKCC) (88). A toxicological priority index (ToxPi) was calculated for each chemical based on the bioactivities of each assay in each TKCC. Further, Becker et al (2017) (89) used the IARC binning of ToxCast assays and cancer designations by USEPA Office of Pesticides Program (OPP) Cancer Assessment Review Committee (CARC) as descriptors for machine learning models to classify chemicals as carcinogens, and ultimately concluded that ToxCast could not classify chemicals as carcinogens. Associating *in vitro* screening data with cancer-related outcomes, and understanding whether this is feasible and informative with ToxCast data, is an active area of research.

Thus, a challenge remains: the above strategies heavily rely on expert knowledge to establish biological links between gene targets and complex outcomes, based on previous understanding of the etiology and progression of these outcomes; however, expert knowledge is limited due to the reliance on low-throughput manual literature review, and cancer etiology, especially the role environmental chemicals play, is not well understood and may benefit from new information. Data-driven strategies can be considered as support for interoperability continues. A wealth of gene information is available that may be relevant to cancer etiology or other complex phenotypes that may be difficult for an expert to identify. For example, Chapter 2 describes a resource that links genes to Medical Subject Headings (MeSH), or keywords in literature. This resource is called Entity MeSH Co-occurrence Network (EMCON) (32) and can be used to identify genes that are linked to complex disorders like breast cancer. EMCON was also used as a data stream in Grashow et al. (2018) (90) as part of a comprehensive gene prioritization framework to identify a breast cancer gene panel. The impetus for this type of approach was linking dose-response gene expression profiles with adverse outcomes of interest. The traditional approach to analyze gene expression results is gene set enrichment analysis (GSEA), where pathways or other concepts are identified from overrepresented differentially-expressed genes in reference gene sets that are primarily manually curated (91). Relevant gene sets for understanding links between chemical exposures and complex phenotypes are not readily available. Most gene sets are available through Molecular Signatures Database (MSigDB) (91–93) or other resources for GSEA like Enrichr (94,95). A commonly used resource that links genes to disease is Online Mendelian Inheritance in Man® (OMIM) (96), which links genetic variants to disease; however, variants that have been linked to complex phenotypes have primarily been identified in genome wide association studies (GWAS) and are not always easy to mechanistically characterize.

A well-curated resource that attempts to link chemical exposures to disease is Comparative Toxicogenomics Database (CTD) (35). CTD curates qualified chemical-gene interactions from literature and integrates gene-disease relationships from primarily OMIM. The chemical-disease links are inferred according to the overlapping genes between chemical-gene interactions and gene-disease relationships (97). CTD is a high-quality resource but dependent on manual curation, which is low-throughput. Manual curation efforts cannot keep pace with the rate of publication, which highlights a need for alternative methods for data extraction like Named Entity Recognition (NER) (98,99). A comprehensive resource of bioassay information is PubChem (69). Bioassay information is crowdsourced, and deposition of information is generalized in order to store heterogenous data in the single resource. PubChem allows adherence to standards, specifically Bioassay Ontology (BAO) (100,101), but is not enforced. PubChem is the largest single resource for chemical bioassay information, however, much of the deposited information does not adhere to a data formatting standard beyond what is enforced in the deposition templates, which creates problems with interoperability like aggregating bioactivities by target. More work still needs to be considered for increased interoperability to utilize all available information for chemical safety evaluation.

To identify environmental exposures that could potentially influence susceptibility to a complex disease, more complex gene networks, not comprised of only variants, may be important to identify. Other manual curation efforts of literature continue for gene and gene function information (102); proteins (103,104); pathways (105–108); diseases (109); and chemicals (35,82,110). Aggregating this curated information with chemical dose-response information from new technologies targeting different levels of biological information, while also considering interoperability, could lead to rapid putative AOP development and development of robust chemical hazard computational models.

Conclusion

The current regulatory framework for toxicology is becoming more flexible to keep pace with modern public health needs. However, a major hurdle is data interoperability. Overcoming these challenges will enable researchers to interrogate available data to better understand the existing knowledge landscape identifying gaps in our understanding of environmental toxicity and how it influences complex disease. Initial efforts in toxicology to promote interoperability demonstrate immense progress and promise, yet, for continued success, more work is needed in development and adherence to CVs and data formatting standards as well as implementing modern infrastructures to support the massive amounts of data generated and subsequent analytics.

Scope of dissertation

The focus of this dissertation is on data interoperability and how efforts to increase data interoperability benefit toxicology and advance our understanding of how chemical exposures affect complex disease. Chapter 2 highlights a use case in leveraging curated information from literature to extract knowledge about a complex disease, breast cancer that would not have been accessible from expert review alone. The data is from articles in PubMed, which is a well-known resource supporting FAIR data guidelines to support scientific research. The articles are manually curated by PubMed indexers to extract information on genes as well identifying keywords called Medical Subject Headings (MeSH) that are a part of a CV, which is also one of the vocabularies in UMLS. With this information and the use of networks, a novel bioinformatic tool was built called Entity MeSH Co-occurrence Network (EMCON) that can be queried to identify putative genes linked to any outcome of interest that has been published within the curated set of literature. EMCON helps overcome issues in data interoperability between many sources of gene and gene-disease relationships. When considering applications within toxicology, this approach can be used to identify important targets of interest when screening chemicals as well as linking bioactivities to complex adverse events and disease.

Chapter 3 describes a major update to Toxicity Reference Database (ToxRefDB). The most significant update to support interoperability is establishing a ToxRefDB-specific CV and mapping the terms to UMLS. This work exposes points of integration, therefore data from ToxRefDB can be consumed and used with other datasets.

Finally, Chapter 4 utilizes work from the previous chapters to answer relevant questions about the use of ToxCast for identifying chemical hazard for cancer and complex disease. A major critique of the utility of ToxCast for building computational models for complex disease has been the lack of biological coverage within the dataset. In fact, this critique is the impetus to move forward with dose-response transcriptomics studies that capture many more targets. Recent efforts have attempted to use ToxCast to identify the cancer hazard for chemicals and a relevant question is: "does ToxCast have relevant biological coverage of cancer?" Because of the previous work, two data integration approaches were possible. One, involves the use of EMCON alone to identify ToxCast gene targets that are linked to cancer or cancer processes. The second uses both EMCON and ToxRefDB to identify genes linked to specific cancer outcomes observed in animal toxicity studies. The chapter concludes that ToxCast has several gene targets linked to cancer or cancer processes, but identifying what amount of biological coverage is necessary to identify chemical cancer hazard remains difficult. More work is needed that supports interoperability to continue building resources and investigating questions like those described in this dissertation.

CHAPTER 2: NOVEL APPLICATION OF NORMALIZED POINTWISE MUTUAL INFORMATION (NPMI) TO MINE BIOMEDICAL LITERATURE FOR GENE SETS ASSOCIATED WITH DISEASE[1]

Introduction

Rapid technological advances in biomedical and life sciences have led to an inundation of heterogeneous information across these fields. For instance, high-throughput technologies like RNA-Seq and other transcriptomics methods generate large amounts of gene-related data such that underlying biological processes can be illuminated through pathway enrichment or association (91,111). Linking these functional genomic data to well-characterized ubiquitous diseases such as breast cancer (112) could provide opportunities to derive additional etiological insight, generate new hypotheses, identify critical genes and pathways, and develop novel therapeutics. However, while the current volume of genetic, experimental, toxicological and other data presents an incredible opportunity for biomedical and basic science to make great knowledge gains, many challenges remain in understanding how this information can be best integrated and queried to produce valuable insight.

Currently, there is no research precedent on how to link genetic and toxicological data to complex disease phenotypes. For example, given that breast cancer will affect one in eight U.S. women and that susceptibility is shaped by both genetic and environmental factors (113–118), it is worthwhile to query publicly available data resources to better understand how risk factors like chemical exposures initiate biological changes to increase disease risk (119–121). Such an approach represents a departure from conventional toxicological strategies, in which a single chemical exposure is investigated as the driver of adverse effects, rather than considering other components of risk, e.g. genetic and lifestyle factors (122). As an alternative, the strategy outlined in this study aligns with the Adverse Outcome Pathway (AOP)

[1] This chapter previously appeared as an article in Computational Toxicology. The original citation is as follows: Watford, S., Grashow, R., De La Rosa, V., Rudel, R., Paul-Friedman, K., Martin, M. (2018). Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis. *Computational Toxicology, 7*, 46-57.

conceptual framework that focuses on aggregating information on perturbed systems across levels of biological organization (70,123). However, development of AOPs faces the same challenges in linking molecular initiating events (MIE), subsequent key events (KE), and adverse outcomes (AOs) in that the most relevant MIEs of KEs for a given AO may not be known, and it may be difficult to link specific risk factors such as chemical exposures to the AO of concern. With respect to breast cancer, it is possible that many chemicals or mixtures contribute to risk through many mechanisms; therefore, it may be more pertinent to work backwards from the AO to better understand the etiology and more effectively identify KEs and MIEs that may lead to increased breast cancer risk (124). However, without a comprehensive data science resource that can integrate gene identifiers or related information on early KEs in toxicity or disease with AOs, the considerable amount of information already available from academic, public, and private sector research may not be fully leveraged for hypothesis generation regarding mechanisms of toxicity or disease. The goal of the work presented herein is to provide just this kind of resource that can provide a putative, ranked linkage between an AO of concern and a given "entity," i.e., a gene, biological process, or chemical; the result of using this new tool is a ranked list of potentially relevant entities that can be evaluated in follow-up screening, representing a quantitative approach to literature review and hypothesis generation.

Currently there are several high-profile, publicly-available efforts in toxicology developed to explore how chemicals perturb biological systems, including the US Environmental Protection Agency's Toxicity Forecaster (ToxCast) (125) and the larger, interagency collaboration Tox21 (9). The high-throughput bioactivity results from these research programs have been used for chemical screening efforts of regulatory importance, like the Endocrine Disruptor Screening Program (EDSP) (126). These data have also been useful in research and development of putative AOPs, wherein the chemical-target interactions from high-throughput screening can be used as MIEs (34). However, even with this considerable amount of information, linking high-throughput screening data to AOs like diseases can be challenging without integration with information at various levels of biological complexity that consider toxicity. Another effort in linking chemical exposures to disease is the Comparative Toxicogenomics Database (CTD). In CTD, chemical-gene interactions are manually curated from published articles and then connected to diseases via inference (35,97). The chemical-disease inference is based on overlapping genes for a given chemical-gene pair where the disease-related genes are either manually curated from articles or pulled in from another publicly available resource, Online Mendelian Inheritance in Man (OMIM) (127). While these disease sources are helpful for exploring gene-disease associations for inherited variants, they do not consider genes involved with the initiation and progression of a disease from non-inherited (i.e. environmental) risk factors. Finally, massive data generation efforts for toxicogenomics are ongoing with applications like the Connectivity Map (CMAP) (33) and the S1500+ from Tox21 (128,129); in these efforts, analysis of gene expression changes resulting from chemical exposures are being used for drug discovery and repositioning as well as understanding chemical mechanisms of toxicity. With these large data generation activities underway, it is more important than ever that toxicologists have access to a tool that can enable ranked associations between gene identifiers or early KEs and possible AOs; such a tool would require integration of multiple types and sources of data or information.

The most substantial source of biological and biomedical information is PubMed, a freely available database managed by the National Library of Medicine (NLM) that contains over 27 million scientific articles that are indexed by medical subject headings (MeSH terms) (130,131). MeSH terms are arranged in a hierarchical tree with parent-child relationships such that each parent encompasses the concepts of each of its descendants, i.e. child MeSH terms are narrower in scope than their broader-scoped parents. For example, as seen in Figure 2.1, “Ductal, Carcinoma, Breast” has one immediate parent from two branches: “Breast Neoplasms”. “Carcinoma, Ductal, Breast” is a narrower concept than its parent and other ancestors in the tree. MeSH terms cover topics across all the articles within PubMed, but the most relevant topics to the work presented here are those on diseases, symptoms, processes, and related biological and chemical entities, which are well-represented in MeSH terms. These MeSH terms need to be linked with gene identifiers, which requires some additional consideration as gene identifiers are not automatically tagged to articles in PubMed.

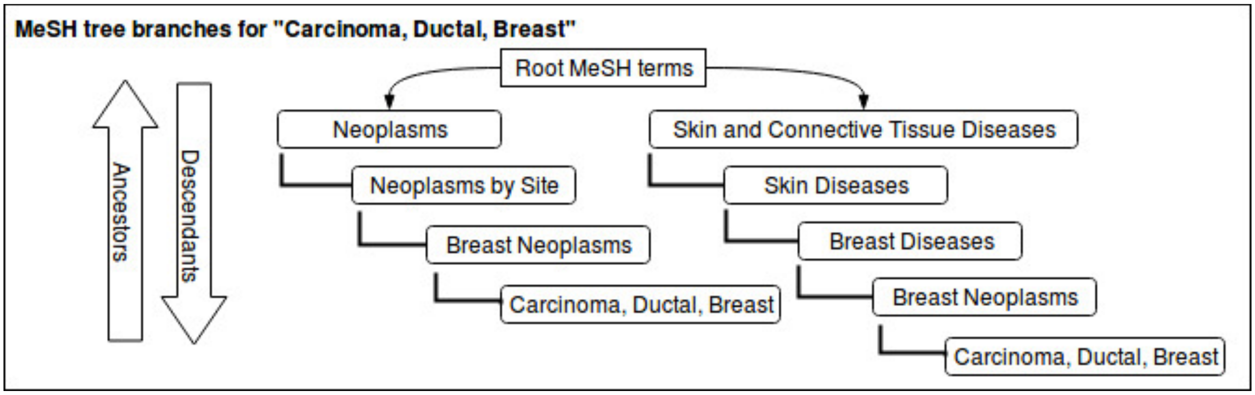

MeSH tree branches for "Carcinoma, Ductal, Breast"
Root MeSH terms
Ancestors
Descendants
Neoplasms
Neoplasms by Site
Breast Neoplasms
Carcinoma, Ductal, Breast
Skin and Connective Tissue Diseases
Skin Diseases
Breast Diseases
Breast Neoplasms
Carcinoma, Ductal, Breast

Figure 2.1: MeSH tree branches for "Carcinoma, Ductal, Breast"
Shown are two branches for the MeSH term "Carcinoma, Ductal, Breast". These branches have two root MeSH terms: "Neoplasms" and "Skin and Connective Tissue Diseases". Preceding MeSH terms (i.e. traversal towards a root MeSH term) are ancestors, where descendants are the MeSH following (i.e. traversal away from a root MeSH term). The depth of a MeSH term corresponds to the number of ancestors it has with depth increasing with traversal away from the root MeSH term.

Although some genes can be specifically identified by MeSH terms, not all genes are represented, especially since gene identifiers are species-specific. Systematic approaches for tagging genes that use strategies like named entity recognition (NER) have been implemented(98). However, despite the most successful efforts(99), no global approach exists where all genes can be systematically identified across all articles in PubMed. In lieu of a global systematic approach, we can rely on numerous manual curation efforts with publicly available resources that tag articles with relevant unique gene identifiers (GeneID). Although manual curation efforts are low throughput, the quality of mappings is higher, especially in resources built based on their manual curation efforts, including CTD and Universal Protein Resource (UniProt/SwissProt)(132). These manually curated resources offer valuable information alone, but also have great potential for discovery if tied together into one larger resource that also includes CTD's chemical-gene interactions and UniProt's protein-specific topics.

Here we describe a novel and transferrable methodology, producing a resource known as the Entity MeSH Co-occurrence Network (EMCON), that integrates several resources to develop a network connecting genes to MeSH terms. EMCON uses ranked associations between genes and MeSH terms to produce ranked gene sets for hypothesis generation and testing. The utility of EMCON was demonstrated by evaluating genes putatively linked to breast carcinogenesis, an example highly relevant to public health. Given the breadth of ongoing research on chemicals known to increase the frequency of breast tumors in animals and humans(113–118), important information about breast cancer mechanisms and risk could be uncovered by integrating already existing data housed in resources mentioned above. In this example, MeSH terms were selected that represent important processes in carcinogenesis and, more specifically, breast carcinogenesis. These MeSH terms were used to query EMCON and retrieve a ranked list of genes. Previously, Silent Spring Institute (SSI) created a list of nearly 300 genes as a reference gene set for breast carcinogenesis through expert literature review (ELR)(90). For the purposes of this study, that reference gene list was used to measure relevance of the EMCON search results. This work demonstrates a novel application of NPMI that critically informs hypothesis generation regarding genes that may be involved in breast carcinogenesis. EMCON may be useful in prioritization and selection of gene sets for transcriptomic experiments and/or articles to be manually reviewed for reference information. The methods described here are transferrable to any disease or AO of interest and could be tailored to myriad biomedical or life science research questions.

Methods

The overall workflow detailed in this paper is represented in Figure 2.2. First, we integrated seven resources, including gene2pubmed(102,133), Gene Reference into Function (GeneRIF)(102,134), CTD(35,135), UniProt/SwissProt(132,136), Reactome(137,138), Rat Genome Database (RGD)(139,140), and Mouse Genome Informatics (MGI)(141,142), to develop a network of naive GeneID-MeSH associations. Parent MeSH terms are less specific than their child terms, and as such, associations with parent MeSH terms may give less specific insight into the AO (see Figure 2.1); these parent MeSH terms should be mapped to all articles mapped to their child terms to reflect this relationship. Accordingly, we normalized the MeSH term frequencies to reflect the hierarchy within the MeSH tree so that broader terms appeared in associations more frequently while narrower, more specific terms showed up less often. This resulted in the less specific parent MeSH terms being mapped more promiscuously. Lastly, the GeneID-MeSH associations were ranked using an association measure called normalized pointwise mutual information (NPMI) (143). NPMI is commonly used in text mining for collocation extraction to identify words that co-occur together more than expected by random chance; the NPMI for any given association is a continuous value between -1 and 1. An NPMI greater than zero indicates a co-occurrence with greater probability than chance, with increasing significance of the probability as the NPMI value approaches 1. A positive NPMI does not indicate the direction of association (positive or negative association) between the GeneID and MeSH term. The final resource of ranked GeneID-MeSH associations is represented by EMCON: a scalable, queryable resource for retrieval of a ranked list of genes for a specific topic covered within MeSH terms.

Integration of biomedical text resources

To build a network relating genes to MeSH terms, we first identified biomedical databases that manually link genes or gene products to relevant articles. These databases included Comparative Toxicogenomics Database (CTD) (35,135), gene2pubmed (102,133), Gene Reference into Function (generif)(102,134), Universal Protein Resource/Swiss-Prot (UniProt) (132,136), Reactome (137,138), Rat Genome Database (RGD) (139,140), Mouse Genome Informatics (MGI) (141,142). Each of these resources provide cross-references to Entrez Gene (GeneID) along with a PubMed Identifier (PMID) that uniquely identifies a gene and PubMed article respectively. Entrez Gene is a resource managed by the National Center for Biotechnology Information (NCBI) providing unique identifiers for genes and linking information to genes (biological function, gene products, sequences, etc.) as this type of information is discovered (102). PubMed is also managed by NCBI and is the largest resource of freely accessible biomedical text with over 27 million citations from a variety of sources including peer-reviewed, biomedical journals. Each of the resources listed above can be integrated into a single resource that links PMIDs with GeneIDs (Figure 2.2A).

Next GeneIDs were linked to concepts across PubMed via Medical Subject Headings (MeSH terms;Figure 2.2A). Articles within PubMed are both manually and automatically tagged with MeSH terms, which is a controlled vocabulary of over 27,000 keywords structured in hierarchical trees used to categorize the concepts covered in an article(131). For example a publication titled "Estrogen receptor variant messenger RNA lacking exon 4 in estrogen-responsive human breast cancer cell lines" (144) has been tagged with MeSH term "Breast Neoplasms", "Receptors, Estrogen", "RNA, Messenger", and others. This is exemplified in Figure 2.3, which shows how GeneIDs are mapped to MeSH terms. This article has also been manually tagged with the gene estrogen receptor alpha (ESR1). Combined with another article(145) that is also manually tagged with ESR1 and a few overlapping MeSH terms, the gene ESR1 now has two articles that support a relationship to the MeSH term “Molecular Sequence Data”, “Amino Acid Sequence”, “Receptors, Estrogen”, “RNA, Messenger”, and “Tumor Cells, Cultured” (Figure 2.3). As more articles are added to the network, the number of supporting articles for a GeneID-MeSH association grows. Integration of these resources yielded a network of naive GeneID-MeSH associations (Figure 2.2A).

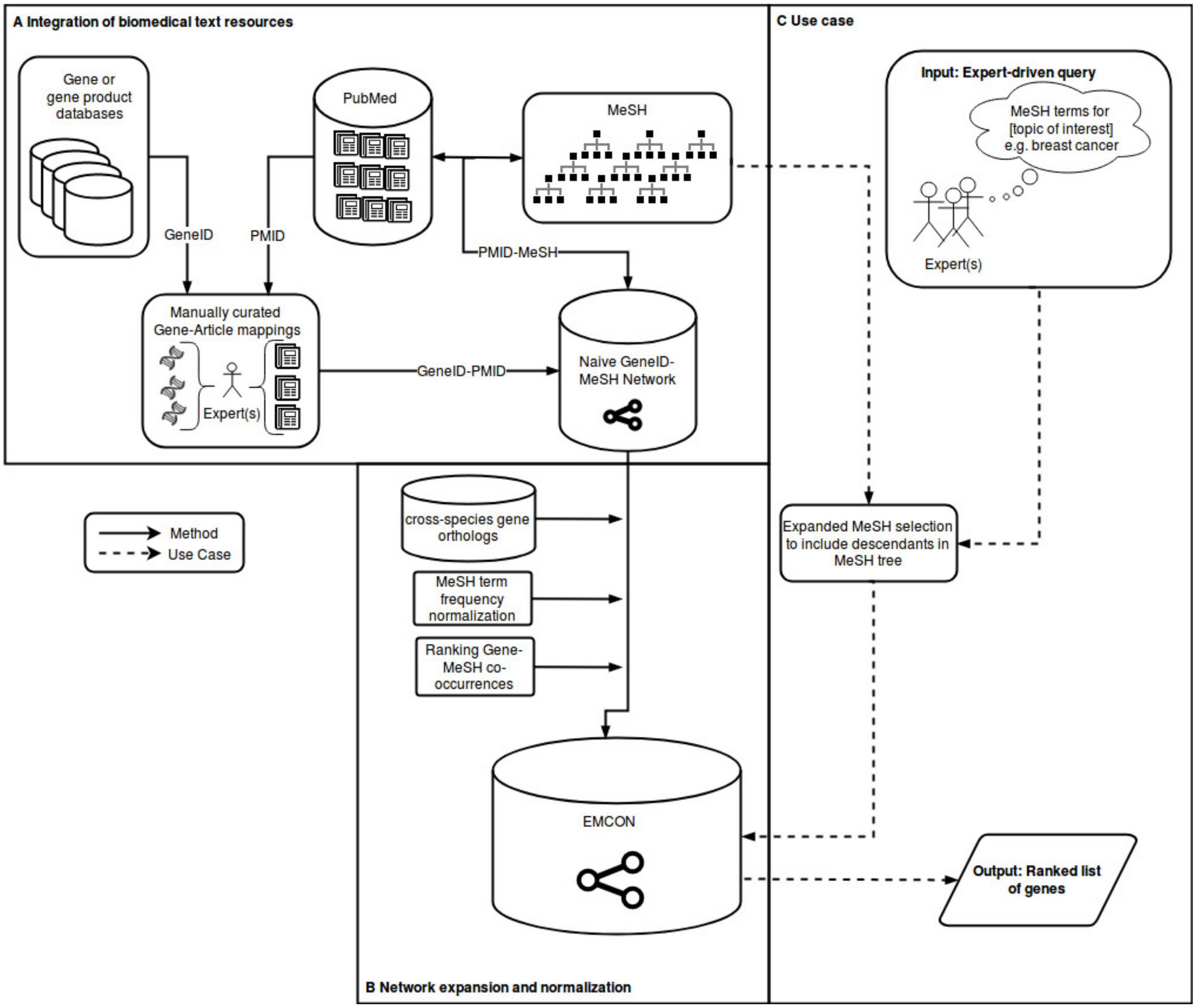

A Integration of biomedical text resources
Gene or gene product databases
PubMed
MeSH
GeneID
PMID
PMID-MeSH
Manually curated Gene-Article mappings
Expert(s)
GeneID-PMID
Naive GeneID-MeSH Network
C Use case
Input: Expert-driven query
MeSH terms for [topic of interest] e.g. breast cancer
Expert(s)
Method
Use Case
cross-species gene orthologs
MeSH term frequency normalization
Ranking Gene-MeSH co-occurrences
Expanded MeSH selection to include descendants in MeSH tree
EMCON
Output: Ranked list of genes
B Network expansion and normalization

Figure 2.2: Workflow for building Entity MeSH Co-occurrence Network (EMCON)
EMCON is created by (A) integration of biomedical resources, specifically manually annotated datasets of GeneID-PMID mappings. These data are combined with PMID-MeSH mappings to create a naive GeneID-MeSH network. (B) Next, the naive GeneID-MeSH network is expanded by mapping orthologous genes followed by MeSH term frequency normalization. The GeneID-MeSH associations are then ranked to generate the final EMCON resource. (C) EMCON can be queried with specific use cases where experts identify MeSH terms important to a topic of interest. Those selected MeSH terms are expanded to include descendants, which is the full set of MeSH terms used to query EMCON. The final output is a ranked list of genes ranked according to overrepresentation of the topic of interest.

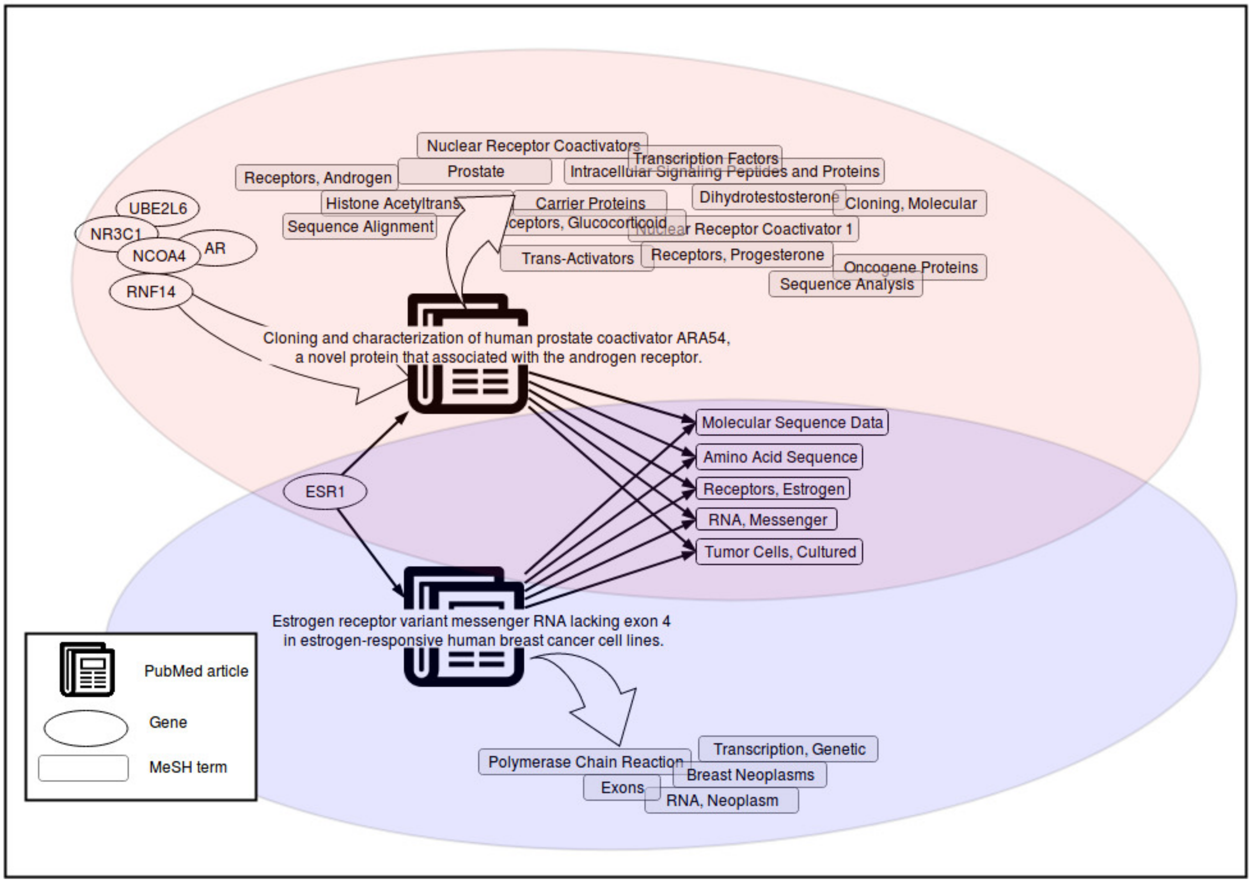


Figure 2.3: Example of the naive GeneID-MeSH network
The naive GeneID-MeSH network consists of GeneIDs that have been manually tagged to articles within PubMed, which are connected to MeSH terms.

Cross-species gene orthologs

Several of the resources above do not exclusively contain information on human genes, but for this work, we are only concerned with human genes. To maximize the number of articles and avoid excluding those that do not have human genes from the network altogether, we identified human orthologous genes. We assumed that the topics from an article about non-human orthologs are relevant to humans. We utilized UniProt Reference Clusters (UniRef), specifically UniRef50, to identify non-human proteins that have a human reference protein with a similar sequence(104). Proteins were mapped back to the naive GeneID-MeSH network via GeneID cross-references from UniProt/SwissProt. Then all articles tagged with the non-human GeneIDs were mapped back to the reference human GeneID from the corresponding similarity cluster (Figure 2.2B).

Medical Subject Heading (MeSH term) Frequency Normalization

At the top of the MeSH hierarchical structure are sixteen root MeSH terms, such as “Anatomy”, “Diseases”, “Chemicals and Drugs”, and “Phenomena and Processes”. These broader terms maintain parent-child relationships in that each parent MeSH term branches into more specific "child" MeSH terms that fall under the umbrella of the broader “parent.” MeSH terms are not limited to any one branch, which means that MeSH terms can have multiple parents. For example, “Breast Neoplasms” has the parent “Neoplasms by Site” as well as “Breast Diseases” (Figure 2.4A). To reflect this hierarchical structure so that broader MeSH terms are mapped more promiscuously to articles than narrower MeSH terms, we normalized the frequency of MeSH terms (Figure 2.2B) by mapping the ancestors of a MeSH term back to articles of the descendants (Figure 2.4B). This ensures that broader, parent MeSH terms are mapped at higher or comparable frequencies than their narrower, more specific descendants. This same type of normalization can be seen in gene sets for hierarchical pathway datasets like that of Reactome(137,138). This normalization prevents skewing of results towards broader MeSH terms, which may have been inconsistently mapped to articles, and enables identification of more specific associations with child terms within the MeSH tree, which cuts down on the overall noise to identify the most relevant associations. This normalization is defined in equations $M(m')$ and $p(m')$ in Table 2.1.

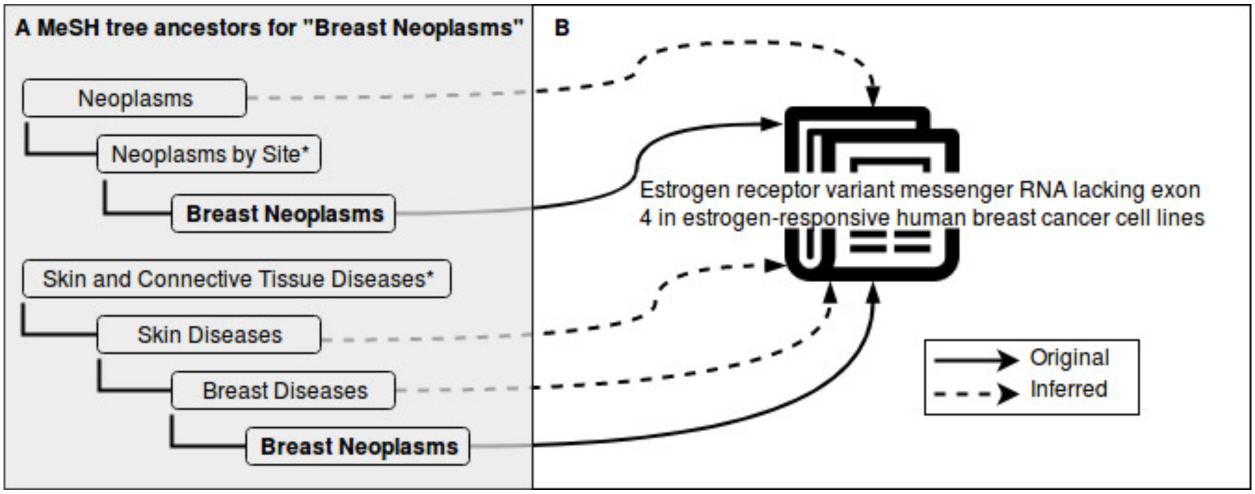

A MeSH tree ancestors for "Breast Neoplasms"
Neoplasms
Neoplasms by Site*
Breast Neoplasms
Skin and Connective Tissue Diseases*
Skin Diseases
Breast Diseases
Breast Neoplasms
B
Estrogen receptor variant messenger RNA lacking exon 4 in estrogen-responsive human breast cancer cell lines
Original
Inferred

Figure 2.4: Example of MeSH term frequency normalization
(A) Shown are the same branches from Figure 2.1. The MeSH term “Breast Neoplasms” has a total of five ancestors with two root MeSH Terms: “Neoplasms” and “Skin and Connective Tissue Diseases.” (B) All the ancestors for a given a MeSH term are subsequently tagged to each article of a specific MeSH term. The original mapped associations are indicated by solid arrows and inferred associations as part of our MeSH term normalization are indicated by dashed arrows.
*MeSH terms only used for structuring the MeSH tree and not used for tagging articles.

Equation	Description
$C = \{c_1, \dots c_n\}$	a set of co-occurrences, where c is a GeneID-MeSH term co-occurrence that is unique by PMID and n is the total number of co-occurrences
$G(g)$	the number of co-occurrences of C that contain gene, g
$M(m)$	the number of co-occurrences of C that contain MeSH term, m
$M(m')$	the number of co-occurrences of C that contain m and all the descendants of m
T$(g; m)$	the subset of C that contains co-occurrences with both g and m
$p(g) = \frac{\lvert G(g) \rvert}{n}$	the probability of g occurring
$p(m) = \frac{\lvert M(m) \rvert}{n}$	the probability of m occurring based on frequencies before MeSH term frequency normalization
$p(m') = \frac{\lvert M(m') \rvert}{n}$	the probability of m and all the descendants of m occurring based on frequencies after MeSH term frequency normalization
$p(g; m) = \frac{\lvert T(g; m) \rvert}{n}$	the probability of g and m co-occurring
$pmi(g; m) = log(\frac{p(g; m)}{p(g)p(m')})$	pointwise mutual information for a given g and m
$npmi(g; m) = \frac{pmi(g; m)}{-log(p(g, m))}$	normalized pointwise mutual information for a given g and m

Table 2.1: Equations for ranking GeneID-MeSH co-occurrences

Ranking Gene-MeSH co-occurrences

The naive GeneID-MeSH network only contains associations between a gene and a MeSH term based on the frequency with which those two entities occur together in an article. To extract meaningful co-occurrences of a GeneID and a MeSH term, we calculated a rank measure called normalized pointwise mutual information (NPMI), which is the normalized variant of pointwise mutual information (PMI) (Table 2.1). PMI is a rank measure commonly used in text mining for collocation extraction, i.e., identifying words that co-occur together more than random indicating a shared meaning like “hot tea” and “crystal clear”. Because PMI is a rank measure, there is no level of significance or accepted cutoff to use for co-occurring terms; however, the normalized variant, NPMI, calculates a continuous value between -1 and 1 where -1 is interpreted as no co-occurrence, 1 is interpreted as perfect co-occurrence, and 0 is interpreted as co-occurrence at random(143). These interpretations can be made about GeneID-MeSH co-occurrences because GeneIDs were tagged to articles independent of MeSH terms. Also, NPMI is biased in that low frequency co-occurrences are ranked higher(143). To reduce the potential for spurious or less-replicable co-occurrences to drive this bias, GeneID-MeSH associations with less than three PubMed articles were excluded from the network. This cutoff was chosen based on assumptions that can be made about positive reporting of results due to publication bias(146). We assumed that for a GeneID-MeSH association with three or more PubMed articles that support the association, it was likely that at least two of the articles reported positive results for a relationship between the GeneID and MeSH term.

Table 2.1 summarizes the equations needed to calculate NPMI. The probability of a MeSH term m and all the descendants of m occurring ($p(m')$) will increase the denominator of PMI resulting in an overall lower NPMI for broader terms since frequency is increased for a given MeSH term based on its descendants. This adjustment decreases the overall ranks of MeSH terms with many descendants whereas more specific MeSH terms ranked higher. The final network was filtered to include only GeneID-MeSH associations with NPMI > 0, which indicates that each association present exceeds the associations expected from random chance (Figure 2.2B).

MeSH Terms for breast carcinogenesis

MeSH terms that comprehensively capture the use case of breast carcinogenesis were needed to query EMCON and retrieve a ranked list of relevant genes. As described in Grashow et al.(90), experts selected seventeen MeSH terms that encompassed concepts from seminal papers on carcinogenesis(88,147,148) and breast carcinogenesis(118), including: Neovascularization, pathologic; Neovascularization, physiologic; Apoptosis; Cell cycle; Epigenomics; DNA damage; DNA repair; Growth hormone; Cell survival; Immune system; Inflammation; Breast; Breast Diseases; Oxidative stress; Cell proliferation; Gonadal steroid hormones; and Xenobiotics. These seventeen MeSH terms alone do not necessarily reflect the full scope of the concept they represent, therefore the full query also includes all descendants of these MeSH terms for a total of 214 MeSH terms to represent breast carcinogenesis. Clearly, some of these concepts may be related to cancer phenotypes more broadly, and some may be more specific for breast carcinogenesis.

Relevance of retrieved gene list

For the topic of breast carcinogenesis, a reference gene set of 287 genes was compiled through expert literature review (ELR) by Silent Spring Institute as described in Grashow et al.(90) , including: (1) gene targets for quantitative nuclease protection assays in ToxCast Phase I; (2) genes responsive to nuclear receptors of interest (estrogen, progesterone, androgen, and aryl hydrocarbon receptors); (3) genes included in Qiagen microarray panels designed to probe pathways relevant to breast cancer (estrogen receptor signaling, breast cancer, DNA repair, DNA damage, growth factors, cellular stress response); (4) important genes in breast cancer based on key literature reports(113,149–151); (5) genes listed as related to breast cancer in curated databases (OMIM, CTD); and, (6) genes listed by partners at NCATS Chemical Genomics Center (NCGC) as important in cytotoxicity response (Figure 2.2C). Potential housekeeping genes were chosen from previous reports in MCF-7 cells(152–154). This ELR gene set was used as a reference gene set to measure the relevance of the retrieved gene list from EMCON to the topic of breast carcinogenesis.

The EMCON search was conducted 214 times to generate one gene list for each MeSH term in the search query. The final gene list was obtained by averaging the NPMI rank per gene in the set of 214 iterations. Relevance to breast carcinogenesis of the final gene list from EMCON was measured by comparing the mean rank of the ELR gene set to the distribution of mean ranks of 1000 randomly generated gene sets of the same length as the ELR gene set of 287 genes. The retrieved gene list was considered relevant to breast carcinogenesis if the mean rank for the ELR gene set is higher than the distribution of mean ranks for randomly generated gene sets yielding an empirical p-value < 0.01. We used an empirical p-value because the comparison dataset is simulated, i.e. it was not derived using reference gene sets from other disorders. We felt that the best comparison would be against random gene sets rather than make inferences about how similar or dissimilar disorders may be based on respective genes. Recall was calculated as the fraction of ELR genes retrieved in the final list produced by EMCON, where the ELR gene set was considered a standard to evaluate the gene list produced by EMCON. Precision scores were calculated based on expert assessment of relevance of the top five genes for each of the seventeen selected MeSH terms. This expert assessment involved manual review of the literature that resulted in the GeneID-MeSH association to classify the association as true positive or false positive.

Comparison to a similar tool

EMCON's performance was compared to Génie, another literature-based gene prioritization approach(155). Génie uses a naive linear Bayesian classifier in conjunction with a Fisher's exact test to produce a list of genes ranked by false discovery rate (FDR). We compared our method with that of Génie by using a Spearman rank correlation of the ELR gene set from the EMCON search results with search results from Génie. Two gene sets were obtained from Génie: one using only the MeSH term "Breast Neoplasms" and another using all 214 MeSH terms used to query EMCON.

Computational and statistical analyses

All data were downloaded as flat files from their respective sources (Table 2.2). Python 3.6+(156) was used to parse the files and import into MongoDB 3.4+(157). All methods were implemented using MongoDB's aggregate pipeline or python packages pandas 0.20+(158), numpy(159), and numba(160). All code is available via iPython notebooks(161) at https://github.com/USEPA/CompTox-HTTr-EMCON.

Results

Entity MeSH co-occurrence network (EMCON)

Seven resources that manually tag PubMed articles with GeneIDs were identified (Table 2.2) and integrated into a single resource containing GeneID-PMID associations. Subsequently, MeSH terms were incorporated to generate a naive GeneID-MeSH network. Most of the genes in the naive GeneID-MeSH network are not human, but many produce proteins with high similarity to human protein orthologs, such that the information from non-human genes may be relevant to human-related research questions. To boost the number of articles mapped to human genes, UniRef50 clusters were used to identify human orthologous genes to increase the human relevant articles from around 500,000 to nearly 900,000. Next, the MeSH term frequency was normalized by mapping MeSH term ancestors back to articles to which their descendants were already mapped. Finally, GeneID-MeSH were ranked using NPMI to create a final network called Entity MeSH Co-occurrence Network (EMCON). EMCON is comprised of nearly 14 million associations, and, when filtered to require an article count > 2, the associations were dramatically reduced with 3.56 million remaining associations. The GeneID-MeSH associations in EMCON have article counts ranging from three to 10,276. The NPMI scores range from -0.5 to 0.7 with a mean of 0.025 (Figure 2.5) and 2.13 million GeneID-MeSH associations with NPMI > 0.

Gene and gene product databases	Number of articles	Number of GeneIDs in articles	Number of species across GeneID s
gene2pubmed (133)	1,062,713	5,565,651	12,782
Gene Reference into Function (GeneRIF) (134)	705,441	90,329	1,913
Comparative Toxicogenomics Database (CTD) (135)	58,180	43,298	76
Universal Protein Resource (UniProt/Swiss-Prot) (136)	950,989	5,156,248	12,555
Reactome (137) 38	15,650	11,110	9
Rat Genome Database (RGD) (139) 40	834,585	87,874	7
Mouse Genome Informatics (MGI) (141)	181,519	42,020	1
Total Unique	1,238,879	7,074,406	14,126

Table 2.2: Manually curated resources used to construct EMCON
A total of seven resources that manually tag GeneID's to articles within PubMed were integrated to serve as the initial dataset for building EMCON. Over 1.2 million articles make up the naive GeneID-MeSH network with over 7 million genes for over 14K species.

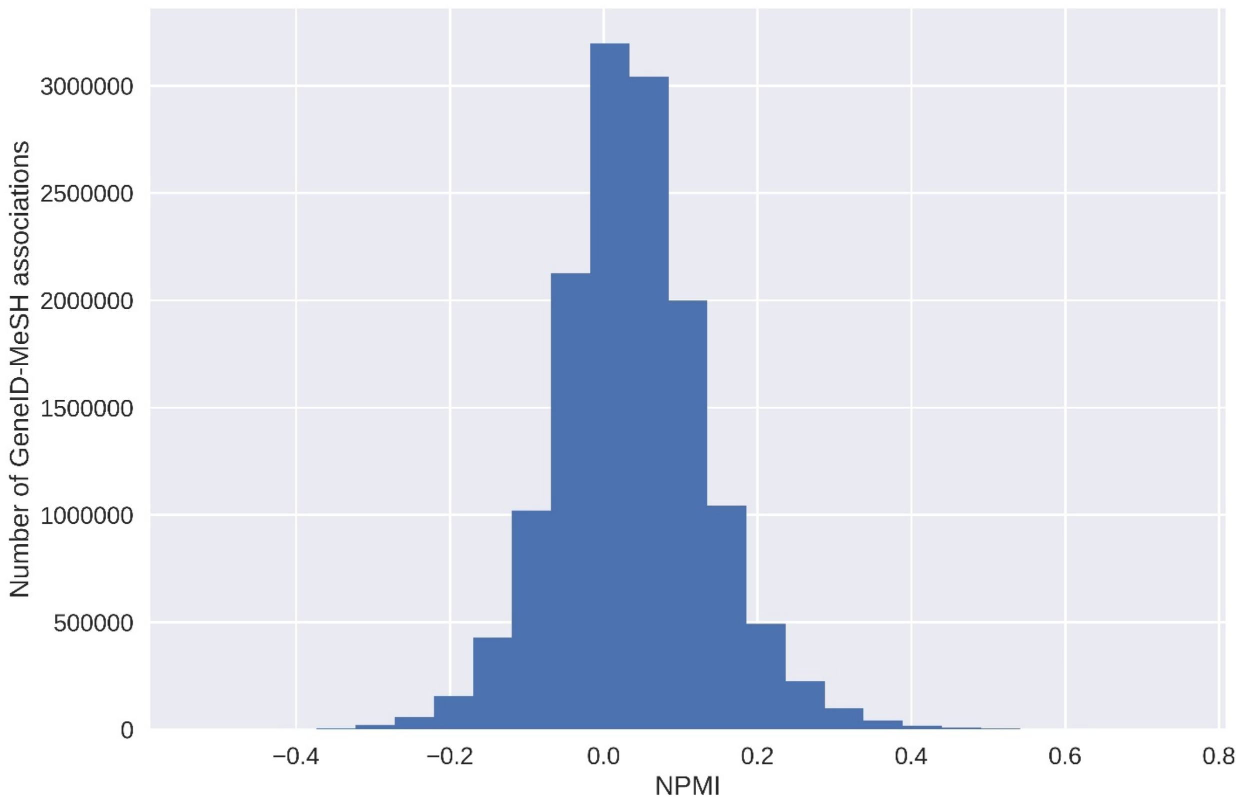


Figure 2.5: EMCON NPMI distribution.

MeSH term frequency normalization

The MeSH term frequency normalization (represented as *p(m');* See Methods) increased the promiscuity of MeSH tree terms based on descendants within the hierarchical trees, via mapping MeSH terms back to the articles of their descendants. The probability of a given MeSH term occurring in the naive GeneID-MeSH network (*p(m))* increased with the number of descendants present in the network. This increase in promiscuity for broader MeSH terms corresponds to decreased NPMI for associated genes. Figure 2.6 demonstrates the probability of a MeSH term occurring before (*p(m)*) and after (*p(m')*) frequency normalization; the probability of the MeSH term co-occurring with the gene of interest (*p(g,m)*); and the associated NPMI scores for the GeneID-MeSH co-occurrences for two MeSH branches, “Cell Cycle” and “Skin Diseases”.

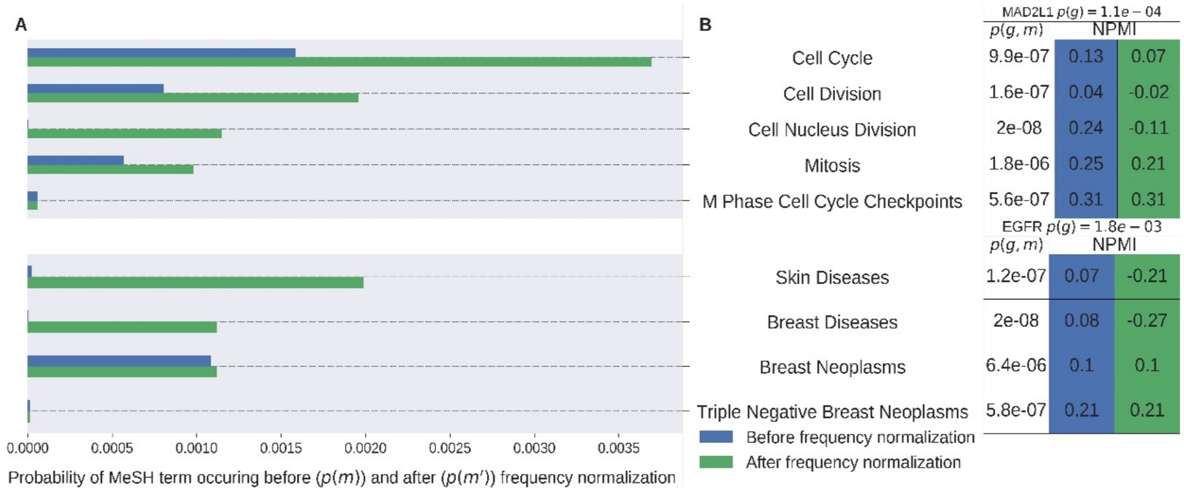


Figure 2.6: MeSH frequency normalization for two branches
Two branches from the MeSH hierarchical tree were used to demonstrate how the annotation bias correction alters the probability of a MeSH term occurring (*p(m)*) along with the resulting NPMI with a relevant gene. These values correspond with the depth of a given MeSH term in the hierarchical tree.

First, *p(m)* and *p(m')* were compared for MeSH terms in the “Cell Cycle” and “Skin Diseases” branches (Figure 2.5A). *p(m)* did not inversely decrease with the depth of the MeSH hierarchical tree for “Cell Cycle” or “Skin Diseases.” This relationship implied that “Breast Neoplasms” was broader than “Skin Diseases” because the *p(m)* for “Breast Neoplasms” (*p(m)*=0.001) was greater than the *p(m)* for “Skin Diseases” (*p(m)*=2.3e-5). However, after MeSH term frequency normalization, *p(m')* decreased as the depth increased for a given branch. For example, the *p(m')* for “M Phase Cycle Checkpoints,” a term representing increased depth within the “Cell Cycle” branch, was less than the *p(m')* values associated with its ancestors. Figure 2.6A also shows that despite *p(m')* decreasing as depth increased within the “Cell Cycle” branch, the MeSH term “Cell Nucleus Division” was nearly absent from the network altogether with a *p(m')*=2e-6. Similarly, following frequency normalization, the probability of the MeSH term “Skin Diseases” occurring in the gene-curated literature was greater than the probability of observing “Triple Negative Breast Neoplasms.”

Increases in *p(m')* correlated with decreases in NPMI, as illustrated in Figure 2.6B. In other words, for more promiscuous MeSH terms, the GeneID-MeSH term co-occurrence for that term was less likely to be specifically relevant for the specific topic of interest. When looking at the association between the MeSH branch, “Skin Diseases,” with epidermal growth factor receptor (EGFR), we see that the broader MeSH terms “Skin Diseases” and “Breast Diseases” had an NPMI < 0 (Figure 2.5B), indicating that these MeSH terms were less relevant specifically to breast carcinogenesis. The NPMI scores for the MeSH terms “Breast Neoplasms” and “Triple Negative Breast Neoplasms” co-occurring with EGFR remained at 0.1 and 0.21, respectively, because the *p(m')* remained relatively similar to *p(m)*. The NPMI decreased for most MeSH terms with MAD2L1 where “Cell Division” and “Cell Nucleus Division” drop below zero, which are excluded from EMCON. However, the NPMI scores for “Mitosis” and “M Phase Cell Cycle Checkpoints” remained above 0, therefore these associations were preserved. Despite the decreased NMPI for “Cell Cycle” and MAD2L1, these associations remained above 0 and were also preserved.

By normalizing the MeSH term frequency, we reduced the noise introduced into the network to retrieve more specific and useful GeneID-MeSH co-occurrences. This network cleaning approach assured that broader terms would not be ranked higher than more specific terms. The net impact is that less-specific MeSH terms will have lower NPMIs; many articles relate to “Skin Diseases” or “Breast Neoplasms,” but these articles may have little association with “Triple Negative Breast Neoplasms.” MeSH terms more closely associated with the pathological finding of interest such as“Triple Negative Breast Neoplasms,” will have a greater NPMI due to closer association. For all gene ID-MeSH co-occurences for a given branch, the NPMI will increase with depth; i.e., the lowest descendant MeSH term-gene co-occurrence will have the greatest NPMI. Thus, the most salient associations will be quantitatively identified.

Relevance of search results to breast cancer

Genes related to the topic of breast carcinogenesis were retrieved from EMCON using seventeen expert-selected MeSH terms that represent concepts from seminal papers on the topic of specifically breast carcinogenesis(118) and carcinogenesis in general(88,147,148). These seventeen MeSH terms were expanded to include all descendants in the MeSH trees to ensure the full scope is represented within the selection. The final list of MeSH terms totals 214, which were used to query EMCON and retrieve a final list of 14,811 genes.

Relevance of the EMCON-returned genes to breast cancer was evaluated by comparing the mean rank of the ELR gene set to the distribution of the mean ranks of randomly generated gene sets

(Figure 2.6). The random gene sets were generated by randomly selecting 287 genes, which is the length of the ELR gene set. The average rank of the ELR gene set was clearly distinguished from the random gene set distribution (empirical p-value << 0.01). Using the ELR gene set, recall from EMCON search results was 0.983. Precision was calculated by manually assessing the relevance of the top five genes with the corresponding MeSH term. The average precision across the seventeen selected MeSH terms was 0.87 (Table 2.3). We then looked at the top MeSH terms related to well-studied, breast cancer genes: *BRCA1*, *BRCA2*, *ESR1*, *ESR2,* and *PGR* (Table 2.4). The MeSH terms retrieved are all specific to breast cancer or molecules linked to breast cancer like “Progesterone” and “Estradiol”.

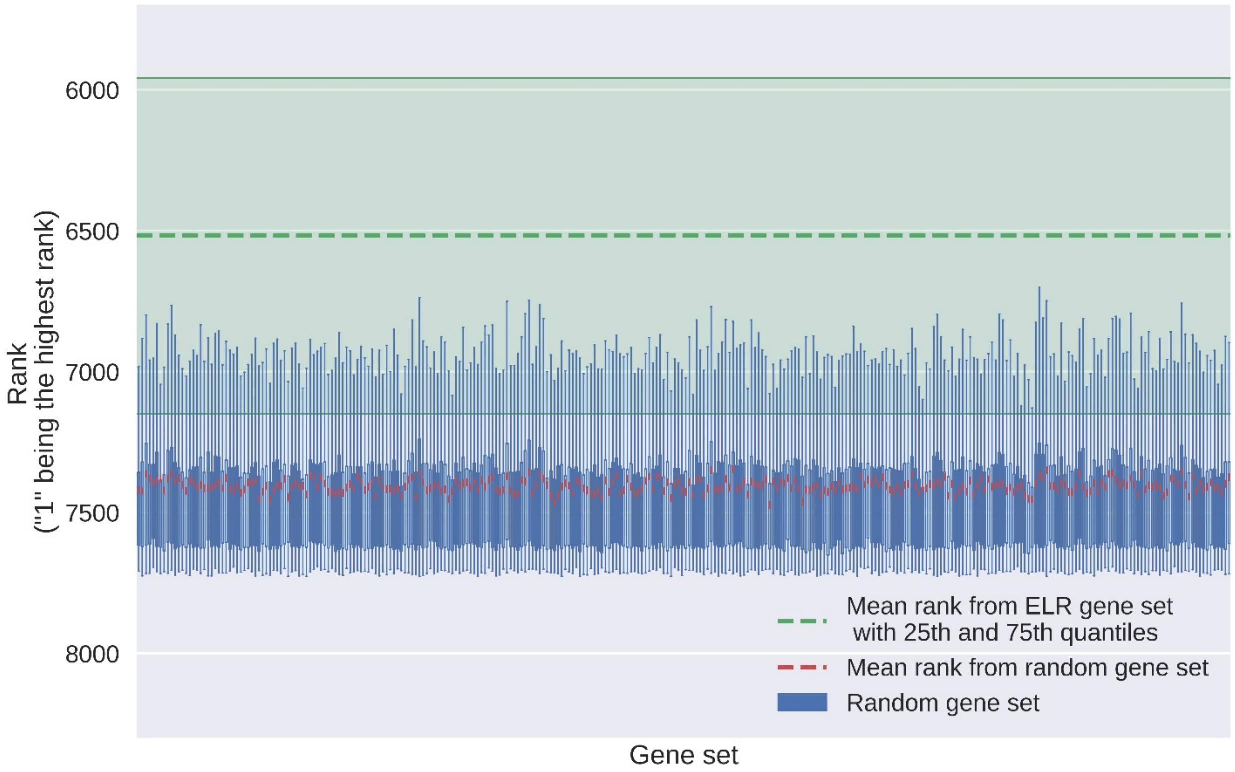


Figure 2.7: Comparison of mean rank of ELR, breast cancer-specific gene set to random gene sets within EMCON search results
The ELR gene set is, on average, ranked higher than any of the mean ranks for randomly generated gene sets of the same length. Shown are 300 representative random gene sets from a total of 1000. The mean rank across all the random gene sets is 7405.

MeSH name	**Top five genes (gene symbol)**	**Precision**
Neovascularization, Pathologic	*VEGFA, KDR, ANGPT2, ANGPT4, VASH1*	1
Neovascularization, Physiologic	*KDR, FLT1, TEK, ANGPT1, EPHB4*	1
Apoptosis	*CASP3, BAX, CASP9, BCL2, CASP8*	1
Cell Cycle	*CDK2, CDK1, CCNE1, CCNA2, CDKN1B*	1
Epigenomics	*PARP12, DNMT3A, TET3, GREB1, KAT8*	0.6
DNA Damage	*ATR, CHEK1, ATM, MDC1, DDB2*	1
DNA Repair	*RAD51, XRCC1, XPC, ERCC2, XPA*	1
Growth Hormone	*CSHL1, GH1, GH2, GHR, CSH1*	1
Cell Survival	*ARIH2OS, CASP3, BAD, BCL2, BCL2L1*	0.8
Immune System	*LAT2, CLEC4E, ARL4C, CLEC6A, NAV1*	0.6
Inflammation	*NLRP3, CRP, PYDC1, SPATA31E1, NLRP13*	0.8
Breast	*SCGB3A1, WISP3, PTK6, SCGB2A1, SCGB2A2*	1
Breast Diseases	*TBX3, IGFBP3, TP63, TP73, IGF1*	1
Oxidative Stress	*CAT, GSR, NFE2L2, SOD2, GPX1*	1
Cell Proliferation	*CCND1, FOXM1, CDKN1B, YAP1, CDKN1A*	1
Gonadal Steroid Hormones	*SEMG2, ACRV1, HSD17B1, SEMG1, HSD17B3*	1
Xenobiotics	*NR1I3, NR1I2, ACSM2A, ACSM2B, NAT1*	1
		0.870588

Table 2.3: Manual Precision for 17 selected MeSH terms
Relevance of the five top ranked genes for each of the seventeen selected MeSH terms relevant to breast carcinogenesis was evaluated by performing a literature search of through Entrez Gene. Gene symbols in red were not explicitly related to the corresponding MeSH term.

GeneID	Gene Symbol	MeSH Term
672	*BRCA1*	Breast Neoplasms
		Triple Negative Breast Neoplasms
		Breast Neoplasms, Male
		Hereditary Breast and Ovarian Cancer Syndrome
		Carcinoma, Ductal, Breast
675	*BRCA2*	Breast Neoplasms, Male
		Hereditary Breast and Ovarian Cancer Syndrome
		Breast Neoplasms
		Triple Negative Breast Neoplasms
2099	*ESR1*	Estradiol
		Fibrocystic Breast Disease
		Estrogens, Conjugated (USP)
		Breast Neoplasms
2100	*ESR2*	Estradiol
		Estrogens, Conjugated (USP)
5241	*PGR*	Progesterone

Table 2.4: Top MeSH terms for genes BRCA1, BRCA2, ESR1, ESR2, and PGR from EMCON
Five breast cancer-related genes were used to search EMCON. Shown are the top-ranking co-occurring MeSH terms.

Comparison to Génie

We searched Génie with two different queries to obtain breast cancer-related genes to compare to EMCON results. The Spearman rank correlation for the results from the query with all 214 MeSH terms is 0.561 (Figure 2.7) with a recall for the ELR gene set of 0.718. When using only the MeSH term “Breast Neoplasms” to retrieve breast cancer-related genes, the Spearman rank correlation drops to 0.451 (Figure 2.7) and the recall for the ELR gene set also drops to 0.641.

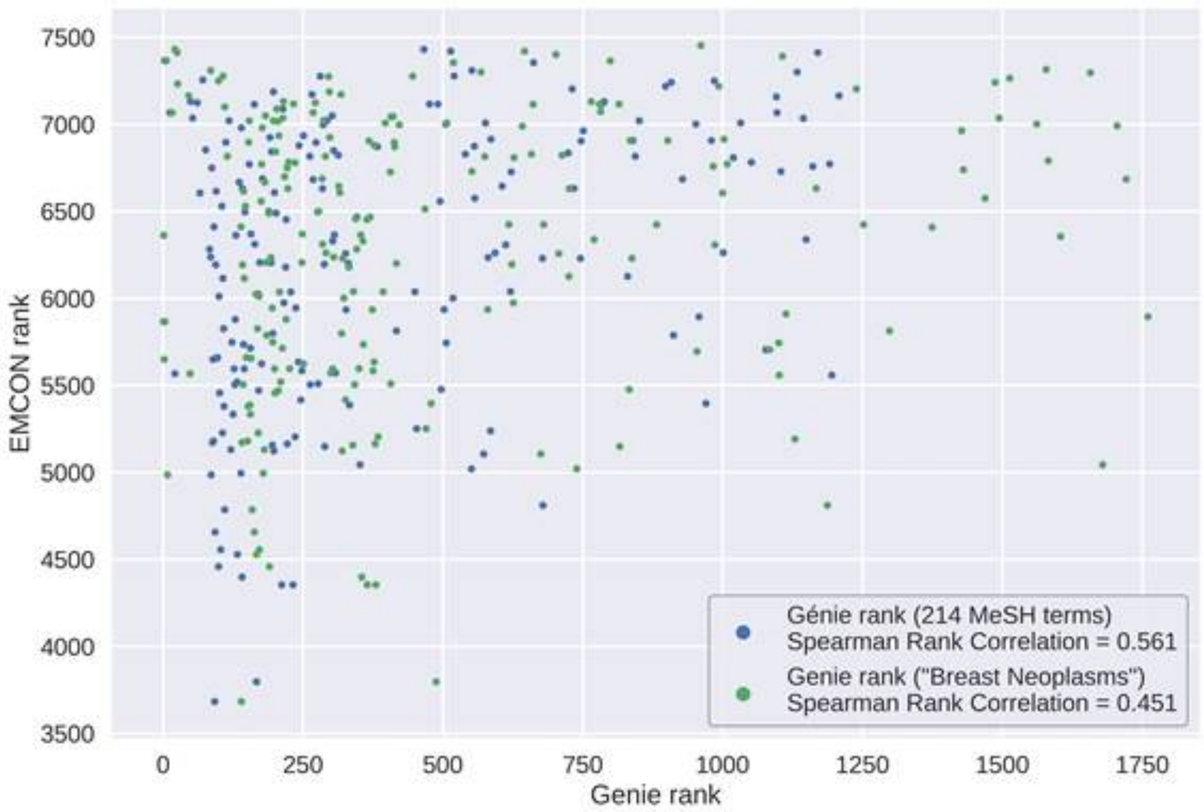


Figure 2.8: The rank comparison of the ELR gene set from EMCON and Génie
We obtained the two Génie gene sets by searching with 214 breast cancer-related MeSH terms and with only “Breast Neoplasms”. The correlation of the rank comparisons was similar across the two queries.

Discussion

We have developed an accessible and scalable resource called EMCON that is comprised of ranked associations between genes and MeSH terms. This novel tool is a needed public health and toxicology resource that enables connection of an AO of concern with hypothetical MIE or KE information, thus improving development of putative AOPs and providing an empirical approach to hypothesis generation. EMCON was developed via integration of multiple data sources and subsequent computation of the rank of specific associations. In the example herein, a ranked list of genes putatively related to breast carcinogenesis was defined using EMCON for use in hypothesis testing. The performance of EMCON in this example was evaluated in three ways: (1) comparison of the mean rank of the ELR gene set compared to randomly generated gene sets from the EMCON search results using the expert-selected MeSH terms related to breast carcinogenesis; (2) evaluation of the recall and precision of the EMCON search results using the ELR-derived gene set as a standard; and, (3) comparison to results for the 214 breast carcinogenesis-related MeSH terms from an existing tool, Génie. These three evaluations demonstrated that EMCON performed well for the use case of defining genes linked to MeSH terms. Within the EMCON search results, the ELR gene set for breast carcinogenesis ranked, on average, higher than any randomly generated gene set based on NPMI. Further, EMCON demonstrated excellent manually assessed precision (0.87) and recall (0.983) using the ELR gene set, and the EMCON results correlated with results from Génie, with some differences noted based on different methodological choices. Overall, the results presented herein suggest this is a valuable tool for hypothesis generation, providing critical support for the building of AOPs and AOP networks in addition to advancing research in biological and biomedical fields.

EMCON was constructed to better utilize existing information in systematic information extraction of information used in hypothesis generation. EMCON was built by first integrating heterogeneous resources that map genes to articles containing information across a multitude of topics from PubMed. Then protein similarity clusters from UniRef50 were used to identify articles with similar, non-human genes to be mapped to the corresponding human gene. MeSH term frequency was normalized by mapping MeSH term ancestors back to articles of their descendants so that MeSH frequencies correspond to the depth of the MeSH tree. Lastly, GeneID-MeSH associations were ranked using NPMI. For construction of EMCON, we utilized several resources that manually curate PubMed articles with genes relevant to the content of the article. Each curation effort prioritizes articles based on specific areas of interest: pathways (Reactome), proteins (UniProt), chemical-gene/gene product interactions (CTD), etc. The total number of articles from all the resources totals to almost 1 million out of 27 million articles within PubMed. Without the development of systematic information extraction or tagging efforts like NER [30], researchers are forced to rely on manual approaches. The mappings from manual efforts may be higher quality than those derived from potential systematic approaches, but the throughput is low. Also, each resource is biased towards a specific topic, so specific topics of interest may not be well represented in EMCON.

An ongoing limitation of any data mining approach using manually curated information from PubMed is that curation efforts are not standardized, as demonstrated by curation of certain GWAS and gene expression profiling studies. These curated studies have varying number of genes mapped without an explanation of whether it was all genes included in the panels, only variants identified, or only those with differential expression. This lack of standardization or clarification introduces noise into the network. However, the article count cutoff used in this approach (see Methods) filtered out much of this noise, reducing the total set of associations by 75%. Noise in the set of GeneID-MeSH associations was also reduced through MeSH term frequency normalization, which is similar to approaches used in overrepresentation analysis like gene set enrichment analysis (GSEA)(91). The Reactome gene set available for GSEA or similar pathway analysis methods is normalized in the same manner, i.e., genes from the child pathways are all annotated to the parent pathways as well(137). A further limitation of only using manually curated information is applicability to certain use cases. In this work, we explored breast cancer, which has a lot of literature in the curated space, but other diseases or outcomes may not be associated with any curated data. In moving forward with this work, systematic approaches can be developed to extract relevant information from articles to fill in gaps in knowledge. Although a particular disease or outcome may have limited information, EMCON could be used if more broad MeSH terms could be connected to these topics.

The universe of possible human-relevant GeneID-MeSH associations was expanded by using UniProt Reference 50 (UniRef50) clusters to map non-human genes to corresponding human orthologs. Human genes are overrepresented in the curated PubMed literature accounting for nearly 50% of the articles with thousands of other species accounted for the remaining articles. However, genes from other research conducted in model species may be relevant to human pathogenesis, at least at the level of hypothesis generation. UniRef50 was used because the protein clusters included the cross-species orthologs whereas UniRef90 and UniRef100 are typically clusters of same-species protein isoforms. Homologene is a resource that also clusters cross-species gene orthologs together(162) and is used by similar methods in literature-based gene prioritization(155). However, Homologene has not been updated since the last release in 2014. UniRef50 is regularly updated and supports many other efforts in proteomics work, and thus presented a clear choice for use in EMCON. It is possible that by expanding the network to include human orthologs, we introduced noise by including genes that may not be relevant for human pathogenesis. This aspect could potentially be explored in further analyses, especially since these articles on the human orthologs can be easily identified and then removed if deemed irrelevant. This type of noise would not necessarily detract from the utility of EMCON for hypothesis generation.

Normalized pointwise mutual information or NPMI was chosen as the association measure because of the defined threshold of NPMI>0.0 is interpreted as having dependent co-occurrence. Similar measures exist like Fisher's exact, log-likelihood ratio, and Pearson's chi-square (163), but despite their similar use in ranking, do not have defined thresholds to distinguish between independent and dependent co-occurrence. Similar co-occurrence measures to identify GeneID-MeSH associations are implemented in Gene2MeSH(164) and MeSHOPs(165). However, Gene2MeSH does not normalize the MeSH term frequency, and both lack cross species similarity mappings and use Fisher's exact test to identify and rank MeSH terms associated with a gene. There is currently no consensus on which association measure works best because each measure can outperform the other depending the dataset(163,166). For the purposes of this paper, NPMI was chosen because a continuous rank measure could be more easily incorporated into other methods like the gene prioritization workflow implemented in Grashow et al.(90). It is possible that the best use of an association measure with this dataset may be a combination of the previously listed measures. However, this work demonstrated the utility of NPMI for this problem where previous work has only focused on use of Fisher's exact(164,165) or more complex machine learning methods(155).

The utility of EMCON was demonstrated within the scope of breast carcinogenesis. Breast carcinogenesis was chosen as a use case because of the large amount of information available on the topic due to major public health interest. Seventeen MeSH terms were selected by experts that, along with their descendants, represented important characteristics of breast carcinogenesis. A total of 214 MeSH terms were used to query EMCON and retrieve ranked lists of genes where the NPMI was averaged across all MeSH terms for a final ranked list of genes. The MeSH tree was not considered when MeSH terms were selected, so the depth of each MeSH term varies. Due to the inclusion of descendants with MeSH terms at varying depths, MeSH terms like "Immune System" are overrepresented with 69 descendants, while MeSH terms like "Epigenomics" are underrepresented with 0 descendants. This over- and underrepresentation introduces bias in the results for breast cancer. It is possible the MeSH term selection process can be improved with a systematic, data-driven approach rather than a manual approach. The MeSH term selection could also be improved by consideration of the parent-child relationships in the MeSH tree and the MeSH-MeSH association from common publications(167).

Comprehensive gene sets for complex disorders are typically comprised of variants derived from genome-wide association studies (GWAS). However, for this work, we wished to include a more heterogeneous set of genes that may be related to processes seen in carcinogenesis. A breast cancer-specific gene set compiled through ELR was used to evaluate the relevance of the retrieved gene list to the topic of breast carcinogenesis. Using the ELR gene set as a standard, the final breast cancer gene list from EMCON had a recall of 0.983, and the ELR gene set ranked well above randomly generated gene sets of the same length (empirical $p<<0.01$) indicating that the higher-ranking genes from EMCON are likely relevant to breast carcinogenesis. This was further demonstrated by manually assessing precision of the top five genes for the seventeen selected MeSH terms (average precision=0.87). The MeSH terms that did not have a precision of 1 had either very few descendants (“Cell Survival” and “Epigenomics”) or had many descendants (“Immune System” and “Inflammation”; Figure 2.9). MeSH terms with few descendants could represent newer topics with fewer relevant articles or could represent cases wherein few genes have been specifically annotated to the topic, e.g., “Epigenomics”. MeSH terms with many descendants could have genes promiscuously mapped to them because of the large number of varied topics within the descendants. In both cases, the genes that were not explicitly related to the corresponding MeSH term were not well annotated, demonstrated low article count relative to the other top-ranked genes, or may have resulted from promiscuous mapping of geneID to MeSH. This type of false positive is an artifact of using NPMI since rare co-occurrences (GeneID-MeSH associations with low article counts) are artificially ranked higher. The article count cutoff could be raised to a more conservative number to remove these types of associations and tune EMCON to the specific research application based on the level of specificity required. When evaluating MeSH terms related to well-known breast cancer-related genes (*BRCA1*, *BRCA2*, *ESR1*, *ESR2*, and *PGR*), the topics were all specific to breast cancer in that they related to breast tissue-specific tumors or molecules like estradiol and progesterone.

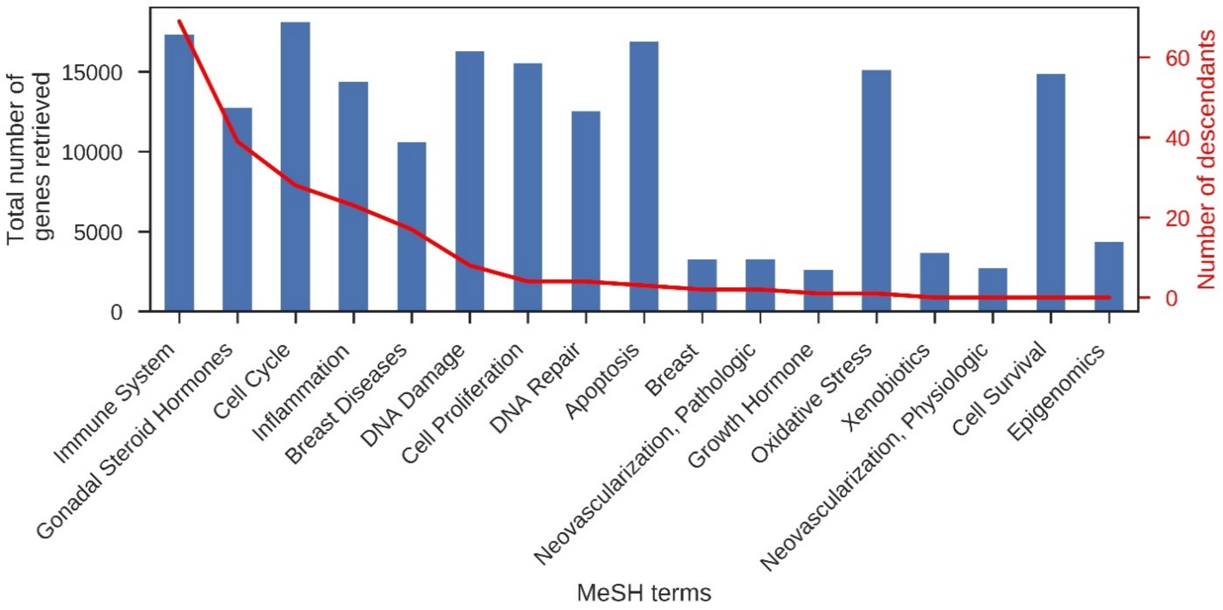


Figure 2.9: Number of genes retrieved for the 17 selected MeSH terms
The number of genes reflects the genes mapped to select MeSH term along with those genes also mapped to the descendants.

EMCON's performance was compared to results from Génie(155). The Spearman rank correlations indicate that EMCON has a strong positive correlation with a more complicated method. The recall values for the ELR gene set in both result sets from Génie were much lower than EMCON; Génie did not retrieve all genes identified as breast cancer-relevant through expert review. The differences in recall between the two tools may be due to some key differences in the function of Génie; unlike EMCON, Génie relies on gene2pubmed and GeneRIF rather than all available sources of curated GeneID-PMID mappings and does not correct for MeSH term tagging frequency. EMCON is further distinguished from Génie as a standalone, easily searchable resource that is scalable to include updates from any of the included resources.

Other previous efforts in data mining to link genes to disease have included a variety of implementations, including GeneDistiller(168), Endeavor(169), and many more further outlined in Moreau and Tranchevent(170). However, many of these resources have not been updated or maintained, are commercial products, or have limited accessibility for further customized integration. Further distinguishing EMCON from these resources is the use of MeSH term frequency normalization, orthologous genes, and the ease of scaling to include other relevant resources. Finally, EMCON is compatible with previous efforts at putative AOP development, but clearly different in its approach. Putative AOPs have been developed using frequent itemset mining(34) based on shared chemicals in ToxCast and CTD, in an effort to identify MIEs and KEs that may be relevant. In contrast, EMCON works in the opposite direction, i.e. starting with the AO and its associated MeSH terms, with the goal of finding possible targets for an MIE or KE related to an AO of interest.

EMCON has been used as one of several data streams for a gene prioritization project to identify breast cancer gene sets for investigating the molecular mechanisms of mammary carcinogens(90). Other potential applications of EMCON include defining reference gene sets for high throughput transcriptomics chemical screening efforts(128,129). It can be used alongside traditional pathway analysis, as it provides a means of linking differentially expressed genes to perturbations at higher levels of biological organization (tissue, organ, body, etc.). Also, due to the scalability of EMCON, other data can easily be incorporated from chemical and toxicity resources like PubChem(171), the Toxicity Reference Database(28), and ToxCast. These additional resources would provide associations between chemicals and biological entities (genes, pathways, *in vitro* and *in vivo* toxicity endpoints), further expanding the utility of EMCON for hypothesis generation, chemical hazard identification and prioritization, and putative AOP development. Ultimately, EMCON provides a scalable, comprehensive resource to strengthen empirical experimental design and systematic literature review via prioritization of hypotheses based on GeneID-MeSH associations. EMCON is a bioinformatic tool for public health and biomedical sciences that leverages the existing body of information on putative gene-outcome relationships to support research and improve health outcomes.

Summary

Entity MeSH Co-occurrence network (EMCON) was created using a novel data integration pipeline and available information on genes extracted from research articles. The resulting resource, EMCON, can be queried to retrieve a ranked list of genes linked to any topic that is captured in literature. To demonstrate the utility of EMCON, genes linked to breast cancer were retrieved and success was measured by comparing the average rank of a list of known breast cancer genes to randomly generated gene lists. The results show that the known breast cancer genes are collectively ranked higher than any list of random genes indicating the results returned from EMCON are, in fact, relevant to breast cancer. The resource can be extended for any topic of interest that is covered within biomedical literature. This work demonstrates resources that support interoperability i.e., curating information to exposure points of integration, opens up new paths to investigate biology.

CHAPTER 3: TOXREFDB VERSION 2.0: IMPROVED UTILITY FOR PREDICTIVE AND RETROSPECTIVE TOXICOLOGY ANALYSES[2]

Introduction

With an increasing need for rapid screening and prioritization of chemicals for hazard and risk evaluations, researchers are developing new strategies for predicting chemical toxicity. In accordance with these efforts, the Toxicity Forecaster (ToxCast) research program (11) has been developed by the U.S. Environmental Protection Agency (EPA) to assist in the realization of the National Research Council's (NRC) vision for improving rapid assessment of the hazard potential of many chemicals for human, animal, and environmental health (5,172). These efforts served as an impetus to develop the Toxicity Reference Database (ToxRefDB), a digital resource of *in vivo* toxicity study results. ToxRefDB comprises information from over fifty years of *in vivo* toxicity data, largely from summaries of studies performed in accordance with US EPA Office of Chemical Safety and Pollution Prevention (OCSPP) 870 series health effects test guidelines. The database includes information for over 1,000 chemicals, and is being used as a primary source of validation for continued efforts of the ToxCast program(28,29), as well as for numerous predictive and retrospective analyses(25,27,173,174). The utility of ToxRefDB to predictive toxicology is clear; it has been used as the basis for validation of new approach methods (NAMs) to identify specific adverse outcomes of interest(24,26,175,176), as a retrospective benchmark for predictive performance of NAMs (27,173,177,178), and in evaluation of the reproducibility and interpretation of observed *in vivo* outcomes(179,180). ToxRefDB has been used for a wide variety of applications across industry, government, and academia, with 41 other publications citing either Martin et al. (2009) (28) or Martin et al. (2009) (29) in PubMed (APPENDIX 1) as of October 2018. Using ToxRefDB to develop an understanding of the reproducibility and variability in *in vivo* toxicity testing

[2] This chapter has been submitted to the journal Reproductive Toxicology for publication in 2019. Watford, S., Pham, L., Wignall, J., Shin, R., Martin, M., Paul-Friedman, K. (2019). *ToxRefDB version 2.0: Improved Utility for Predictive and Retrospective Toxicology Analyses*..

clearly supports development of baseline expectations for NAMs that promise to assist with rapid prioritization and screening level assessments(181–185). Thus, ToxRefDB represents a seminal resource for predictive toxicology applications, and lessons learned from the initial implementation have been addressed in a major re-development that we describe herein as ToxRefDB version 2.0.

To understand this re-development, it is necessary to further describe previous development and the evolution of ToxRefDB. The first version of ToxRefDB (ToxRefDB v1) initially captured basic study design, dosing, qualitative information for effects, and point of departures (PODs) from summaries of roughly 400 chemicals tested in over 4,000 registrant-submitted toxicity studies, known as data evaluation records (DERs), from the U.S. EPA's Office of Pesticide Programs (OPP). These studies adhered to Office of Chemical Safety and Pollution Prevention (OCSPP) 870 series Health Effects testing guidelines. As this resource was intended to serve as training information in understanding the utility of NAMs like those employed in ToxCast(9,125), the chemical selection for ToxRefDB was originally prioritized to maximize the overlap with ToxCast phase 1 chemicals (ToxCast ph1v1)(1), which were compiled based on commercial availability, solubility in dimethyl sulfoxide, chemical structural features suggesting diversity, and the availability of *in vivo* data, with the result that pesticide active ingredients comprised a high percentage of the ToxRefDB and ToxCast ph1v1 libraries. Expanded efforts in data collection and curation, driven by an attempt to cover as much of the primary ToxCast chemical library as possible, increased the chemical and biological coverage of ToxRefDB v1.3 to over 5,900 i*n vivo* toxicity studies from additional sources, including the National Toxicology Program (NTP), peer-reviewed primary research articles, and pharmaceutical preclinical toxicity studies, among others, for a total of over 1,000 chemicals. As an update to ToxRefDB v1, ToxRefDB v1.3 was released in 2014 to the public as three spreadsheets that consolidated information on adverse effects from the database as well as study citations(186,187).

Though ToxRefDB is unique in its public availability, level of curation, and coverage of chemicals and study types, since the initial release of ToxRefDB v1 in 2009 and through subsequent updates, challenges have surfaced surrounding the extraction, storage, and maintenance of heterogeneous *in vivo* toxicity information. Several stakeholders commented on challenges in using ToxRefDB with respect to the vocabulary used to describe effects, including concerns about grouping effects as "neoplastic" or "non-neoplastic,"(188), as well as the need to be able to integrate the data from ToxRefDB with other public databases(189). Further concerns about the need for negatives, i.e. chemicals tested and shown to be negative for a specific endpoint or effect, to form balanced datasets for predictive modeling(24,25,180,190) strongly relate to the need for an updated vocabulary and determination of which effects are measured in a given study. The desire for more quantitative dose-response information is obvious, given that benchmark dose modeling(191) may provide POD estimates less dependent on specific dose selection. Developmental and reproductive effects, involving complex study designs with multiple generations, also appeared to require a more complex database structure to distinguish effect levels between generations(175). A more nebulous problem that is common to all databases that seek to make legacy information computationally accessible is minimizing data entry error rate. While error rates never reach zero, they could be improved through standardized form-based data extraction with additional layers of quality assurance (QA)(192).

In this work, we further describe the challenges realized through the use and release of ToxRefDB v1, and how these challenges have been addressed to date with development of ToxRefDB v2, including a detailed description of the new content in ToxRefDB v2. The goal of ToxRefDB v2 is to provide a public database that better supports the needs of predictive toxicology by increasing the qualitative and quantitative information available and by facilitating the interoperability of legacy *in vivo* hazard information with other tools and databases. Recognizing that predictive toxicology will require iterative efforts to build computational resources like ToxRefDB, work to generate ToxRefDB v2 has been conducted primarily in three main areas:

- Aggregation of complex and heterogeneous study designs;
- Controlled vocabulary for accurate data extraction, aggregation, and integration; and,
- Quantitative data extraction, including quality assurance and efforts to reduce error rate.

This work represents a significant advancement in increasing the richness of information available for predictive and retrospective analyses from ToxRefDB.

Meeting Challenges in ToxRefDB 2.0

ToxRefDB Overview

Like ToxRefDB v1, ToxRefDB v2 contains summary information for over 5,900 studies labeled “acceptable” for data extraction purposes only, i.e. source document was readable and study design was clear, from six main subsources: DERs from the US EPA OPP (OPP DER), a subset of available NTP study reports (NTP), the open literature (OpenLit), donated pharmaceutical industry studies (pharma), and other (Other; including unpublished submissions and unknown sources) (Figure 3.1A). The study types included in ToxRefDB v2 cover the same study designs as ToxRefDB v1: chronic (CHR; 1-2 year exposures depending on species and study design) bioassays conducted predominantly in rats, mice, and dogs; subchronic (SUB; 90 day exposures) bioassays conducted predominantly in rats, mice, and dogs; subacute (SAC; 14-28 day exposures depending on the source and guideline) bioassays conducted predominantly in rats, mice, and dogs; developmental toxicity studies (DEV) conducted predominantly in rats and rabbits; multigeneration reproductive toxicity studies (MGR) conducted predominantly in rats; reproductive (REP) toxicity studies conducted largely in rats; developmental neurotoxicity (DNT) studies conducted predominantly in rats; and a small number of studies with designs characterized as acute (ACU), neurological (NEU), or “other” (OTH) (Figure 3.1B). Though ToxRefDB v2 contains summary data from roughly the same number of studies and chemicals as ToxRefDB v1, substantial additions that increase the utility of these data have been made.

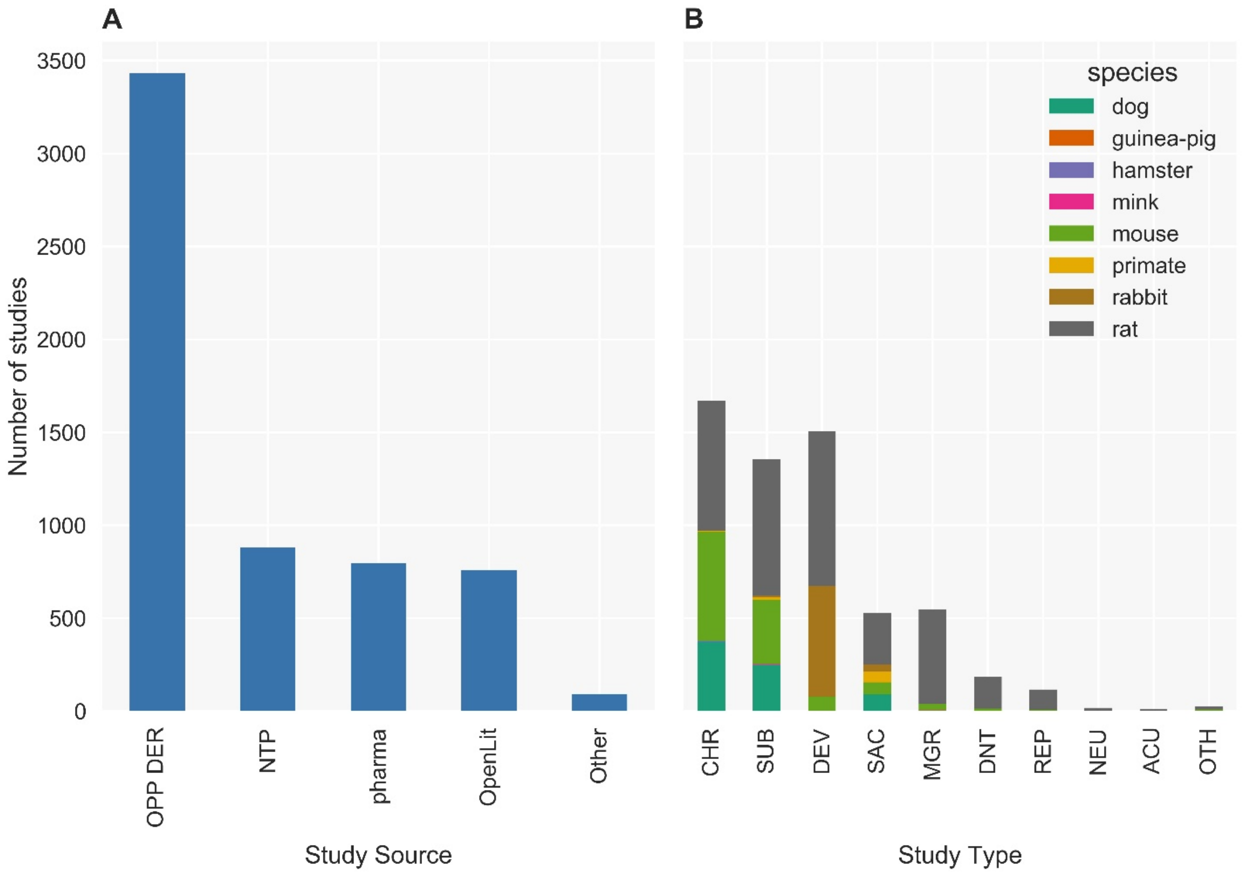


Figure 3.1: Number of studies by study type and species in ToxRefDB v2
(A) ToxRefDB contains over 5,900 animal toxicity studies from a variety of sources include Office of Pesticides Programs Data Evaluation Records (OPP DER), National Toxicology Program study reports (NTP), pharmaceutical preclinical testing (pharma), open literature (OpenLit), and others (Other). (B) The study designs include chronic (CHR), sub-chronic (SUB), developmental (DEV), subacute (SAC), multigeneration reproductive (MGR), developmental neurotoxicity (DNT), reproductive (REP), neurotoxicity (NEU), acute (ACU), and other (OTH) for numerous species, but mostly for rat, mouse, rabbit, and dog.

Aggregation of complex and heterogeneous study designs

Animal studies are designed to address specific hypotheses, with flexibility in study design required to potentially reduce cost, time, and the number of animals needed(193). However, this flexibility presents challenges in structuring the study-related information, so when designing a database to capture both study design and adverse effect-related information, a structure that allows for that flexibility is necessary. An example of the needed flexibility in terms of archiving information on many treatment groups becomes apparent in consideration of the MGR study in rats(194) (Figure 3.2). The addition of quantitative data required a reorganization and expansion of the previous database. Thus, the number of tables and their connections have been significantly increased in ToxRefDB v2 to enable archiving of information from heterogeneous study designs that also may have additions or deletions of treatment groups or doses needed to thoroughly investigate toxicity and complete a study (Figure 3.3).

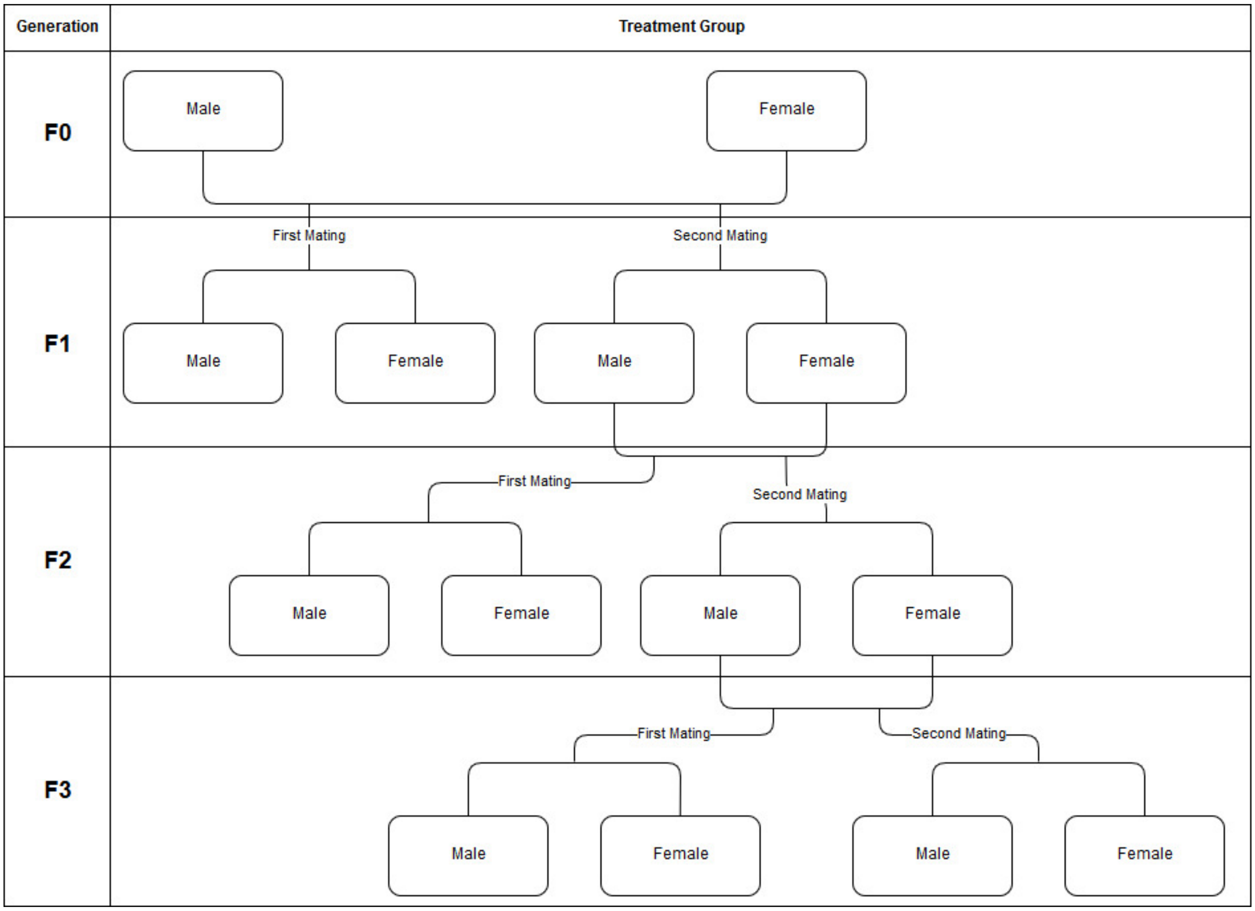


Figure 3.2: Three generation MGR example
This example demonstrates that within the MGR study design, there could be 14 treatment groups, which would then need to be multiplied by the number of doses used in the study. Many of the study designs in ToxRefDB have the potential for the addition of interim, recovery, and satellite groups in order to investigate findings of interest. Even though the guidance in the associated MGR guideline (194) does not require an F3 generation, many studies will report findings from at least a “first mating” treatment group of the F3 generation.

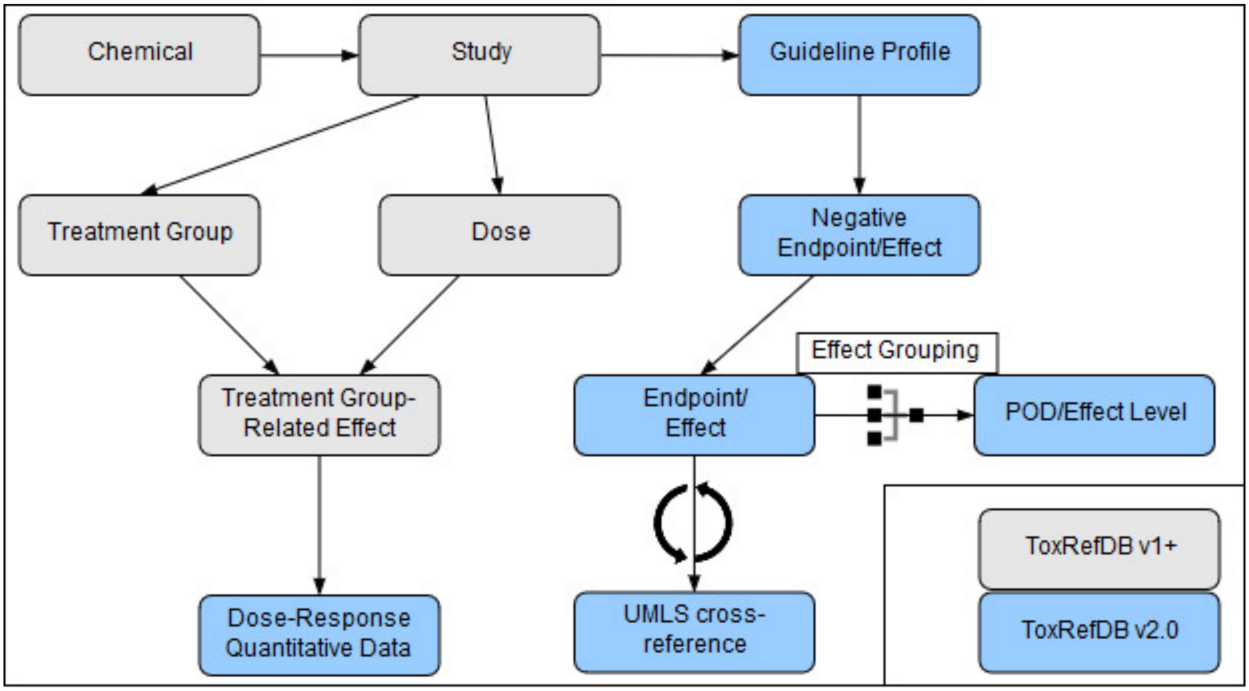


Figure 3.3: ToxRefDB general schema with changes made from ToxRefDB v1 to ToxRefDB v2 Highlighted in blue are the additions to the generic schema to accommodate the updates and additional features for ToxRefDB v2. These include tables to capture the dose-response, quantitative data; guideline profiles for the inference workflow to determine negative endpoints and effects; UMLS cross-references; and effect groupings for systematically calculating PODs and associated effect levels.

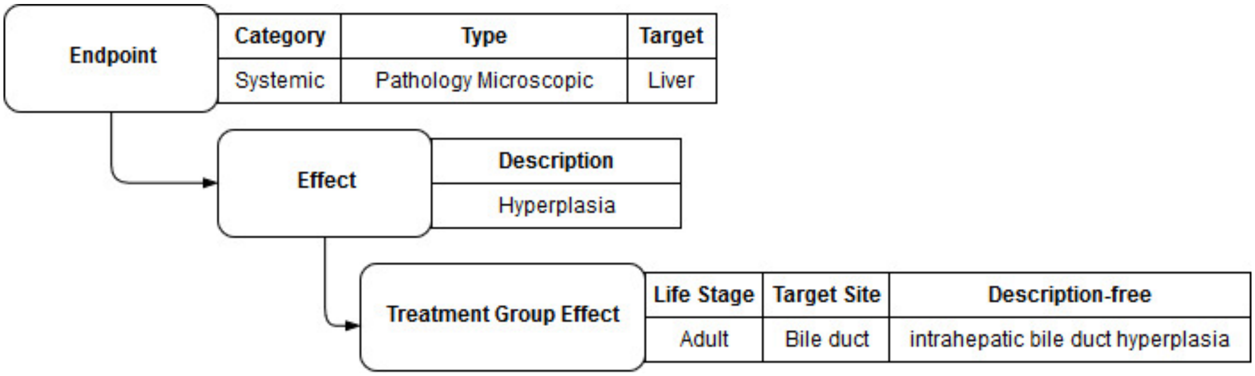
Controlled effect vocabulary for accurate data extraction, aggregation, and integration

Controlled effect vocabulary

A controlled effect vocabulary is critical for any resource aggregating information across a diverse set of sources for efficient retrieval and to enforce semantics, especially within biology (195). ToxRefDB exemplifies this need as it is used for modeling efforts and retrospective analysis. One of the most significant challenges in extracting and/or integrating *in vivo* toxicity studies is the lack of adherence to controlled vocabularies. Inconsistencies in vocabulary arise both as advancements are made to better understand adverse effects in the fields of pathology and toxicology and through preferential terminology in reporting due to differences among experts (196). These inconsistencies can also be seen across studies adhering to the same guideline but conducted years apart. Without adherence to a standard vocabulary that is actively updated and maintained, these studies can only be manually integrated in a way that is unreliably subjective (197). The current landscape of pathology terminology appears to be growing, but with more terminologies for specific species and lesion types available via the Society of

Toxicologic Pathology (STP) as part of the International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) project (198). Perhaps most evolved and ready for current use are solutions in the medical science field that can be seen through the adoption of electronic medical records and electronic health records for reporting adverse events, including data reporting from clinical trials for pharmaceuticals and medical devices (199,200). In fact, current efforts are underway to develop international standards for capturing data from clinical trials, which includes non-clinical data. These efforts are led by the Clinical Data Interchange Standards Consortium (CDISC), where collaboration between international regulatory agencies and their stakeholders is fostered to develop standards for digital submission of clinical trial data (201,202).

Originally, ToxRefDB vocabulary for endpoints distinguished between non-neoplastic and neoplastic lesions, which conformed to the vocabulary used by NTP (79). Improving the controlled endpoint vocabulary for ToxRefDB was a particular challenge because the terminology found in OCSPP guidelines or NTP study specifications may not necessarily match the reported pathology, clinical chemistry, and toxicology study results, where terminology is sometimes more specific. Guideline language needs to be flexible and lasting, rather than overly prescriptive, but this needed flexibility also leads to potential mismatching of information across studies. One demonstrative example is provided by the terminology of the guideline requirement for OCSPP 870.4100, “full histopathology on the organs and tissues…of all rodents and nonrodents in the control and high-dose groups, and all rodents and nonrodents that died or were killed during the study” (203), which doesn’t distinguish between non-neoplastic and neoplastic lesion types nor detail all possible histological findings that could be observed, e.g., hypertrophy, adenoma, fatty changes. In ToxRefDB v1, the effect vocabulary was generally standardized and hierarchically structured into broader categories called endpoints (Figure 3.4 and further described below). Effects were grouped into categories like carcinogenic, neoplastic, and non-neoplastic pathology, organ weight, etc. This categorization was maintained for ToxRefDB v2, however, as mentioned before, the terminology for endpoints reported in the studies did not match the terminology in the corresponding guidelines and specifications. This was problematic for two primary reasons: (1) identifying the correct endpoint within a guideline is required to determine whether or not it was negative and, (2) the endpoint terminology relied on determination of the contribution of a given endpoint to a non-neoplastic or neoplastic process rather than allowing the user to define what effects might be related to cancer phenotypes or other adverse outcomes.



Category	Type	Target
Systemic	Pathology Microscopic	Liver

Description
Hyperplasia

Life Stage	Target Site	Description-free
Adult	Bile duct	intrahepatic bile duct hyperplasia

Figure 3.4: Example of the controlled effect terminology in ToxRefDB v2
An example of the terminology hierarchy is demonstrated for an effect described as “intrahepatic bile duct hyperplasia”. The finding is recorded as the “effect description free”, which is the wording used in the study report. The remaining fields are part of the ToxRefDB controlled terminology. The endpoint category is systemic, the endpoint type is pathology microscopic, the endpoint target is the liver, the effect description is hyperplasia, and the specific observation of “intrahepatic bile duct hyperplasia” was made in the adult life-stage at the specific target site, the bile duct.

The terminology for both endpoints and effects was standardized to better reflect the terminology used in both the OCSPP guidelines as well as what was reported in the summaries in DERs. The primary change made in ToxRefDB v2 is that for the endpoint category “systemic,” the tissue pathology endpoint types are now “pathology microscopic” and “pathology gross,” with no *a priori* suggestion of whether the observation relates to specific cancer or non-cancer related adverse outcomes. Further, duplicative endpoints were standardized, reducing the number of endpoints from approximately 500 to 400. The number of effects remained the same as they were re-binned into the most relevant endpoint. Though the endpoint and effect terminology in ToxRefDBv2 is not comprehensive for all *in vivo* toxicity studies, it captures the observations from the studies and study types currently within ToxRefDB. Each effect can be further qualified to include life stage, direction of effect (increase, decrease, neutral), target site, and exact terms from the source document used to capture the effect (a field called “effect description free”) (Figure 3.4).

The endpoint category “neurological” was not updated and has been left out of the release of ToxRefDB v2. The corresponding effects for that endpoint are associated with 18 NEU and 185 DNT studies. However, the study design, dosing, and treatment group information is still available for these studies in the current 2014 release of ToxRefDB v1.3 (187). The neurological terminology is still under development, with the intention to extend the controlled terminology and extract this information to be available in future updates.

Enabling semantic interoperability

Adopting a controlled terminology for ToxRefDB is beneficial for data extraction and data retrieval, but we can also extend the use to enable semantic interoperability across similar resources archiving *in vivo* toxicology data. This will allow interoperability with other sources that also capture *in vivo* toxicology data like Chemical Effects of Biological Systems (CEBS) (61), International Uniform Chemical Information Database (IUCLID) (83), and eTox (80). We identified resources like CDISC that actively maintain and update controlled vocabularies for all aspects of nonclinical studies. Specifically, we were interested in the terminology developed in Standards for Exchange of Nonclinical Data (CDSIC-SEND) and Study Data Tabulation Model (CDISC-SDTM). These vocabularies are maintained by National Cancer Institute Thesaurus (NCIt), which is a subset of the National Library of Medicine’s (NLM) Unified Medical Language System (UMLS) (78). UMLS is a semantic network linking over 150 terminology resources (CDISC being one of those resources) within the biomedical domain. A UMLS concept is uniquely identified by a concept code. These concept codes were mapped to the controlled terminology defined in ToxRefDB on a manual basis, using the UMLS Terminology Services and NCI Thesaurus browsers. Figure 3.5 describes the completeness of the mapping for endpoints and effects and the coverage from CDISC-SEND and CDISC-SDTM. By cross-referencing ToxRefDB terminology with UMLS, a crosswalk to any other resources that adhere to any of the terminology resources maintained within UMLS is enabled.

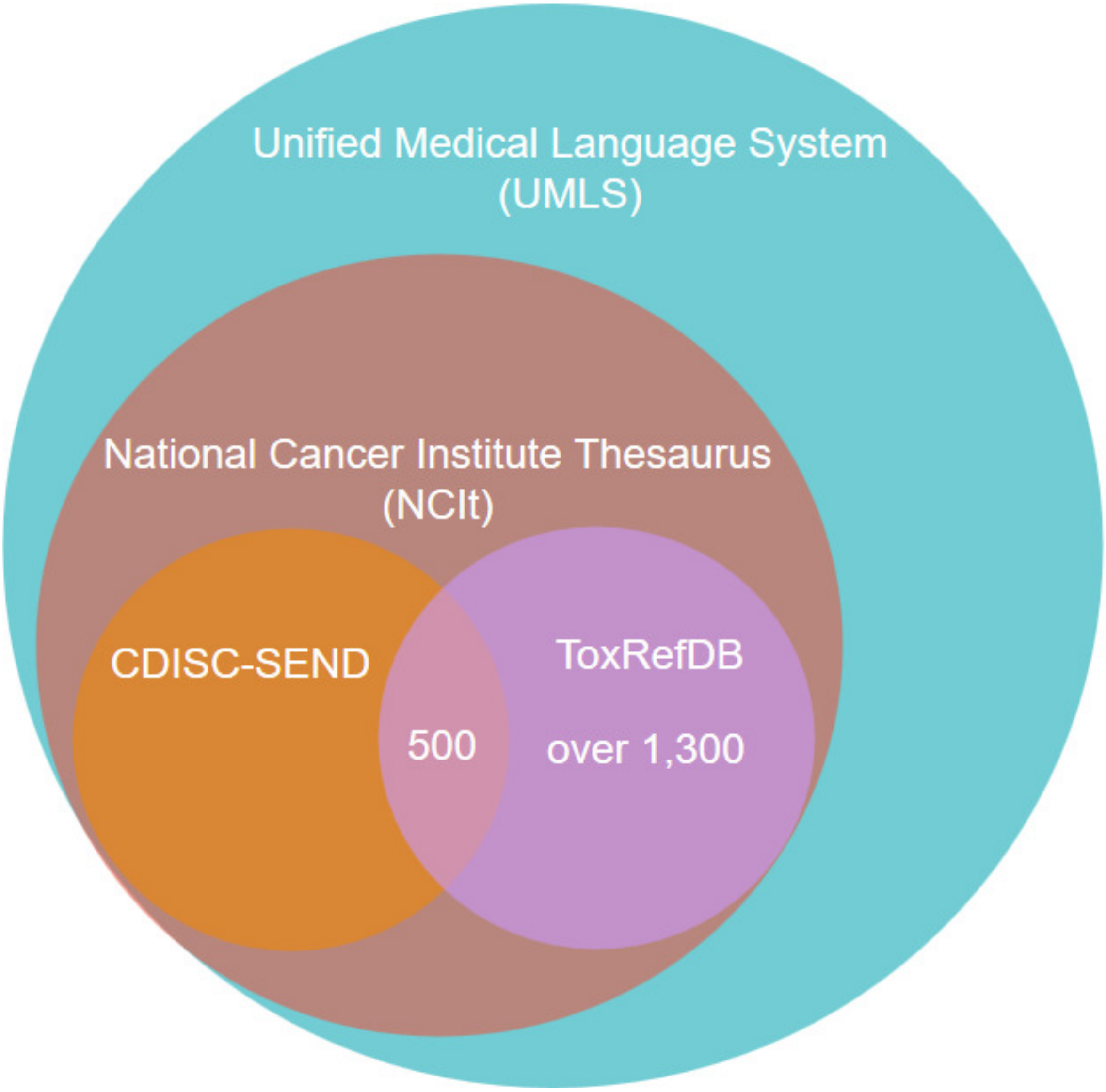

Unified Medical Language System
(UMLS)
National Cancer Institute Thesaurus
(NCIt)
CDISC-SEND
500
ToxRefDB
over 1,300

Figure 3.5: Terminology sources and membership of UMLS concept codes cross-referenced to ToxRefDB endpoints and effects
Over 1,800 UMLS concept codes were mapped to endpoints and effects in ToxRefDB. Only 500 of those concept codes are a part of the CDISC-SEND terminology. All of the concept codes are a part of vocabularies within both National Cancer Institute Thesaurus (NCIt) as well as UMLS.

Quantitative data extraction, quality assurance, and efforts to reduce error rate

<u>Study extraction process</u>

Initially, studies available in ToxRefDB v1 lacked quantitative, dose response information; the quantitative information and its application is described in the next section in more detail but served as a strong impetus to motivate re-extraction of the studies in ToxRefDB. This task initially proceeded using an Excel file-based extraction. However, there were faults in this process that required manual corrections after uploading study extractions to the ToxRefDB MySQL database, including: inconsistent comments, different number of animals for the same treatment group, and added effects outside of the controlled terminology. Thus, following initial attempts with Excel-based extraction, an Access database file was generated from the MySQL database for each study (Figure 3.6). The Access database files featured several improvements, including: standardized options for more consistent reporting in some fields, such as the units on time and dose, dose-treatment group, and effect information; checkbox reporting for observation status on each endpoint and effect; and a log for tracking changes and facilitating QA. Nearly 32% of the studies were extracted using Excel-based approach, with the remaining studies extracted using the Access database approach. Switching to Access database files significantly reduced errors and increased standardization of reporting.

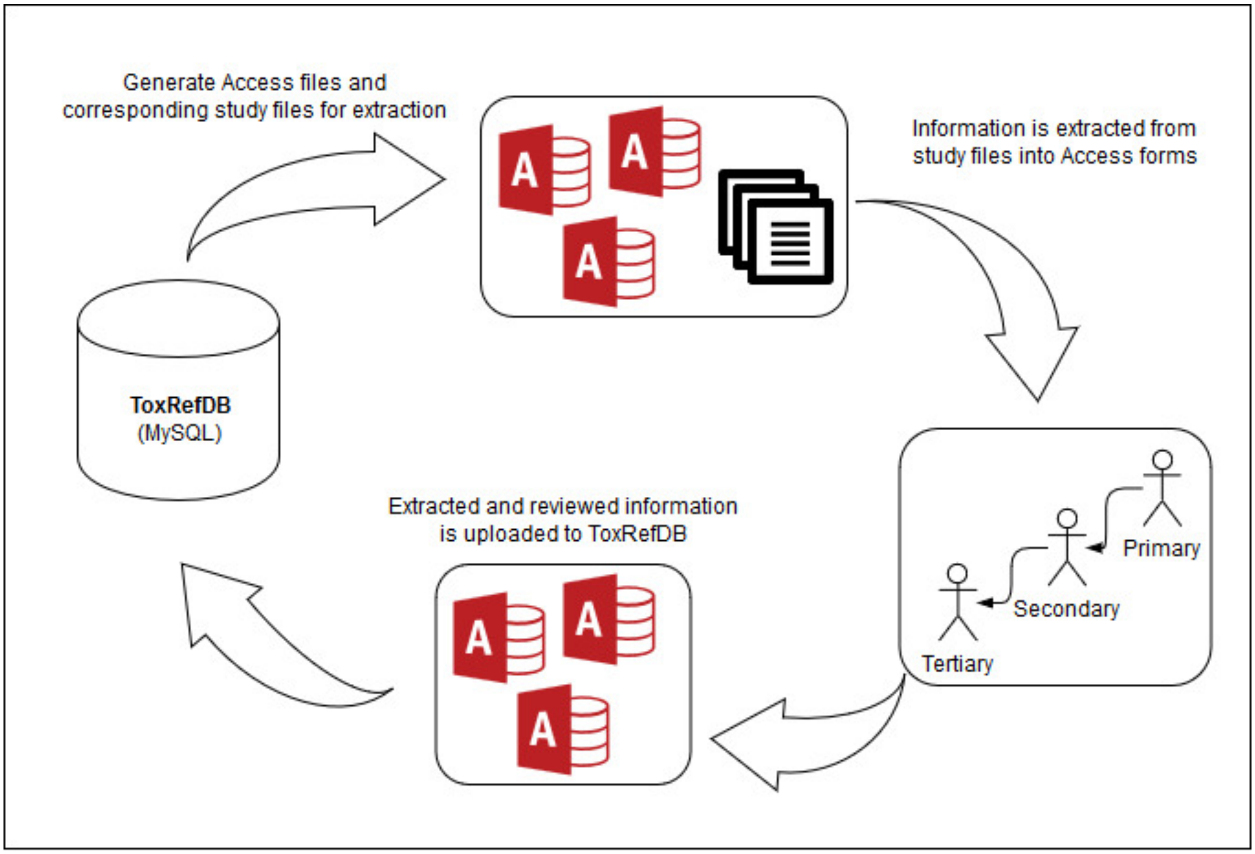

Generate Access files and
corresponding study files for extraction
Information is extracted from
study files into Access forms
ToxRefDB
(MySQL)
Primary
Secondary
Tertiary
Extracted and reviewed information
is uploaded to ToxRefDB

Figure 3.6: Data extraction and review workflow
Access databases are generated for each study and batched to data extractors with the corresponding source files. The data in the Access databases are curated with additional data extracted from the source files with up to three levels of review. The Access databases are batched back and the data is imported back into the MySQL database.

Guidance for data extraction was stratified first according to study type (e.g., CHR, SUB, DEV, MGR) then by study source (e.g., OPP DER and NTP) because of the differences in both study design and adverse effects required for reporting as stated in guidelines. The process used to extract study information was also an important aspect of QA efforts for ToxRefDB v2. First, a primary reviewer extracted study, dose, treatment group, effect, and endpoint observation information, per standard operating procedures provided to the reviewer. The instructions detailed how to review the toxicological data and extract it from the original data sources consistently across reviewers using the Access database. This was reviewed by a second, senior reviewer, who was asked to review all extracted information as if they were extracting it again and, also, to review the comment log from the primary reviewer. Finally, if either the primary or secondary reviewer noted that it was necessary, an additional senior toxicologist reviewed the comment logs, extracted information, and resolved any conflicts or questions prior to finalization of the extraction. The final, tertiary review occurred for approximately 10% of the studies. All reviewers were trained in the procedures prior to reviewing studies. For release of ToxRefDBv2, the full quantitative data extraction for all CHR and SUB studies were completed, with quantitative data extraction completed for many other study types and sources as well (additional quantitative data will be added in updates to the version 2 release). Table 3.1 lists the current number of studies with quantitative data extracted by study type and source.

A	Study type	Study source	Number of studies extracted
	CHR	NTP	347
	CHR	OPP DER	1,079
	CHR	OpenLit	9
	DEV	NTP	10
	DEV	OPP DER	958
	DEV	OpenLit	1
	DEV	Other	6
	MGR	OPP DER	345
	MGR	OpenLit	1
	MGR	Other	20
	SAC	NTP	59
	SAC	OPP DER	25
	SUB	NTP	247
	SUB	OPP DER	769
	SUB	OpenLit	6
	Total		3,882

B	Study type	Number of chemicals
	ACU	10
	CHR	663
	DEV	710
	DNT	124
	MGR	458
	NEU	18
	OTH	18
	REP	77
	SAC	191
	SUB	659
	Total chemicals	1142

Table 3.1: Extraction progress as of ToxRefDB v2 release
Over 65% of the studies have been curated with dose-response, quantitative data extracted. Priority was given to chronic (CHR) and sub-chronic (SUB) OPP DERs and NTP study reports, which are completed. The remaining studies are predominantly from OpenLit and pharma. A) The number of studies per source by study type. B) The number of chemicals per study type.

Critical Effect Determination

ToxRefDB v1 and v2 have several effect levels stored; treatment-related effects define lowest effect levels (LELs) and no effect levels (NELs), whereas critical effect designations define the lowest observable adverse effect and no observable adverse effect levels (LOAELs, NOAELs). A critical effect level is defined as the dose at which a treatment-related effect is deemed to have toxicological significance. Critical effects are typically used to define the study-level POD for regulatory toxicology applications. Not all studies within ToxRefDB v1 had been assessed for critical effects, or the critical effects had not been extracted. For extractions of OPP DER files, the critical effect was simply captured from the source document as previously identified by toxicologists who had reviewed the original study file. If for a given source document, particularly those from sources other than OPP DER files, critical effect information was lacking, senior toxicologists trained to extract information for ToxRefDB reviewed the study and determined the critical effects and critical effect levels. For each study, the reviewers determined the critical effect and lowest observed adverse effect level(s) (LOAEL) using a weight-of-evidence (WoE) approach (204), like the approach used to evaluate registrant-submitted studies for generation of DERs. Using this approach, the identification of potential critical effects from a given study was determined based on statistical significance, considerations of biological relevance, and consistency across multiple endpoints (in the presence or absence of statistical significance) to select the appropriate LOAEL value(s) and the overall study LOAEL. The WoE evaluation included review of all pertinent information so that the full impact of biological plausibility and coherence was adequately considered. This approach involves weighing individual lines of evidence and combining the entire body of evidence to make an informed judgment. Judgment about the WoE involved considerations of the quality and adequacy of data, and consistency of responses induced by the agent in question. The WoE judgment required combined input of relevant disciplines. Generally, no single factor determined the overall weight; all potential factors were judged in combination. The results of these reviews were recorded along with appropriate rationales and can be found in ToxRefDB v2.

Quality assurance and quality control efforts to reduce error rate

Error rate is an inherent problem for legacy databases as much of the source information was entered manually, so human errors resulting from transcription are impossible to completely avoid (192). However, as part of the ToxRefDB v2 effort, increased QA measures to promote greater fidelity of the information captured, which included numerous quality control (QC) checks to ensure data integrity were implemented. First, studies were extracted utilizing a defined QA process, with multiple levels of review and Access form-based entry (described previously) to prevent extraction errors as described above. Upon upload into ToxRefDB v2, these extractions were required to pass QC measures because, although the Access database files enforce the MySQL database constraints as well as use of the controlled terminology to minimize data entry error, logical errors can persist. We checked a series of potential logical errors after the extracted was uploaded through the import script. These errors were identified by defining a series of tests up front that must resolve to a particular answer. Below are some of the logical errors that were flagged using a QC check following import of the Access database files to the MySQL database:

- Dose level numbering did not correspond to the total number of doses;
- Duplication of concentration/dose values, including two control doses;
- No concentration and no dose adjusted value for a reported effect (possible extraction error or possibly that the effect was qualitatively reported);
- The critical effect level is at a dose below where treatment-related effects were observed; and/or,
- The control was incorrectly identified as a critical effect level.

Any of these issues that could not be resolved systematically were flagged to undergo a second round of extraction and QA to correct. Though QC is an ongoing and evolving process, these QC checks are serving as an improvement to the overall database and database development process.

An additional ongoing problem for reporting quantitative data from clinical or related laboratory findings is unit standardization (205,206). No guidance is provided on how to report findings in the OCSPP guidelines nor from any other sources, so units were extracted exactly as they were presented in the reports. The units were standardized by eliminating duplicate entries for the same units that were originally entered differently or with typographical errors. Units were only standardized, so no conversions were made. Further work must be undertaken to further standardize units and define conversions that can be systematically automated.

In order to understand how increased quality assurance and quality control may have affected quantitative information in ToxRefDB, a comparison of study level LEL and LOAEL values for 3,446 studies between ToxRefDB v1.3 and v2 was conducted. This evaluation showed ~95% concordance for the LELs and ~90% for LOAELs between ToxRefDB v1.3 and ToxRefDB v2. Though the values in v1.3 and v2 were largely concordant, addition of critical effect review for studies that previously lacked a critical effect and/or error correction account for the minute differences. The magnitude of the differences ranged from 0.1 log10-mg/kg/day to 2.4 log10-mg/kg/day, with an average difference of 0.52 log10-mg/kg/day for LEL and 0.57 log10-mg/kg/day average difference for LOAELs.

Distinctions between negative effects and not tested effects

Many study sources only report information on the adverse effects, and data extracted in ToxRefDB v1 reflected this (i.e. only contained data for positive or treatment-related effects). These values were reported as lowest effect levels (LELs) or lowest observable adverse effect levels (LOAELs), with no effect and no observable adverse effect levels (NELs, NOAELs) inferred as the next lowest dose, respectively. A positives-only database presented a major challenge for predictive modeling applications that require balanced training sets of positive and negative findings: the user was left to infer negatives from the database without the guidance of what was tested and reported for the study based on its adherence (or non-adherence) to a guideline. Finding a solution to systematic and accurate inference of negatives involved leveraging the new controlled effect terminology to match the OCSPP guidelines (described above) and annotating endpoints as required, triggered, or recommended. Required endpoints are always tested according to the guidelines, whereas triggered endpoints are required under specific circumstances, e.g. if a chemical is known to perturb a specific system based on information from previous studies. A recommended endpoint is not always tested but are mentioned as important in the guidelines. All other endpoints not explicitly mentioned in the guidelines were assumed to be not required. The collections of endpoint annotations for guidelines are referred to as guideline profiles.

These guideline profiles enable assumptions about whether an endpoint was tested for a given study based on which guideline the study followed. A majority of the studies (58%) described in ToxRefDB are based on OPP DERs, which summarize registrant-submitted data in accordance with OCSPP series 870 Health Effects Testing Guidelines as seen in Table 3.2. Additionally, though not strictly referred to as guidelines, study specifications for SAC, SUB, and CHR studies from NTP (207) were also reviewed and developed into guideline profiles to allow for their inclusion in determination of negatives. Developmental and reproductive studies from the NTP were not included in guideline profile development at this time due to the assumption that these studies may have been highly customized based on the experimental need, and as such inference of negatives may not lead to accurate conclusions (personal communication, John Bucher and Paul Foster). Because the studies included in ToxRefDB span decades, we also included guideline profiles for updated guidelines. For example, since testing requirements were added to the MGR guideline (OCSPP 870.3800) in 1998 (194), the MGR study type has two associated guideline profiles: one for studies conducted before 1998 and another for studies conducted in 1998 and later. All of the guideline profiles were reviewed by an independent senior toxicologist familiar with the guideline and guidance documents.

Observations were recorded and confirmed in the data extraction process for each to reflect concordance with guideline profiles, deviations, endpoints that were measured following a trigger, etc. An observation is defined as the testing and reporting status of a given endpoint in the study. Extractors made decisions about testing and reporting status as described in Table 3.3, where for example endpoints that were reported as tested can be differentiated from endpoints that are assumed to be tested based on the guideline profile. The important result of the development of these guideline profiles is that missing or not tested data can now be distinguished from negative (tested with no effect seen) for a large fraction of the studies described in ToxRefDB v2. The inference workflow to determine negative effects based on observations and guideline profiles is described in Figure 3.7.

Guideline number	Guideline name	Study Type in ToxRefDB
870.3100	90-day Oral Toxicity in Rodents	SUB
870.3150	90-day Oral Toxicity in Nonrodents	SUB
870.3250	90-day Dermal Toxicity	SUB
870.3465	90-Day Inhalation Toxicity	SUB
870.3550	Reproduction/Development Toxicity Screening Test	REP
870.3700	Prenatal Developmental Toxicity Study	DEV
870.3800	Reproduction and Fertility Effects	MGR
870.4100	Chronic Toxicity	CHR
870.4200	Carcinogenicity	CHR
870.4300	Combined Chronic Toxicity/Carcinogenicity	CHR
870.6200	Neurotoxicity Screening Battery	NEU
870.6300	Developmental Neurotoxicity Study	DNT
870.3050	28-day Oral Toxicity in Rodents	SAC

Table 3.2: OCSPP 870 series health effects guidelines in ToxRefDB

Tested status	Reported status	Example
Tested	Reported	The endpoint was SPECIFICALLY written in the text of the study source indicating that data was collected (default if required by the guideline for that study type)
Not tested	Reported	The endpoint was SPECIFICALLY written in the text of the study source indicating that data was NOT collected, even if required by the guideline
Tested	Not reported	The endpoint was NOT specifically written in the text of the study source, however other evidence indicates it can be deduced that it was tested (or was required by the guideline to be tested)
Not tested	Not reported	The endpoint was NOT specifically written in the text of the study source and is not required by the guideline, so we assume that the endpoint was not collected in this study

Table 3.3: Observations for guideline profiles
The tested status indicates if the endpoint was evaluated or not by the given study. The reported status indicates if the testing status was reported in the given study. Combining the tested and reported status yields the observation status for the specific endpoint of interest on a study-by-study basis.

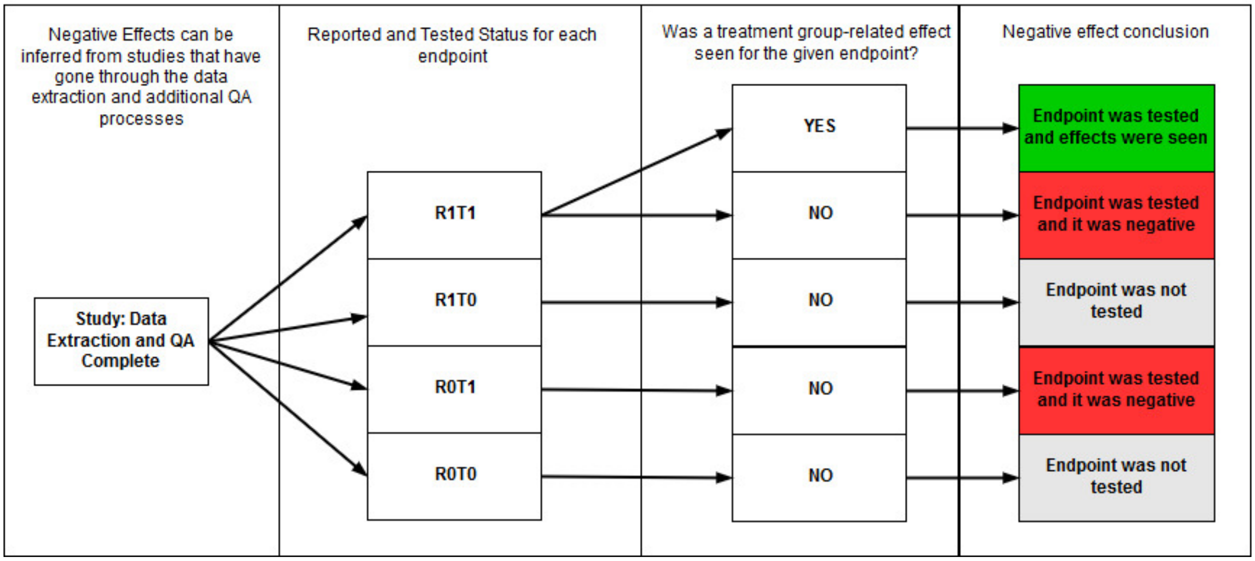

Negative Effects can be inferred from studies that have gone through the data extraction and additional QA processes
Reported and Tested Status for each endpoint
Was a treatment group-related effect seen for the given endpoint?
Negative effect conclusion
Study: Data Extraction and QA Complete
R1T1
R1T0
R0T1
R0T0
YES
NO
NO
NO
NO
Endpoint was tested and effects were seen
Endpoint was tested and it was negative
Endpoint was not tested
Endpoint was tested and it was negative
Endpoint was not tested

Figure 3.7: Inference workflow to determine negative effects
Four steps are taken to systematically infer true positives and negatives: (1) study extracted completely; (2) application of the observation status; (3) determination of the effect seen (yes/no) on the basis of statistically significant findings; (4) conclusion, with true positive (green), true negative (red), not tested (orange), and inconclusive (gray) as possible outcomes.

Study Reliability (ToxRTool)

A majority of the studies referenced within ToxRefDB were extracted via summaries from OPP DERs, and these studies typically follow OCSPP 870 series Health Effects Testing Guidelines; however, as ToxRefDB was expanded, other studies were summarized from various sources, including: NTP, pharma, OpenLit, and Other. NTP and pharma studies were considered guideline-like, as a study guideline or specification that these studies resembled could be identified, but OpenLit studies were not assumed to conform to any guideline. Therefore, all open literature studies were assessed for reliability and guideline adherence. The Toxicological Data Reliability Assessment Tool (ToxRTool) was adapted for this assessment(208). ToxRTool is an Excel application that includes questions across 5 criteria with numerical responses that are summed to lead to a Klimisch score: a score ranging from 1-4 that captures an overall assessment of reliability(209). The ToxRTool was adapted specifically in the following ways for this project:

- Added Guideline Adherence Score (an initial question for the reviewer regarding the study's adherence to or consistency with OCSPP guidelines with a five-point rating scale) further described in Table 3.4.
- Added "Context of Tool and Rationale/Intent for Study" field (an open-text field to insert the purpose of the study quality review to address the concern raised by Segal et al. (2015) (210) that the intended purpose of the ToxRTool-facilitated review could influence evaluations).
- Added additional scoring notes (to help the reviewers assign scores consistently).
- Added option for "0.5" rating for selected criteria (for some questions considered more subjective than others, if the reviewer concluded the question was partially fulfilled).

A total of 522 OpenLit studies were assessed with the ToxRTool with scores ranging from 8 to 23 with 23 being the highest score. The majority of the studies reviewed for ToxRefDB v2 corresponded to Klimisch quality scores of 1 (ToxRTool score of ≥ 18) or 2 (ToxRTool score of 13-18). The ToxRTool scores could be used as a quality flag both to qualify and prioritize studies for the extraction process, or by users who are performing reviews of information on a single chemical basis.

Score	Description
5	Adheres to modern* OECD/EPA guideline for repeat-dose toxicity studies (explicitly stated by authors; broad endpoint coverage and ability to assess dose-response)
4	Adheres to an existing or previous guideline (explicitly stated by authors; previous version of OECD/EPA guidelines or FDA guidelines)
3	Not stated to adhere to guideline but guideline-like in terms of endpoint coverage and ability to assess dose-response (e.g., NTP). Please see Quick Guide to EPA Guidelines for Chronic and Subchronic studies. In this table, you can easily assess whether the study was guideline-like in terms of the animals used (species, sex, age, number), dosing requirements, and reporting recommendations.
2	Unacceptable adherence to guideline (intended to adhere to guideline but had major deficiencies)
1	Unacceptable (no intention to be run as a guideline study, purely open literature or specialized study)

Table 3.4: Guideline adherence scoring added to ToxRTool
Note that many of the studies extracted, particularly from sources like the NTP and open literature, were never intended to adhere to a guideline and as such “unacceptable” in this case only refers to their guideline adherence and not the study design itself.
*A study is considered as adhering to “modern” OECD/EPA guidelines if it was published after 1998, which is the date that many Health Effect 870 series guidelines were re-published.

Extensions of ToxRefDB v2 updates for research applications

Systematic calculation of point of departures (PODs) and related effect levels

Related to the new ToxRefDB v2 controlled effect terminology is the application of this terminology for calculation of PODs and related effect levels for various modeling and retrospective analyses in the predictive toxicology realm. For purposes of predictive toxicology, PODs can be computed per chemical (i.e., lowest dose that produced effects or adverse effects across all study types included in the database) or per study (i.e., lowest dose that produced effects or adverse effects in a given study of interest). PODs computed by chemical could be broken down into a POD for some combination of effects in a POD “category,” e.g., the lowest dose that produced effects or adverse effects on developmental or reproductive effects as a group. Acknowledging that the specific application may define the appropriate aggregation of the effect data in ToxRefDB for calculation of PODs, ToxRefDB v2 (Figure 3.3) enables definition of the list of effects to be grouped together, followed by storage of the PODs calculated based on that list. A collection of effect groupings is referred to as an effect profile. An initial set of effect profiles were created to define custom grouping of effects from the study, treatment groups, and effects. For example, all developmental effects, across studies, could be combined to give a POD, or minimum LOAEL or LEL value, for developmental effects. The NEL and NOAEL are designated as the next lowest doses from the LEL and LOAEL, respectively. A complication in providing PODs is that not every effect is necessarily of toxicological significance and may not correspond to the critical effect level as reviewed by toxicologists. In the case that no effects of toxicological significance were observed for a given category of effects, or by study, the LOAEL is greater than the highest dose tested and the NOAEL is greater than or equal to the highest dose tested for that effect (i.e., a “free-standing NOAEL”). For all POD types, including NEL, NOAEL, LEL, and LOAEL, a qualifer (<, >, or =) is provided to assist with quantitative interpretation of these values.

The effect profiles are an important feature addition and address problems previously highlighted(189); essentially, the endpoints and effects in ToxRefDB can be grouped a number of ways, which may lead to differing interpretations. However, there is no single way to create POD values via grouping of effects, as differing interpretations may be equally valid for divergent applications of the data. The two effect profiles currently available in ToxRefDB v2 are summarized in Table 3.5 for clarity, with the expectation that as use of the database grows, additional effect profiles can be added. It should be noted that these effect profiles, and the POD values generated in using them, are for research purposes and do not necessarily reflect POD values that may be used in chemical safety evaluations.

First, effect level data were grouped by study type, endpoint category, and life stage. This first effect profile produced POD values for each study type, life stage, and endpoint category combination. This first effect profile was used to calculate effect levels for the CompTox Dashboard (17).

A second effect profile was also employed, where PODs were calculated for each endpoint category-endpoint type pairing, except in the case of the systemic endpoint category, where PODs were reported for each endpoint targets (i.e., organs). This second effect profile produced POD values for cholinesterase, developmental, and reproductive endpoint categories; hematology, in-life observation, and urinalysis endpoint types; and organ-specific endpoint targets (e.g., liver). Either of these effect groupings and associated effect levels may be useful for research purposes as a meaningful way of considering many pieces of information for a chemical at one time.

<table>
<tr><th colspan="4">Effect profile id
Description</th><th>Example output</th></tr>
<tr><td rowspan="2">1</td><td colspan="3">Endpoints are grouped by study type, life stage, and endpoint category to produce a POD.</td><td rowspan="2">NEL, LEL, NOAEL, and LOAEL will be presented for combinations, e.g.:

MGR/systemic/adult
CHR/cholinesterase/adult
DEV/systemic/fetal
DEV/reproductive/adult-pregnancy</td></tr>
<tr><td>Study type
• SAC
• SUB
• CHR
• DEV
• MGR
• OTH</td><td>Endpoint category
• Cholinesterase
• Reproductive
• Developmental
• Systemic</td><td>Lifestage
• Adult
• Adult_pregnancy
• Juvenile
• Fetal</td></tr>
<tr><td rowspan="4">2</td><td colspan="3">Endpoints are grouped according to endpoint category, endpoint type, or endpoint target. The endpoint target is used for systemic pathology endpoints, i.e. if the effects are organ-level, the POD is reported for the organ system.</td><td rowspan="4">NEL, LEL, NOAEL, and LOAEL will be presented for combinations, e.g.:

Cholinesterase
Reproductive
Developmental
Systemic/liver
Systemic/clinical chemistry</td></tr>
<tr><td colspan="3">Endpoint category
• Cholinesterase
• Reproductive
• Developmental</td></tr>
<tr><td colspan="3">For the systemic endpoint category, either the endpoint type or organ are used</td></tr>
<tr><td colspan="3">Endpoint Type:
• Clinical chemistry
• Hematology
• In life observation
For pathology microscopic and gross, and organ weights, the organ name is used, e.g.: liver, kidney, heart</td></tr>
</table>

Table 3.5: Effect profiles in ToxRefDB v2 for POD computation
Two effect profiles group effects for computation of POD values, i.e. NEL, LEL, NOAEL, and LOAEL values, which can be used in research applications.

Benchmark Dose (BMD) Modeling for ToxRefDB

Quantitative dose-response modeling yields PODs for research applications that are less dependent on the doses selected for a given study, and ensures that dose values selected correspond to similar levels of effects across studies (211). Though there are many possible approaches to curve-fitting(16,212), the US EPA Benchmark Dose Modeling Software (BMDS)(213–215) has become the canonical tool for use in toxicology regulatory and research applications(212,216–218). Using BMDS to fit the quantitative response data in ToxRefDB provides modeled values, e.g., benchmark dose (BMD) values, using the default recommendations from the BMDS guidance(191). In ToxRefDB v2, we report the results from the largest use of batch processing with BMDS v2.7 employed to date, using a Python package(219) (Pham et al). The objectives in reporting this demonstration within the database are (1) to promote the use of BMD values in development of predictive toxicology solutions for regulatory applications; and (2) to demonstrate the feasibility of large-scale BMDS analysis of legacy toxicology information. The BMD values reported are in no way intended to reflect any regulatory decision-making on a single chemical basis.

Over 92,000 quantitative dose-response datasets, i.e. data from chemical-effect pairs, from complete study extractions with at least 3 non-control dose levels in ToxRefDB v2 were filtered to yield datasets amenable to modeling using BMDS. For each dose group of a study, BMDS analysis requires the dose, N, and dichotomous incidence or the continuous effect level mean and variance. In large part due to inconsistent or incomplete reporting of variance for continuous responses, only about one-third of the total datasets were available for BMDS modeling (nearly 28,000). For each modeled dose-response, the data were grouped according to the response type, i.e., continuous response, continuous response for organ and body weights, dichotomous response, or dichotomous cancer response. The response type guided selection of the models and benchmark response (BMR) used in the automated analysis, as shown in Table 3.6, as recommended by the BMDS guidance(191). The effect terminology corresponding to cancer is available in APPENDIX 2.

The current BMD table in ToxRefDB v2 holds results for nearly 28,000 datasets. This includes BMD models with 1 standard deviation for continuous data, 10% relative deviation for organ/body weight, and 5 and 10% BMR for dichotomous data. Almost 90% of the datasets were successfully modeled and have at least one recommended model and therefore a BMD value (Figure 3.8). Currently, there are 627 unique chemicals (as indicated by CAS registry number) with at least one modeled BMD value. The lower 95% confidence limit on the BMD value estimate, known as the BMDL, is also stored in the database. However, some recommended models are associated with cautions for using the BMD and BMDL values that were auto-generated by BMDS (v2.7). For example, there may be warnings to indicate a large distance between the BMD and BMDL, or that a computed BMDL is likely imprecise because the model has not converged. All warnings related to model recommendations are stored in the “logic cautions” column of the “bmd” table in the database and should be considered by the end user of the data. The rates at which recommended models were achieved for each data type and BMD model type are illustrated in Figure 3.9.

One hypothesis in modeling these dose-response data is that BMDL values will tend to be lower than the discrete NOAEL or NEL values for a given study-level effect. Indeed, most of the recommended BMDL values are less than the stored NOAEL and NEL values for that effect. For datasets (continuous and dichotomous) successfully modeled using a BMR of 5% extra risk (dichotomous) or 5% increase relative to control mean estimate (continuous), 98% of the BMDL values were less than the corresponding NOAEL values, and 66.4% of the BMDL values were less than the corresponding NEL values. For the datasets with a BMR of 10% extra risk (dichotomous) or 10% increase relative to control mean estimate (continuous), 95.7% of the BMDL values were lower than the NOAEL values, and 48.5% of the BMDLs were lower than the corresponding NEL values. This is mostly consistent with previous works showing that modeled BMDLs are a more conservative estimation of PODs than the statistically derived NOAELs(220,221). Similarly, we also compared BMD values to LEL and LOAEL values. For the datasets that used a BMR of 5% extra risk (dichotomous) or 5% increase relative to control mean estimate (continuous), BMD values were less than their corresponding LEL or LOAEL values approximately 89% of the time. For the datasets that used a BMR of 10% extra risk (dichotomous) or 10% increase relative to control mean estimate (continuous), the BMD values were less than the corresponding LEL values 80% of the time and less than the corresponding LOAEL values 81% of the time.

All study type and BMDS-amenable data were used for this exercise to model the largest dataset possible. Some caution needs to be taken when evaluating the BMD models, particularly for studies where multiple generations are evaluated. Ideally, for MGR and DEV studies, a nested model should be used to calculate BMDs for the litters and, if needed, a correction for the degree of variability or sample size adjustment. However, due to the availability of information from source files, and the data structure in ToxRefDB v2, information from individuals in each litter were not available. Therefore, the summary data and statistics for litters were used for BMD modeling.

Data type	Number of datasets available for BMDS	Benchmark responses used	Have Recommended BMD Model	Chemicals (n) with a Recommended BMD Model / Total
Cancer	1 170	5%	1 101	243 / 247
		10%	1 107	246 / 247
Non-cancer dichotomous	17 318	5%	16 059	609 / 612
		10%	16 165	611 / 612
Continuous body/organ weight	9 268	10% relative deviation	4 151	416 / 430
Continuous non-body/organ weight		1 standard deviation	3 006	284 / 300

Table 3.6: Number of datasets in ToxRefDB v2 for BMDS modeling
Each dataset consists of all doses, number of test animals, effect values, and variance information if available. The table shows the number of modeled datasets, the BMR used, the number of recommended models, and number of chemicals with a recommended BMD model by type of data. The chemical counts are by data type, as the same chemical can have data in multiple data types. The “total” number of chemicals refers to the total number of chemicals associated with the datasets available for BMDS.

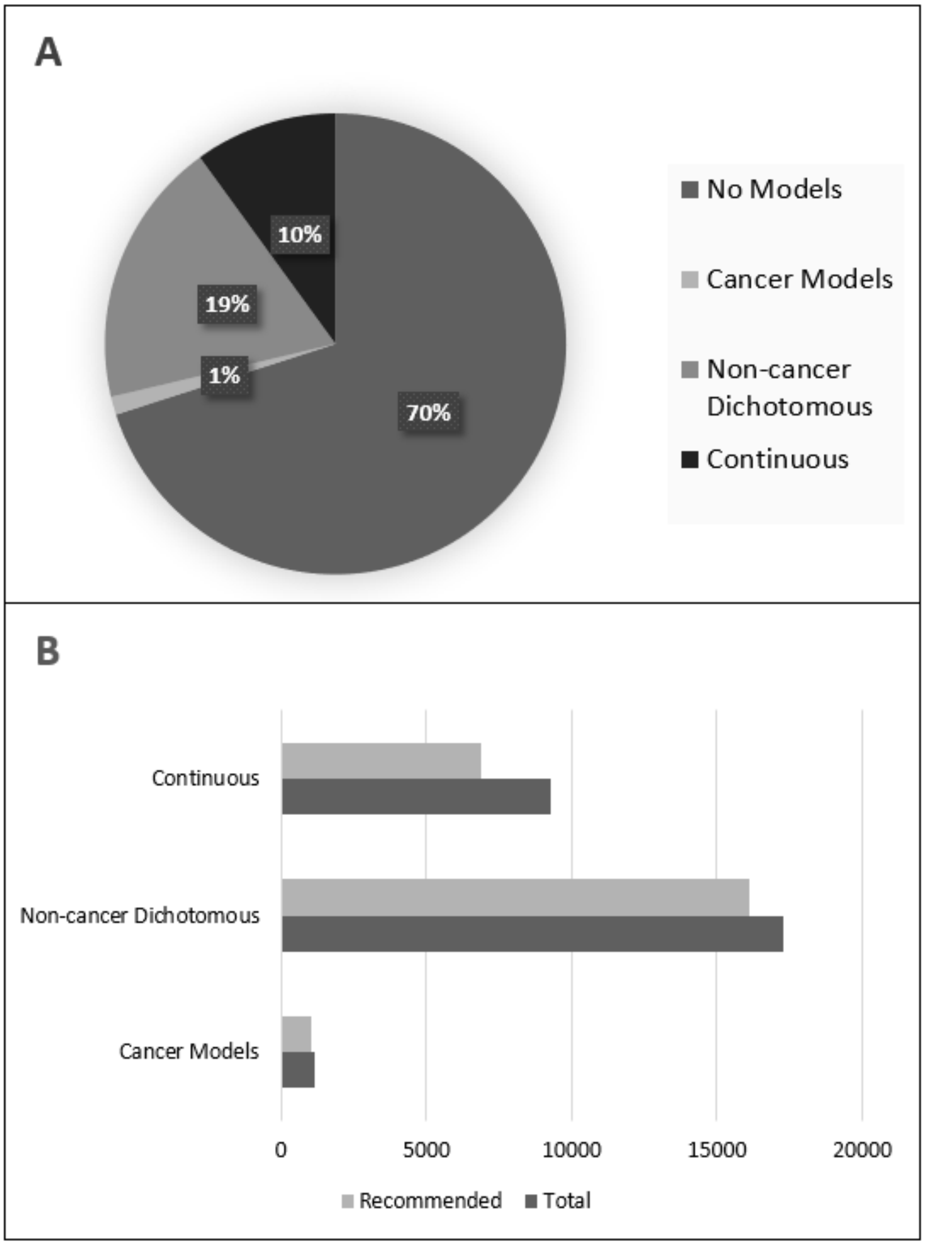


Figure 3.8: Proportion of effects in ToxRefDB that can be modeled
(A) Of the 92,646 quantitative dose response datasets, approximately one third met data quality filters for BMDS. (B) There were 27,756 datasets that met data quality filters for using BMDS. Of the ~28k datasets, 87% produced a recommended model result. The percentage of the data corresponding to each data type (10% BMR for dichotomous) that yielded winning models are shown.

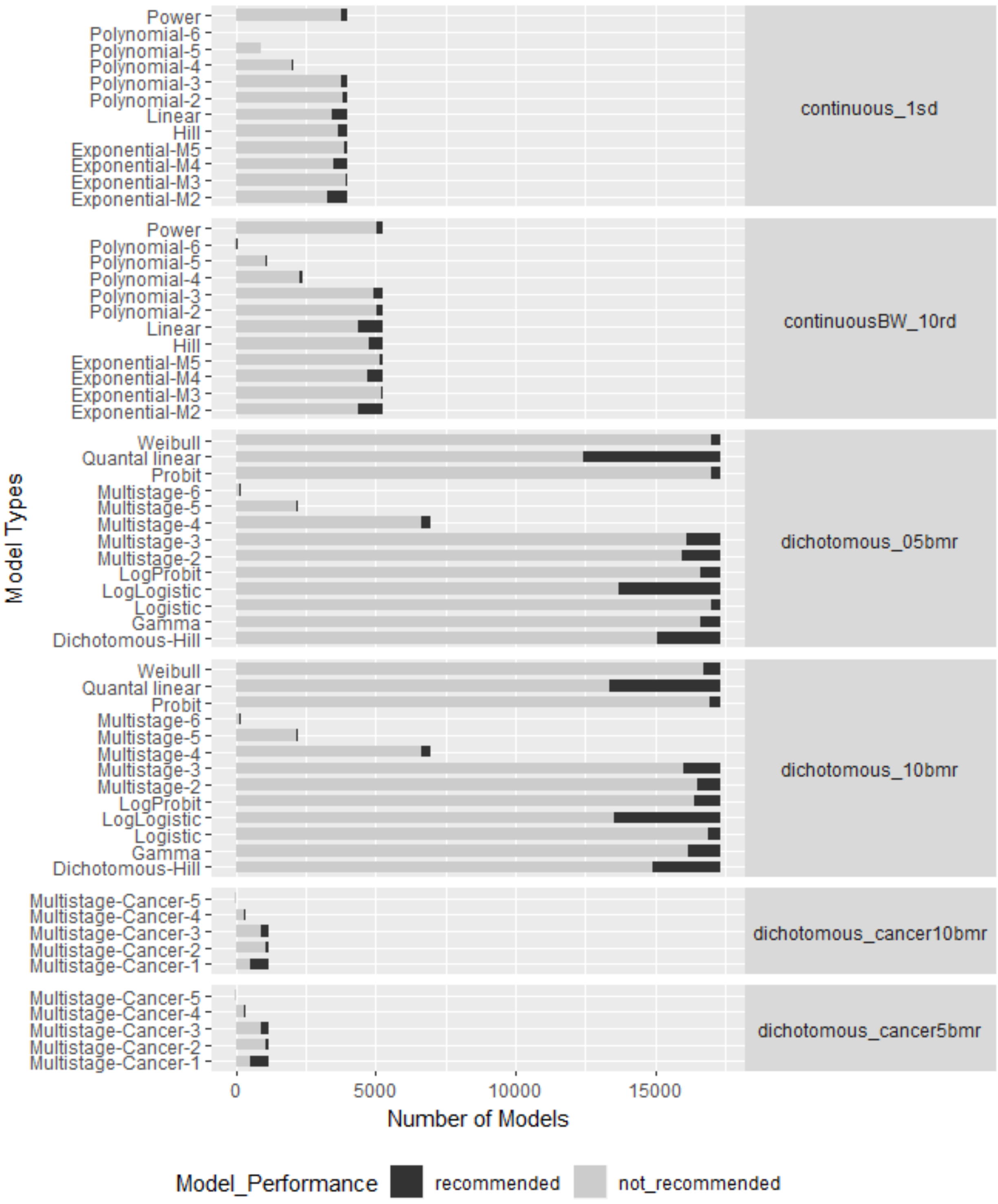


Figure 3.9: The number and type of models for each dataset
The stacked bars for each model indicates the number of models that were or were not recommended. For each dataset type, the BMR was also indicated with their label. A 10 % relative deviation was used as the BMR for the continuous datasets for the body weight and organ weights. All other continuous datasets used a BMR of 1 standard deviation. The dichotomous datasets used a 5 and a 10% BMR.

Data Integration

It is increasingly apparent that many toxicology research questions will require the integration of public data resources, both with those containing the same types of information, as well as with other databases to connect different kinds of information. For example, with increasing global interest in finding rapid chemical screening alternatives like the use of ToxCast to build predictive models, the need for linking *in vitro* effects to outcomes observed *in vivo* is essential(222). To connect the ToxRefDB endpoint and effect terminology with other resources, the ToxRefDB terminology was standardized and cross-referenced to the United Medical Language System (UMLS). UMLS cross-references enable mapping of *in vivo* pathological effects from ToxRefDB to PubMed (via Medical Subject Headings or MeSH terms) that may be relevant for toxicological research and systematic review. This enables linkage to any resource that is also connected to PubMed or indexed with MeSH. For example, Entity MeSH Co-occurrence Network (EMCON)(32), a resource to retrieve ranked lists of genes for a given topic, can be used to identify genes related to adverse effects observed in ToxRefDB. Subsequently, ToxCast can be integrated since the intended targets are mapped to Entrez gene IDs. The result of updating the terminology in ToxRefDB v2 and linking to the UMLS concepts is that ToxRefDB may be used to better anchor or compare to *in vitro* data like ToxCast data, or to other *in vivo* databases of toxicological information, like those available from eChemPortal(223), e-TOX(80), or others. Integration of these data resources is a major hurdle toward to evaluating the reproducibility and biological meaning of both traditional, legacy toxicity information and the data from NAMs.

Conclusions

ToxRefDB has served as a seminal resource for *in vivo* toxicity studies with broad applications in predictive modeling, retrospective analysis, and validation of *in vitro* chemical screening results. Although robust in scope for capturing effect information, early versions of ToxRefDB only contained data for positive findings and thus were limited by a lack of distinction between tested and not tested effects. Additionally, the terminology concerning endpoints did not adhere to a standardized classification system. Moreover, specific effect information and quantitative, dose-response information were needed to support predictive toxicology questions. To address these issues, ToxRefDB has undergone extensive updates that include extraction of additional information (quantitative data as well as observations about tested endpoints), data standardization, and quality assurance measures to maintain data integrity. With these updates, the utility of ToxRefDB can be extended to myriad applications, and our process can serve as a reference for other resources aggregating similar information. The features added in this release of ToxRefDB v2 support ongoing efforts to use these data to train predictive models and also to evaluate the reproducibility and variability in existing animal-based approaches for safety testing used for model training and performance evaluation. The MySQL database and all associated summary flat files are available at ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Animal_Tox_Data/current/. Further documentation, code, and examples are available at https://github.com/USEPA/CompTox-ToxRefDB.

Summary

Toxicity Reference Database (ToxRefDB) underwent significant updates to extract more information from animal toxicity studies, ensure quality, and support interoperability. Of the updates, the most relevant to the overall theme of this dissertation is the establishment of a controlled vocabulary (CV) for reporting of adverse events as well as mapping the CV to Unified Medical Language Systems (UMLS), a semantic network comprised of over 150 biomedical vocabularies. The mappings support interoperability by exposing a point of integration with any resource that also maps to UMLS; e.g., outcomes related to cancer can be rapidly identified and connected to other information related to those outcomes, such as articles in PubMed labeled with relevant Medical Subject Heading terms (MeSH). ToxRefDB can now be included in data integration pipelines to make use of animal toxicity studies in new investigative approaches to understand how chemicals influence complex disease.

CHAPTER 4: EVALUATING THE CANCER- RELATED BIOLOGICAL COVERAGE OF TOXICITY FORECASTER (TOXCAST): IDENTIFICATION OF KNOWLEDGE GAPS AND IMPLICATIONS FOR CHEMICAL SCREENING

Introduction

With increasing global interest in utilizing new approach methodologies (NAMs) (6,7,224) for rapid screening and prioritization of chemicals for safety applications(37), agencies including the US Environmental Protection Agency (USEPA), the US Food and Drug Administration (FDA), the National Intitules of Environmental Health Sciences (NIEHS), Health Canada, the European Chemicals Agency (ECHA), the European Food Safety Authority (EFSA), and many others (225) are actively generating and aggregating data to support the development of computational models to predict adverse outcomes (AOs). Among these efforts are USEPA's Toxicity Forecaster (ToxCast) program (8,11) that has produced *in vitro* toxicity information on over 3,000 chemicals and 400 gene targets. The resulting information from these efforts has been used to develop a number of computational models. However, adoption by regulators has been slow with questions about the relevance of mechanistic data within the context of human-related AOs like cancer both due to lack of clear links between mechanisms and AOs as well as relevant biological coverage. The work in this publication seeks to address the questions on the biological coverage of ToxCast assay information for specific AOs using a computational approach, rather than an expert-driven approach, to better understand the challenges in using the information for chemical safety applications. The result is a demonstration of a new computational approach that is relatable to any AO.

Currently, only the ToxCast estrogen receptor model (23) has been adopted for regulatory applications; However, an androgen receptor model (21) and steroidogenesis model (22) are also being considered for use in screening chemicals as part of the Endocrine Disruptor Screening Program (EDSP). Chemical-mediated endocrine disruption has been well-characterized mechanistically with clear links between exposures and human-relevant AOs (226,227). For example, exposure to diethylstilbestrol (DES), which is a xenoestrogen, has been linked to breast cancer as well a number of other AOs from *in* *utero* exposure (114). However, other human-relevant AOs are complex without clear links between chemical exposures and a resulting phenotype as is the case for many types of cancer (228) and metabolic disorders (229). Also, many environmental chemicals do not have well defined targets (12), so establishing clear links between exposure and outcome can be difficult. Both promiscuous bioactivity of chemicals and lack of concordant responses between *in vitro* and *in vivo* studies is evident in attempts at building computational models for complex AOs like reproductive (173) or developmental (27) toxicity, hepatotoxicity (24), and even cancer (25). From these models, despite being considered successful, questions remain about why some *in vitro* targets would be predictive for the corresponding AO.

Findings from both Cox et al. (2016) (85) and Becker et al. (2017) (89) echo these concerns and propose that a "more rigorous mode-of-action pathway-based framework to organize, evaluate, and integrate mechanistic evidence with animal toxicity"(89) is needed. This approach aligns with the Adverse Outcome Pathway (AOP) framework that can be used to organize relevant biological events across levels of biological organization(70). The typical structure of an AOP begins with an upstream molecular initiating event (MIE), in which a chemical interacts with a biological molecule. An MIE is then connected to a series of downstream key events (KE) by key event relationships (KER) that contains evidence showing a progression from one KE to the next. Finally, the AOP ends with an adverse outcome (AO), which is a phenotype that is of toxicological importance.

The interest in establishing AOPs for regulatory use is to not only organize existing toxicity data but also to serve as a screening and prioritization approach. For example, if evidence links a data-poor chemical (i.e. no animal toxicity data is available) to early KEs within an established AOP, then assumptions can be made about the hazard potential based on the evidence along an already established AOP. The major challenge in developing AOPs is identifying and filling gaps of information. Current approaches in toxicity testing capture information at either end of the AOP, (i.e. MIE to early KE or only the AO). For example, ToxCast provides information about chemical-biological interactions predominantly within the cell. The next steps would be to extrapolate from those bioactivities to tissue level effects like hypertrophy or some other lesion that would be measurable *in vivo*. However, this is difficult since current approaches to capture toxicity information at higher levels of biological organization are low throughput and costly. Assays targeting this level of organization are rapidly under development(49), but still remain medium to low throughput with applications as a validation component in a tiered screening approach rather than data generation as is the case for high throughput approaches. Toxicity data generated at the AO are from animal toxicity studies, which are low throughput, but a massive amount of animal toxicity data exists within publications and regulatory documents. Thus, USEPA's Toxicity Reference Database (ToxRefDB)(13) serves to aggregate animal toxicity data into a digital repository to support development of computational models.

ToxRefDB aggregates animal toxicity information for over 1,000 chemicals that are mostly pesticides. The information is primarily from registrant submitted studies to USEPA's Office of Pesticides Programs (OPP) and has been extracted from the summary reports called data evaluation records (DERs). The information available in ToxRefDB are effects measured *in vivo* that would be considered complex AOs within the context of an AOP.

The previous attempts to link *in vitro* and *in vivo* effects using ToxCast and ToxRefDB respectively have primarily consisted of identifying activities for the same chemical across these two domains(25), yet we understand the limitations of such an analysis: it is unclear if observed concordance is due to the same mechanism(s) operating *in vitro* and *in vivo*, and conversely, it is unclear if lack of observed concordance is due to lack of biological or mechanistic coverage, e.g. aspects of absorption, distribution, metabolism, and excretion, or other aspects of the *in vitro* screening approach. Additionally, there are known factors that contribute to uncertainty in interpretation of the *in vitro* and *in vivo* data and increase the chances for discordant results, including: compensatory mechanisms *in vivo* akin to the "tipping points" investigated for *in vitro* effects(50), high-throughput screening assay technology domains of applicability(1,23,67,230,231), and *in vivo* study-level design features like strain used, dose selection, etc.(178,232). Thus, a biological connection between the *in vitro* and *in vivo* resources is needed to understand if more assays or screening information would improve predictions for complex AOs (86,89).

The complex AO we focus on in this work is cancer. Cancer is a multifactorial adverse outcome where susceptibility is influenced by lifestyle, genetic variants, and chemical exposures. Identification of chemical carcinogens is predominantly based on available human and animal data that associates exposure of a chemical with cancer (233). Human and animal data can be stronger when chemical modes of action (MOA) tie the link between exposure and outcome together as is the case for DNA adduct formation in nasal mucosa from inhaled formaldehyde that leads to increased risk of nasopharyngeal cancer(234). However, associations are not always so clear as exemplified with formaldehyde exposures and links to leukemia. Despite being classified as carcinogenic to humans (Group 1) by International Agency for Research on Cancer (IARC)(228), links between formaldehyde exposure and leukemia are abundant from epidemiological studies(235), but mechanistic links are still being investigated. Currently mechanistic data is being used to more rapidly identify carcinogenicity of a chemical by binning ToxCast assays into Ten Key Characteristics of Carcinogens (TKCC) (86–88) and subsequently evaluating total bioactivity observed in each of those assays for a given chemical. However, this use of chemical-centric data serves to identify carcinogens with more direct links between exposure and outcome and not those with an unknown number of chained events that culminate in a measurable *in vivo* effect(5,236). With the example of formaldehyde exposures and leukemia, if critical events leading to leukemia can be identified, indirect links between the known formaldehyde mechanisms and leukemia could be made. Here we present a data integration approach to evaluate the biological coverage of ToxCast for cancer by identifying putative upstream KEs from cancer thus bridging the gap between *in vitro* and *in vivo* data. In evaluating the biological coverage of ToxCast for cancer, we can further identify the potential utility of this data in further building predictive models.

The work presented here extends previous work from Watford et al. (32) and Grashow et al. (90) where a putative gene set for breast cancer was retrieved from a resource called Entity MeSH Co-occurrence Network (EMCON). EMCON is built using biomedical literature and contains links between Entrez gene identifiers (GeneIDs)(102) and Medical Subject Headings (MeSH terms) (131). MeSH terms are keywords used to index articles within PubMed, the largest publicly available resource of biomedical literature. The genes retrieved from searching EMCON serve as a point of integration to link *in vitro* targets either directly to a corresponding search concept or indirectly through other concepts like pathways. We construct two search strategies to build queries to retrieve gene sets from EMCON. The first strategy is similar to that described in Watford et al. (32) where cancer concepts (CCs) like evading apoptosis were identified from seminal publications (88,118,147,148). A second search strategy is used in this approach that directly utilizes the cross-referenced observed toxicity effects from ToxRefDB. The observed toxicity effects were cross-referenced to Unified Medical Language System (UMLS), which is a semantic network of over 150 biomedical terminologies(78). MeSH is one of those terminologies thus we can identify MeSH terms for each ToxRefDB observed toxicity effect. We identify the cancer-related effects and corresponding MeSH terms from ToxRefDB stratified by organ (i.e. liver cancer effects) that serve as queries for EMCON. The resulting gene sets are validated using an online tool called Enrichr(95) that performs gene set enrichment analysis (GSEA) (91). GSEA identifies topics from reference libraries that have significantly overlapping gene sets. Four reference libraries were chosen to validate the cancer-related gene sets retrieved in this work: Gene Ontology (GO) biological process(237,238), GO molecular function(237,238), Reactome(138,239), and Kyoto Encyclopedia of Genes and Genomes (KEGG)(105–107). With gene sets retrieved and validated for CCs, we can elucidate ToxCast coverage across CCs as well as highlight gaps in the coverage by quantifying the gene overlap (direct association) as well the topic overlap (indirect association) from the four reference libraries.

Our analysis aims to estimate the biological coverage of ToxCast for cancer, addressing limitations in current knowledge about indirect links between chemical mechanisms and *in vivo* adverse outcomes. Some of the key conclusions of this research include a greater understanding of why ToxCast may currently be inadequate for unsupervised prediction of cancer on the basis of biological coverage, and the identification of gene sets for specific cancer-related effects observed in a rich database that may have forward utility in feature selection for models that use transcriptomic data to predict cancer. Ultimately, this work suggests that a combination of unsupervised and expert approaches may be needed to use the current data resource for prediction of possible cancer-related outcomes.

Methods

Overview of approach

In this work, gene sets linked to cancer concepts (CCs) are retrieved from Entity MeSH Co-occurrence Network (EMCON) through two search strategies (Figure 4.1A): one using an expert driven approach to identify CCs and a second using cancer-related observed effects from ToxRefDB. The resulting gene sets are validated (Figure 4.1B) and subsequently used to estimate the biological coverage of ToxCast for cancer (Figure 4.1C).

For search strategy #1 CCs are identified by experts along with corresponding seed MeSH terms. This approach is similar to that described in Watford et al(32) for identification of a breast cancer-related gene set, but instead of using the MeSH tree for expanding the MeSH term selection, a MeSH co-occurrence network is used. A MeSH co-occurrence is any two MeSH terms that have been tagged to the same article(167,240). Figure 4.2 shows the workflow for building the MeSH co-occurrence network and is further described in the following section. The seed MeSH terms are used to query the MeSH co-occurrence network and retrieve all other relevant MeSH terms for a CC. For search strategy #2, cancer-related observed effects were identified from Toxicity Reference Database (ToxRefDB). The effects in ToxRefDB are cross-referenced to Unified Medical Language System (UMLS) concepts from which relevant MeSH terms can also be identified(13). This initial MeSH term selection was not expanded in this strategy because the effects were grouped together at the organ level. For example, effects like “hepatocelluar carcinoma” and “hepatocellular adenoma” are grouped together under “liver cancer”. Next, for both search strategies, the full set of MeSH terms are used to query EMCON and retrieve a single gene set per concept (e.g. “Angiogenesis” or “Liver cancer”). Each search strategy applies a separate set of filters (further described in following sections) to identify only the relevant genes. Next the gene sets were validated (Figure 4.1B) by manually assessing the relevance of enriched topics for a corresponding CC using a gene set enrichment tool called Enrichr (95). Finally, the cancer-related biological coverage of ToxCast was evaluated by quantifying the ToxCast assay endpoint gene target overlap with five different datasets (Figure 4.1C): CC gene sets, Reactome pathways, KEGG pathways, GO biological processes, and GO molecular functions.

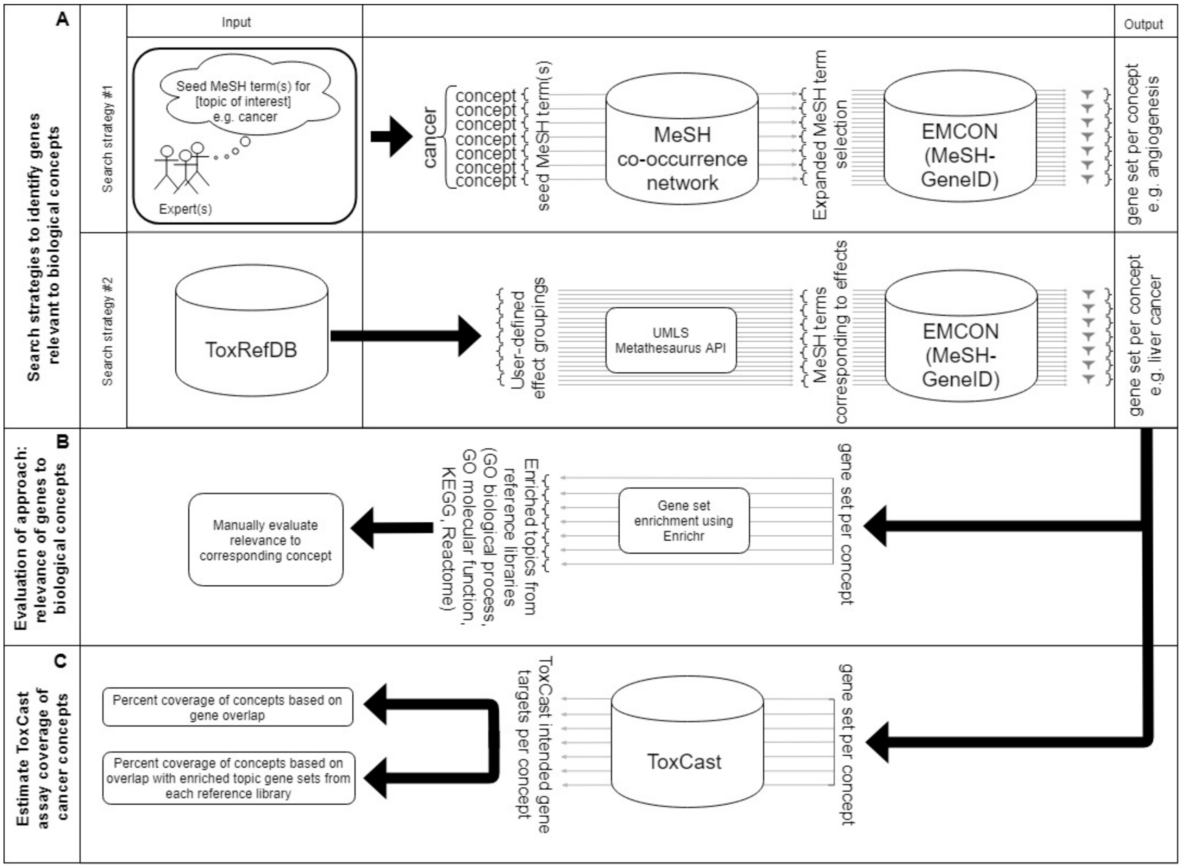

A
Search strategies to identify genes relevant to biological concepts
Input
Output
Search strategy #1
Seed MeSH term(s) for [topic of interest] e.g. cancer
Expert(s)
cancer
concept
concept
concept
concept
concept
concept
concept
seed MeSH term(s)
MeSH co-occurrence network
Expanded MeSH term selection
EMCON (MeSH-GeneID)
gene set per concept e.g. angiogenesis
Search strategy #2
ToxRefDB
User-defined effect groupings
UMLS Metathesaurus API
MeSH terms corresponding to effects
EMCON (MeSH-GeneID)
gene set per concept e.g. liver cancer
B
Evaluation of approach: relevance of genes to biological concepts
gene set per concept
Gene set enrichment using Enrichr
Enriched topics from reference libraries (GO biological process, GO molecular function, KEGG, Reactome)
Manually evaluate relevance to corresponding concept
C
Estimate ToxCast assay coverage of cancer concepts
gene set per concept
ToxCast
ToxCast intended gene targets per concept
Percent coverage of concepts based on gene overlap
Percent coverage of concepts based on overlap with enriched topic gene sets from each reference library

Figure 4.1: Overall approach for identifying cancer-related gene sets to estimate biological coverage of ToxCast for cancer
(A) First, two search strategies are used to identify cancer concepts (CCs) used to query EMCON. Search strategy #1 employs an expert-driven approach to identify seed MeSH terms that correspond to CCs and then undergo MeSH term expansion using a MeSH co-occurrence network. Search strategy #2 uses the organ-level, cancer-related observed effects identified from ToxRefDB where corresponding MeSH terms were retrieved from the UMLS metathesaurus. The queries from both search strategies comprise MeSH terms that are used to retrieve a gene set for each CC from EMCON. (B) Next, the resulting gene sets are validated using an online tool called Enrichr that was used to perform GSEA with four reference libraries. (C) Finally, ToxCast biological coverage was quantified as percent overlap with the CC-derived gene sets as well as with the enriched topics for each CC from Enrichr.

Building the MeSH co-occurrence network

A previous limitation of identifying gene sets using EMCON was expanding the MeSH term selection using the MeSH tree (32). Although the MeSH tree implies a relationship between parent and child MeSH terms, that relationship is not necessarily reflected in the article tagging. For example, the MeSH term “Diabetes Mellitus, Type 2” is closely related to the MeSH term “Insulin resistance” because insulin resistance is a key phenotype in type 2 diabetes (241), yet many “Insulin resistance” and “Diabetes Mellitus Type 2” do not belong to any of the same branches within the MeSH hierarchical tree. To account for this limitation, a MeSH co-occurrence network was built and used to expand the MeSH term selection. A MeSH co-occurrence is any two MeSH terms that are tagged to the same article (167,242). The MeSH co-occurrence network was built using the 2017 Medline Baseline Repository (MBR) raw data files, specifically the file named “MH_items” (243). This file provides a snapshot of MeSH term to PMID mappings (Figure 4.2) from 2016 and is updated annually. This distribution is used because it provides only the identifiers we need to build the MeSH co-occurrence network. Next, MeSH term frequencies were normalized according to the MeSH tree by tagging all ancestors to articles that are tagged with their descendants (Figure 4.2). This ensures that the tagging frequency increases while traversing the MeSH tree towards a root MeSH term and that broader MeSH terms have higher frequencies than their more specific descendants. This normalization was previously described in Watford et al. (32). Next, each co-occurrence is identified by taking every two ~~combinations of~~ MeSH terms for an article and keeping only the MeSH co-occurrences that have more than 2 articles to support the association. Finally, the MeSH co-occurrences are ranked using normalized pointwise mutual information (NPMI), which is the normalized variant of pointwise mutual information that ranks association on a -1 to 1 scale(143). NPMI values of zero or less indicate an association that may have occurred by chance, and NPMI values of greater than zero indicate increasing strength of the association as NPMI approaches 1. The MeSH co-occurrence network is filtered to include only MeSH co-occurrences with NPMI > 0.

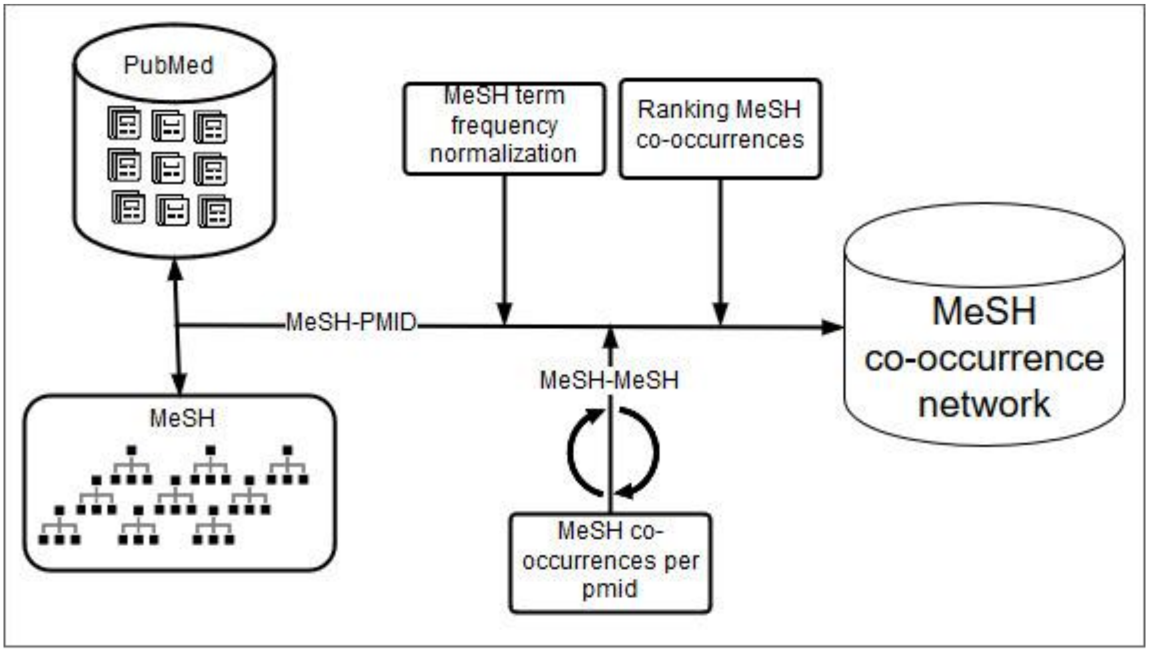

PubMed
MeSH term frequency normalization
Ranking MeSH co-occurrences
MeSH-PMID
MeSH co-occurrence network
MeSH-MeSH
MeSH
MeSH co-occurrences per pmid

Figure 4.2: Building the MeSH co-occurrence network
The MeSH co-occurrence network was built by first normalizing the MeSH term frequency so that the frequency of each MeSH term is proportional to the number of descendants the MeSH term has according to the MeSH term tree. Next MeSH co-occurrences were identified. A MeSH term co-occurrence is any two MeSH terms that have been annotated to the same article. Finally, MeSH term co-occurrences were ranked using normalized pointwise mutual information (NPMI). Any MeSH term co-occurrences with NPMI < 0 were excluded from the final network.

Selecting carcinogenesis concepts and corresponding MeSH terms

Cancer concepts (CCs) and the corresponding MeSH terms were identified from previous work to identify a breast cancer gene set(90). CCs are defined as critical processes that characterize the pathogenesis of cancer and were identified from seminal publications that investigate cancer (88,118,147,148). For this work, the topic of mammary tissue was excluded as the CCs of interest are not limited to specific breast cancer-related mechanisms. MeSH terms are needed to retrieve gene sets from EMCON, so corresponding MeSH terms were selected that best represent each CC (Table 4.1). These MeSH terms alone do not fully represent the CC, but rather serves as a seed MeSH term to retrieve all closely-associated MeSH terms. The MeSH co-occurrence network was queried with the seed MeSH terms to retrieve a comprehensive set of MeSH terms that characterize the CC. Finally, only MeSH terms with greater than the 95th percentile of the NPMI distribution were kept.

Carcinogenesis Concept	MeSH term (DUI, i.e. unique identifier)
Angiogenesis	Neovascularization, Pathologic (D009389)
	Neovascularization, Physiologic (D018919)
Apoptosis	Apoptosis (D017209)
Cell cycle	Cell cycle (D002453)
Epigenetics	Epigenomics (D057890)
Genotoxicity	DNA damage (D004249)
	DNA repair (D004260)
Growth hormones	Growth hormone (D013006)
Immortalization	Cell survival (D002470)
Immunomodulation	Immune system (D007107)
Inflammation	Inflammation (D007249)
Oxidative stress	Oxidative stress (D018384)
Proliferation	Cell proliferation (D049109)
Steroid hormones	Gonadal steroid hormones (D012739)
Xenobiotic metabolism	Xenobiotics (D015262)

Table 4.1: Carcinogenesis concepts and the corresponding MeSH term(s)

ToxRefDB cancer-related effects and corresponding MeSH terms

Toxicity Reference Database (ToxRefDB) is a digital resource of results from animal toxicity studies [25]. ToxRefDB reporting of observations from *in vivo* studies is structured as a hierarchy, with endpoints as a parent category for effects. For example, the endpoint “systemic pathology microscopic liver” is the parent for the effect “liver hyperplasia.” Each endpoint and effect has been cross-referenced with Unified Medical Language System (UMLS), which is a semantic network of over 150 biomedical vocabularies(78). One of the vocabularies maintained in the semantic network is the MeSH vocabulary. The MeSH terms for the subset of cancer-related effects and corresponding endpoints were retrieved using the UMLS representational state transfer (REST) application programming interface (API)(244). Each effect can link to multiple UMLS concept codes, and multiple effects may link to a single endpoint, so each endpoint has multiple MeSH terms mapped to it. For example, MeSH terms “Liver”, “Adenoma”, “Carcinoma, Hepatocellular”, and others are mapped to the endpoint “systemic microscopic pathology liver”. These MeSH terms are used to retrieve gene sets relevant to the endpoint. The cancer-related effects and endpoints are available in APPENDIX 2.

Gene set retrieval

To identify putative gene sets for CCs, Entity MeSH Co-occurrence Network (EMCON) was queried with the expanded MeSH term selections that were derived from the methods described above. EMCON is a co-occurrence network of GeneID-MeSH associations derived from curated biomedical literature resources and ranked using NPMI(143). Like the MeSH co-occurrence network, EMCON only retains co-occurrences with NPMI > 0. The MeSH term selections were used to query EMCON and retrieve associated genes where only the GeneID-MeSH co-occurrences within the 95th percentile of the NPMI distribution for the corresponding MeSH term set were kept.

Validation of gene sets

To verify that the gene sets for each CC were, in fact, returning relevant genes, gene set enrichment analysis (GSEA)(91) was performed to retrieve enriched topics from four different sources called reference libraries: Gene Ontology (GO) biological process, GO molecular function, KEGG, and Reactome. GSEA is commonly used to analyze genomic and transcriptomic data to interpret the findings of the experimental results. Reference libraries are typically built manually by reviewing literature and constructing hierarchical relationships between extracted topics. The four reference libraries that were chosen to validate the CC gene sets are widely used and are not confined to a specific topic (i.e. confined to a cell line like MCF7). GSEA was performed using Enrichr, which is an online tool that performs enrichment analysis for a submitted gene set across over 130 reference libraries. Enrichr produces four values that indicates enrichment of a topic for a given gene set and reference library. A p-value is given as a result of either Fisher's Exact or hypergeometric mean. An adjusted p-value is also calculated that corrects for multiple hypothesis testing. The z-score is the distance away the observed rank is from the expected rank. The converted score is the product of the log of the p-value and z-score. In this work, a topic was enriched if the adjusted p-value is < 0.05. Each CC gene set was submitted to Enrichr through the RESTful API(245). Finally, the relevance of the top five enriched topics to the corresponding CC was manually determined by reviewing literature associated with the enriched topic and either cancer or the CC. Overall precision and CC-specific precision was calculated using the relevance determinations.

Linking ToxCast assay endpoint targets to cancer concepts

ToxCast assay endpoints are uniquely identified by an assay endpoint identifier (aeid)(16) and cover a wide range of biological concepts with genes identified as the intended targets for most assays. Assays that measure general cytotoxicity or a complex process (i.e. proliferation) are not necessarily annotated with an intended gene target. ToxCast assays and corresponding genes were downloaded from(246). The intended gene targets are used as a point of integration with each CC via the gene set retrieved from EMCON. ToxCast can be linked to a CC either by the gene set retrieved from EMCON or by the gene set of an enriched topic from Enrichr. The gene sets from each reference library are made available for download from Enrichr(247).

Evaluating cancer-related biological coverage of ToxCast

With gene sets and enriched topics identified for each CC, the biological coverage of ToxCast can be evaluated in two approaches (Figure 4.1C): (1) direct coverage via gene overlap with EMCON gene sets and (2) indirect coverage via gene overlap with enriched topics. Indirect coverage was included because even though ToxCast may not have a target within the retrieved gene set to assess direct coverage, ToxCast genes may play a role in the enriched topics that include pathways and biological processes related to the CC, which implies the assay endpoint may still be relevant to a CC. Direct coverage was quantified as the percent overlap with the gene set for each CC. Indirect coverage was quantified as the percent overlap with the enriched topics for each CC. Because reference libraries arrange topics hierarchically, gene sets are either supersets or subsets of each other. For example, the pathway “Estrogen-dependent gene expression”(248) has the parent “ESR mediated signaling”(249), so the gene set for “Estrogen-dependent gene expression” is a superset of the “ESR mediated signaling” gene set because each gene of the child’s gene set is present in the parent’s gene set. The consequence of this pattern in quantifying coverage is bias for enriched topics from a single branch because each member of the branch would be mapped to the same CC and ToxCast assay endpoint gene targets resulting in duplicative associations. To avoid this bias, enriched topics were only kept if the gene set was not a superset of any other gene set within their respective reference library. Finally, ToxCast is only considered to overlap with an enriched topic if at least five ToxCast intended gene targets overlap with the enriched topic's gene set. This threshold is called the membership cutoff. The reasoning for the cutoff is addressed in the results.

Results

MeSH co-occurrence network

The MeSH co-occurrence network has over 76.5 million unique co-occurrences with 61% having more than two articles supporting the co-occurrence and 26% having an NPMI > 0.0. When applying both filters, the network is reduced by 83% with 13 million unique MeSH co-occurrences remaining.

Cancer-related gene sets

For search strategy #1, 13 CCs were identified along with corresponding seed MeSH terms (Table 4.2). For search strategy #2, 34 organ-level endpoints were identified to have cancer-related effects. However, many of the endpoints had extremely poor coverage with ToxCast. Liver had the most coverage and is the only cancer-related endpoint from ToxRefDB in the figures and tables. A summary of the remaining 33 organ-level and cancer-related gene sets along with the ToxCast overlap is available in APPENDIX 3.

The number of MeSH terms retrieved from the MeSH co-occurrence network for each CC varied widely with a minimum of 2 for "Immunomodulation" to a maximum of 70 for "Oxidative Stress" (Table 4.2). The number of genes returned for each CC also varied and is not correlated with the number of MeSH terms used for the query (rank correlation-coefficient 0.372).

The NPMI distributions for both the MeSH-MeSH and GeneID-MeSH co-occurrences for all CCs are shaped differently. Figure 4.3 shows these distributions for the "Immunomodulation" (Figure 4.3A) and "Angiogenesis" (Figure 4.3B) CCs. Only 37 genes are retrieved for the "Immunomodulation" gene set (Figure 4.3A) while 313 genes are retrieved for the "Angiogenesis" gene set. Both MeSH-MeSH and GeneID-MeSH NPMI distributions for "Angiogenesis" are shifted towards zero which means many of the co-occurrences for "Angiogenesis" have less supporting (either as number of articles or promiscuously co-occurring with other entities) for the associations.

Search strategy	Cancer concept	Number of associated MeSH terms	Number of associated genes
#1	Angiogenesis	30	313
	Apoptosis	29	2025
	Cell cycle	7	711
	Epigenetics	11	221
	Genotoxicity	37	907
	Growth hormones	39	195
	Immortalization	7	825
	Immunomodulation	2	37
	Inflammation	20	467
	Oxidative stress	70	2511
	Proliferation	61	3623
	Steroid hormones	41	212
	Xenobiotic metabolism	35	336
#2	Liver cancer (ToxRefDB)	16	478

Table 4.2: The number of MeSH terms used to query EMCON and the total number of genes retrieved from EMCON for the cancer concept

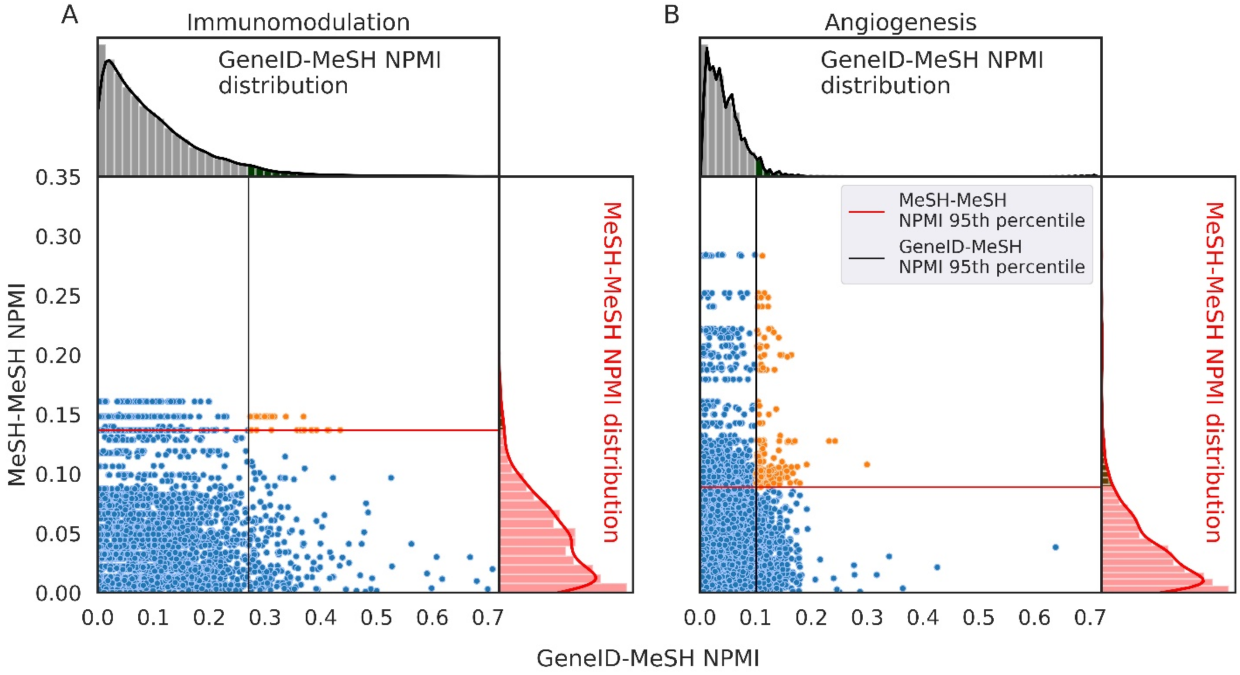


Figure 4.3: Filtering gene sets for cancer concepts “Immunomodulation” and “Angiogenesis”
The GeneID-MeSH NPMI distributions and 95th percentile cutoffs are shown in black, while the MeSH-MeSH NPMI distributions and 95th percentile cutoffs are shown in red. Each point represents two MeSH terms and a GeneID, along with the corresponding NPMI values. The points in blue represent associations that are filtered out because they fall below the 95th percentile cutoffs, while the points in orange represent associations that are used to identify relevant GeneIDs for a final gene set. (A) For the CC “Immunomodulation”, only 37 GeneIDs are above the cutoffs. (B) For the CC “Angiogenesis”, 313 GeneIDs are above the cutoffs.

Validating gene sets

The gene sets were validated by calculating precision for the top five enriched topics from four reference libraries: GO biological process, GO Molecular Function, KEGG, and Reactome. The enriched topics from each reference library were identified using Enrichr's REST API(245). The relevance of an enriched concept to a CC was determined by reviewing abstracts retrieved from a PubMed or related search. For example, for the CC "Epigenetics" and enriched concept from GO biological process "chromatin assembly", PubMed was searched with the query "epigenetics and chromatin assembly" which returns nearly 800 articles. Many of these articles not only link the CC and enriched concept, but also link both topics to cancer (250,251). The overall precision score is 0.975. "Growth hormones" had the lowest precision of any CC at 0.9 (Table 4.3). APPENDIX 4 provides the relevance decisions made for each enriched topic.

	GO Biological Process	GO Molecular Function	KEGG	Reactome	Average Precision
Proliferation	1.0	1.0	1.0	1.0	1.00
Angiogenesis	1.0	1.0	1.0	1.0	1.00
Oxidative Stress	1.0	1.0	1.0	1.0	1.00
Apoptosis	1.0	1.0	1.0	1.0	1.00
Genotoxicity	1.0	1.0	1.0	0.8	0.95
Immortalization	1.0	1.0	0.8	1.0	0.95
Cell Cycle	1.0	1.0	1.0	1.0	1.00
Xenobiotic Metabolism	1.0	1.0	1.0	1.0	1.00
Inflammation	1.0	0.8	1.0	1.0	0.95
Epigenetics	1.0	1.0	1.0	1.0	1.00
Steroid Hormones	1.0	0.8	1.0	1.0	0.95
Growth Hormones	1.0	1.0	0.8	0.8	0.90
Immunomodulation	1.0	1.0	1.0	1.0	1.00
Liver Cancer (ToxRefDB)	1.0	1.0	1.0	1.0	1.00
Overall precision					0.975

Table 4.3: Precision of enriched topics retrieved from each reference library for each CC

Estimating ToxCast biological coverage of cancer

A gene set was retrieved from EMCON for each CC and validated using Enrichr to retrieve a set of enriched topics from four reference libraries. ToxCast biological coverage is quantified as percent overlap with the gene set for each CC and the percent enriched topics with overlapping genes greater than the membership cutoff. The membership cutoff of four is approximately the 95th percentile for each reference library as indicated by the black dashed line in Figure 4.4A. With a membership cutoff of 0 (at least 1 gene is present in the enriched topic’s gene set), the average percent coverage from the four reference libraries would be 65% with KEGG having the highest percent coverage of 82%.

As shown in Figure 4.4B, the gene set overlap alone is low with a mean coverage of 6%, but when expanded for enriched topics and at a membership cutoff of 4, the coverage increases to an average of 32%. No ToxCast intended gene targets overlap with the “Epigenetics” gene set, but they overlap with enriched topics from all four reference libraries. The genes in the “Epigenetics” gene set account for protein-coding genes for histone modifications and DNA methylation, which were not considered in selection of ToxCast targets. ToxCast targets were selected by both availability of assay technologies as well as targeting endocrine disruption and metabolic activity. ToxCast has the highest coverage from KEGG for each CC; however, this may be misleading because KEGG has the lowest number of topics at 278 as well as the largest gene sets per topic at an average of 80.6 (Table 4.4).

For this approach, no other aspects are considered like assay technology or cell line, which means the average cancer-related biological coverage with a membership cutoff of 4 is 32% is the maximum coverage that will only decrease as other parameters are considered. Because of this, the biological coverage is considered low and insufficient for the CCs identified which also includes all organ-level effects from ToxRefDB. The enriched topics with the most overlapping genes from each reference library are in Table 4.5. The enriched topic with the highest percent coverage is “Nuclear Receptor transcription pathway (R-HSA-383280)” from Reactome with 47 ToxCast genes present in the gene set of 51 genes.

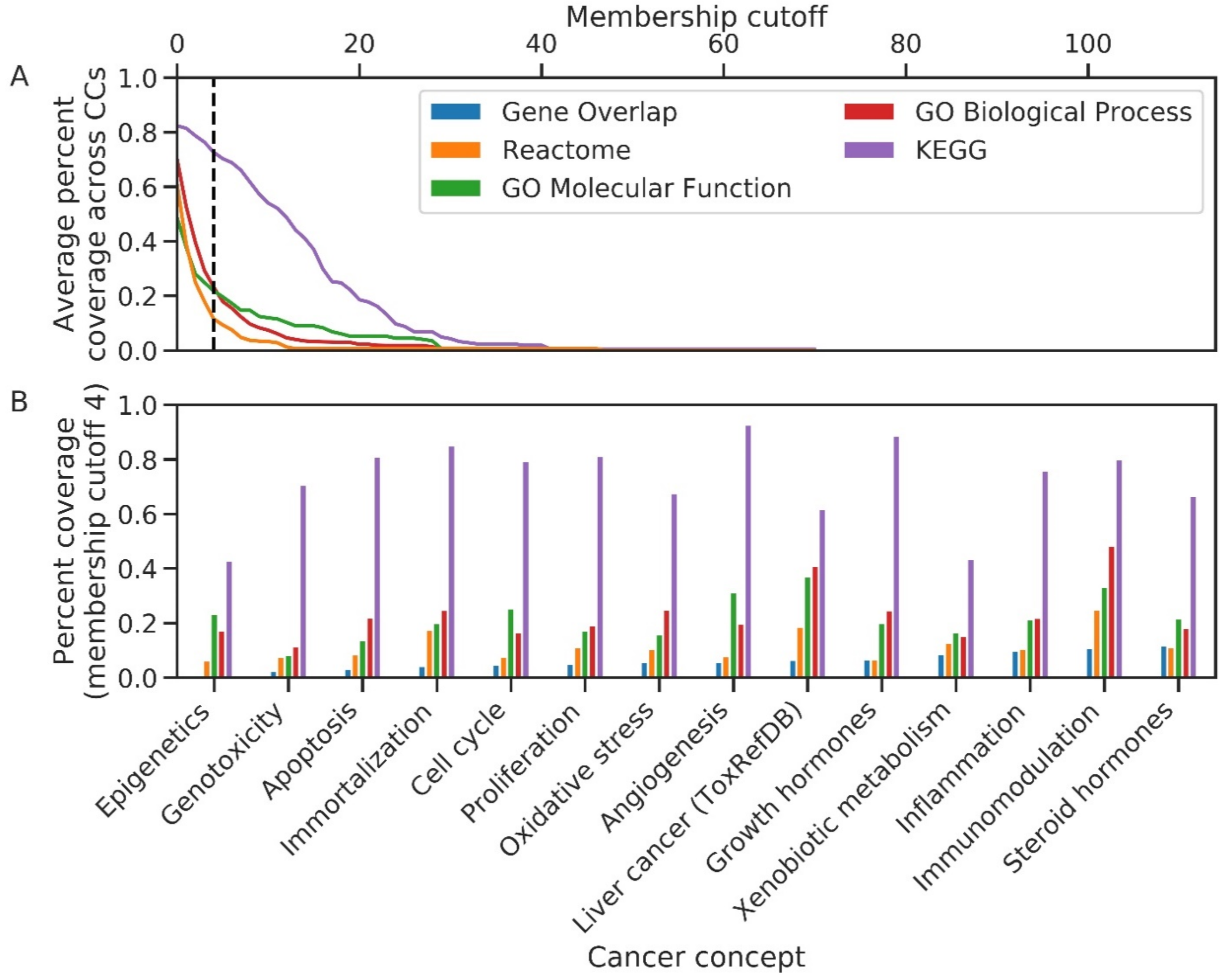


Figure 4.4: The ToxCast coverage for cancer-related gene sets and enriched topics from four gene set reference libraries

(A) Shown is the average percent coverage across all CCs for each reference library as the membership cutoff increases. The black dashed line is the membership cutoff of 4 genes and is approximately the 95th percentile for each reference library. (B) Also, shown in this figure is the percent of the ToxCast intended gene targets that overlap with the gene sets per CC and the percent of enriched topics identified using Enrichr from four reference libraries: GO biological process, GO molecular function, KEGG, and Reactome.

Reference library	Number of topics	Mean gene set length per topic
GO biological process	3076	23.3
GO molecular function	697	23.5
Reactome	699	23.7
KEGG	278	80.6

Table 4.4: Summary statistics for reference libraries

Reference library	Enriched topic (identifier)	ToxCast genes present in gene set	Total number of genes in gene set
KEGG	Neuroactive ligand-receptor interaction (hsa04080)	71	277
GO Biological Process	positive regulation of nucleic acid-templated transcription (GO:1903508)	51	503
GO Molecular Function	RNA polymerase II core promoter proximal region sequence-specific DNA binding (GO:0000978)	35	263
Reactome	Nuclear Receptor transcription pathway (R-HSA-383280)	47	51

Table 4.5: Enriched topics with the highest number of ToxCast genes present in gene set

Discussion

In this work, we presented novel methods 1) to identify gene sets linked to cancer concepts (CCs) and specifically to cancer-related animal toxicity endpoints from ToxRefDB and 2) to estimate the cancer-related biological coverage of ToxCast. The goal was to understand if unsupervised approaches could be used to relate ToxCast information to prediction of cancer. Two search strategies were implemented to identify cancer-related gene sets. The first approach, search strategy #1, is an expert driven approach where seed MeSH terms were selected that best represent a concept of interest (e.g. evading apoptosis). Next, the seed MeSH terms were used to query the MeSH co-occurrence network to retrieve closely associated MeSH terms that serve as the full query for EMCON and retrieve gene sets relevant to the CCs. The second approach, search strategy #2, uses the UMLS cross-references to cancer-related animal toxicity endpoints from ToxRefDB to identify relevant MeSH terms to query EMCON and retrieve corresponding gene sets. Gene set enrichment analysis (GSEA) was performed on each gene set using Enrichr to identify enriched topics from four reference libraries: GO biological process, GO molecular function, KEGG, and Reactome. Precision was calculated by manually determining the relevance of the top five enriched topics from each reference library and CC. The precision was 0.975 with only seven of 280 associations not directly relevant. Finally, the biological coverage of ToxCast for each CC was estimated directly by ToxCast intended gene target overlap with each CC gene set and indirectly by percent enriched topics with at least five ToxCast intended gene targets overlapping with each enriched topic’s gene set. The average direct coverage across each CC is only 6% while the average indirect coverage across each CC is 32%, which is considered a maximum coverage that will decrease when considering other parameters like membership cutoff. Because of these dependencies, the cancer-related biological coverage of ToxCast is consider poor, implying the use of ToxCast for unsupervised prediction of cancer is limited. However, the unsupervised approaches here to identify ToxCast data that may be relevant to prediction of cancer could be reviewed by experts in support of weight-of-evidence analysis for carcinogenesis.

Co-occurrence networks are used to represent knowledge in datasets whether it's from social media (252), language (253), or biomedical concepts (42,167,254) that can subsequently be used to support information retrieval, analytics, and predictive modeling like link prediction for knowledge discovery, i.e. identify future associations. In this work, two co-occurrence networks, MeSH co-occurrences and MeSH-GeneID co-occurrences (EMCON), were used to retrieve cancer-related gene sets. Related efforts have been previously used to identify gene sets from co-occurrence networks like that from Gene2MeSH (164) and MeSH Overrepresentation Profiles (MeSHOPs) (165), but combining this information with MeSH co-occurrences has not been done before. This approach yields a high precision of 0.975 of the enriched topics from GSEA indicating that relevant genes to each CC, including the cancer-related ToxRefDB observed effects, are retrieved. The success of this approach supports the notion that our current methodologies can be improved upon with integrating more data for richer datasets, yet complexity in managing and analyzing this information increases exponentially (42,51). One of the major challenges in this approach was properly managing numerous external resources (i.e. PubMed and resources supporting EMCON) in a single infrastructure to support the analysis. Further work could incorporate many other entities (e.g. pathways, chemicals, proteins, and experimental results), but, without both a standard data model (255–257) and a modern technology stack to support the increased complexity, this approach becomes chaotic. Also, EMCON maintains the limitations previously described in Watford et al. (2018) (32): EMCON is built using only manually curated biomedical resources. EMCON contains information from nearly 700K PubMed articles, but much more information about genes remains in articles that have yet to be curated.

The number of associated MeSH terms retrieved from the MeSH co-occurrence network was not correlated with the number of genes retrieved from EMCON as indicated by the rank correlation coefficient of 0.372. Other factors that may contribute to the number of MeSH terms or genes retrieved is the number of descendants a MeSH term has according to the MeSH hierarchical tree. For example, the CC with the fewest number of genes is “Immunomodulation” with only 37 genes. The seed term used to represent “Immunomodulation” is “Immune System”, which has 69 descendants. Immune system function is a broad topic that plays a role is nearly all aspects of biology and is related to many of the other CCs used in this approach like “Oxidative stress”, which has only 2 descendants, and “Apoptosis”, which has only 3 descendants. Though this approach improved upon the original expert-driven search strategy from initial implementation of EMCON for breast cancer research (32,90), limitations remain in selection of the MeSH terms that are most relevant.

Gene set enrichment analysis (GSEA) was performed by Enrichr to validate that the cancer-related genes, were, in fact, being identified from EMCON. GSEA relies on reference libraries of gene sets that are specific to a topic. Of the four reference libraries used in this approach, only two, Reactome and KEGG, arrange the gene sets according to the relationships between genes or provide pathway diagrams. Both GO reference libraries are a collection of annotations from GO consortium participants. Each reference library varied by the number of topics as well as the average gene set length per topic. KEGG varied the most with the lowest number of lowest-level (determined by gene set subsets) topics at 278 and the highest average gene set length at 80.6. Reference library selection is crucial in interpreting GSEA results (258) as exemplified in the contrasting results in Figure 4.4B. Without comparison to other reference libraries, we might have concluded that ToxCast cancer-related biological coverage is sufficient for building predictive models in cancer from KEGG alone. However, when identifying the KEGG enriched topic with the most overlapping ToxCast intended gene targets (Neuroactive ligand-receptor interaction), we find that it is a non-specific topic of 277 genes that wouldn’t necessarily be useful in interpreting results from a predictive model. Both GO reference libraries show similar results. In contrast, the Reactome enriched topic with the most overlapping ToxCast intended gene targets (Nuclear Receptor transcription pathway) is well characterized with only 51 genes. However, this includes estrogen receptor and androgen receptor activity, which is already the primary use case in building predictive models using ToxCast (21,23). These models comprise multiple assays but only account for two genes each (ER model (23): ESR1 and ESR2, AR model (21): AR and SRC). All genes except for SRC are present in Reactome’s “Nuclear Receptor transcription pathway”, although, SRC-mediated interactions are mentioned in the description of a reaction in the pathway (259). Because ToxCast intended gene targets cover many other genes involved with the “Nuclear Receptor transcription pathway”, a more comprehensive computational network model combining both existing ER and AR models could be developed to include more reactions that are downstream KEs for endocrine disruption AOs. Both KEGG and Reactome would be great sources to look further into for applicability of building computational network models similar to the AR and ER models because the genes are arranged in a pathway with evidence linking each reaction together similar to linking KEs. This would require manual review of the enriched topics and expert knowledge on the ToxCast assay technologies to determine relevance, and each enriched topic has been linked to cancer with supporting evidence from EMCON. This approach can be expanded to include other topics besides cancer.

Despite possible applications to further utilize ToxCast for building predictive models for cancer mentioned above, the overall cancer-related biological coverage of ToxCast is considered poor. This conclusion was made primarily on the average percent coverage across each CC and reference libraries, which was 32% despite having higher coverage when lowering the membership cutoff (Figure 4.4A). Cancer, along with the cancer concepts selected in this work, is a broad topic that involves complex biological interactions across nearly all known domains of biology. The ToxCast assays comprise chemical bioactivity information on over 350 gene targets, but account for numerous cell types, species, and technologies limiting the applicability to few domains. Ongoing efforts to generate a dose-response high throughput transcriptomics (HTTr) information(33,128,129) will eliminate some parameters because the technology will be consistent as well eliminate questions on relevant biological coverage by evaluating the whole genome. However, questions still remain on how to analyze dose-response transcriptomic data linking the changes in gene expression to adverse outcomes (i.e. linking genotype to phenotype) (112).

The approach in this work has myriad applications in computational toxicology including information retrieval and organization for systematic review(260), putative linkage of key events (KEs) to fill knowledge gaps in existing Adverse Outcome Pathways(261), and analysis of high throughput transcriptomics (HTTr) or gene expression profiling results(33,128,129). Identifying gene sets specific to a concept allows for linking other topics like pathways to that concept as well. This provides a way to organize available information to aid in manual review process like that of annotating ToxCast assays or identifying linked information for a systematic review. Traditional pathway analysis of gene expression results relies on GSEA. Finally, the methods presented here allow for retrieval of gene sets relevant to any topic that could be used as a reference library to link the observed changes to downstream phenotypes.

Summary

Chapter 2 introduced a resource called EMCON created using a data integration pipeline that links genes to any topic in literature. Chapter 3 highlights a major update to ToxRefDB that exposes a point of integration so that the resource can be included in data integration pipelines. The work presented in this chapter utilized work from both chapters to investigate the biological coverage of ToxCast specifically for cancer. Two approaches were taken to link ToxCast gene targets to cancer or cancer processes. The first queries EMCON to retrieve genes linked to important processes in cancer, which are then subsequently linked to the ToxCat genes. The second approach identifies organ-specific cancer genes by integrating ToxRefDB. Both direct and indirect methods were used to calculate biological coverage. The direct method quantifies the number of ToxCast genes are linked to each cancer process or organ-specific cancer observed from ToxRefDB. The indirect method incorporates pathways and processes from other resources like Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome. The indirect biological coverage is calculated as the number of pathways or processes that are linked to both ToxCast gene targets and corresponding cancer processes. According to direct biological coverage, ToxCast is relevant to most of the cancer processes identified in this work except for epigenetics. The indirect coverage shows higher biological coverage indicating that ToxCast has relevant biological coverage for all cancer processes. Despite having relevant biological coverage for cancer using this method, more work is needed to understand the implications of the links identified in this work. Manual curation of ToxCast is needed like that from the International Agency for Research on Cancer (IARC), but the data integration strategy here can aid in identifying relevant information to manually review. More work to understand etiologies of complex disease and the influence of environmental chemical exposures is needed in order to answer questions about the necessary biological targets to probe using NAMs to enable robust predictions of adverse outcomes.

CHAPTER 5: CONCLUSIONS, PERSPECTIVES, AND FUTURE DIRECTIONS

Interest in the development and use of NAMs for chemical safety evaluation will continue to grow as new technologies become available for massive data generation. The employment of this information in regulatory applications is still being considered as progress is made in building robust computational models. A prominent challenge exists in linking *in vitro* bioactivities to adverse outcomes (AOs) of interest. To understand the current data landscape and identify gaps in knowledge, legacy data as well as newly generated data must be able to be integrated. Thus, efforts to support interoperability across different information systems must be prioritized. The work outlined in this dissertation described methods for increasing data interoperability so resources can be combined in data integration and analysis pipelines as a new approach to investigate obstacles in toxicology.

First, in Chapter 2, a resource called Entity MeSH Co-occurrence Network (EMCON) that links genes to any topic in literature was created from information manually extracted from literature. The utility of EMCON was demonstrated by identifying genes linked to breast cancer and shown to retrieve relevant genes to the topic of interest, namely breast cancer. Second, in Chapter 3, Toxicity Reference Database (ToxRefDB) was extensively updated to increase the richness of the data provided and to ensure quality and support interoperability. A controlled vocabulary (CV) was established for the reporting of adverse events observed in animal toxicity studies. The CV was mapped to Unified Medical Language System (UMLS), a semantic network of over 150 biomedical vocabularies that allows for easy integration with other resources that are also mapped to UMLS or any of the contained vocabularies. Finally, in Chapter 4, the work from the previous two chapters was utilized to investigate the biological coverage of ToxCast for cancer. ToxCast is comprised of a number of assay endpoints that employ measures of genes relevant for all of the cancer processes identified in this work, except for epigenetics. When including pathways that contain the genes associated with ToxCast assay endpoints as an indirect measurement of biological coverage, the coverage of cancer concepts was higher and also included epigenetics. Despite having relevant biological coverage for cancer using this method, the use of ToxCast for building robust computational models for chemical cancer hazard classification may still not be feasible at this time.

The data integration workflow presented in Chapter 2 is an extension of other strategies for integrating the same information. The workflow and the use of normalized pointwise mutual information (NPMI) to identify relevant GeneID-MeSH associations were shown to be successful, supporting the assertion that bioinformatic tools like EMCON that bridge gaps in data interoperability will critically inform hypothesis generation and further research. However, caveats in this approach remain. The available information was limited to curated literature. Of the over 28 million articles in PubMed, less than one million were able to be integrated into EMCON. As manual curation and extraction efforts continue, there’s a need to develop automatic data extraction pipelines. The rate of publishing will always outpace the rate of manual curation, so unless automatic data extraction of relevant information exists, no one resource relying on information extracted from literature can be considered comprehensive.

For legacy information databases like ToxRefDB, manual curation and extraction of data is the only solution for many of the documents due to the format of the source files, which are not amenable to computational processing. The updates to ToxRefDB exemplify a good approach to ensure data quality and completeness. The use of a CV enforced consistency and reduced error rates. Although, manual extraction can never accomplish an error rate of zero due to unavoidable human error, quality checks can be put into place to identify errors derived from manual entry. Unfortunately, many legacy information systems with relevant toxicity data are in need of updates similar to ToxRefDB, and document management and an underlying infrastructure to support such updates are not prioritized. Without an emphasis on document management and building modern and flexible infrastructures, legacy systems will continue to grow in silos. Updating information systems infrastructure and prioritizing interoperability are feasible as exemplified by the recent changes to FDA’s data submission standards (262). FDA worked closely with an international data standards organization, CDISC, to define data formatting standards for global exchange of clinical and nonclinical information. Currently, companies that submit information on drugs to the FDA for review must comply with strict data formatting standards. A relevant use case in toxicology for the benefits of data formatting standards is the analysis of ToxCast data. Chemical dose-response data is collected from several vendors in various formats and may limit insights using established data analysis pipelines (e.g., tcpl). Developing data formatting standards for chemical dose-response information could speed up the delivery and analysis of results by enabling easier integration into other systems for the development of automated pipelines.

A significant application of this work is the use in organizing information for automatic binning and subsequent manual review for tasks like that of IARC in identifying ToxCast assays relevant to any of the TKCC. In Chapter 4, cancer processes were defined by consulting seminal publications in cancer; however, IARC only considered the TKCC. The data integration and binning methods from Chapter 4 can be updated to focus on TKCC to identify putative links between ToxCast assays and the TKCC. The underlying information that supports the associations can then be manually reviewed. The information available for review includes PubMed articles that support the links between ToxCast assay gene targets and TKCCs, enriched pathways and terms from the indirect biological coverage, and the evidence (i.e. research articles) linking genes to each pathway and term. This data organization approach yields a more comprehensive review of available information than a literature search alone and can be applied to any concept beyond cancer.

A further application of the work presented in this dissertation is a new route for exploring biological information within existing applications like the Comptox Chemicals Dashboard and the AOP wiki. Currently within the Comptox Chemicals Dashboard, information about assays is accessible if some aspect about the assay is known like the gene target or assay design. However, many users are interested in which assays are related to complex diseases, a feature that is not currently available. Within the AOP wiki, the information has been incorporated by crowdsourcing. With an underlying resource like EMCON, a biological search and subsequent exploration modules can be designed for browsing. In the AOP wiki, genes, gene products (e.g. proteins), pathways, and related concepts can be putatively added to gaps in existing AOPs for manual review. Also possible is condensing manual curation efforts. For example, AOPs for each TKCC or other relevant cancer process can be added automatically into the AOP wiki from EMCON. The AOP could then serve as a data curation and extraction tool used by IARC to continuously and transparently build evidence linking ToxCast assays to relevant cancer processes. Crowdsourcing or internal manual curation can continue to improve quality, but, if interoperability is considered when designing the application, information can continuously be incorporated whenever available. Continued development of resources like EMCON is iterative because information can be incrementally added by manual review and identifying and integrating other available, relevant resources.

Resources like EMCON can also be used to analyze information from new data streams like high-throughput transcriptomics. Identifying environmental chemical biological targets from gene expression data remains an active field of research as the understanding of genotype-phenotype relationships becomes more complex. Instead of traditional methods like GSEA and using reference gene set libraries to link pathways or other processes, a graph-based approach like those used to understand genetic susceptibility to complex disease (263–266) could be implemented. A network like EMCON could serve as a starting point to develop new methods for analyzing the dose-response gene expression profiles from high-throughput transcriptomics. As more information is added and curated, the analysis methods can iteratively be explored and improved upon. Table 5.1 further presents challenges faced by specific user groups or organizations while describing how EMCON can be applied to address their respective challenges. An IARC use case is further described in Figure 5.1 depicting how the original workflow can be augmented with EMCON incorporating a data-driven approach to identifying assays relevant to consider in evaluating carcinogenic potential of chemicals.

The path forward in the toxicology must include consideration of how data can be integrated and used, coinciding with expansion of the methods and approaches used in regulatory toxicology to investigate the impact chemical exposures have on human health. As federal agencies begin to prioritize good data management and modernize existing infrastructures, new analytical methods will continue to become available to explore not just how chemicals affect biological systems, but, also, deepen our understanding of human disease.

User group /Organization	Challenge	EMCON application
IARC	Bin *in vitro* assays into TKCC to evaluate mechanistic bioactivity per chemical	IARC has been increasingly relying on mechanistic information for identifying carcinogenic potential of chemicals for subsequent grouping. The ToxCast assays have been identified as a source of mechanistic data to be used in a weight-of-evidence approach by binning assays into TKCC. The original workflow relies only on experts to manually bin the assays into TKCC, but EMCON can be utilized as a data-driven approach to bin assays into TKCC while also pulling in relevant articles to review. Experts can select MeSH that represent each TKCC to retrieve both gene sets and relevant enriched topics from relevant reference libraries like Reactome or Gene Ontology. The links between the gene or enriched topic and one of the TKCC is supported with literature that can be reviewed. Expert involvement is still required because prior knowledge about the assay platform and any targets is needed as well as expertise to review literature. Figure 5.1 shows where EMCON can be incorporated in the existing workflow for using ToxCast as a source of mechanistic data to evaluate carcinogenic potential of chemicals.
EPA/Integrated Risk Information System (IRIS)	Identify relevant literature and related evidence for building chemical assessments	The IRIS program is tasked with building chemical assessments that include evidence supporting links between a given chemical exposure and specific health outcomes through a process called systematic review. Information is initially collected from a broad literature search from a number of different databases and then iteratively filtered using approaches like topic modeling. Manual review of literature is required, so maximizing retrieval of only relevant articles to minimize time spent reviewing irrelevant articles is a goal. EMCON can be used as one of the databases to retrieve literature as well as pulling in associated meta-data (e.g. linked pathways or other topics) as additional evidence. For this application, EMCON would be best utilized if chemicals were added the network. Experts can select MeSH that best account for health outcomes of interest pulling out relevant links to genes as well as chemicals. All links are supported by articles that can be further

		reviewed and considered as evidence in the assessment.
NTP, EPA/NCCT	Analyze dose-response transcriptomics data streams from S1500+ and HTTr projects by linking changes in gene expression to adverse outcomes	Both S1500+ and HTTr projects aim to generate dose-response transcriptomics information. Among the benefits of this approach are use of a single platform and ability to collect information from human cell lines. This approach has the potential to generate a massive amount of information, yet understanding how to analyze this information and linking the changes in gene expression to relevant adverse outcomes remains a challenge. Many of the environmental chemicals screened do not have well-defined molecular targets and, therefore, a gene expression profile may be difficult to extract and subsequently link to any adverse outcome. EMCON can be used to generate gene sets for any adverse outcome of interest. Traditional GSEA can be used, but would eliminate the need for manually curated reference libraries that may not capture relevant information about toxicity pathways. Experts select MeSH for adverse outcomes of interest or select toxicity outcomes from ToxRefDB to identify gene sets that serve as a reference for GSEA to analyze results from S1500+ and HTTr.

Table 5.1 Example use cases for using EMCON to address ongoing efforts in toxicology

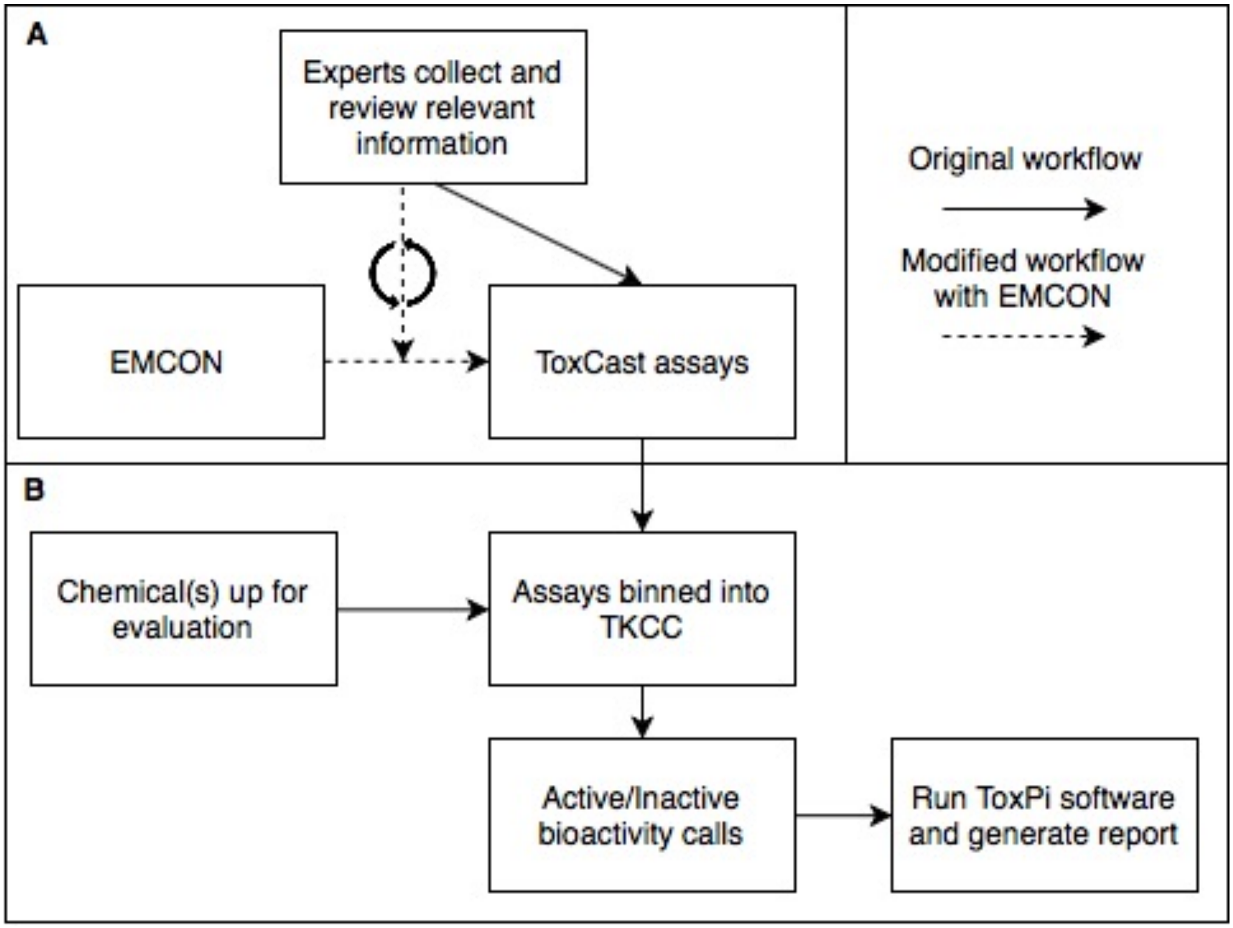

A
Experts collect and review relevant information
EMCON
ToxCast assays
Original workflow
Modified workflow with EMCON
B
Chemical(s) up for evaluation
Assays binned into TKCC
Active/Inactive bioactivity calls
Run ToxPi software and generate report

Figure 5.1 Example IARC workflow for using ToxCast to identify carcinogenic potential of chemicals
A) The original workflow relies experts with prior knowledge about the ToxCast assays as well as relevant expertise to review literature to bin each of the assays into one or more of the TKCC. EMCON can be used to modify this workflow and putatively bin many more assays into the TKCC. Iterative manual review is still required to eliminate false positives. B) The remaining original workflow would not have to be modified, but more assays could be utilized in this weight-of-evidence approach.

APPENDIX 1: USEPA PRODUCTS SUPPORTING COMPUTATIONAL TOXICOLOGY

Name	Description
ToxCast (invitrodb)	invitrodb is a database that stores the processed outputs and resulting data from the complete analysis pipeline (16) from different high-throughput technologies used for screening chemicals. The full MySQL database as well as a number of summary files are available from EPA's ftp website (267).
ToxVal	ToxVal is a database that contains Toxicity Values, or values that may be used within regulatory applications, derived from a number of sources (268).
ToxRefDB	Toxicity Reference Database (ToxRefDB) is a publicly available database of animal toxicity studies that primarily adhere to guideline studies (13). The recent update will include increased accessibility to the resource that includes access to the full database, summary files, example code, and a user's guide.
CPDat	Chemical Products Database (CPDat) stores information on chemicals in various products (68,269). The information is accessible through the Comptox Chemistry Dashboard (17).
DSSTox	Distributed Structure Searchable Toxicity Database (DSSTox) stores high quality chemical information and serves as the primary resource supporting the Comptox Chemistry Dashboard.
httk	High-throughput toxicokinetics (httk) is an R software package that implements models to calculate the dose a species would have been exposed to given the bioactivities in ToxCast (15).
ACToR	Aggregated Computational Toxicology Resource (ACToR) was a web application and database providing access to toxicity information across numerous domains. As other tools and applications have been developed, the ACToR project's primary focus has been development of a web services API called actorws (65).
tcpl	ToxCast pipeline (tcpl) is an R package used to model dose-response curves from ToxCast (16).
ToxCast Dashboard	The ToxCast Dashboard is a web application that exclusively provides access to the chemical and dose-response information for the ToxCast assays (270).
EDSP21 Dashboard	The EDSP21 Dashboard is a web application that provides access to the ToxCast data that was used to develop computational models to support the Endocrine Disruptor Screening Program (EDSP). The information available in this tool is the 2015 snapshot coinciding with the public notice on EDSP in the federal register (271,272).
CompTox Chemistry Dashboard	The Comptox Chemistry Dashboard provides toxicity and chemical information from a majority of the resources available within the EPA and many external resources. The web application serves as a step towards increasing interoperability across all existing resources currently supporting EPA's computational toxicology efforts (17).

APPENDIX 2: METADATA FOR PUBMED ARTICLES CITING EITHER MARTIN ET AL. (2009) (28) OR MARTIN ET AL. (2009) (29) AS OF AUGUST 8, 2018

PubMed identifier (PMID)	Article title	Journal title	Publication type(s)
19479008	The toxicity data landscape for environmental chemicals.	Environmental health perspectives	Journal Article;Review
20056575	Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook.	Environmental health perspectives	Congresses;Research Support, Non-U.S. Gov't
20368123	In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project.	Environmental health perspectives	Journal Article;Research Support, U.S. Gov't, Non-P.H.S.
20421935	Simulating microdosimetry in a virtual hepatic lobule.	PLoS computational biology	Journal Article;Research Support, U.S. Gov't, Non-P.H.S.
20483702	Computational toxicology: realizing the promise of the toxicity testing in the 21st century.	Environmental health perspectives	Congresses
20572635	Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research.	Journal of chemical information and modeling	Journal Article;Research Support, N.I.H., Extramural;Research Support, U.S. Gov't, Non-P.H.S.
20702588	A novel framework for predicting in vivo toxicities from in vitro data using optimal methods for dense and sparse matrix reordering and logistic regression.	Toxicological sciences : an official journal of the Society of Toxicology	Journal Article;Research Support, N.I.H., Extramural;Research Support, U.S. Gov't, Non-P.H.S.
20826373	Endocrine profiling and prioritization of environmental chemicals using ToxCast data.	Environmental health perspectives	Journal Article;Research Support, U.S. Gov't, Non-P.H.S.
21339822	Using nuclear receptor activity to stratify hepatocarcinogens.	PloS one	Journal Article
21538556	Genetic toxicology in the 21st century: reflections and future directions.	Environmental and molecular mutagenesis	Journal Article;Review
21556171	Combined toxic exposures and human health: biomarkers of exposure and effect.	International journal of environmental	Journal Article;Research

		research and public health	Support, Non-U.S. Gov't;Review
21666745	Evaluation of 309 environmental chemicals using a mouse embryonic stem cell adherent cell differentiation and cytotoxicity assay.	PloS one	Journal Article;Research Support, U.S. Gov't, Non-P.H.S.
21745399	Simulating quantitative cellular responses using asynchronous threshold Boolean network ensembles.	BMC systems biology	Journal Article
21788198	Environmental impact on vascular development predicted by high-throughput screening.	Environmental health perspectives	Evaluation Studies;Journal Article;Research Support, U.S. Gov't, Non-P.H.S.
22387746	Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data.	Toxicological sciences : an official journal of the Society of Toxicology	Journal Article;Research Support, N.I.H., Extramural;Research Support, U.S. Gov't, Non-P.H.S.
22405527	Toxic environmental chemicals: the role of reproductive health professionals in preventing harmful exposures.	American journal of obstetrics and gynecology	Journal Article;Research Support, N.I.H., Extramural;Research Support, Non-U.S. Gov't
22408426	Aggregating data for computational toxicology applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System.	International journal of molecular sciences	Journal Article;Research Support, U.S. Gov't, Non-P.H.S.;Review
23056181	Integrating constitutive gene expression and chemoactivity: mining the NCI60 anticancer screen.	PloS one	Journal Article
23086837	From QSAR to QSIIR: searching for enhanced computational toxicology models.	Methods in molecular biology (Clifton, N.J.)	Journal Article;Research Support, Non-U.S. Gov't;Review
23603051	A C. elegans screening platform for the rapid assessment of chemical disruption of germline function.	Environmental health perspectives	Journal Article;Research Support, N.I.H., Extramural;Research Support, Non-U.S. Gov't

23603828	Improving the human hazard characterization of chemicals: a Tox21 update.	Environmental health perspectives	Evaluation Studies;Journal Article;Review
23844697	Advancing human health risk assessment: integrating recent advisory committee recommendations.	Critical reviews in toxicology	Journal Article;Review
23958734	Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework.	Toxicological sciences : an official journal of the Society of Toxicology	Journal Article;Research Support, Non-U.S. Gov't
24415822	THE INTERACTIVE DECISION COMMITTEE FOR CHEMICAL TOXICITY ANALYSIS.	Journal of statistical research	Journal Article
24950175	Profiling animal toxicants by automatically mining public bioassay data: a big data approach for computational toxicology.	PloS one	Journal Article;Research Support, N.I.H., Extramural;Research Support, Non-U.S. Gov't
24972337	Multigenerational exposure to dietary zearalenone (ZEA), an estrogenic mycotoxin, affects puberty and reproduction in female mice.	Reproductive toxicology (Elmsford, N.Y.)	Journal Article;Research Support, N.I.H., Extramural
25326588	Identification of Environmental Chemicals Associated with the Development of Toxicant-associated Fatty Liver Disease in Rodents.	Toxicologic pathology	Journal Article;Research Support, N.I.H., Extramural;Review
25821157	Models of germ cell development and their application for toxicity studies.	Environmental and molecular mutagenesis	Journal Article;Research Support, N.I.H., Extramural;Research Support, Non-U.S. Gov't;Review
25836969	Predicting the future: opportunities and challenges for the chemical industry to apply 21st-century toxicity testing.	Journal of the American Association for Laboratory Animal Science : JAALAS	Journal Article
25984295	Reproductive toxicity and meiotic dysfunction following exposure to the pesticides Maneb, Diazinon and Fenarimol.	Toxicology research	Journal Article
26106137	Assessing the carcinogenic potential of low-dose exposures to	Carcinogenesis	Journal Article;Research

	chemical mixtures in the environment: focus on the cancer hallmark of tumor angiogenesis.		Support, N.I.H., Extramural;Research Support, Non-U.S. Gov't;Review
26496690	Developmental Effects of the ToxCast™ Phase I and Phase II Chemicals in Caenorhabditis elegans and Corresponding Responses in Zebrafish, Rats, and Rabbits.	Environmental health perspectives	Journal Article;Research Support, U.S. Gov't, Non-P.H.S.;Research Support, N.I.H., Extramural;Research Support, N.I.H., Intramural
26506572	Health effects of toxicants: Online knowledge support.	Life sciences	Journal Article;Research Support, N.I.H., Intramural
26662846	Systems Toxicology of Male Reproductive Development: Profiling 774 Chemicals for Molecular Targets and Adverse Outcomes.	Environmental health perspectives	Journal Article
26863090	Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008-2014.	ALTEX	Journal Article;Research Support, N.I.H., Extramural;Research Support, Non-U.S. Gov't
27884602	A data-driven weighting scheme for multivariate phenotypic endpoints recapitulates zebrafish developmental cascades.	Toxicology and applied pharmacology	Journal Article
28531190	Real-time cell toxicity profiling of Tox21 10K compounds reveals cytotoxicity dependent toxicity pathway linkage.	PloS one	Journal Article
29075892	Predicting in vivo effect levels for repeat-dose systemic toxicity using chemical, biological, kinetic and study covariates.	Archives of toxicology	Journal Article
29155963	Generating Modeling Data From Repeat-Dose Toxicity Reports.	Toxicological sciences : an official journal of the Society of Toxicology	Journal Article
30090397	Improving the prediction of organism-level toxicity through integration of chemical, protein target and cytotoxicity qHTS data.	Toxicology research	Journal Article

30090410	QSAR modeling for predicting reproductive toxicity of chemicals in rats for regulatory purposes.	Toxicology research	Journal Article

APPENDIX 3: SUBSET OF TOXREFDB ENDPOINTS AND EFFECTS CONSIDERED CANCER-RELATED FOR RESEARCH PURPOSES

Endpoint category	Endpoint type	Endpoint target	Effect description
systemic	pathology microscopic	bone marrow	lymphoma malignant
systemic	pathology gross	full gross necropsy	lipoma
systemic	pathology microscopic	thyroid gland	adenocarcinoma
systemic	pathology microscopic	epididymis	mesothelioma nos
systemic	pathology microscopic	uterus	interstitial stromal tumor
systemic	pathology microscopic	clitoral gland	adenoma
systemic	pathology microscopic	testes	adenoma
systemic	pathology microscopic	ovary	granulosa cell tumor
systemic	pathology microscopic	liver	hepatocholangiocarcinoma
systemic	pathology microscopic	epididymis	sarcoma
systemic	pathology microscopic	uterus	polyp adenomatous
systemic	pathology microscopic	kidney	neoplasm nos
systemic	pathology microscopic	zymbal's gland	squamous cell carcinoma
systemic	pathology microscopic	stomach	hemangiosarcoma
systemic	pathology microscopic	skin	fibrous histiocytoma
systemic	pathology microscopic	preputial gland	neoplasm nos
systemic	pathology microscopic	stomach	adenoma
systemic	pathology microscopic	intestine large	adenoma/carcinoma combined
systemic	pathology microscopic	liver	sarcoma
systemic	pathology microscopic	lymph node	leukemia
systemic	pathology microscopic	pancreas	hemangiosarcoma
systemic	pathology gross	full gross necropsy	mesothelioma malignant
systemic	pathology microscopic	bone	osteosarcoma
systemic	pathology microscopic	uterus	histiocytic sarcoma
systemic	pathology microscopic	seminal vesicle	adenoma
systemic	pathology microscopic	skin	subcutaneous mass
systemic	pathology microscopic	lung	histiocytic sarcoma
systemic	pathology gross	full gross necropsy	neoplasm nos
systemic	pathology microscopic	skin	adenocarcinoma
systemic	pathology microscopic	ovary	hemangiosarcoma
systemic	pathology microscopic	mammary gland	fibroadenoma
systemic	pathology microscopic	skin	papilloma
systemic	pathology microscopic	skin	lipoma
systemic	pathology microscopic	blood	leukemia mononuclear

systemic	pathology microscopic	skin	adenoma
systemic	pathology microscopic	thymus	lymphoma nos
systemic	pathology microscopic	lung	carcinoma nos
systemic	pathology microscopic	brain	glioma nos
systemic	pathology microscopic	mammary gland	adenoacanthoma
systemic	pathology microscopic	lymph node	lymphoma malignant
systemic	pathology microscopic	skin	basal cell carcinoma
systemic	pathology microscopic	stomach	fibrosarcoma
systemic	pathology microscopic	harderian gland	adenocarcinoma
systemic	pathology microscopic	ovary	sex cord stromal tumor, benign
systemic	pathology gross	full gross necropsy	lymphoma nos
systemic	pathology microscopic	lung	adenoma/carcinoma combined
systemic	pathology microscopic	zymbal's gland	carcinoma
systemic	pathology microscopic	intestine large	carcinoma
systemic	pathology microscopic	spleen	leukemia lymphocytic
systemic	pathology microscopic	kidney	lymphoma malignant
systemic	pathology microscopic	pituitary gland	acidophil adenoma
systemic	pathology microscopic	liver	cholangioma
systemic	pathology microscopic	testes	mesothelioma nos
systemic	pathology microscopic	mammary gland	adenocarcinoma
systemic	pathology microscopic	adrenal gland	pheochromocytoma malignant
systemic	pathology microscopic	skin	trichoepithelioma
systemic	pathology microscopic	skin	carcinoma
systemic	pathology microscopic	parathyroid gland	adenoma
systemic	pathology microscopic	bone marrow	leukemia
systemic	pathology microscopic	skin	basal cell adenoma
systemic	pathology microscopic	pharynx	squamous cell papilloma
systemic	pathology microscopic	nose	squamous cell carcinoma
systemic	pathology microscopic	intestine small	carcinoma
systemic	pathology microscopic	liver	adenoma/carcinoma combined
systemic	pathology microscopic	liver	hemangiosarcoma
systemic	pathology microscopic	urinary bladder	papilloma
systemic	pathology microscopic	mammary gland	adenoma/carcinoma combined
systemic	pathology microscopic	mesentery	mesothelioma nos
systemic	pathology gross	full gross necropsy	leukemia mononuclear

systemic	pathology microscopic	lung	alveolar/bronchiolar adenoma
systemic	pathology microscopic	kidney	sarcoma
systemic	pathology microscopic	mammary gland	mixed tumor nos
systemic	pathology microscopic	stomach	neoplasm nos
systemic	pathology microscopic	brain	astrocytoma malignant
systemic	pathology microscopic	skin	fibroma
systemic	pathology microscopic	uterus	deciduoma nos
systemic	pathology microscopic	spleen	fibroma
systemic	pathology microscopic	zymbal's gland	adenoma
systemic	pathology microscopic	stomach	leiomyosarcoma
systemic	pathology microscopic	adrenal gland	pheochromocytoma nos
systemic	pathology microscopic	spleen	hemangiosarcoma
systemic	pathology microscopic	liver	leukemia lymphocytic
systemic	pathology microscopic	adrenal gland	hemangiosarcoma
systemic	pathology microscopic	blood	neoplasm nos
systemic	pathology microscopic	liver	neoplasm nos
systemic	pathology microscopic	nose	mixed tumor malignant
systemic	pathology microscopic	liver	neoplastic nodule
systemic	pathology microscopic	testes	neoplasm nos
systemic	pathology microscopic	skin	mesothelioma malignant
systemic	pathology microscopic	gallbladder	adenoma
systemic	pathology microscopic	uterus	polyp
systemic	pathology microscopic	uterus	adenoma/carcinoma combined
systemic	pathology microscopic	mammary gland	carcinoma
systemic	pathology microscopic	skin	keratoacanthoma
systemic	pathology microscopic	ovary	granulosa-theca tumor nos
systemic	pathology microscopic	peritoneum	neoplasm nos
systemic	pathology microscopic	thyroid gland	neoplasm nos
systemic	pathology microscopic	liver	mixed tumor nos
systemic	pathology microscopic	lacrimal gland	lymphoma malignant
systemic	pathology microscopic	tongue	papilloma
systemic	pathology microscopic	skin	rhabdomyosarcoma
systemic	pathology microscopic	epididymis	mesothelioma benign
systemic	pathology microscopic	pituitary gland	adenoma
systemic	pathology microscopic	oral mucosa	squamous cell carcinoma
systemic	pathology microscopic	nerve	sarcoma
systemic	pathology microscopic	skin	hemangioma
systemic	pathology microscopic	nose	neoplasm nos
systemic	pathology microscopic	urinary bladder	lymphoma nos
systemic	pathology microscopic	liver	mixed tumor malignant

systemic	pathology microscopic	preputial gland	adenoma/carcinoma combined
systemic	pathology microscopic	blood	lymphoma malignant
systemic	pathology microscopic	skin	schwannoma nos
systemic	pathology gross	full gross necropsy	osteosarcoma
systemic	pathology microscopic	intestine small	adenoma
systemic	pathology microscopic	liver	adenocarcinoma
systemic	pathology microscopic	liver	hepatocellular carcinoma
systemic	pathology microscopic	liver	hepatoblastoma
systemic	pathology microscopic	blood	leukemia lymphocytic
systemic	pathology microscopic	tongue	carcinoma
systemic	pathology microscopic	liver	adenoma
systemic	pathology microscopic	liver	cholangiocarcinoma
systemic	pathology microscopic	salivary glands	lymphoma malignant
systemic	pathology microscopic	lung	adenocarcinoma
systemic	pathology microscopic	heart	schwannoma malignant
systemic	pathology microscopic	stomach	squamous cell carcinoma
systemic	pathology microscopic	oral mucosa	carcinoma nos
systemic	pathology microscopic	uterus	adenocarcinoma
systemic	pathology microscopic	brain	astrocytoma nos
systemic	pathology microscopic	intestine large	adenoma
systemic	pathology microscopic	stomach	squamous cell papilloma
systemic	pathology microscopic	harderian gland	carcinoma
systemic	pathology microscopic	intestine large	hemangiosarcoma
systemic	pathology microscopic	peritoneum	mesothelioma benign
systemic	pathology microscopic	lymph node	hemangioma
systemic	pathology microscopic	heart	lymphoma malignant
systemic	pathology microscopic	nose	adenocarcinoma
systemic	pathology microscopic	bone	adenoma
systemic	pathology microscopic	lymph node	lymphoma nos
systemic	pathology microscopic	kidney	adenoma/carcinoma combined
systemic	pathology microscopic	spleen	fibrosarcoma
systemic	pathology microscopic	blood	lymphoma malignant histiocytic
systemic	pathology microscopic	skin	squamous cell carcinoma
systemic	pathology microscopic	intestine small	neoplasm nos
systemic	pathology microscopic	pituitary gland	carcinoma
systemic	pathology microscopic	ovary	mixed tumor benign
systemic	pathology microscopic	ureter	papilloma
systemic	pathology microscopic	thymus	thymoma nos
systemic	pathology microscopic	thyroid gland	cystadenoma

systemic	pathology microscopic	intestine large	squamous cell carcinoma
systemic	pathology gross	full gross necropsy	rhabdomyosarcoma
systemic	pathology microscopic	uterus	polyp stromal
systemic	pathology microscopic	lung	alveolar/bronchiolar carcinoma
systemic	pathology microscopic	tongue	squamous cell papilloma
systemic	pathology microscopic	thyroid gland	adenoma/carcinoma combined
systemic	pathology microscopic	esophagus	squamous cell papilloma
systemic	pathology microscopic	clitoral gland	adenoma/carcinoma combined
systemic	pathology microscopic	preputial gland	squamous cell carcinoma
systemic	pathology microscopic	spleen	sarcoma
systemic	pathology microscopic	intestine small	polyp adenomatous
systemic	pathology microscopic	nose	polyp
systemic	pathology microscopic	thymus	lymphoma malignant
systemic	pathology microscopic	skin	squamous cell papilloma
systemic	pathology microscopic	pituitary gland	adenoma/carcinoma combined
systemic	pathology gross	full gross necropsy	lymphoma malignant
systemic	pathology microscopic	skeletal muscle	sarcoma
systemic	pathology microscopic	thyroid gland	mixed tumor nos
systemic	pathology microscopic	blood	leukemia
systemic	pathology microscopic	ovary	neoplasm nos
systemic	pathology microscopic	blood vessel	hemangioma
systemic	pathology gross	full gross necropsy	leukemia granulocytic
systemic	pathology microscopic	kidney	transitional epithelial carcinoma
systemic	pathology microscopic	lung	squamous cell carcinoma
systemic	pathology microscopic	spleen	neoplasm nos
systemic	pathology microscopic	skin	fibrous histiocytoma benign
systemic	pathology microscopic	adrenal gland	adenoma
systemic	pathology microscopic	testes	interstitial cell tumor benign
systemic	pathology microscopic	mesentery	hemangiosarcoma
systemic	pathology microscopic	kidney	papilloma
systemic	pathology microscopic	stomach	carcinosarcoma
systemic	pathology microscopic	uterus	mesothelioma malignant
systemic	pathology microscopic	intestine small	adenoma/carcinoma combined
systemic	pathology microscopic	lung	carcinoma
systemic	pathology microscopic	seminal vesicle	carcinoma

systemic	pathology microscopic	pancreas	adenoma
systemic	pathology microscopic	preputial gland	adenoma
systemic	pathology microscopic	urinary bladder	carcinoma
systemic	pathology microscopic	pancreas	adenoma/carcinoma combined
systemic	pathology microscopic	liver	carcinoma nos
systemic	pathology microscopic	uterus	carcinoma
systemic	pathology microscopic	stomach	papilloma
systemic	pathology microscopic	uterus	sarcoma
systemic	pathology microscopic	lung	mixed tumor nos
systemic	pathology gross	full gross necropsy	adenoma
systemic	pathology microscopic	liver	hemangioma
systemic	pathology microscopic	adrenal gland	pheochromocytoma benign
systemic	pathology gross	full gross necropsy	hemangiosarcoma
systemic	pathology microscopic	urinary bladder	squamous cell carcinoma
systemic	pathology microscopic	kidney	adenocarcinoma
systemic	pathology microscopic	skin	fibrosarcoma
systemic	pathology microscopic	kidney	lymphoma nos
systemic	pathology microscopic	harderian gland	adenoma/carcinoma combined
systemic	pathology microscopic	seminal vesicle	carcinoma nos
systemic	pathology microscopic	liver	carcinoma
systemic	pathology microscopic	liver	hepatocholangioma
systemic	pathology microscopic	adrenal gland	adenoma/carcinoma combined
systemic	pathology gross	full gross necropsy	hemangioma
systemic	pathology microscopic	testes	interstitial cell tumor nos
systemic	pathology microscopic	ovary	lymphoma malignant
systemic	pathology microscopic	bone marrow	leukemia mononuclear
systemic	pathology microscopic	lymph node	leukemia granulocytic
systemic	pathology microscopic	lung	cystic keratinizing epithelioma
systemic	pathology microscopic	brain	adenoma
systemic	pathology microscopic	mammary gland	mixed tumor malignant
systemic	pathology microscopic	liver	histiocytic sarcoma
systemic	pathology microscopic	liver	leukemia mononuclear
systemic	pathology microscopic	nose	carcinoma
systemic	pathology microscopic	adrenal gland	pheochromocytoma complex
systemic	pathology microscopic	ovary	adenoma
systemic	pathology microscopic	kidney	carcinoma

systemic	pathology microscopic	skin	hemangiosarcoma
systemic	pathology microscopic	preputial gland	carcinoma
systemic	pathology microscopic	uterus	adenoma
systemic	pathology microscopic	intestine large	polyp adenomatous
systemic	pathology microscopic	gallbladder	papilloma
systemic	pathology microscopic	prostate	adenoma
systemic	pathology microscopic	thyroid gland	adenoma
systemic	pathology microscopic	pancreas	adenocarcinoma
systemic	pathology microscopic	gallbladder	hemangioma
systemic	pathology microscopic	intestine large	lipoma
systemic	pathology microscopic	ureter	adenoma
systemic	pathology microscopic	blood vessel	hemangiosarcoma
systemic	pathology microscopic	uterus	squamous cell carcinoma
systemic	pathology microscopic	spinal cord	astrocytoma nos
systemic	pathology microscopic	eye	adenoma
systemic	pathology microscopic	stomach	carcinoma in situ
systemic	pathology microscopic	stomach	adenocarcinoma
systemic	pathology microscopic	heart	hemangiosarcoma
systemic	pathology microscopic	ovary	tubulostromal adenoma
systemic	pathology microscopic	pancreas	lymphoma nos
systemic	pathology microscopic	ovary	interstitial stromal tumor
systemic	pathology microscopic	stomach	carcinoma
systemic	pathology gross	full gross necropsy	histiocytic sarcoma
systemic	pathology microscopic	intestine small	mixed tumor nos
systemic	pathology microscopic	liver	hepatocellular adenoma
systemic	pathology microscopic	mesentery	hemangioma
systemic	pathology microscopic	uterus	hemangioma
systemic	pathology microscopic	oral mucosa	squamous cell papilloma
systemic	pathology microscopic	urinary bladder	leiomyosarcoma
systemic	pathology microscopic	pharynx	carcinoma
systemic	pathology microscopic	zymbal's gland	adenoma/carcinoma combined
systemic	pathology microscopic	spleen	leukemia mononuclear
systemic	pathology microscopic	intestine large	adenocarcinoma
systemic	pathology microscopic	intestine small	sarcoma
systemic	pathology microscopic	nose	sarcoma
systemic	pathology microscopic	ovary	granular cell tumor malignant
systemic	pathology microscopic	spleen	osteosarcoma
systemic	pathology microscopic	urinary bladder	transitional epithelial carcinoma
systemic	pathology microscopic	mammary gland	fibroma

systemic	pathology microscopic	vagina	squamous cell carcinoma
systemic	pathology microscopic	intestine small	adenocarcinoma
systemic	pathology microscopic	thyroid gland	carcinoma
systemic	pathology microscopic	ear	squamous cell papilloma
systemic	pathology microscopic	skin	sarcoma
systemic	pathology microscopic	stomach	polyp adenomatous
systemic	pathology microscopic	mammary gland	adenoma
systemic	pathology microscopic	mammary gland	carcinosarcoma
systemic	pathology microscopic	blood	lymphoma nos
systemic	pathology gross	full gross necropsy	sarcoma
systemic	pathology microscopic	nose	papilloma
systemic	pathology microscopic	peritoneum	sarcoma
systemic	pathology microscopic	ovary	granular cell tumor benign
systemic	pathology microscopic	ovary	granulosa cell tumor benign
systemic	pathology microscopic	clitoral gland	carcinoma
systemic	pathology microscopic	nose	rhabdomyosarcoma
systemic	pathology microscopic	ovary	luteoma
systemic	pathology microscopic	lung	adenoma
systemic	pathology microscopic	nose	adenoma
systemic	pathology microscopic	lymph node	hemangiosarcoma
systemic	pathology microscopic	testes	mesothelioma malignant
systemic	pathology microscopic	blood	histiocytic sarcoma
systemic	pathology gross	full gross necropsy	fibrosarcoma
systemic	pathology microscopic	lung	neoplasm nos
systemic	pathology microscopic	kidney	adenoma
systemic	pathology gross	full gross necropsy	lymp lymphoma malignant lymphocytic
systemic	pathology microscopic	harderian gland	adenoma
systemic	pathology microscopic	uterus	neoplasm nos
systemic	pathology microscopic	ovary	cystadenoma
systemic	pathology microscopic	ovary	carcinoma nos
systemic	pathology microscopic	[other]	histiocytic sarcoma
systemic	pathology microscopic	[other]	hemangioma
systemic	pathology microscopic	[other]	hemangiosarcoma
systemic	pathology microscopic	[other]	leukemia granulocytic
systemic	pathology microscopic	[other]	lymphoma malignant
systemic	pathology microscopic	[other]	leukemia mononuclear

APPENDIX 4: CANCER-RELATED ENDPOINTS FROM TOXREFDB AND THE CORRESPONDING TOXCAST INTENDED GENE TARGETS AND ASSAY ENDPOINTS

Endpoint category	Endpoint type	Endpoint target	Number of ToxCast intended gene targets	Number of AEIDs
systemic	pathology microscopic	prostate	1	9
systemic	pathology microscopic	blood vessel	1	2
systemic	pathology microscopic	thymus	1	2
systemic	pathology microscopic	harderian gland	1	4
systemic	pathology microscopic	pancreas	1	1
systemic	pathology microscopic	mesentery	1	2
systemic	pathology microscopic	oral mucosa	1	4
systemic	pathology microscopic	skeletal muscle	1	2
systemic	pathology microscopic	adrenal gland	2	2
systemic	pathology microscopic	gallbladder	2	9
systemic	pathology microscopic	zymbal's gland	2	16
systemic	pathology microscopic	seminal vesicle	2	11
systemic	pathology microscopic	preputial gland	2	16
systemic	pathology microscopic	blood	2	2
systemic	pathology microscopic	thyroid gland	2	5
systemic	pathology microscopic	spleen	3	8
systemic	pathology microscopic	brain	3	4
systemic	pathology microscopic	[other]	3	8
systemic	pathology gross	full gross necropsy	3	5
systemic	pathology microscopic	kidney	4	19
systemic	pathology microscopic	bone marrow	4	11
systemic	pathology microscopic	stomach	4	12
systemic	pathology microscopic	intestine large	4	20
systemic	pathology microscopic	lymph node	4	8
systemic	pathology microscopic	nose	4	18
systemic	pathology microscopic	testes	4	4
systemic	pathology microscopic	ovary	5	8
systemic	pathology microscopic	vagina	5	25
systemic	pathology microscopic	urinary bladder	6	9
systemic	pathology microscopic	uterus	7	40
systemic	pathology microscopic	lung	7	25
systemic	pathology microscopic	mammary gland	8	42
systemic	pathology microscopic	skin	10	22
systemic	pathology microscopic	liver	31	68

APPENDIX 5: RELEVANCE DECISIONS FOR TOP TEN TERMS FROM EACH REFERENCE LIBRARY

Term	Relevance (0/1)	Carcinogenesis concept (CC)	Reference library
regulation of endothelial cell chemotaxis to fibroblast growth factor (GO:2000544)	1	Angiogenesis	GO_Biological_Process_2018
fibroblast growth factor receptor signaling pathway (GO:0008543)	1	Angiogenesis	GO_Biological_Process_2018
Activation of gene expression by SREBF (SREBP)_Homo sapiens_R-HSA-2426168	1	Liver_cancer_(ToxRefDB)	Reactome_2016
regulation of angiogenesis (GO:0045765)	1	Angiogenesis	GO_Biological_Process_2018
cellular response to fibroblast growth factor stimulus (GO:0044344)	1	Angiogenesis	GO_Biological_Process_2018
activin-activated receptor activity (GO:0017002)	1	Angiogenesis	GO_Molecular_Function_2018
growth factor activity (GO:0008083)	1	Angiogenesis	GO_Molecular_Function_2018
Alcoholism_Homo sapiens_hsa05034	1	Epigenetics	KEGG_2016
fibroblast growth factor receptor binding (GO:0005104)	1	Angiogenesis	GO_Molecular_Function_2018
APC/C:Cdc20 mediated degradation of Securin_Homo sapiens_R-HSA-174154	1	Apoptosis	Reactome_2016
APC/C:Cdc20 mediated degradation of Securin_Homo sapiens_R-HSA-174154	1	Cell_cycle	Reactome_2016
APC/C:Cdc20 mediated degradation of Securin_Homo sapiens_R-HSA-174154	1	Proliferation	Reactome_2016
Ascorbate and aldarate metabolism_Homo sapiens_hsa00053	1	Growth_hormones	KEGG_2016
Autodegradation of Cdh1 by Cdh1:APC/C_Homo sapiens_R-HSA-174084	1	Apoptosis	Reactome_2016
Autodegradation of Cdh1 by Cdh1:APC/C_Homo sapiens_R-HSA-174084	1	Cell_cycle	Reactome_2016
C3HC4-type RING finger domain binding (GO:0055131)	1	Immortalization	GO_Molecular_Function_2018

CARD domain binding (GO:0050700)	1	Inflammation	GO_Molecular_Function_2018
DEx/H-box helicases activate type I IFN and inflammatory cytokines production_Homo sapiens_R-HSA-3134963	1	Inflammation	Reactome_2016
DNA biosynthetic process (GO:0071897)	1	Oxidative_stress	GO_Biological_Process_2018
DNA helicase activity (GO:0003678)	1	Epigenetics	GO_Molecular_Function_2018
DNA-dependent ATPase activity (GO:0008094)	1	Apoptosis	GO_Molecular_Function_2018
negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle (GO:0051436)	1	Apoptosis	GO_Biological_Process_2018
anaphase-promoting complex-dependent catabolic process (GO:0031145)	1	Apoptosis	GO_Biological_Process_2018
cell cycle G2/M phase transition (GO:0044839)	1	Apoptosis	GO_Biological_Process_2018
positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition (GO:0051437)	1	Apoptosis	GO_Biological_Process_2018
DNA-dependent ATPase activity (GO:0008094)	1	Epigenetics	GO_Molecular_Function_2018
DNA-dependent ATPase activity (GO:0008094)	1	Oxidative_stress	GO_Molecular_Function_2018
double-strand break repair via nonhomologous end joining (GO:0006303)	1	Oxidative_stress	GO_Biological_Process_2018
kinase binding (GO:0019900)	1	Apoptosis	GO_Molecular_Function_2018
damaged DNA binding (GO:0003684)	1	Apoptosis	GO_Molecular_Function_2018
Nucleotide excision repair_Homo sapiens_hsa03420	1	Apoptosis	KEGG_2016
Dual Incision in GG-NER_Homo sapiens_R-HSA-5696400	1	Oxidative_stress	Reactome_2016
p53 signaling pathway_Homo sapiens_hsa04115	1	Apoptosis	KEGG_2016
Cell cycle_Homo sapiens_hsa04110	1	Apoptosis	KEGG_2016
Apoptosis_Homo sapiens_hsa04210	1	Apoptosis	KEGG_2016

Dual incision in TC-NER_Homo sapiens_R-HSA-6782135	1	Oxidative_stress	Reactome_2016
Activation of NF-kappaB in B cells_Homo sapiens_R-HSA-1169091	1	Apoptosis	Reactome_2016
Processing of DNA double-strand break ends_Homo sapiens_R-HSA-5693607	1	Apoptosis	Reactome_2016
Energy dependent regulation of mTOR by LKB1-AMPK_Homo sapiens_R-HSA-380972	1	Immortalization	Reactome_2016
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression_Homo sapiens_R-HSA-427389	1	Epigenetics	Reactome_2016
negative regulation of G2/M transition of mitotic cell cycle (GO:0010972)	1	Cell_cycle	GO_Biological_Process_2018
regulation of hematopoietic progenitor cell differentiation (GO:1901532)	1	Cell_cycle	GO_Biological_Process_2018
exodeoxyribonuclease activity, producing 5'-phosphomonoesters (GO:0016895)	1	Apoptosis	GO_Molecular_Function_2018
anaphase-promoting complex-dependent catabolic process (GO:0031145)	1	Cell_cycle	GO_Biological_Process_2018
positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition (GO:0051437)	1	Cell_cycle	GO_Biological_Process_2018
Fanconi anemia pathway_Homo sapiens_hsa03460	1	Genotoxicity	KEGG_2016
FGFR2c ligand binding and activation_Homo sapiens_R-HSA-190375	1	Angiogenesis	Reactome_2016
kinase binding (GO:0019900)	1	Cell_cycle	GO_Molecular_Function_2018
FGFR4 ligand binding and activation_Homo sapiens_R-HSA-190322	1	Angiogenesis	Reactome_2016
FoxO signaling pathway_Homo sapiens_hsa04068	1	Apoptosis	KEGG_2016

FoxO signaling pathway_Homo sapiens_hsa04068	1	Immortalization	KEGG_2016
frizzled binding (GO:0005109)	1	Cell_cycle	GO_Molecular_Function_2018
Oocyte meiosis_Homo sapiens_hsa04114	1	Cell_cycle	KEGG_2016
Glucagon-type ligand receptors_Homo sapiens_R-HSA-420092	1	Growth_hormones	Reactome_2016
Cell cycle_Homo sapiens_hsa04110	1	Cell_cycle	KEGG_2016
The role of GTSE1 in G2/M progression after G2 checkpoint_Homo sapiens_R-HSA-8852276	1	Cell_cycle	Reactome_2016
glucuronate metabolic process (GO:0019585)	0	Growth_hormones	GO_Biological_Process_2018
Glucuronidation_Homo sapiens_R-HSA-156588	0	Growth_hormones	Reactome_2016
Hepatitis B_Homo sapiens_hsa05161	0	Immortalization	KEGG_2016
Hippo signaling pathway_Homo sapiens_hsa04390	1	Cell_cycle	KEGG_2016
DNA replication-independent nucleosome organization (GO:0034724)	1	Epigenetics	GO_Biological_Process_2018
chromatin silencing at rDNA (GO:0000183)	1	Epigenetics	GO_Biological_Process_2018
negative regulation of gene expression, epigenetic (GO:0045814)	1	Epigenetics	GO_Biological_Process_2018
ATP-dependent chromatin remodeling (GO:0043044)	1	Epigenetics	GO_Biological_Process_2018
chromatin assembly (GO:0031497)	1	Epigenetics	GO_Biological_Process_2018
HTLV-I infection_Homo sapiens_hsa05166	1	Cell_cycle	KEGG_2016
HTLV-I infection_Homo sapiens_hsa05166	1	Proliferation	KEGG_2016
Integrin cell surface interactions_Homo sapiens_R-HSA-216083	1	Angiogenesis	Reactome_2016
interstrand cross-link repair (GO:0036297)	1	Oxidative_stress	GO_Biological_Process_2018
nucleosomal DNA binding (GO:0031492)	1	Epigenetics	GO_Molecular_Function_2018
Lysine degradation_Homo sapiens_hsa00310	1	Epigenetics	KEGG_2016
Transcriptional misregulation in	1	Epigenetics	KEGG_2016

cancer_Homo sapiens_hsa05202			
MAPK signaling pathway_Homo sapiens_hsa04010	1	Angiogenesis	KEGG_2016
Melanoma_Homo sapiens_hsa05218	1	Angiogenesis	KEGG_2016
microtubule plus-end binding (GO:0051010)	1	Cell_cycle	GO_Molecular_Function_2018
HDACs deacetylate histones_Homo sapiens_R-HSA-3214815	1	Epigenetics	Reactome_2016
mitochondrial electron transport, NADH to ubiquinone (GO:0006120)	1	Oxidative_stress	GO_Biological_Process_2018
DNA methylation_Homo sapiens_R-HSA-5334118	1	Epigenetics	Reactome_2016
RMTs methylate histone arginines_Homo sapiens_R-HSA-3214858	1	Epigenetics	Reactome_2016
PRC2 methylates histones and DNA_Homo sapiens_R-HSA-212300	1	Epigenetics	Reactome_2016
interstrand cross-link repair (GO:0036297)	1	Genotoxicity	GO_Biological_Process_2018
non-recombinational repair (GO:0000726)	1	Genotoxicity	GO_Biological_Process_2018
DNA replication (GO:0006260)	1	Genotoxicity	GO_Biological_Process_2018
double-strand break repair via nonhomologous end joining (GO:0006303)	1	Genotoxicity	GO_Biological_Process_2018
transcription-coupled nucleotide-excision repair (GO:0006283)	1	Genotoxicity	GO_Biological_Process_2018
DNA helicase activity (GO:0003678)	1	Genotoxicity	GO_Molecular_Function_2018
DNA-directed DNA polymerase activity (GO:0003887)	1	Genotoxicity	GO_Molecular_Function_2018
DNA-dependent ATPase activity (GO:0008094)	1	Genotoxicity	GO_Molecular_Function_2018
damaged DNA binding (GO:0003684)	1	Genotoxicity	GO_Molecular_Function_2018
single-stranded DNA binding (GO:0003697)	1	Genotoxicity	GO_Molecular_Function_2018
Base excision repair_Homo sapiens_hsa03410	1	Genotoxicity	KEGG_2016
Alcoholism_Homo sapiens_hsa05034	1	Genotoxicity	KEGG_2016

Viral carcinogenesis_Homo sapiens_hsa05203	1	Genotoxicity	KEGG_2016
mTORC1-mediated signalling_Homo sapiens_R-HSA-166208	1	Immortalization	Reactome_2016
Nucleotide excision repair_Homo sapiens_hsa03420	1	Genotoxicity	KEGG_2016
Nonhomologous End-Joining (NHEJ)_Homo sapiens_R-HSA-5693571	1	Genotoxicity	Reactome_2016
Dual incision in TC-NER_Homo sapiens_R-HSA-6782135	1	Genotoxicity	Reactome_2016
NAD+ ADP-ribosyltransferase activity (GO:0003950)	1	Apoptosis	GO_Molecular_Function_2018
Processing of DNA double-strand break ends_Homo sapiens_R-HSA-5693607	1	Genotoxicity	Reactome_2016
Resolution of Sister Chromatid Cohesion_Homo sapiens_R-HSA-2500257	1	Genotoxicity	Reactome_2016
cAMP-mediated signaling (GO:0019933)	1	Growth_hormones	GO_Biological_Process_2018
JAK-STAT cascade involved in growth hormone signaling pathway (GO:0060397)	1	Growth_hormones	GO_Biological_Process_2018
negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle (GO:0051436)	1	Cell_cycle	GO_Biological_Process_2018
positive regulation of multicellular organism growth (GO:0040018)	1	Growth_hormones	GO_Biological_Process_2018
response to nutrient levels (GO:0031667)	1	Growth_hormones	GO_Biological_Process_2018
insulin receptor binding (GO:0005158)	1	Growth_hormones	GO_Molecular_Function_2018
neuropeptide hormone activity (GO:0005184)	1	Growth_hormones	GO_Molecular_Function_2018
protein-hormone receptor activity (GO:0016500)	1	Growth_hormones	GO_Molecular_Function_2018
insulin-like growth factor receptor binding (GO:0005159)	1	Growth_hormones	GO_Molecular_Function_2018
insulin-like growth factor II binding (GO:0031995)	1	Growth_hormones	GO_Molecular_Function_2018
Neuroactive ligand-receptor	0	Growth_hormones	KEGG_2016

interaction_Homo sapiens_hsa04080			
NF-kappa B signaling pathway_Homo sapiens_hsa04064	1	Inflammation	KEGG_2016
Ovarian steroidogenesis_Homo sapiens_hsa04913	1	Growth_hormones	KEGG_2016
Steroid hormone biosynthesis_Homo sapiens_hsa00140	1	Growth_hormones	KEGG_2016
NOD1/2 Signaling Pathway_Homo sapiens_R-HSA-168638	1	Immortalization	Reactome_2016
NOD-like receptor signaling pathway_Homo sapiens_hsa04621	1	Inflammation	KEGG_2016
Prolactin receptor signaling_Homo sapiens_R-HSA-1170546	1	Growth_hormones	Reactome_2016
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)_Homo sapiens_R-HSA-381426	1	Growth_hormones	Reactome_2016
Growth hormone receptor signaling_Homo sapiens_R-HSA-982772	1	Growth_hormones	Reactome_2016
Pentose and glucuronate interconversions_Homo sapiens_hsa00040	1	Growth_hormones	KEGG_2016
regulation of autophagy (GO:0010506)	1	Immortalization	GO_Biological_Process_2018
autophagy of nucleus (GO:0044804)	1	Immortalization	GO_Biological_Process_2018
mitochondrion disassembly (GO:0061726)	1	Immortalization	GO_Biological_Process_2018
positive regulation of autophagy (GO:0010508)	1	Immortalization	GO_Biological_Process_2018
regulation of macroautophagy (GO:0016241)	1	Immortalization	GO_Biological_Process_2018
phosphatidylcholine-sterol O-acyltransferase activator activity (GO:0060228)	0	Inflammation	GO_Molecular_Function_2018
death domain binding (GO:0070513)	1	Immortalization	GO_Molecular_Function_2018
kinase binding (GO:0019900)	1	Immortalization	GO_Molecular_Function_2018
cysteine-type endopeptidase activity	1	Immortalization	GO_Molecular_Function_2018

involved in execution phase of apoptosis (GO:0097200)			
phosphatidylinositol-4,5-bisphosphate 3-kinase activity (GO:0046934)	1	Angiogenesis	GO_Molecular_Function_2018
phosphatidylinositol-4,5-bisphosphate 3-kinase activity (GO:0046934)	1	Oxidative_stress	GO_Molecular_Function_2018
phospholipase activator activity (GO:0016004)	1	Immunomodulation	GO_Molecular_Function_2018
Longevity regulating pathway - mammal_Homo sapiens_hsa04211	1	Immortalization	KEGG_2016
p53 signaling pathway_Homo sapiens_hsa04115	1	Immortalization	KEGG_2016
Apoptosis_Homo sapiens_hsa04210	1	Immortalization	KEGG_2016
Regulation of TP53 Activity through Phosphorylation_Homo sapiens_R-HSA-6804756	1	Immortalization	Reactome_2016
PI3K-Akt signaling pathway_Homo sapiens_hsa04151	1	Angiogenesis	KEGG_2016
PI3K-Akt signaling pathway_Homo sapiens_hsa04151	1	Proliferation	KEGG_2016
Platelet degranulation_Homo sapiens_R-HSA-114608	1	Angiogenesis	Reactome_2016
Macroautophagy_Homo sapiens_R-HSA-1632852	1	Immortalization	Reactome_2016
monocyte chemotaxis (GO:0002548)	1	Immunomodulation	GO_Biological_Process_2018
chemokine-mediated signaling pathway (GO:0070098)	1	Immunomodulation	GO_Biological_Process_2018
neutrophil chemotaxis (GO:0030593)	1	Immunomodulation	GO_Biological_Process_2018
B cell activation (GO:0042113)	1	Immunomodulation	GO_Biological_Process_2018
B cell receptor signaling pathway (GO:0050853)	1	Immunomodulation	GO_Biological_Process_2018
C-C chemokine binding (GO:0019957)	1	Immunomodulation	GO_Molecular_Function_2018
positive regulation of protein kinase B signaling (GO:0051897)	1	Angiogenesis	GO_Biological_Process_2018
Processing of DNA double-strand break ends_Homo sapiens_R-HSA-5693607	1	Oxidative_stress	Reactome_2016

chemokine activity (GO:0008009)	1	Immunomodulation	GO_Molecular_Function_2018
CCR chemokine receptor binding (GO:0048020)	1	Immunomodulation	GO_Molecular_Function_2018
Hematopoietic cell lineage_Homo sapiens_hsa04640	1	Immunomodulation	KEGG_2016
Chemokine signaling pathway_Homo sapiens_hsa04062	1	Immunomodulation	KEGG_2016
B cell receptor signaling pathway_Homo sapiens_hsa04662	1	Immunomodulation	KEGG_2016
Primary immunodeficiency_Homo sapiens_hsa05340	1	Immunomodulation	KEGG_2016
Cytokine-cytokine receptor interaction_Homo sapiens_hsa04060	1	Immunomodulation	KEGG_2016
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell_Homo sapiens_R-HSA-198933	1	Immunomodulation	Reactome_2016
Prostanoid ligand receptors_Homo sapiens_R-HSA-391908	1	Immunomodulation	Reactome_2016
TNFs bind their physiological receptors_Homo sapiens_R-HSA-5669034	1	Immunomodulation	Reactome_2016
CD22 mediated BCR regulation_Homo sapiens_R-HSA-5690714	1	Immunomodulation	Reactome_2016
Chemokine receptors bind chemokines_Homo sapiens_R-HSA-380108	1	Immunomodulation	Reactome_2016
regulation of interleukin-6 production (GO:0032675)	1	Inflammation	GO_Biological_Process_2018
positive regulation of interleukin-1 beta secretion (GO:0050718)	1	Inflammation	GO_Biological_Process_2018
negative regulation of defense response (GO:0031348)	1	Inflammation	GO_Biological_Process_2018
neutrophil degranulation (GO:0043312)	1	Inflammation	GO_Biological_Process_2018
regulation of inflammatory response (GO:0050727)	1	Inflammation	GO_Biological_Process_2018
Proteasome_Homo sapiens_hsa03050	1	Cell_cycle	KEGG_2016

proteasome-activating ATPase activity (GO:0036402)	1	Cell_cycle	GO_Molecular_Function_2018
interleukin-1 receptor binding (GO:0005149)	1	Inflammation	GO_Molecular_Function_2018
Toll-like receptor binding (GO:0035325)	1	Inflammation	GO_Molecular_Function_2018
chemokine activity (GO:0008009)	1	Inflammation	GO_Molecular_Function_2018
Toll-like receptor signaling pathway_Homo sapiens_hsa04620	1	Inflammation	KEGG_2016
protein homodimerization activity (GO:0042803)	1	Immortalization	GO_Molecular_Function_2018
Pertussis_Homo sapiens_hsa05133	1	Inflammation	KEGG_2016
protein homodimerization activity (GO:0042803)	1	Oxidative_stress	GO_Molecular_Function_2018
Cytokine-cytokine receptor interaction_Homo sapiens_hsa04060	1	Inflammation	KEGG_2016
protein-lysine N-methyltransferase activity (GO:0016279)	1	Epigenetics	GO_Molecular_Function_2018
Proteoglycans in cancer_Homo sapiens_hsa05205	1	Proliferation	KEGG_2016
The NLRP3 inflammasome_Homo sapiens_R-HSA-844456	1	Inflammation	Reactome_2016
Chemokine receptors bind chemokines_Homo sapiens_R-HSA-380108	1	Inflammation	Reactome_2016
Platelet degranulation_Homo sapiens_R-HSA-114608	1	Inflammation	Reactome_2016
lipid hydroxylation (GO:0002933)	1	Liver_cancer_(ToxRefDB)	GO_Biological_Process_2018
fatty acid beta-oxidation using acyl-CoA oxidase (GO:0033540)	1	Liver_cancer_(ToxRefDB)	GO_Biological_Process_2018
omega-hydroxylase P450 pathway (GO:0097267)	1	Liver_cancer_(ToxRefDB)	GO_Biological_Process_2018
glucuronate metabolic process (GO:0019585)	1	Liver_cancer_(ToxRefDB)	GO_Biological_Process_2018
epoxygenase P450 pathway (GO:0019373)	1	Liver_cancer_(ToxRefDB)	GO_Biological_Process_2018
CoA hydrolase activity (GO:0016289)	1	Liver_cancer_(ToxRefDB)	GO_Molecular_Function_2018
glucuronosyltransferase activity (GO:0015020)	1	Liver_cancer_(ToxRefDB)	GO_Molecular_Function_2018
acyl-CoA oxidase activity (GO:0003997)	1	Liver_cancer_(ToxRefDB)	GO_Molecular_Function_2018

oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen (GO:0016712)	1	Liver_cancer_(ToxRefDB)	GO_Molecular_Function_2018
steroid hydroxylase activity (GO:0008395)	1	Liver_cancer_(ToxRefDB)	GO_Molecular_Function_2018
Rap1 signaling pathway_Homo sapiens_hsa04015	1	Angiogenesis	KEGG_2016
Ras signaling pathway_Homo sapiens_hsa04014	1	Angiogenesis	KEGG_2016
Metabolism of xenobiotics by cytochrome P450_Homo sapiens_hsa00980	1	Liver_cancer_(ToxRefDB)	KEGG_2016
Chemical carcinogenesis_Homo sapiens_hsa05204	1	Liver_cancer_(ToxRefDB)	KEGG_2016
Drug metabolism - cytochrome P450_Homo sapiens_hsa00982	1	Liver_cancer_(ToxRefDB)	KEGG_2016
regulation of hematopoietic progenitor cell differentiation (GO:1901532)	1	Apoptosis	GO_Biological_Process_2018
CYP2E1 reactions_Homo sapiens_R-HSA-211999	1	Liver_cancer_(ToxRefDB)	Reactome_2016
Recycling of bile acids and salts_Homo sapiens_R-HSA-159418	1	Liver_cancer_(ToxRefDB)	Reactome_2016
PPARA activates gene expression_Homo sapiens_R-HSA-1989781	1	Liver_cancer_(ToxRefDB)	Reactome_2016
Glucuronidation_Homo sapiens_R-HSA-156588	1	Liver_cancer_(ToxRefDB)	Reactome_2016
Regulation of RAS by GAPs_Homo sapiens_R-HSA-5658442	1	Angiogenesis	Reactome_2016
Resolution of Sister Chromatid Cohesion_Homo sapiens_R-HSA-2500257	1	Cell_cycle	Reactome_2016
Retinol metabolism_Homo sapiens_hsa00830	1	Liver_cancer_(ToxRefDB)	KEGG_2016
RHO GTPases Activate Formins_Homo sapiens_R-HSA-5663220	1	Cell_cycle	Reactome_2016

cellular response to oxidative stress (GO:0034599)	1	Oxidative_stress	GO_Biological_Process_2018
RHO GTPases Activate Formins_Homo sapiens_R-HSA-5663220	0	Genotoxicity	Reactome_2016
RNA polymerase II core promoter sequence-specific DNA binding (GO:0000979)	1	Immunomodulation	GO_Molecular_Function_2018
RNA polymerase II distal enhancer sequence-specific DNA binding (GO:0000980)	1	Epigenetics	GO_Molecular_Function_2018
glutathione transferase activity (GO:0004364)	1	Oxidative_stress	GO_Molecular_Function_2018
damaged DNA binding (GO:0003684)	1	Oxidative_stress	GO_Molecular_Function_2018
SCF-beta-TrCP mediated degradation of Emi1_Homo sapiens_R-HSA-174113	1	Apoptosis	Reactome_2016
Drug metabolism - cytochrome P450_Homo sapiens_hsa00982	1	Oxidative_stress	KEGG_2016
Chemical carcinogenesis_Homo sapiens_hsa05204	1	Oxidative_stress	KEGG_2016
Metabolism of xenobiotics by cytochrome P450_Homo sapiens_hsa00980	1	Oxidative_stress	KEGG_2016
Non-alcoholic fatty liver disease (NAFLD)_Homo sapiens_hsa04932	1	Oxidative_stress	KEGG_2016
Senescence-Associated Secretory Phenotype (SASP)_Homo sapiens_R-HSA-2559582	1	Oxidative_stress	Reactome_2016
Steroid hormone biosynthesis_Homo sapiens_hsa00140	1	Liver_cancer_(ToxRefDB)	KEGG_2016
Oxidative Stress Induced Senescence_Homo sapiens_R-HSA-2559580	1	Oxidative_stress	Reactome_2016
Systemic lupus erythematosus_Homo sapiens_hsa05322	1	Epigenetics	KEGG_2016
transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding (GO:0001077)	1	Cell_cycle	GO_Molecular_Function_2018
anaphase-promoting complex-dependent	1	Proliferation	GO_Biological_Process_2018

catabolic process (GO:0031145)			
cell cycle G2/M phase transition (GO:0044839)	1	Proliferation	GO_Biological_Process_2018
regulation of stem cell differentiation (GO:2000736)	1	Proliferation	GO_Biological_Process_2018
regulation of hematopoietic progenitor cell differentiation (GO:1901532)	1	Proliferation	GO_Biological_Process_2018
positive regulation of nucleic acid-templated transcription (GO:1903508)	1	Proliferation	GO_Biological_Process_2018
DNA-dependent ATPase activity (GO:0008094)	1	Proliferation	GO_Molecular_Function_2018
RNA polymerase II core promoter proximal region sequence-specific DNA binding (GO:0000978)	1	Proliferation	GO_Molecular_Function_2018
transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding (GO:0001077)	1	Proliferation	GO_Molecular_Function_2018
kinase binding (GO:0019900)	1	Proliferation	GO_Molecular_Function_2018
phosphatidylinositol-4,5-bisphosphate 3-kinase activity (GO:0046934)	1	Proliferation	GO_Molecular_Function_2018
Transcriptional regulation of white adipocyte differentiation_Homo sapiens_R-HSA-381340	1	Inflammation	Reactome_2016
Focal adhesion_Homo sapiens_hsa04510	1	Proliferation	KEGG_2016
Cell cycle_Homo sapiens_hsa04110	1	Proliferation	KEGG_2016
type II transforming growth factor beta receptor binding (GO:0005114)	1	Angiogenesis	GO_Molecular_Function_2018
Viral carcinogenesis_Homo sapiens_hsa05203	1	Epigenetics	KEGG_2016
Processing of DNA double-strand break ends_Homo sapiens_R-HSA-5693607	1	Proliferation	Reactome_2016
Autodegradation of Cdh1 by Cdh1:APC/C_Homo sapiens_R-HSA-174084	1	Proliferation	Reactome_2016

Oxidative Stress Induced Senescence_Homo sapiens_R-HSA-2559580	1	Proliferation	Reactome_2016
Viral carcinogenesis_Homo sapiens_hsa05203	1	Oxidative_stress	KEGG_2016
Senescence-Associated Secretory Phenotype (SASP)_Homo sapiens_R-HSA-2559582	1	Proliferation	Reactome_2016
sodium-independent organic anion transport (GO:0043252)	1	Steroid_hormones	GO_Biological_Process_2018
progesterone metabolic process (GO:0042448)	1	Steroid_hormones	GO_Biological_Process_2018
C21-steroid hormone metabolic process (GO:0008207)	1	Steroid_hormones	GO_Biological_Process_2018
flavonoid glucuronidation (GO:0052696)	1	Steroid_hormones	GO_Biological_Process_2018
glucuronate metabolic process (GO:0019585)	1	Steroid_hormones	GO_Biological_Process_2018
aryl sulfotransferase activity (GO:0004062)	1	Steroid_hormones	GO_Molecular_Function_2018
inorganic anion exchanger activity (GO:0005452)	0	Steroid_hormones	GO_Molecular_Function_2018
ketosteroid monooxygenase activity (GO:0047086)	1	Steroid_hormones	GO_Molecular_Function_2018
glucuronosyltransferase activity (GO:0015020)	1	Steroid_hormones	GO_Molecular_Function_2018
sodium-independent organic anion transmembrane transporter activity (GO:0015347)	1	Steroid_hormones	GO_Molecular_Function_2018
Ascorbate and aldarate metabolism_Homo sapiens_hsa00053	1	Steroid_hormones	KEGG_2016
Drug metabolism - cytochrome P450_Homo sapiens_hsa00982	1	Steroid_hormones	KEGG_2016
Chemical carcinogenesis_Homo sapiens_hsa05204	1	Steroid_hormones	KEGG_2016
Metabolism of xenobiotics by cytochrome P450_Homo sapiens_hsa00980	1	Steroid_hormones	KEGG_2016
Steroid hormone biosynthesis_Homo sapiens_hsa00140	1	Steroid_hormones	KEGG_2016
Common Pathway of Fibrin Clot	1	Steroid_hormones	Reactome_2016

Formation_Homo sapiens_R-HSA-140875			
Hormone ligand-binding receptors_Homo sapiens_R-HSA-375281	1	Steroid_hormones	Reactome_2016
Transport of organic anions_Homo sapiens_R-HSA-879518	1	Steroid_hormones	Reactome_2016
Androgen biosynthesis_Homo sapiens_R-HSA-193048	1	Steroid_hormones	Reactome_2016
Glucuronidation_Homo sapiens_R-HSA-156588	1	Steroid_hormones	Reactome_2016
flavonoid glucuronidation (GO:0052696)	1	Xenobiotic_metabolism	GO_Biological_Process_2018
glucuronate metabolic process (GO:0019585)	1	Xenobiotic_metabolism	GO_Biological_Process_2018
glycosaminoglycan metabolic process (GO:0030203)	1	Xenobiotic_metabolism	GO_Biological_Process_2018
aminoglycan metabolic process (GO:0006022)	1	Xenobiotic_metabolism	GO_Biological_Process_2018
epoxygenase P450 pathway (GO:0019373)	1	Xenobiotic_metabolism	GO_Biological_Process_2018
steroid hydroxylase activity (GO:0008395)	1	Xenobiotic_metabolism	GO_Molecular_Function_2018
aryl sulfotransferase activity (GO:0004062)	1	Xenobiotic_metabolism	GO_Molecular_Function_2018
glucuronosyltransferase activity (GO:0015020)	1	Xenobiotic_metabolism	GO_Molecular_Function_2018
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen (GO:0016712)	1	Xenobiotic_metabolism	GO_Molecular_Function_2018
glutathione transferase activity (GO:0004364)	1	Xenobiotic_metabolism	GO_Molecular_Function_2018
Retinol metabolism_Homo sapiens_hsa00830	1	Xenobiotic_metabolism	KEGG_2016
Steroid hormone biosynthesis_Homo sapiens_hsa00140	1	Xenobiotic_metabolism	KEGG_2016
Metabolism of xenobiotics by cytochrome P450_Homo sapiens_hsa00980	1	Xenobiotic_metabolism	KEGG_2016
Drug metabolism - cytochrome P450_Homo sapiens_hsa00982	1	Xenobiotic_metabolism	KEGG_2016

Chemical carcinogenesis_Homo sapiens_hsa05204	1	Xenobiotic_metabolism	KEGG_2016
CYP2E1 reactions_Homo sapiens_R-HSA-211999	1	Xenobiotic_metabolism	Reactome_2016
Miscellaneous substrates_Homo sapiens_R-HSA-211958	1	Xenobiotic_metabolism	Reactome_2016
Synthesis of Leukotrienes (LT) and Eoxins (EX)_Homo sapiens_R-HSA-2142691	1	Xenobiotic_metabolism	Reactome_2016
Fatty acids_Homo sapiens_R-HSA-211935	1	Xenobiotic_metabolism	Reactome_2016
Glucuronidation_Homo sapiens_R-HSA-156588	1	Xenobiotic_metabolism	Reactome_2016

REFERENCES

1. Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem Res Toxicol [Internet]. 2016;29(8):1225–51. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27367298

2. Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, et al. The toxicity data landscape for environmental chemicals. Env Heal Perspect [Internet]. 2009;117(5):685–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19479008

3. Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, et al. The exposure data landscape for manufactured chemicals. Sci Total Env [Internet]. 2012;414:159–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22104386

4. Shimkus JM. Frank R. Lautenberg Chemical Safety for the 21st Century Act [Internet]. United States Statutes at Large. 2014. p. 448–513. Available from: https://www.congress.gov/114/plaws/publ182/PLAW-114publ182.pdf

5. National Research Council. Toxicity Testing in the 21st Century: A Vision and a Strategy [Internet]. Washington, DC: The National Academies Press; 2007. 216 p. Available from: https://www.nap.edu/catalog/11970/toxicity-testing-in-the-21st-century-a-vision-and-a

6. ECHA. New Approach Methodologies in Regulatory Science. In Helsinki; 2016. Available from: https://echa.europa.eu/documents/10162/22816069/scientific_ws_proceedings_en.pdf

7. OCSPP U, USEPA. Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program [Internet]. 2018. Available from: https://www.epa.gov/sites/production/files/2018-06/documents/epa_alt_strat_plan_6-20-18_clean_final.pdf

8. Thomas RS, Paules RS, Simeonov A, Fitzpatrick SC, Crofton KM, Casey WM, et al. The US Federal Tox21 Program: A strategic and operational plan for continued leadership. ALTEX [Internet]. 2018;35(2):163–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29529324

9. Tice RR, Austin CP, Kavlock RJ, Bucher JR. Improving the human hazard characterization of chemicals: a Tox21 update. Env Heal Perspect [Internet]. 2013;121(7):756–65. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23603828

10. Using 21st Century Science to Improve Risk-Related Evaluations [Internet]. Washington (DC); 2017. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28267305

11. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci [Internet]. 2007;95(1):5–12. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16963515

12. Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, et al. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Env Heal Perspect [Internet]. 2010;118(4):485–92. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20368123

13. Watford S, Pham L, Wignall J, Shin R, Martin M, Paul-Friedman K. ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses. Prep.

14. Judson RS, Martin MT, Egeghy P, Gangwal S, Reif DM, Kothiya P, et al. Aggregating data for computational toxicology applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System. Int J Mol Sci [Internet]. 2012;13(2):1805–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22408426

15. Pearce RG, Setzer RW, Strope CL, Wambaugh JF, Sipes NS. httk: R Package for High-Throughput Toxicokinetics. J Stat Softw [Internet]. 2017;79(4):1–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30220889

16. Filer DL, Kothiya P, Setzer RW, Judson RS, Martin MT. tcpl: the ToxCast pipeline for high-throughput screening data. Bioinformatics [Internet]. 2017;33(4):618–20. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27797781

17. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminform [Internet]. 2017;9(1):61. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29185060

18. Bioplanet [Internet]. [cited 2018 Dec 28]. Available from: https://tripod.nih.gov/bioplanet/

19. Tox21 Toolbox [Internet]. Vol. 2018. Available from: https://ntp.niehs.nih.gov/results/tox21/tbox/

20. Bell SM, Phillips J, Sedykh A, Tandon A, Sprankle C, Morefield SQ, et al. An Integrated Chemical Environment to Support 21st-Century Toxicology. Environ Health Perspect [Internet]. National Institute of Environmental Health Science; 2017 [cited 2019 Jan 15];125(5):054501. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28557712

21. Kleinstreuer NC, Ceger P, Watt ED, Martin M, Houck K, Browne P, et al. Development and Validation of a Computational Model for Androgen Receptor Activity. Chem Res Toxicol [Internet]. 2017;30(4):946–64. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27933809

22. Haggard DE, Karmaus AL, Martin MT, Judson RS, Setzer RW, Paul Friedman K. High-Throughput H295R Steroidogenesis Assay: Utility as an Alternative and a Statistical Approach to Characterize Effects on Steroidogenesis. Toxicol Sci [Internet]. 2018;162(2):509–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29216406

23. Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, et al. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. Toxicol Sci [Internet]. 2015;148(1):137–54. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26272952

24. Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, et al. Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. Chem Res Toxicol [Internet]. 2015;28(4):738–51. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25697799

25. Kleinstreuer NC, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, et al. In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. Toxicol Sci [Internet]. 2013;131(1):40–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23024176

26. Kleinstreuer N, Dix D, Rountree M, Baker N, Sipes N, Reif D, et al. A computational model predicting disruption of blood vessel development. PLoS Comput Biol [Internet]. 2013;9(4):e1002996. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23592958

27. Sipes NS, Martin MT, Reif DM, Kleinstreuer NC, Judson RS, Singh A V, et al. Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. Toxicol Sci [Internet]. 2011;124(1):109–27. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21873373

28. Martin MT, Mendez E, Corum DG, Judson RS, Kavlock RJ, Rotroff DM, et al. Profiling the reproductive toxicity of chemicals from multigeneration studies in the toxicity reference database. Toxicol Sci [Internet]. 2009;110(1):181–90. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19363143

29. Martin MT, Judson RS, Reif DM, Kavlock RJ, Dix DJ. Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. Env Heal Perspect [Internet]. 2009;117(3):392–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19337514

30. Pigliucci M. Genotype-phenotype mapping and the end of the “genes as blueprint” metaphor. Philos Trans R Soc L B Biol Sci [Internet]. 2010;365(1540):557–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20083632

31. Lehner B. Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet [Internet]. 2013;14(3):168–78. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23358379

32. Watford SM, Grashow RG, De La Rosa VY, Rudel RA, Friedman KP, Martin MT. Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis. Comput Toxicol [Internet]. 2018; Available from: http://www.sciencedirect.com/science/article/pii/S246811131830029X

33. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell [Internet]. 2017;171(6):1437–1452 e17. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29195078

34. Oki NO, Edwards SW. An integrative data mining approach to identifying adverse outcome pathway signatures. Toxicology [Internet]. 2016;350–352:49–61. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27108252

35. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res [Internet]. 2017;45(D1):D972–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27651457

36. Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. Systematic review and evidence integration for literature-based environmental health science assessments. Env Heal Perspect [Internet]. 2014;122(7):711–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24755067

37. Kavlock RJ, Bahadori T, Barton-Maclaren TS, Gwinn MR, Rasenberg M, Thomas RS. Accelerating the Pace of Chemical Risk Assessment. Chem Res Toxicol [Internet]. 2018;31(5):287–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29600706

38. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data [Internet]. The Author(s); 2016 Mar 15;3:160018. Available from: https://doi.org/10.1038/sdata.2016.18

39. Chen M, Mao S, Liu Y. Big Data: A Survey. Mob Networks Appl [Internet]. 2014;19(2):171–209. Available from: https://doi.org/10.1007/s11036-013-0489-0

40. Raja K, Patrick M, Gao Y, Madu D, Yang Y, Tsoi LC. A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. Int J Genomics [Internet]. 2017;2017:6213474. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28331849

41. Jagodnik KM, Koplev S, Jenkins SL, Ohno-Machado L, Paten B, Schurer SC, et al. Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop. J Biomed Inf [Internet]. 2017;71:49–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28501646

42. Bui AAT, Van Horn JD, Consortium NBKC. Envisioning the future of "big data" biomedicine. J Biomed Inf [Internet]. 2017;69:115–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28366789

43. Sagiroglu S, Sinanc D. Big data: A review. In: 2013 International Conference on Collaboration Technologies and Systems (CTS). 2013. p. 42–7.

44. Begoli E, Horey J. Design Principles for Effective Knowledge Discovery from Big Data. In: 2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture. 2012. p. 215–8.

45. Affymetrix Standards Program [Internet]. [cited 2019 Jan 15]. Available from: http://www.affymetrix.com/about_affymetrix/outreach/standards_program/standards-program.affx

46. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res [Internet]. 2002 Jan 1 [cited 2019 Feb 6];30(1):207–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11752295

47. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res [Internet]. 2012 Nov 26 [cited 2019 Feb 6];41(D1):D991–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23193258

48. Merrick BA, Paules RS, Tice RR. Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective. Int J Biotechnol [Internet]. 2015;14(1):7–27. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27122658

49. Harrill J. Development and Use of a High Content Imaging-Based Phenotypic Profiling Assay for Chemical Bioactivity Screening. 2018; Available from: https://epa.figshare.com/articles/Development_and_Use_of_a_High_Content_Imaging-Based_Phenotypic_Profiling_Assay_for_Chemical_Bioactivity_Screening/7003181

50. Shah I, Setzer RW, Jack J, Houck KA, Judson RS, Knudsen TB, et al. Using ToxCast^{TM} Data to Reconstruct Dynamic Cell State Trajectories and Estimate Toxicological Points of Departure. Environ Health Perspect [Internet]. 2016 Jul [cited 2019 Jan 16];124(7):910–9. Available from: https://ehp.niehs.nih.gov/doi/10.1289/ehp.1409029

51. NIH STRATEGIC PLAN FOR DATA SCIENCE. [cited 2019 Jan 15]; Available from: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

52. Mahony C, Currie R, Daston G, Kleinstreuer N, van de Water B. Highlight report: 'Big data in the 3R's: outlook and recommendations', a roundtable summary. Arch Toxicol [Internet]. Springer Berlin Heidelberg; 2018 Feb 16 [cited 2019 Jan 24];92(2):1015–20. Available from: http://link.springer.com/10.1007/s00204-017-2145-0

53. Judson RS, Kavlock RJ, Setzer RW, Cohen Hubal EA, Martin MT, Knudsen TB, et al. Estimating Toxicity-Related Biological Pathway Altering Doses for High-Throughput Chemical Risk Assessment. Chem Res Toxicol [Internet]. 2011 Apr 18 [cited 2019 Jan 24];24(4):451–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21384849

54. Judson R, Houck K, Martin M, Knudsen T, Thomas RS, Sipes N, et al. In vitro and modelling approaches to risk assessment from the U.S. Environmental Protection Agency ToxCast programme. Basic Clin Pharmacol Toxicol [Internet]. 2014;115(1):69–76. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24684691

55. Barton-Maclaren T, Gwinn M, Thomas R, Kavlock R, Rasenberg M. INSIGHT: New Approaches to Chemical Assessment—a Progress Report [Internet]. 2019 [cited 2019 Jan 24]. Available from: https://news.bloombergenvironment.com/environment-and-energy/insight-new-approaches-to-chemical-assessmenta-progress-report

56. Fitzpatrick JM, Patlewicz G. Application of IATA – A case study in evaluating the global and local performance of a Bayesian network model for skin sensitization. SAR QSAR Environ Res [Internet]. 2017 Apr 3 [cited 2019 Jan 24];28(4):297–310. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28423913

57. Fitzpatrick JM, Roberts DW, Patlewicz G. An evaluation of selected (Q)SARs/expert systems for predicting skin sensitisation potential. SAR QSAR Environ Res [Internet]. 2018 Jun 3 [cited 2019 Jan 24];29(6):439–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29676182

58. Roberts DW, Patlewicz G. Non-animal assessment of skin sensitization hazard: Is an integrated testing strategy needed, and if so what should be integrated? J Appl Toxicol [Internet]. 2018 Jan [cited 2019 Jan 24];38(1):41–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28543848

59. US EPA O. Integrated Risk Information System. [cited 2019 Jan 24]; Available from: https://www.epa.gov/iris

60. NCBI. PubMed [Internet]. [cited 2019 Jan 24]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/

61. Lea IA, Gong H, Paleja A, Rashid A, Fostel J. CEBS: a comprehensive annotated database of toxicological data. Nucleic Acids Res [Internet]. 2017;45(D1):D964–71. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27899660

62. eChemPortal - Substance Search [Internet]. [cited 2019 Feb 6]. Available from: https://www.echemportal.org/echemportal/substancesearch/substancesearchlink.action

63. US EPA O. Provisional Peer-Reviewed Toxicity Values (PPRTVs). [cited 2019 Feb 6]; Available from: https://www.epa.gov/pprtv

64. The Carcinogenic Potency Project (CPDB) [Internet]. [cited 2019 Feb 6]. Available from: https://toxnet.nlm.nih.gov/cpdb/index.html

65. ACToR Web Services [Internet]. [cited 2019 Jan 21]. Available from: https://actorws.epa.gov/actorws/

66. Reif DM, Martin MT, Tan SW, Houck KA, Judson RS, Richard AM, et al. Endocrine Profiling and Prioritization of Environmental Chemicals Using ToxCast Data. Environ Health Perspect [Internet]. National Institute of Environmental Health Science; 2010 Dec [cited 2019 Jan 24];118(12):1714–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20826373

67. Rusyn I, Sedykh A, Low Y, Guyton KZ, Tropsha A. Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data. Toxicol Sci [Internet]. 2012;127(1):1–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22387746

68. Dionisio KL, Phillips K, Price PS, Grulke CM, Williams A, Biryol D, et al. The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. Sci Data [Internet]. Nature Publishing Group; 2018 Jul 10 [cited 2019 Jan 24];5:180125. Available from: http://www.nature.com/articles/sdata2018125

69. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res [Internet]. 2019 Jan 8 [cited 2019 Feb 7];47(D1):D1102–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30371825

70. Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, et al. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Env Toxicol Chem [Internet]. 2010;29(3):730–41. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20821501

71. Brockmeier EK, Hodges G, Hutchinson TH, Butler E, Hecker M, Tollefsen KE, et al. The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment. Toxicol Sci [Internet]. Oxford University Press; 2017 [cited 2019 Feb 7];158(2):252–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28525648

72. Kleinstreuer NC, Sullivan K, Allen D, Edwards S, Mendrick DL, Embry M, et al. Adverse outcome pathways: From research to regulation scientific workshop report. Regul Toxicol Pharmacol [Internet]. Academic Press; 2016 Apr 1 [cited 2019 Feb 7];76:39–50. Available from: https://www.sciencedirect.com/science/article/pii/S0273230016300071?via%3Dihub

73. Wittwehr C, Aladjov H, Ankley G, Byrne HJ, de Knecht J, Heinzle E, et al. How Adverse Outcome Pathways Can Aid the Development and Use of Computational Prediction Models for Regulatory Toxicology. Toxicol Sci [Internet]. Oxford University Press; 2017 [cited 2019 Feb 7];155(2):326–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27994170

74. Tollefsen KE, Scholz S, Cronin MT, Edwards SW, de Knecht J, Crofton K, et al. Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). Regul Toxicol Pharmacol [Internet]. Academic Press; 2014 Dec 1 [cited 2019 Feb 7];70(3):629–40. Available from: https://www.sciencedirect.com/science/article/pii/S0273230014002141?via%3Dihub

75. Juberg DR, Knudsen TB, Sander M, Beck NB, Faustman EM, Mendrick DL, et al. FutureTox III: Bridges for Translation. Toxicol Sci [Internet]. Oxford University Press; 2017 [cited 2019 Feb 7];155(1):22–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27780885

76. Buesen R, Chorley BN, da Silva Lima B, Daston G, Deferme L, Ebbels T, et al. Applying ’omics technologies in chemicals risk assessment: Report of an ECETOC workshop. Regul Toxicol Pharmacol [Internet]. Academic Press; 2017 Dec 1 [cited 2019 Feb 7];91:S3–13. Available from: https://www.sciencedirect.com/science/article/pii/S027323001730274X?via%3Dihub

77. Adverse Outcome Pathways, Molecular Screening and Toxicogenomics - OECD [Internet]. [cited 2019 Jan 21]. Available from: http://www.oecd.org/chemicalsafety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm

78. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res [Internet]. 2004;32(Database issue):D267-70. Available from: https://www.ncbi.nlm.nih.gov/pubmed/14681409

79. Cesta MF, Malarkey DE, Herbert RA, Brix A, Hamlin 2nd MH, Singletary E, et al. The National Toxicology Program Web-based nonneoplastic lesion atlas: a global toxicology and pathology resource. Toxicol Pathol [Internet]. 2014;42(2):458–60. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24488020

80. Briggs K, Barber C, Cases M, Marc P, Steger-Hartmann T. Value of shared preclinical safety studies - The eTOX database. Toxicol Rep [Internet]. 2015;2:210–21. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28962354

81. Ravagli C, Pognan F, Marc P. OntoBrowser: a collaborative tool for curation of ontologies by subject matter experts. Bioinformatics [Internet]. Oxford University Press; 2017 Jan 1 [cited 2019 Jan 31];33(1):148–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw579

82. Shapiro AJ, Antoni S, Guyton KZ, Lunn RM, Loomis D, Rusyn I, et al. Software Tools to Facilitate Systematic Review Used for Cancer Hazard Identification. [cited 2019 Jan 31]; Available from: https://doi.org/10.1289/EHP4224.

83. Heidorn CJ, Rasmussen K, Hansen BG, Norager O, Allanou R, Seynaeve R, et al. IUCLID: an information management tool for existing chemicals and biocides. J Chem Inf Comput Sci [Internet]. 2003;43(3):779–86. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12767136

84. IUCLID format - IUCLID [Internet]. [cited 2019 Jan 31]. Available from: https://iuclid6.echa.europa.eu/format

85. Anthony Tony Cox L, Popken DA, Kaplan AM, Plunkett LM, Becker RA. How well can in vitro data predict in vivo effects of chemicals? Rodent carcinogenicity as a case study. Regul Toxicol Pharmacol [Internet]. 2016;77:54–64. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26879462

86. Chiu WA, Guyton KZ, Martin MT, Reif DM, Rusyn I. Use of high-throughput in vitro toxicity screening data in cancer hazard evaluations by IARC Monograph Working Groups. ALTEX [Internet]. 2018;35(1):51–64. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28738424

87. Guyton KZ, Rusyn I, Chiu WA, Corpet DE, van den Berg M, Ross MK, et al. Application of the key characteristics of carcinogens in cancer hazard identification. Carcinogenesis [Internet]. 2018;39(4):614–22. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29562322

88. Smith MT, Guyton KZ, Gibbons CF, Fritz JM, Portier CJ, Rusyn I, et al. Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. Env Heal Perspect [Internet]. 2016;124(6):713–21. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26600562

89. Becker RA, Dreier DA, Manibusan MK, Cox LAT, Simon TW, Bus JS. How well can carcinogenicity be predicted by high throughput "characteristics of carcinogens" mechanistic data? Regul Toxicol Pharmacol [Internet]. 2017;90:185–96. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28866267

90. Grashow RG, De La Rosa VY, Watford SM, Ackerman JM, Rudel RA. BCScreen: A gene panel to test for breast carcinogenesis in chemical safety screening. Comput Toxicol [Internet]. 2018;5(Supplement C):16–24. Available from: http://www.sciencedirect.com/science/article/pii/S2468111317300580

91. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A [Internet]. 2005;102(43):15545–50. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16199517

92. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics [Internet]. Oxford University Press; 2011 Jun 15 [cited 2019 Jan 24];27(12):1739–40. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr260

93. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst [Internet]. Cell Press; 2015 Dec 23 [cited 2019 Jan 24];1(6):417–25. Available from: https://www.sciencedirect.com/science/article/pii/S2405471215002185

94. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles G V, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics [Internet]. 2013;14:128. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23586463

95. Kuleshov M V, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res [Internet]. 2016;44(W1):W90-7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27141961

96. Online Mendelian Inheritance in Man, OMIM® [Internet]. Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; Available from: https://omim.org/

97. King BL, Davis AP, Rosenstein MC, Wiegers TC, Mattingly CJ. Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. PLoS One [Internet]. 2012;7(11):e46524. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23144783

98. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics [Internet]. 2005;6 Suppl 1:S1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15960821

99. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, et al. Overview of BioCreative II gene mention recognition. Genome Biol [Internet]. 2008;9 Suppl 2:S2. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18834493

100. Vempati UD, Przydzial MJ, Chung C, Abeyruwan S, Mir A, Sakurai K, et al. Formalization, Annotation and Analysis of Diverse Drug and Probe Screening Assay Datasets Using the BioAssay Ontology (BAO). PLoS One [Internet]. Public Library of Science; 2012 [cited 2019 Feb 7];7(11):e49198. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049198

101. Abeyruwan S, Vempati UD, Küçük-McGinty H, Visser U, Koleti A, Mir A, et al. Evolving BioAssay Ontology (BAO): modularization, integration and applications. J Biomed Semantics [Internet]. 2014 [cited 2019 Feb 7];5(Suppl 1):S5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25093074

102. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res [Internet]. 2015;43(Database issue):D36-42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25355515

103. Pundir S, Martin MJ, O’Donovan C. UniProt Protein Knowledgebase. Methods Mol Biol [Internet]. 2017;1558:41–55. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28150232

104. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics [Internet]. 2015;31(6):926–32. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25398609

105. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res [Internet]. 2000;28(1):27–30. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10592173

106. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res [Internet]. 2016;44(D1):D457-62. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26476454

107. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res [Internet]. 2017;45(D1):D353–61. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27899662

108. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res [Internet]. 2016;44(D1):D481-7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26656494

109. Online Mendelian Inheritance in Man, OMIM® [Internet]. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). [cited 2019 Jan 30]. Available from: https://www.omim.org/

110. Judson R, Thomas RS, Baker N, Simha A, Howey XM, Marable C, et al. Workflow for Defining Reference Chemicals for Assessing Performance of In Vitro Assays. ALTEX [Internet]. 2018 Dec 17 [cited 2019 Jan 31]; Available from: https://www.altex.org/index.php/altex/article/view/1168

111. Fridley BL, Jenkins GD, Biernacka JM. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. PLoS One [Internet]. 2010;5(9). Available from: https://www.ncbi.nlm.nih.gov/pubmed/20862301

112. Papatheodorou I, Oellrich A, Smedley D. Linking gene expression to phenotypes via pathway information. J Biomed Semant [Internet]. 2015;6:17. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25901272

113. Nelson ER, Wardell SE, Jasper JS, Park S, Suchindran S, Howe MK, et al. 27-Hydroxycholesterol links hypercholesterolemia and breast cancer pathophysiology. Science (80-) [Internet]. 2013;342(6162):1094–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24288332

114. Hoover RN, Hyer M, Pfeiffer RM, Adam E, Bond B, Cheville AL, et al. Adverse health outcomes in women exposed in utero to diethylstilbestrol. N Engl J Med [Internet]. 2011;365(14):1304–14. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21991952

115. Hamajima N, Hirose K, Tajima K, Rohan T, Calle EE, Heath Jr. CW, et al. Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. Br J Cancer [Internet]. 2002;87(11):1234–45. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12439712

116. Chlebowski RT, Manson JE, Anderson GL, Cauley JA, Aragaki AK, Stefanick ML, et al. Estrogen plus progestin and breast cancer incidence and mortality in the Women’s Health Initiative Observational Study. J Natl Cancer Inst [Internet]. 2013;105(8):526–35. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23543779

117. Shioda T, Rosenthal NF, Coser KR, Suto M, Phatak M, Medvedovic M, et al. Expressomal approach for comprehensive analysis and visualization of ligand sensitivities of xenoestrogen responsive genes. Proc Natl Acad Sci U S A [Internet]. 2013;110(41):16508–13. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24062438

118. Schwarzman MR, Ackerman JM, Dairkee SH, Fenton SE, Johnson D, Navarro KM, et al. Screening for Chemical Contributions to Breast Cancer Risk: A Case Study for Chemical Safety Evaluation. Env Heal Perspect [Internet]. 2015;123(12):1255–64. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26032647

119. Pirone JR, D’Arcy M, Stewart DA, Hines WC, Johnson M, Gould MN, et al. Age-associated gene expression in normal breast tissue mirrors qualitative age-at-incidence patterns for breast cancer. Cancer Epidemiol Biomarkers Prev [Internet]. 2012;21(10):1735–44. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22859400

120. Parada Jr. H, Sun X, Fleming JM, Williams-DeVane CR, Kirk EL, Olsson LT, et al. Race-associated biological differences among luminal A and basal-like breast cancers in the Carolina Breast Cancer Study. Breast Cancer Res [Internet]. 2017;19(1):131. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29228969

121. Wu Y, Ding Y, Tanaka Y, Zhang W. Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. Int J Med Sci [Internet]. 2014;11(11):1185–200. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25249787

122. Gwinn MR, Axelrad DA, Bahadori T, Bussard D, Cascio WE, Deener K, et al. Chemical Risk Assessment: Traditional vs Public Health Perspectives. Am J Public Heal [Internet]. 2017;107(7):1032–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28520487

123. Perkins EJ, Antczak P, Burgoon L, Falciani F, Garcia-Reyero N, Gutsell S, et al. Adverse Outcome Pathways for Regulatory Applications: Examination of Four Case Studies With Different Degrees of Completeness and Scientific Confidence. Toxicol Sci [Internet]. 2015;148(1):14–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26500288

124. IBCERCC. Breast Cancer and the Environment: Prioritizing Prevention. In IBCERCC (Interagency Breast Cancer and Environmental Research Coordinating Committee); 2013. Available from: https://www.niehs.nih.gov/about/assets/docs/breast_cancer_and_the_environment_prioritizing_prevention_508.pdf

125. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, et al. Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. Chem Res Toxicol [Internet]. 2012;25(7):1287–302. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22519603

126. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Env Heal Perspect [Internet]. 2016;124(7):1023–33. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26908244

127. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res [Internet]. 2015;43(Database issue):D789-98. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25428349

128. Merrick BA, Paules RS, Tice RR. Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective. Int J Biotechnol [Internet]. NIH Public Access; 2015 [cited 2019 Jan 15];14(1):7–27. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27122658

129. Mav D, Shah RR, Howard BE, Auerbach SS, Bushel PR, Collins JB, et al. A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. PLoS One [Internet]. 2018;13(2):e0191105. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29462216

130. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res [Internet]. 2017; Available from: https://www.ncbi.nlm.nih.gov/pubmed/29140470

131. NLM. Chapter 11 Relationships in Medical Subject Headings. 2016.

132. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res [Internet]. 2017;45(D1):D158–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27899622

133. NLM. Gene [Internet]. 2018 [cited 2018 Jan 5]. Available from: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz

134. NLM. Gene Reference into Function (GeneRIF) [Internet]. 2017 [cited 2017 Nov 30]. Available from: https://www.ncbi.nlm.nih.gov/gene/about-generif

135. CTD. CTD downloads [Internet]. 2017 [cited 2017 Nov 30]. Available from: http://ctdbase.org/downloads/

136. UniProt. UniProt downloads [Internet]. 2017 [cited 2017 Nov 30]. Available from: http://www.uniprot.org/downloads

137. Reactome. Reactome downloads [Internet]. 2018 [cited 2018 Jan 5]. Available from: https://reactome.org/

138. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res [Internet]. 2017;46(D1):D649–55. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29145629

139. RGD. RGD downloads [Internet]. 2018 [cited 2018 Nov 30]. Available from: ftp://ftp.rgd.mcw.edu/pub/data_release/

140. Shimoyama M, De Pons J, Hayman GT, Laulederkind SJ, Liu W, Nigam R, et al. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Res [Internet]. 2015;43(Database issue):D743-50. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25355511

141. MGI. MGI download [Internet]. 2017 [cited 2017 Nov 30]. Available from: http://www.informatics.jax.org/downloads/reports/index.html

142. Eppig JT, Smith CL, Blake JA, Ringwald M, Kadin JA, Richardson JE, et al. Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. Methods Mol Biol [Internet]. 2017;1488:47–73. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27933520

143. Bouma G. Normalized (pointwise) mutual information in collocation extraction. Proc Int Conf Ger Soc Comput Linguist Lang Technol. 2009;

144. Pfeffer U, Fecarotta E, Castagnetta L, Vidali G. Estrogen receptor variant messenger RNA lacking exon 4 in estrogen-responsive human breast cancer cell lines. Cancer Res [Internet]. 1993;53(4):741–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7916651

145. Kang HY, Yeh S, Fujimoto N, Chang C. Cloning and characterization of human prostate coactivator ARA54, a novel protein that associates with the androgen receptor. J Biol Chem [Internet]. 1999;274(13):8570–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10085091

146. Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Med Res Methodol [Internet]. 2009;9:79. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19941636

147. Goodson 3rd WH, Lowe L, Carpenter DO, Gilbertson M, Manaf Ali A, Lopez de Cerain Salsamendi A, et al. Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: the challenge ahead. Carcinogenesis [Internet]. 2015;36 Suppl 1:S254-96. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26106142

148. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell [Internet]. 2011;144(5):646–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21376230

149. Bastien RR, Rodriguez-Lescure A, Ebbert MT, Prat A, Munarriz B, Rowe L, et al. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC Med Genomics [Internet]. 2012;5:44. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23035882

150. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature [Internet]. 2012;490(7418):61–70. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23000897

151. Latimer JJ, Johnson JM, Kelly CM, Miles TD, Beaudry-Rodgers KA, Lalanne NA, et al. Nucleotide excision repair deficiency is intrinsic in sporadic stage I breast cancer. Proc Natl Acad Sci U S A [Internet]. 2010;107(50):21725–30. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21118987

152. Chen L, Xiao Z, Meng Y, Zhao Y, Han J, Su G, et al. The enhancement of cancer stem cell properties of MCF-7 cells in 3D collagen scaffolds for modeling of cancer and anti-cancer drugs. Biomaterials [Internet]. 2012;33(5):1437–44. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22078807

153. Chua SL, See Too WC, Khoo BY, Few LL. UBC and YWHAZ as suitable reference genes for accurate normalisation of gene expression using MCF7, HCT116 and HepG2 cell lines. Cytotechnology [Internet]. 2011;63(6):645–54. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21850463

154. Jacob MD, Audas TE, Uniacke J, Trinkle-Mulcahy L, Lee S. Environmental cues induce a long noncoding RNA-dependent remodeling of the nucleolus. Mol Biol Cell [Internet]. 2013;24(18):2943–53. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23904269

155. Fontaine JF, Priller F, Barbosa-Silva A, Andrade-Navarro MA. Genie: literature-based gene prioritization at multi genomic scale. Nucleic Acids Res [Internet]. 2011;39(Web Server issue):W455-61. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21609954

156. Guido. Python tutorial, Technical Report CS-R9526. 1995.

157. MongoDB [Internet]. Available from: https://www.mongodb.com/

158. pandas: Python Data Analysis Library [Internet]. 2012. Available from: http://pandas.pydata.org/

159. Van Der Walt S, Colbert C, Varoquaux G. The NumPy array: a structure for efficient numerical computation. 2011; Available from: http://arxiv.org/abs/1102.1523

160. Lam S, Pitrou A, Seibert S. Numba: A LLVM-based Python JIT Compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC [Internet]. Austin, Texas: ACM; 2015. Available from: http://dx.doi.org/10.1145/2833157.2833162

161. Pérez F, Granger B. IPython: A System for Interactive Scientific Computing. Comput Sci Eng [Internet]. IEEE; 2007;9(3):21–9. Available from: http://dx.doi.org/10.1109/mcse.2007.53

162. Homologene [Internet]. Vol. 2017. Available from: https://www.ncbi.nlm.nih.gov/homologene

163. Pecina P. Lexical association measures and collocation extraction. Lang Resour Eval [Internet]. 2010;44(1):137–58. Available from: https://doi.org/10.1007/s10579-009-9101-4

164. Ade AS, Wright ZC, States DJ. Gene2MeSH [Internet]. Ann Arbor (MI): National Center for Integrative Biomedical Informatics; 2007. Available from: http://gene2mesh.ncibi.org

165. Cheung WA, Ouellette BFF, Wasserman WW. Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs). BMC Bioinformatics [Internet]. 2012;13(1):249. Available from: http://dx.doi.org/10.1186/1471-2105-13-249

166. Pearce D. A Comparative Evaluation of Collocation Extraction Techniques [Internet]. International Conference on Language Resources and Evaluation. 2002. p. 1530–6. Available from: http://www.lrec-conf.org/proceedings/lrec2002/

167. Kastrin A, Rindflesch TC, Hristovski D. Large-scale structure of a network of co-occurring MeSH terms: statistical analysis of macroscopic properties. PLoS One [Internet]. 2014;9(7):e102188. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25006672

168. Seelow D, Schwarz JM, Schuelke M. GeneDistiller--distilling candidate genes from linkage intervals. PLoS One [Internet]. 2008;3(12):e3874. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19057649

169. Tranchevent LC, Ardeshirdavani A, ElShal S, Alcaide D, Aerts J, Auboeuf D, et al. Candidate gene prioritization with Endeavour. Nucleic Acids Res [Internet]. 2016;44(W1):W117-21. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27131783

170. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet [Internet]. 2012;13(8):523–36. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22751426

171. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res [Internet]. 2016;44(D1):D1202-13. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26400175

172. Collins FS, Gray GM, Bucher JR. Toxicology. Transforming environmental health protection. Science (80-) [Internet]. 2008;319(5865):906–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18276874

173. Martin MT, Knudsen TB, Reif DM, Houck KA, Judson RS, Kavlock RJ, et al. Predictive model of rat reproductive toxicity from ToxCast high throughput screening. Biol Reprod [Internet]. 2011;85(2):327–39. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21565999

174. Theunissen PT, Beken S, Cappon GD, Chen C, Hoberman AM, van der Laan JW, et al. Toward a comparative retrospective analysis of rat and rabbit developmental toxicity studies for pharmaceutical compounds. Reprod Toxicol [Internet]. 2014;47:27–32. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25517003

175. Knudsen TB, Martin MT, Kavlock RJ, Judson RS, Dix DJ, Singh A V. Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. Reprod Toxicol [Internet]. 2009;28(2):209–19. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19446433

176. Thomas RS, Black MB, Li L, Healy E, Chu TM, Bao W, et al. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. Toxicol Sci [Internet]. 2012;128(2):398–417. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22543276

177. Novotarskyi S, Abdelaziz A, Sushko Y, Korner R, Vogt J, Tetko I V. ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. Chem Res Toxicol [Internet]. 2016;29(5):768–75. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27120770

178. Truong L, Ouedraogo G, Pham L, Clouzeau J, Loisel-Joubert S, Blanchet D, et al. Predicting in vivo effect levels for repeat-dose systemic toxicity using chemical, biological, kinetic and study covariates. Arch Toxicol [Internet]. 2018;92(2):587–600. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29075892

179. Hill 3rd T, Nelms MD, Edwards SW, Martin M, Judson R, Corton JC, et al. Editor's Highlight: Negative Predictors of Carcinogenicity for Environmental Chemicals. Toxicol Sci [Internet]. 2017;155(1):157–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27679563

180. Judson RS, Martin MT, Patlewicz G, Wood CE. Retrospective mining of toxicology data to discover multispecies and chemical class effects: Anemia as a case study. Regul Toxicol Pharmacol [Internet]. 2017;86:74–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28242142

181. Casati S, Aschberger K, Barroso J, Casey W, Delgado I, Kim TS, et al. Standardisation of defined approaches for skin sensitisation testing to support regulatory use and international adoption: position of the International Cooperation on Alternative Test Methods. Arch Toxicol [Internet]. 2018;92(2):611–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29127450

182. Covell DG. Integrating constitutive gene expression and chemoactivity: mining the NCI60 anticancer screen. PLoS One [Internet]. 2012;7(10):e44631. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23056181

183. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model [Internet]. 2010;50(7):1189–204. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20572635

184. Sutton P, Woodruff TJ, Perron J, Stotland N, Conry JA, Miller MD, et al. Toxic environmental chemicals: the role of reproductive health professionals in preventing harmful exposures. Am J Obs Gynecol [Internet]. 2012;207(3):164–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22405527

185. Zhao F, Li R, Xiao S, Diao H, El Zowalaty AE, Ye X. Multigenerational exposure to dietary zearalenone (ZEA), an estrogenic mycotoxin, affects puberty and reproduction in female mice. Reprod Toxicol [Internet]. 2014;47:81–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24972337

186. NCCT. ReadMe for Animal Toxicity Study ToxRefDB files [Internet]. 2018. Available from: https://figshare.com/articles/ReadMe_for_Animal_Toxicity_Study_ToxRefDB_files/6062548

187. NCCT. Animal Toxicity Studies: Effects and Endpoints (Toxicity Reference Database - ToxRefDB files) [Internet]. 2018. Available from: https://figshare.com/articles/Animal_Toxicity_Studies_Effects_and_Endpoints_Toxicity_Reference_Database_-_ToxRefDB_files_/6062545

188. Janus E. Concerns of CropLife America Regarding the Application and Use of the U.S. EPA's Toxicity Reference Database. Environ Health Perspect [Internet]. National Institute of Environmental Health Sciences; 2009;117(10):A432–A432. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2897210/

189. Plunkett LM, Kaplan AM, Becker RA. Challenges in using the ToxRefDB as a resource for toxicity prediction modeling. Regul Toxicol Pharmacol [Internet]. 2015;72(3):610–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26003516

190. Liu J, Patlewicz G, Williams AJ, Thomas RS, Shah I. Predicting Organ Toxicity Using in Vitro Bioactivity Data and Chemical Structure. Chem Res Toxicol [Internet]. 2017;30(11):2046–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28768096

191. USEPA. Benchmark dose technical guidance [Internet]. Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum; 2012. Available from: https://www.epa.gov/risk/benchmark-dose-technical-guidance

192. Wahi MM, Parks D V, Skeate RC, Goldin SB. Reducing errors from the electronic transcription of data collected on paper forms: a research data case study. J Am Med Inf Assoc [Internet]. 2008;15(3):386–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18308994

193. Majid A, Bae ON, Redgrave J, Teare D, Ali A, Zemke D. The Potential of Adaptive Design in Animal Studies. Int J Mol Sci [Internet]. 2015;16(10):24048–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26473839

194. USEPA. Health Effects Test Guidelines: OPPTS 870.3800 Reproduction and Fertility Effects [Internet]. 1998. Available from: https://www.regulations.gov/document?D=EPA-HQ-OPPT-2009-0156-0018

195. Courtot M, Juty N, Knupfer C, Waltemath D, Zhukova A, Drager A, et al. Controlled vocabularies and semantics in systems biology. Mol Syst Biol [Internet]. 2011;7:543. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22027554

196. Ward JM, Schofield PN, Sundberg JP. Reproducibility of histopathological findings in experimental pathology of the mouse: a sorry tail. Lab Anim (NY) [Internet]. NIH Public Access; 2017 Mar 22 [cited 2019 Jan 19];46(4):146–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28328876

197. Wolf JC, Maack G. Evaluating the credibility of histopathology data in environmental endocrine toxicity studies. Environ Toxicol Chem [Internet]. 2017 Mar [cited 2019 Jan 19];36(3):601–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27883231

198. Society of Toxicologic Pathology. International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) [Internet]. Available from: https://www.toxpath.org/inhand.asp

199. Evans RS. Electronic Health Records: Then, Now, and in the Future. Yearb Med Inf [Internet]. 2016;Suppl 1:S48-61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27199197

200. Moreno-Conde A, Moner D, Cruz WD, Santos MR, Maldonado JA, Robles M, et al. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inf Assoc [Internet]. 2015;22(4):925–34. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25796595

201. Kaufman L, Gore K, Zandee JC. Data Standardization, Pharmaceutical Drug Development, and the 3Rs. ILAR J [Internet]. 2016;57(2):109–19. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28053065

202. Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med [Internet]. 2009;48(5):408–13. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19621114

203. USEPA. Health Effects Test Guidelines: OPPTS 870.4100 Chronic Toxicity [Internet]. 1998. Available from: https://www.regulations.gov/document?D=EPA-HQ-OPPT-2009-0156-0019

204. Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, et al. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA J [Internet]. 15(8):e04971. Available from: https://efsa.onlinelibrary.wiley.com/doi/abs/10.2903/j.efsa.2017.4971

205. Simpson D. Units for reporting the results of toxicological measurements. Ann Clin Biochem [Internet]. 1980;17(6):328–31. Available from: https://www.ncbi.nlm.nih.gov/pubmed/7212606

206. Zegers I, Schimmel H. To Harmonize and Standardize: Making Measurement Results Comparable. Clin Chem [Internet]. 2014;60(7):911. Available from: http://clinchem.aaccjnls.org/content/60/7/911.abstract

207. NTP. Specifications for the conduct of studies to evaluate the toxic and carcinogenic potential of chemical, biological and physical agents in laboratory animals for the national toxicology program (NTP) [Internet]. RTP, NC; 2006. Available from: http://ntp.niehs.nih.gov/ntp/Test_Info/FinalNTP_ReproSpecsMay2011_508.pdf

208. Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, et al. “ToxRTool”, a new tool to assess the reliability of toxicological data. Toxicol Lett [Internet]. 2009;189(2):138–44. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19477248

209. Klimisch HJ, Andreae M, Tillmann U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol [Internet]. 1997;25(1):1–5. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9056496

210. Segal D, Makris SL, Kraft AD, Bale AS, Fox J, Gilbert M, et al. Evaluation of the ToxRTool’s ability to rate the reliability of toxicological data for human health hazard assessments. Regul Toxicol Pharmacol [Internet]. 2015;72(1):94–101. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25777839

211. Haber LT, Dourson ML, Allen BC, Hertzberg RC, Parker A, Vincent MJ, et al. Benchmark dose (BMD) modeling: current practice, issues, and challenges. Crit Rev Toxicol [Internet]. 2018;48(5):387–415. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29516780

212. Hardy A, Benford D, Halldorsson T, Jeger Michael J, Knutsen Katrine H, More S, et al. Update: use of the benchmark dose approach in risk assessment. EFSA J [Internet]. Wiley-Blackwell; 2017;15(1):e04658. Available from: https://doi.org/10.2903/j.efsa.2017.4658

213. Davis JA, Gift JS, Zhao QJ. Introduction to benchmark dose methods and U.S. EPA’s benchmark dose software (BMDS) version 2.1.1. Toxicol Appl Pharmacol [Internet]. 2011;254(2):181–91. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21034758

214. USEPA. Benchmark Dose Software (BMDS) [Internet]. 2.2 R65 [B. Research Triangle Park, NC: National Center for Environmental Assessment; 2011. Available from: http://www.epa.gov/NCEA/bmds/index.html

215. USEPA. Benchmark Dose Software (BMDS). Version 2.2 [Internet]. 2012. Available from: http://www.epa.gov/NCEA/bmds/index.html

216. Fournier K, Tebby C, Zeman F, Glorennec P, Zmirou-Navier D, Bonvallot N. Multiple exposures to indoor contaminants: Derivation of benchmark doses and relative potency factors based on male reprotoxic effects. Regul Toxicol Pharmacol [Internet]. 2016;74:23–30. Available from: http://dx.doi.org/10.1016/j.yrtph.2015.11.017

217. Gephart LA, Salminen WF, Nicolich MJ, Pelekis M. Evaluation of subchronic toxicity data using the benchmark dose approach. Regul Toxicol Pharmacol [Internet]. 2001;33(1):37–59. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11259178

218. Wignall JA, Shapiro AJ, Wright FA, Woodruff TJ, Chiu WA, Guyton KZ, et al. Standardizing benchmark dose calculations to improve science-based decisions in human health assessments. Env Heal Perspect [Internet]. 2014;122(5):499–505. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24569956

219. Pham L, Watford S, Paul-Friedman K, Fostel J, Wignall J, Shapiro A. Python BMDS: A Python interface library and webserver for the canonical EPA dose-response modeling software. Prep.

220. Faustman EM, Allen BC, Kavlock RJ, Kimmel CA. Dose-response assessment for developmental toxicity. I. Characterization of database and determination of no observed adverse effect levels. Fundam Appl Toxicol [Internet]. 1994 Nov [cited 2019 Jan 19];23(4):478–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7867899

221. Filipsson AF, Sand S, Nilsson J, Victorin K. The benchmark dose method--review of available models, and recommendations for application in health risk assessment. Crit Rev Toxicol [Internet]. 2003 [cited 2019 Jan 19];33(5):505–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14594105

222. Yoon M, Blaauboer BJ, Clewell HJ. Quantitative in vitro to in vivo extrapolation (QIVIVE): An essential element for in vitro-based risk assessment. Toxicology [Internet]. 2015;332:1–3. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25680635

223. OECD. The Global Portal to Information on Chemical Substances, eChemPortal [Internet]. 2014. Available from: http://www.oecd.org/chemicalsafety/risk-assessment/echemportalglobalportaltoinformationonchemicalsubstances.htm

224. Canada H. Science Approach Document Threshold of Toxicological Concern (TTC)-based Approach for Certain Substances [Internet]. 2016. Available from: https://www.ec.gc.ca/ese-ees/default.asp?lang=En&n=326E3E17-1

225. Casey W, Jacobs A, Maull E, Matheson J, Clarke C, Lowit A. A new path forward: the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and National Toxicology Program’s Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM). J Am Assoc Lab Anim Sci [Internet]. 2015;54(2):170–3. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25836963

226. Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, Hauser R, Prins GS, Soto AM, et al. Endocrine-disrupting chemicals: an Endocrine Society scientific statement. Endocr Rev [Internet]. 2009;30(4):293–342. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19502515

227. Tilghman SL, Nierth-Simpson EN, Wallace R, Burow ME, McLachlan JA. Environmental hormones: Multiple pathways for response may lead to multiple disease outcomes. Steroids [Internet]. 2010;75(8–9):520–3. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20466011

228. IARC. Formaldehyde. In: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans [Internet]. 2012. p. 401–35. Available from: https://monographs.iarc.fr/wp-content/uploads/2018/06/mono100F-29.pdf

229. Heindel JJ, Blumberg B, Cave M, Machtinger R, Mantovani A, Mendez MA, et al. Metabolism disrupting chemicals and metabolic disorders. Reprod Toxicol [Internet]. 2017;68:3–33. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27760374

230. Hartung T, Daston G. Are in vitro tests suitable for regulatory use? Toxicol Sci [Internet]. 2009;111(2):233–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19617452

231. Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des [Internet]. 2007;13(34):3494–504. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18220786

232. Pham L, Watford S, Pradeep P, Judson R, Grulke C, Setzer R, et al. Estimating the limits on alternative model predictivity for system effects by defining variability in in vivo toxicity studies in Toxicity Reference Database (ToxRefDB). Prep.

233. IARC. Preamble: IARC monographs on the evaluation of carcinogenic risks to humans. In Lyon, France; 2006. Available from: https://monographs.iarc.fr/wp-content/uploads/2018/06/CurrentPreamble.pdf

234. Swenberg JA, Moeller BC, Lu K, Rager JE, Fry RC, Starr TB. Formaldehyde carcinogenicity research: 30 years and counting for mode of action, epidemiology, and cancer risk assessment. Toxicol Pathol [Internet]. 2013;41(2):181–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23160431

235. Zhang L, Freeman LE, Nakamura J, Hecht SS, Vandenberg JJ, Smith MT, et al. Formaldehyde and leukemia: epidemiology, potential mechanisms, and implications for risk assessment. Env Mol Mutagen [Internet]. 2010;51(3):181–91. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19790261

236. Cohen SM, Arnold LL. Chemical carcinogenesis. Toxicol Sci [Internet]. 2011;120 Suppl:S76-92. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21147961

237. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet [Internet]. 2000;25(1):25–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10802651

238. The Gene Ontology C. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res [Internet]. 2017;45(D1):D331–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27899567

239. Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, et al. Reactome graph database: Efficient access to complex pathway data. PLoS Comput Biol [Internet]. 2018;14(1):e1005968. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29377902

240. MEDLINE Co-Occurrences (MRCOC) Files [Internet]. 2016. Available from: https://ii.nlm.nih.gov/MRCOC.shtml

241. Lillioja S, Mott DM, Spraul M, Ferraro R, Foley JE, Ravussin E, et al. Insulin resistance and insulin secretory dysfunction as precursors of non-insulin-dependent diabetes mellitus. Prospective studies of Pima Indians. N Engl J Med [Internet]. 1993;329(27):1988–92. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8247074

242. Indexing Initiative [Internet]. MEDLINE Co-Occurrences (MRCOC) Files. US National Library of Medicine; Available from: https://ii.nlm.nih.gov/MRCOC.shtml

243. Medicine USNL of, editor. MEDLINE Baseline Repository Raw Data Files [Internet]. 2017. Available from: https://mbr.nlm.nih.gov/Download/RawData/2017/

244. UMLS API Technical Documentation. Available from: https://documentation.uts.nlm.nih.gov/rest/home.html

245. Enrichr Help Center: API Documentation [Internet]. Available from: http://amp.pharm.mssm.edu/Enrichr/help#api

246. ToxCast and Tox21 Summary Files. 2018; Available from: https://figshare.com/articles/ToxCast_and_Tox21_Summary_Files/6062479

247. Enrichr Libraries [Internet]. Available from: http://amp.pharm.mssm.edu/Enrichr/#stats

248. Estrogen-dependent gene expression [Internet]. Reactome Pathway Knowledgebase; Available from: https://reactome.org/content/detail/R-HSA-9018519

249. ESR mediated signaling [Internet]. Reactome Pathway Knowledgebase; Available from: https://reactome.org/content/detail/R-HSA-8939211

250. Shu XS, Li L, Tao Q. Chromatin regulators with tumor suppressor properties and their alterations in human cancers. Epigenomics [Internet]. 2012;4(5):537–49. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23130835

251. Kumar R, Li DQ, Muller S, Knapp S. Epigenomic regulation of oncogenesis by chromatin remodeling. Oncogene [Internet]. 2016;35(34):4423–36. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26804164

252. Martincic-Ipsic S, Mocibob E, Perc M. Link prediction on Twitter. PLoS One [Internet]. 2017;12(7):e0181079. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28719651

253. Goh WP, Luke K-K, Cheong SA. Functional shortcuts in language co-occurrence networks. PLoS One [Internet]. Public Library of Science; 2018;13(9):e0203025. Available from: https://doi.org/10.1371/journal.pone.0203025

254. Kastrin A, Rindflesch TC, Hristovski D. Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-based Discovery. Methods Inf Med [Internet]. 2016;55(4):340–6. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27435341

255. Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: a technical report. J Am Med Inf Assoc [Internet]. 2008;15(2):130–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18096907

256. Pareja-Tobes P, Tobes R, Manrique M, Pareja E, Pareja-Tobes E. Bio4j: a high-performance cloud-enabled graph-based data platform. bioRxiv [Internet]. 2015; Available from: http://biorxiv.org/content/early/2015/03/20/016758.abstract

257. Messina A, Pribadi H, Stichbury J, Bucci M, Klarman S, Urso A. BioGrakn: A Knowledge Graph-Based Semantic Database for Biomedical Sciences. In: Barolli L, Terzo O, editors. Complex, Intelligent, and Software Intensive Systems. Cham: Springer International Publishing; 2018. p. 299–309.

258. Bateman AR, El-Hachem N, Beck AH, Aerts HJWL, Haibe-Kains B. Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. Sci Rep [Internet]. The Author(s); 2014;4:4092. Available from: https://doi.org/10.1038/srep04092

259. Formation of NR-MED1 Coactivator Complex [Internet]. Reactome Pathway Knowledgebase; Available from: https://reactome.org/content/detail/R-HSA-376419

260. Angrish MM, Allard P, McCullough SD, Druwe IL, Helbling Chadwick L, Hines E, et al. Epigenetic Applications in Adverse Outcome Pathways and Environmental Risk Evaluation. Env Heal Perspect [Internet]. 2018;126(4):45001. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29669403

261. Edwards SW, Tan YM, Villeneuve DL, Meek ME, McQueen CA. Adverse Outcome Pathways-Organizing Toxicological Information to Improve Decision Making. J Pharmacol Exp Ther [Internet]. 2016;356(1):170–81. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26537250

262. Study Data Standards - Study Data for Submission to CDER and CBER. [cited 2019 Feb 2]; Available from: https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm587508.htm

263. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci [Internet]. National Academy of Sciences; 2007 May 22 [cited 2019 Feb 3];104(21):8685–90. Available from: https://www.pnas.org/content/104/21/8685

264. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nat Rev Genet [Internet]. NIH Public Access; 2014 Jan [cited 2019 Feb 3];15(1):34–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24296534

265. Silverman EK, Loscalzo J. Network medicine approaches to the genetics of complex diseases. Discov Med [Internet]. NIH Public Access; 2012 Aug [cited 2019 Feb 3];14(75):143–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22935211

266. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Front cell Dev Biol [Internet]. Frontiers Media SA; 2014 [cited 2019 Feb 3];2:38. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25364745

267. Toxicology ENC for C. ToxCast Database (invitroDB) [Internet]. 2018. Available from: https://epa.figshare.com/articles/ToxCast_Database_invitroDB_/6062623

268. US EPA O. Chemistry Dashboard Help: Toxicity Values. [cited 2019 Feb 1]; Available from: https://www.epa.gov/chemical-research/chemistry-dashboard-help-toxicity-values

269. US EPA O. Chemical and Products Database (CPDat). [cited 2019 Feb 1]; Available from: https://www.epa.gov/chemical-research/chemical-and-products-database-cpdat

270. US EPA O. ToxCast Dashboard. [cited 2019 Feb 1]; Available from: https://www.epa.gov/chemical-research/toxcast-dashboard

271. USEPA. Use of High Throughput Assays and Computational Tools; Endocrine Disruptor Screening Program; Notice of Availability and Opportunity for Comment. 2015.

272. US EPA OCSPP. Endocrine Disruptor Screening Program (EDSP) in the 21st Century. [cited 2019 Feb 1]; Available from: https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-edsp-21st-century