ABSTRACT

WILLIAMS-DEVANE, CLARLYNDA RAYNELL. Towards a Toxico-Chemogenomic Future: The Transformation of Public Gene Expression Data and Consideration for its Use. (Under the Direction of Ann Richard.)

The term "toxico-chemogenomics" is used to convey extension of toxicogenomics to more broadly survey gene expression changes across chemical space. Moving towards an improved, publicly available toxico-chemogenomics capability requires not only common data standards and protocols across public resources, but also broad data coverage within the chemical, genomics and toxicological information domains, and transparent and functional linkages of Internet data resources. The first goal of this project was to assess the current extent of standardization, interoperability, and chemical indexing of public genomics resources with respect to toxico-chemogenomics utility. Focusing on the largest of these public data resources – Gene Expression Omnibus (GEO) and ArrayExpress -- the second goal was to chemically index the full experimental content of these repositories to assess the current coverage of chemical exposure-related microarray experiments in relation to chemical space and toxicology, and to make these data accessible in relation to other publicly available, chemically-indexed toxicological information. Current standards for chemical annotation within ArrayExpress and GEO are presently inadequate to this task, such that development of new methodologies to mine the author-submitted content was required. A series of automated Perl programs were utilized along with extensive manual review to transform the raw experiment/study descriptions and text files into a standardized chemically-indexed inventory of microarray experiments in both resources. These files and top-level experiment annotations allowed for identification of all current chemical-associated experimental content as well as the subset of chemical exposure-related (or "Treatment") content deemed most relevant to toxicogenomics in the GEO Series and ArrayExpress Repository experiment

inventories. With chemical exposure experiments suitably indexed by chemical structure, it is possible for the first time to assess the breadth of chemical study space represented in these databases, as well as the overlapping chemical content, and to begin to assess the sufficiency of data for making chemical similarity inferences. Chemical indexing of public genomics databases is also the first step towards integrating chemical, toxicological and genomics data into predictive toxicology by providing linkages across public resources. The main products of this effort include the following: (1) published, downloadable and structure-searchable DSSTox Structure-Index (Locator) files for both the GEO Series (GEOGDS) and ArrayExpress Repository (ARYEXP), containing standard chemical fields for the unique chemical "Treatment" subset, accompanied by URLs to AccessionID experiment pages in GEO and ArrayExpress; (2) published, downloadable DSSTox Aux data files for GEOGDS and ARYEXP providing a chemical-experiment pair index to all chemical-associated content in each resource and containing 14 standard genomics fields (e.g., Experiment_Title, Experiment_Description, Experiment_ArrayType, Species, Number_Samples, etc.) and source-specific fields extracted from each resource (e.g., MIAME_Protocol, MIAMI_Factors, etc. for ArrayExpress); and (3) incorporation of the "Treatment" chemical-experiment pair index with URLs linked directly to AccessionID pages for GEO and ArrayExpress into the National Center for Biotechnology Information (NCBI) PubChem resource. The secondary product of this effort is a methodology discussion about the proper use of public microarray data with a demonstrative analysis of how one might use the newly identified public microarray data.

Towards a Toxico-Chemogenomic Future: The Transformation of Public Gene Expression Data and Consideration for its Use

by
ClarLynda Raynell Williams-DeVane

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2009

APPROVED BY:

David Muddiman Co-Chair of Advisory Committee	Dahlia Nielsen Co-Chair of Advisory Committee
Ann Richard	Steffen Heber

Jose Alonso

DEDICATION

I dedicate this first to my husband. Without your sacrifice, support, and love I would have never made it through this journey.

From the moment I saw you, I wanted to meet you.

From the moment I met you, I wanted to know you.

From the moment I knew you, I was in love with you.

From the moment I loved you, I wanted to share my life with you.

And from this moment to that moment, and all the moments to come…I will love you with all of my heart.

Second, I would like to dedicate this to my parents, grandparents, and brother because your faith in me and my abilities have helped me survive this journey. Your sacrifice to help me become the first in our family to go to college and then to graduate school will never be forgotten.

BIOGRAPHY

ClarLynda Raynell Williams-DeVane was born on March 25, 1981 to Clarence Ray and Linda Darnell Williams in Grifton, North Carolina. ClarLynda has a younger brother, Cory Williams. From a young age, ClarLynda loved anything with buttons and numbers. She was fascinated with calculators and cash registers. She was an odd child who loved to *read* math books. She went on to play basketball and softball for most of her middle and high school years. She attended Ayden-Grifton high school in Littlefield, NC for her freshman and sophomore years and the North Carolina of School of Science and Mathematics (NCSSM) in Durham, NC for her Junior and Senior year. While at NCSSM, ClarLynda feel deeper in love with mathematics and science. So much so, that she went on to major in mathematics at North Carolina Central University (NCCU) in Durham, NC. During her freshman year at NCCU, ClarLynda was fortunate enough to earn an internship at the Environmental Protection Agency (EPA) in Research Triangle Park, NC. She developed visual basic macros to help scientist more efficiently analyze their data. Through several other internships at the EPA, ClarLynda ended up in the lab of Dr. Ann Richard. ClarLynda worked in Dr. Richard's computational chemistry lab on the DSSTox project for the remainder of her matriculation at NCCU. She graduated Summa Cum Laude from North Carolina Central University. ClarLynda was encouraged to go to graduate school after graduation. After searching for a program that fit her unique background of mathematics, science, database creation; ClarLynda decided on attending the bioinformatics program at North Carolina State University in the summer of 2003. In the meantime ClarLynda had fallen in love with her best friend and they married on December 18, 2004. ClarLynda continued graduate school at NCSU and found her way back to Dr. Richard's lab to complement her previous work on the DSSTox project with her newfound bioinformatic knowledge. After graduation ClarLynda will continue on as a Post-Doc at the Environmental Protection Agency.

ACKNOWLEDGEMENTS

I would like to first thank GOD and my Lord and Savior Jesus Christ for his grace and mercy throughout this journey.

I would like to thank my research advisor, Dr. Ann Richard for her dedication and patience in helping me to develop from a teenager into a scientist, which sometime meant being a mom as much as an advisor. I would also like to thank the other members of my committee for their willingness to serve and patience and belief in my research. . I would like to thank Erik Griffis for his invaluable assistance this summer.

I would like to thank Dr. Maritja Wolf for her time and dedication to this project. Your willingness to teach and continue to produce such high quality and well respected work is invaluable.

I would like to thank Dr. Jennifer Fostel for her support, guidance, and help in refining my ideas.

I would like to thank my best friend, Chirnese, for her support throughout this journey. Who knew that those late nights in January 2003 with me applying to graduate schools and you applying to law schools at the last minute would end like this?

I would like to thank all of my teachers and professors who saw the best in me. Thank you for your encouragement, your high expectations, and your honesty. I would also like to thank one very special professor who only saw the worst in me. Thank you for your persistence in assuring my failure, I am a better, more determined person because of you.

Lastly, I would like to thank all of my fellow graduate students and officemates at the EPA. You have made me think, made me laugh, and strive to be a better person and scientist.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

TOXICOGENOMICS

Conventional toxicology investigates cellular and animal responses to chemical treatment through domain-specific bioassay studies (e.g., chronic, developmental, reproductive, neurological, immunological, etc), typically mapping a single chemical to one or a few toxicological endpoints. Microarray technologies, in contrast, detect genome-wide perturbations resulting from a chemical treatment, and measure response variables that probe a large number of genes and gene pathways potentially underlying multiple toxicological endpoints. In 1999, Nuwaysir et al. proposed a novel approach to toxicology, toxicogenomics that would take advantage of the strengths of both toxicology and microarray or other gene expression technologies. A typical toxicogenomics experiment bridges these technologies and approaches, focusing on treatment-related effects of a single chemical and attempting to relate gene expression changes to one or more a toxicological endpoint(s) (Ahuja et al. 2007; Gomase and Tagore 2008; Hamadeh et al. 2002; Hirabayashi et al. 2002). Repetition of these experiments for a series of chemicals simultaneously probes the structure-activity basis for both the genomics and toxicological responses (Goetz et al. 2006). Hence, in addition to their use for exploring underlying mechanisms, toxicogenomics studies that more broadly sample chemical and biological effects have the potential to be used to derive compound-induced gene expression patterns, or activity profiles, that are predictive of a toxicological response across chemical space.

Toxicogenomics is the study of the effect of a compound or stressor on an organism, as measured using conventional toxicology as well as other 'omic' technologies, mainly transcriptomics. Transcriptomics is "a global way of looking at gene-expression patterns that can involve measurements of thousands of genes simultaneously with microarrays or measurements of small numbers of genes that are facilitated by global sequence information from EST or genome-

sequencing projects" (Nature Glossary 2008). The goal of toxicogenomics is to explore the relationship between gene expression changes and the biological parameters resulting from toxicology experiments. Measurement of Toxicity and Gene Expression are the key components of toxicogenomics.

TOXICOLOGY

GENERAL TOXICOLOGY

Toxicology in its simplest terms refers to a science that deals with poisons, any substance that causes a harmful effect when administrated to a living organism by accident or intentionally. Poisons can be chemical, physical or biological agents that in small doses may cause no adverse effect. The characteristic property of a poison is that the dose makes the poison; i.e. a particular dose of an agent causes the adverse effect. Toxicology is specifically involved with the study of the adverse effects of these agents. The study of toxicology is a complex field that takes into account the cascade of events that occur as a result of a particular poison. A typical in vivo toxicology study follows the poison through the organism's system from exposure to distribution and metabolism and finally to the toxic endpoint. *In vitro* toxicology experiments attempt to mimic these changes through the use of cell or tissue cultures in lieu of whole organisms. There are several dimensions to a toxicological study, i.e., dose, time, end point, and route of administration.

Toxicology is studied from many perspectives and on many levels. Biochemical and molecular toxicology focuses on the adverse effects of a poison at the biochemical and molecular levels (Hodgson 2004). Behavioral toxicology focuses on observing the adverse effects of a poison on the behavior of an organism. Nutritional toxicology focuses how the adverse effects of a poison are affected by the nutrition of the organism. Carcinogenesis focuses on the global view of all events that

lead to cancer as a result of exposure to a poison; events are considered from the chemical, biochemical and molecular perspectives. Teratogenesis is similar to carcinogenesis in its perspective, but focuses on events that lead to adverse developmental events. Mutagenesis focuses on the adverse effect of a poison to genetic material and, thus, inherited properties resulting from mutations caused by a poison. Finally, organ toxicity focuses on adverse effects of a poison to a particular target organ. Organ level toxicity is perhaps the most familiar of all toxicological measurements and includes: neurotoxicity (brain), hepatotoxicity (liver), nephrotoxicity (kidney), reproductive toxicity (reproductive organs), cardiac toxicity (heart), and pulmonary toxicity (lung).

Measurement of Toxicity

The measurement of toxicity is also complex because it must take into account several parameters; i.e. the route of administration, the dose, the time of exposure, and the endpoint. A researcher may see up to a tenfold difference in toxicity measurements based simply on the route of administration (Hodgson 2004). According to the Food and Drug Administration (FDA), there are more than 111 specific routes of administration (FDA 2008). The routes of administration are broadly classified as enternal, pulmonary, topical, and parentaral (Barnes 2006). Enternal administration refers to administration through the gastrointestinal tract, i.e. mouth. Pulmonary administration refers to administration through the respiratory system; i.e. inhalation. Topical administration refers to administration directly to the body part, i.e. skin. Parenteral administration refers to administration through any route other than the gastrointestinal or respiratory tracts mainly through injection. Each of these routes of administration may be used in the observation or measurement of local or systematic effects. Local effects occur at the site of administration, whereas systemic effects occur as the substance travels through the organism. The systemic effects of a particular poison involve all the events leading to toxicity *in vivo*, from uptake, distribution, and metabolism to target interaction and

excretion. Mode-of-action refers to the overall cascade of events leading to the toxicity endpoint, whereas mechanism-of-action refers more specifically to a key determining event that occurs during the cascade of events after the uptake of the poison (Hodgson 2004).

There are several traditional methods of toxicity measurement that focus on measuring toxicity in living organisms. Traditional toxicity measurements primarily focus on whole-mammalian animals, *in vivo* (Sklarew 1993). Analytical toxicology focuses on the identification and assay of toxic chemicals and their metabolites. Epidemiology focuses on extrapolating of experimental results to the human population specifically in reference to human disease. Biomathematics and statistics deal with the analysis of data, determination of significance, and formulation of risk estimates for predictive models. Bioassay testing involves the use of living systems to estimate toxicity effects. Most toxicity tests examine specific adverse effects referred to as endpoints. Sample endpoints are acute systematic toxicity, skin irritation/corrosion, eye irritation/corrosion, skin sensitization, repeated dose toxicity, reproductive and developmental toxicity, genotoxicity, carcinogenicity, neurotoxicity, and ecotoxicity (Alternative Toxicology 2008a). Pharmokinetics and metabolism also play an integral role in determining adverse outcome pertaining to each of these endpoints. Other toxicity test are more general and measure the general effects of a chemical after a single dose, i.e., acute exposure, or after multiple-doses , i.e., repeated dose exposure (Hodgson 2004). Finally, toxicity studies can have a temporal dimension , with *in vivo* studies spanning one month termed acute, studies spanning three months termed sub acute, and studies generally lasting more than three months termed long-term chronic studies (Alttox.org).

Alternative toxicity methods consist of the use of methods alternative to the use of the live mammalian species. Alternative toxicity methods range from structure-activity relationship (SAR) modeling to *in vitro* assays employing cell and tissue cultures. SAR methods, which rely solely on

computational methods, fit into a broad *in silico* category. *In silico* methods seek to define relationships based on existing data and computational models that can be used to predict toxicity. SAR specifically defines relationships between the biological, physical, and toxicological properties of the poison and the chemical structure in order to predict toxicity (Alternative Toxicology 2008b). *In vitro* methods primarily focus on using cell and tissue cultures to measure toxicity in ways that are comparable to *in vivo* results. Newer high-throughput screening (HTS) methods can run a large number of samples (hundreds to thousands) through an *in vitro* assay in a quick, automated process.

Many *in vitro* studies are accompanied by other methods such as 'omics' methods, which include , e.g., genomics, proteomics, metabonomics, transcriptomics, glycomics, and lipomics(Alternative Toxicology 2008c). Genomics is the study of genes and their function (Alternative Toxicology 2008c). Proteomics is the study of proteins. Metabonomics is the study of molecules involved in cellular metabolism. Transcriptomics is the study of the mRNA. Glycomics and lipoemics are the study of cellular carbohydrates and lipids, respectively. These technologies can be used to evaluate global changes involving cellular DNA or RNA. Other alternative methods involve the use of non-mammalian model organisms such as bacteria, yeast, *c. elegans*, and *e. coli*, or vertebrates such fish, amphibians, reptiles, and birds, where there are conserved characteristics that produce results that can be more easily extrapolated to mammalian species (Sklarew 1993).

The computational analysis of the many facets, and sometimes large dimensions of alternative toxicity results is referred to as computational biology. Bioinformatics is a computational biology method that involves managing and analyzing large amounts of biological data using advanced computing techniques (HGP 2003). Bioinformatics refers to approaches ranging from multidimensional correlation approaches to approaches that generate maps of cellular and physiological pathways and responses. As Bayat (2002) put it, "Bioinformatics is used to abstract

knowledge and principles from large-scale data, to present a complete representation of the cell and the organism, and to predict computationally systems of high complexity, such as the interaction networks in cellular processed and phenotypes of whole organism" (Bayat, 2002). Systems biology is an integration of data across all levels of complexity into a systems view of biological and pathological processes (Mcgee 2006). The National Center for Biotechnology Information (NCBI) offer several public databases and tools for the integration of genomics and bioinformatic methods (NCBI 2008). The primary 'omic' technology currently employed in toxicology is gene expression, giving rise to the field of toxicogenomics.

GENE EXPRESSION

GENERAL GENE EXPRESSION

On the molecular level, toxicity can be better understood by observing changes in gene expression as a result of exposure to a poison(s). In 1958, Watson and Crick proposed the Central Dogma Theory which states generally that DNA is transcribed into RNA and RNA is translated into proteins (Figure 1-1) (Watson and Crick 1953). Deoxyribose Nucleic Acid (DNA) is a double stranded helix macromolecule residing in the nucleus of each cell of a high organism that stores the basic genetic code for the organism. Through a process called transcription, an enzyme called Riobonucleic Acid (RNA) polymerase makes a copy of one of the DNA strands, producing a strand of RNA. Further, through a process called translation, messenger RNA or mRNA carries the genetic instructions to the ribosome. Ribosomes read the genetic code in the mRNA and build a protein based on these instructions. This process is the fundamental foundation of gene expression.

MEASURE OF GENE EXPRESSION ACTIVITY

There are many different measures of gene expression which focus on the functional units that are produced based on DNA, mainly RNA and proteins. The expression of a gene is often measured by the amount of mRNA. The use of mRNA concentration as an indication of gene expression, however, can be misleading because of post-transcriptional regulation such as alternative splicing and production of miRNA. This may skew the correlation between gene expression and mRNA concentration such that an increase in mRNA concentration may not correspond to an increase in expression of a particular gene (Gygi et al. 1999). Since proteins are thought of as the functional product of DNA, direct measurement of proteins could be considered to be the best measure of gene expression. However, mRNA measurement is technologically more practical and feasible at present and, hence, is most often employed as a measure of gene expression and, more specifically, gene activity. The concentration of a single mRNA or it's transcript can be estimated by a myriad of techniques. Quantitative measurements of mRNA primarily deal with the idea of hybridization, i.e., the ability of a single-stranded nucleic acid to form a double helix with another single strand of complementary base sequence. It is possible that two single-stranded nucleic acid molecules that are not fully complementary may also hybridize, but the greater the complementarity, the stronger the binding (Causton et al. 2003).

One of the more widely used techniques involves the use of polymerase chain reactions (PCR), which is defined as the amplification of a region of DNA using primers that flank the region and repeated cycles of DNA polymerase action (Weaver 2005). Real-time PCR (RT-PCR) or quantitative PCR (qPCR), also called RT-qPCR is a combination of three steps: 1) the reverse transcription (RT)-dependent conversion of RNA into cDNA, 2) the amplification of the cDNA using PCR, and 3) the detection and quantification of amplification producers in real time (Gibson et al. 2006). RT-qPCR

has become the "gold standard" of mRNA quantification (Bustin 2000; Ginzinger 2002). However, in 2006, Nolan et al. noted that there were many problems and inconsistencies that needed to be addressed, ranging from the standardization of protocols to the variation in data analysis methods. However, RT-qPCR is still considered a standard methodology and is often used for the validation of high-throughput techniques for the quantification of mRNA measure.

Gene Expression can be divided into two categories: 1) where samples provide information on genes, and 2) where genes provide information about the sample. This discussion focuses on the latter category. A gene expression profile, or signature, can be thought of as a precise, reproducible molecular definition of the cell in a specific state, i.e., a snapshot of sorts (Young 2000). A large reference collection of profiles can be used to compare gene expression data to identify similar patterns (Ganter et al. 2005; NCI60 2008). Gene expression has proved to be a highly robust 'reporter' of biological status for a wide range of samples under a variety of conditions, with the result that microarray technologies are utilized extensively within industry (Causton et al. 2003). With the advent of whole organism sequencing of the genomes of several model organisms in 1997, as well as large portions of the genomes of mammals and humans, microarrays were employed to study large-scale gene expression in the context of a high-throughput, genome-wide assay (DeRisis et al. 1997). In general, mRNA samples are labeled, hybrydized to an array, and scanned to measure the abundance of the label. There are two types of microarray technologies: 1) spotted microarrays, and 2) oligonucleotide microarrays. Spotted microarrays have cDNA probes that correspond to mRNAs. The probes are synthesized and then spotted onto glass arrays. Spotted microarrays have the advantage of being low cost and easily customizable for each experimental study; however, spotted microarrays are not thought to have the same sensitivity as commercial oligonucleotide microarrays (Bammler et al. 2005). Oligonucleotide microarrays have short sequence probes that

match parts of the sequence of known or predicted open reading frames. Oligonucleotide arrays differ from traditional arrays in that they are printed with short oligonucleotide sequences (probe sets) that are designed to represent one gene in lieu of the entire gene sequence, as in spotted arrays. Oligonucleotide arrays are said to be more sensitive and expensive. Gene expression can be detected using two microarray technologies: 1) two-color or two-channel arrays and 2) one-color one channel arrays. Two color arrays use two different fluorophores to label two comparable samples (Shalon et al. 1996) (Figure 2-1); Cy3 and Cy5 dyes are typically used. The intensities of the fluorescent dyes are measured during the scanning process as an indication of the gene expression of one sample relative to the other sample for the represented genes. Conversely, single color arrays are used to estimate absolute gene expression. Each sample requires a single array; hence, the expense of a one color array experiment is twice that of a two color array experiment. However, there is a higher level of quality with one color arrays because each sample is tested independently and not in conjunction with another sample. The choice of microarray technologies depends on both the specific needs of the project, such as the duration and types of comparisons that are to be made, as well as the budget for the project. Each technology has its own advantages and disadvantages. However, the ability to capture a snapshot of global changes in gene expression, genome-wide is of great advantage in toxicology where one is most often interested in higher order phenotypic changes associated with adverse outcomes. For this reason, one channel arrays have gained popularity within the field of toxicology (Vrana et al. 2003).

There are other high-throughput methods such as Serial Analysis of Gene Expression (SAGE) that use short sequence tags (10-14 base pairs) to identify gene expression patterns. SAGE has three primary principles: 1)a short sequence tag is obtained from a unique position within each transcript, 2) sequence tags are linked together to form long serial molecules that can be cloned and sequenced,

and 3) the number of times a particular tag is observed corresponds to the expression level of the corresponding transcript(Velculescu et al. 1995). SAGE is a digital method and its sensitivity is dependent on the number of tags sequences. SAGE's power lies in its ability to identify novel genes that are expressed under certain conditions; however SAGE is difficult to carry out on a routine basis (Høgh and Nielsen 2008). Public SAGE data can be found in public gene expression repositories.

CHEMOGENOMICS

A natural extension of gene expression measurement is chemogenomics, where one begins to look across a series of transcriptomic experiments for chemical effects. The term "chemogenomics" was introduced to more generally encompass the overlap of genomics technologies with treatment-related chemical effects on biological systems, including both toxicity-related and therapeutic effects (Ganter et al. 2005). Chemogenomics adds a top-most chemical layer to data organization, with broad chemical coverage of standardized-protocol experiments a key requirement for discerning activity patterns that can be confidently extrapolated across chemical space. This approach and its implementation are perhaps best exemplified by the Iconix DrugMatrix^{R} database and applications (Ganter et al. 2005; Fielden and Halbert 2007). The Iconix database consists of data generated for a single species (rat), treated by more than 600 compounds in 7 tissue types, representing upwards of 3200 different drug-dose-time-tissue combinations. The database covers five different domains of data: microarray, clinical chemistry, hematology, organ weight, and histopathology, and was built using a common microarray platform and stringent experimental protocols and standards for data generation and processing. Iconix has applied this proprietary commercial database to screen new drug leads for potential toxicity, to characterize mechanisms of toxicity, and to determine the probable efficacy of new drugs early in the drug development process.

IN SILICO TOXICOGENOMICS

Much like chemogenomics, *In silico* meta-analysis methods combine data from existing toxicological experiments with gene expression data to generate new, and to confirm existing hypotheses of the effect of a compound treatment. However, as has been demonstrated with the Iconix database, it is the effective coupling of traditional toxicity endpoint measures (e.g., animal/organ/tissue adverse outcomes) with gene expression data, spanning sufficiently broad endpoint and chemical space, and adhering to common data standards and protocols, that enables derivation of biomarkers or "signatures" for use in toxicity prediction (Ganter et al. 2005; Fielden and Halbert 2007). This spectrum of overlapping, related, and confirming information domains substantially increases the power of inferences that can be made towards better understanding the effect of a given compound, or for extrapolating the results in chemical or biological endpoint space. Specific toxicology domains also have begun to realize the importance and power of integrated data types. The Birth Defects Systems Manager (BDSM) for developmental toxicity integrates microarray data with a broad spectrum of developmental experimental data to derive developmental gene signatures. BDSM also includes specific tools for the integrated analysis of diverse data types (Knudsen et al. 2005; Singh et al. 2005; Singh et al. 2007). Beyond toxicology there are efforts to integrate databases across domains, i.e. proteomics, genomics, and drug molecules (Segota et al. 2008; Waters et al. 2003).

In recent years, there have been significant advances in several areas – microarray databases, toxicology data models, quantitative high-throughput screening (qHTS), and chemically indexed bioassay data – that, taken as a whole, have the potential to greatly enhance toxicogenomics capabilities in the public domain. Several public initiatives are promoting the use of toxicity data standards and data models across toxicity study areas, and populating these with legacy toxicity data from government archive studies (Yang et al. 2006a, 2006b; Richard et al. 2008; Martin et al. in

press) and the scientific literature (Julien et al., 2004; ILSI DevTox, 2008). Additional efforts are more specifically focused on the general improvement of toxicity data standards in relation to toxicogenomics experiments (Fostel et al. 2005; Burgoon 2007; Fostel et al. 2007; Fostel 2008). These varied collaborative efforts are significantly increasing the accessibility and utility of standardized *in vivo* bioassay reference data for data-mining applications and for anchoring new predictive technologies (Yang et al. 2008; Dix et al. 2007; Kavlock et al. 2008; Martin et al. in press). Moving beyond legacy data, the National Toxicology Program (NTP) HTS and EPA ToxCast™ projects (Xia et al. 2008; Dix et al., 2007; Tice et. al 2007), in collaboration with the National Institutes of Health (NIH) Chemical Genomics Center (NCGC) (Collins et al. 2008), are generating new data for thousands of chemicals of toxicological/environmental relevance in hundreds of *in vitro* cell-based and biochemical assays, and anchoring these results to *in vivo* bioassay data where possible (Houck and Kavlock 2008; Dix et al. 2007). As part of this effort, the EPA ToxCast™ project is also generating microarray data for several hundred (mostly pesticidal) compounds for which *in vivo* toxicity data are available.

Corresponding advances in cheminformatics are complementing these various toxicity data enrichment efforts with large-scale chemical structure indexing of public Internet data resources and bioassay data (Richard et al. 2006). The NIH PubChem Project (PubChem 2007) is the most prominent of these public efforts and currently provides full public access and searchability through millions of compounds and hundreds of qHTS bioassays, the latter primarily associated with the NIH Molecular Libraries Screening Initiative (MLSI) (Parker et al. 2006). The ChemSpider Project (Williams 2008; ChemSpider 2008) is an even larger public chemical indexing service, incorporating the full PubChem chemical inventory along with a vast number of other chemical inventories, structures, and properties, and aspiring to be “the richest single source of structure-based chemistry

information" but without direct hosting of bioassay data. More specifically focused on toxicology, the Distributed Structure-Searchable Toxicity (DSSTox) Data Network project (Richard et al. 2008; EPA DSSTox 2008a) and the new Aggregated Computational Toxicology Resource (ACToR) (Judson et al. 2008; EPA ACToR 2008], both being developed within EPA's National Center for Computational Toxicology (NCCT), are enhancing linkages to toxicologically relevant data through standardized chemical structure indexing of a broad array of public toxicity-related data, as well as providing chemical indexing and cheminformatics support for the NTP HTS and EPA ToxCast™ projects. The European Bioinformatics Institute (EBI) Chemicals of Biological Importance (ChEBI) project is a growing resource providing a chemical ontology for biological effects (Degtyarenko et al. 2008).

In the genomics area, publishing requirements for the deposition of raw or processed microarray data to a few main publicly available database repositories, coupled with broad adoption of MIAME (Minimum Information About a Microarray Experiment) standards for data reporting, are increasing the comparability, utility and breadth of these resources (Anderle et al.2003; Ball et al. 2004; Fielden and Kolaja 2006; Larsson and Sandberg 2006]. In addition, crucial to the success of *in silico* meta-analysis methods, progress is being made in providing external programmatic access to the major public microarray data resources, allowing third parties to automatically extract and reformulate data to enhance informatics and data mining capabilities [Parkinson et al. 2007; Barrett et al. 2007, Wheeler et al. 2008; EBI 2008b; NCBI GEO 2008b). In this regard, several public initiatives are being developed that utilize and expand programmatic access to the major public microarray data resources: EzArray (Zhu at al. 2008a); GEOmetadb (Zhu et al. 2008b); MaRe (Ivliev et al. 2008), GEOquery (Sean and Meltzer, 2007), SeqExpress(Boyle 2005), ArrayQuest (Argraves et al. 2005).

TOXICO-CHEMOGENOMICS

RESEARCH PROBLEM

From the previous discussion, it is clear that the potential for toxicogenomics investigation in the public domain through *in silico* methods lies within several isolated resources. These resources currently represent a patchwork of disconnected or loosely connected inventories and capabilities, covering different areas of chemical space and toxicological/biological activity focus, and having different goals, degrees of standardization, public data accessibility, relational data-mining ability, and utility for toxicogenomics investigation. In addition, there is little to no linkage between these resources and other publicly available toxicology related data resources. Thus, there is a need for data integration between these resources to further toxicogenomic investigation in the public domain.

Requirements for data integration were recognized early on (Waters et al. 2003) and, to varying degrees, have informed development of publicly available microarray databases (Mattes et al. 2004; Salter 2005). Given the increasing size and general nature of such databases, and the fact that the global gene expression in response to a chemical treatment has potential relevance to many aspects of toxicology, these resources have the potential to support toxicogenomics investigations. Regulatory agencies, such as the Food and Drug Administration (FDA) and the Environmental Protection Agency (EPA) have issued guidelines for incorporating toxicogenomics data on a case by case basis with other data to support new drug evaluation reviews or environmental risk assessments (U.S. EPA Science Policy Council 2004; U.S. FDA Center for Drug Evaluation and Research 2005). However, the inability to easily access, mine and analyze relevant publicly available toxicogenomics data for the chemical space of interest effectively limits the utility of available public microarray data resources for regulatory applications (Bhogal et al. 2005). As the commercial sector employs

toxicogenomic investigation to address increasingly sophisticated problems and challenges, at high economical cost, regulatory, academic, and other public interests must attempt to keep pace, relying upon the public domain of resources for furthering toxicogenomics capabilities.

MY PROPOSED SOLUTION

In the present research project, I propose to extend the concept of chemogenomics into the public realm of toxicology in order to enhance toxicogenomic studies. I introduce here the expanded term "toxico-chemogenomics" to convey the extension of toxicogenomics to more broadly survey gene expression changes across chemical space, while maintaining a toxicological focus. Creating linkages across multiple domains of toxicological inquiry can be used to build public toxico-chemogenomics data sets for toxicogenomic exploration. Moving towards an improved, publicly available toxico-chemogenomics capability requires not only common data standards, protocols, and relational read-across capability, but also broad data coverage within the chemical, genomics and toxicological information domains. In addition, given the proliferation of Internet resources, federated nature of these data, and inability of any single provider to adequately cover all data and information domains, these efforts require transparent and functional linkages of Internet data resources to be optimally useful.

In this research project I will:

1. Identify potential resources for a sufficient toxico-chemogenomic capability, i.e. gene expression resources, toxicological resources, and other chemically indexed sources.
2. Survey identified resources for chemical indexing and standardization for interoperability.
3. Develop methods to address deficiencies identified during the survey of identified resources.

4. Consider the limitations and potential of the resulting toxico-chemogenomics capability, specifically in reference to the use of public microarray data.

The goal of this project is to clearly identify major, public toxico-chemogenomic resources including gene expression, toxicology, and other chemically indexed biologically relevant resources and provide functional linkages between them as well as interoperable standardization between gene expression resources to increase the public toxico-chemogenomic capability.

RESEARCH CHAPTERS

Chapter 2- Survey of the Chemical Landscape of Public Gene Expression Databases for Toxicogenomic Applications

In this chapter, I will discuss the results of a survey of toxicogenomic resources from the standpoint of interoperability, standardization and chemical indexing, specifically gene expression resources. This will serve to highlight limitations of existing resources, particularly with regard to lack of sufficient chemical annotation of treatment-related microarray experiments in the two largest public microarray repositories. Having defined the problem, I will propose adoption of a set of Standard Genomics Fields that will index and enable cross comparison of experimental content in the main public microarray resources, including from a chemical perspective.

Chapter 3- Chemical Indexing of Experiments in GEO and ArrayExpress

In this chapter, I will implement the proposed Standard Genomics fields from Chapter 2 and discuss the methods used to chemically index the two main public gene expression repositories, ArrayExpress and GEO. The chemical index files resulting from this work allow, for the first time, assessment of the total chemical-experiment content, and the chemical treatment-related experimental content within the two repositories. Major products resulting from this project are curated chemical index

files for ArrayExpress and GEO published in coordination with the DSSTox Database Network project: ARYEXP, ARYEXP_Aux, GEOGSE and GEOGSE_Aux. I will also discuss the incorporation of these files into the DSSTox Structure Browser and PubChem, and enhanced linkages created to other chemically indexed data as a result of this research project.

Chapter 4- Considerations for the analysis and use of public gene expression-microarray data

In this chapter, I will discuss the considerations of public microarray data associated with toxico-chemogenomics. A review of gene expression analysis is presented. In addition, a demonstrative analysis is described evaluating statistical decision-making and processing at each step that occurs when integrating multiple gene expression experiments. Insights are provided into the use of public gene expression data.

Chapter 5- Conclusions and Future Work

Here the lessons learned and works completed are summarized. Implications of this research project are presented.

REFERENCES

Ahuja YR, Vijayalakshmi V, Polasa K. 2007. Stem cell test: A practical tool in toxicogenomics. Toxicology 231(1):1-10; doi: 10.1016/j.tox.2006.11.060.

Alternative Toxicology. 2008a. Toxicity Endpoints & Tests. Available: http://www.alttox.org/ttrc/toxicity-tests/ [accessed 12 November 2008]

Alternative Toxicology. 2008b. (Q)SAR. A review of QSAR as an alternative toxicity testing method. Available: http://www.alttox.org/ttrc/emerging-technologies/qsar [accessed 16 November 2008].

Alternative Toxicology. 2008 c. –Omics, Bioinformatics & Computational Biology: A Review of Omics technologies as an alternative for traditional toxicity measurement. Available: http://www.alttox.org/ttrc/emerging-technologies/-omics [accessed 16 October 2008].

Anderle P, Duval M, Dradhici S, Kuklin A, Littlejohn TG, Medrano JF et al. 2003.Gene expression databases and data mining. BioTechniques:36-44.

Argraves GL, Jani S, Barth JL, Argraves WS. 2005. ArrayQuest: A web resource for the analysis of DNA microarray data. BMC Bioinformatics 6:287; doi: 10.1186/1471-2105-6-287.

Ball C, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P et al. 2004. Standards for microarray data: An open letter. Environ Health Perspect 112(12):A666-7.

Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, Cunningham ML, Deng S, Dressman HK, Fannin RD, Farin FM, Freedman JH, Fry RC, Harper A, Humble MC, Hurban P, Kavanagh TJ, Kaufmann WK, Kerr KF, Jing L, Lapidus JA, Lasarev MR, Li J, Li YJ, Lobenhofer EK, Lu X, Malek RL, Milton S, Nagalla SR, O'malley JP, Palmer VS, Pattee P, Paules RS, Perou CM, Phillips K, Qin LX, Qiu Y, Quigley SD, Rodland M, Rusyn I, Samson LD, Schwartz DA, Shi Y, Shin JL, Sieber SO, Slifer S, Speer MC, Spencer PS, Sproles DI, Swenberg JA, Suk WA, Sullivan RC, Tian R, Tennant RW, Todd SA, Tucker CJ, Van Houten B, Weis BK, Xuan S, Zarbl H; Members of the Toxicogenomics Research Consortium. 2005. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods. 2: 351–356. doi:10.1038/nmeth0605-477a

Barnes, Donald. 2006. Pharmacology, Chapter 1: Introduction to Toxicology. Cambridge, Massachusetts: Elsevier Health Sciences, 1-18.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C et al. 2007. NCBI GEO: Mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35(Database issue):D760-5; doi: 10.1093/nar/gkl887.

Bayat A. 2002. Science, medicine, and the future of Bioinformatics. BMJ. 324:1018-1022.

Bhogal N, Grindon C, Combes R, Balls M. 2005. Toxicity testing: Creating a revolution based on new technologies. Trends Biotechnol 23(6):299-307; doi: 10.1016/j.tibtech.2005.04.006.

Boyle J. 2005. Gene-Expression Omnibus integration and clustering tools in SeqExpress. Bioinformatics 21(10):2550-2551; doi: 10.1093/bioinformatics/bti355.

Brazma A, Parkinson H, ArrayExpress team E. 2006b. ArrayExpress service for reviewers/editors of DNA microarray papers. Nat Biotechnol 24(11):1321-1322; doi: 10.1038/nbt1106-1321.

Burgoon LD, Boutros PC, Dere E, Zacharewski TR. 2006. dbZach: A MIAME-compliant toxicogenomic supportive relational database. Toxicol Sci 90(2):558-568; doi: 10.1093/toxsci/kfj097.

Burgoon LD, Zacharewski TR. 2007. dbZach toxicogenomic information management system. Pharmacogenomics 8(3):287-291; doi: 10.2217/14622416.8.3.287.

Burgoon LD. 2007. Clearing the standards landscape: The semantics of terminology and their impact on toxicogenomics. Toxicol Sci 99(2):403-412; doi: 10.1093/toxsci/kfm108.

Bustin, S.A. 2000. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J. Mol. Endocrinol. 25, 169-193.

Causton, H.C., Quackenbush, J., Brazma, Alvis. 2003. Microarray Gene Expression Data Analysis: A Beginner's Guide. Malden, MA: Blackwell Publishing. 1-130.

Collins FS, Gray GM, Bucher JR. 2008. Transforming environmental health protection. Science 319(5865):906-907.

Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ. 2008. Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res; doi: 10.1093/nar/gkn580.

Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M. 2008. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344-D350.

DeRisi, J.L. , Iyer, V.R., Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science. 278: 680-686.

DevTox. 2005. Developmental Toxicology Nomenclature Homepage. Available:http://www.devtox.org/index.htm [accessed 7 October 2008].

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci 95(1):5-12; doi: 10.1093/toxsci/kfl103.

EBI (European Bioinformatics Institute). 2008b. ArrayExpress Programmatic Access. Available: http://www.ebi.ac.uk/microarray/doc/help/programmatic_access.html (ArrayExpress) [accessed 7 October 2008].

EPA (Environmental Protection Agency) ACToR. 2008. Aggregated Computational Toxicology Resource (ACTor) Homepage. Available: http://www.epa.gov/actor/ [accessed 7 October 2008].

EPA (Environmental Protection Agency) DSSTox. 2008a. Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network Homepage. Available: http://www.epa.gov/ncct/dsstox/index.html [accessed 7 October 2008].

EPA (Environmental Protection Agency) Science Policy Council. 2004. EPA 100/B-04/002- Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA Final. Available: http://www.epa.gov/osa/genomics.htm [accessed 7 October 2008].

FDA (Food and Drug Administration) Center for Drug Evaluation and Research. 2005. MAPP 4180.3 – Processing and Reviewing Voluntary Genomic Data Submissions (VGDs). Available: http://www.fda.gov/Cder/genomics/VGDS.htm [accessed 7 October 2008].

FDA (Food and Drug Administration). 2008. The 111 routes of administration for substances. Available: http://fda.gov/cder/dsm/DRG/drg00301.htm [accessed 12 November 2008].

Fielden MR, Halbert DN. 2007. Iconix biosciences, inc. Pharmacogenomics 8(4):401-405; doi: 10.2217/14622416.8.4.401.

Fielden MR, Kolaja KL. 2006. The state-of-the-art in predictive toxicogenomics. Curr Opin Drug Discov Devel 9(1):84-91.

Fostel J, Choi D, Zwickl C, Morrison N, Rashid A, Hasan A et al. 2005. Chemical Effects in Biological Systems--data dictionary (CEBS-DD): A compendium of terms for the capture and integration of biological study design description, conventional phenotypes, and 'omics data. Toxicol Sci 88(2):585-601; doi: 10.1093/toxsci/kfi315.

Fostel JM, Burgoon L, Zwickl C, Lord P, Corton JC, Bushel PR et al. 2007. Toward a checklist for exchange and interpretation of data from a toxicology study. Toxicol Sci 99(1):26-34; doi: 10.1093/toxsci/kfm090.

Fostel JM. 2008. Towards standards for data exchange and integration and their impact on a public database such as CEBS (Chemical Effects in Biological Systems). Toxicol Appl Pharmacol; doi: 10.1016/j.taap.2008.06.015.

Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA et al. 2005. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. J Biotechnol:219-244.

Gibson, U.E., Heidi, C.A., Williams, .P.M. 1996. A novel method for real time quantitative RT-PCR. Genome Res. 6, 995-1001.

Ginzinger, D.G. 2002. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp. Hematol.: 30:503-512.

Goetz AK, Bao W, Ren H, Schmid JE, Tully DB, Wood C et al. 2006. Gene expression profiling in the liver of CD-1 mice to characterize the hepatotoxicity of triazole fungicides. Toxicol Appl Pharmacol 215(3):274-284; doi: 10.1016/j.taap.2006.02.016.

Gomase VS, Tagore S, Kale KV. 2008. Microarray: An approach for current drug targets. Curr Drug Metab 9(3):221-231.

Gyri, S.P., Rochon, Y, Franza, B.R., Aebersold, R. 1999. Correlation between protein and mRNA abundance in yeast. Molecular and Cellular Biology. 19:1720-1730.

Hamadeh HK, Amin RP, Paules RS, Afshari CA. 2002. An overview of toxicogenomics. Curr Issues Mol Biol 4(2):45-56.

Hirabayashi Y, Inoue T. 2002. Toxicogenomics--a new paradigm of toxicology and birth of reverse toxicology. Kokuritsu Iyakuhin Shokuhin Eisei Kenkyusho Hokoku (120)(120):39-52.

Hodgson, Ernest. A textbook of Modern Toxicology, Third Edition, Chapter 1: Introduction to Toxicology. Hoboken, New Jersey: John Wiley & Sons, Inc, 3-12.

Houck KA, Kavlock RJ. 2008. Understanding mechanisms of toxicity: Insights from drug discovery research. Toxicol Appl Pharmacol 227(2):163-178; doi: 10.1016/j.taap.2007.10.022.

Human Genome Project (HGP). 2003. Genome glossary: A glossary for the human genome provided by the US Department of Energy (DOE). Available: http://www.ornl.gov/sci/techresources/Human_Genome/glossary/glossary.shtml [accessed 17 November 2008].

ILSI (International Life Sciences Institute) DevTox. 2008. Improving the Use of Toxicity Data in Statistically Based Structure-Activity Relationship Models for Developmental Toxicity, http://rsi.ilsi.org/Projects/devtoxsar.htm [accessed 19 October 2008].

Ivliev AE, 't Hoen PA, Villerius MP, den Dunnen JT, Brandt BW. 2008. Microarray retriever: A web-based tool for searching and large scale retrieval of public microarray data. Nucleic Acids Res; doi: 10.1093/nar/gkn213.

Judson R, Richard A, Dix D, Houck K, Elloumil F, Martin M, Cathey T, Transue T, Spencer R, Wolf M. ACToR – Aggregated Computational Toxicology Resource, Toxicol Appl Pharmacol. 2008, in press.

Julien, E., Willhite, C. C., Richard, A. M., and DeSesso, J. M. 2004. Challenges in constructing statistically-based SAR models for developmental toxicity. Birth Defects Res. Part A 70:902–911.

Kavlock RJ, Ankley G, Blancato J, Breen M, Conolly R, Dix D et al. 2008. Computational toxicology--a state of the science mini review. Toxicol Sci 103(1):14-27; doi: 10.1093/toxsci/kfm297.

Knudsen KB, Singh AV, Knudsen TB. 2005. Data input module for birth defects systems manager. Reprod Toxicol 20(3):369-375; doi: 10.1016/j.reprotox.2005.04.002.

Larsson O, Sandberg R. 2006. Lack of correct data format and comparability limits future integrative microarray research. Nat Biotechnol 24(11):1322-1323; doi: 10.1038/nbt1106-1322.

Martin MT, Judson RS, Reif DM, Dix DJ. In press. Classifying Chemicals Based on Toxicity Profiles from the U.S. EPA ToxRef Database. Environ Health Perspect.

Mattes WB, Pettit SD, Sansone SA, Bushel PR, Waters MD. 2004. Database development in toxicogenomics: Issues and efforts. Environ Health Perspect. 112(4): 495-505.

McGee, P. 2006. Genomics 2007: The next step in the sequence: An article on the next steps after the sequencing of genomes. Available: http://www.dddmag.com/resequencing-genes-could-payoff.aspx [accessed 16 November 2008].

Nature Glossary . 2008. A glossary of Genetic Terms. Available: http://www.nature.com/nrg/journal/v6/n4/glossary/nrg1575_glossary.htm [accessed 20 October 2008].

NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus). 2008b. GEO Programmatic Access. Available: http://www.ncbi.nlm.nih.gov/projects/geo/info/geo_paccess.html [accessed 7 October 2008].

NCBI (National Center for Biotechnology Information). 2008. Homepage for the National Center for Biotechnology Information. Available: http://ncbi/nlm/nih/gov [accessed 16 November 2008].

NCI60. 2008. A listing dataset of several thousand compounds tested on 60 cancer cell lines. Available: http://dtp.nci.nih.gov/docs/cancer/cancer_data.html [accessed 17 November 2008].

Nolan, T., Hands, R., Bustin, S. 2006. Quantification of mRNA using real-time RT-PCR. Nature Protocols: 1(3):1559-1582.

NRC (National Research Council). 2008. Toxicity Testing in the Twenty-first Century: A Vision and a Strategy. Committee on Toxicity and Assessment of Environmental Agents (Online Book). Available: http://www.nap.edu/catalog/11970.html [accessed 16 November 2008]

Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. 1999. Microarrays and toxicology: The advent of toxicogenomics. Mol Carcinog 24(3):153-159.

Parker CN, Shamu CE, Kraybill B, Austin CP, Bajorath J. 2006. Measure, mine, model, and manipulate: The future for HTS and chemoinformatics? Drug Discov Today 11(19-20):863-865; doi: 10.1016/j.drudis.2006.08.006.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A et al. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35(Database issue):D747-50; doi: 10.1093/nar/gkl995.

PubChem. 2008. PubChem HomePage. Available: http://pubchem.ncbi.nlm.nih.gov [accessed 7 October 2008].

Richard A, Yang C, Judson R. 2008. Toxicity data informatics: Supporting a new paradigm for toxicity prediction. Tox Mech Meth 18:103-118.

Richard AM, Gold LS, Nicklaus MC. 2006. Chemical structure indexing of toxicity data on the internet: Moving toward a flat world. Curr Opin Drug Discov Devel 9(3):314-325.

Salter AH. 2005. Large-scale databases in toxicogenomics. Pharmacogenomics 6(7):749-754; doi: 10.2217/14622416.6.7.749.

Sean D, Meltzer PS. 2007. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23(14):1846-1847; doi: 10.1093/bioinformatics/btm254.

Segota I, Bartonicek N, Vlahovicek K. 2008. MADNet: Microarray database network web server. Nucleic Acids Res 36(Web Server issue):W332-5; doi: 10.1093/nar/gkn289.

Shalon D, Smith SJ, Brown PO 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6: 639–645. doi:10.1101/gr.6.7.639

Singh AV, Knudsen KB, Knudsen TB. 2005. Computational systems analysis of developmental toxicity: Design, development and implementation of a Birth Defects Systems Manager (BDSM). Reprod Toxicol 19(3):421-439; doi: 10.1016/j.reprotox.2004.11.008.

Singh AV, Rouchka EC, Rempala GA, Bastian CD, Knudsen TB. 2007. Integrative database management for mouse development: Systems and concepts. Birth Defects Res C Embryo Today 81(1):1-19; doi: 10.1002/bdrc.20089.

Sklarew, Myra. 1993. Toxicity Testing in Animals: Alternative Models. Available: http://www,ehponline.org/docs/1193/101-4/focus.html [accessed 16 November 2008].

Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology. 98:503-517.

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, et al. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnol 26(8):889-896.

Tice RR, Fostel J, Smith CS, Witt K, Freedman JH, Portier CJ, et al. 2007. The National Toxicology Program high throughput screening initiative: current status and future directions [Abstract]. Toxicologist 46:246.

Vrana KE, Freeman WM, Aschner M. 2003. Use of microarray technologies **in** *toxicology* research. Neuro Toxicology. 24:321-332.

Wagner BK, Kitami T, Gilbert TJ, Peck D, Ramanathan A, Schreiber SL et al. 2008. Large-scale chemical dissection of mitochondrial function. Nat Biotechnol 26(3):343-351; doi: 10.1038/nbt1387.

Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A et al. 2003. Systems toxicology and the chemical effects in biological systems (CEBS) knowledge base. EHP Toxicogenomics 111(1T):15-28.

Watson, J.D. and Crick, F.H.C. 1953. Molecular structure of the nucleic acids: A structure for deoxyribose nucleic acid. Nature. 171:737-38.

Weaver, R. Molecular Biology: Third Edition. 2005. New York, New York: McGraw-Hill. 1-423.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V et al. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36(Database issue):D13-21; doi: 10.1093/nar/gkm1000.

Williams AJ. 2008. Public chemical compound databases. Curr Opin Drug Discov Devel 11:393-404.

Xia M, Huang R, Witt KL, Southall N, Fostel J, Cho MH et al. 2008. Compound cytotoxicity profiling using quantitative high-throughput screening. Environ Health Perspect 116(3):284-291; doi: 10.1289/ehp.10727.

Yang C, Arnby CH, Arvidson K, Aveston S, Benigni R, Benz RD et al. 2008. Understanding genetic toxicity through data mining: The process of building knowledge by integrating multiple genetic toxicity databases. Tox Mech Meth 18(277):295.

Yang C, Benz RD, Cheeseman MA. 2006a. Landscape of current toxicity databases and database standards. Curr Opin Drug Discov Devel 9(1):124-133.

Yang C, Richard AM, Cross KP. 2006b. The art of data mining the minefields of toxicity databases to link chemistry to biology. Curr Comput-Aided Drug Design 2(2):135-150.

Young, R.A. 2000. Biomedical discovery with DNA arrays. Cell. 102, 9-16.

Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. 2008b. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus (GEO). Bioinformatics. in press; doi:10.1093/bioinformatics/btn520

Zhu Y, Zhu Y, Xu W. 2008a. EzArray: A web-based highly automated affymetrix expression array data management and analysis system. BMC Bioinformatics 9:46; doi: 10.1186/1471-2105-9-46.

FIGURES

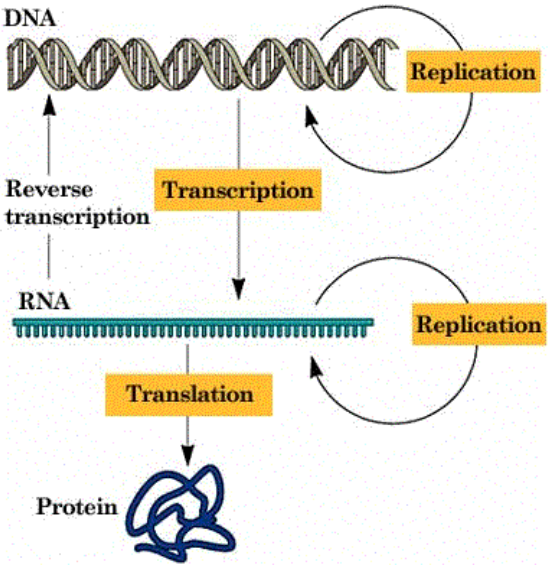


Figure 1-1 Central Dogma Theory. The Central Dogma Theory explains the transition of information between DNA, mRNA, and protein [Adapted from http://cats.med.uvm.edu/cats_teachingmod/microbiology/courses/genomics/images_new/1_centraldogma_wisc_13.jpg].

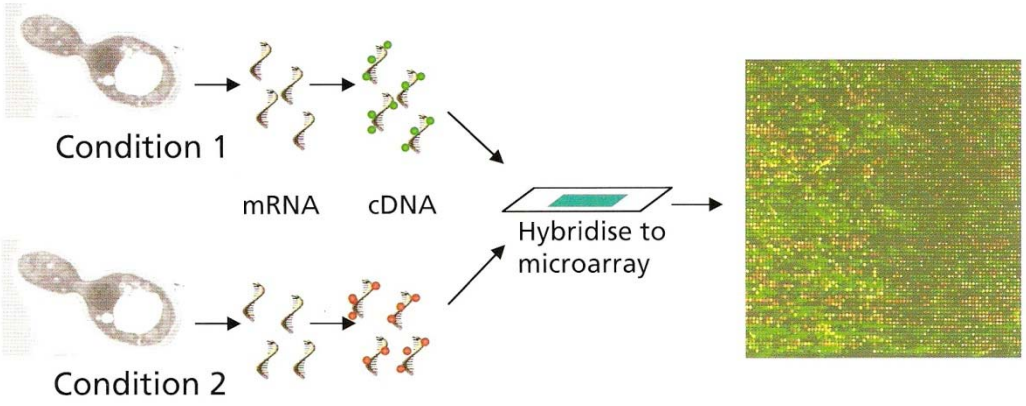


Figure 1-2. A typical 2 color, Cy3 and Cy5, microarray experiment, where representative mRNA from condition 1 cells is labeled with a green fluorescent dye and representative mRNA from Condition 2 cells from condition 2 is labeled with a red fluorescent dye. Both labeled mRNA extracts are hybridized to the same microarray and scanned [Adapted from Causton et al. 2003 page 84].

CHAPTER 2

SURVEY OF THE CHEMICAL LANDSCAPE OF PUBLIC GENE EXPRESSION DATABASES FOR TOXICOGENOMIC APPLICATIONS

ABSTRACT

Toxicogenomics experiments that more broadly probe chemical space have the potential to reveal gene expression patterns that are predictive of a toxicological response. A publicly available toxicogenomics capability for supporting predictive toxicology depends on availability of gene expression data for chemical treatment scenarios, the ability to locate and aggregate such information by chemical, and broad data coverage within chemical, genomics and toxicological information domains. This capability also depends on common genomics standards, protocol description, and functional linkages of diverse public Internet data resources. We present a survey of public genomics resources from these vantage points and conclude that, despite progress in many areas, the current state of many public microarray databases is inadequate for supporting interoperability and linkages across diverse data domains in support of toxicogenomics studies. In particular, lack of adoption of minimal chemical annotation standards, difficulty in locating experiments pertaining to chemical treatment, and inability to assess the chemical scope of current public microarray data inventories limits use of these resources for toxicogenomics in support of predictive toxicology. We propose a set of standard genomics fields for indexing of experimental study records, aligned with MIAME guidelines, that will enable cross referencing and comparison of total experimental content in the largest public microarray resources, Gene Expression Omnibus (GEO) and ArrayExpress. In addition, we propose chemical indexing standards be adopted by public microarray data repositories to enable assessment of chemical diversity and coverage of microarray experiments, as well as to facilitate linkages to external, chemically indexed bioassay information.

INTRODUCTION

Toxico-chemogenomics represents a paradigm shift applied to the field of toxicogenomics. Traditionally toxicogenomics approaches generally involve a single study that surveys the adverse effects of a poison on an organism(s) through toxicity testing as well as supporting omics technologies, mainly transcriptomics or gene expression (Gomase et al. 2008; Hamadeh et al. 2002; Hirabayshi and Inoue 2002). On the other hand, toxico-chemogenomics represents the creation of toxicogenomic datasets through linkages across chemical space within pubic toxicological, biological, and gene expression resources. This paradigm shift is necessary in order to address the shortage of public toxicogenomics data for the purpose of toxicogenomic investigation, regulatory science, methodology development, and contributions to predictive toxicology. The requirements for creating a toxico-chemogenomics public capability, as discussed in Chapter 1, are common data standards, protocols, and relational read-across capability, as well as broad coverage within the chemical, gene expression, and toxicological information domains. Given the central and prominent role of the chemical to these endeavors and essential need for functional linkages across publicly available chemically indexed resources , the key components of a successful toxico-chemogenomics effort are standardized interoperability and chemical indexing of public gene expression and toxicological resources.

In the toxicological area, significant progress has been achieved in the standardization and chemical indexing of toxicological resources over the past several years. The Distributed Structure-Searchable Toxicity (DSSTox) database network has been a particularly influential contributor to these developments, providing standardized, interoperable, chemically indexed toxicity databases and resources. As an undergraduate research associate, I helped to conceptualize, and was a major coauthor and contributor to the initial DSSTox project proposal in early 2002. The project was

designed to specifically address the disconnect and lack of chemical indexing and interoperability among toxicity data resources, particularly as these limitations impeded progress in structure-activity relationship (SAR) toxicity prediction modeling (Richard and Williams 2002; Richard et al. 2002). We particularly noted the need for a way to "read-across" public toxicity database networks on the basis of standardized chemical structure description. The standardized chemical indexing of available toxicity data at that time posed a unique problem. In spite of the fact that a large percentage of toxicity experiments primarily focus on the adverse effects of one or a few chemicals, there was a lack of attention paid to the publication of details about the chemical substance and very few publications included chemical structures. In addition there was a great deal of unstructured textual content associated toxicological results centered on a chemical or class of chemicals that did not lend itself to standardized comparisons of experimental results for different chemicals.

The primary challenge for the DSSTox effort was to represent standardized summary activity values from the textual information and to combine this with an accurate and unique representation of the chemical structure. Toxicity data resources are primary indexed by chemical name or chemical abstracts registry number (CAS-RN 2008) that are non-unique representations of the chemical, i.e., there can be more than one CAS-RN or chemical name for the same chemical. Chemical structure provides a unique and scientifically meaningful description, which is essential when looking across multiple resources to group toxicity results for the same or structurally similar chemicals. A chemical structure can be represented either in 2-dimensional or 3-dimensional form as an MDL Molfile, a publicly available industry-standard file format that holds the information about the atoms, bonds, connectivity and coordinates of the molecule (MDL 2008). The related structure-data file (SD file) format is used to store the chemical Molfile along with textual data fields (i.e., toxicological information) for a collection of molecules. To increase the interoperability of the toxicity data

resources, a set of DSSTox standard chemical fields were developed to represent the tested substance and its associated chemical structure in a common uniform way across all available toxicity data resources (Table 2-1).

Gene expression resources, being newer onto the scientific scene than traditional toxicology resources, have more fully embraced public availability, Internet hosting, and some level of standardization with respect to experimental description. These resources, however, are not typically chemically indexed and, therefore, present a more difficult problem than toxicological resources. Even when chemical treatment is a primary objective of the experiment the chemical layer associated with the experimental data has been entirely missing or largely neglected in the genomics field. Standardized chemical indexing is essential for aggregating data and systematically relating chemical property and effects information across the diverse data domains contributing to toxicogenomics. Furthermore, the ability to query, relate, and aggregate information by chemical and across chemical space is essential to the goal of chemical screening and toxicity assessment (Dix et al., 2007; Yang et al., 2008; Richard et al., 2008).

In this chapter, I will broadly survey the current state of public microarray resources from the above vantage points. Despite progress in many areas, I determined the present state of public microarray databases to be inadequate for supporting interoperability and linkages across diverse data domains in support of toxicogenomics study. Particularly noteworthy is the lack of chemical standards or indexing of the largest public microarray resources and, as a result, the effective isolation of these resources from the growing inventories of chemically indexed bioassay information of potential relevance to toxicology (Richard et al., 2006). I propose here a set of standard genomics fields for indexing of experimental study records (with regard to species, protocol, raw data availability, etc.) that will enable cross referencing and comparison of total experimental content in the largest public

microarray resources. In addition, I propose chemical indexing standards be applied to public microarray data repositories to enable assessment of chemical diversity and coverage of microarray experiments, as well as to facilitate linkages to external, chemically indexed bioassay information. In chapter 3 of this dissertation, I implement these standards within the two largest public microarray resources: Gene Expression Omnibus (GEO) and ArrayExpress (Table 2-2) and describe the methods and challenges associated with this task.

METHODS

For the present purpose of assessing the relevance of public microarray resources to toxicogenomics and predictive toxicology, I considered the extent of experimental content pertaining to chemical treatment scenarios. These are cases in which study of gene expression changes induced by chemical treatment constituted the primary goal of the experiment. As a measure of interoperability between data resources, I examined the standardization of terminology and data accessibility, as well as the formatting of data, paying particular attention to specification of experimental protocols, such as animal/tissue/cell treatment, RNA extraction, microarray preparation, data import/export, and analysis. As a measure of chemical indexing, I examined experimental content pertaining to chemical treatment across public genomics resources, sufficiency of labeling for annotation of chemical exposure experiments, and inclusion of chemical identifier information and chemical structure annotation.

Over 42 public Internet resources housing microarray data of potential toxicogenomics relevance were initially identified from various categories (Microarray World list of databases, http://www.microarryaworld.com/DatabasePage.html; Leung's Microarray Software Comparison Page, http://ihome.cuhk.edu.hk/~b400559/arraysoft_public.html). Of these, the two primary public

resources, the National Center for Biotechnology Information's (NCBI) GEO and the European Bioinformatics Institute's (EBI) ArrayExpress, serve as the main repositories of gene expression data published in the scientific literature (Table 2-2). In addition, the Center for Information Biology Gene Expression (CIBEX) database serves as a repository of gene expression data published in scientific literature, but does not yet contain a significant number of experiments. Four additional public genomics resources of potential toxicogenomics relevance were identified that contain or reference data gathered from chemical exposure experiments in one or more laboratories: the Environment, Drugs, and Gene Expression (EDGE) database, the Comparative Toxicogenomics Database (CTD), the Chemical Effects in Biological Systems (CEBS) knowledgebase, the Public Expression Profiling Resources (PEPR) database, and (Table 2-3). These resources differ greatly in terms of the types of data available and intended usage, and each has different degrees of standardization and read-across capability.

For initial inventory purposes, extraction of experimental description fields, and evaluating the location of chemical information within ArrayExpress and GEO, I used available web search options and programmatic access tools within each system. For ArrayExpress, a bulk download of all the data housed in the repository from the main web site was undertaken with a wildcard query in the *accession number* query box (i.e., to retrieve all experiments). For GEO, all data were downloaded from the GEO homepage in the GSE Series form. In addition, customized queries were constructed through programmatic access within ArrayExpress (XML format; http://www.ebi.ac.uk/microarray/doc/help/programmatic_access.html) and GEO (Entrez Utilities; http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html).

RESULTS

For present purposes, genomics resources are divided into Primary Genomics Resources and Secondary Genomics Resources. Primary genomics resources consist of the three MGED-approved gene expression repositories: ArrayExpress, GEO, and CIBEX (Table 2-2). Secondary genomics resources consist of other public gene expression databases/repositories (Table 2-3) of particular interest for toxicogenomics. A selection of public cheminformatics resources of potential relevance and utility for building a public toxico-cheminformatics capability are listed in Table 2-4.

Primary genomics resources (Table 2-2) serve as officially sanctioned repositories for public gene expression data referred to in the scientific literature (Salter, 2005; Mattes et al., 2004). The CIBEX database is included for completeness sake, but is small and currently of little relevance to toxicogenomics (Tateno and Ikeo, 2004). GEO and ArrayExpress are the largest primary public repositories of gene expression data associated with the scientific literature. In addition, they both are MIAME-supportive databases, meaning that they accept all information about an experiment set forth by the MIAME guidelines; however, they do not actually require this information.

Secondary genomics resources contain genomic-related data but generally have more limited content and are designed for more specialized purposes and applications (Table 2-3). The Comparative Toxicogenomics Database (CTD) is worthy of brief mention for its toxicogenomics relevance, but is not a traditional genomics database (Davis et al., 2008). Rather, it is a database of curated relationships between chemicals, genes, and diseases mined from journal articles, providing linkages to genomics data in other public resources, specifically EDGE, and text-mineable access to the toxicogenomic literatureCEBS is unique in its ability to characterize both the toxicological and genomics (transcriptomic, proteomic, and metabolomic) aspects of toxicogenomics experiments (Waters et al., 2008). However, it is not included as a primary genomics resource here because of its

more specific focus on environmentally relevant biological data. Each of the secondary genomics resources provides valuable information and capabilities, and most of the data in these resources have also been deposited in either GEO or ArrayExpress.

A selection of cheminformatics resources is listed in Table2-4 to highlight the diversity and range of resources and capabilities that can potentially contribute to toxicogenomic investigations (Mattingly, 2008). This list includes chemically indexed resources ranging from local installation, toxicogenomics database platforms, such as ArrayTrack, to resources spanning millions of compounds and containing thousands of *in vitro* and *in vivo* bioassay results, such as NCBI's PubChem. These varied resources are rapidly expanding and increasing linkages to biological and toxicological datadbZach, a laboratory tool offered for local installation, is of specific interest as a modular MIAME-compliant, toxicogenomic-supportive relational database designed to facilitate data integration, analysis, and sharing in support of toxicogenomics studies (Burgoon et al., 2006). dbZach consists of several subsystems for the standardization of all data elements of a toxicogenomics experiment as well as traditional toxicological experiments and, additionally, has built-in functionality for data import and export of both raw data and processed data.

Table 2-5 presents a comparison of the primary and secondary genomics resources with respect to the types of gene expression data stored, toxicological focus, the formats of data available for download (raw or processed), the ability to query data, the ability to import or export experimental data, and programmatic access to all or a large portion of the database. Table 2-6 compares the same resources from the standpoint of being chemically indexed (i.e., chemical identifiers are required and entered in standard fields), MIAME-supportive, and standardized with respect to various experimental descriptions. Additional details on selected resources of greatest interest for toxicogenomics are provided below.

ARRAYEXPRESS REPOSITORY

ArrayExpress is a public user-depositor data repository and MIAME-supportive public archive of microarray data consisting of two parts - ArrayExpress Repository and the ArrayExpress Data Warehouse (Brazma et al., 2006; Parkinson et al., 2007; Rustici et al., 2008). The ArrayExpress Repository currently exceeds 6900 experiments and was developed in response to the MGED society's recommendation for a public MIAME-compliant archive of microarray/gene expression data and stores experimental data throughout the publication process. It is indexed by Experiment Array Design, and Protocol, and experiments can be queried by Keyword, Experimental Accession Number, Species, Experiment Type and Factors, Author, Laboratory, and Publication information (http://www.ebi.ac.uk/microarray-as/aer/entry). Repository data are cataloged, assessed for completeness, and assigned a MIAME score that represents the degree of MIAME compliance (Brazma et al., 2006). The ArrayExpress Data Warehouse is based on more limited processed data results from the ArrayExpress Repository, containing 740 Expression Profiles currently, and allows users to browse curated datasets from both a gene- and/or experiment-centric view (Parkinson et al., 2007).

ArrayExpress is not chemically indexed, nor does it contain additional information about the chemical tested other than the infrequently provided Chemical Abstracts Service Registry Number (CASRN) or Chemicals of Biological Interest (ChEBI) number (http://www.ebi.ac.uk/chebi/). Chemical information may be located in the user-supplied protocols and free-text experimental description, or can be searched with the advanced query tools from ArrayExpress, including a keyword or text search in the Description field in "Query for Experiments". These can also be combined with specifications of <Experiment type>= "compound treatment" or "dose response", but these latter annotations are optional and not consistently applied by depositors to all chemical treatment experiments in the database. Chemical information can also be embedded within the

ArrayExpress Sample-Data Relationship File (SDRF)

(http://tab2mage.sourceforge.net/docs/sdrf.html).

The ArrayExpress Accession Number Code, TOXM, can be used to identify toxicogenomics experiments. However, the designation must be assigned by the data submitter and, as a result, tends to be under-utilized (fewer than 25 experiments are assigned this code). Where the TOXM label for a toxicogenomics experiment is used, however, typically more chemical identifier information, such as a CASRN and/or a ChEBI number, is provided by the submitter along with additional information recommended by the MIAME/Tox initiative (http://www.ebi.ac.uk/tox-miamexpress), i.e., minimum information about a toxicogenomics microarray experiment.

GEO

GEO is a large public user-depositor data repository (containing over 10,000 experiments at the time of this writing) housed at the National Center for Biotechnology Information (NCBI) (Barrett et al., 2007; Wheeler et al., 2008). GEO is a MIAME-supportive platform for the public use and dissemination of gene expression data generated by high-throughput methodologies. In GEO, raw and/or processed data can be exported through the ftp site as well as through the main GEO Series website (Table 2-2). User information, however, is entered using a free-text format that is subsequently curated. GEO allows for a wide range of informed queries with the Preview/Index window, where users can select data based on choices for each attribute of the experiment. The GEO repository has three key components: "Platform", "Sample", and "Series". "Platform" provides a description of the array used in the experiment, as well as a data table defining the array template (Barrett et al., 2007). The data table contains hybridization measurements for each element of the corresponding platform. "Sample" provides a description of the biological source and the experimental protocols. "Series" defines a set of related samples considered to be part of a study and

describes the overall study aim and design. GEO has a complex, hierarchical structure that works with the Entrez tools and utilities, allowing one to query by submitter, organism, platform, sample type, sample titles, and release date. Similar to ArrayExpress, GEO hosts a smaller warehouse-type addition named "GEO Datasets and Profiles" containing processed, curated datasets that can be explored from both a gene- and/or experiment-centric view (Barrett and Edgar, 2006).

Also, similar to Array Express, GEO is not chemically indexed nor does it consistently contain information about the chemical tested. Of the more than 9000 experiments currently housed in GEO, there is no easy or reliable way to identify a chemical exposure-related experiment. Chemical names can be located in the submitter-deposited GEO Data Series (GDS) record fields Title, Summary, Citation, or Samples field, and are not consistently present in any single field. Chemicals names are provided by the submitter, are rarely accompanied by CASRN or ChEBI identifiers, and do not undergo curation or review. Hence, as is the case with ArrayExpress, the chemical names are highly variable, prone to errors and misspellings, and frequently incorporate abbreviations.

CIBEX

The Center for Information Biology gene Expression (CIBEX) database is a Japanese gene expression MIAME supportive user-depositor system (Tateno and Ikeo, 2004). It is included in this discussion for completeness, but currently does not contain significant chemical content. CIBEX's primary objective is to serve as a home for the data of experimenters from Asian countries. CIBEX is one of the three main MGED-approved repositories, but has grown at a rate much slower than that of ArrayExpress and GEO. However, the experimental protocol and detail standardization should be noted in CIBEX, where each record contains all MIAME details stored in an accompanying document. There is also a high level of curation and collaboration between CIBEX administrators and depositors. This allows for missing information to be identified before publication, as well as for

a higher level of standardization and accuracy. Currently, CIBEX contains 31 experiments, only one of which is a chemical exposure experiment. Whereas chemicals are clearly denoted in this single example (CBX14), standards for chemical indexing should be incorporated as the resource is expanded.

EDGE

The EDGE database (Hayes et al., 2005) is a closed (i.e., not open to public user-deposits of data), curated system designed for the comparison, analysis and distribution of toxicogenomics information in a relational format. EDGE is chemical treatment-centric and chemically indexed, with a toxicological focus. All experiments were performed in the Bradfield Laboratory using a standardized protocol involving custom cDNA arrays of minimally redundant hepatic clones, chosen through chemical exposure experiments with known hepatic toxicants: 2,3,7,8, tetrachlorodibenzo-*p* dioxin (TCDD), cobalt chloride, and phenobarbital. Non-redundancy of hepatic clones was verified in the 3' and 5' direction using RefSeq, UniGene, and GenBank. The experimental conditions include 22 chemical treatments, 4 control treatments, and 1 environmental stressor (fasting) over 1 mutant (circadian wild-type control). All chemical treatments were chosen for the express purpose of investigating hepatotoxicity in mice.

Despite its small size and limited focus, EDGE incorporates a high level of standardization and comparability across species, array, experimental protocol, and experimental details, and demonstrates how a fully relational database built on such data can facilitate toxicogenomics investigation. Although EDGE offers optimal standardization and relational elements for proof-of-concept in a public toxicogenomics database, EDGE is not a user-depositor system and currently lacks the tissue, species, and chemical diversity necessary for broader toxicogenomics exploration.

PEPR

Similar to EDGE, the PEPR database (Chen et al., 2004) is a closed, curated system designed to serve as a public resource of gene expression profile data generated in the same laboratory, using the same chip type for three species, and subject to the same quality and procedural controls. PEPR does not have a toxicological or toxicogenomics focus, but is aimed at providing a standardized warehouse for the analysis of time-series data. The high degree of standardization within PEPR grants users comparability across arrays without laboratory and array bias, much like EDGE. PEPR adheres to quality control and standard operating procedures (QC/SOP) and is indexed by Principle Investigator (PI), Tissue type, Experiment, and Organism, but has a very few chemical treatment-related experiments and lacks relational searching capabilities. However, the time-series query analysis tool (SGQT) enables the novel generation of graphs and spreadsheets showing the action of any transcript of interest over time. PEPR also differs from EDGE in the extensive data export options offered for both raw and processed data (Chen et al. 2004) (Table 2-6). Data can be exported from PEPR as raw image files (.dat), processed image files (.cel) and interpretation files (.txt). PEPR also has external links to GEO, where PEPR data are mirrored through an automated export/import process.

In PEPR, chemical information is stored in free-text fields such as the title, description, and array titles, similar to ArrayExpress and GEO. At the time of this writing, PEPR contained 72 experiments, of which 10 were determined to be chemical/environmental exposure experiments. Hence, PEPR currently covers very limited chemical space, but the SGQT tool for analysis of time-series microarray data, as well as the standardized chemical exposure experiments are of potential value for toxicogenomics studies.

CEBS

CEBS is a public user-depositor data repository with an explicit toxicological and toxicogenomics focus (Waters et al., 2008). CEBS can accommodate study design, timeline, clinical chemistry and histopathology findings, as well as microarray and proteomics data. Each experiment in CEBS pertains to a chemical/environmental exposure or a genetic alteration in reference to clinical or environmental studies. CEBS has a complementary functional component known as the Biomedical Investigation Database (BID) (https://dir-apps.niehs.nih.gov/arc/), which is a relational database used to load and curate study data prior to exporting to public CEBS. BID also aids in the capture and display of novel data, including PCR, and toxicogenomic-relevant fields, as used in ArrayExpress's TOXM designation http://www.ebi.ac.uk/tox-miamexpress). With its specific focus on toxicogenomics, attention to chemical indexing, and addition of the relational searching capability in the BID system, CEBS is becoming a leading toxicogenomics resource. CEBS is currently indexed by study and subject characteristics, such as environmental, chemical, or genetic stressor and stressor protocol, and includes observations on rat, mouse, and *Caenorhabditis elegans* (*C. elegans*).

The relational searching capability of CEBS allows users to look across toxicogenomics studies in a variety of ways, much like the EDGE database. However, unlike EDGE, CEBS is a user-depositor system striving for much larger coverage of chemical space in relation to chemical treatment experiments. To support robust relational searching for toxicogenomics, CEBS has the added task of capturing and systematizing user-deposited data pertaining to a study. This is accomplished by implementing a standardized and controlled "toxicogenomic" vocabulary, defined in part by the CEBS-Data Dictionary (CEBS-DD), and mapping of non-standard representations to a common format (Fostel et al., 2005). CEBS bridges the gap between an open-access, user-depositor system and a relational, curated database by instituting a high degree of standardization and data controls that

extend beyond MIAME guidelines. Public data can be imported into CEBS by the curation staff, and raw and/or processed data can be exported through the CEBS web-interface system.

CEBS is incorporating standards for chemical indexing in collaboration with the EPA Distributed Structure-Searchable Toxicity Database project (http://www.epa.gov/ncct/dsstox/) and, through this collaboration, will incorporate linkages to external public resources, such as PubChem. Additionally, local structure-searching capabilities will be accessed through the DSSTox Structure-Browser, linking users to chemically-indexed web pages on the CEBS website (Richard et al., 2008).

DISCUSSION

GEO and ArrayExpress are, currently, the largest and most important public repositories of gene expression data. Since publishing requirements in Europe and the US require data associated with microarray studies to be deposited in one of these two public resources, they will contain significant content of potential use for toxicogenomics. These two resources are similar to each other in programmatic access and general format, and have significant experimental and chemical overlapping content – by our current estimation, ArrayExpress and GEO have more than 3700 general experiments in common (or greater than 40% of their total content). Yet despite the central importance of these two resources, other than cases where experimental content is deposited in both systems, there are no formal connections between the two databases or their current capabilities, no mechanism for comparing overall experimental content, and no way to assess or uniformly access the chemical treatment-related experimental inventory.

Web-based queries and programmatic access were used in the present study to extract current experimental content from the ArrayExpress Repository and GEO Series, and to identify corresponding experiment annotation fields in the two systems that could be mapped to a common

field to enable comparisons across the two inventories. I propose a set of 14 Standard Genomics Fields in Table 2-7 to serve this purpose and to confer "read-across" capability across the two inventories. All but two of these fields map to existing data fields in both GEO and ArrayExpress and, thus, are straightforward to implement. One field, "Experiment_URL", contains a URL link to enable outside Internet access directly to the experiment accession summary page in either ArrayExpress or GEO. The last field, "Chemical_StudyType", pertains to the currently missing chemical annotation layer for gene expression experiments in both resources. Prior to introducing formal standards for chemical annotation in GEO and ArrayExpress (which would include, e.g., standardized chemical names and structures), chemical-associated experiments first must be identified and labeled within the two resources. The proposed field "Chemical_StudyType" and it's allowed entries (e.g., Reference, Treatment, Vehicle, etc.) would serve to label an experiment according to the purpose of the chemical in relation to the experiment. This would also serve to isolate the subset of chemical associated experiments of greatest potential interest for toxicogenomics study, i.e. "Chemical_StudyType = Treatment".

Finally, to underscore the inadequacies of the current chemical information layer in GEO and ArrayExpress, I present a single example. In ArrayExpress (website accessed on 14November2008), a keyword search for "estrogen" in the website query form yielded 131 hits, with some of these referring to the estrogen receptor or estrogen-mediated mechanisms, others finding the words "phytoestrogen" or "xenoestrogen". These include an undetermined subset of experiments focused on gene expression changes resulting from estrogen treatment. A search for the abbreviation "E2" yields 702 hits, only a few of which pertain to estrogen treatment. Similarly, a search for "estradiol" yields 71 hits, whereas a search for "17 beta estradiol" yields 36 hits and the more correct "17-beta estradiol" yields only 4 hits.

The full extent of the potential problem of using non-standardized chemical names for searching can be inferred from the National Library of Medicine's (NLM) ChemID resource (http://chem2.sis.nlm.nih.gov/chemidplus/chemidlite.jsp), which lists 150 synonyms for "estradiol", including numerous trade names (e.g., Altrad) and valid chemical names (e.g., 3,17-epidihydorxyestratriene).

CONCLUSIONS

It is helpful to envision idealized elements of a public toxicogenomics capability, and then to survey available resources and their suitability for this purpose. User-assigned designations such as "TOXM" in ArrayExpress, to label an explicitly purposed toxicogenomics experiment, are useful but a much larger portion of the experimental inventory of ArrayExpress pertaining to chemical treatment has potential utility for toxicogenomics.

In summary, the ability to effectively "read-across" varied genomics data resources, assessing chemical coverage and combining data for similar experimental protocols, is a necessary prerequisite to determining the sufficiency of public data for toxicogenomics analysis. Equally important is the ability to flexibly search and mine data across genomics, toxicology and chemical domains to bring all relevant information to bear on the problem of toxicity prediction. The present survey highlights deficiencies in microarray experiment data deposition requirements and standards with regard to chemistry and chemical-treatment related experiments that, if better addressed, could greatly facilitate chemical annotation and data integration efforts in the future. I proposed a set of Standard Genomics Fields that could be applied to GEO and ArrayExpress to bridge between the two resources and to facilitate comparisons and incorporation of their content into other resources in a standardized way. In the following chapter of this dissertation, I implement these standards within GEO and

ArrayExpress, describe the challenges therein and the methods employed, and illustrate the ways in which newly indexed chemical-experimental content within the two databases can be assessed, explored, and linked to external chemically-indexed bioassay resources.

REFERENCES

Ball C, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Icahn C, Parkinson H, Quackenbush J, Ringwald M, Sansone SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N; Microarray Gene Expression Data (MGED) Society. (2004) Standards for microarray data: An open letter. *Environ. Health Perspect.* **112,** A666-A667.

Barrett T, Edgar R. 2006. Gene expression omnibus: Microarray data storage, submission, retrieval, and analysis. Methods Enzymol 411:352-369; doi: 10.1016/S0076-6879(06)11019-8.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. 2007. NCBI GEO: Mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35(Database issue):D760-5; doi: 10.1093/nar/gkl887.

Bhogal N, Grindon C, Combes R, Balls M. 2005. Toxicity testing: Creating a revolution based on new technologies. Trends Biotechnol 23(6):299-307; doi: 10.1016/j.tibtech.2005.04.006.

Boyle J. 2005. Gene-Expression Omnibus integration and clustering tools in SeqExpress. Bioinformatics 21(10):2550-2551; doi: 10.1093/bioinformatics/bti355.

Brazma A, Kapushesky M, Parkinson H, Sarkans U, Shojatalab M. 2006. Data storage and analysis in ArrayExpress. Methods Enzymol 411:370-386; doi: 10.1016/S0076-6879(06)11020-4.

Burgoon LD. 2007. Clearing the standards landscape: The semantics of terminology and their impact on toxicogenomics. Toxicol Sci 99(2):403-412; doi: 10.1093/toxsci/kfm108.

Burgoon LD, Boutros PC, Dere E, Zacharewski TR. 2006. dbZach: A MIAME-compliant toxicogenomic supportive relational database. Toxicol Sci 90(2):558-568; doi: 10.1093/toxsci/kfj097.

Chen J, Zhao P, Massaro D, Clerch LB, Almon RR, DuBois DC, Jusko WJ, Hoffman EP. 2004. The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface. Nucleic Acids Res 32(Database issue):D578-81; doi: 10.1093/nar/gkh003.

Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ. 2008. Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res; doi: 10.1093/nar/gkn580.

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci 95(1):5-12; doi: 10.1093/toxsci/kfl103.

Fielden MR, Kolaja KL. 2006. The state-of-the-art in predictive toxicogenomics. Curr Opin Drug Discov Devel 9(1):84-91.

Fostel J, Choi D, Zwickl C, Morrison N, Rashid A, Hasan A et al. 2005. Chemical Effects in Biological Systems--data dictionary (CEBS-DD): A compendium of terms for the capture and integration of biological study design description, conventional phenotypes, and 'omics data. Toxicol Sci 88(2):585-601; doi: 10.1093/toxsci/kfi315.

Fostel JM, Burgoon L, Zwickl C, Lord P, Corton JC, Bushel PR, Cunningham M, Fan L, Edwards SW, Hester S, Stevens J, Tong W, Waters M, Yang C, Tennant R. 2007. Toward a checklist for exchange and interpretation of data from a toxicology study. Toxicol Sci 99(1):26-34; doi: 10.1093/toxsci/kfm090.

Fostel JM. 2008. Towards standards for data exchange and integration and their impact on a public database such as CEBS (Chemical Effects in Biological Systems). Toxicol Appl Pharmacol; doi: 10.1016/j.taap.2008.06.015.

Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, Brady L, Browne LJ, Calvin JT, Day GJ, Breckenridge N, Dunlea S, Eynon BP, Furness LM, Ferng J, Fielden MR, Fujimoto SY, Gong L, Hu C, Idury R, Judo MS, Kolaja KL, Lee MD, McSorley C, Minor JM, Nair RV, Natsoulis G, Nguyen P, Nicholson SM, Pham H, Roter AH, Sun D, Tan S, Thode S, Tolley AM, Vladimirova A, Yang J, Zhou Z, Jarnagin K. 2005. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. J Biotechnol:219-244.

Goetz AK, Bao W, Ren H, Schmid JE, Tully DB, Wood C, Rockett JC, Narotsky MG, Sun G, Lambert GR, Thai SF, Wolf DC, Nesnow S, Dix DJ. 2006. Gene expression profiling in the liver of CD-1 mice to characterize the hepatotoxicity of triazole fungicides. Toxicol Appl Pharmacol 215(3):274-284; doi: 10.1016/j.taap.2006.02.016.

Gomase VS, Tagore S, Kale KV. 2008. Microarray: An approach for current drug targets. Curr Drug Metab 9(3):221-231.

Hamadeh HK, Amin RP, Paules RS, Afshari CA. 2002. An overview of toxicogenomics. Curr Issues Mol Biol 4(2):45-56.

Hayes KR, Vollrath AL, Zastrow GM, McMillan BJ, Craven M, Jovanovich S, Rank DR, Penn S, Walisser JA, Reddy JK, Thomas RS, Bradfield CA. 2005. EDGE: A centralized resource for the comparison, analysis, and distribution of toxicogenomic information. Mol Pharmacol 67(4):1360-1368; doi: 10.1124/mol.104.009175.

Hirabayashi Y, Inoue T. 2002. Toxicogenomics--a new paradigm of toxicology and birth of reverse toxicology. Kokuritsu Iyakuhin Shokuhin Eisei Kenkyusho Hokoku (120)(120):39-52.

Ivliev AE, 't Hoen PA, Villerius MP, den Dunnen JT, Brandt BW. 2008. Microarray retriever: A web-based tool for searching and large scale retrieval of public microarray data. Nucleic Acids Res; doi: 10.1093/nar/gkn213.

Larsson O, Sandberg R. 2006. Lack of correct data format and comparability limits future integrative microarray research. Nat Biotechnol 24(11):1322-1323; doi: 10.1038/nbt1106-1322.

Martin MT, Judson RS, Reif DM, Dix DJ. 2008. Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. Environ Health Perspect; doi: 10.1289/ehp.0800074.

Mattes WB, Pettit SD, Sansone SA, Bushel PR, Waters MD. 2004. Database development in toxicogenomics: Issues and efforts. Environ Health Perspect 112:495-505.

Mattingly CJ. 2008. Chemical Databases for Environmental Health. Toxicol Lett; doi:10.1016/j.toxlet.2008.10.003

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35(Database issue):D747-50; doi: 10.1093/nar/gkl995.

Richard A, Yang C, Judson R. 2008. Toxicity data informatics: Supporting a new paradigm for toxicity prediction. Tox Mech Meth 18:103-118.

Richard AM, Gold LS, Nicklaus MC. 2006. Chemical structure indexing of toxicity data on the internet: Moving toward a flat world. Curr Opin Drug Discov Devel 9(3):314-325.

Rustici G, Kapushesky M, Kolesnikov N, Parkinson H, Sarkans U, Brazma A. 2008. Data storage and analysis in ArrayExpress and expression profiler. Curr Protoc Bioinformatics Chapter 7:Unit 7.13; doi: 10.1002/0471250953.bi0713s23.

Salter AH. 2005. Large-scale databases in toxicogenomics. Pharmacogenomics 6(7):749-754; doi: 10.2217/14622416.6.7.749.

Tateno Y, Ikeo K. 2004. International public gene expression database (CIBEX) and data submission. Tanpakushitsu Kakusan Koso 49(17 Suppl):2678-2683.

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK Jr, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Novère NL, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ Jr, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnol 26(8):889-896.

Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A et al. 2003. Systems

toxicology and the chemical effects in biological systems (CEBS) knowledge base. EHP

Toxicogenomics 111(1T):15-28.

Waters M, Stasiewicz S, Merrick BA, Tomer K, Bushel P, Paules R, Stegman N, Nehls G, Yost KJ, Johnson CH, Gustafson SF, Xirasagar S, Xiao N, Huang CC, Boyer P, Chan DD, Pan Q, Gong H, Taylor J, Choi D, Rashid A, Ahmed A, Howle R, Selkirk J, Tennant R, Fostel J. 2008. CEBS--chemical effects in biological systems: A public data repository integrating study design and toxicity data with microarray and proteomics data. Nucleic Acids Res 36(Database issue):D892-900; doi: 10.1093/nar/gkm755.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36(Database issue):D13-21; doi: 10.1093/nar/gkm1000.

Yang C, Benz RD, Cheeseman MA. 2006a. Landscape of current toxicity databases and database standards. Curr Opin Drug Discov Devel 9(1):124-133.

Yang C, Richard AM, Cross KP. 2006b. The art of data mining the minefields of toxicity databases to link chemistry to biology. Curr Comput-Aided Drug Design 2(2):135-150.

Zhu Y, Zhu Y, Xu W. 2008. EzArray: A web-based highly automated affymetrix expression array data management and analysis system. BMC Bioinformatics 9:46; doi: 10.1186/1471-2105-9-4

TABLES

Table 2-1. Hyperlinked List and Definitions of DSSTox Standard Chemical Fields as found in DSSTox. Chemical Standard fields are used as a standardized vocabulary for the description of chemicals and chemical structures.

Field Name	Allowed entries	Field Definition
STRUCTURE	*Molecule represented as molfile*	2D (or 3D) "mol" file coordinates for defined molecular structure. **STRUCTURE_Shown** field relates content of **STRUCTURE** field to actual tested substance and **TestSubstance_...** fields. **STRUCTURE** field directly corresponds to, and is used to generate the content of the remaining **STRUCTURE_...** fields. **STRUCTURE** field entry is *blank* only when no reasonable or representative 2D structure can be provided, as in some cases when **TestSubstance_Description** entry is "mixtureor formulation" or "unspecified or multiple forms".
DSSTox_RID	*# (integer)*	DSSTox Record ID (RID) is number uniquely assigned to each DSSTox record across all DSSTox files, regardless of Test Substance characteristics or **STRUCTURE** field content, i.e. no two DSSTox records share a **DSSTox_RID**. It is used to centrally manage DSSTox data file information and to register DSSTox data file records in PubChem.
DSSTox_CID	*# (integer)*	DSSTox Chemical ID number uniquely assigned to a particular **STRUCTURE** and "**STRUCTURE**-content" fields across all DSSTox databases (see More on DSSTox Standard Chemical Fields). Different CID numbers will be assigned if two **STRUCTURE** records are substantively different, e.g., different chemical, salt or complex form, or stereochemical isomer. DSSTox records with the same CID number will share the contents of all DSSTox **STRUCTURE**-content fields, except for **STRUCTURE_Shown**, which depends on the relationship to the **TestSubstance_Description**.

Table 2-1. (Continued).

DSSTox_ Generic_SID	*# (integer)*	Records with the same **DSSTox_Generic_SID** (Generic Substance ID) will share all DSSTox Standard Chemical Fields, including **STRUCTURE**. Field distinguishes at the level of "Test Substance" across all DSSTox data files, most often corresponding to the level of CASRN distinction, but not always. Different **DSSTox_Generic_SID** numbers will be assigned to the same **STRUCTURE** record if, e.g., the **TestSubstance_Description** differs in the data record, i.e. one is "single chemical compound", the other is "mixture or formulation", or in cases where explicit information on Test Substance grade or purity is available (e.g., technical grade, etc). **DSSTox_Generic_SID** does not, however, distinguish DSSTox test substance records that differ in experimental settings only by lot/batch/plate location, etc.
DSSTox_FileID	*# (integer) Text*	Sequential ID number is assigned to each record in data file, with values ranging from 1 to n, where n=total # of records in the data file. ID number is followed by an underscore and then the abbreviated DSSTox SDF standard file name with version, e.g., 1_CPDBAS_v4a. Field entry provides a unique record identifier for every DSSTox data record and is updated whenever a new version or revision of DSSTox SDF data file is generated.
STRUCTURE _ Formula	*Text*	Empirical formula of displayed **STRUCTURE**. *blank* if **STRUCTURE_Shown** entry is "no structure"
STRUCTURE _ MolecularWeight	#	Molecular weight or molar mass (atomic mass units) of displayed **STRUCTURE** . *blank* if **STRUCTURE_Shown** entry is "no structure"

Table 2-1. (Continued).

STRUCTURE_ChemicalType	defined organic/ inorganic/ organometallic/ no structure/	Nature of chemical displayed in **STRUCTURE** field: “defined organic” = defined chemical structure containing carbon but not organometallic, i.e. containing no metal or metalloid atom other than simple salt alkali (I) or alkaline earth (II) metals (Na, K, Mg, Ca, etc.); “inorganic” = defined chemical structure containing no carbon; “organometallic” = operationally defined as a chemical structure containing carbon and any metal or metalloid atom other than alkali (I) or alkaline earth (II) metals that occur in simple salts; “no structure” indicates **STRUCTURE** field is *blank*; only used when **TestSubstance_Description** = “undefined mixture” , “unspecified or multiple forms” , or "macromolecule".
STRUCTURE_TestedForm_DefinedOrganic	parent/ salt, complex/ , Na, K, HCl, Cl, H2O, Ca, H2SO4, acetate, bis, etc.	Tested form of chemical displayed in **STRUCTURE** field only for **STRUCTURE_ChemicalType** = “defined organic”. Operational definitions of allowable entries as follows: “parent” = single defined organic chemical entity, without counter ions or complexed chemical entities; “salt” = simple ionic salts of defined organics with alkali (I) or alkaline earth (II) metal (e.g., Na, K, Mg, Ca) or halide (e.g., Cl) counter ions; “complex” = any defined organic with associated acid, base, or hydrate. *blank* if **STRUCTURE_Shown** entry is "no structure" or if **STRUCTURE_ChemicalType** entry is other than “defined organic”. Following the field entry “salt” or “complex”, the counter ion or complexed chemical moiety is listed in abbreviated form, e.g.: Na, K, HCl, Cl, H2O, Ca, H2SO4, acetate, etc.;“bis” signifies parent structure occurs twice in complex, etc.

Table 2-1. (Continued**).**

STRUCTURE_ Shown	tested chemical/ active ingredient in formulation/ representative isomer in mixture/ representative component in mixture/ monomer of polymer/ general form of chemical/ no structure/ , simplified to parent	Identifies relationship of the graphical structure displayed in the **STRUCTURE** field to the actual tested chemical substance : “tested chemical” - structure displayed is the actual form of the chemical tested; “active ingredient in formulation” - the tested form of the chemical substance was a mixture or formulation and only the active ingredient is displayed in the **STRUCTURE** field; “representative isomer in mixture” - the structure shown is one isomer in a test substance consisting of a mixture of isomers (e.g., cis, trans, Z, E); “representative component in mixture” - the structure shown is a major component in a test substance consisting of a mixture of distinct chemical substances; “monomer of polymer” - the structure shown is a small repeating subunit of a polymer or macromolecule; “general form of chemical” - chemical record contains toxicity data fields summarized from multiple experiments, where either multiple tested forms (e.g., salts or complexes) of the chemical were evaluated, or where the tested form of the chemical is not specified or ambiguous; “no structure” - when no reasonable or representative structure can be provided, as when **TestSubstance_Description** entry is “mixture or formulation” or “unspecified or multiple forms”. “, simplified to parent” - for desalted files, only occurs as a comma-separated modifier to another field entry; used when **STRUCTURE_ChemicalType**=”defined organic” and **STRUCTURE_TestedForm_DefinedOrganic** =”salt” or “complex”; and signifies that **STRUCTURE** is being represented in its desalted, neutral or protonated forms, without counter ions or complexed chemical entities. An exception is quaternary ammonium ions, which are represented in positively charged state with salt counter ion removed.

Table 2-1. (Continued).

TestSubstance_ChemicalName	*Text*	Common or trade name of chemical listed in original Source database. Field entry corresponds to **TestSubstance_CASRN**. If **STRUCTURE_Shown** = "tested chemical", field entry corresponds directly to contents of **STRUCTURE** field.
TestSubstance_CASRN	######-##-#/ NOCAS/	Chemical Abstracts Service (CAS) Registry Number of the tested substance, formatted **000000-00-0**. In general, corresponds to **TestSubstance_ChemicalName**. If **STRUCTURE_Shown** = "tested chemical", field entry corresponds directly to **STRUCTURE**. "NOCAS" indicates CAS registry number was unavailable from original Source data table or was not found.
TestSubstance_Description	single chemical compound/ macromolecule/ mixture or formulation/ unspecified or multiple forms/	"single chemical compound" = pure, neat or approximately pure single chemical compound (could be parent, salt or complex) with defined molecular structure; "macromolecule" = polymer, protein, DNA, or other large biomolecular species; "mixture or formulation" = test substance consists of more than one chemical compound, which may be fully or partially characterized, or consists of an active ingredient in an unspecified formulation,or the individual chemical components are not known; "unspecified or multiple forms" = either the exact nature of the test substance is unknown or the test results refer to more than a single form of the test substance (e.g., multiple salt forms or derivatives of a parent chemical).

Table 2-1.(Continued).

ChemicalNote	*Text,* ammonium, stereochem, tautomers, parent [*CASRN*], CAS replicate, replicate 2D, replicate parent, etc	Note provides additional information related to tested substance, e.g., when uncertainty exists in chemical name or CAS number, parent structure is "ammonium" ion, tautomeric forms are known to exist, mixture characteristics are known, "stereochem" information is known (e.g., *cis, trans, Z, E*, R, S), CAS of parent salt or complex is known, common chemical name synonym, etc.
STRUCTURE_ ChemicalName_ IUPAC	*Text*	IUPAC (International Union of Pure and Applied Chemistry) refers to standardized nomenclature of organic chemistry. IUPAC chemical names are generated automatically from **STRUCTURE** using the ACD/Name generation software (ACD Labs, see LogFile for version) or obtained as a systematic name from other chemical sources (see DSSTox Chemical Information Quality Review Procedures). *blank* if **STRUCTURE_Shown** entry is "no structure"
STRUCTURE_ SMILES	*Text*	SMILES (**S** implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) molecular text code of displayed **STRUCTURE** . *blank* if **STRUCTURE_Shown** entry is "no structure"

Table 2-1. (Continued).

STRUCTURE_ Parent_SMILES	*Text*	SMILES (**S** implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) molecular text code of displayed **STRUCTURE** unless **STRUCTURE_TestedForm_DefinedOrganic** entry is either "salt" or "complex", in which case field entry corresponds to parent structure in desalted or neutralized (protonated) form, without salt counter ions or complexed moieties. **STRUCTURE_Parent_SMILES** only provided for **STRUCTURE_ChemicalType** = "defined organic". *blank* if **STRUCTURE_Shown** entry is "no structure"
STRUCTURE_ InChI	*Text*	InChI = IUPAC (International Union of Pure and Applied Chemistry) NIST (National Institutes of Standards and Technology) Chemical Identifier, a unique, standardized, text-based code for molecular structure. InChI codes were generated automatically from the **STRUCTURE** using the publicly available NIST/IUPAC InChI code generator program (see Log File of DSSTox database for code version). InChI codes encapsulate essential chemical structural information and can be used for text, web-based, chemical structure searching. If **STRUCTURE_Shown** entry is "no structure", InChI default entry is: InChi=1//
STRUCTURE_ InChIKey	*Text*	"InChIKey" is a fixed-length (25-character) condensed digital representation of the InChI Identifier that can be used to facilitate structure look-up; the full InChI is required for structure-regeneration. For more information, see **STRUCTURE_InChI** and More on InChI. If **STRUCTURE_Shown** entry is "no structure", InChIKey default entry is: MOSFIJXAXDLOML-UHFFFAOYAM
Substance_modify_yyyymmdd	*Numeric*	Sortable numeric date assigned to every unique substance in the DSSTox inventory (i.e., every unique DSSTox_Generic_SID) indicating the most recent date of modification of the structure or Standard Chemical Fields. (Note the date does not apply to changes in the Source-specific toxicity data fields.) yyyymmdd = year, month, day (e.g., 20081021 = 21 October 2008)

Table 2-2. Primary Genomic Expression Resources. Primary Gene Expression Resources are defined as those identified by the Microarray Gene Expression (MGED) Data Society as public repositories for public gene expression data. Primary gene expression resources house gene expression data referenced in scientific journals.

Database	Source/URL	Summary	Local Installation/Data Download	Public Data Deposition	Programmatic Access	Public Queries
ArrayExpress (AE)	European Bioinformatics Institute (EBI); www.ebi.uk.ac/ArrayExpress	A public repository for high-throughput functional genomics data that serves as a MIAME supportive public archive of microarray data[Brazma et al. 2006a, 2006b Parkinson et al. 2007]	YES/YES	YES	YES (XML)	YES
Gene Expression Omnibus (GEO)	National Center for Biotechnology Information (NCBI), National Insitutes of Health; www.ncbi.nlm.nih.gov/geo	A MIAME-complaint public repository for freely disseminating fully annotated raw and processed microarray data and other forms of high-throughput data generated by the scientific community [Barrett et al. 2007; Wheeler at al. 2008].	NO/YES	YES	YES (E-Utilities)	YES
Center for Information Biology Gene Expression Database (CIBEX)	DNA Data Bank of Japan (DDBJ), National Institute of Genetics; http://cibex.nig.ac.jp/	A MIAME-complaint public repository for high-throughput experimental data in gene expression research including microarray-based experiments measuring mRNA, serial analysis of gene expression (SAGE tags), and mass spectrometry proteomic data from Asian countries [Tateno and Ikeo 2004].	NO/YES	YES	NO	YES

Table 2-3. Secondary Gene Expression Resources. Secondary gene expression resources contain gene expression data similar to primary gene expression resources. However, secondary gene expression resources have limited scope or do not allow public data deposition.

Database	Source/URL	Summary	Local Installation/ Data Download	Public Data Deposition	Public Queries
Environment, Drug, Gene Expression Database (EDGE)	McArdle Laboratory for Cancer Research, University of Wisconsin-Madison; http://edge.oncology.wisc.edu	A centralized resource for the Comparison, Analysis, and Distribution of Toxicogenomic Information used to house the results of a standardized set of microarray reagents and reproducible protocols to simplify the analysis of liver gene expression in the mouse model [Hayes et al. 2005].	NO / NO	NO	YES (pre-formed queries)
Comparative Toxicogenomics Database (CTD)	Mount Desert Island Biological Laboratory; http://ctd.mdibl.org	A database of scientifically reviewed and curated information on chemicals relevant genes and proteins, and their interactions in vertebrates and invertebrates integrating sequence, reference, species, microarray, and general toxicology information to provide a unique, centralized resource for toxicogenomic research[Davis et al. 2008; Mattingly et al. 2006a, 2006b].	NO / NO*(does not contain primary data)	NO	YES
Chemical Effects in Biological Systems (CEBS)	National Institute of Environmental Health Sciences; http://www.niehs.nih.gov/research/resources/databases/cebs/index.cfm	An integrated public repository for toxicogenomic data, including the study design and timeline, clinical chemistry and histopathology finds and microarray and proteomics data [Waters et al. 2008].	* in progress/YES	YES	YES
Public Expression Profiling Resources (PEPR)	Center for Genetic Medicine Research; http://pepr.cnmcresearch.org/	A public data warehouse of human, rat, and mouse Affymetrix GeneChip expression profiles, generated in the same laboratory and subject to the same quality and procedural controls [Chen et al. 2004].	NO / YES	NO	YES

Table 2-4. Chemoinformatics Resources for Toxicogenomics. Chemoinformtics resources are different than gene expression resources in that they do not actually contain gene expression data. Chemoinformatics resources are chemically indexed resources that help to enable linkages between public chemical and biological resources.

Database	Source/URL	Summary	Local Installation/ Data Download	Supported Data Types	Included Query Capability	Included Analysis Tools
ArrayTrack	US Food and Drug Administration (FDA); http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/arraytrack_webaccess.htm	An integrated suite designed for management, analysis and interpretation of microarray data including 3 integrated components: a MIAME-supportive database that stores and annotates essential experiment information, analysis tools providing the ability to search, filter, apply statistical operations and graphically visualize data, and several libraries that provide gene annotation, protein function and pathway information directly hyperlinked within the data analysis process [Tong et al, 2003; Tong et al. 2004].	YES / NO	Gene Expression, Toxicological Parameters, Protein Function, Pathway Information	NO	YES
CHemical Entities of Biological Importance (ChEBI)	European Bioinformatics Institute (EBI); http://www.ebi.ac.uk/chebi	A dictionary of molecular entities focused on "small" chemical compounds, natural or synthetic products used to intervene in the processed of living organisms; macromolecules are excluded as a rule [Degtyarenko et al. 2008].	NO / NO	Chemical structure and properties	YES	NO
ChemBank	Chemical Biology Program and Platform at the Broad Institute of Harvard and MIT; http://chembank.broad.harvard.edu/	ChemBank is a public, web-based system designed to provide a chemoinformatics knowledge environment that includes freely available data derived from small molecules and small-molecule screens and resources for studying these data. ChemBank provides storage of raw screening data, rigorous definition of screening experiments in terms of statistical hypothesis testing, and metadata-based organization of screening experiments into projects involving collections of related assays [Seiler et al. 2008; Wagner et al. 2008].	NO / YES	Small molecule raw screening data	YES	YES

Table 2-4. (Continued)

ChemDB	Institute for Genomics and Bioinformatics and Department of Computer Science, University of California at Irvine (UCI); http://cdb.ics.uci.edu	A public database of small molecules and related chemoinformatics resources built using digital catalogs of venders and public resources [Chen et al. 2007].	NO / YES	Information from data resources, calculated properties	YES	YES
ChemSpider	ChemZoo: http://chemspider.com	A chemical search engine providing access to millions of chemical structures and properties, and integration to a large number of online services, including PubChem [Williams 2008].	NO / NO	Chemical structure, properties, links to other resources	YES	NO
DbZach	Department of Biochemistry & Molecular Biology, Michigan State University http://dbzach.fst.msu.edu	A modular MIAME-Complaint toxicogenomic Supportive Relational Database with associated data insertion, retrieval, and mining tools that manages traditional toxicology and complementary toxicogenomic data to facilitate comprehensive data integration, analysis, and sharing.	YES/NO	Chemical Structure, Omics Data, PCR, Toxiciology	YES	YES
DrugBank	Departments of Computing Science and Biological Sciences, University of Alberta; http://redpoll.pharmacy.ualberta.ca/drugbank/	A database resource that combines detailed drug (i.e., chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e., sequence, structure, and pathway) information [Wishart et al. 2006; Wishart et al. 2008].	NO / YES	Chemical, Drug Target	YES	NO
Distributed Structure-Searchable Toxicity (DSSTox) Data Network	US Environmental Protection Agency (EPA); http://www.epa.gov/ncct/dsstox/	A public resource offering downloadable, structure-searchable, standardized chemical structure files associated with toxicity data, and spanning approx 120K substances, offering structure-searchability and links to other public cheminformatic resources, such as PubChem and ChemSpider [Richard et al. 2006; Richard et al. 2008].	NO / YES	Chemical structure and summary toxicity information	YES	NO
PubChem	National Center for Biotechnology Information (NCBI); http://pubchem.ncbi.nlm.nih.gov/	A public repository of information on small molecules with three components: PubChem Compound, Substance, and Bioassay, spanning more than 10 million substances and 500 bioassays; serves as primary data repository for the NIH Molecular Libraries Roadmap project	NO / YES	Chemical Structure, BioAssay Data, Links to other data	YES	NO

Table 2-5. Data Description for Primary and Secondary Gene Expression Resources. Gene Expression Resources vary greatly in diversity. This table describes the current characteristics of primary and secondary gene expression resources for the purpose of understanding the current state of primary and secondary resources. CEBS by far shows the greatest amount of diversity and capabilities.

Database	Type(s) of Gene Expression Data Stored				Other Data Types		Web Interface				
	Spotted Arrays	**Oligonucleotide Arrays**	**PCR**	**SAGE**	**Other 'Omic Data**	**Toxicological Focus/ Toxicological Data**	**Array Images**	**Analytical Results**	**Data Input/Data Query**	**Data Export / Import**	**Public Programmatic Access**
ArrayExpress	Y	Y	N	N	N	N	Y	Y (ArrayExpress Data Warehouse)	Y / Y	Y / Y	Y
GEO	Y	Y	N	Y	N	N	Y	Y (GEO Profiles)	Y / Y	Y / Y	Y
CIBEX	Y	Y	N	N	N	N	Y	N	N / Y	Y / Y	N
dbZach	Y	Y	Y	N	Y	Y	Y	N	N / Y	Y / N	N
EDGE	N	Y	N	N	N	Y	N	Y	N / Y	N / N	N
CTD	N	N (contains linkages to microarray data where available)	N	N	N	N (contains linkages to tox data where available)	N	N	N / Y	Y / N	N
CEBS	**Y**	**Y**	**Y**	**N**	**Y**	**Y**	**Y**	**Y**	**N / Y**	**Y / Y**	**N**
PEPR	N	Y	N	N	N	N	N	Y	Y* in laboratory only / Y	Y / N	N

Table 2-6. Standardization and Indexing of Gene Expression Resources. A presentation of the current state of the standardization and chemical indexing is shown in order to identify deficiencies. Secondary gene expression resources may have limited applicability and content but have had more success in standardization and indexing.

Database	Indexed by Chemical	MIAME Supportive	Standardized				Relational Searching
			Species	Array Information	Experimental Protocol	Experimental Details	
ArrayExpress	N	Y	N	Y	N	N	N
GEO	N	Y	Y	Y	N	N	N
CIBEX	N	Y	Y	Y	Y	Y	N
EDGE	**Y**	**N**	**Y**	**Y**	**Y**	**Y**	**Y**
CTD	Y	N	Y	N/A	N/A	N	Y
CEBS	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
PEPR	N	Y	Y	Y	Y	Y	N

Table 2-7. Standard Genomics Fields incorporated into DSSTox Auxiliary files (indexed by Experiment) for ArrayExpress and GEO. Standard genomic fields are used to increase the interoperability of primary and secondary gene expression resources. Mapping from ArrayExpress and GEO are shown to explain where the information originates.

Field Name	**Subsections**(allowed entries) [1]	**Field Type**[2]	**Description**	**Mapping from ArrayExpress**[3]	**Mapping from GEO**[4]
Experiment_Accession		alphanumeric	A unique combination of an informative prefix and number used to identify each dataset.	"experiment accnum"	"GSE" with prefix *added during Annotation Process*[5]
Experiment_Alternative Accession		alphanumeric	An alternate accession number used to access data in a different way. Example: GEO files that have been imported into ArrayExpress have a GSE#### (GEO Series) secondary Accession number allowing users to find the same data in GEO)	"secondary accession"	"GDS"
Experiment_IdNumber		numeric	A unique identification number for each experiment within each database. This number, if queried, takes the user directly to data information page where more information is provided not about the experiment but the data available for the experiment.	"id"	GSE "Id"
Experiment_Title		text	The title of the experiment.	"name"	"title"
Experiment_Description		text	A free-text, user-submitted description of the experiment or dataset.	"description"	"summary"
Experiment_URL		hyperlink	URL that links directly to the Source Experimental Download Page.	*Added during Annotation Process*	*Added during Annotation Process*
Experiment_PubMed_Information		numeric	A unique number that links users to each publication associated with each experiment or dataset. This number can be queried within PubMed to yield each article	"bibliography"	"PubMedIds"
Experiment_Publication Date		numeric	Date indicating when the dataset was released to the public or published.	"releasedate"	"PDAT"
Species		text	Species as listed by the user.	"species"	"taxon"

Table 2-7. (Continued).

Number_Samples		numeric	Number of samples used within a microarray experiment or dataset.	"samples"	"n_samples"
Experiment_Array Accession		numeric	An accession number for each array design or platform.	arraydesign count "identifier"	"GPL" with prefix *added during Annotation Process*
Experiment_ArrayType		text	Details about the platform used or details about data other than raw data that users have submitted.	"array"	gpl "title"
Experiment_ArrayTitle		text	The user submitted title of the Array/Platform used in the experiment.	arraydesign count "name"	GPL "summary"
Chemical_StudyType		text	A DSSTox-assigned designation of the role of the identified chemical in the given experiment. Allowed entries are listed as Subsections to this field (e.g., Reference, Treatment, Vehicle, …) Note: Terms may occur in combination (denoted with "AND) since a chemical may be used in more than one role for an experiment. For example, trans-hydroxytamoxifen and Raloxifene were tested separately and together in the study of breast cancer. Therefore trans-hydroxytamoxifen would be listed as Combination_TreatmentANDTreatment.	*Added during Annotation Process*	*Added during Annotation Process*
	Reference	text	A chemical used to mimic a biological or environmental situation, where the purpose of the experiment is to learn more about the effects of the biological or environmental situation. Example: the use of streptozocin to induce Diabetes Mellitus in a study investigating the genomic effects of diabetes.	*Added during Annotation Process*	*Added during Annotation Process*
	Treatment	text	A chemical is the primary focus of the experiment or study, where the purpose of the experiment is to understand the transcriptomic effects of the chemical.	*Added during Annotation Process*	*Added during Annotation Process*
	Vehicle	text	A chemical used to aid the administration of the treatment to the organism, such as Dimethyl Sulfoxide (DMSO).	*Added during Annotation Process*	*Added during Annotation Process*

Table 2-7. (Continued).

	Combination_Treatment	text	Multiple chemicals used together for treatment purposes. See description of treatment above.	*Added during Annotation Process*	*Added during Annotation Process*
	Media	text	A chemical used in the maintenance of the tissue culture or sample conditions, such as Phosphate Buffered Saline (PBS).	*Added during Annotation Process*	*Added during Annotation Process*
	Not_Enough_ Information	text	Sufficient information is not present in the experimental description to determine the role of the chemical.	*Added during Annotation Process*	*Added during Annotation Process*

[1] Subsections to the Chemical_StudyType field have allowed entries: Reference, Treatment, etc, with linkage text "AND" used for combinations (e.g., TreatmentANDReference).
[2] Field Type indicates the type of field entry, restricted for processing purposes.
[3] Indicates the corresponding data field name or entry in ArrayExpress via programmatic access. Text provided in quotations corresponds directly to the text of the resulting programmatic access document. Terms not in quotations indicate the hierarchical structure of the programmatic access document. The quoted portion of these terms indicates the level from which the information was extracted. (See EBI 2008b). Example: arraydesign count "identifier"-indicates that the identifier portion of the array design count section of the programmatic access document was used.
[4] Indicates the corresponding data field name or entry in GEO. Text provided in quotations corresponds directly to the text of the resulting programmatic access document. Terms not in quotations indicate the database entry method for programmatic access. (See NCBI GEO 2008b). Example: GPL "id"- indicates that data was accessed using the GEO Platform (GPL) database entry method and resulting id field was used.
[5] Fields were added during the current annotation process described in METHODS. Currently these fields and field entries cannot be extracted directly through programmatic access in the format provided.

CHAPTER 3

CHEMICAL INDEXING OF EXPERIMENTS IN GEO AND ARRAYEXPRESS

ABSTRACT

A publicly available toxico-chemogenomics capability requires common data genomics standards and protocols, broad data coverage within chemical, genomics and toxicological information domains, and transparent and functional linkages of public Internet data resources. Focusing on the two largest resources – Gene Expression Omnibus (GEO) and ArrayExpress -- the goal was to chemically index the experimental content of these repositories to identify all chemical exposure-related microarray experiments of potential toxicogenomics value, and to make these data accessible in relation to other publicly available chemically-indexed toxicological information. Current standards for chemical annotation within ArrayExpress and GEO were found to be inadequate to this task. A series of automated Perl scripts and extensive manual review were employed to transform raw experiment descriptions and text files into a standardized chemically-indexed inventory of experiments in both resources. These files and top-level summary annotations allowed for identification of all current chemical-associated experimental content as well as the chemical exposure-related (or "Treatment") content of greatest potential value to toxicogenomics investigation. With these chemical index files, it is possible for the first time to assess the breadth and overlap of chemical study space represented in these databases, and to begin to assess the sufficiency of data for chemical similarity inferences. Chemical indexing of public genomics databases is also the first step towards integrating chemical, toxicological and genomics data into predictive toxicology by providing linkages across public resources.

INTRODUCTION

Toxicogenomics in support of predictive toxicology requires the ability to anchor gene expression results to traditional toxicological bioassays or effects measures for a series of chemicals, derive useful inferences from those linkages, and apply the resulting gene expression patterns, perhaps along

with other types of data, to estimating the potential toxicity of chemicals for which bioassay results are unavailable. Such a capability depends upon the availability of gene expression data derived from chemical treatment scenarios, as well as broad data coverage across chemical, genomics and toxicological information domains to support predictive inferences. Hence, centrally important to this task is the ability to easily locate and aggregate all relevant data pertaining to the chemical or chemicals of interest. Given the chemical-centered nature of the problem, the solution requires standardized, chemical-centric accessibility to data at all levels.

In Chapter 2, I surveyed current public microarray resources from the standpoint of interoperability and formal chemical indexing for use in toxicogenomics and predictive toxicology. By far, the two largest and most important public repositories of microarray data are the U.S. National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) (Brazma et al. 2006a, 2006b; Parkinson et al. 2007; Rustici et al. 2008) and the European Bioinformatics Institute's (EBI) ArrayExpress (Barrett et al. 2007; Wheeler et al. 2008; NCBI GEO 2008a). Both GEO and ArrayExpress are officially approved by the Microarray Gene Expression Data (MGED) Society (MGED 2008a), support MIAME (Minimum Information About a Microarray Experiment) standards for data reporting (MGED 2008b), and serve as primary repositories of microarray data associated with genomics publications in the scientific literature. I concluded that these resources; however, do not contain sufficient interoperability, in terms of standards and linkages or formal chemical indexing, to support a public toxicogenomics capability at present. At the time of this writing, there is no formal chemical annotation of experiments to be able to assess or efficiently locate chemical treatment content within these resources, no way to effectively search the chemical space covered by the experimental content, and no way to compare chemical experimental content between the two

resources. Hence, these resources are effectively isolated from the rapidly growing inventory of chemically indexed bioassay data on the Internet that can contribute to toxicogenomic analysis.

At the time of this writing, GEO contains over 9,000 user-deposited microarray experiments, whereas ArrayExpress contains over 6000 experiments, many of which originated from GEO. Given the wide range of applications of genomics technologies in the published literature, and that both GEO and ArrayExpress are public user-depositor resources, they are expected to contain a wide range and diversity of experimental content with respect to array type, species, protocol, experiment design and objectives, etc. Also, considering the general nature of toxicology in relation to biology, any microarray experiment in which gene expression changes resulting from a chemical treatment are being investigated is of potential value for toxicogenomics investigations, even if the experiment is not labeled as such

Chemical text mining is a technology that makes it possible to discover patterns and trends semi-automatically from large collections of unstructured text based on technologies such as natural language processing, information retrieval, information extraction, data mining, and chemical names (Uramoto et al. 2004). Typical chemical text mining occurs in the biomedical literature where there is some form of chemical names in addition to biological text using methods such as named entity recognition (NER) (Cohen and Hersh 2004). Alternatively, other methods use a standardized ontology such as ChEBI with pattern matching tools(Rebholz-Schuhmann et al. 2008). Commercial applications have been marketed that automate chemical text mining (Accelrys 2008) for locating text resembling chemical names in documents. After chemical names are identified, there are several chemical name-to-structure applications that convert a chemical name to a chemical structure (ACDLabs 2008). Ideally, these technologies could be employed to chemically and structurally index resources such as GEO or ArrayExpress. However, each of these technologies requires a minimal

level of textual consistency and accuracy which are not applicable to ArrayExpress and GEO at present due insufficient annotation of chemical information. Unlike journal articles or other types of biomedical literature, gene expression records deposited into public repositories are not peer reviewed for sufficient information.

Currently, chemical treatment-related experiments are located within GEO and ArrayExpress through their on-line query interfaces by keyword search of non-standardized (user-provided) chemical names (Figure 3-1 and 3-2) (EBI 2008b, NCBI 2008c). This web-accessible search is an impractical means for generating a chemical index for the following reasons: 1) a priori knowledge of the chemical inventory in the resource is needed to locate all experiments; 2) chemical names contain varied spellings (a high percentage of misspellings), punctuation, and/or abbreviations; and 3) chemical names are not unique, i.e., a large number of trade names, abbreviations and synonyms would have to be included in the search to locate all records. Table 3-1 shows a subset of names that would be necessary to correctly identify all estrogen-related experiments contained within ArrayExpress. Neither GEO nor ArrayExpress offers a standardized, searchable data field for chemical name, nor does either resource have minimal requirements for chemical annotation of experiments deposited by submitters. Furthermore, without standardized chemical structure annotation of these microarray resources, there is no ability to identify potentially relevant experiments pertaining to closely related chemical structures.

In Chapter 2, I proposed a set of standard genomics fields for indexing of experimental study records (with regard to species, protocol, raw data availability, etc.) to enable cross referencing and comparison of total experimental content in the two largest public microarray resources, GEO and ArrayExpress (Table 2-7). In addition, I proposed adoption of chemical indexing standards (Table 2-1) to enable assessment of chemical content within microarray repositories, as well as to facilitate

linkages to external, chemically indexed bioassay information (Table 2-4). In this chapter, I present implementation of these standards within GEO and ArrayExpress.

METHODS

For the purpose of assessing and comparing the content of GEO and ArrayExpress, it is important to consider the different means used to index their experimental content. In the ArrayExpress Repository, an experimental Accession Number is used to identify each "experiment". In this resource, an experiment is defined as a set of related studies, and experiments may span multiple laboratories, chemicals, species, and platforms. In GEO, there are also multiple accession-indexing mechanisms, GSE and GDS being of most importance (Barrett et al. 2007; NCBI GEO 2008d). A GSE is a GEO Series Accession identifier that defines a set of related samples considered to be part of a single experiment. A GDS is a GEO Datasets Accession identifier that provides an experiment-centric view of the data in GEO. A GEO Dataset is a curated version of GEO Series where samples are organized and formatted in a manner conducive to and appropriate for further data analysis, i.e. same species, laboratory, and platform. GDS Accession numbers are also linked to data-mining tools. A result of these varied experiment definitions and indexing schemes is that the mapping of chemical to experiment can vary between resources. For the present study, GEO Series provides the most complete inventory of current experiments within GEO and these are also most closely aligned with ArrayExpress Repository experiments. Hence, the ArrayExpress Repository and GEO Series repository were the focus of the present chemical indexing efforts.

For present chemical-indexing purposes, a chemical exposure (or treatment-related) microarray experiment (or study) is broadly defined as a study in which the cells, tissues, or whole organisms were treated with a defined chemical, chemical mixture, or natural substance, DNA was extracted,

and the gene expression changes resulting from this treatment were investigated with microarray technologies. Whether the chemical to which the system is exposed is a known toxicant, potential toxicant, natural substance, or therapeutic agent need not be distinguished since the measured outcome is the same, i.e., treatment-related gene expression changes. However, experiments in which the chemical treatment is secondary to the primary purpose of the experiment (e.g., treatment with prophylactic antibiotics for maintaining tissue culture conditions) or where study of chemical exposure-induced effects was not the primary purpose of the experiment (e.g., treatment with streptozocin to induce Diabetes Mellitus for investigating the effects of diabetes) require further annotation and review. These cases of chemical-experiment associations are labeled to indicate the role of the chemical as other than “treatment". The Standard Genomics Field, Chemical_StudyType, and its allowed entries (treatment, vehicle, media, reference, etc) is employed for this chemical labeling of experiments (Table 2-7).

The primary goal of the present effort was to create an accurate chemical index file for ArrayExpress and GEO Series experimental content, with particular focus on identifying chemical treatment-related content. Given the dynamic, user-depositor nature of these resources, with frequent updates, our chemical index file would, by necessity, reflect a static, dated inventory. The current chemical annotation process took place over the course of a year and, thus, involved several iterations of methods and updates of GEO and ArrayExpress content. With each interation, I attempted to use manually curated, older versions of chemical index files to refine and validate automated processes for producing updated files most efficiently and accurately. A description of the methods used for assessing the total chemical-associated experimental content, and the subset of treatment-related experimental content in the two major public resources, GEO and ArrayExpress, and the procedure for constructing chemical index files follows.

ARRAYEXPRESS REPOSITORY

Due to its large size (over 6800 experiments at the time of this writing), limited and unstructured chemical annotation, and dynamic content (updated regularly with new experiments) (Figure 3-3), the review and annotation process for ArrayExpress involved several iterativesteps for identification and characterization of chemical treatment-related experiments within the main database repository (EBI 2008a).

After communication with the ArrayExpress development team, a bulk download of all the data housed in the repository from the main web site was undertaken with a wildcard query in the accession number query box (i.e., to retrieve all experiments) in Figure 3-1. Initially, each experimental record was individually reviewed and a preliminary index of the following information was constructed: Experimental Accession Number (EAN), Chemical Name, Identification CAS RN and/or ChEBI (EBI Chemical Entities of Biological Interest) Number (if included), Species, and <Indications of a Chemical Exposure Record>. The latter field included any detail deemed as potentially useful for discerning whether a record pertained to a chemical exposure experiment, e.g., designations in the ArrayExpress <Experimental_Type> field, such as, "compound treatment" or "dose". This preliminary chemical index was used to identify true chemical exposure experiments, to infer the minimum information necessary to identify such records from within ArrayExpress, and to build an automated indexing capability. The initial manual curation process required the reading of 711 experimental records (the entire content as of the October 2005) over a period of two weeks. As discussed above, each record had to be meticulously annotated. It was apparent that an automated process was necessary to improve the efficiency of these methods.

To automate the curation process it was necessary to define a set of inclusion criteria from the <Indications of Chemical Exposure> field recorded during the manual curation process. The goal of

the automated method was to 1) decrease the number of records that had to be manually curated 2) to insert markers to identify relevant information within the records to make the manual curation process more efficient 3) have 0 false negatives, chemical exposure experiments not identified and 4) identify newly added chemical exposure experiments. There is a level of expert knowledge that is applied during the manual curation process; therefore the automated process was not intended to completely replace the manual curation process. The efficiency of the automated method was measured through the decrease in the number of non-chemical exposure records (false positives) that were manually curated. The initial manual curation identified 639 non-chemical exposure records of the total 711 experiments.

Due to the fact that there were so few chemical exposure experiments (72) during the first chemical exposure manual annotation, a second complete manual curation of the current public content of ArrayExpress as of April 2006 was completed to build the knowledge base of < Indications of Chemical Exposure Records>. During the second complete manual curation, 136 chemical exposure experiments were identified. Using the inclusion criteria 299 candidate chemical exposure records were identified from a total of 1363 experiments. Each record was assigned a 0 (false positive) or a 1 (true positive). Each of 136 chemical exposure records were included in the candidate chemical exposure records. Entries in the <Indications of a Chemical Exposure Records>, referred to as indications from this point forth, were sorted and duplicates were removed. For each record a 1 (present) and 0 (not present) was added for each indication. Inclusion criteria were selected by identifying the combination of indications that were present for true positive records and not present for false positives. The best combination was: <Experimental_Type>=compound treatment, <Experimental_Type=dose, and <Experimental_Factor>=compound. These selection criteria seem obvious, but are in fact too general, i.e., identifying too many non-chemical exposure experiments as

chemical exposure experiments (i.e., false positives), and too specific, i.e., not correctly identifying all chemical exposure experiments as such (i.e., false negatives). An example of a falsely identified positive record using this filter is provided by the experiment "E-GEOD-1977," which is a radiation experiment with label <Experiment_Type>=compound treatment (in which the word "compound" does not refer to a chemical, but rather to a compound vs. simple treatment). On the other hand, an experiment not identified by the above filter criteria, but that can be classified as a chemical exposure experiment, is "E-SMDB-1896", identified as <Experiment_Type>=unknown where the Submitter's description states: "To directly characterize the effects of IFN-alpha in humans, we used microarrays to profile gene expression in peripheral blood mononuclear cells (PBMCs) from hepatitis C patients treated with IFN-alpha." Optimally, the inclusion criteria would not identify non-chemical exposure experiments. However, a combination of indications that resulted in 0 false positives was not possible. The best combination of indications still resulted in the identification of all 299 candidate records which included the 136 true positive records. Therefore, this method based only on the inclusion criteria resulted in a 55% (163/299) false positive error rate. All 136 chemical exposure records were identified leading to a 0% false negative rate.

Finally, using the inclusion criteria described above, an automated method was implemented to update the previously identified chemical exposure records with new chemical exposure records added to ArrayExpress since the previous annotation and curation. The automated method was implemented by first completing a bulk download of the public ArrayExpress content as of September 2008 in XML format. A text-mining program was written in Perl (Perl 2008), due to its particular suitability for text expression matching, and consists of several components: raw data clean-up; data extraction from plain text to tabular entries; identification of chemical-related information through pattern matching in Experimental_Type factors and user descriptions; creation of

a list of records with textual markers to easily identify chemical information; removal of any records already included in previous chemical annotations of ArrayExpress; and extraction of pertinent information for top-level indexing, including the Experiment_Accession_Number, Experiment_Type, Species, Chip_Type, and Details. Due to the necessity to review free-text description fields in ArrayExpress for primary chemical identifier information, output from the Perl script required a final step of visual inspection and manual review to extract a chemical name and any additional chemical identification information (such as CASRN). In addition, details that allowed for definitive assignment of the record as a chemical exposure experiment were recorded, and were iteratively incorporated into the Perl scripts to refine the inclusion criteria and reduce false positives.

The first and second manual curation methods required 1,363 records be manually annotated for identification of only 136 chemical exposure records. Using automated methods in combination with manual methods for the same dataset, only 299 records required manual annotation for the identification of 136 chemical exposure records resulting in a 55% false positive rate and a 0% false negative rate. After the complete automated method was refined only 1,617 chemical exposure records required manual annotation to identify 1,352 chemical exposure records resulting in a false positive rate of 16%. A false negative rate of 1% was estimated by manually reviewing 10% of records identified as non-chemical exposure. The false negative records were not "Treatment" records; therefore, this is an acceptable false negative rate. Great improvement was seen as a result of the refinement of the automated methods to account for experiments annotated or reviewed in previous annotations. As ArrayExpress continues to grow, a 16% false positive rate will decrease the amount of time necessary to update the chemical exposure experiments.

GEO *SERIES*

Similar to ArrayExpress, GEO is a large public resource (over 9000 series at the time of this writing), and has limited and unstructured chemical annotation, and dynamic user-deposited content (Figure 3-4). Hence, a manual method similar to that employed in the review of ArrayExpress was initially required. All data were downloaded from the GEO homepage in the GSE series form. Each of the Series was manually reviewed for chemical content and this information was used to construct an index of the chemical content that included the Series Accession Number, Chemical Name, Identification CAS-RN (if included), Species, and <Indications of a Chemical Exposure Record>. As in ArrayExpress, the <Indications of a Chemical Exposure Record> field contained details to aid in discerning whether a record pertained to a true chemical exposure experiment or not. From this chemical index, the first chemical annotation of GEO was completed. This method took a great deal of time, lacked efficiency, and did not represent a feasible long-term solution for indexing a live, growing public resource. As was done with ArrayExpress, this manually curated chemical index was used to test and refine automated curation approaches. GEO, however, significantly differs in its organization and content from ArrayExpress, thus, requiring significantly different procedures for indexing chemical treatment-related experiments.

Several automated methods were developed using NCBI Entrez Programming Utilities (E-Utilities) (NCBI E-Utilities 2008; NCBI E-Utilities 2004), i.e., tools that provide programming access to NCBI data outside of the regular web query interface. All automated methods use GDS Accession numbers, which index experiments in GEO to support data-mining access and are directly comparable to ArrayExpress accession numbers. The first method used the Medical Subject Headings (MeSH) vocabulary mapping of the GEO database, accessed from the Preview/Index option, as a query tool to retrieve the GEO experiments in a systematic way (NLM MeSH 2005). MeSH terms were indexed

and selected based on review of the manually constructed chemical index generated previously. A Perl script reformatted the MeSH terms for use in the E-utility scripts, and a modified NCBI Perl/E-Utility Script was used to submit searches based on these MeSH terms. To enable automated parsing of results, the E-search E-utility was used to send the MeSH terms query to the Backend NCBI-GEO server, and a second command retrieved a textual summary of each of the records. Additional Perl scripts converted the GEO records to XML format, parsed the XML output into readable text, and exported the candidate chemical exposure records to a new file, viewable in Microsoft Excel. This method returned all previously identified chemical exposure records, along with a large number of false positives, i.e., experiments incorrectly identified as chemical exposure records. As was done with ArrayExpress, keywords used in the script were edited to be more selective. The final selection criteria consisted of a title containing the following keywords and word fragments: “compound”, “treat”, “expos”, “dos”, and “chemical” or “drug”. Although producing fewer false positives, these criteria still retrieved several hundred records that did not qualify as chemical exposure experiments according to our definition, records that had to be located and eliminated from further consideration by tedious manual review.

To achieve greater efficiency, a second automated method more specifically targeted the chemical content of GEO using the chemical supplement of the MeSH vocabulary. An XML version of the chemical supplement was downloaded from the main MeSH search page (NLM MeSH 2005) and a Perl script was used to write the chemical terms to a separate file to be used as a query within the E-utility script. The E-Utility Script was edited to query these MeSH Terms against MeSH Terms internally mapped to GEO DataSets accession numbers (GDSxxx), with GDS numbers written to an XML file. Since GDS accession numbers refer to curated sets of GEO sample data rather than to individual experiments, it was difficult to compare results to, or improve upon the previous curation

method. Hence, at this stage, neither automated method was sufficiently robust to extract a chemical index of GEO chemical exposure-related experiments without a significant amount of manual curation and review.

To circumvent these problems, a third method was developed to parse through a complete XML version of the GEO Series database, downloaded using E-utilities, with a series of Perl scripts. As shown in Figure 3-4, GEO records have several components and we found it necessary to capture all aspects to extract the chemical name with certainty. In the example in Figure 3-4, the chemical name is only listed in the "Samples" title field, whereas in other instances the name and indications of a chemical exposure experiment occur only in the user-deposited description (see Figure 3-4). Whereas previously, we relied exclusively on parsing the latter description, the current method searches through the complete GEO record to identify chemical exposure experiments based on the previously identified keywords and word fragments ("compound", "treat", "expos", "dos", "chemical" or "drug"), with addition of the keyword "response". This method produced a 0% false negative rate and a 40% false positive rate, based on post analysis, which was an improvement from a 94% false positive rate using the manual method. To put this in context, at the time of this analysis, GEO contained 9567 series/experiments, of which 2364 were identified as pertaining to chemical exposure treatment-related experiments. Of these, 941 were false positives and 1423 were determined to be actual chemical exposure experiments. Reviewing 2364 records as opposed to 9567 records provided a significant gain in efficiency, although the false positive rate remains high. Further efforts to filter out false positive records with the keywords "radiation", "UV", "cold", "temperature", "gamma" or "heat" were unsuccessful since a number of experiments combined chemical treatment with temperature or radiation exposure, (e.g., GSE9463: Chemical toxicity of thorium in Saccharomyces cerevisiae). Future work may focus on additional means for increasing the efficiency of this method,

including modifying scripts to capture data in a more complete format, such as provided by Simple Omnibus Format in Text (SOFT) (NCBI GEO 2008c).

The above efforts evaluated relatively simple automated text-mining and keyword searching in free-text description fields for use in constructing chemical index files. A number of more sophisticated text-mining algorithms are available that potentially could have been applied to this task. However, given the highly variable nature of the text description field for microarray experiments, the lack of standardization and poor accuracy of chemical names (i.e., high incidence of errors and abbreviations), the lack of uniformity or consistency in the textual context for those names, and subtleties associated with discerning chemical use categories (i.e., treatment, reference, media, etc), etc, we deemed these more sophisticated tools unsuited to the present challenge.

Chemical Index Files

The main result of the above process was to produce a static, preliminary chemical index for chemical-associated microarray experiments in ArrayExpress and GEO. These preliminary index files took the form of a list of minimal chemical identifiers (most often chemical names and much less commonly including CASRN and/or ChEBI identifiers) directly extracted from the user-deposited information in these two resources. These chemical-experiment index files subsequently underwent a rigorous cleanup and chemical quality review to reconcile source-provided chemical information (i.e., chemical name is a valid name and agrees with CASRN or other identifier), to identify the relationship of the chemical to the experiment (e.g., treatment, vehicle, reference, etc.) and, for those experiments identified as "Treatment" by our definition, to add DSSTox Standard Chemical Fields (EPA DSSTox 2008b, 2008c).

DSSTox Standard Chemical Fields are divided into two categories: "TestSubstance" fields are uniquely indexed by generic substance identifier (DSSTox_Generic_SID) and include chemical name, CASRN (if available), and test substance description (e.g., single chemical compound, macromolecule, mixture or formulation, etc.) (Table 2-1, EPA DSSTox 2008c). Where the TestSubstance can be reasonably represented by a molecular structure, "STRUCTURE" fields are provided. STRUCTURE fields are uniquely indexed by compound identifier (DSSTox_CID) and include a standard, 2 dimensional "Molfile" representation of the chemical structure assigned to the substance (Dalby et al. 1992), several fields automatically derived from the Molfile structure (i.e., molecular weight, formula, IUPAC name, SMILES, SMILES_Parent, InChI, InChIKey), chemical type (i.e., defined organic, inorganic, organometallic), the tested form of the chemical if it is a defined organic (e.g., parent, Na salt, HCl complex, etc.) and, finally, a field indicating the relationship of the STRUCTURE to the TestSubstance (i.e., tested chemical, active ingredient of formulation, representative isomer in mixture, etc.). Such fields allow for standardized representation of both the test substance and the chemical structure in relation to any experiment or test record.

RESULTS & DISCUSSION

The major results of the above chemical indexing efforts were the creation and on-line publication of two DSSTox Structure-Index Locator files, GEOGSE and ARYEXP, and two DSSTox Auxiliary files, GEOGSE_Aux and ARYEXP_Aux, based on experimental microarray content extracted from GEO Series and ArrayExpress Repository, respectively, on September 20, 2008. All 4 files contain the full complement of DSSTox Standard Chemical Fields for each chemical record contained therein (Table 2-1). The Auxiliary data files are indexed by experiment and contain the full inventory of chemical-experiment pairs (one experiment per chemical) contained within GEO Series

and ArrayExpress Repository. These files include the full complement of 14 Standard Genomics Fields proposed earlier (Table 2-7), and include experiments for all Chemical_StudyType categories defined previously (i.e., Treatment, Reference, Media, Vehicle, etc.). In addition, since the Auxiliary data files are indexed by experiment, they can accommodate additional Source-specific content from either ArrayExpress or GEO pertaining to further experimental annotation provided in the two systems.

Whereas the Auxiliary data files can contain multiple records (for different experiments) pertaining to the same chemical substance (e.g., estradiol), the DSSTox Structure Index Locator files are indexed by unique chemical substance (i.e., one chemical mapped to multiple experiments) and contain only the Chemical_StudyType = "Treatment" (or combination Treatment AND…) subset of the total chemical-experiment inventory. In addition to the full complement of Standard Chemical Fields, these files include only a "Locator" field that contains one or more URLs that are linked, or indexed, to each of the microarray experiments associated with the particular chemical substance in either GEO or ArrayExpress.

These two sets of files (DSSTox Structure-Index Locator files and Auxiliary data files) enable, for the first time, an assessment of the chemical landscape associated with the largest public repositories of microarray experiments associated with the scientific literature, i.e. GEO Series and ArrayExpress Repository as of September 20, 2008. Table 3-2 provides a breakdown of the current chemical-associated experimental content within the GEO Series and ArrayExpress Repository according to Chemical_StudyType categories (Table 2-7). Of the 6346 total ArrayExpress experiments downloaded, 2365 chemical-associated experiment records were identified by the procedures outlined in the Methods section, corresponding to a total of 1011 unique chemical test substances (Table 3-2). Similarly, of the 9957 GEO Series experiments, 2381 chemical-experiment

records were identified, corresponding to a total of 1064 unique chemical test substances (Table 3-2). Hence, nearly 37% of ArrayExpress Repository experiments and 11% of GEO Series experiments are determined to be chemical-associated microarray experiments.

Table 3-3 provides a breakdown of the "Treatment"-associated experimental content within the GEO Series and ArrayExpress Repository according to the DSSTox chemical classification categories. Of the 1835 total "Treatment"-associated experiments in the ArrayExpress Repository, 1282 experiments (or 70% of the total) are associated with a "defined organic" chemical test substance. These are generally small molecular weight (<500 amu) organic chemicals for which a chemical structure can be assigned, and that tend to be of greatest interest for environmental toxicology and structure-activity relationship models and inferences. GEO Series contains a similarly high percentage of "Treatment" experiments associated with a defined organic chemical test substance, i.e. 1544/2134, or 72%. The above indicators give a rough sense of the size of the inventory of microarray experiments associated with defined organics in the public domain. Within ArrayExpress, the chemical that maps to the largest number of chemical-experiments is "estradiol", occurring in 53 experiments, 44 of which are classified as "Treatment" experiments.

Table 3-3 also provides indications of the size of the chemical space associated with these "Treatment" experiments. Of the total number of unique chemical test substances associated with the "Treatment" category of experiments in ArrayExpress Repository (i.e., 887), 628, or 71%, correspond to defined organics. A similar percentage applies to GEO, i.e., 751/1014, or 74% of unique chemical test substances associated with "Treatment" experiments correspond to defined organics. This indicates a relatively large number of unique defined organic chemicals, which implies a broad chemical diversity associated with public microarray experiments.

Of the 1835 "Treatment" experiments in ArrayExpress Repository, I determined that a small number, only 25, are explicitly labeled as "TOXM" experiments. As mentioned in Chapter 2, TOXM is a user-selected designation in ArrayExpress for a "toxicogenomics" experiment (EBI 2008c). For these experiments, MIAME/TOX guidelines (MGED 2008c) apply and users are encouraged (but not required) to provide additional details pertaining to the toxicogenomics experiment. These particular 25 experiments in ArrayExpress map to 66 unique chemical test substances, indicating that an explicitly labeled toxicogenomics experiment is more likely to be associated with chemical treatment and, additionally, to consider multiple chemical treatments.

Whereas the experimental content in ArrayExpress appears to have grown exponentially since its inception in 2003 (Figure 3-5), the total number of chemical exposure, or treatment-related experiments has increased at a more steady rate to current levels, comparable to the steady growth of toxicogenomics publications (estimated at approx 100 per year) over the past 5 years. In contrast, there has been a negligible increase in TOXM-designated toxicogenomics content in ArrayExpress during the same period of time (Figure 3-5). Hence, some submitters performing major toxicological experiments use the TOXM-designation for their data, but the label is rarely used and does not reflect the larger relevance of non-TOXM, "Treatment" experiments to toxicogenomics.

COMPARISON OF GEO AND ARRAYEXPRESS CONTENT

DSSTox Standard Chemical Fields and the additional Standard Genomics Fields are designed to be sufficiently generic for application to multiple sources and diverse resources. In contrast, Source-specific content fields have been included within the published DSSTox Auxiliary data file (Table 3-4 and 3-5). In the case of ArrayExpress, there are a number of easily extracted field characteristics that can be affiliated with each experiment, including Array/Platform type, Species and

the MIAME Score and its five subcategories: Array or Platform information, Factor information, Raw Data information, Processed Data information, and Protocol information (Table 2-7). The latter annotations are particularly valuable for assessing the sufficiency of the experimental data for reanalysis. The distribution of MIAME scores within the "Treatment" content in the ArrayExpress Repository is particularly illuminating. A total MIAME Score of 5 indicates that all components of the MIAME compliance criteria have been included by the submitter. Only 18% (or 216) of the chemical exposure experiments have all 5 components of information, whereas 50% (or 596) have 4 components of information (Table 3-6). Most noteworthy, however, for this subset of "Treatment" experiments, raw data information is missing for 29% (or 347), processed data is missing for 11% (or 131), and protocol is missing for 21% (or 291) (Table 3-6). Given that these are essential experimental components for the reanalysis of gene expression data, these numbers further reduce the number of chemical treatment experiments within ArrayExpress that are potentially useful for broader toxicogenomics investigation.

The ArrayExpress Repository has experienced steep growth over the last year, largely as a result of the integration of GEO experimental content (Figure 3-5). ArrayExpress files with E-GEOD-XXXX accession numbers mirror GEO Series entries and currently represent more than 50% of the chemical exposure experiments in ArrayExpress (Figure 3-6). Figure 3-6 also shows that the total number of chemical-experiment pairs (a pair being a 1:1 mapping of chemical to experiment) and total number of "Treatment" chemical-experiment pairs identified in the current study are comparable between ArrayExpress and GEO, with greater than 50% overlap of chemical-experiment pairs in all categories.

Since a significant portion of the "Treatment" experiments represented in GEO Series are also included in the ArrayExpress Repository, it was possible to create a snap-shot content analysis of

GEO chemical exposure experiments, similar to that of all ArrayExpress chemicals (Table 3-7). Only 11% (or 81) of the GEO records in ArrayExpress have a MIAME Score of 5; however 56% (or 415) have a MIAME score of 4. A much greater percentage, 45% (or 335) of GEO records in ArrayExpress, have corresponding Raw Data, whereas 100% (or 745) have Processed Data. GEO Series content has grown exponentially since its inception; however, as was seen with the ArrayExpress Repository, the chemical treatment content lags behind and is difficult to identify. Similar to ArrayExpress, GEO serves its purpose as a repository but falls short of supporting toxicogenomics investigation on a number of fronts, most glaringly in providing no chemical indexing. Unlike ArrayExpress, GEO Series also provides no MIAME scoring of content, so it is difficult to assess sufficiency of annotation across the entire inventory. In addition, since user-deposited information is largely free-text and unaccompanied by controlled vocabulary entries, extensive curation is required and crucial information for experimental replication is often found to be missing. GEO Datasets provides a valuable complement to GEO Series and adds useful functionality, post analysis, and relational query capability across experimental parameters. However, the creation of these Datasets within GEO is labor-intensive, lags significantly behind the total Series inventory, and is not well coordinated with other public efforts and data repositories.

Figure 3-7 presents the overlap of the unique chemical content pertaining to the "Treatment" chemical-experiment category. Assessment of chemical overlap between GEO and ArrayExpress DSSTox files was determined on the basis of TestSubstance using the DSSTox_Generic_SID identifiers. The steroids, estradiol and dexamethasone, are associated with the largest numbers of microarray experiments in both cases, and the largest number of shared experiments as well. Other test substances most commonly associated with experiments in either GEO or ArrayExpress include

Ethanol, 2,3,7,8-TCDD (Tetrachlorobenzo-p-dioxane) Retinoic Acid, and Trichostatin, each of which has relevance to the field of toxicology.

DSSTox Chemical Index Files & Linkages

As discussed above, this project resulted in the creation of 4 different files, ARYEXP, ARYEXP_Aux, GEOGSE, and GEOGSE_Aux. All 4 files are available for download from the DSSTox website (http://www.epa.gov/ncct/dsstox/) in both Excel and Structure-Data File (SDFile) formats (Figures 3-8 through 3-11). ARYEXP_Aux contains a total of 2365 chemical-experiment records (with 44 source fields), corresponding to 1011 unique chemical substances. Of these 2365 chemical-experiment pairs, 1835 were identified as "Treatment" and these map to 887 unique chemical records in the ARYEXP file. The GEOGSE_Aux file contains a total of 2381 chemical-experiment records (with 18 source fields), corresponding to 1064 unique chemical substances. Of these 2381 chemical-experiment pairs, 2134 were identified as "Treatment" and these map to 1014 unique chemical records in the GEOGSE file.

DSSTox Structure-Index files for GEO and ArrayExpress enable, for the first time, an examination of the chemical diversity and coverage of the GEO Series and ArrayExpress Repository experiments. We found significant numbers of experiments in both resources mapped to families of similar chemicals, as well as to a broad diversity of chemical structures, spanning a wide range of toxicologically relevant chemical functional hierarchies and classes (Figure 3-12). The availability of several microarray experiments within a class of similar chemicals, such as shown in Figure 3-12 for benzopyrans, increases the likelihood of finding sufficiently similar protocol experiments for toxicogenomics reanalysis and structure-activity inferences.

A further metric of toxicological relevance of the current GEO and ArrayExpress chemical inventory is provided by the overlap of the chemicals in GEO and ArrayExpress with the current DSSTox inventory, which includes more than 7000 unique chemicals and spans a variety of environmentally and toxicologically relevant chemical inventories and data sets from the various sources (EPA DSSTox 2008e). A total of 551 unique chemical test substances in the GEO and/or ArrayExpress DSSTox Structure-Index Locator files overlap with one or more of the 11 previously published DSSTox Data Files (EPA DSSTox 2008e), and there are a total of 1294 overlapping instances. Of these overlapping instances, 3 chemical substances (Bisphenol A, di(2-ethylhexyl) phthalate or DEHP, and dibutylphthalate) occur in 8 DSSTox Data Files, 13 chemical substances occur in 7 DSSTox Data Files, and a total of 309 chemical substances occur in 2 or more DSSTox Data Files. The GEOGSE file has 468 overlapping occurances with the 11 previously published DSSTox files and ARYEXP has 317. Examining the structure overlap of the GEOGSE and ARYEXP files individually with the 11 previously published DSSTox files, one DSSTox file (National Toxicology Program Bioassay On-line database) has 214 and 170 overlapping chemicals, respectively. Further, 261 GEOGSE chemicals and 194 ARYEXP chemicals occur in at least 2 DSSTox data files(Figures 3-13 and 3-14). These numbers indicate that significant numbers of chemicals of potential toxicological concern, for which additional *in vitro* or *in vitro* data possibly exist, are contained in both the ArrayExpress Repository and GEO Series and accompanied by microarray data.

The GEOGSE and ARYEXP Structure-Index Locator files with associated experiment Accession ID URLs have been incorporated into the public DSSTox Structure-Browser (http://www.epa.gov/dsstox_structurebrowser/), allowing these files to be structure-searched in the context of the full DSSTox Data File inventory of over 7000 unique substances. Figure 3-15 shows a

screen shot of the Browser's main search page, in which the chemical "estradiol" is entered for either a text or a structure search. Figure 3-16 shows a screen shot of the result of this structure search, where the "exact match" for estradiol in the file ARYEXP was chosen and the Substance Results page is shown. A user can link directly from this search results page to any one of the listed URLs for experiment summary pages within ArrayExpress associated with the structure for estradiol, regardless of the spelling or representation of that name within the actual ArrayExpress experiment descriptions. Additional data for estradiol and its structurally similar analogs can be accessed through this browser, and link-outs are also provided to "External Resources", such as PubChem, ACToR (EPA ACToR 2008; Judson et al. 2008), and ChemSpider (Chemspider 2008) from the indicated substance.

The GEOGSE and ARYEXP Structure-Index Locator files have also been deposited in PubChem as chemical-experiment pairs, i.e. one experiment per chemical. PubChem substances for these two files can be located with main text keyword searches that include "arrayexpress", "gene expression", "NCBI GEO", etc. All substances in either file can be retrieved in this manner, with the results for ARYEXP shown in Figure 3-17. We have provided PubChem with URLs such that a user can link directly from the PubChem substance listing to the corresponding ArrayExpress experiment summary page as shown in this Figure. In addition, users can access the full PubChem inventory of bioassay information and linkages for each substance and related substances.

Furthermore, all files will be incorporated within CEBS to enhance their utility for toxicogenomics and to be searchable in the larger context of the CEBS relational database environment. Finally, all files will be made available to NCBI's GEO, and EBI's ArrayExpress and ChEBI projects in the hopes that the curated and quality reviewed chemical annotation content can be directly utilized. Figure 3-18 illustrates the use of public linkages to aggregate a wide range of

potentially relevant toxico-chemogenomics data for a particular chemical, e.g., acetaminophen, now represented in a uniform, standardized way in relation to chemical structure. Additionally, structure-searching allows for the aggregation of data on structural analogs to be located from within or across data resources, a capability that allows users to begin to address the issue of data gaps and to explore structure-activity associations within the genomics data world.

CONCLUSIONS

With the chemical indexing of the two largest public repositories of microarray data, i.e. GEO and ArrayExpress, it is now possible to: 1) assess the chemical landscape of microarray experiments associated with the scientific literature; 2) structurally search the content of ArrayExpress and GEO and locate experiments pertaining to particular structures or structural classes; 2) compare the content of ArrayExpress and GEO; and 3) view gene expression data in relation to other data types such as toxicological data in DSSTox and other biological and chemical data in PubChem.

Practically speaking, public microarray data repositories cannot limit their content to include only experiments adhering to strict common protocol standards and traditional model organisms. A public data resource can, however, strive for completeness and accuracy of experimental annotations and to provide user-access to raw data for reanalysis. Also, whereas standardization and chemical indexing of toxicogenomics experiments at the time of publication is the ideal, if pertinent information is collected at the time of data deposition, this can be accomplished post-publication with manual or automated text-mining and reformatting. With such steps, it becomes possible to assess the chemical coverage of public gene expression databases, to link data for common or similar chemicals across information domains, including toxicology, as well as to gather data from comparable experiments,

possibly performed in different labs and species, that can begin to serve as the basis for meta-analysis or structure-activity hypotheses.

The process of identifying chemical treatment-related experiments in the two major public microarray repositories, ArrayExpress Repository and GEO Series, was time-consuming and difficult to automate, emphasizing the woefully inadequate chemical annotation of these data resources. These efforts also serve to highlight deficiencies in microarray experiment data deposition requirements and standards with regard to chemistry and chemical-treatment related experiments that, if better addressed, could greatly facilitate chemical annotation and data integration efforts in the future. In the course of this work, I implemented a set of Standard Genomics Fields, most of which map to existing fields from both GEO and ArrayExpress to provide a bridge between the two resources and to facilitate comparisons and incorporation of their content into other resources in a standardized way.

It is hoped that the current exercise to create, publish, and link chemical index files has had two primary impacts: 1) to highlight deficiencies in the current chemical annotation and curation methods within ArrayExpress and GEO that particularly impact toxicogenomics applications of these resources; and 2) to show the way forward in terms of what is needed and what potential benefits can be derived by incorporating robust chemical annotation and linkages of chemical treatment-related content to the public resources. Recently improved coordination of the EBI ArrayExpress and ChEBI projects is a significant step forward and should immediately benefit from incorporation of the newly generated ArrayExpress chemical-experiment index file, as well as the addition of the corresponding GEO index file. However, as is apparent from past failures, it is not sufficient to recommend that users add CASRN or ChEBI identifiers at the time of data submission unless more stringent efforts to require this information are instituted. In addition, we strongly recommend adoption of the “Chemical_StudyType” categories, or something comparable, for each chemical-

associated study or experiment deposited into GEO and ArrayExpress. Finally, improved coordination with external efforts to provide linkages to ArrayExpress and GEO experiments based on chemical structure are needed to ensure that externally indexed content is kept current.

REFERENCES

Burgoon LD. 2007. Clearing the standards landscape: The semantics of terminology and their impact on toxicogenomics. Toxicol Sci 99(2):403-412; doi: 10.1093/toxsci/kfm108.

CDISC (Center for Data Interchange Standards Consortium). 2008. CDISC Homepage. Available: http://www.cdisc.org/models/send/v2.3/index.html [accessed 21 October 2008].

Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A et al. 2008. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344-D350.

DevTox. 2005. Developmental Toxicology Nomenclature Homepage. Available:http://www.devtox.org/index.htm [accessed 7 October 2008].

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci 95(1):5-12; doi: 10.1093/toxsci/kfl103.

EBI (European Bioinformatics Institute). 2008b. ArrayExpress Programmatic Access. Available: http://www.ebi.ac.uk/microarray/doc/help/programmatic_access.html (ArrayExpress) [accessed 7 October 2008].

EBI (European Bioinformatics Institute). 2008d. ArrayExpress Guide to Accession Codes. Available: http://www.ebi.ac.uk/microarray/doc/help/accession_codes.html [accessed 7 October 2008].

EBI (European Bioinformatics Institute). 2008f. Sample Data Relationship File (SDRF) Explanation Page. Available: http://tab2mage.sourceforge.net/docs/sdrf.html [accessed 10 October 2008].

Fostel J, Choi D, Zwickl C, Morrison N, Rashid A, Hasan A et al. 2005. Chemical Effects in Biological Systems--data dictionary (CEBS-DD): A compendium of terms for the capture and integration of biological study design description, conventional phenotypes, and 'omics data. Toxicol Sci 88(2):585-601; doi: 10.1093/toxsci/kfi315.

Fostel JM, Burgoon L, Zwickl C, Lord P, Corton JC, Bushel PR et al. 2007. Toward a checklist for exchange and interpretation of data from a toxicology study. Toxicol Sci 99(1):26-34; doi: 10.1093/toxsci/kfm090.

Larsson O, Sandberg R. 2006. Lack of correct data format and comparability limits future integrative microarray research. Nat Biotechnol 24(11):1322-1323; doi: 10.1038/nbt1106-1322.

Martin MT, Judson RS, Reif DM, Dix DJ..2008. Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. Environ Health Perspect; doi: 10.1289/ehp.0800074.

MIBBI (Minimum Information for Biological and Biomedical Investigations) 2008. MIBBI Homepage. Available: http://mibbi.org/index.php/Main_Page [accessed 21 October 2008].

Richard A, Yang C, Judson R. 2008. Toxicity data informatics: Supporting a new paradigm for toxicity prediction. Tox Mech Meth 18:103-118.

Richard AM, Gold LS, Nicklaus MC. 2006. Chemical structure indexing of toxicity data on the internet: Moving toward a flat world. Curr Opin Drug Discov Devel 9(3):314-325.

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, et al. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnol 26(8):889-896.

Accelrys. 2008. Pipeline Pilot Chemistry and Excel ChemMining Collection. Available: http://accelrys.com/events/webinars/chemial-text-mining/index.php [accessed 20 October 2008].

ACDlabs 2008. Chemical Name to Structure Conversion Tool. Available: www.acdlabs.com/products/name_lab/ [accessed 20 October 2008].

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C et al. 2007. NCBI GEO: Mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35(Database issue):D760-5; doi: 10.1093/nar/gkl887.

Brazma A, Kapushesky M, Parkinson H, Sarkans U, Shojatalab M. 2006a. Data storage and analysis in ArrayExpress. Methods Enzymol 411:370-386; doi: 10.1016/S0076-6879(06)11020-4.

Brazma A, Parkinson H, ArrayExpress team E. 2006b. ArrayExpress service for reviewers/editors of DNA microarray papers. Nat Biotechnol 24(11):1321-1322; doi: 10.1038/nbt1106-1321.

ChemSpider. 2008. ChemSpider HomePage. Available: http://chemspider.com [accessed 7 October 2008].

Cohen, AM and Hersh, WR. 2005, A survey of current work in biomedical text mining. Brief Bioinform 6(1):57-71

Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier, DL, Leland BA, Laufer J. 1992. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited, J Chem Inf Comput Sci 32:244-255.

EBI (European Bioinformatics Institute) ChEBI (Chemicals of Biological Interest). 2008. CheBI Homepage. Available: http://www.ebi.ac.uk/chebi/ [accessed 27 October 2008].

EBI (European Bioinformatics Institute). 2008a. ArrayExpress Homepage. Available: http://www.ebi.ac.uk/arrayexpress [accessed 7 October 2008].

EBI (European Bioinformatics Institute). 2008b. ArrayExpress Advanced Query Page. Available: http://www.ebi.ac.uk/microarray-as/aer/entry [accessed 7 October 2008].

EBI (European Bioinformatics Institute). 2008c. Tox-MIAMExpress for ArrayExpress Homepage. Available: http://www.ebi.ac.uk/tox-miamexpress [accessed 7 October 2008].

EPA (Environmental Protection Agency) ACToR. 2008. Aggregated Computational Toxicology Resource (ACTor) Homepage. Available: http://www.epa.gov/actor/ [accessed 7 October 2008].

EPA (Environmental Protection Agency) DSSTox. 2008a. Distributed Structure-Searchable Toxicity (DSSTox) Database Network - Homepage. Available: http://www.epa.gov/ncct/dsstox/index.html [accessed 7 October 2008].

EPA (Environmental Protection Agency) DSSTox. 2008b. Distributed Structure-Searchable Toxicity (DSSTox) Database Network - Chemical Information Quality Review Procedures. Available: http://www.epa.gov/ncct/dsstox/ChemicalInfQAProcedures.html [accessed 7 October 2008].

EPA (Environmental Protection Agency) DSSTox. 2008c. Distributed Structure-Searchable Toxicity (DSSTox) Database Network - More Information on DSSTox Standard Fields. Available: http://www.epa.gov/ncct/dsstox/MoreonStandardChemFields.html [accessed 7 October 2008].

EPA (Environmental Protection Agency) DSSTox. 2008d. Distributed Structure-Searchable Toxicity (DSSTox) Database Network - Structure Browser. Available: http://www.epa.gov/dsstox_structurebrowser/ (DSSTox 2004d) [accessed 7 October 2008].

EPA (Environmental Protection Agency) DSSTox. 2008e. Distributed Structure-Searchable Toxicity (DSSTox) Database Network - Structure Data Files. Available: http://www.epa.gov/ncct/dsstox/DataFiles.html [accessed 7 October 2008].

Judson R, Richard A, Dix D, Houck K, Elloumil F, Martin M et al. 2008. ACToR – Aggregated Computational Toxicology Resource, Toxicol Appl Pharmacol doi:10.1016/j.taap.2007.12.037.

MGED (Microarray Gene Expression Data Society). 2008a. MGED Homepage. Available: http://www.mged.org [accessed 7 October 2008].

MGED (Microarray Gene Expression Data Society). 2008c. MGED recommended Minimum Information About a Microarray Experiment – MIAME for Toxicogenomics (MIAME/Tox). Available: http://www.mged.org/MIAME1.1-DenverDraft.DOC [accessed 7 October 2008].

MGED (Microarray Gene Expression Data) Society. 2008b. MGED Recommended Minimum Information About a Microarray Experiment (MIAME) Checklist. Available: http://www.mged.org/Workgroups/MIAME/miame.html [accessed 7 October 2008].

NCBI (National Center for Biotechnology Information) E-Utilities (Entrez Utilities). 2008. Entrez Utilities Homepage. Available: http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html [accessed 7 October 2008].

NCBI (National Center for Biotechnology Information) E-Utilities (Entrez Utilities). 2004. Entrez Utilities Newsletter announcement. Available: http://www.ncbi.nlm.nih.gov/Web/Newsltr/SummerFall04/sumfall04.pdf [accessed 7 October 2008].

NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus). 2008a. GEO Homepage. Available: http://www.ncbi.nlm.nih.gov/geo [accessed 7 October 2008].

NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus). 2008b. GEO Programmatic Access. Available: http://www.ncbi.nlm.nih.gov/projects/geo/info/geo_paccess.html [accessed 7 October 2008].

NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus). 2008c. EO Information on SOFT format. Available: http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html#BatchupdatesinSOFT [accessed 7 October 2008].

NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus). 2008d. GEO Dataset File Example for download of PEPR database. Available: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds [accessed 7 October 2008]

NCBI (National Center for Biotechnology Information) PubChem. 2008. PubChem HomePage. Available: http://pubchem.ncbi.nlm.nih.gov [accessed 7 October 2008].

NLM (National Library of Medicine) MeSH (Medical Subject Headings). 2008. MeSH Homepage. Available: http://www.ncbi.nlm.nih.gov/sites/entrez?db=mesh [accessed 7 October 2008].

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A et al. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35(Database issue):D747-50; doi: 10.1093/nar/gkl995.

Perl. 2008. Perl Programming Language Homepage. Available: http://www.perl.com [accessed 7 October 2008].

Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. 2008. Text processing through Web services: calling Whatizit. Bioinform 24(2):296-298.

Rustici G, Kapushesky M, Kolesnikov N, Parkinson H, Sarkans U, Brazma A. 2008. Data storage and analysis in ArrayExpress and expression profiler. Curr Protoc Bioinformatics Chapter 7:Unit 7.13; doi: 10.1002/0471250953.bi0713s23.

Uramoto N, Matsuzawa H, Nagano T, Mura Kami A, Takeuchi H, Takeda K, 2004. A text-mining system for knowledge discovery from biomedical documents. IBM Systems Journal. 43(3):516-533.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V et al. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36(Database issue):D13-21; doi: 10.1093/nar/gkm1000.

TABLES

Table 3-1. List of ArrayExpress experiments involving treatment with 17β-estradiol, showing the varied chemical name spellings and synonyms extracted from the Submitter's descriptions.

Accession Number	Listed Chemical Name
E-MEXP-132	estrogen
E-MAXD-39	oestradiol
E-GEOD-4025	estradiol (E2)
E-GEOD-848	E2
E-NASC-65	estradiol
E-SMDB-1443	Beta-estradiol
E-TABM-269	Estradiol
E-AFMX-12	17 beta-estradiol (E2)
E-AFMX-13	Estrogen
E-GEOD-1045	Estradiol
E-GEOD-1153	Estradiol E2
E-GEOD-2195	17beta-estradiol
E-GEOD-2251	17beta-estradiol
E-GEOD-2292	17beta-estradiol
E-GEOD-2889	17beta-estradiol
E-GEOD-3529	Estradiol(E2)
E-GEOD-4668	17beta-estradiol
E-MEXP-1053	Estradiol
E-TABM-231	Estradiol
E-GEOD-2225	17 B-estradiol
E-GEOD-3013	Oestradiol (Oestrogen)
E-TABM-275	Estrogen
E-MEXP-1147	17-beta estradiol (E2)
E-GEOD-6219	17-estradiol (E2)
E-GEOD-628	Estradiol
E-GEOD-1819	17beta-estradiol
E-GEOD-1839	17beta-estradiol

Table 3-2. Classification of chemically indexed genomics experiments in ArrayExpress and GEO by Chemical_StudyType.

Database[1]	Total # of Experiments[2]	Total # of Chemical-Experiment Records[3]	Total # of Unique Chemicals[4]	Chemical_StudyType Classification [5] Breakdown of Total # of Chemical-Experiment Records (Unique Chemicals) [6]						
				Treatment[7]	Reference	Vehicle	Media	Combination Treatment[7]	Multiple Classifications	Other
ArrayExpress Repository	6346	2365	1011	1609(810)	266(157)	138(26)	111(68)	109(91)	118(83)	14(10)
GEO Series	9957[8]	2381	1064	1951(838)	152(60)	81(48)	14(14)	72(38)	111(67)	0(0)

Table 3-2. (Continued) Footnotes

[1] All numbers relate to database content extracted on 20 September 2008.

[2] Total number of experiments contained in the public resource (also corresponds to the number of unique Accession IDs).

[3] Number of unique Chemical-Experiment pairs extracted from the Total # of Experiments prior to determination of the Chemical_StudyType Classification (Table 2-7) where some experiments in the Total # of Experiments map to no chemicals, and some experiments involving multiple chemicals map to more than one Chemical-Experiment record.

[4] Total number of unique chemicals (i.e., no chemical identity is duplicated) identified in the total group of Chemical-Experiment records, irrespective of Chemical_StudyType Classification.

[5] Refer to definitions of Chemical_StudyType Classifications in the Standard Genomics Fields listing in Table 2-7.

[6] Number of Chemical-Experiment Records corresponding to each Chemical_StudyType category (with corresponding number of unique chemicals in parentheses), where for the purposes of this table, one record is assigned to one category and if the chemical is used for different purposes within one experiment (e.g., TreatmentANDReference), it is assigned to the "Multiple Classifications" category.

[7] Number of Chemical-Experiment Records (with corresponding number of unique chemicals in parentheses) out of the total group of Chemical-Experiment Records that are associated with the "Treatment" category according to the criteria for a chemical-exposure scenario set forth in this paper; any record labeled as "Treatment" or "CombinationTreatment" (alone or in combination with other Chemical_StudyType labels, e.g. TreatmentANDReference), are included in the final DSSTox chemical index file.

[8] Total number of GEO series experiments pulled from the GDS system is less than 9957 due to a backlog of GEO series into the GDS system. It is estimated that approximately 6000 GEO series were extracted from the GDS system for chemical indexing.

Table 3-3. Classification of chemically indexed "Treatment" genomics experiments in ArrayExpress and GEO by DSSTox Chemical Classification.

Database[1]	Total # of Chemical-Experiment Records [2]	Total # of "Treatment" Chemical-Experiment Records [3]	Total # of Unique Chemicals [4]	DSSTox Chemical Classification [5] Breakdown for "Treatment" Chemical-Experiment Records (Unique Chemicals) [6]			
				No structure [7]	Defined Organic	Inorganic	Organometallic
ArrayExpress Repository	2365	1835	887	373(179)	1282(628)	153(60)	27(20)
GEO Series	2381	2134	1014	346(173)	1544(751)	210(71)	34(19)

Table 3-3. (Continued) Footnotes

[1] All numbers relate to database content extracted on 20 September 2008.

[2] See Table 3-2.

[3] Total number of Chemical-Experiment records assigned to any "Treatment" Chemical_StudyType (e.g., Treatment, CombinationTreatment, Treatment&Reference, etc.) according to the criteria for a chemical-exposure scenario set forth in this chapter.

[4] Total number of unique chemicals (i.e., no chemical identity is duplicated) identified in the total group of "Treatment" Chemical-Experiment Records.

[5] See Table 2-1

[6] Number of "Treatment" Chemical-Experiment Records corresponding to each Chemical Classification category (with corresponding number of unique chemicals in parentheses), where each record maps to a single chemical classification and the list of unique chemicals for this "Treatment" subset of experiments constitutes the final DSSTox chemical index.

[7] Number of "Treatment" Chemical-Experiment Records (with corresponding # of unique chemicals) where the chemical is identified, but not assigned to a DSSTox chemical structure, e.g., this can be an undefined mixture, polymer, or macromolecule.

Table 3-4: GEOGSE_Aux source-specific fields.

Field	Format	Description
Experiment_SupplementalFile	text	If supplemental data is submitted, the format of the data
Sample_AccessionNumber	text	GSMXXX sample Accession Numbers
Sample_Title	text	Title of each of the samples
Experiment_ArrayTechnologyType	text	Type of Gene Expression Technology

Table 3-5. Additional ArrayExpress Source-Specific Fields contained in the ARYEXP Auxiliary chemical-experiment index file

Field Name	Field Type	Description
Experiment_ArrayCount	numeric	Number of Arrays submitted by submitter
Experiment_ArrayId	numeric	Id Number of Array
Experiment_BiassayAccessionCode	text	Accession Code for Bioassay http://ebi.ac.uk/microarray/doc/help/accession_codes.html
Experiment_BioassayCount	text	Number of Bioassays
Experiment_BioassayDataFormat	text	Delimiter used for Bioassay results. Example: whitespace
Experiment_BioassayId	text	Bioassay Id Number
Experiment_BioassayTitle	text	Title of Derived Bioassay
Experiment_DesignFile_PNG	text	Experiment Design File in Potable Network Graphics (PNG) Format
Experiment_DesignFile_SVG	text	Experiment Design File as a Scalable Vector Graphics (SVG) File (
Experiment_DesignType	text	Experiment Type Example: Dose Response
Experiment_DescriptionId	numeric	Id Number for Description
Experiment_Factors	text	List of Experimental Factors Example: Dose
Experiment_SampleAttributes	text	List of Sample Attributes Example: Developmental Stage
Experiment_Submitter	text	Contact Information for Submitter if Provided
Link_ProcessedDataFiles	url	Link to Download Processed Data
Link_RawDataFiles	url	Link to Download Raw Data
Link_SampleDataRelationshipFile	url	Tab-delimited Detailed Sample Annotation
Link_TwoColumnData	url	Tab-delimited Sample Annotation

Table 3-5. (Continued)

MIAME_ArrayDesign_Score	numeric	Specific information about the design of the array or the platform used was submitted, 1, or not submitted, 0, with the experiment by the submitter.
MIAME_Factors_Score	numeric	A list of experimental factors was submitted, 1, or not submitted, 0, with the experiment by the submitter. Factors might include information on the cell line or particular compounds and dose information used in the experiment.
MIAME_ProcessedData_Score	numeric	Processed data was submitted, 1, or not submitted, 0, with the experiment by the submitter.
MIAME_RawData_Score	numeric	Raw data was submitted, 1, or not submitted, 0, with the experiment by the submitter.
MIAME_Protocol		Specific information about the experimental protocols used in the experiment was submitted, 1, or not submitted, 0, with the experiment by the submitter.
MIAME_Total	numeric	The Total MIAME score ranges from 0 to 5 and is a sum of the independent scores of the five subcomponent scores, each of which takes on the value of either 0 or 1 (absent or present).
Number BioassayDerived	numeric	Number of Boassay Entries that can be computationally derived.
Number_BadCubes	numeric	Number of Boassay Entries that cannot be computationally derived.
Number_CelFiles_Affymetrix	numeric	Number of Cel Files Submitted
Number_Hybridizations	numeric	Number of Hybridizations Submitted

Table 3-6. Characteristics of the ArrayExpress Repository pertaining to 1181 "Treatment" Chemical Experiment Records for Unique Chemicals (based on data extracted on 20 September 2008).

Characteristics	Major Characteristic Value	Number (%) of Chemical Exposure Experiments
Array/Platform	Affymetrix	861 (73%)
	Agilent	82 (7%)
	Stanford Microarray Database (SMD)	37 (3%)
	Complete Arabidopsis Transcriptome MicroArray (CATMA)	29 (2%)
	Sanger Institute	10 (<1%)
	GE Health Care	5 (<1%)
	Illumina	5 (<1%)
	Codelink	5 (<1%)
	Other	141 (12%)
	Not Listed	4 (<1%)
Species	Homo sapiens	377 (32%)
	Mus musculus	317 (27%)
	Rattus	173 (15%)
	Arabidopsis	159 (13%)
	Saccharomyces cerevisiae	55 (5%)
	Drosophila melanogaster	13 (1%)
	Escherichia coli	13 (1%)
	Other	74 (6%)
MIAMEScore_Total [a]	5	216 (18%)
	4	596 (50%)
	3	309 (26%)
	2	55 (5%)
	1	6 (1%)
MIAMEScore_Array [b]	0	78 (7%)
	1	1103 (93%)
MIAMEScore_Factor [c]	0	551 (47%)
	1	630 (53%)
MIAMEScore_RawData [d]	0	347 (29%)
MIAMEScore_ProcessedData [e]	0	131 (11%)
	1	1050 (89%)
MIAMEScore_Protocol [f]	0	291 (21%)
	1	890 (79%)

Table 3.6. (Continued) Footnotes

[a] The Total MIAME score ranges from 0 to 5 and is a sum of the independent scores of the five subcomponent scores, each of which takes on the value of either 0 or 1 (absent or present).
[b] Specific information about the design of the array or the platform used was submitted , 1, or not submitted, 0, with the experiment by the submitter. Included Array information is assigned an Array Accession number (see Table 2-7) within ArrayExpress.
[c] A list of experimental factors was submitted, 1, or not submitted, 0, with the experiment by the submitter. Factors might include information on the cell line or particular compounds and dose information used in the experiment.
[d] Raw data was submitted,1, or not submitted, 0, with the experiment by the submitter.
[e] Processed data was submitted, 1, or not submitted, 0, with the experiment by the submitter.
[f] Specific information about the experimental protocols used in the experiment was submitted, 1, or not submitted, 0, with the experiment by the submitter. Included Protocol information is assigned a Protocol Accession number within ArrayExpress.

Table 3-7. Characteristics of the GEO Series pertaining to 745 "Treatment" Chemical Experiment Records for Unique Chemicals, extracted from data mirrored in ArrayExpress (based on data extracted on 20 September 2008).

Characteristics	Major Characteristic Value	Number (%) of Chemical Exposure Experiments
Array/Platform	Affymetrix	691 (93%)
	Agilent	54 (7%)
Species	Homo sapiens	264(35%)
	Mus musculus	220 (30%)
	Arabidopsis	76 (10%)
	Rattus	126 (17%)
	Saccharomyces cerevisiae	15(2%)
	Other	44 (6%)
MIAMEScore_Total[a]	5	81(11%)
	4	415 (56%)
	3	211 (28%)
	2	38 (5%)
	1	0 (0%)
MIAMEScore_Array	0	0 (0%)
	1	745 (100%)
MIAMEScore_Factor	0	421 (57%)
	1	324(43%)
MIAMEScore_RawData	0	335 (45%)
	1	410 (55%)
MIAMEScore_ProcessedData	0	0 (0%)
	1	745 (100%)
MIAMEScore_Protocol	0	195 (26%)
	1	550 (74%)

[a-f] See legends for Table 2-10.

FIGURES

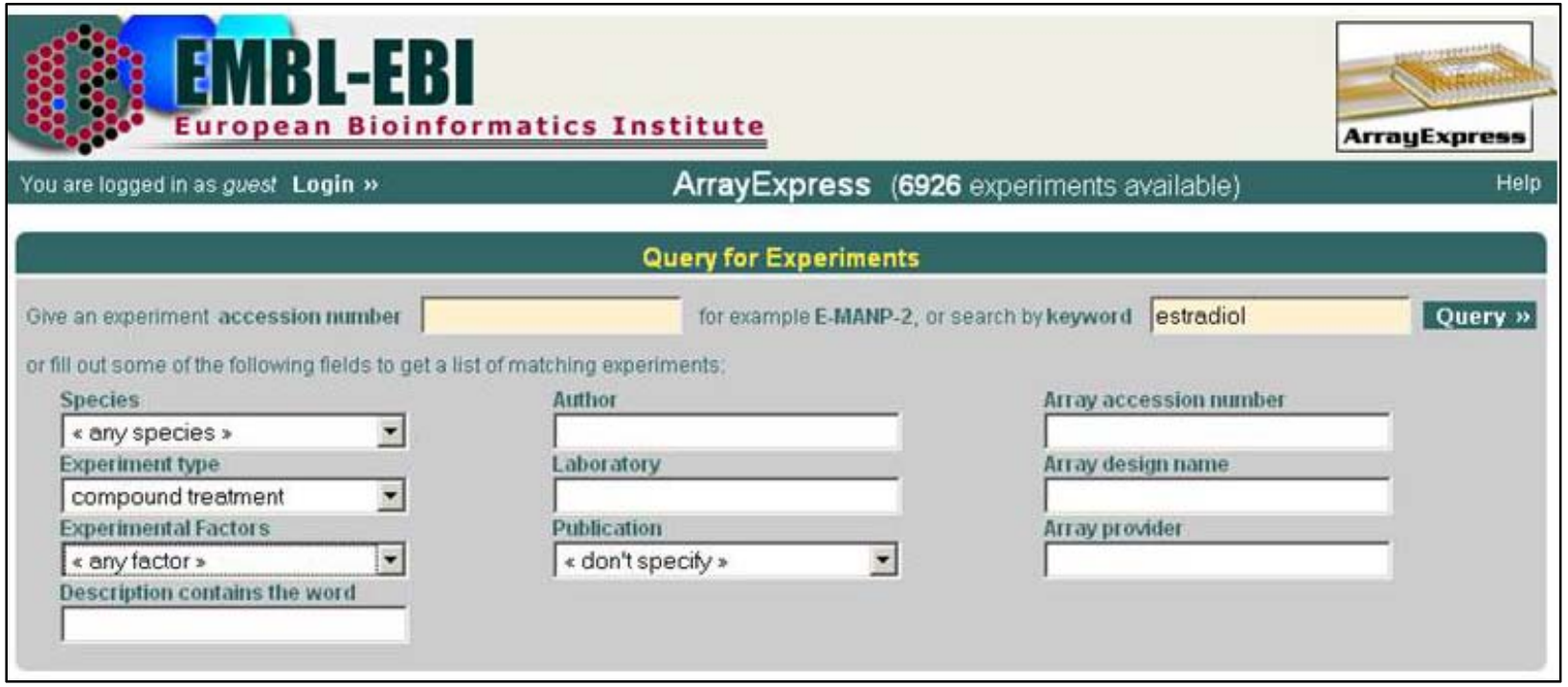


Figure 3-1. Advanced query interface for ArrayExpress showing an example of a keyword search for "estradiol" coupled with <Experiment type> = "compound treatment", and <Experimental Factors> = "dose" with the latter categories applied to the experiment by the data submitter (site accessed on 03 October 2008).

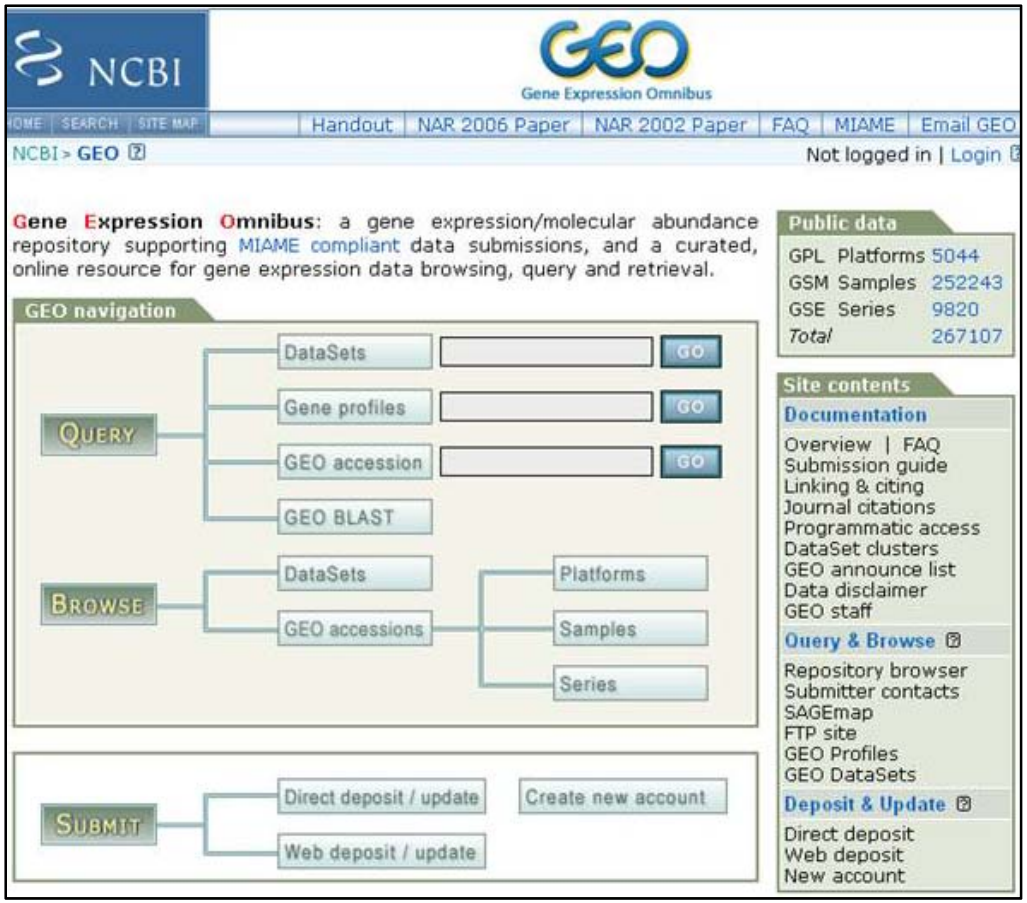


Figure 3-2. GEO home navigation screen showing site contents and options for browsing and querying the data repository (site accessed on 03 October 2008).

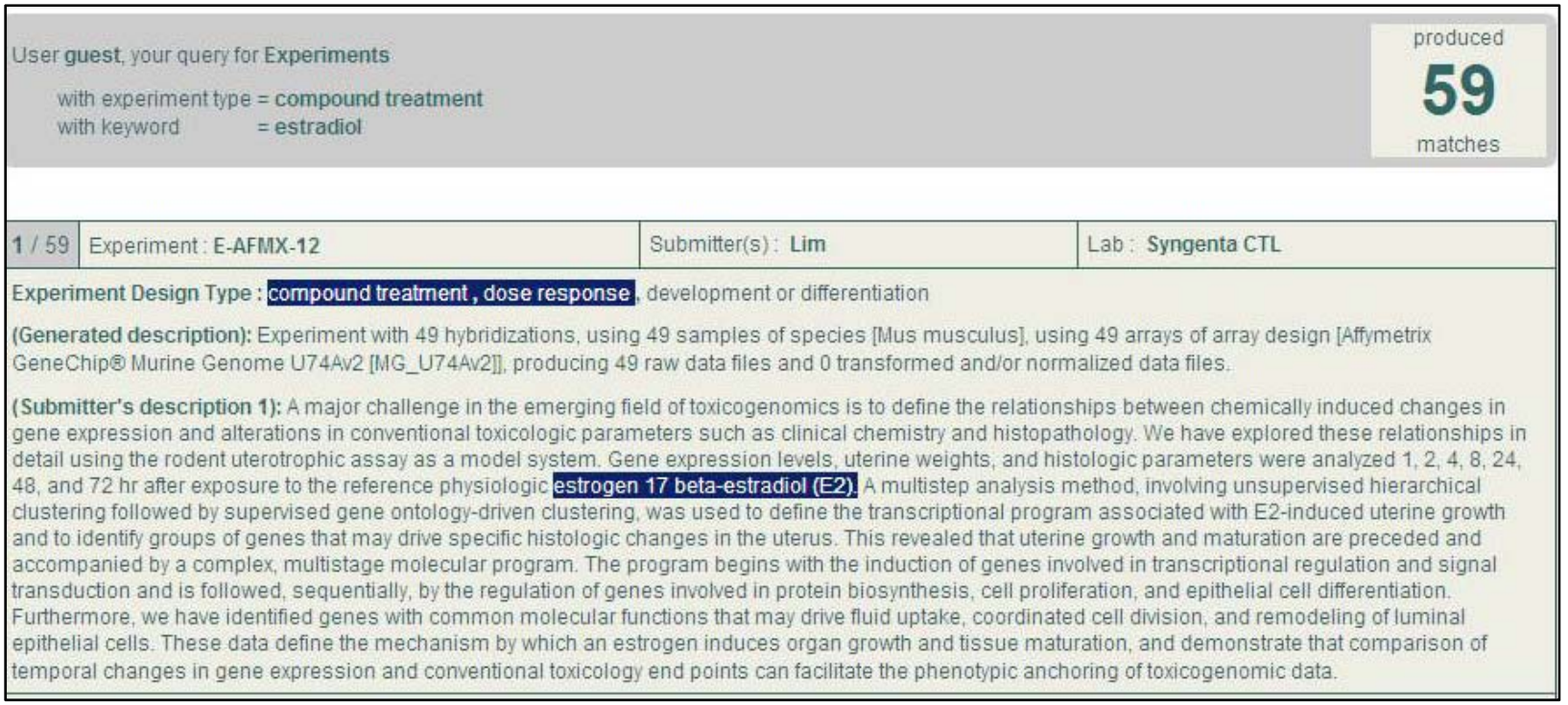
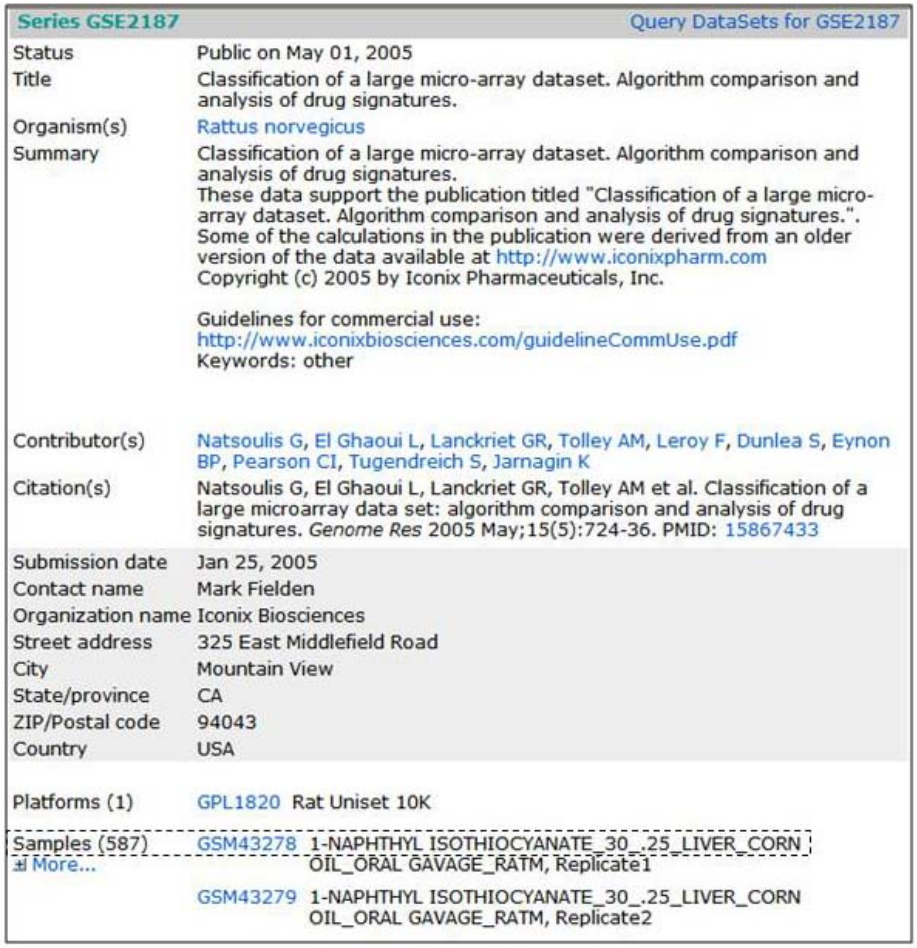


Figure 3-3. The first of 59 hits returned based on the ArrayExpress query shown in Figure 1, i.e., a keyword search for "estradiol" coupled with <Experiment type> = "compound treatment", showing the chemical name embedded in the Submitter's description (site accessed on 03 October 2008).



Series GSE2187	Query DataSets for GSE2187
Status	Public on May 01, 2005
Title	Classification of a large micro-array dataset. Algorithm comparison and analysis of drug signatures.
Organism(s)	Rattus norvegicus
Summary	Classification of a large micro-array dataset. Algorithm comparison and analysis of drug signatures. These data support the publication titled "Classification of a large micro-array dataset. Algorithm comparison and analysis of drug signatures.". Some of the calculations in the publication were derived from an older version of the data available at http://www.iconixpharm.com Copyright (c) 2005 by Iconix Pharmaceuticals, Inc. Guidelines for commercial use: http://www.iconixbiosciences.com/guidelineCommUse.pdf Keywords: other
Contributor(s)	Natsoulis G, El Ghaoui L, Lanckriet GR, Tolley AM, Leroy F, Dunlea S, Eynon BP, Pearson CI, Tugendreich S, Jarnagin K
Citation(s)	Natsoulis G, El Ghaoui L, Lanckriet GR, Tolley AM et al. Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res* 2005 May;15(5):724-36. PMID: 15867433
Submission date	Jan 25, 2005
Contact name	Mark Fielden
Organization name	Iconix Biosciences
Street address	325 East Middlefield Road
City	Mountain View
State/province	CA
ZIP/Postal code	94043
Country	USA
Platforms (1)	GPL1820 Rat Uniset 10K
Samples (587)	GSM43278 1-NAPHTHYL ISOTHIOCYANATE_30_.25_LIVER_CORN OIL_ORAL GAVAGE_RATM, Replicate1
More...	GSM43279 1-NAPHTHYL ISOTHIOCYANATE_30_.25_LIVER_CORN OIL_ORAL GAVAGE_RATM, Replicate2

Figure 3-4. A sample GEO DataSets (GDS) record showing the various field categories and field entries for the experimental series, GSE2187, with a chemical name shown in the "Samples" field (site accessed on 03 October 2008).

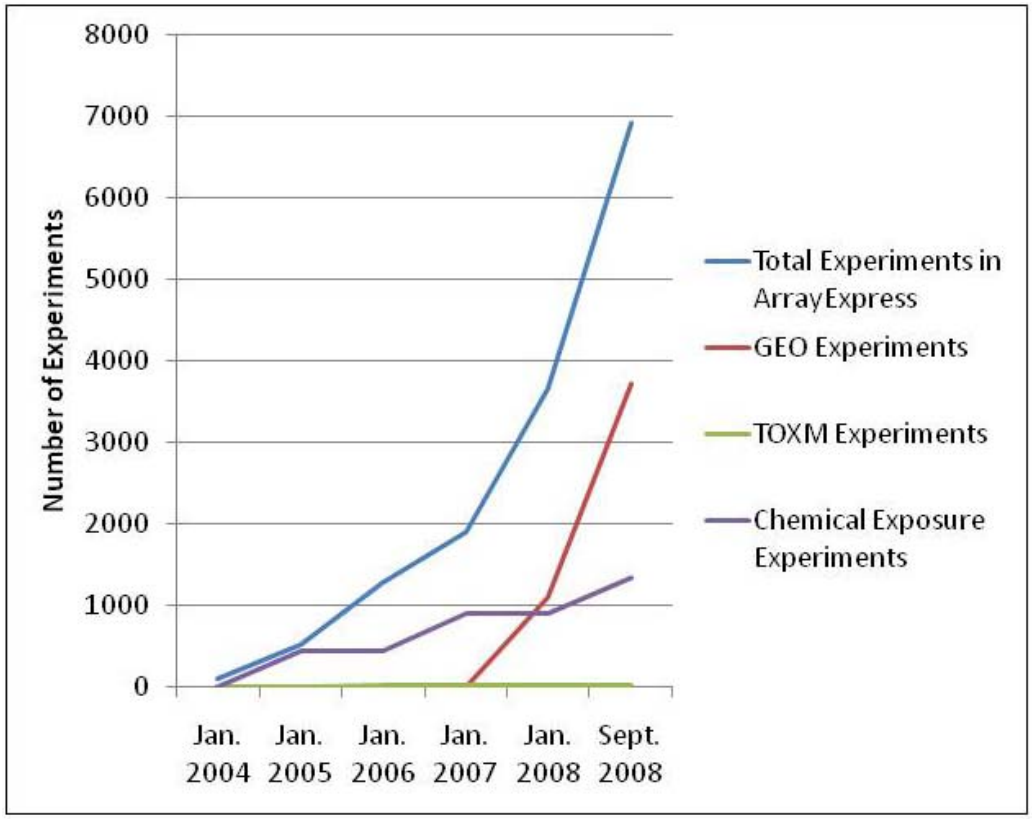


Figure 3-5. Growth of genomics content in ArrayExpress Repository since its inception in 2003, with the formal TOXM-designated toxicogenomic content growing at negligible rate, but the chemical exposure-related content identified though this project significantly larger.

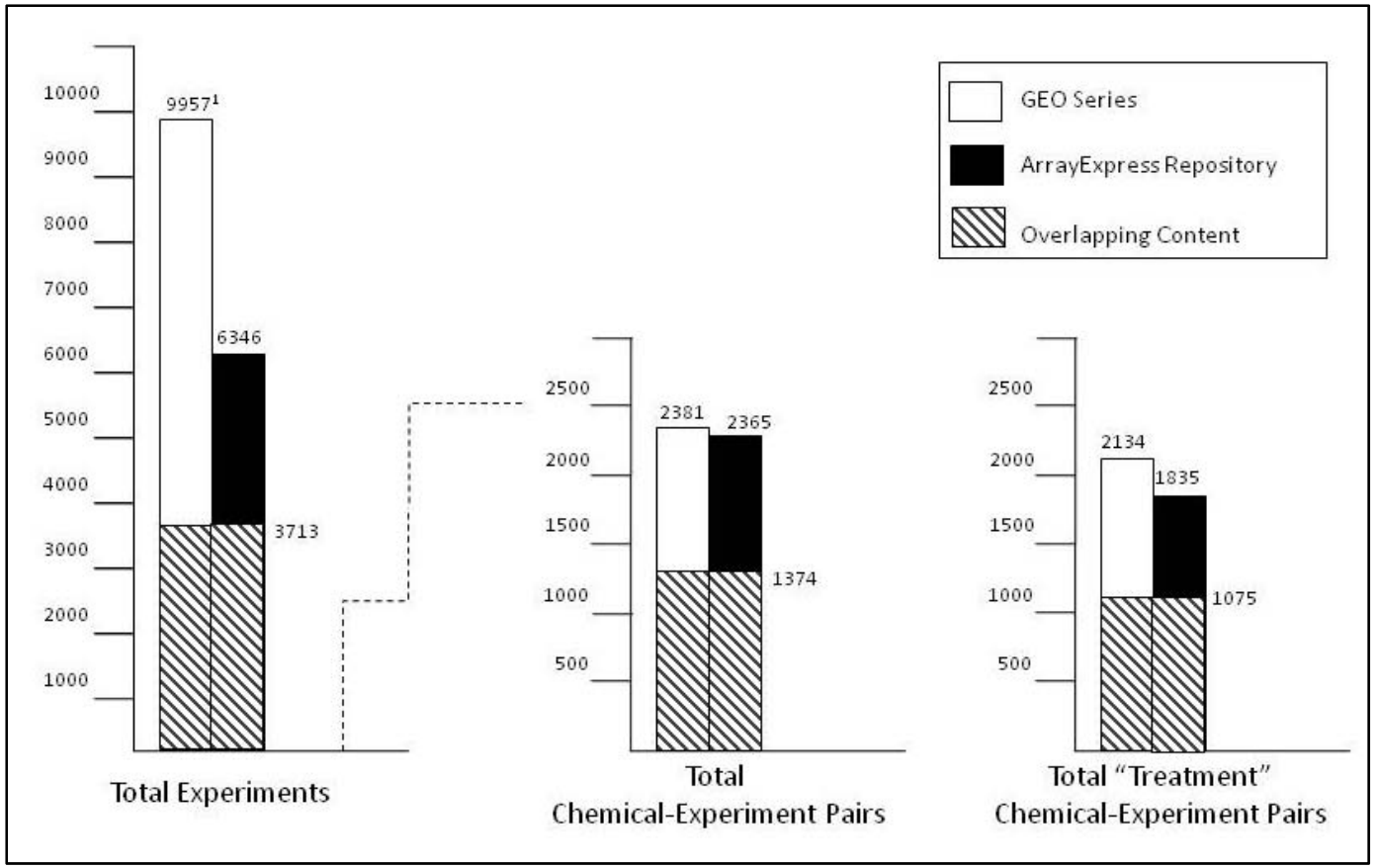


Figure 3-6. Comparison of numbers of GEO Series and ArrayExpress Repository experiments, chemical-experiment pairs, and "Treatment" chemical-experiment pairs, also showing overlapping content between the two systems; refer to totals and legends in Tables 3-1 and 3-2 (current as of 20 September 2008).

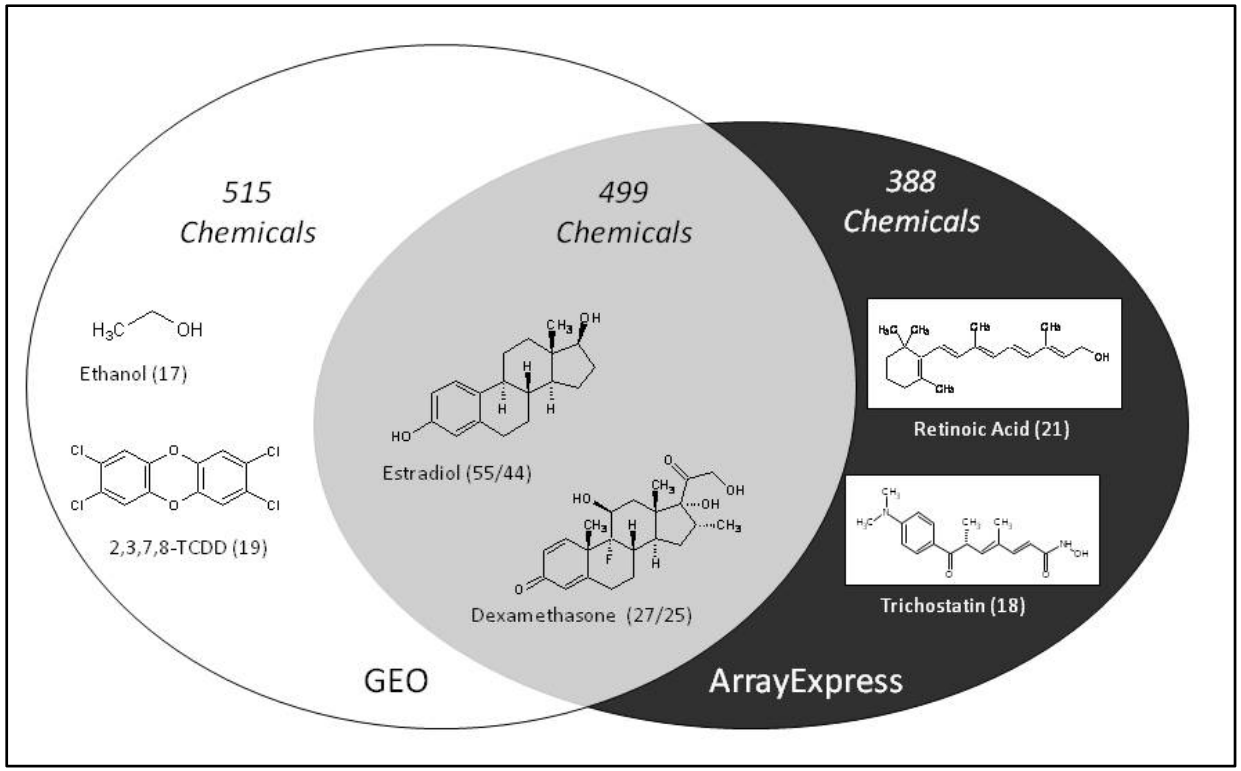


Figure 3-7. Comparison of the total sets of unique chemicals pertaining to Treatment Chemical-Experiment pairs in ArrayExpress Repository and GEO Series from the DSSTox GEOGDS and ARYEXP data files; shown in each section are the chemicals mapping to the largest number of "Treatment" Chemical-Experiments in each case , with the number of experiments shown in parentheses (GEO/AE) (current as of 20 September 2008).

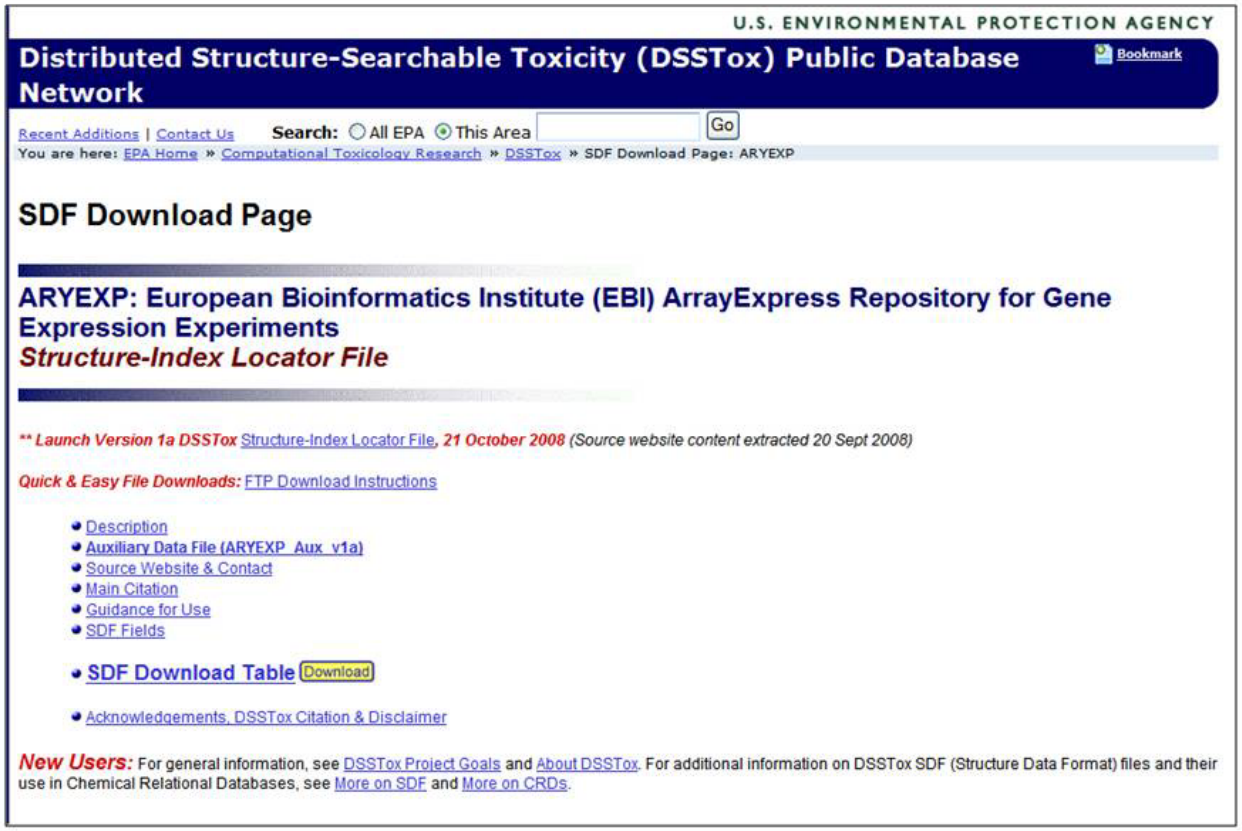


Figure 3-8. Screen shot of DSSTox ARYEXP Download Page (http://www.epa.gov/dsstox/ sdf_aryexp.html) showing links to more information about ARYEXP and the SDF download table (site accessed on November 7, 2008).

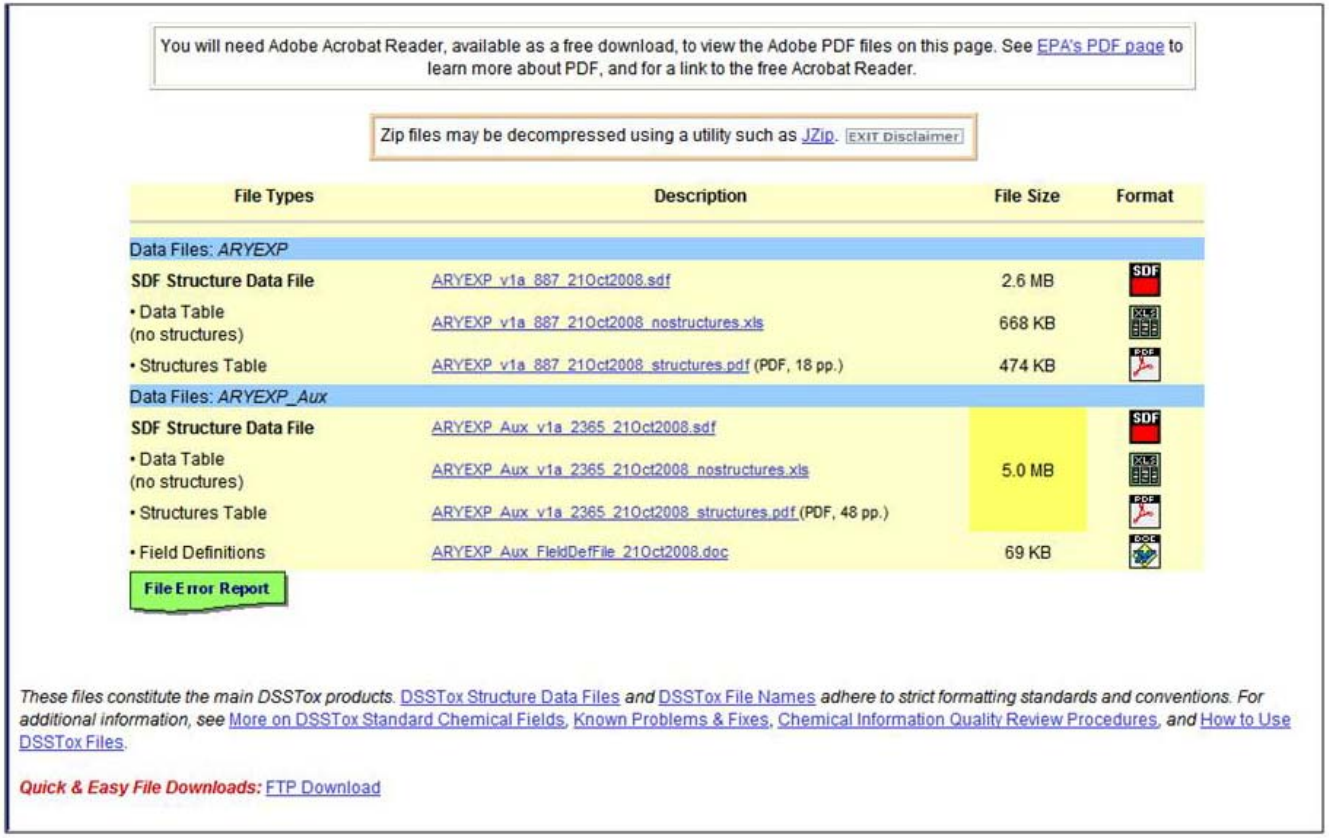


Figure 3-9. Screen shot of DSSTox ARYEXP Download Table (http://www.epa.gov/dsstox/ sdf_geogse.html) showing the available files and formats. ARYEXP and ARYEXP_Aux are available as a SDF structure file, Microsoft Excel Data Table, and a PDF Structure Table. The field definition file explaining each of the fields in the Auxiliary file is available as a Microsoft Word document (site accessed on November 7, 2008).

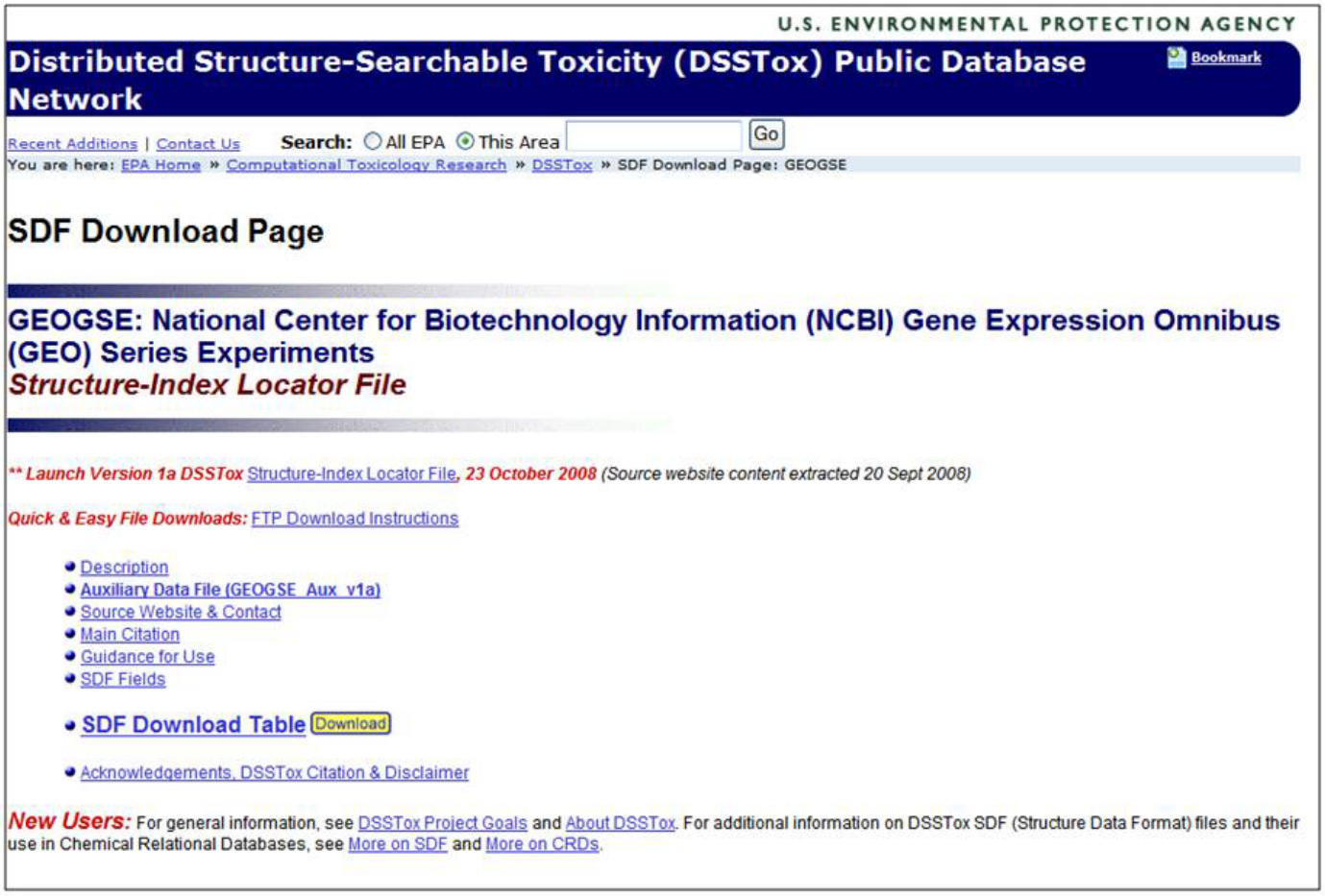


Figure 3-10. Screen shot of DSSTox GEOGSE Download Page (http://www.epa.gov/dsstox/sdf_geogse.html) showing links to more information about GEOGSE and the SDF download table (site accessed on November 7, 2008).

Figure 3-11. Screen shot of DSSTox GEOGSE Download Table (http://www.epa.gov/ dsstox/sdf_geogse.html) showing the available files and formats. GEOGSE and GEOGSE_Aux are available as a SDF structure file, Microsoft Excel Data Table, and a PDF Structure Table. The field definition file explaining each of the fields in the Auxiliary file is available as a Microsoft Word document (site accessed on November 7, 2008).

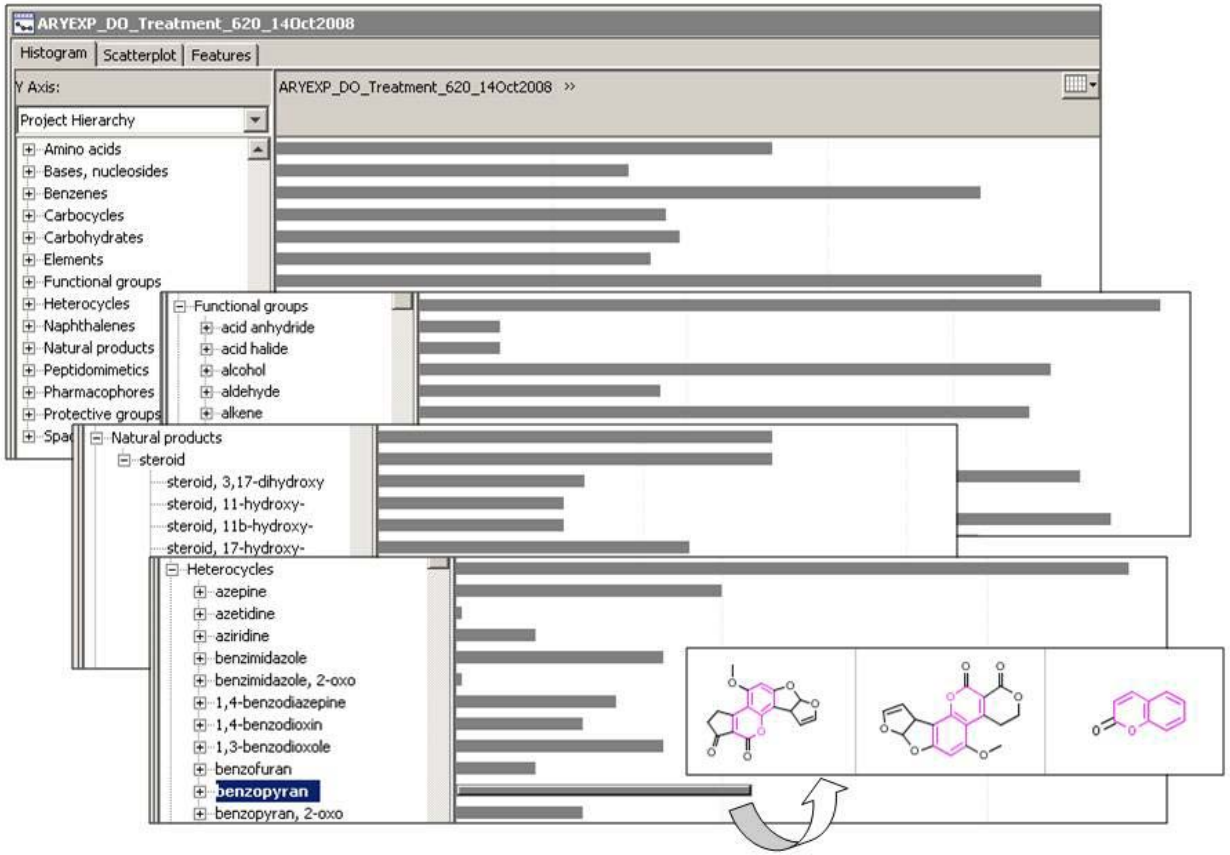


Figure 3-12. Mapping of DSSTox ARYEXP structure file (unique, defined organics) to Leadscope (Leadscope, Inc, 2008) chemical hierarchy, showing incidence breakdowns by Functional group, Natural products – steroid class, and Heterocycles, showing 3/13 compounds in the benzopyran class (length of bars are proportional to number of structures in class).

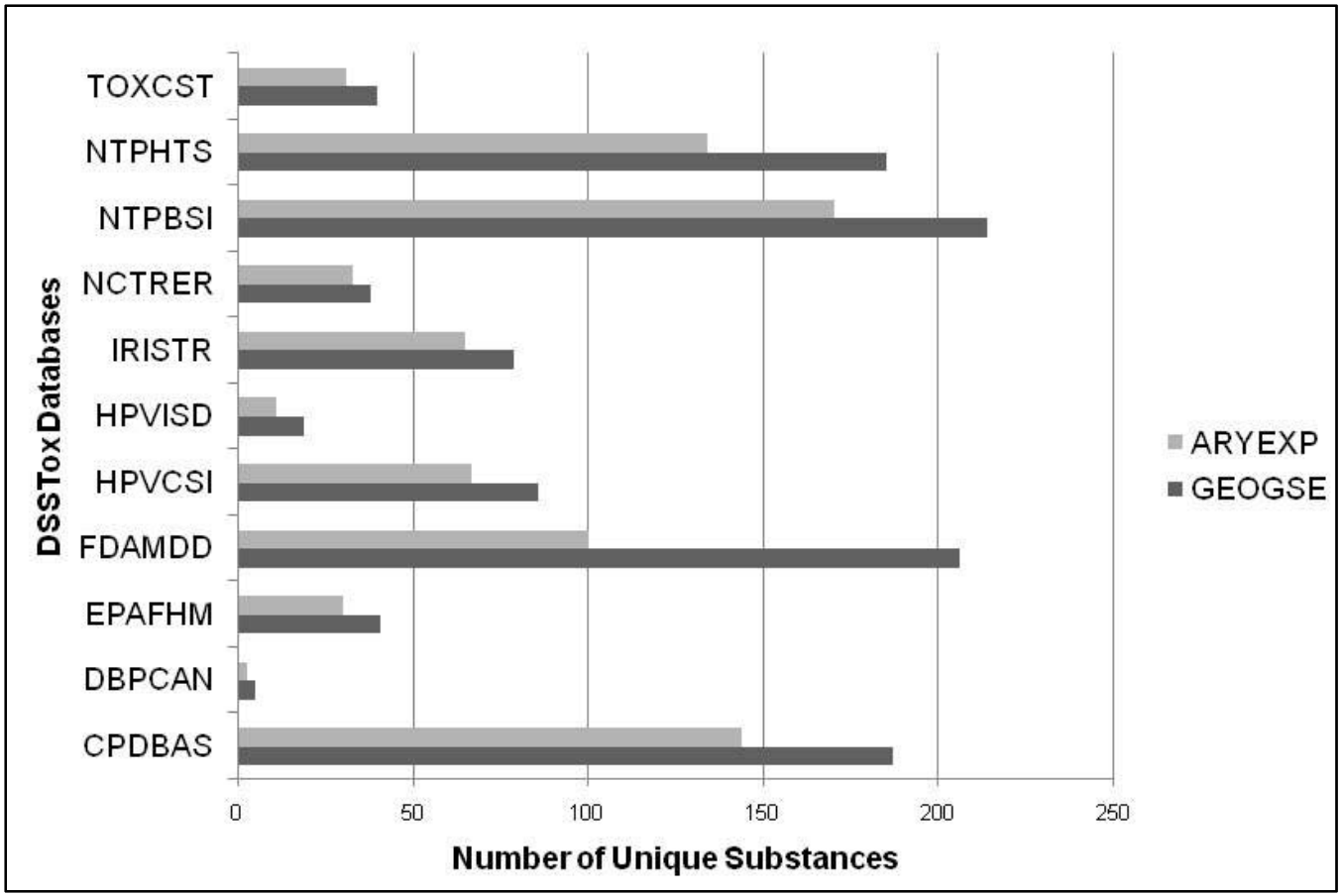


Figure 3-13. ARYEXP and GEOGSE Unique Substances Overlap with 11 previously published DSSTox databases shown as a metric of toxicological relevance of ARYEXP and GEOGSE chemical substances.

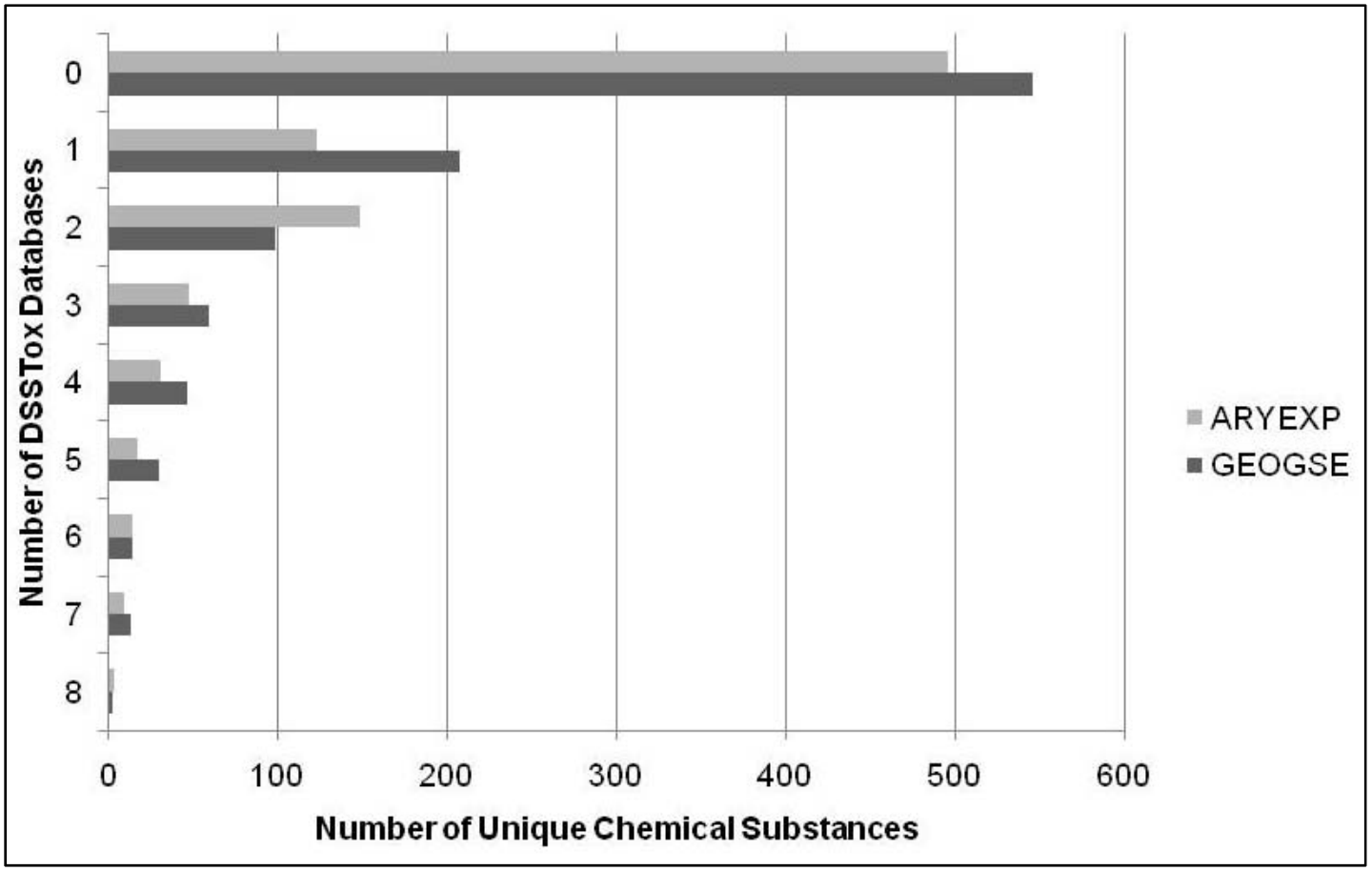


Figure 3-14. ARYEXP and GEOGSE Unique Substances Overlap with Multiple DSSTox databases shown here as a further metric of the toxicological relevance of ARYEXP and GEOGSE. There are 261 chemicals in GEOGSE and 194 in ARYEXP that overlap 2 or more (2,3,4,5,6,7,or 8) DSSTox databases.

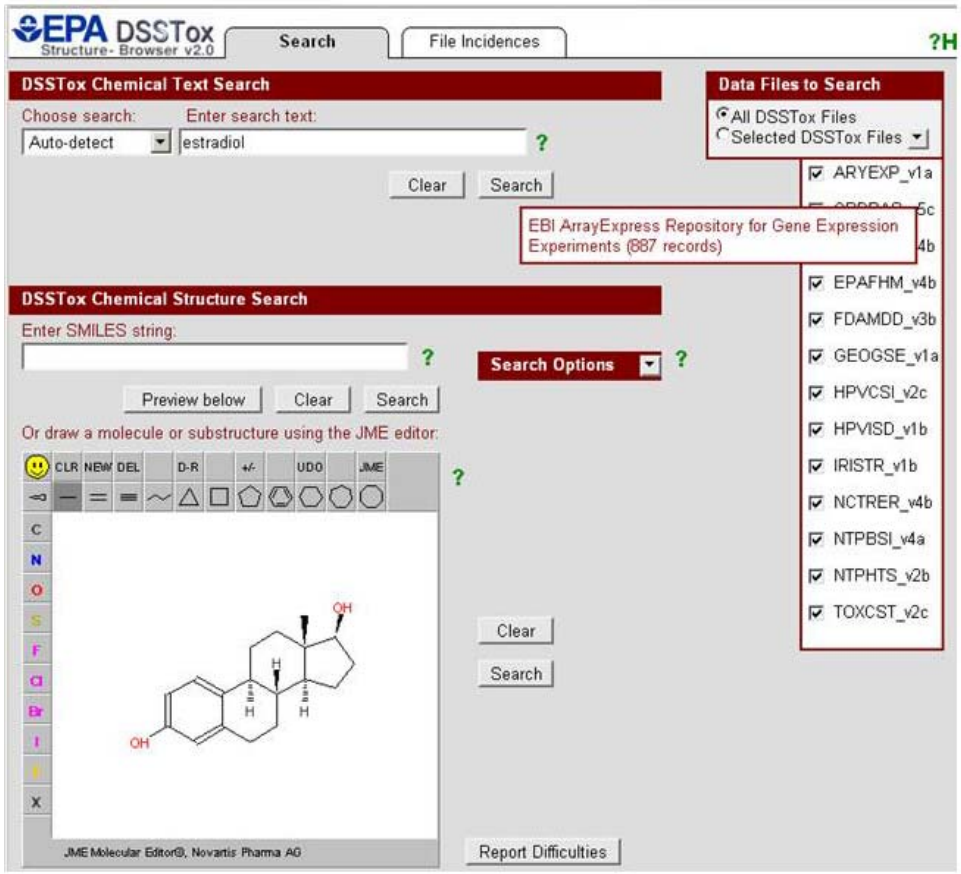


Figure 3-15. Screen shot of DSSTox Structure-Browser (http://www.epa.gov/dsstox_structurebrowser/) showing drawn text entry and drawn structure for "estradiol", either of which could be submitted for search across all DSSTox Data Files, including ARYEXP and GEOGSE (site accessed on November 7, 2008).

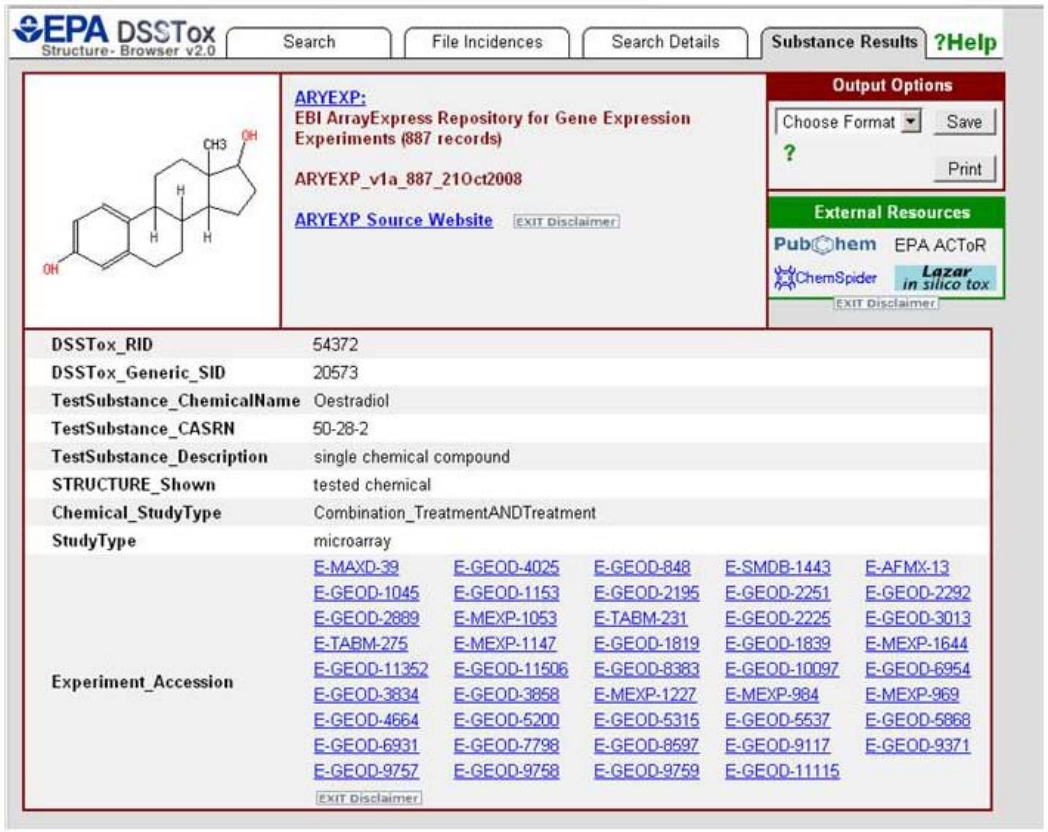


Figure 3-16. Screen shot of DSSTox Structure-Browser (http://www.epa.gov/dsstox_structurebrowser/) showing Substance Results for "exact match" to structure search for estradiol within the ARYEXP data file, in which a large number of ArrayExpress experiment Accession ID URLs are listed, each linked to the corresponding experiment summary page within ArrayExpress; page also shows link-outs to External Resources for this substance (site accessed on November 7, 2008).

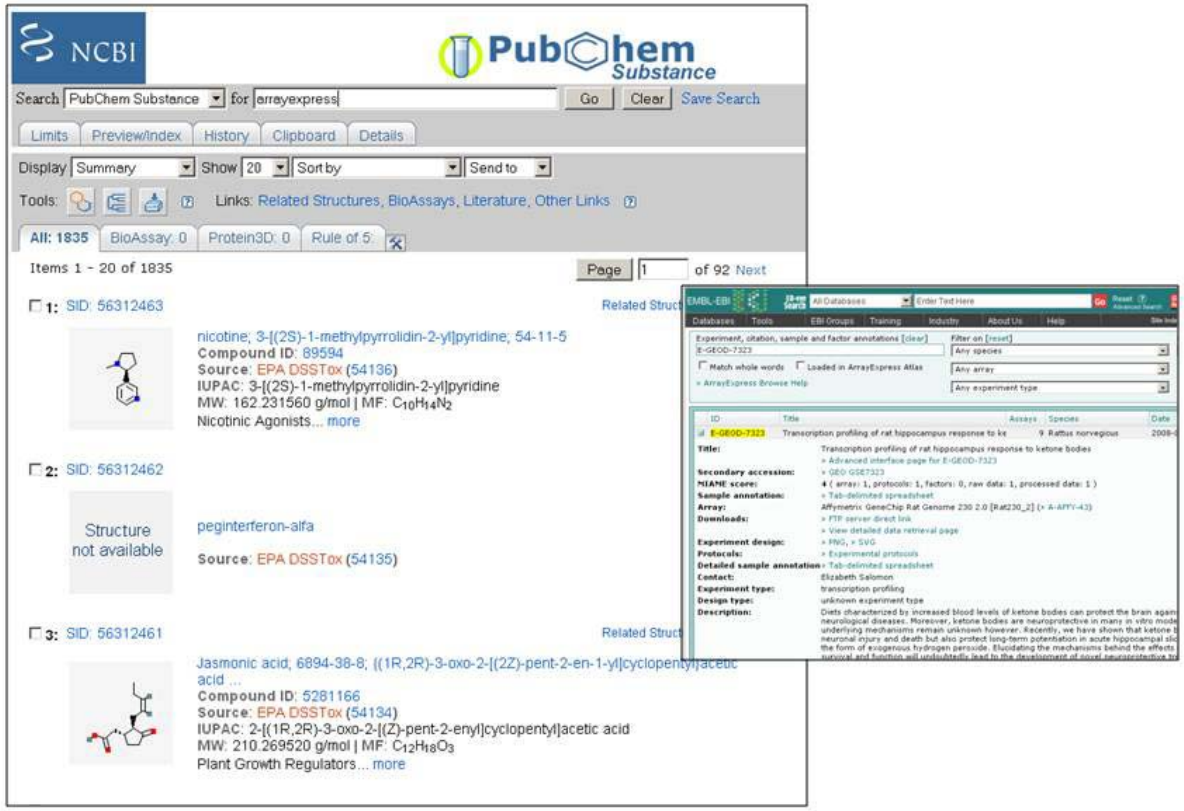


Figure 3-17. Screen shot of PubChem substance listing for ARYEXP, indicating 1835 substances retrieved with keyword search for "arrayexpress" from main PubChem search page, with each substance ID linked to a corresponding experiment summary page within ArrayExpress directly from PubChem (site accessed on 07 November 2008).

CHAPTER 4

CONSIDERATIONS FOR THE ANALYSIS AND USE OF PUBLIC GENE EXPRESSION-MICROARRAY DATA

ABSTRACT

The use and analysis of microarray data is without a clear consensus. It is for this reason that a large effort has been put forth by various agencies and consortiums to come to a consensus. It is without question that far less is known about the use of public microarray data because only within the last few years has the amount of public data grown to a level adequate for further analysis. There are still questions of the comparability of microarray data across platforms and laboratories. In this chapter the use of a method applied to other microarray data projects will be used to attempt to better determine the proper use and analysis of public microarray data. Specifically the Percent of Overlapping Genes will be used to measure the consistency of differentially expressed gene list within a laboratory and across laboratories. This exercise demonstrates a first step in understanding the statistical decision making necessary for the use of public microarray data. Research Questions: Is the consistency of the DEG list affected by the choice of normalization packages, Robust Multichip Average (RMA) or Guanine Cytosine Robust Multi-Array Analysis (GCRMA) methods?

- Is there a high rate of consistency of DEG lists within and across laboratories without any multiple testing correction?
- Does the selection of error control, FDR or FWER, affect the consistency of the DEG list among genes within and across laboratories?
- Does the alpha (α) level (e.g. 0.01, 0.05, 0.10, and 0.25) affect the consistency of DEG lists within and across laboratories?
- Are internal Affymetrix control genes consistent within and across laboratories?

The result of this study is that in most situations that an increase in power does not lead to a loss of consistency.

INTRODUCTION

As more public gene expression data are made available, the ability to understand and properly analyze these data poses increasing challenges. As seen in previous chapters, gene expression data are a primary component of toxicogenomics or toxico-chemogenomics data sets. Methods used for

locating public chemical-related gene expression data sets have been described in detail. An analysis of the chemical content, both individually and in reference to public chemically indexed resources, has also been shown. It is in this chapter that attention is shifted from identifying, annotating and analyzing the chemical content of public gene expression repositories to considering how to use the newly annotated and identified data.

There are partial solutions proposed to enable researchers to look across multiple domains of data in a systematic fashion, i.e. the ability to anchor microarray data to toxicological results (Paananen and Wong. 2006). However, a complete solution has yet to be proposed. For this reason, in commercial enterprises, such as drug discovery in the pharmaceutical industry, great controls have been instituted to maintain strict protocols on limited platforms, species, and laboratory environment. This luxury is not afforded to users of public data. The unique quality of each public data set is that it is the product of a study designed to address a specific hypothesis or set of hypotheses, but it may have broader applicability to other hypotheses with the addition of new methodologies or information. The ability to combine microarray data across platform, laboratory and species is the topic of much debate (Tan et al. 2003; Shi et al. 2006, Chen et al. 2007). In general, there is little clear consensus about how individual microarray experiments should be conducted or how the resulting data should be analyzed. Several organizations and consortiums have been formed to address these issues, including: the Microarray Gene Expression Data Society (MGED)-Minimum Information About a Microarray Experiment (MIAME), the External RNA Controls Consortium (ERCC)-Universal RNA Standards, National Institutes of Standards and Technology (NIST) metrology, and the Microarray Quality Control (MAQC) Project, the last to not only consider standards for analyzing data, but standards for the entire microarray experiment and how microarray data should be submitted concordant to journal publications. The MAQC efforts can be directed towards future standards adoption, but are not

directly applicable to existing public microarray data. When using public microarray data, the opportunity to conduct the experiment under strict protocols favorable to the secondary user's needs does not exist and such users cannot control for non-experimental factors in the laboratory setting. Most of the proposed standardization efforts include data analysis on the same RNA sample or some spike-in controls (Tan et al. 2003).

The focus of this analysis project is to begin to understand the statistical decision-making that occurs in the integration of multiple public gene expression experiments to determine if various methods set forth for the comparison of standardized gene expression data are appropriate for non-standardized public gene expression data. Most recent theories about cross-platform and cross-laboratory microarray data analysis take the whole experiment into account and attempt to solve the problem several steps at a time (Draghici et al. 2006, Quackenbush 2002; Shi et al 2006; Irizarry et al. 2005; Tan et al. 2003). With the exception of evaluations of the normalization methods, researchers seldom take into account the statistical decision-making and processing at each step that occurs when integrating multiple gene expression experiments. Here I attempt to begin to dissect the analysis process one step at a time with the careful selection of public microarray data from ArrayExpress and the extension of existing methodologies to public data. Given its particular applicability to evaluating the reproducibility of microarray experiments, the Percent of Overlapping Genes (POG) method is used in the present study to explore public microarray data (Shi et al 2008).

POG

The POG method for evaluating public genomic data derives from recent work on the reproducibility of microarray experiments set forth within the MAQC project, where reproducibility is used as a metric for determining the usability of microarray data. The ability to reproduce scientific results is paramount in *good* science. Therefore, if microarray experiments cannot be reproduced, the resulting

data are not reliable. Many factors arise that may affect the reproducibility of microarray experiments, including: laboratory bias, species extrapolation, experimental variation, and analysis methods. Although data reproducibility evaluates one's ability to *recreate* experimental results, the methods used to evaluate reproducibility also provide a measure of comparability across data generated in similar experiments. Consequently methods for data reproducibility are used in this analysis for the comparison of several public gene expression datasets to begin to dissect the decision-making process and the impact of statistical processing.

The MAQC project sets forth the use of the percentage of overlapping genes (POG) in a differentially expressed gene (DEG) list as a metric for the measurement of reproducibility of microarray experiments. DEG lists were determined using fold change (FC) as ranking criterion, in addition to a non-stringent P cutoff. According to Shi et al. (2008), the best way to select FC and p-value cutoff is to identify genes as a function of both the nature of the data and the tradeoff between sensitivity and specificity. There is not one *correct* answer; therefore, further evaluation is necessary in terms of gene function, and biological pathways (Guo et al. 2006). The case is made that the use of P-value without the consideration of FC makes the tradeoff of reproducibility, sensitivity, and specificity more pronounced (Shi et al. 2008). However, there is still much debate about the suitability of the POG MAQC method as an adequate solution to this problem (Chen et al. 2007). Researchers continue to use methods that rely on P-value alone in the determination of a DEG list. In this chapter the use of an overlapping gene list will be extended to measure consistency among microarray methods. Here, consistency is defined as the percent of overlapping genes between two treatments from two different sources.

MICROARRAY HYPOTHESIS TESTING

A primary goal of a toxicogenomic microarray experiment is to identify which genes behave differentially under different conditions and to assess confidence in the results. In general there are two statistical approaches for the analysis of differentially expressed genes: classical hypothesis testing; and Bayesian hypothesis testing. These approaches differ in the assumptions made about the parameters, such as population mean (Dudoit et al 2002; Ge et al. 2003; Efron et al. 2001). There are advantages and disadvantages to both methods; however classical hypothesis testing was used in prior studies that are employed in the design of this study (Shi et al. 2008).

In classical hypothesis testing there is a null hypothesis and an alternative hypothesis. Specifically, in gene expression, the null hypothesis is that gene *x* is not differentially expressed across the two conditions. The result of a hypothesis test is an absurdity probability called a p-value that relates very closely to the false positive rate (FPR) to be discussed later (Wit and McClure 2004). The lower the probability value, the stronger the evidence is against the null hypothesis. If the p-value is less than a certain threshold or alpha (α) level, then the null hypothesis is rejected in favor of the alternative hypothesis, under which gene *x* is differentially expressed.

Multiple testing issues arise because in microarray experiments there is one test for each gene on an array. The larger the number of hypotheses, the more likely the researcher is to find extreme differential expression scores or test statistics leading to a low p-value, even if all the null hypotheses are true. If each hypothesis is rejected at some fixed posterior probability or fixed p-value and the number of hypotheses grows, then it becomes more and more likely that at least one null hypothesis will be falsely rejected. In the context of microarray experiments, it means that we would say that a gene is differentially expressed when it is truly not. This is a problem if the user is interested in controlling the overall error rate.

There are four components of classical hypothesis testing: the hypotheses (discussed above), the test statistic, the error rate, and the decision rule used to control the error rate. The test statistic provides a summary of the data used to evaluate the truth of the hypotheses. One of the most commonly used test statistic is the t-statistic used for comparing two sample tests, which is the difference between averages divided by the standard deviation.

For each pair of hypotheses, there are two types of error, a false positive or a false negative. A false positive (FP) is where the null hypothesis is wrongly rejected and a false negative (FN) is when the null hypothesis is wrongly accepted. These error rates are dependent on one another. The individual false positive rate is controlled by the α significance level. By reducing the significance rate allowing fewer false positives, the power of the test is also reduced. Power is equal to 1 minus the false negative rate (FNR). Power is further defined as the probability that the test will reject a false null hypothesis. Traditional methods focus on controlling the proportion of expected false positives or the false positive rate (FPR),

$$FPR = E(\frac{F_P}{m_0}),$$

where F_P is the number of false positives and m_0 is the total number of true null hypotheses(Table 4-1). It should be noted here that the true false positive and false negative rates are unknown; therefore, they must be estimated. The familywise error rate (FWER) is the probability that one null hypothesis is falsely rejected among all other hypotheses, or the probability that among all genes that are not differentially expressed, at least one is incorrectly labeled as differentially expressed:

$$FWER = P(F_p > 0).,$$

where F_P is the number of false positives. The FWER is often controlled for in experiments, but this can be a conservative approach with extremely large numbers of hypotheses, as is the case with microarray experiments. Table 4-1 shows a description of false positives and negatives in the context of differentially expressed genes.

Strong FWER control of the Type I error rate under any combination of true and false hypotheses. Weak FWER control is control of the Type I error rate only when all the null hypotheses are true. As an alternative to strong FWER controls, there are weak FWER control methods such as the Benjamini and Hochberg or the Benjamini and Yekutieli false discovery rate (FDR) methods, i.e., the expected number of non-differentially expressed genes among the declared differentially expressed genes, or the number of falsly rejected hypothesis tests among all rejected hypothesis tests,

$$FDR = E\left(\frac{F_P}{S}\right),$$

where S is the total number of rejected hypothesesand F_P is the number of false positives. In cases where all null hypothesis tests are true, the FWER is approximately equivalent to the FDR. FDR controls are more suited for exploratory purposes and FWER controls are more suited for confirmatory purposes (Wit and McClure, 2004). FDR control methods are considered advantageous in certain situations because of the increase in power over FWER control methods, the expected proportion of truly differentially expressed genes that are correctly identified as being differentially expressed.

Lastly, the decision rule is a method to translate observed test statistics into a set of binary decisions, such that the error rate of choice is controlled at a preset level. The p-value controls the FPR by comparing it to a pre-specified significance level α for a single hypothesis test, the probability of a false positive less than α. The alpha (α) level can be set at any number; however, it must be chosen

before the experiment is conducted. The p-value translates the value of the test statistic into a probability that expresses the absurdity of the null hypothesis (Wit and McClure 2004). P-values close to 0 indicate that the null hypothesis is absurd and should be rejected in favor of the alternative hypothesis. A p-value less than α for a pair of hypotheses leads to the rejection of the null hypothesis. In the context of gene expression, this means that a gene should be declared significantly expressed. The p-value translates to how likely the observed data would be if the null hypothesis were actually true.

In the present analysis, an extension of percent of overlapping gene (POG) is used to evaluate the consistency of differentially expressed genes (DEG). For this purpose, the following research questions will be addressed:

- Is the consistency of the DEG list affected by the choice of normalization packages, Robust Multichip Average (RMA) or Guanine Cytosine Robust Multi-Array Analysis (GCRMA) methods?
- Is there a high rate of consistency of DEG lists within and across laboratories without any multiple testing correction?
- Does the selection of error control, FDR or FWER, affect the consistency of the DEG list among genes within and across laboratories?
- Does the alpha (α) level (e.g. 0.01, 0.05, 0.10, and 0.25) affect the consistency of DEG lists within and across laboratories?
- Are internal Affymetrix control genes consistent within and across laboratories?

In this chapter, a study is presented to address each of these research questions.

METHODS

DATA SELECTION

At the time of the selection of microarray data, only the ArrayExpress Repository had been fully chemically indexed and annotated. A small set of prospective datasets were identified (Figure 4-1). From 2365 chemical-experiment pairs (See Chapter 3), a dataset consisting of chemical treatment experiments involving the chemical "hydroxyurea" was the most complete and appropriate for this study. An important criterion used in the data selection was that the chemical have toxicogenomic relevance. As previously discussed, within ArrayExpress this distinction can be established made with the use of the TOXM accession codes (http://www.ebi.ac.uk/microarray-as/aer/entry). There are other toxicogenomic relevant experiments in the larger database; however, for this exercise I was able to identify a perspective dataset within the TOXM listing of experiments. The most difficult obstacle to overcome in constructing an analysis dataset from these public data was identifying datasets with sufficiently common cell types, species, platform, and endpoints.

The resulting datasets, E-TOXM-6 and E-TOXM-17, are from two genotoxicity experiments that span three laboratories/performers: Sanofi Aventis, Procter & Gamble, and Astra-Zeneca over seven different doses, and three different time points (www.ebi.ac.uk/microarry-as/ae/browse.html?keywords=E-TOXM-17, www.ebi.ac.uk/microarry-as/ae/browse.html?keywords=E-TOXM-16). The original purpose of these experiments was to determine if laboratory variability precluded the identification of biological functions impacted by hydroxyurea. The results were that the laboratory variability did not preclude the identification of the biological function. (Müller at al. 2005).

When selecting the specific arrays to be used, there were several considerations of sufficiency for inclusion: 1) at least one biological replicate, 2) an adequate control with at least one replicate at that

time period, and 3) a comparable array with the same time-dose treatment level from a different laboratory or the same laboratory at a different time . The Aventis performer submitted two datasets for this project that resulted from experiments performed at two different times periods, in years 2001 and 2003, labeled Aventis17 and Aventis 6, respectively. Data for Astra-Zeneca were not included in this analysis because Astra-Zeneca submitted arrays that lacked sufficiency, as discussed above. Additional P&G arrays for two time periods, 4 hours and 24 hours (after a twenty hour recovery period), and 4 doses, low, low-mid, mid-high, and high, were sufficient for inclusion, creating 8 within-lab comparisons, 8 across-lab comparisons, and 4 within- and across-lab comparisons /contrasts (Figure 4-2). An exception was made for the P & G 4hr control so that multiple lab comparisons could be made.

DATA PROCESSING PLAN

The data processing plan shown in figure 4-3a and b indicates the steps for general processing of each microarray experiment. Image processing is the first general step in data processing in a microarray experiment; however, in this data processing plan, I will begin with data normalization for the Affymetrix cel file.

NORMALIZATION

Normalization is the next step after image processing in removing systematic errors or bias as a result of experimentation. There are methods in which normalization is computed along with the actual identification of differential expressed genes. One such method is an analysis of variance method developed by R.A Fisher in 1923 and adapted to microarrays by (Kerr and Churchill 2001). The major disadvantage of this method, among others, is that data must be normalized again for another type of exploration or analysis of the data. In this discussion, a brief overview of normalization

methods are given with special attention shown for methods recommended in Wit and McClure (2004).

The transformation of data is a logical first step. Gene expression values represent the amount of observed reflection of the dye molecule attached to the mRNA. The amount of reflection is thought to be partially linear. There is most often non-linearity in areas of relatively high intensities and low intensities and where there are dye-effects. Dye effects are due to the different physical properties of dyes used in two-color or two-channel arrays, mainly Cy3 and Cy5. Scaling gene expression data can sometimes help to overcome these problems. There are four scaling perspectives to consider: rank scale, original scale, scale corrected only at areas of low and high intensities, and the logarithmic scale. In general the original scale usually produces linearity of signal intensities with exception of areas of low and high intensities. Other scaling or transformation methods have been developed (Irizarry et al. 2003a,b; Huber et al. 2002;Rocke and Durbin 2001); however, logarithmic transformation is perhaps the best solution because it retains the original signal while transforming areas of non-linearity. It has the added advantage of expressing fold-change differences in gene expression as an absolute difference, e.g., a two-fold change increase and decrease show the same magnitude of change. For example, on the logarithmic scale, a 2-fold increase or decrease is +/- 0.7 and on the original scale it is (+) 50 or (-) 25. Most commercial arrays use housekeeping genes that, in addition to internal purposes for scanning, can be used as control genes in normalization because for these housekeeping genes, also known as endogenous spike-in genes, the concentration and location of the genes is often known. Arrays may also have exogenous mRNA, from species different than that of the experiment species, that is expected to show no changes in gene expression, but that may have limited usefulness because of cross hybridization.

Normalization methods are dependent on the set of genes that the researcher deems most appropriate. There are methods that are based on housekeeping genes where no change in expression level is expected, or based on exogenous spike-in controls where there should be little to no hybridization (Lee et al. 2002, Thellin et al, 1999). As discussed previously, these methods are not dependent on expression levels. It is assumed that there are no changes in the expression of housekeeping genes and that there are no cross-hybridizations, which is not always the case (Causton et al. 2003). There are methods that use all genes, but they have the disadvantage of having no independent confirmation and can be skewed if the majority of the genes on the array are differentially expressed. Other methods select an unbiased set of genes by ordering the genes by signal from each spot, placing them in order based on expression level, and using only those within a fixed window, i.e., 25^{th} percentile to 75^{th} percentile, or within a certain number of standard deviations of the mean (Schudt et al. 2001; Tseng et al. 2001). There is no right answer to all experimental situations; the best solution depends on the specific technology and experimental system.

The normalization process has several steps, as outlined by Wit and McClure (2004). The first step, spatial normalization, adjusts for location of spots. A general approach is to take the log ratio; however, this approach assumes that the spatial effects are multiplicative and affect each channel similarly. Other methods do not make this assumption, but have their own advantages and/or disadvantages (Yang et al. 2002b). The second step, background correction, adjusts the signal intensity for background noise. There are several methods for background correction, ranging from simple subtraction methods (Eisen 1999; Wolkenhauer et al. 2002) to more sophisticated methods (Efron et al. 2001; Kooperberg et al. 2002; Irizarry et al. 2003b; Rocke and Durbin 2001; Huber et al. 2002). It is the general consensus of the more sophisticated methods that, instead of simply subtracting the local background intensities, these methods use the global background intensities as

measured by the mean or median of all empty spot values and reset the negative values to zero. Next, dye-effect normalization is used to correct the artifacts from the dye due to the physical properties of Cy5 and Cy3; dye effect is not discussed here because single channel arrays are used in this study. There are two important critera for the choice of normalization methods: 1) genes expected to be approximately equal across both dyes, and 2) sufficiently large sample size.

Once the effect of location, background noise, and dye-effect have been corrected or accounted for, it is possible to normalize across multiple arrays within- and/or across- conditions. Methods that normalize within- and across-condition concurrently are known as global array normalization methods. These methods bring the mean or median intensity for each array to a fixed quantity and/or adjust the scale to another fixed value (Kerr et al. 2000; Wolfinger et al. 2001; Yang et al. 2002b). The major disadvantages of these methods are that they assume linearity. There are several within-replication methods that normalize all of the arrays within one condition and arrays across multiple conditions. There is little consensus in the microarray community about the best normalization method, and there may not in-fact be a single preferred solution (Allison et al. 2005; Imbeaud and Auffray 2005; Park et al. 2003). Quantile normalization is an accepted normalization method that can be adapted for within- and across-condition normalization and will be used in this study (Wit and McClure 2004).

There are several normalization package methods for the complete normalization of Affymetrix data, ranging from spatial to across array/condition normalization. These include the Mas 5.0 method developed by Affymetrix (www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf) and the RMA method (Bolstad et al. 2003). There is little consensus about the best method to use; however, there is some consensus about methods that don't perform as well as desired (Wit and McClure 2003). For this reason, in this study Mas 5.0 was not used; instead RMA and an

improvement on RMA, known as GCRMA, were used. All normalization methods were implemented using R Bioconductor (www.bioconductor.org) (See Appendix A for Commented R Code). RMA is a method of adjusting gene expression values that fits a robust, linear model to the probe-level data, analyzing each hybridized chip in the context of other chips in the experiment. The first step is model-based background correction, which neutralizes the effects of background noise and processing artifacts. The second step is quantile normalization that aligns expression values to a common scaled value. The last step of RMA is an interactive median-polishing procedure that summarizes the data and generates a single expression value for each probe set (Irizarry et al. 2003). GCRMA is a refined version of the RMA algorithm that replaces the model used in the background correction stage with a more sophisticated computation. The latter uses each probe's sequence information to adjust the measured intensity for the effects of non-specific binding due to the differences in bond strength between the two types of base pairs. GCRMA also takes into account the optical noise present in data acquisition to achieve an even greater accuracy and sensitivity. The two steps of the RMA algorithm following background correction remain unchanged in CGRMA. RMA was implemented in stages to view the different normalization steps, and GCRMA was implemented as a package. It should be noted that both the GCRMA and RMA procedures produce log 2 transformed data, as discussed above. Box plots and histograms of data quality before and after normalization show that all artifacts are removed and no outliers or systematic errors remain (Figures 4-4 a-b). Normalization was completed for each dataset, A17, A6, and PG, independently, because most artifacts and noise occurred as a result of a specific laboratory procedure or piece of equipment.

Experimental Analysis Plan

The experimental analysis plan to address the proposed research questions was implemented in R (Figure 4-3a-b). In addition to no testing correction, three multiple testing correction methods/options

are used to adjust the decision rules for the t-statistic: Benjamini and Hochberg 1995 (FDR BH), Benjamini and Yekutieli 2001 (FDR BY), and Holm 1979 (FWER). Five α levels were used: 0, 0.01, 0.05, 0.10, and 0.25. In addition, Affymetrix housekeeping genes (genes beginning with AFFY) are tested for consistency across multiple contrasts, control verses each treatment(4-2).

RESULTS AND DISCUSSION

RMA VS GCRMA

RMA and GCRMA resulted in nearly identical consistency results, as shown in Figures 4-5 a-c. Within the Aventis laboratory, a small variation in consistency is seen throughout; however one particular array shows greater inconsistency: Aventis low-dose 24-hour period. There is greater than 20% variation between the GCRMA and RMA methods, with GCRMA showing greater consistency (Table 4-2). Normalization results were revisited to attempt to identify reasons why the low-dose 24-hour period treatment groups between the Aventis 6 and Aventis 17 groups might show lower consistency. No reason for the variation could be identified. However, it was shown that the GCRMA method showed greater consistency for this treatment group and should, therefore, be used instead of RMA. Hence, for the remainder of this study, results will be presented using the GCRMA method.

NO MULTIPLE TESTING CORRECTION

Figures 4-6 a-c shows that consistency decreases within the Aventis laboratory as the α level or the expected percentage of false positives increases. This is anticipated because there are expected to be fewer true positives as the α level increases. In the original study, no multiple testing correction was used and an α level of 0.0001 was used, meaning that with 12,422 test genes (12488 standard Affymetrix GeneChip® **MGU**-74A mouse array without 66 AFFY control genes) there is expected

to be less than one false positive. The biological signal was proven to be retained across arrays, but there was no evaluation of power. As the alpha level increases, the consistency decreases dramatically, even at the 0.01 level. Therefore, some form of multiple testing correction might improve this result.

FDR vs FWER

Figures 4-7 a-c compare the consistency between controls of the family-wise error rate and the false discovery rate. As discussed above, the FDR is the expected proportion of false discoveries amongst the rejected hypotheses, and the FWER is the probability of erroneously rejecting even one of the true null hypotheses. The control of the FWER at some level, α, requires each of the individual m tests to be conducted at lower levels, as in the Bonferroni procedure where α level is divided by the number of tests performed (Holm 1979). The power to detect a specific hypothesis while controlling the FWER is greatly reduced when the number of hypotheses in the family increases. The incurred loss of power, even in the medium size problems, has led many practitioners to neglect multiplicity control altogether; however, as shown above, this does not always lead to consistent results.

Two different false discovery rates were compared to evaluate the consistency between results obtained by the two methods at different error control rates. The results for FWER and FDR BY were very surprising: both FWER and FDR BY showed similar consistency at each alpha level for AFFY control genes for each treatment group. . As with the other methods, as power increased or the p-value cutoff increased, the consistency of the FDR methods decreased. No change in consistency was noted with the FWER method. A small decrease in consistency was noted with the FDR BY method and a large decrease in consistency was noted with the FDR BH method. The Benjamini and Hochberg 1995 FDR control method is the original FDR method for multiple independent hypotheses. The Benjamini and Yekutieli FDR control BY method is equivalent to the

Benjamini and Hochberg method for dependent hypotheses. In the general case, independence cannot be assumed about gene expression and microarray experiments; therefore, the Benjamini and Yekutieli should be the appropriate method to control the FDR for gene expression experiments. The results of this study show that there is greater consistency when using the appropriate FDR BYmethod. In addition to the general microarray case, in this study multiple contrasts are made between treatment groups and one control; therefore, independence cannot be assumed.

On the larger scale, the results between the FWER and FDR methods are comparable and almost equivalent using the BH FDR method. The control of the FDR is less stringent than the FWER, leading to large gain of power with the use of the FDR method. The difference between the rejection regions at each alpha- level is quite large, which leads to the difference in results. For example, if there were 18 hypotheses being tested, a simple approximation of the rejection region at the 0.05 α level for the FWER control method for the ninth hypothesis would be $\alpha/m\text{-}i+1$, where m is the total number of hypotheses, and i is the ith test, which in this case would be values less than 0.005. In comparison using the Benjamini and Hochberg method, the same rejection region at the 0.05 level would be values less than $(i/m)\alpha \approx .025$. The overall message is that, for exploratory methods, a great increase of power is seen with the use of FDR without a loss of consistency between methods. Results in the following sections are shown for the FDR 1995 method and the FWER method.

ALPHA LEVELS 01, 0.05, 0.10, 0.25

Figures 4-8 a-c show that as the alpha level increases, the consistency decreases. The choice of alpha value controls the probability of being wrong if the null hypothesis is rejected for each hypothesis and is dependent on the purpose of the experiment. In the context of studies that use of multiple public microarrays to generate new hypothesis, a higher alpha level is acceptable because the user wants to

discover interesting genes of biological significance. As shown in above example, the alpha value is also used in the calculation of the rejection region. The results of this study show that even at $\alpha=0.25$ there is still consistency greater than 96% using the FWER method and greater than 90% using the FDR method. For exploratory purposes, an alpha level of 0.10 or less would be appropriate to apply across public microarray experiments without loss of consistency.

AFFYMETRIX INTERNAL CONTROL GENES

Figures 4-9a-c show that when using a multiple testing correction method, internal housekeeping genes are consistent across all arrays. This is a reassuring result for normalization methods that depend on internal control genes. It is also reassuring that for alpha levels less than or equal to 0.10, the consistency of these genes are the nearly equal.

CONCLUSION

The use of public microarray data to derive new hypotheses or to test existing theories is necessary and responsible. The ability to reuse legacy data to learn new information is an invaluable resource in the public domain and can enhance the value of these data. This study demonstrates that there are methods that retain consistency across laboratories and within laboratories. It is suggested that, whether certainty or exploration is desired, there are methods that offer higher levels of consistency. The biological implications are not considered here; however, as shown in the original Müeller et al. (2005) paper, the biological message is retained across multiple laboratories and time periods.

A study similar to this should be conducted using other data sets. There are further considerations for the use of public microarray data that have not been addressed such as multiple platforms, species, tissues and cell types. This study represents an important first step towards understanding how to

effectively use microarray data from public resources. An interesting next step for this project would be to add in fold change criterion and evaluate the biological consistency of these data by exploring the use of pathway level analysis methods such as the Pathway Level Analysis of Gene Expression (PLAGE) using singular value decomposition (Tomfohr et al. 2005; http://dulci.org/pathways/) and the Gene Enrichment Analysis (GSEA) method from the Broad Institute (Subramanian et al. 2007; http://www.broad.mit.edu/gsea). The present study, while limited in scope, offers useful insight into the statistical decision making necessary for the use of public microarray data.

REFERENCES

Allison DB, Cui Cui X, Page GP. Sabripour M. 2005. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 7:55-65

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series* B, **57**, 289–300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high densitiy oligonucleotide array data based on bias and variance. Bioinformatics 19(2):185-93.

Causton, H.C., Quackenbush, J., Brazma, Alvis. 2003. Microarray Gene Expression Data Analysis: A Beginner's Guide. Malden, MA: Blackwell Publishing. 1-130.

Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA. 2007. Reproducibility of microarray data: A further analysis of microarray quality control (MAQC) data. BMC Bioinformatics 8:412; doi: 10.1186/1471-2105-8-412.

Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ. 2008. Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res; doi: 10.1093/nar/gkn580.

Draghici S, Khatri P, Eklund AC, Szallasi Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet 22(2):101-109; doi: 10.1016/j.tig.2005.12.005.

Dudoit S, Yang T, Callow M, Speed T. 2002. Stattistical methods for identifying differentially expressed genes in replicated CDNA microarray experiments. Statistica Sinica 12(1):111-39.

Efron B, Tibshirani R, Storey JD, Tusher V. 2001. Empirical Bayes analysis of a microarray experiment Am Stat Assoc 96(454): 1151-60.

Eisen MB. 1999. ScanAlyze, User Manual. Stanford University, Stanford. Available: http://rana.lbl.gov/manuals/ScanAlyzeDoc.pdf. [accessed 20 Nov 2008].

Ge YC, Dudoit S, Speed TP. 2003. Resampling-based multiple testing for microarray data analysis. Test 12(1). 1-77.

Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L et al. 2006. Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nat Biotechnol 24(9):1162-1169.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Huber W, con Heydebreck A, Sültmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinfrmatics. 18:s96-s104.

Imbeaud S and Auffray C. 2005. The 39 steps in gene expression profiling: critical issues and proposed best practices for microarray experiments. Drug Discov. Today. 10(17):1175-1182.

Irizarry RA, Gautier L, Cope LM. 2003a. An R Package for analyses of Affymetrix oligonucleotide arrays. The Analysis of Gene Expression Data (eds. Parmigiani G, Garrett ES, Irizarry RA), Statistics for Biology and Health. New York:Springer-Verlag 102-19.

Irizarry RA, Hobbs B, Colin F, Beaazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003b. Exploration, normalization, and summaries of high denisity oligonucleotide array probe level data. Biostatistics 4(2):249-64.

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC et al. 2005. Multiple-laboratory comparison of microarray platforms. Nat Methods 2(5):345-350; doi: 10.1038/nmeth756.

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8(1):118-127; doi: 10.1093/biostatistics/kxj037.

Kerr MK and Churchill GA. 2001. Statistical design and the analysis of gene expression microarray data. Genet. Res. 77:123-8.

Kerr Mk, Martin M, Churchill GA. 200. Analysis of variance for gene expression microarray data. J. Comput Biol. 7(6):819-37.

Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T. 2002. Improved background correction for spotted DNA microarrays. J Comput Biol 9(1):55-66.

Lee PD, Sladek R, Greenwood CMT, Hudson TJ. 2002. Control genes and cariability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res. 12:292-297.

MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA et al. 2006. The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24(9):1151-1161; doi: 10.1038/nbt1239.

Müller A, Boitier E, Hu T, Carr GJ, Le Fèvre A, Marchandeau J, Flor M, Jefferson F, Aardema MJ, Thybaud V. 2005. Laboratory Variability Does not Preclude Identification of Biological Functions Impacted by Hydroxyurea. Environ Mol Mutagen. 46:221-235.

Paananen J, Wong G. 2006. Integration of genomic data for pharmacology and toxicology using internet resources. SAR QSAR Environ Res 17(1):25-36; doi: 10.1080/10659360600562053.

Park T, Yi, S, Kang S, Less, S, Less, Y, Simon R. 2003. Evaluation of normalization methods for microarray data. BMC Bioinformatics 4(33):3-16

Quackenbush J. 2002. Microarray data normalization and transformation. Nat Genet 32 Suppl:496-501; doi: 10.1038/ng1032.

Reiner A, Yekutieli D, Benjamini Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19(3):368-375.

Rocke DM and Durbin B. 2001. A model for measurement error for gene expression arrays. J. Bioinform. Comput. Biol. 8:557-69.

Schadt EE, Li C, Ellis B, Wong WH. 2001. Feature extraction and normalization algorithms for high-density oligonucleotide gen expression array data. J Cell Biochem. 37:120-125.

Shi L, Jones WD, Jensen RV, Harris SC, Perkins RG, Goodsaid FM et al. 2008. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. BMC Bioinformatics 9 Suppl 9:S10; doi: 10.1186/1471-2105-9-S9-S10.

Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. 2007. GSEA-P: A desktop application for gene set enrichment analysis. Bioinformatics 23(23):3251-3253; doi: 10.1093/bioinformatics/btm369.

Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. 2007. GSEA-P: A desktop application for gene set enrichment analysis. Bioinformatics 23(23):3251-3253; doi: 10.1093/bioinformatics/btm369.

Tan PK, Downey TJ, Spitznagel EL,Jr, Xu P, Fu D, Dimitrov DS et al. 2003. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 31(19):5676-5684.

Tenenbaum JD, Walker MG, Utz PJ, Butte AJ. 2008. Expression-based pathway signature analysis (EPSA): Mining publicly available microarray data for insight into human disease. BMC Med Genomics 1(1):51; doi: 10.1186/1755-8794-1-51.

Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A , Heinen E. 1999. Housekeeping genes as internal standards: use and limits. J Biotechnol 75:291-295.

Tomfohr J, Lu J, Kepler TB. 2005. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics 6:225; doi: 10.1186/1471-2105-6-225.

Tseng GC, Oh MK, Rohlin L, Liao, JC, Wong WH. 2001. Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acid Res. 29:2549-2557.

Wit EC, and McClure JD. 2004. Statistics for Microarrays Design, Analysis and Inference. West Sussex, England: John, Wiley and Sons, LTD. 13-203.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. 2001. Assessing Gene Significance frin cDNA microarray expression data via mixed models. J Comput Biol 6(6):625-37.

Wolkenhauer O, Moeller-Levet C, Sanchez-Cabo F. 2002. The curse of normalization. Comp Funct Genomics 3(4):375-9

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30(4):e15.

TABLES

Table 4-1. Description of False Positives (F_P) and False Negatives (F_N) in the context of Differentially Expressed Genes.

	Declared Non-Differentially Expressed	**Declared Differentially Expressed**	**Total**
# of non-differentially expressed genes	T_N	F_P	m_0
# of differentially expressed genes	F_N	T_p	$m-m_0$
Total	m-S	S	m

Table 4-2. Difference in consistency rates between the RMA and GCRMA methods of normalization with each multiple testing correction method at each alpha level. Each entry is the result of subtracting the consistency using the RMA normalization method and the consistency using the GCRMA normalization method. A negative indicates that the GCRMA method outperformed the RMA method, the AVENTIS low 24 treatment group demonstrates the improvement by the GCRMA method from the RMA method.

ALL GENES	None 0	Holm 0.01	FDR BH 0.01	FDR BY 0.01	None 0.05	Holm 0.05	FDR BH 0.05	FDR BY 0.05	None 0.10	Holm 0.10	FDR BH 0.10	FDR BY 0.10	None 0.25	Holm 0.25	FDR BH 0.25	FDR BY 0.25
AVENTIS low4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AVENTIS lowmid4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AVENTIS midhigh 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AVENTIS high 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AVENTIS low 24	-0.12	0	-0.02	-0.04	-0.24	0	-0.03	-0.10	-0.28	0	-0.04	-0.17	-0.25	0	-0.07	-0.40
AVENTIS lowmid 24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AVENTIS midhigh 24	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0.01	0	0
AVENTIS high 24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PG A17 lowmid 4	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0
PG A6 lowmid 4	0	0.01	0	0	0	0.01	0	0	0	0.02	0	0	0	0.01	0	0
PG A17 high 4	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0
PG A6 high 4	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0

Table 4-2. (Continued)

PG A17 lowmid 24	0	0	0	0	0	0	0	0	0	-0.01	0	0	0	0	0	0
PG A6 lowmid 24	0	0	0	0	0	0	0	0	0	-0.01	0	0	0	0	0	0
PG A17 high 24	0	0	0	0	0	-0.01	0	0	0	-0.01	0	0	0	0	0	0
PG A6 high 24	0	0	0	0	0	-0.01	0	0	0	-0.01	0	0	0	0	0	0
lowmid 4	0	0.01	0	0	0	0.01	0	0	0	0.02	0	0	0	0.01	0	0
high 4	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0	0	0.01	0	0
lowmid 24	0	0	0	0	0	0	0	0	0	-0.01	0	0	0	0	0	0
high 24	0	0	0	0	0	-0.01	0	0	0	-0.01	0	0	0	0	0	0

FIGURES

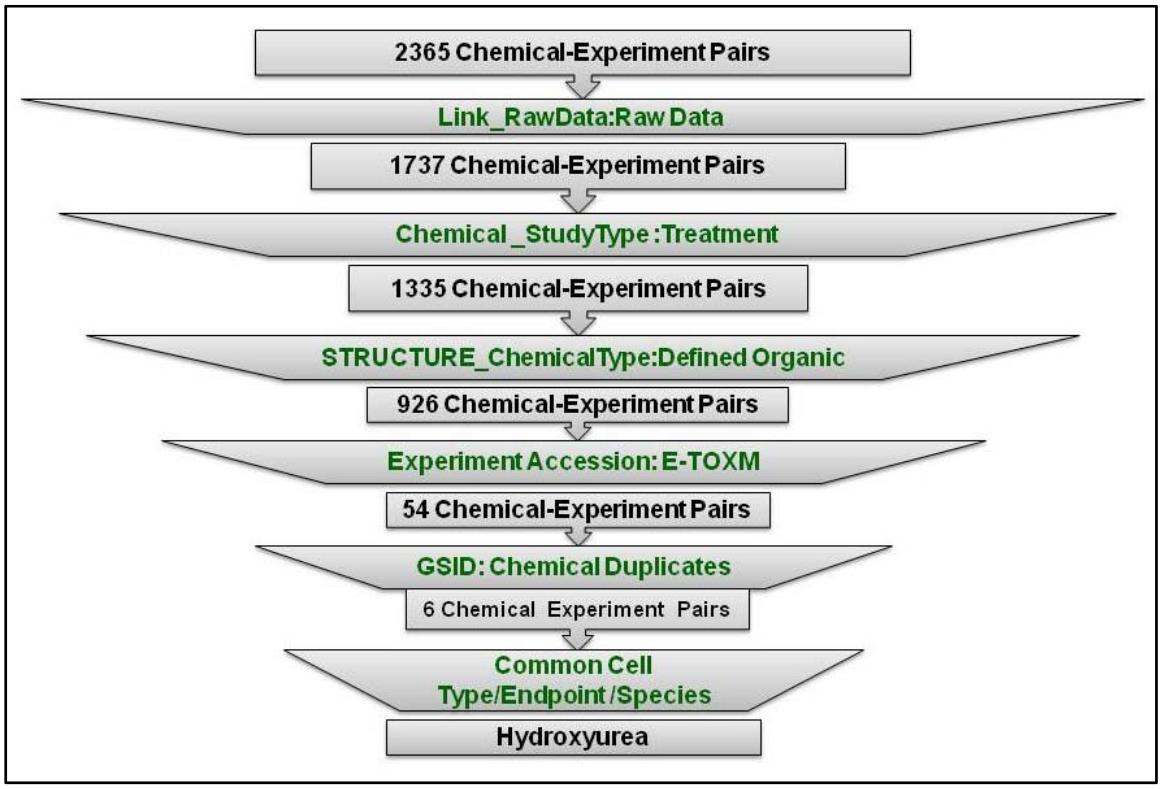


Figure 4-1. The Data Selection Paradigm. Data was selected from the ArrayExpress Chemical Index.

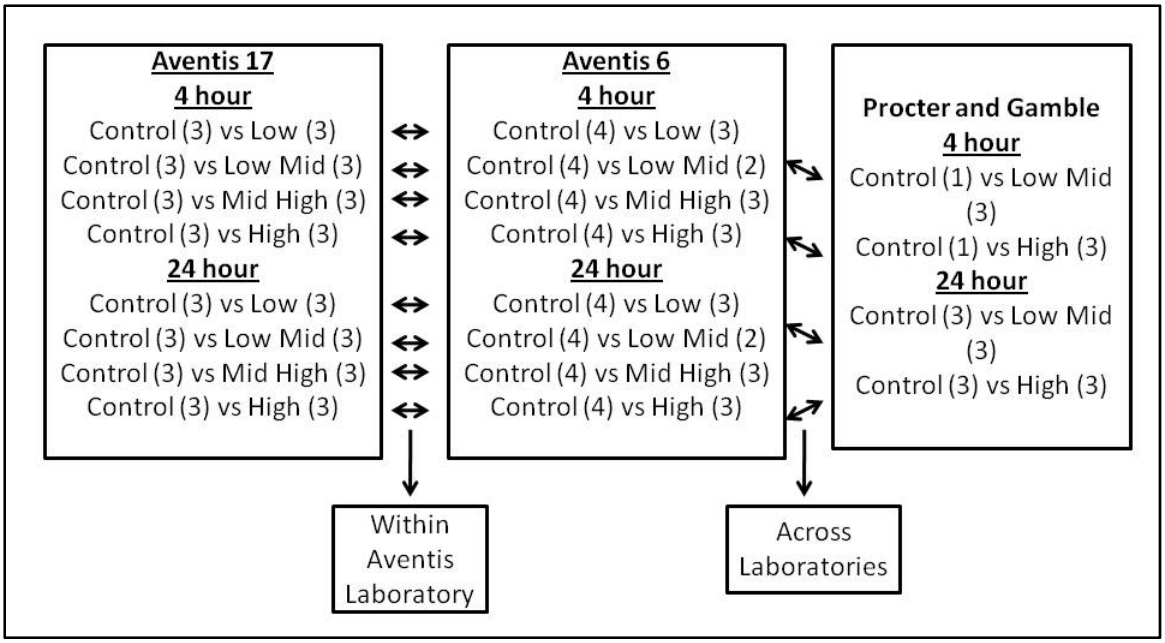


Figure 4-2. Experimental Design Plan. Numbers in parenthesis represent the number of arrays. Included contrast represent those that are minimally sufficient. An exception was made for the Procter and Gamble 4hr control.

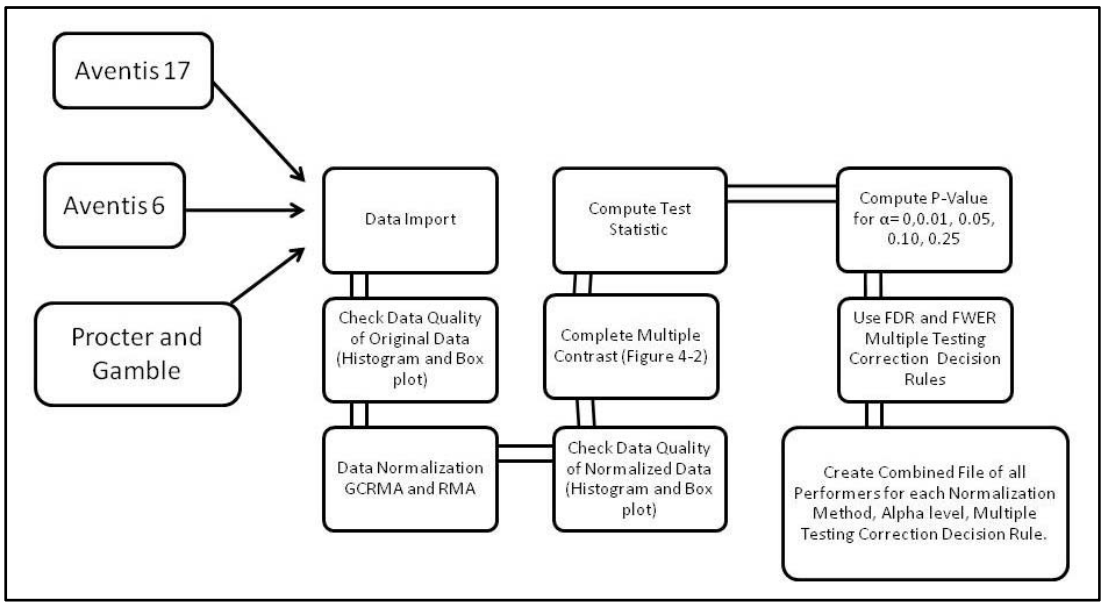


Figure 4-3a. Data Analysis Plan as implemented using R libraries affy, biobase, gcrma, and limma. See Apendix A for the complete R Scripts.

Gene	A17 24hr Control High RMA FDR α=0.05	A6 24hr Control High RMA FDR α=0.05	PG 24hr Control High RMA FDR α=0.05	AVENTIS 24hr High RMA FDR α=0.05	A17 PG 24hr High RMA FDR α=0.05	A6 PG 24hr High RMA FDR α=0.05	A6 A17 PG 24hr High RMA FDR α=0.05
Gene 1	1	1	1	1	1	1	1
Gene 2	0	0	-1	1	0	0	0
Gene 3	1	-1	1	0	1	0	0
Gene 4	1	0	0	0	0	1	0
AFFY-Gene 1	0	0	0	1	1	1	1
AFFY Gene 2	0	0	0	1	1	1	1
Affy Gene 3	1	-1	1	0	1	0	0
Total Number of Genes	4	4	4				
Total Overlap				2	2	2	1
Total Number of AFFY Genes	3	3	3	2	3	2	2

Within Aventis Laboratory Calculation: Consistency = Total Aventis Overlap / Total Number of Genes

Across Labortories: Consistency = Total (A17 PG) or (A6 PG) Overlap / Total Number of Genes

Within and Across Labortories: Consistency = Total (A17 PG A6) Overlap / Total Number of Genes

Output Decision Rule: 1= Significant (Upregulated) -1=Significant (Down Regulated) 0= No Change

Computed Overlap between performers: 1= Performers have same call 0= Performers have different calls

Figure 4-3b. Data Analysis Plan for the Computation of Consistency Within and Across Laboratories.

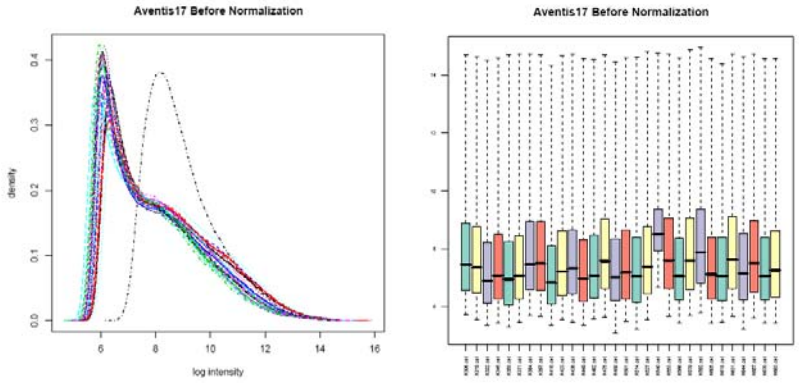


Figure 4-4a. Data from Aventis 17 Arrays before Normalization. There appears to be one outlier indicated in both the histogram and box plot.

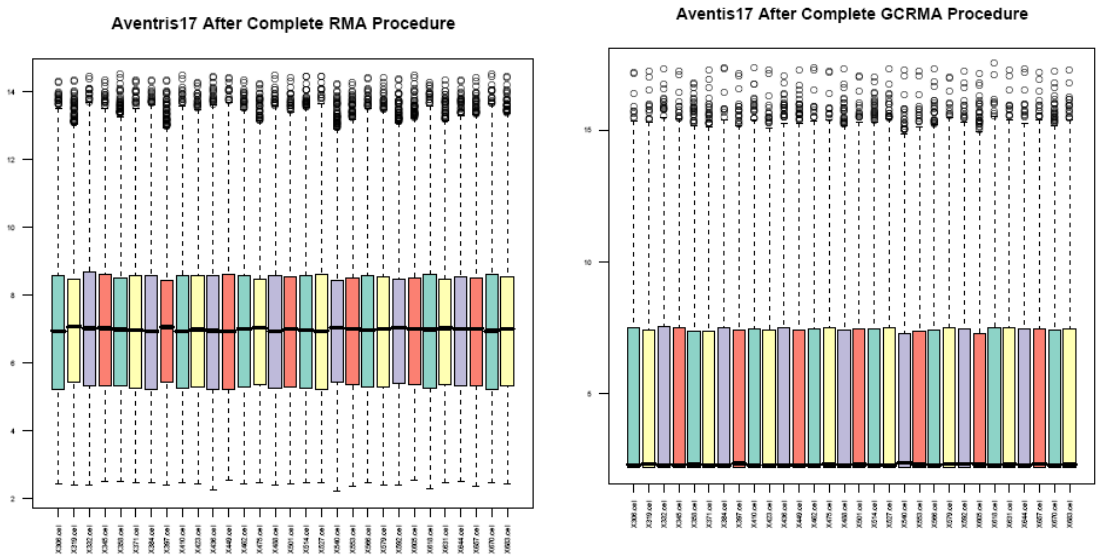


Figure 4-4b. Data from Aventis 17 After GCRMA and RMA Normalization. There appears to be no major outliers or systematic effects.

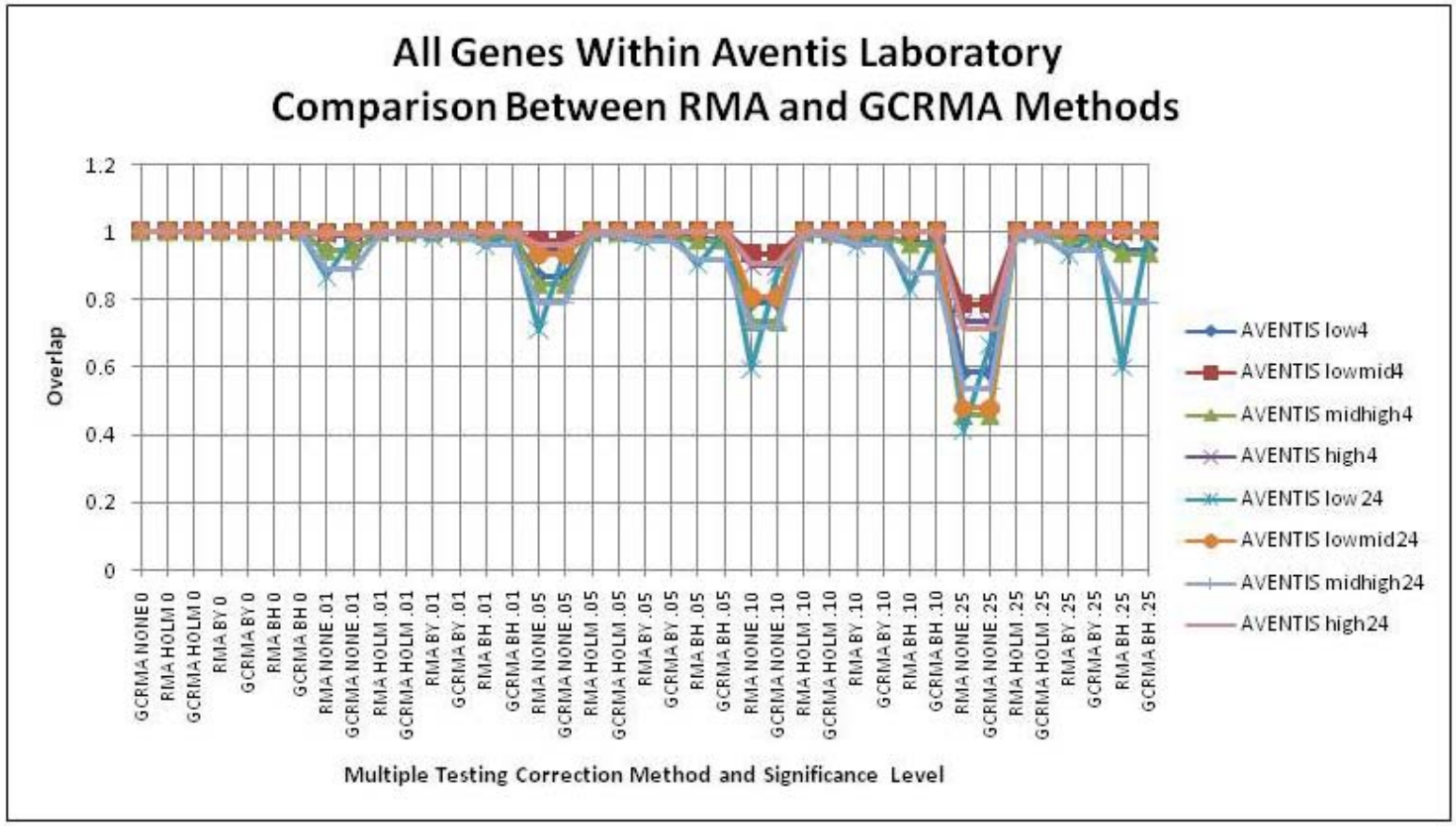


Figure 4-5a. Comparison of the Consistency within the Aventis laboratory between the RMA method and GCRMA Methods. Overall and in the most variable treatment group, AVENTIS low 24, GCRMA shows greater consistency.

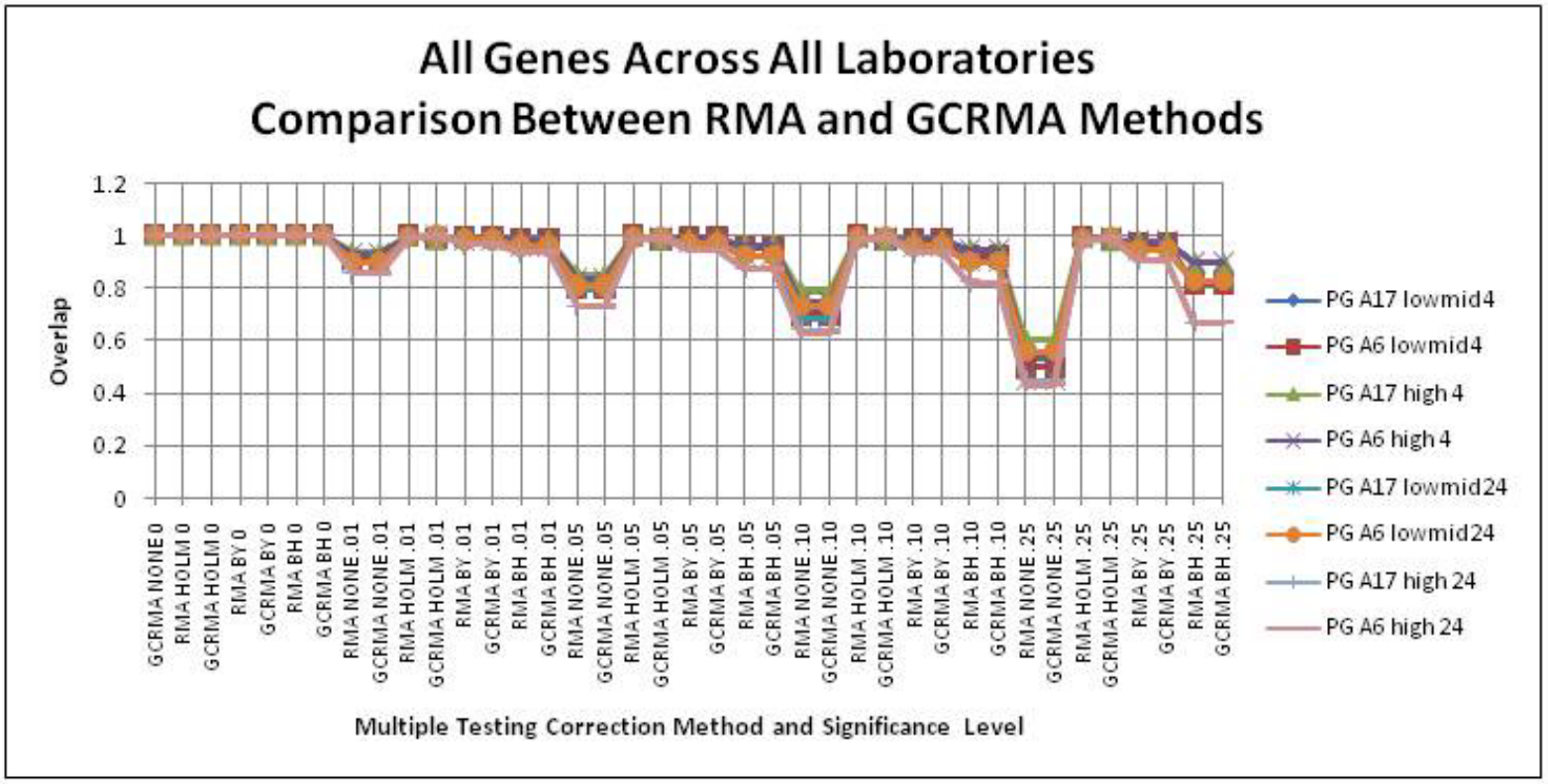


Figure 4-5b. Comparison of the Consistency across Aventis and Procter and Gamble laboratories between the RMA method and GCRMA Methods. Overall GCRMA shows greater consistency.

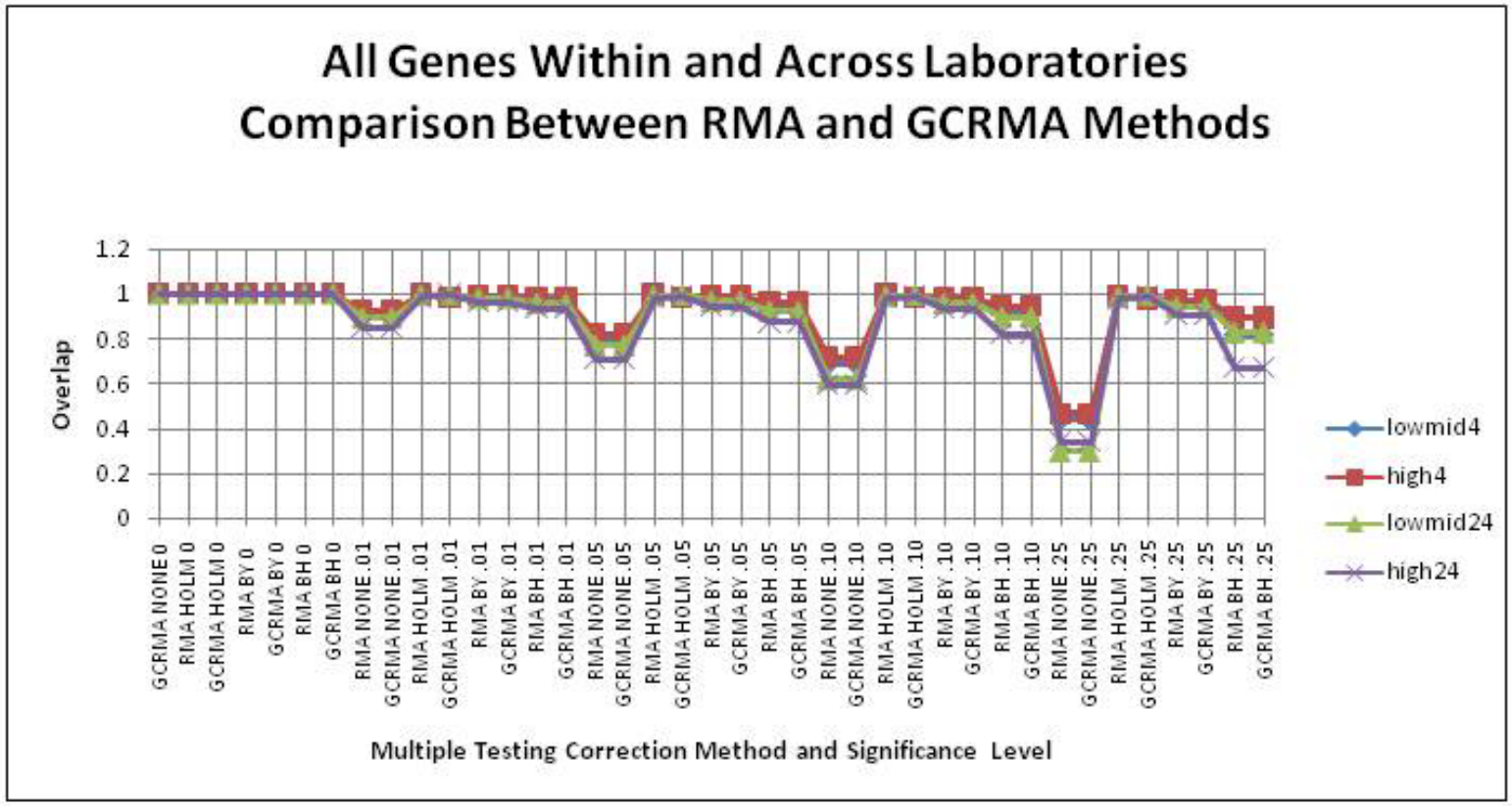


Figure 4-5c. Comparison of the Consistency within and Across Aventis and Procter and Gamble laboratories between the RMA method and GCRMA Methods. Overall GCRMA shows greater consistency.

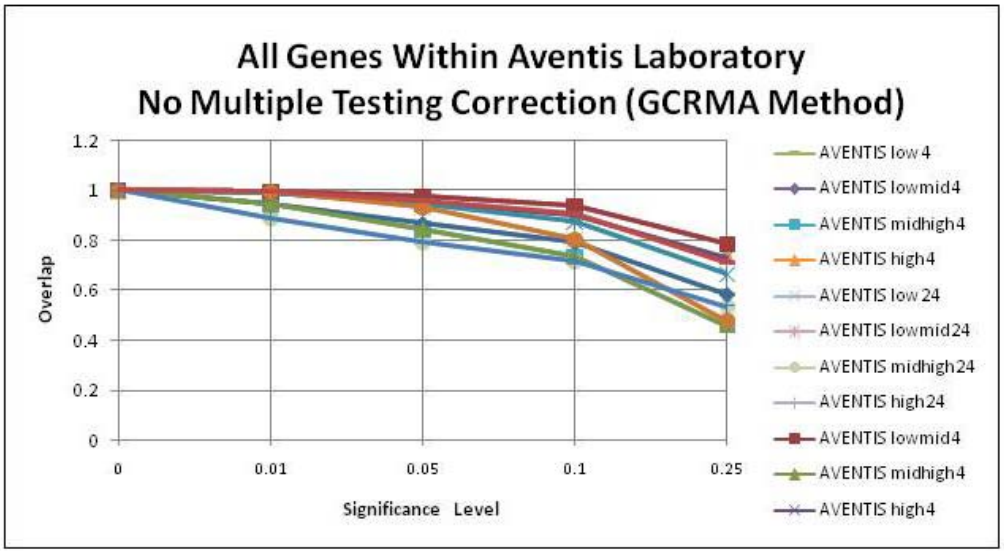


Figure 4-6a. Consistency within the Aventis laboratory with no Multiple Testing Correction. A dramatic drop in consistency is seen as the Significance Level increases.

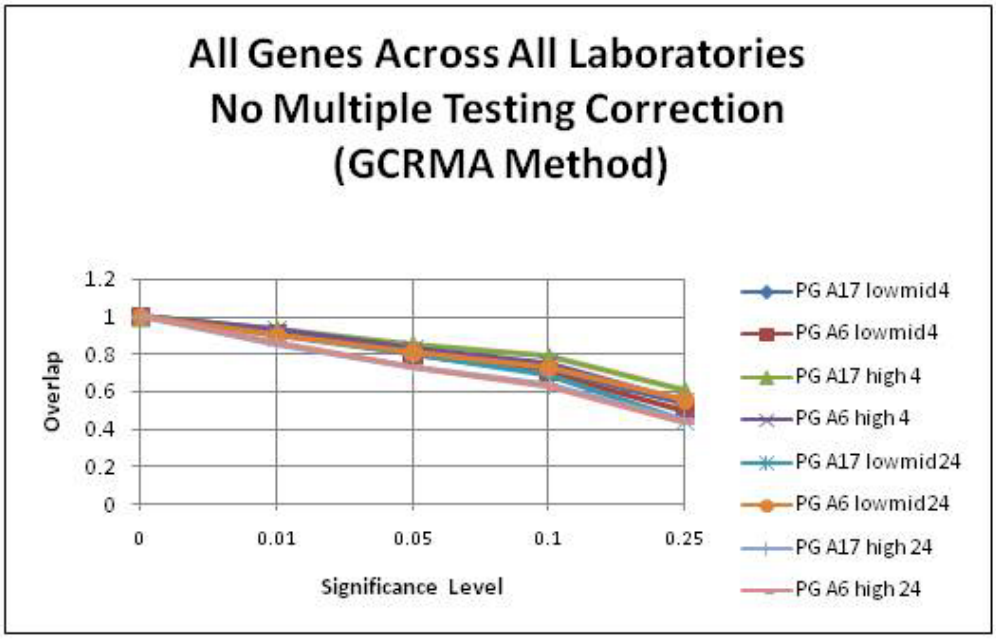


Figure 4-6b. Consistency across Aventis and Procter and Gamble laboratories with no Multiple Testing Correction. A dramatic drop in consistency is seen as the Significance Level increases.

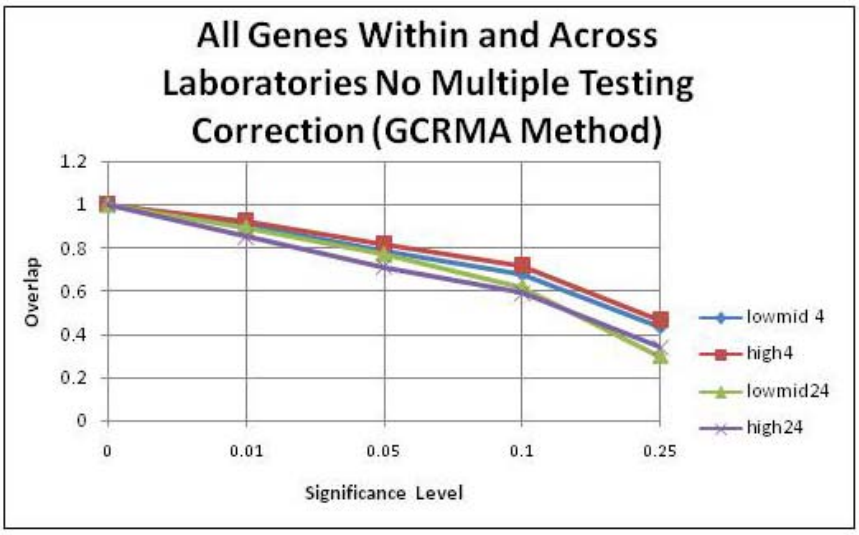


Figure 4-6c. Consistency within and across Aventis and Procter and Gamble laboratories with no Multiple Testing Correction. A dramatic drop in consistency is seen as the Significance Level increases.

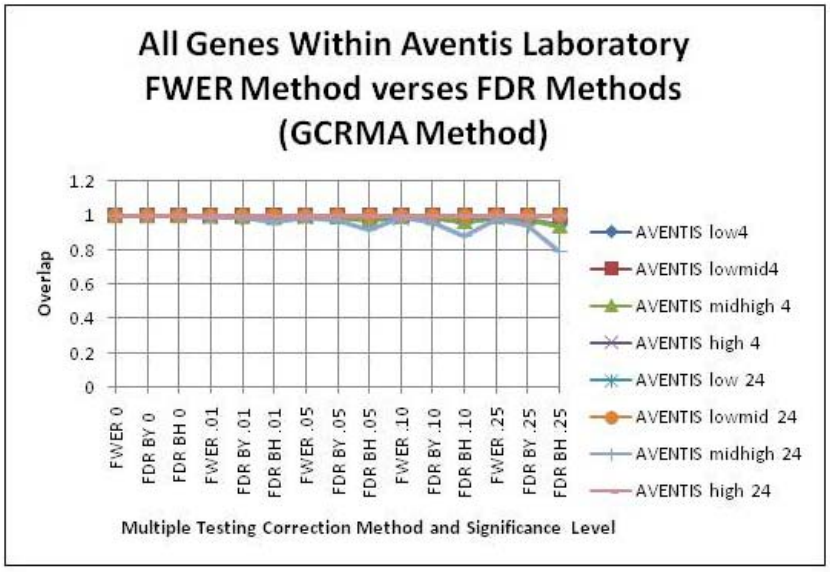


Figure 4-7a. Comparison of the Consistency within the Aventis laboratory between FWER control and FDR controls. FDR BY is the Benjamini and Yekutieli method for dependent tests and FDR BH is the Bejamini and Hochberg method for independent tests. FWER and FDR By show high levels of consistency. The FDR BH should not be used for microarray studies similar to this study.

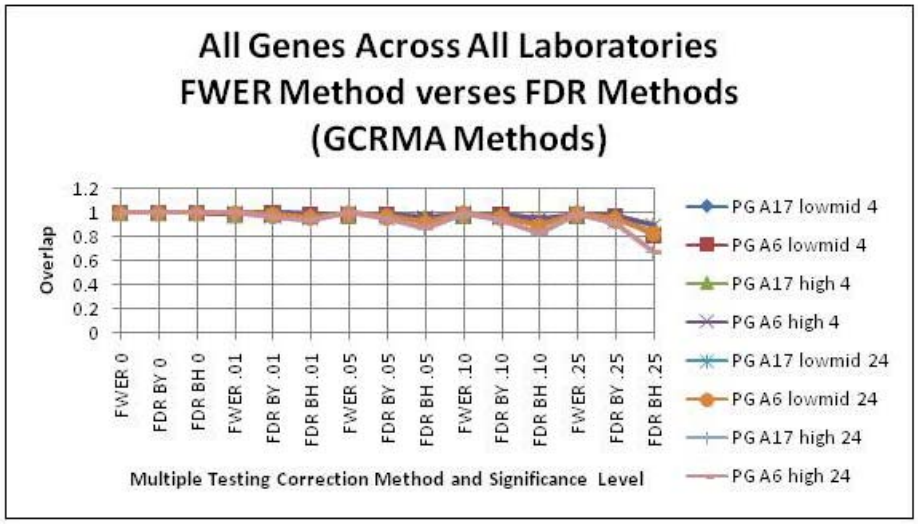


Figure 4-7b. Comparison of the Consistency across Aventis and Procter and Gamble laboratories between FWER control and FDR controls. FWER and FDR By show high levels of consistency. The FDR BH should not be used for microarray studies similar to this study.

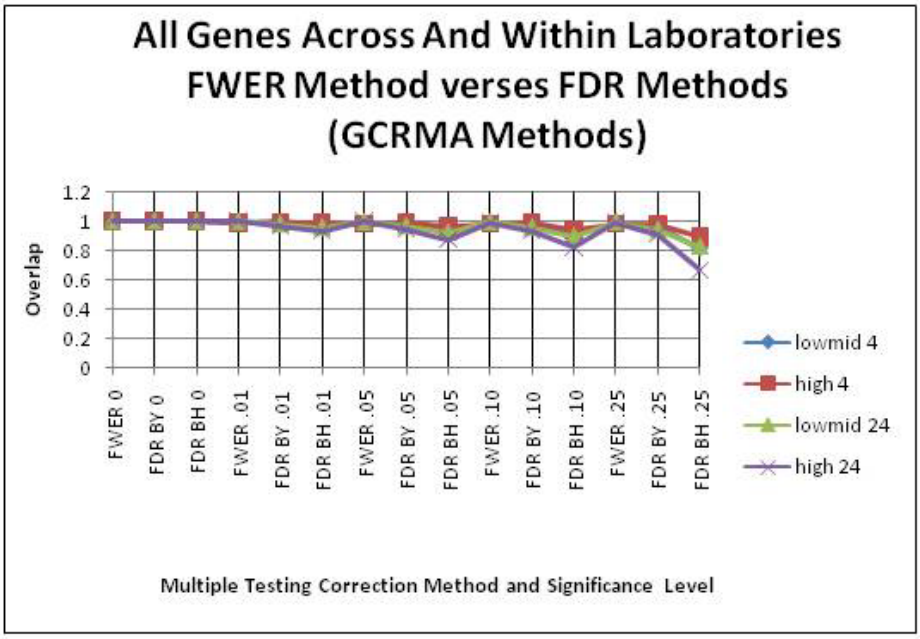


Figure 4-7c. Comparison of the Consistency within and across Aventis and Procter and Gamble laboratories between FWER control and FDR controls. FWER and FDR By show high levels of consistency. The FDR BH should not be used for microarray studies similar to this study.

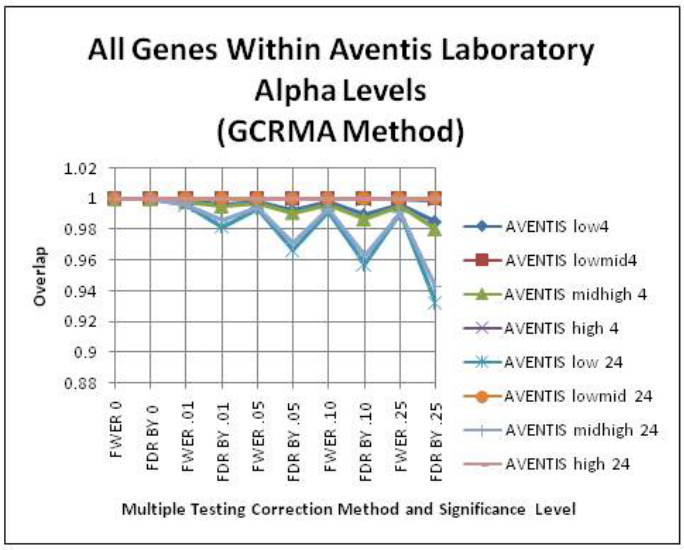


Figure 4-8a. Comparison of the Consistency within the Aventis laboratory at each significance level. As α increases the consistency decreases with the use of FDR control methods, however the increase is not large. The smallest consistency is still greater than 90%.

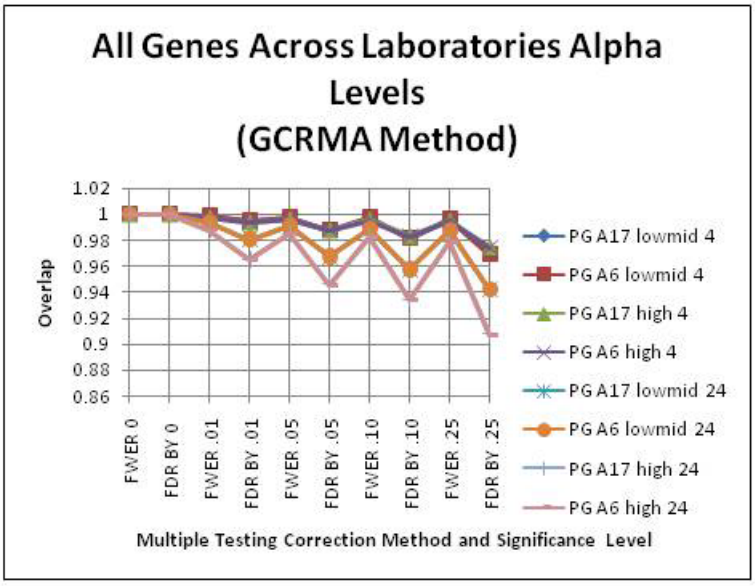


Figure 4-8b. Comparison of the Consistency across Aventis and Procter and Gamble laboratories at each significance level. As α increases the consistency decreases with the use of FDR control methods, however the increase is not large. The smallest consistency is still greater than 90%.

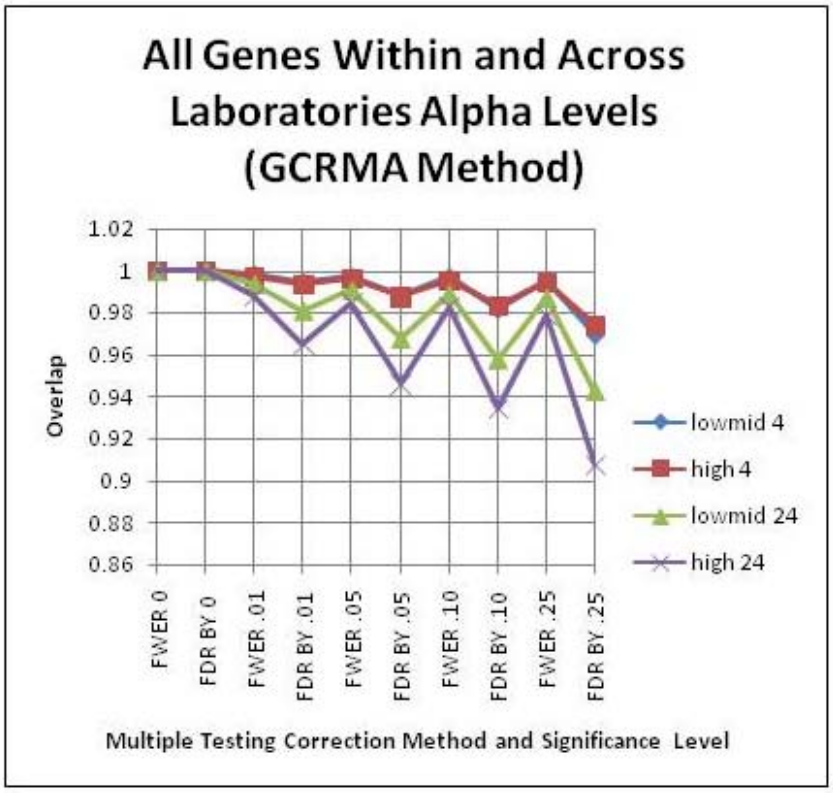


Figure 4-8c. Comparison of the Consistency within and across Aventis and Procter and Gamble laboratories at each significance level. As α increases the consistency decreases with the use of FDR control methods, however the increase is not large. The smallest consistency is still greater than 90%.

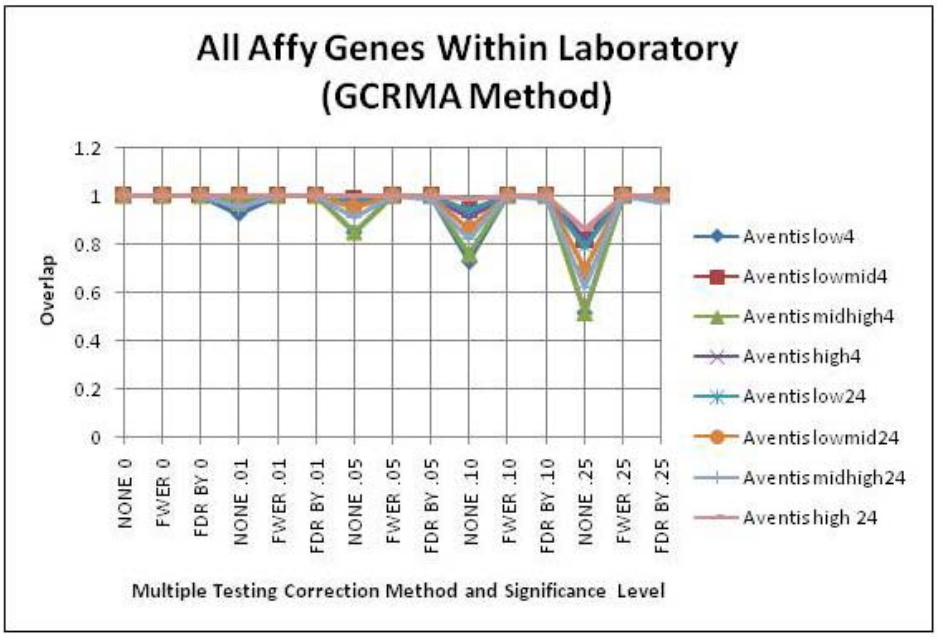


Figure 4-9a. Consistency within the Aventis laboratory for AFFY Control Genes. When a multiple testing correction method is used AFFY genes are Consistent within the Aventis Laboratory.

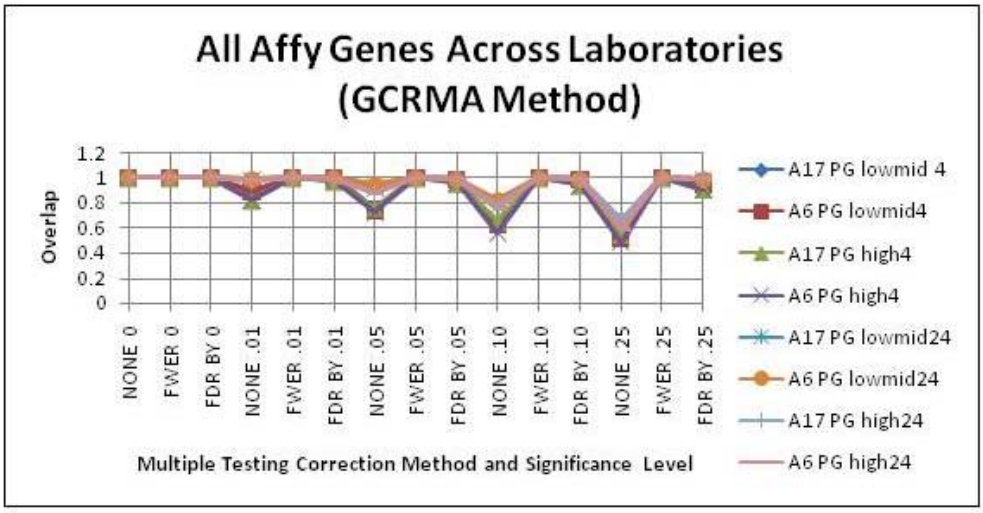


Figure 4-9b. Consistency across Aventis and Procter and Gambles laboratories for AFFY Control Genes. When a multiple testing correction method is used AFFY genes are consistent across laboratories.

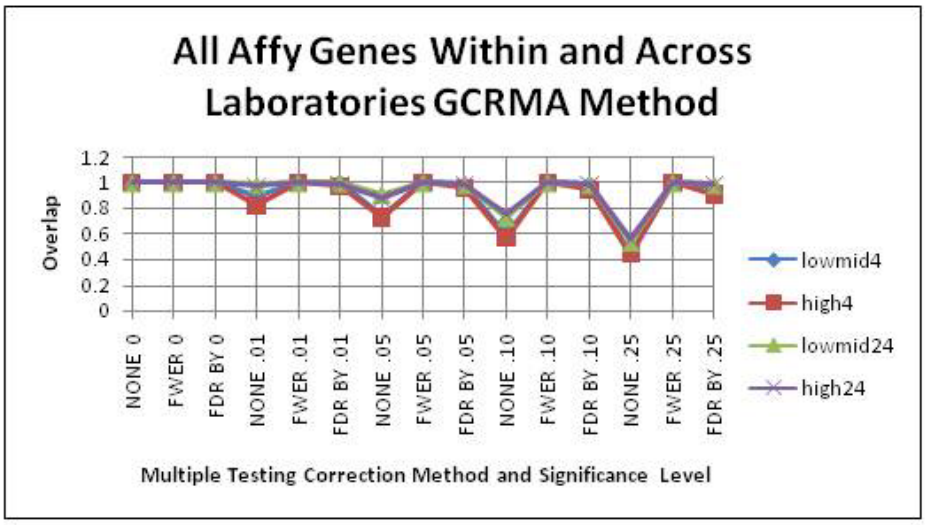


Figure 4-9c. Consistency within and across Aventis and Procter and Gambles laboratories for AFFY Control Genes. AFFY control genes are less consistent when FDR controls methods are used within and across laboratories.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In this research project, a paradigm shift in the field of toxicogenomics, towards what I have termed toxico-chemogenomics, has been described. Towards this objective, public gene expression resources have been identified, static versions have been chemically annotated and standardized, these annotated files have been published on-line in downloadable and structure-searchable form, and the files have been integrated into a public, multi-domain, chemically indexed resource. The two major public gene expression resources that served as the primary focus of these efforts are the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) and the European Bioinformatics Institute's (EBI) ArrayExpress. As of September 20, 2008 the standardized, chemically indexed and annotated versions of GEO and ArrayExpress experimental content were published in coordination with the U.S. Environmental Protection Agency's Distributed Structure-Searchable Toxicity (DSSTox) Database Network project. The resulting Structure-Index Locator files and Auxiliary files, ARYEXP and GEOGSE, were published in downloadable file formats on the DSSTox website and incorporated into the DSSTox Structure-Browser, allowing ArrayExpress and GEO to be structurally searched in the context of representative toxicological data contained in DSSTox. In addition, the Structure Index Locator files for ArrayExpress and GEO were incorporated into the NCBI PubChem system, allowing ArrayExpress and GEO experiments to be structure-located and structure-searched in the context of diverse biological and chemical data for millions of compounds and thousands of bioassays. These files also enabled, for the first time, assessment of the chemical content and coverage of GEO and ArrayExpress pertaining to a subclass of experiments of high relevance to toxicogenomics, i.e. "Treatment"-related experiments involving defined organic chemicals, many of which are of potential toxicological interest.

The path forward is clear for the ARYEXP and GEOGSE files. The information used to chemically annotate these files, as well as the resulting files, are being provided to both EBI's ArrayExpress and

NCBI's GEO teams, along with the supporting publications detailing the annotation process and results. The chemical-experiment inventory in these two systems is current in our files as of September 20, 2008. The total experimental content in these two "live" user-depositor repositories has increased approximately 5% since that time. We have laid out a clear motivation for chemical annotation of these resources and demonstrated the enhanced capabilities that can result from such annotation. We have also documented the tedious and inefficient process, involving automated and manual curation efforts, that was required for the present annotation effort. Our annotation process is far from the ideal and it should go without saying that a clear objective of the present effort is to convince the major parties (GEO and ArrayExpress) that this process should never have to be repeated. Immediate adoption of minimal chemical annotation requirements accompanying user-deposits of microarray experiment deposits could greatly facilitate chemical annotation updates in the future (although quality review will likely still be necessary).

GEO and PubChem are both projects within the NCBI, and better coordination between these two resources has been a long-sought goal for both. In recent discussions, the PubChem staff has already indicated an interest in highlighting the newly deposited DSSTox GEO and ArrayExpress content within PubChem to their users, including a plan to create a new biological data category of "microarray data" to highlight this new content on PubChem Compound pages. The PubChem staff is also willing to help champion the cause of instituting better chemical annotation standards within both NCBI and EBI, using our recent efforts and files as significant leveraging. Furthermore, within EBI, we learned that the Chemicals of Biological Interest (ChEBI) chemical ontology project has very recently instituted cross-referencing of ChEBI chemical information to ArrayExpress experiments, albeit for only a handful of experiments (i.e., 4) where ChEBI identifiers were provided by submitters. The present project has, in effect, delivered to ChEBI not only the entire curated and

quality reviewed structure inventory of ArrayExpress experiments linked to experiment accession identifiers, but also the entire curated inventory for GEO. The capability currently exists within ChEBI to immediately incorporate our published structure-index files into their system. Through ChEBI and PubChem, we will have effectively enlisted strong advocates for encouraging more formal chemical annotation requirements to be instituted within both GEO and ArrayExpress. These requirements could be as minimal as a required field presented to the depositor for entry of a chemical name pertaining to the experiment, ideally accompanied by the Chemical_StudyType field entries proposed herein. Until such time as these recommendations are adopted, quarterly updates incorporating chemical annotation of new experimental content in GEO and ArrayExpress will be carried out using the chemical annotation methods described herein.

ARYEXP and GEOGSE are structure-index files and do not contain the actually microarray data or microarray annotation files. To encompass these data types, integration of ArrayExpress and GEO into the Chemical Effects in Biological Systems (CEBS) toxicogenomics database is proposed and collaboration with CEBS is currently underway. A future goal is to develop an automated system of porting microarray data and annotation files directly from ArrayExpress and GEO to CEBS, with chemical annotation handled in collaboration with the DSSTox project. In coordination with CEBS, we are also recommending minimal chemical annotation requirements to ontology workgroups involved in MIBBI (Minimum Information for Biological and Biomedical Investigation; http://www.mibbi.org/index.php/Main_Page), in particular the MIAME/Tox and Tox Biology Checklist projects.

A system such as CEBS, where microarray data can be structure–searched and subsequently analyzed will greatly facilitate projects that rely upon aggregation and analysis of public microarray data, such as was presented in Chapter 4 where statistical decision making was evaluated. Beyond purely

statistical studies, with the capabilities of a fully relational database resource such as CEBS, which houses all aspects of toxicological experiments, including genomics, it is possible to more broadly explore the biological implications of public microarray data in the context of public toxicological and biological data. Phenotypic anchoring, where gene expression data is coupled with supporting biological and toxicological data for a range of chemicals, from multiple sources, in an integrated analysis, is the next logical step in the use of public toxico-chemogenomics data. Phenotypic anchoring typically occurs in the context of traditional toxicogenomics data for a single chemical, where gene expression data and supporting toxicological and biological data, such as histopathology and clinical chemistry data from the same experiment are used. However, with the addition of chemical indexing of experiments to create the extended toxico-chemogenomic capabilities, phenotypic anchoring methodologies can be extended to a much broader range of chemically indexed datasets.

Existing methodologies use clustering and/or multivariate and general linear models to couple both the qualitative and quantitative data. With phenotypic anchoring and purely statistical methods applied to toxico-chemogenomics data, there are still several aspects of public microarray data that must be addressed before these data can optimally utilized. How to utilize gene expression results across experiments that may differ with regard to dose, route of administration, and species across labs and experiments, as well as how to account for systematic laboratory and performer bias within a group of experiments, are particularly difficult issues to be resolved. However, the first step in assessing and beginning to grapple with these issues is to have the ability to aggregate and explore data across many dimensions, including, importantly, the chemical dimension in toxico-chemogenomics.

This project represents a first step in the creation of a public toxico-chemogenomic capability in the public domain, and in demonstrating its utility to develop new hypotheses and to explore existing hypotheses. This research was conducted in the context of regulatory science, where the results of this research project can be extended to the support of predictive toxicological modeling efforts. In the future, increased coordination from both ArrayExpress and GEO collaborators, coupled with enhanced linkages between data resources will increase the efficiency of annotation of these resources and enhance public capabilities for data integration in support of toxicogenomics.

APPENDIX A

R SCRIPTS

###Sampling of R Script for FDR BY and BH method s Aventis17 dataset shown.

```
#load libraries
library(affy)
library(gcrma)
library(limma)
library(RColorBrewer)
cols<-brewer.pal(4,"Set3")
library("simpleaffy")
#set work directory to specific data set folder
getwd()
setwd("E:/R_19Nov2008/Aventis_17")
dir()
#load cell files from current working directory
data<-ReadAffy()
#check to make sure complete data set was loaded
data
#open graphical device to store graphical output
pdf("Aventis17graphics.pdf")
#open text file to store output
sink("Aventis17_results.txt")
#Check Data Quality
hist(data,main= "Aventis17 Before Normalization")
boxplot(data, col=cols, main="Aventis17 Before Normalization")
#Background Correction
data.bg<-bg.correct(data,"rma")
#QC Statistics
```

```
data.qc<-qc(data)
#Compute Background Statistics
avbg(data.qc)
#Compute Scale Statistics
sfs(data.qc)
#Percent Present Statistics
percent.present(data.qc)
#Check Quality After Background Adjustment
hist(data.bg, main="Aventis17 After RMA BG Correction")
boxplot(data.bg, col=cols, main="Aventis17 After RMA BG Correction")
#Quantile Normalization
data.rmaNorm<-normalize(data.bg, "quantiles")
#Check Quality After Quantile Normalization
hist(data.rmaNorm, main="Aventis17 After RMA BG Normalization")
boxplot(data.rmaNorm, col=cols, main="Aventis17 After RMA BG Normalization")
#Compute Expression Summaries One Value for Each Gene Represented on Each Chip
data.rmaExpr<-computeExprSet(data.rmaNorm, "pmonly", "medianpolish")
#Compute GCRMA Package Normalization and Summarization
data.gcrmaExpr<-gcrma(data)
#Check final quality
#boxplot(data.gcrmaExpr, main="Aventis17 After Complete RMA Procedure")
#boxplot(data.rmaExpr, col=cols, main="Aventis17 After GCRMA Package")
dev.off()
sink()

####Using RMA
```

```
#make design matrix for data

design<-model.matrix(~-1+factor(c(0,0,0,1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9)))

#assign group names for design matrix

colnames(design)<-c("A17control4", "A17low4", "A17lowmid4", "A17midhigh4",
"A17high4","A17control24", "A17low24", "A17lowmid24", "A17midhigh24", "A17high24")

#fit linear model to data

rma.fit<-lmFit(data.rmaExpr,design)

#make matrix of contrast

contrast.matrix<- makeContrasts(A17control4-A17low4, A17control4-A17lowmid4, A17control4-
A17midhigh4, A17control4-A17high4, A17control24-A17low24, A17control24-A17lowmid24,
A17control24-A17midhigh24, A17control24-A17high24, levels=design)

#compute contrast for linear model

rma.fit2<-contrasts.fit(rma.fit, contrast.matrix)

#Compute Statistics for contrast

rma.fit2<-eBayes(rma.fit2)

#Output Results for RMA

A17rma.result1<-decideTests(rma.fit2, method="separate", adjust.method="none", p.value=0)

write.table(A17rma.result1, "A17RMA.RESULT.NONE.0.txt", sep="\t" )

A17rma.result2<-decideTests(rma.fit2, method="separate", adjust.method="BY", p.value=0)

write.table(A17rma.result2, "A17RMA.RESULT.BY.0.txt", sep="\t")

A17rma.result3<-decideTests(rma.fit2, method="separate", adjust.method="BH", p.value=0)

write.table(A17rma.result3, "A17RMA.RESULT.BH.0.txt", sep="\t")

A17rma.result4<-decideTests(rma.fit2, method="separate", adjust.method="none", p.value=0.01)

write.table(A17rma.result4, "A17RMA.RESULT.NONE.001.txt", sep="\t")

A17rma.result5<-decideTests(rma.fit2, method="separate", adjust.method="BY", p.value=0.01)

write.table(A17rma.result5, "A17RMA.RESULT.BY.001.txt", sep="\t")

A17rma.result6<-decideTests(rma.fit2, method="separate", adjust.method="BH", p.value=0.01)

write.table(A17rma.result6, "A17RMA.RESULT.BH.001.txt", sep="\t")
```

```
A17rma.result7<-decideTests(rma.fit2, method="separate", adjust.method="none", p.value=0.05)
write.table(A17rma.result7, "A17RMA.RESULT.NONE.005.txt", sep="\t")
A17rma.result8<-decideTests(rma.fit2, method="separate", adjust.method="BY", p.value=0.05)
write.table(A17rma.result8, "A17RMA.RESULT.BY.005.txt", sep="\t")
A17rma.result9<-decideTests(rma.fit2, method="separate", adjust.method="BH", p.value=0.05)
write.table(A17rma.result9, "A17RMA.RESULT.BH.005.txt", sep="\t")
A17rma.result10<-decideTests(rma.fit2, method="separate", adjust.method="none", p.value=0.10)
write.table(A17rma.result10, "A17RMA.RESULT.NONE.010.txt", sep="\t")
A17rma.result11<-decideTests(rma.fit2, method="separate", adjust.method="BY", p.value=0.10)
write.table(A17rma.result11, "A17RMA.RESULT.BY.010.txt", sep="\t")
A17rma.result12<-decideTests(rma.fit2, method="separate", adjust.method="BH", p.value=0.10)
write.table(A17rma.result12, "A17RMA.RESULT.BH.010.txt", sep="\t")
A17rma.result13<-decideTests(rma.fit2, method="separate", adjust.method="none", p.value=0.25)
write.table(A17rma.result13, "A17RMA.RESULT.NONE.025.txt", sep="\t")
A17rma.result14<-decideTests(rma.fit2, method="separate", adjust.method="BY", p.value=0.25)
write.table(A17rma.result14, "A17RMA.RESULT.BY.025.txt", sep="\t")
A17rma.result15<-decideTests(rma.fit2, method="separate", adjust.method="BH", p.value=0.25)
write.table(A17rma.result15, "A17RMA.RESULT.BH.025.txt", sep="\t")
####Using GCRMA
#make design matrix for data
design<-model.matrix(~-1+factor(c(0,0,0,1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9)))
#assign group names for design matrix
colnames(design)<-c("A17control4", "A17low4", "A17lowmid4", "A17midhigh4",
"A17high4","A17control24", "A17low24", "A17lowmid24", "A17midhigh24", "A17high24")
#fit linear model to data
gcrma.fit<-lmFit(data.rmaExpr,design)
#make matrix of contrast
```

```
contrast.matrix<- makeContrasts(A17control4-A17low4, A17control4-A17lowmid4, A17control4-
A17midhigh4, A17control4-A17high4, A17control24-A17low4, A17control24-A17lowmid24,
A17control24-A17midhigh24, A17control24-A17high24, levels=design)

#compute contrast for linear model

gcrma.fit2<-contrasts.fit(gcrma.fit, contrast.matrix)

#Compute Statistics for contrast

gcrma.fit2<-eBayes(gcrma.fit2)

#Output Results for GCRMA

A17gcrma.result1<-decideTests(gcrma.fit2, method="separate", adjust.method="none", p.value=0)

write.table(A17gcrma.result1, "A17GCRMA.RESULT.NONE.0.txt", sep="\t")

A17gcrma.result2<-decideTests(gcrma.fit2, method="separate", adjust.method="BY", p.value=0)

write.table(A17gcrma.result2, "A17GCRMA.RESULT.BY.0.txt", sep="\t")

A17gcrma.result3<-decideTests(gcrma.fit2, method="separate", adjust.method="BH", p.value=0)

write.table(A17gcrma.result3, "A17GCRMA.RESULT.BH.0.txt", sep="\t")

A17gcrma.result4<-decideTests(gcrma.fit2, method="separate", adjust.method="none",
p.value=0.01)

write.table(A17gcrma.result4, "A17GCRMA.RESULT.NONE.001.txt", sep="\t")

A17gcrma.result5<-decideTests(gcrma.fit2, method="separate", adjust.method="BY", p.value=0.01)

write.table(A17gcrma.result5, "A17GCRMA.RESULT.BY.001.txt", sep="\t")

A17gcrma.result6<-decideTests(gcrma.fit2, method="separate", adjust.method="BH", p.value=0.01)

write.table(A17gcrma.result6, "A17GCRMA.RESULT.BH.001.txt", sep="\t")

A17gcrma.result7<-decideTests(gcrma.fit2, method="separate", adjust.method="none",
p.value=0.05)

write.table(A17gcrma.result7, "A17GCRMA.RESULT.NONE.005.txt", sep="\t")

A17gcrma.result8<-decideTests(gcrma.fit2, method="separate", adjust.method="BY", p.value=0.05)

write.table(A17gcrma.result8, "A17GCRMA.RESULT.BY.005.txt", sep="\t")

A17gcrma.result9<-decideTests(gcrma.fit2, method="separate", adjust.method="BH", p.value=0.05)

write.table(A17gcrma.result9, "A17GCRMA.RESULT.BH.005.txt", sep="\t")
```

```
A17gcrma.result10<-decideTests(gcrma.fit2, method="separate", adjust.method="none",
p.value=0.10)

write.table(A17gcrma.result10,"A17GCRMA.RESULT.NONE.010.txt", sep="\t")

A17gcrma.result11<-decideTests(gcrma.fit2, method="separate", adjust.method="BY", p.value=0.10)

write.table(A17gcrma.result11,"A17GCRMA.RESULT.BY.010.txt", sep="\t")

A17gcrma.result12<-decideTests(gcrma.fit2, method="separate", adjust.method="BH", p.value=0.10)

write.table(A17gcrma.result12,"A17GCRMA.RESULT.BH.010.txt", sep="\t")

A17gcrma.result13<-decideTests(gcrma.fit2, method="separate", adjust.method="none",
p.value=0.25)

write.table(A17gcrma.result13,"A17GCRMA.RESULT.NONE.025.txt", sep="\t")

A17gcrma.result14<-decideTests(gcrma.fit2, method="separate", adjust.method="BY", p.value=0.25)

write.table(A17gcrma.result14,"A17GCRMA.RESULT.BY.025.txt", sep="\t")

A17gcrma.result15<-decideTests(gcrma.fit2, method="separate", adjust.method="BH", p.value=0.25)

write.table(A17gcrma.result15,"A17GCRMA.RESULT.BH.025.txt", sep="\t")
```

###Sampling of R Script for FWER Holm's method Aventis6 dataset shown.

```
#load libraries

library(affy)

library(gcrma)

library(limma)

library(RColorBrewer)

cols<-brewer.pal(4,"Set3")

library("simpleaffy")

#set work directory to specific data set folder

getwd()

setwd("G:/R_19Nov2008/Aventis_6")

dir()

#load cell files from current working directory
```

```
data<-ReadAffy()
#check to make sure complete data set was loaded
data
#open graphical device to store graphical output
pdf("Aventis6graphics.pdf")
#open text file to store output
sink("Aventis6_results.txt")
#Check Data Quality
hist(data,main= "Aventis6 Before Normalization")
boxplot(data, col=cols, main="Aventis6 Before Normalization")
#Background Correction
data.bg<-bg.correct(data,"rma")
#QC Statistics
data.qc<-qc(data)
#Compute Background Statistics
avbg(data.qc)
#Compute Scale Statistics
sfs(data.qc)
#Percent Present Statistics
percent.present(data.qc)
#Check Quality After Background Adjustment
hist(data.bg, main="Aventis6 After RMA BG Correction")
boxplot(data.bg, col=cols, main="Aventis6 After RMA BG Correction")
#Quantile Normalization
data.rmaNorm<-normalize(data.bg, "quantiles")
#Check Quality After Quantile Normalization
```

```
hist(data.rmaNorm, main="Aventis6 After RMA BG Normalization")

boxplot(data.rmaNorm, col=cols, main="Aventis6 After RMA BG Normalization")

#Compute Expression Summaries One Value for Each Gene Represented on Each Chip

data.rmaExpr<-computeExprSet(data.rmaNorm, "pmonly", "medianpolish")

#Check final quality

#boxplot(data.gcrmaExpr, main="Aventis6 After Complete RMA Procedure")

#boxplot(data.rmaExpr, col=cols, main="Aventis6 After GCRMA Package")

dev.off()

sink()


####Using RMA

#make design matrix for data

design<-model.matrix(~-1+factor(c(0,0,0,0,1,1,1,2,2,3,3,3,4,4,4,5,5,5,5,5,6,6,6,7,7,8,8,8,9,9,9)))

#assign group names for design matrix

colnames(design)<-c("A6control4", "A6low4", "A6lowmid4", "A6midhigh4",
"A6high4","A6control24", "A6low24", "A6lowmid24", "A6midhigh24", "A6high24")

#fit linear model to data

rma.fit<-lmFit(data.rmaExpr,design)

#make matrix of contrast

contrast.matrix<- makeContrasts(A6control4-A6low4, A6control4-A6lowmid4, A6control4-
A6midhigh4, A6control4-A6high4, A6control24-A6low24, A6control24-A6lowmid24, A6control24-
A6midhigh24, A6control24-A6high24, levels=design)

#compute contrast for linear model

rma.fit2<-contrasts.fit(rma.fit, contrast.matrix)

#Compute Statistics for contrast

rma.fit2<-eBayes(rma.fit2)

#Output Results for RMA
```

```
A6rma.result3<-decideTests(rma.fit2, method="hierarchical", adjust.method="holm", p.value=0)
write.table(A6rma.result3, "A6RMA.RESULT.holm.0.txt", sep="\t")
A6rma.result4<-decideTests(rma.fit2, method="hierarchical", adjust.method="holm", p.value=0.01)
write.table(A6rma.result4, "A6RMA.RESULT.holm.001.txt", sep="\t")
A6rma.result9<-decideTests(rma.fit2, method="hierarchical", adjust.method="holm", p.value=0.05)
write.table(A6rma.result9, "A6RMA.RESULT.holm.005.txt", sep="\t")
A6rma.result10<-decideTests(rma.fit2, method="hierarchical", adjust.method="holm", p.value=0.10)
write.table(A6rma.result10, "A6RMA.RESULT.holm.010.txt", sep="\t")
A6rma.result15<-decideTests(rma.fit2, method="hierarchical", adjust.method="holm", p.value=0.25)
write.table(A6rma.result15, "A6RMA.RESULT.holm.025.txt", sep="\t")
dat1<-read.table("A17RMA.RESULT.NONE.0.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.NONE.0.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.NONE.0.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.NONE.0.txt", sep="\t")
##############################################################################
###Sampling of Combiner programs combine data for each analysis combination of normalization type, multiple correction method and alpha level.
dat1<-read.table("A17RMA.RESULT.BY.0.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BY.0.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BY.0.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BY.0.txt", sep="\t")
##############################################################################

dat1<-read.table("A17RMA.RESULT.BH.0.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BH.0.txt", header=TRUE, sep="\t")
```

```
dat3<-read.table("PGRMA.RESULT.BH.0.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BH.0.txt", sep="\t")
##############################################################################
##############################################################################
dat1<-read.table("A17RMA.RESULT.NONE.001.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.NONE.001.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.NONE.001.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.NONE.001.txt", sep="\t")
##############################################################################

dat1<-read.table("A17RMA.RESULT.BY.001.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BY.001.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BY.001.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BY.001.txt", sep="\t")
##############################################################################

dat1<-read.table("A17RMA.RESULT.BH.001.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BH.001.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BH.001.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BH.001.txt", sep="\t")
##############################################################################
##############################################################################
```

```
dat1<-read.table("A17RMA.RESULT.NONE.005.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.NONE.005.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.NONE.005.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.NONE.005.txt", sep="\t")
##############################################################################

dat1<-read.table("A17RMA.RESULT.BY.005.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BY.005.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BY.005.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BY.005.txt", sep="\t")
##############################################################################

dat1<-read.table("A17RMA.RESULT.BH.005.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BH.005.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BH.005.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BH.005.txt", sep="\t")
##############################################################################
##############################################################################
dat1<-read.table("A17RMA.RESULT.NONE.010.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.NONE.010.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.NONE.010.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.NONE.010.txt", sep="\t")
```

```
##############################################################################

dat1<-read.table("A17RMA.RESULT.BY.010.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BY.010.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BY.010.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BY.010.txt", sep="\t")
##############################################################################

dat1<-read.table("A17RMA.RESULT.BH.010.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BH.010.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BH.010.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BH.010.txt", sep="\t")
##############################################################################
##############################################################################
dat1<-read.table("A17RMA.RESULT.NONE.025.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.NONE.025.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.NONE.025.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.NONE.025.txt", sep="\t")
##############################################################################

dat1<-read.table("A17RMA.RESULT.BY.025.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BY.025.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BY.025.txt", header=TRUE, sep="\t")
```

```
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BY.025.txt", sep="\t")
#############################################################################

dat1<-read.table("A17RMA.RESULT.BH.025.txt", header=TRUE, sep="\t")
dat2<-read.table("A6RMA.RESULT.BH.025.txt", header=TRUE, sep="\t")
dat3<-read.table("PGRMA.RESULT.BH.025.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "RMA.BH.025.txt", sep="\t")
#############################################################################
#############################################################################
#*************************************************************************
dat1<-read.table("A17GCRMA.RESULT.NONE.0.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.NONE.0.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.NONE.0.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.NONE.0.txt", sep="\t")
#############################################################################

dat1<-read.table("A17GCRMA.RESULT.BY.0.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BY.0.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BY.0.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BY.0.txt", sep="\t")
#############################################################################
```

```
dat1<-read.table("A17GCRMA.RESULT.BH.0.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BH.0.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BH.0.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BH.0.txt", sep="\t")
#############################################################################
#############################################################################
dat1<-read.table("A17GCRMA.RESULT.NONE.001.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.NONE.001.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.NONE.001.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.NONE.001.txt", sep="\t")
#############################################################################

dat1<-read.table("A17GCRMA.RESULT.BY.001.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BY.001.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BY.001.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BY.001.txt", sep="\t")
#############################################################################

dat1<-read.table("A17GCRMA.RESULT.BH.001.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BH.001.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BH.001.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BH.001.txt", sep="\t")
```

```
##############################################################################
##############################################################################
dat1<-read.table("A17GCRMA.RESULT.NONE.005.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.NONE.005.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.NONE.005.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.NONE.005.txt", sep="\t")
##############################################################################

dat1<-read.table("A17GCRMA.RESULT.BY.005.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BY.005.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BY.005.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BY.005.txt", sep="\t")
##############################################################################

dat1<-read.table("A17GCRMA.RESULT.BH.005.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BH.005.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BH.005.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BH.005.txt", sep="\t")
##############################################################################
##############################################################################
dat1<-read.table("A17GCRMA.RESULT.NONE.010.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.NONE.010.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.NONE.010.txt", header=TRUE, sep="\t")
```

```
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.NONE.010.txt", sep="\t")
##########################################################################

dat1<-read.table("A17GCRMA.RESULT.BY.010.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BY.010.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BY.010.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BY.010.txt", sep="\t")
##########################################################################

dat1<-read.table("A17GCRMA.RESULT.BH.010.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BH.010.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BH.010.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BH.010.txt", sep="\t")
##########################################################################
##########################################################################
dat1<-read.table("A17GCRMA.RESULT.NONE.025.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.NONE.025.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.NONE.025.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.NONE.025.txt", sep="\t")
##########################################################################

dat1<-read.table("A17GCRMA.RESULT.BY.025.txt", header=TRUE, sep="\t")
```

```
dat2<-read.table("A6GCRMA.RESULT.BY.025.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BY.025.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BY.025.txt", sep="\t")
##############################################################################

dat1<-read.table("A17GCRMA.RESULT.BH.025.txt", header=TRUE, sep="\t")
dat2<-read.table("A6GCRMA.RESULT.BH.025.txt", header=TRUE, sep="\t")
dat3<-read.table("PGGCRMA.RESULT.BH.025.txt", header=TRUE, sep="\t")
combined<-cbind(dat1, dat2, dat3)
write.table(combined, "GCRMA.BH.025.txt", sep="\t")
##############################################################################
##############################################################################
```