Designing toxicogenomics to support decision making in environmental toxicology

Jessica D. Ewald

Department of Natural Resource Sciences

McGill University, Montreal

March 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

TABLE OF CONTENTS

LIST OF TABLES V
LIST OF FIGURES VI
ABSTRACT IX
RÉSUMÉ XII
ACKNOWLEDGEMENTS XV
CONTRIBUTION TO ORIGINAL KNOWLEDGE XVII
PREFACE AND CONTRIBUTION OF THE AUTHORS XX
LIST OF ABBREVIATIONS XXIV

CHAPTER 1. GENERAL INTRODUCTION 1

1.1 KNOWLEDGE GAP 1
1.2 OVERALL OBJECTIVE 2
1.3 SPECIFIC AIMS 3

CHAPTER 2. LITERATURE REVIEW 5

2.1 LITERATURE REVIEW 5
2.1.1 Motivation for alternative toxicity testing methods 5
2.1.2 Existing analytical methods for toxicogenomics data 9
2.1.3 Toxicogenomics for ecological risk assessment 13
2.1.4 Outlook for toxicogenomics data use in chemical risk assessment 16
2.2 REFERENCES 19
CONNECTING PARAGRAPH 31

CHAPTER 3. A COMPUTATIONAL ECOSYSTEM FOR RNA-SEQ DATA FROM NON-MODEL ORGANISMS 33

3.1 ABSTRACT 33
3.2 INTRODUCTION 34
3.3 RESULTS 36
3.3.1 Expanded ortholog databases in Seq2Fun version 2.0 37
3.3.2 EcoOmicsDB 40
3.3.3 Case Studies 41
3.4 DISCUSSION 51
3.5 METHODS 52
3.5.1 Ortholog group definition and annotation 52
3.5.2 Other updates to Seq2Fun algorithm 53
3.5.3 Methods for case studies #1 and #2 53
3.5.4 Implementation of web tools 56
3.6 ACKNOWLEDGEMENTS 57
3.7 REFERENCES 57
CONNECTING PARAGRAPH 62

CHAPTER 4. ECOTOXMODULES: CUSTOM GENE SETS TO ORGANIZE AND ANALYZE TOXICOGENOMICS DATA FROM ECOLOGICAL SPECIES 63

4.1 ABSTRACT 64
4.2 INTRODUCTION 65

4.3 MATERIALS AND METHODS 67
4.3.1 Phase I: Data preparation and processing........ 69
4.3.2 Phase II: EcoToxModule definition and characterization........ 70
4.3.3 Phase III: Comparison of EcoToxModule-based analysis to traditional bioinformatics 73
4.4 RESULTS AND DISCUSSION........ 75
4.4.1 Re-annotation of microarray probe sequences........ 75
4.4.2 General comparison of the statistical and functional EcoToxModules........ 76
4.4.3 Case study 1: DES exposures (GSE36465 and GSE36466) 80
4.4.4 Case study 2: WWTP effluent exposures (GSE49098) 82
4.4.5 Statistical EcoToxModules can be used to identify potential biomarkers........ 84
4.4.6 EcoToxModules as tools to help improve the interpretation of toxicogenomics data 87
4.5 SUPPORTING INFORMATION 89
4.6 ACKNOWLEDGEMENTS........ 90
4.7 REFERENCES 90
CONNECTING PARAGRAPH........ 102

CHAPTER 5. FASTBMD: AN ONLINE TOOL FOR RAPID BENCHMARK DOSE-RESPONSE ANALYSIS OF TRANSCRIPTOMICS DATA........ 103

5.1 ABSTRACT 103
5.2 INTRODUCTION 104
5.3 IMPLEMENTATION........ 105
5.4 MATERIALS AND METHODS 106
5.5 RESULTS AND DISCUSSION........ 107
5.6 ACKNOWLEDGEMENTS........ 108
5.7 REFERENCES 108
CONNECTING PARAGRAPH........ 110

CHAPTER 6. ECOTOXXPLORER: LEVERAGING DESIGN THINKING TO DEVELOP A STANDARDIZED WEB-BASED TRANSCRIPTOMICS ANALYTICS PLATFORM FOR DIVERSE USERS 112

6.1 TRANSCRIPTOMICS DATA ARE TRANSFORMATIONAL YET CHALLENGING 113
6.2 DEMOCRATIZING ‘OMICS DATA VIA WEB-BASED VISUAL ANALYTICS........ 114
6.3 OBJECTIVE 117
6.4 DESIGN THINKING 119
6.5 OVERVIEW OF ECOTOXXPLORER 122
6.6 NAVIGATING ECOTOXCHIP ANALYSIS........ 125
6.7 CONCLUDING REMARKS 132
6.8 REFERENCES 133

CHAPTER 7. GENERAL DISCUSSION........ 136

7.1 TOXICOGENOMICS TO SUPPORT DECISION-MAKING........ 136
7.2 DESIGN-THINKING IN CHAPTERS 3-6........ 138
7.3 ADDITIONAL OBSTACLES FACING TOXICOGENOMICS FOR REGULATORY DECISION-MAKING........ 139
7.4 STRENGTHS AND LIMITATIONS OF THIS THESIS........ 140

7.4.1 Limitations *140*
7.4.2 Strengths *142*

CHAPTER 8. GENERAL CONCLUSION 145

MASTER REFERENCE LIST........ 146
APPENDIX: SUPPLEMENTARY INFORMATION FOR CHAPTERS 3-6........ 151
SI FOR CHAPTER 3 151
SI FOR CHAPTER 4 160
SI FOR CHAPTER 5 170
SI FOR CHAPTER 6 183

LIST OF TABLES

Table 2-1: Methods for transcriptomic dose-response analysis. These papers were found by searching “transcriptomic*” AND “dose-response” in Web of Science, yielding 94 results. From this list, the papers with a primary focus of describing a new method or software for transcriptomic dose-response modeling were read and summarized... 11

Table 2-2: Summary of genome annotation resources for selected vertebrate species. In NCBI, there are genome assemblies (genomic DNA – accession ID starts with GCA) and genome annotations (location of features on the genome – accession ID starts with GCF). For species with no annotation report, the genome assembly ID is reported. 15

Table 3-1: Expanded ortholog databases for Seq2Fun version 2.0... 39

Table 3-2: Results from zebrafish DEA and GSA.. 43

Table 3-3: Results from lobster DEA and GSA.. 45

Table 3-4: Top 5 DEGs from case study #2.. 49

Table 4-1: Overview of publicly available microarray data analyzed in Phases II and III.......... 70

Table 6-1: A list of web-based tools for ‘omics data analytics that team members have developed. The user statistics are based on Google Analytics (retrieved in May 2021). The metabolomics and microbiome data analysis tools are species agnostic and applicable to both model and non-model species. The transcriptomics data analysis tools are species dependent, though NetworkAnalyst, FastBMD, and Seq2Fun can support generic species based on KEGG orthologs as well as custom annotations. OmicsAnalyst can integrate different ‘omics data in a species independent manner. 115

Table 6-2: Key features identified through design thinking activities. These design features are described in more detail in the following sections... 122

LIST OF FIGURES

Figure 1-1: The knowledge gap addressed by this thesis is that traditional toxicogenomics results are complex, detailed, and not easily integrated with decision-making processes. The objective of this thesis is to develop statistical methods and software that make analysis and interpretation of toxicogenomics data simple and accessible, particularly for ecological species. 2

Figure 2-1: Timeline of selected key events from the last 30 years related to environmental risk assessment, alternative toxicity testing, and toxicogenomics data. Events are divided into two groups based on whether they were primarily driven by regulatory actions or 'omics/toxicogenomics research. 6

Figure 2-2: a) Comparison of traditional and alternative toxicity testing methods. Traditional toxicity testing relies on exposing whole organisms to chemicals and then observing apical outcomes. Alternative methods aim to predict apical outcomes from measurements at the molecular level. b) Main objectives of whole transcriptome toxicogenomics data analysis. Gene expression is one of the molecular level effects that can be measured within the alternative toxicity testing framework – others include protein and metabolite levels. These methods can be applied to the whole transcriptome, or to a reduced set of genes that adequately cover the biological space of interest........ 8

Figure 3-1: Overview of the computational ecosystem and steps for a typical workflow. 37

Figure 3-2: A) Histogram of the number of genes mapping to each ortholog in the "eukaryotes" database. B) Scatterplot of the number of species and the number of orthologs in a database. ... 40

Figure 3-3: PCA of normalized zebrafish counts tables for a) Kallisto and b) Seq2Fun. This plot was generated by ExpressAnalyst........ 42

Figure 3-4: PCA of normalized lobster counts tables for a) Kallisto and b) Seq2Fun. This plot was generated by ExpressAnalyst........ 45

Figure 3-5: PCA of normalized counts tables for the salamander data, with samples annotated by A) time since amputation, and B) species. This plot was generated by ExpressAnalyst. 48

Figure 4-1: Graphical abstract published in manuscript........ 64

Figure 4-2: A visual representation of the different phases of the analysis workflow. 68

Figure 4-3: Co-expression network dendrogram showing the locations of the EcoToxModule probes. Each leaf in the dendrogram corresponds to one probe. Probes were hierarchically clustered based on their pairwise correlation scores. The first band displays the locations of the clusters detected in the dendrogram, including both predicted coding and non-coding probes. The statistical EcoToxModules consist of the coding probes within the detected clusters. Clusters with a consistent biological theme in their gene set analysis results are highlighted. The second band displays the locations of the probes from the functional EcoToxModules. 76

Figure 4-4: Heatmaps of EcoToxModule eigengene expression for all exposure conditions. Rows correspond to EcoToxModules, columns correspond to the exposure conditions, and squares correspond to the eigengene expression. The eigengene is calculated as the 1st principal component of log2FC of module probes for A: statistical EcoToxModules, and B: functional EcoToxModules. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. P-val < 0.05, FDR). 79

Figure 4-5: EcoToxModule fold changes for case study 1 (DES exposures). EcoToxModule fold changes are calculated as the mean of the abs(log2FC) of all probes in the module for a given exposure condition. Grey bars correspond to the low (1 ng/L), medium (10 ng/L), and high (100 ng/L) exposures to DES and green bars correspond to the same low, medium, and high exposures after a 4 day recovery period, for the A) functional EcoToxModules and B) statistical EcoToxModules. Data are from the GSE36465 and GSE36466 data series. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR). 81

Figure 4-6: EcoToxModule fold changes for case study 2 (WWTP exposures). EcoToxModule fold changes are calculated as the mean of the abs(log2FC) of all probes in the module for a given exposure condition. Bars correspond to exposure to water samples collected upstream of (green), downstream of (grey), and at the effluent discharge point (black) of the WWTP in Ely, Minnesota, USA. Data are from the GSE49098 data series. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR). 83

Figure 4-7: Statistical EcoToxModules that are potential exposure biomarkers. Exposure conditions (*n* = 38) were divided based on whether they had high estrogenic potency (*n* = 8, DES and E2 exposures, red fill) or medium-low estrogenic potency (*n* = 30, all other exposures, blue fill). Fold change, calculated as the mean of abs(log2FC), for each A) statistical EcoToxModules and B) probe within the SM9 EcoToxModule. Statistically significant differences (adj. p-val < 0.05) between the two groups were assessed using the Wilcoxon test for each module and probe. There are 66 probes in SM9, 23 of which correspond to coding sequences. P-values were adjusted using the Bonferroni method. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR). 85

Figure 5-1: FastBMD performance. A) Snapshot of interactive gene, pathway, and transcriptomic BMD visualization in FastBMD for the TRBZ (13 weeks) dataset. B) Comparison of elapsed time during curve-fitting between FastBMD (yellow) and BMDExpress (grey) for 24 datasets. C) OmicBMDs from FastBMD and BMDExpress for 24 datasets. 106

Figure 6-1: Design thinking components and activities therein that underpinned the development and evaluation of EcoToxXplorer. 121

Figure 6-2: Screenshots of EcoToxXplorer's interface design. A) Home page with links to FAQs, tutorials, user accounts and other resources, along with entry points into the three main modules: 1) EcoToxChip Analysis; 2) RNAseq Expression Profiling; and 3) Raw RNAseq

Processing. B) For the EcoToxChip Analysis module, three example datasets are provided to help the user learn how to navigate EcoToxXplorer. C) A navigation panel is provided to inform the user of their current processing step. 124

Figure 6-3: Data processing workflow of EcoToxXplorer. All major steps are listed with key aspects summarized. 125

Figure 7-1: Overview of barriers to toxicogenomics that were addressed by thesis chapters 3 – 6. 137

ABSTRACT

Traditional toxicity testing methods rely on exposing whole organisms to chemicals and observing high-level responses such as mortality and reproduction. These methods are too slow, expensive, and ethically concerning to assess the tens of thousands of legacy and novel substances in need of testing, and since they measure only a small number of endpoints, do not provide much biological insight into the toxicity mechanism. Over the last fifteen years, there has been a growing push to develop new approach methods for toxicity testing that do not use animal exposures. One major objective of this movement is to conduct exposures *in vitro*, measure comprehensive molecular outcomes, and use the molecular data to predict and manage risk to whole organisms. This not only promises to make toxicity testing faster, less expensive, and more humane, but also promises to generate more informative data. Toxicogenomics, the measurement of ‘omics data such as transcriptomics, proteomics, and metabolomics in the context of toxicology, is key to realizing this goal. However, toxicogenomics data is extremely complex and the lack of computing resources, programming skills, advanced statistical training, and knowledge of bioinformatics databases presents barriers to many researchers and regulators. The barriers are particularly pronounced for environmental toxicologists using ecologically relevant species, as there are few bioinformatics resources outside of a small number of model organisms.

The objective of this thesis is to design new statistical methods and corresponding software for analyzing and visualizing toxicogenomics data to support decision-making in the context of environmental toxicology. Many of the additional barriers to using transcriptomics data in ecologically relevant non-model organisms are related to raw data processing and annotation.

Chapter 3 presents a set of computational tools (EcoOmicsAnalyst, ExpressAnalyst, EcoOmicsDB) for producing and analyzing annotated counts tables from raw RNA-seq data from any species, regardless of whether there is a reference genome. Traditional transcriptomics results such as lists of impacted genes and pathways are difficult to integrate into regulatory decision-making processes. Chapter 4 presents EcoToxModules, custom gene sets for summarizing and communicating transcriptomics data that are focused on toxicologically relevant biological processes. Chapter 5 presents FastBMD, software for performing rapid transcriptomics dose-response modeling. Since dose-response results such as benchmark dose values and points-of-departure are already familiar to the toxicology community, this type of analysis is useful because it translates unfamiliar toxicogenomics data into the familiar dose-response framework. Finally, while the cost of acquiring whole-transcriptome data has decreased tremendously over the last few decades, it is still outside the scope of many research programs. EcoToxChips are qPCR arrays with 384 genes for six ecologically relevant species that address this issue because qPCR technology is cost-effective with widespread availability. Chapter 6 presents EcoToxXplorer, software focused on EcoToxChip data processing, analysis, and interpretation.

Together, the chapters in this thesis aim to support the use of toxicogenomics data in decision-making processes by making it more usable and understandable to individual members of the toxicology community, while also enabling standardized workflows that can be easily accessed by many different people in different locations. One important aspect of this is that all software presented in this thesis are web-based, which means that they do not require users to have substantial computing resources or programming skills, and do not require local installation.

Throughout statistical method and software development, a design-thinking framework was used to continuously obtain and incorporate feedback from a large group of stakeholders and potential end-users from academia, government, and industry.

RÉSUMÉ

Les méthodes traditionnelles de test de toxicité reposent sur l'exposition d'organismes entiers à des produits chimiques et sur l'observation de réactions de haut niveau telles que la mortalité et la reproduction. Ces méthodes sont trop lentes, coûteuses et problématiques sur un point de vue éthique pour évaluer les dizaines de milliers de substances héritées et nouvelles qui doivent être testées. De plus, comme elles ne mesurent qu'un petit nombre de paramètres, elles ne fournissent pas beaucoup d'informations biologiques sur le mécanisme de toxicité. Au cours des quinze dernières années, il y a de plus en plus de pressions pour mettre au point de nouvelles méthodes d'analyse de la toxicité qui n'utilisent pas l'expositions des animaux. L'un des principaux objectifs de ce mouvement est de réaliser des expositions « in vitro », de mesurer des résultats moléculaires complets et d'utiliser les données moléculaires pour prévoir et gérer les risques pour les organismes entiers. Cela promet non seulement de rendre les tests de toxicité plus rapides, moins coûteux et plus humains, mais aussi de générer des données plus informatives. La toxicogénomique, c'est-à-dire la mesure des données « omiques » telles que la transcriptomique, la protéomique et la métabolomique dans le contexte de la toxicologie, est essentielle à la réalisation de cet objectif. Cependant, les données toxicogénomiques sont extrêmement complexes et le manque de ressources informatiques, de compétences en programmation, de formation statistique avancée et de connaissances des bases de données bioinformatiques présente des obstacles pour de nombreux chercheurs et régulateurs qui souhaitent utiliser ce type de données. Ces obstacles sont particulièrement prononcés pour les toxicologues environnementaux qui utilisent des espèces écologiquement pertinentes, car il existe peu de ressources bioinformatiques en dehors d'un petit nombre d'organismes modèles.

L'objectif de cette thèse est de concevoir de nouvelles méthodes statistiques et des logiciels correspondants pour analyser et visualiser les données toxicogénomiques afin de soutenir la prise de décision dans le contexte de la toxicologie environnementale. Un grand nombre d’obstacles supplémentaires à l'utilisation des données transcriptomiques chez les organismes non-modèles écologiquement pertinents sont liés au traitement et à l'annotation des données brutes. Le chapitre 3 présente un ensemble d'outils informatiques (EcoOmicsAnalyst, ExpressAnalyst, EcoOmicsDB) permettant de produire et d’analyser des tableaux de comptage annotés à partir de données brutes d'ARN-seq provenant de n'importe quelle espèce, qu'il existe ou non un génome de référence. Les résultats traditionnels de la transcriptomiques, tels que les listes de DEG et de voies impactées, sont difficiles à intégrer dans les processus décisionnels réglementaires. Le chapitre 4 présente les EcoToxModules, des ensembles de gènes personnalisés pour résumer et communiquer des données transcriptomiques axées sur des processus biologiques pertinents sur le plan toxicologique. Le chapitre 5 présente FastBMD, un logiciel permettant de réaliser rapidement une modélisation transcriptomique dose-réponse. Étant donné que les résultats dose-réponse, tels que les valeurs de dose de référence et les points de départ, sont déjà familiers à la communauté toxicologique, ce type d'analyse est utile car il traduit des données toxicogénomiques inconnues dans le cadre dose-réponse familier. Enfin, bien que le coût d'acquisition des données du transcriptome entier ait considérablement diminué au cours des dernières décennies, il n'entre toujours pas dans le cadre de nombreux programmes de recherche. Les EcoToxChips sont des matrices qPCR comprenant 384 gènes pour six espèces écologiquement pertinentes qui répondent à ce problème car la technologie qPCR est rentable et largement disponible. Le chapitre 6 présente EcoToxXplorer, un logiciel axé sur le traitement, l'analyse et l'interprétation des données EcoToxChip.

Ensemble, les chapitres de cette thèse visent à soutenir l'utilisation des données toxicogénomiques dans les processus décisionnels en les rendant plus utilisables et compréhensibles pour les membres individuels de la communauté toxicologique, tout en permettant des flux de travail standardisés qui peuvent être facilement accessibles par de nombreuses personnes différentes dans différents endroits. Un aspect important de cela est que tous les logiciels présentés dans cette thèse sont basés sur le web, ce qui signifie qu'ils ne nécessitent pas que les utilisateurs disposent de ressources informatiques importantes ou de compétences en programmation, et ne nécessitent pas d'installation locale. Tout au long du développement de la méthode statistique et du logiciel, un cadre de conception a été utilisé pour obtenir et intégrer continuellement les commentaires d'un grand groupe d’intervenants et d'utilisateurs finaux potentiels du milieu universitaire, du gouvernement et de l'industrie.

ACKNOWLEDGEMENTS

This thesis involved many people other than me, and I am very thankful to each of them for their part.

First, thanks to my supervisor, Nil. I learned so much from you over the last five years, not just about toxicology and environmental science, but also how to communicate with broad audiences, manage people and projects, and write extremely efficient emails. Looking back on my early presentations and proposals, I can see how huge of an impact you've had on my ability to distill chaotic and complex thoughts into clear phrases and figures (and flowcharts). I also admire how you work in so many different areas of the environmental/public health sciences and how you are continually working to connect people across disciplines.

To Jeff, thank you for going above and beyond what it means to be a committee member. I did not expect to leave my PhD as a software engineer. Working with you and the Xia Lab has been one of the best surprises of the last few years. To Doug, thank you for always making me feel appreciated and valued, and for giving such carefully thought-out feedback, at committee meetings, on manuscripts, and for presentations. Having the three of you as advisors meant that I always looked forward to committee meetings – they ended leaving me feeling supported and not grilled. Thanks also to Jess, Markus, and Natacha. You were never official advisors, but I always felt like I could come to you with any problems that I had, and I learned a lot from all of you.

I feel like I have been part of four different "labs": Basu Lab, Xia Lab, Head Lab, and the EcoToxChip project team. Thanks to everyone in these groups for the meals, conferences, meetings, social outings, coffee breaks, and work sessions. You made my graduate experience infinitely richer. A special thanks to Othman, Peng, Guangyan, Krittika, Emily, Elena, YeonSeon, Derek, and Alper. I spent many hours working with you in the office and over Slack, Zoom, email, and GoToMeeting, and there are huge sections of my thesis that would not exist without you. There are so many others on MacDonald Campus that I feel fortunate to count among my friends – trivia in the Ceilidh, "community dinners" in Ste. Anne, movie nights, bootcamp, running club, MacWIS … my PhD was much more fun because of you.

Finally, thank you to my family: Mom, Dad, Brenda, Dave, Phyl, Lindsay, Gideon, Renee, Aaron, Levi, Adriene, Roland, Pip, Malcolm, and Jesiah. Mom and Dad, thank you for being my first teachers, for raising me to be an independent problem solver, and for passing on a love of plants, animals, and ecosystems. To you and Dave, Brenda, and Phyl – thank you for opening up your homes to me and Jesiah after March 2020. Spending so much more time with family was the biggest silver lining of the pandemic. Also, I'm sorry for the number of times that I invaded your morning coffee with my Zoom meetings – we made it through with only a few "bathrobe behind the webcam" incidents ☺ Last but not least – thank you, Jesiah, for being my partner in all of this. I calculated it out, and you cooked me (and half the grad students in Ste Anne) over 1500 meals since I started my PhD. Moving across the country with me, making me laugh before every stressful event, continually encouraging me to sleep, exercise, and socialize, pushing me to be assertive, being up for anything post-PhD … I couldn't have done this without you, and I am so excited for the future.

CONTRIBUTION TO ORIGINAL KNOWLEDGE

This thesis developed statistical methods and software for processing and analyzing toxicogenomics data from ecological species. Here, novel aspects of the methods and software are highlighted.

- **First web-based software for processing raw RNA-seq data from any eukaryotic species without a reference genome.** The Seq2Fun algorithm was previously published, however developing a web-based platform to support it is still a significant amount of work. FASTQ files are typically between 1-3 GB each and engineering a stable system that can handle simultaneous upload, processing, and storage for multiple users requires balancing bandwidth consumption, memory usage, and storage considerations. My contributions to this software also involved designing the ortholog databases, evaluating their coverage and relevance, demonstrating consistency with traditional methods, adapting software for downstream analysis, and developing the database (EcoOmicsDB) for ortholog evidence lookup. This software (EcoOmicsAnalyst, ExpressAnalyst, and EcoOmicsDB) reduces barriers to extracting value from raw RNA-seq data from non-model organisms, making transcriptomics data analysis accessible to research programs with no prior bioinformatics experience.

- **Demonstration of Seq2Fun for comparative transcriptomics.** In chapter 3, I demonstrate how Seq2Fun makes comparing transcriptomics results across species extremely easy through a case study with three different salamander species. Seq2Fun still recovers the same functional results as obtained by a reference and two *de novo* transcriptomes, but since sequences from all three species can be mapped to the same ID space, it removes the need for extensive ortholog mapping and allows the data to be integrated throughout the entire analysis. This opens up possibilities for much more extensive comparisons of transcriptomics data across more species.

- **First gene set library focused on toxicology for non-model organisms (EcoToxModules).** Many custom gene sets have been developed for model organisms[1] however to my knowledge there are no open-source gene sets available for toxicology research in ecological species. While Ingenuity Pathway Analysis does have Tox Lists and Tox Functions[2], these are only defined for a few model organisms. By focusing on more general biological processes that are familiar to many within the toxicology community, EcoToxModules make overall transcriptomics results understandable by people who have no prior experience with bioinformatics results. This is useful for communicating toxicogenomics results to diverse audiences.

- **Fastest transcriptomics dose-response method that follows the NTP recommended approach.** FastBMD uses previously developed methods for curve fitting in R that have proven to be very fast and adapted them to work with the statistical models recommended by the US National Toxicology Program. In addition, I developed my own method for calculating asymmetric confidence intervals to derive the BMD_l and BMD_u that is significantly faster than unrestricted likelihood profiling. These speed improvements

[1] https://maayanlab.cloud/Enrichr/#libraries

[2] https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/features/tox-lists-tox-functions/

enable hosting FastBMD online, which improves accessibility.

- **Most comprehensive software for toxicogenomics analysis.** To my knowledge, EcoToxXplorer is the only software available that covers raw data QA/QC, processing and normalization, differential analysis, integration with tox-focused knowledgebases, and visual analytics. qPCR is an affordable technology that many research programs already have the instruments and expertise to collect. By providing a standardized processing pipeline and tools for interpretation, EcoToxXplorer contributes towards the effort of making toxicogenomics analysis accessible for all ecotoxicology research programs.

- **Overall, provides user-friendly access to standardized toxicogenomics analysis pipelines.** Together, the software in this thesis provides web-based, programming-free access to standardized toxicogenomics pipelines for transcriptomics data for many different species. This includes raw data processing, differential expression analysis with tox-focused functional analysis, and transcriptomics dose-response analysis. The goal of relying on *in vitro* exposures and other new approach methods for next-generation toxicity testing requires that big molecular data be integrated into the activities of research programs and government agencies worldwide. Software that makes it easy for non-bioinformaticians to draw biological insights from molecular data will help establish toxicogenomics as routine technology, moving the community closer to the aforementioned goal.

PREFACE AND CONTRIBUTION OF THE AUTHORS

This thesis is composed of four original research chapters (Chapters 3-6), each of which have been prepared for submission, submitted to, and/or published in an academic journal. I am the sole first author of two of the manuscripts (Chapters 4 and 5) and co-first author of the other two (Chapters 3 and 6). The other co-first authors were post-doctoral fellows at the time the manuscripts were submitted, and as such, are not contributing the works to a thesis. Details on author CRediT contributions are outlined below.

Manuscript 1. A computational ecosystem for RNA-seq data from non-model organisms. This manuscript is being prepared for submission to *Nature Methods* in March 2022.

Peng Liu *: conceptualization, methodology, software, validation, formal analysis, data curation, writing – original draft, writing – review & editing, visualization

Jessica Ewald *: conceptualization, methodology, software, validation, formal analysis, data curation, writing – original draft, writing – review & editing, visualization

Orcun Hacariz: methodology, software, resources

Elena Legrand: validation, data curation, formal analysis, writing – review & editing

Guangyan Zhou: software, resources

Benjamin de Jourdan: methodology, data collection

Jessica Head: writing – review & editing, supervision, funding acquisition

Niladri Basu: conceptualization, writing – review & editing, supervision, funding acquisition

Jianguo Xia: conceptualization, software, resources, writing – review & editing, supervision, funding acquisition

* These authors contributed equally to the work.

Manuscript 2. EcoToxModules: Custom gene sets to organize and analyze toxicogenomics data from ecological species. This manuscript was published in *Environmental Science and Technology* on February 28, 2020 (volume 54 (7), pages 4376-4387).

Jessica Ewald: conceptualization, methodology, software, validation, formal analysis, data curation, writing – original draft, writing – review & editing, visualization

Othman Soufan: conceptualization, methodology, data curation, writing – review & editing

Doug Crump: conceptualization, data curation, writing – review & editing, funding acquisition

Markus Hecker: conceptualization, data curation, writing – review & editing, funding acquisition

Jianguo Xia: conceptualization, resources, writing – review & editing, funding acquisition

Niladri Basu: conceptualization, methodology, data curation, writing – review & editing, supervision, funding acquisition

Manuscript 3. FastBMD: an online tool for rapid benchmark dose-response analysis of transcriptomics data. This manuscript was published in *Bioinformatics* on August 6, 2020 (volume 37 (7), pages 1035-1036).

Jessica Ewald: conceptualization, methodology, software, validation, data curation, writing – original draft, writing – review & editing, visualization,

Othman Soufan: conceptualization, software, writing – review & editing

Jianguo Xia: conceptualization, software, resources, writing – review & editing, supervision, funding acquisition

Niladri Basu: conceptualization, writing – review & editing, supervision, funding acquisition

Manuscript 4. EcoToxXplorer: Leveraging Design Thinking to Develop a Standardized Web-Based Transcriptomics Analytics Platform for Diverse Users. This manuscript was published in *Environmental Toxicology and Chemistry* on November 11, 2021 (volume 41 (1), pages 21-29). Othman Soufan (led development from fall 2017 – summer 2019) and Jessica Ewald (led development from summer 2019 - present) are co-first authors.

Othman Soufan*: conceptualization, methodology, software, validation, data curation, writing – original draft, writing – review & editing, visualization

Jessica Ewald*: conceptualization, methodology, software, validation, data curation, writing – original draft, writing – review & editing, visualization

Guangyan Zhou: methodology, software, writing – review & editing

Orcun Hacariz: methodology, software, data curation, writing – review & editing

Emily Boulanger: methodology, software, validation, data collection, writing – review & editing

Alper James Alcaraz: methodology, software, validation, data curation, writing – review & editing

Gordon Hickey: conceptualization, resources, writing – review & editing, funding acquisition

Steve Maguire: conceptualization, resources, writing – review & editing, funding acquisition

Guillaume Pain: conceptualization, writing – review & editing

Natacha Hogan: conceptualization, methodology, validation, resources, writing – review & editing, funding acquisition

Markus Hecker: conceptualization, methodology, validation, resources, writing – review & editing, funding acquisition

Doug Crump: conceptualization, methodology, validation, resources, writing – review & editing, funding acquisition

Jessica Head: conceptualization, methodology, validation, resources, writing – review & editing, funding acquisition

Niladri Basu*: conceptualization, methodology, validation, resources, writing – original draft, writing – review & editing, visualization, supervision, project administration, funding acquisition

Jianguo Xia*: conceptualization, methodology, software, resources, writing – original draft, writing – review & editing, supervision, project administration, funding acquisition

* These authors contributed equally to the work.

LIST OF ABBREVIATIONS

Abbreviation	Full spelling
AOP	Adverse outcome pathway
AND	*Ambystomatidae andersoni*
ANOVA	Analysis of variance
BLAST	Basic local alignment search tool
BMD	Benchmark dose
BMD_t	Transcriptomics benchmark dose
CEPA	Canadian Environmental Protection Act
CPU	Central processing unit
Ct	Cycle threshold
DEA	Differential expression analysis
DEG	Differentially expressed gene
DES	Diethylstilbestrol
DNA	Deoxyribonucleic acid
DRA	Dose-response analysis
EC	Effect concentration
ECCC	Environment and Climate Change Canada
ECHA	European Chemicals Agency
EDC	Endocrine-disrupting chemicals
EPA	Environmental Protection Agency
ERA	Ecological Risk Assessment
E2	Estrogen
FAQ	Frequently Asked Question
FDR	False discovery rate
GB	Gigabyte
GEO	Gene Expression Omnibus
GDC	Genomic DNA contamination
GO	Gene Ontology
GO BP	GO biological process
GO CC	GO cellular component
GO MF	GO molecular function
GSA	Gene set analysis
GSEA	Gene set enrichment analysis
KAAS	KEGG Automatic Annotation Server
KB	Kilobyte
KE	Key event
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Ortholog

Log2FC	Log2(fold-change)
MAC	*Ambystomatidae maculatum*
MB	Megabyte
MEX	*Ambystomatidae mexicanum*
MIAME	Minimum information about a microarray experiment
MIE	Molecular initiating event
MOA	Mode of action
NAMs	New Approach Methodologies
NASEM	National Academies of Science, Engineering, and Medicine
NCBI	National Center for Biotechnology Information
NOTEL	No observed transcriptional effect level
NRC	National Research Council
NSERC	National Science and Engineering Research Council of Canada
NTP	National Toxicology Program
OBS	Sodium p-perfluorous nonenoxybenzene sulfonate
OECD	Organisation for Economic Co-operation and Development
ORA	Overrepresentation analysis
PCA	Principal components analysis
PFC	Perfluorochemicals
PFOS	Perfluorooctane sulfonate
PHN	Phenanthrene
POD	Point of departure
PPAR	Peroxisome proliferator-activated receptor
PPC	Positive PCR control
qPCR	Quantitative polymerase chain reaction
RAM	Random access memory
RDX	1,3,5-Trinitroperhydro-1,3,5-triazine
REACH	Registration, Evaluation, Authorisation, and restriction of CHemicals
ROS	Reactive oxygen species
RTC	Reverse transcription control
SETAC	Society for Environmental Toxicology and Chemistry
SRA	Sequence read archive
TRF	Transcriptomics reporting framework
TSCA	Toxic Substances Control Act
UI/UX	User interface and user experience
WAF	Water accommodated fraction
WGCNA	Weighted gene co-expression network analysis
WHO	World Health Organization
WWTP	Wastewater treatment plant
3Rs	Replacement, Reduction, and Refinement

CHAPTER 1. GENERAL INTRODUCTION

1.1 KNOWLEDGE GAP

Traditional methods for analyzing gene expression data have contributed to the generation of new biological and toxicological knowledge, but the results they produce are not easily integrated into regulatory decision-making processes (Figure 1-1) (Villeneuve and Garcia-Reyero 2011). One factor is that practical applications of toxicogenomics data requires integration of knowledge from three different disciplines: bioinformatics, toxicology, and regulatory decision-making (NASEM, 2017). Proper analysis requires advanced programming skills, statistical training, and knowledge of bioinformatics resources (Ankley *et al.*, 2006; Fent and Sumpter, 2011), appropriate interpretation requires deep knowledge of toxicological mechanisms, and real-world application requires familiarity with and connections to regulatory processes. Developing the capacity to extract value from toxicogenomics data requires significant investments of time and money (Balbus and Environmental Defense, 2005; ECETOC 2007). Thus, there is an urgent need for simplified statistical methods and user-friendly software that make toxicogenomics analysis, interpretation, and application more accessible to more people. In 2020, the US EPA identified increased investment in software and databases as one of four main objectives in their five-year blueprint for computational toxicology (Thomas *et al.*, 2019). While this call is for tools to support both human and ecological risk assessment, it is particularly important to develop solutions for ecologically relevant species, given the relative paucity of genomics resources compared to mammalian model organisms.

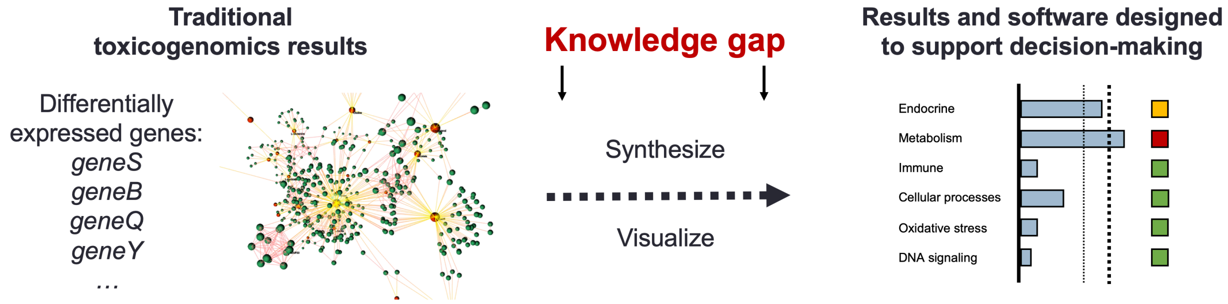


Figure 1-1: The knowledge gap addressed by this thesis is that traditional toxicogenomics results are complex, detailed, and not easily integrated with decision-making processes. The objective of this thesis is to develop statistical methods and software that make analysis and interpretation of toxicogenomics data simple and accessible, particularly for ecological species.

1.2 OVERALL OBJECTIVE

The overall objective of this thesis is to design new statistical methods and corresponding software for analyzing and visualizing toxicogenomics data to support decision-making in the context of environmental toxicology. *Statistical methods* refer to bioinformatics methods expressed as R functions and *corresponding software* refers to web-based interfaces that make the statistical methods developed in this thesis, as well as other popular R packages and command-line software, accessible to a wide audience through careful design of user-friendly interfaces. The focus on *supporting decision-making* means that the design of the new methods and software will take an adopter-centric view by utilizing the principles of design-thinking, aiming to make toxicogenomics workflows simple and compatible with existing regulatory processes. The goal of doing this work in the *context of environmental toxicology* means that this thesis will analyze data from and design software for ecologically relevant species wherever possible. In some cases, this thesis presents data from mammalian model organisms, however this was only done when a comparable dataset could not be found from an ecologically relevant species.

1.3 Specific aims

Chapter 3: Develop web-based tools for quantifying, analyzing, and interpreting raw RNA-seq reads from any eukaryotic species, regardless of whether it has a reference transcriptome.

- Build web-based tools for the following tasks:
    - EcoOmicsAnalyst for raw RNA-seq data quantification (Kallisto for species with a reference transcriptome, Seq2Fun for those without)
    - ExpressAnalyst for statistical analysis of RNA-seq counts tables
    - EcoOmicsDB for retrieving detailed information on the Seq2Fun ortholog groups
- Update the Seq2Fun algorithm by improving the efficiency of the core algorithm and expanding the underlaying protein ortholog database
- Demonstrate key workflows with two case studies

Chapter 4: Develop custom gene sets for organizing and analyzing toxicogenomics data from ecologically relevant species

- Define two types of gene sets:
    - Statistical EcoToxModules, using co-expression analysis of a large microarray dataset
    - Functional EcoToxModules, using known molecular pathways from the KEGG database that are organized by ecotoxicology experts
- Evaluate the ability of both EcoToxModule sets to capture trends in differential expression analysis results
- Demonstrate how EcoToxModules can be used to interpret toxicogenomics data through two case studies

Chapter 5: Develop a web-based software for transcriptomics dose-response analysis

- Develop a set of R functions to perform transcriptomic dose-response analysis, following the US National Toxicology Program's recommended methods
- Develop FastBMD, a web-interface to make transcriptomic dose-response analysis accessible to more people
- Evaluate FastBMD by performing transcriptomic dose-response analysis on 24 datasets using both FastBMD and BMDExpress, another software, and comparing the performance

Chapter 6: Develop a web-based platform for analyzing EcoToxChip data in the context of environmental toxicology

- Design a standardized pipeline for QA/QC, normalization, and differential expression analysis of EcoToxChip qPCR data
- Curate toxicology-focused knowledgebases for interpretation of EcoToxChip results
    - Manually expand EcoToxModule annotations of EcoToxChip genes
    - Pull and curate Adverse Outcome Pathway annotations of EcoToxChip genes from the AOP wiki
- Develop visualizations and an automatic analysis report to aid in communication of EcoToxChip results

CHAPTER 2. LITERATURE REVIEW

2.1 LITERATURE REVIEW

2.1.1 MOTIVATION FOR ALTERNATIVE TOXICITY TESTING METHODS

Assessing the potential toxic properties of chemicals and released wastes is an essential part of ecological risk assessments (ERAs) that governments and businesses conduct worldwide (Embry *et al*. 2014; Pastoor *et al.* 2014). Toxicity must be assessed in a range of contexts, including setting standards for new and legacy chemicals, ensuring regulatory compliance of treated wastewater, and prioritizing and monitoring the remediation of contaminated sites. Exposure to chemicals at environmentally relevant levels has been linked to adverse outcomes of regulatory concern in ecological species (Cizmas *et al.,* 2015; Guigueno and Fernie, 2017; WHO, 2012). Additionally, pollution has been estimated to be responsible for 16% of premature human deaths worldwide (Landrigan *et al.* 2018). Thus, assessing and responding to the risks posed by contaminant exposure is vital for protecting both human and ecological health.

Traditional toxicity testing methods cannot meet the need to assess the enormous number of untested chemicals because they rely on exposing thousands of animals and measuring apical outcomes, which is expensive, time consuming, and of ethical concern (Andersen and Krewski 2008; *Burden et al.,* 2015). In traditional chemical risk assessment, animals are exposed to a pure chemical at varying concentrations according to an established test method and endpoints of interest are measured (Embry *et al.*, 2014; Pastoor *et al.*, 2014). Dose-response models are fit to the observations to determine the benchmark dose (BMD), which is used to establish an exposure level with an acceptable amount of risk. The BMD is extrapolated to protect other species, life stages, and vulnerable sub-populations using uncertainty or safety factors (Celander *et al*. 2011; Forbes *et al.* 2001). In comparison to human health risk assessment, ERAs face the additional challenges of assessing risk to many species, for both pure chemicals and the complex mixtures of contaminants found in the environment.

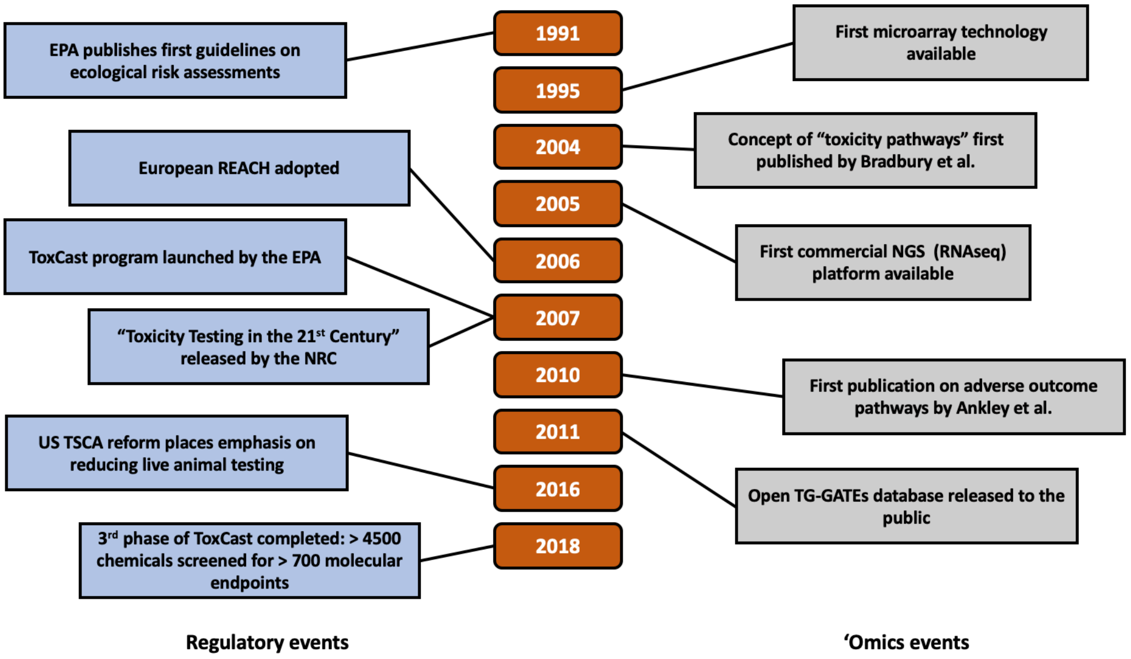

EPA publishes first guidelines on ecological risk assessments
1991
First microarray technology available
1995
European REACH adopted
2004
Concept of "toxicity pathways" first published by Bradbury et al.
2005
ToxCast program launched by the EPA
2006
First commercial NGS (RNAseq) platform available
"Toxicity Testing in the 21st Century" released by the NRC
2007
2010
First publication on adverse outcome pathways by Ankley et al.
US TSCA reform places emphasis on reducing live animal testing
2011
2016
Open TG-GATEs database released to the public
3rd phase of ToxCast completed: > 4500 chemicals screened for > 700 molecular endpoints
2018
Regulatory events
'Omics events

Figure 2-1: Timeline of selected key events from the last 30 years related to environmental risk assessment, alternative toxicity testing, and toxicogenomics data. Events are divided into two groups based on whether they were primarily driven by regulatory actions or 'omics/toxicogenomics research.
Note – EPA: Environmental Protection Agency; NGS: Next generation sequencing; REACH: Registration, Evaluation, Authorisation, and Restriction of Chemicals; TSCA: Toxic Substances Control Act; NRC: National Research Council

In 2007, the U.S. National Research Council (NRC) proposed a new toxicity testing paradigm that would shift to in vitro exposures and harness systems biology to assess toxicity from molecular outcomes, resulting in a large reduction of live animal exposures (Figure 2-1) (Krewski *et al.* 2007). This new paradigm is organized around the idea that any organism-level adverse outcome, also called an apical outcome, is downstream of some initial molecular or cellular-level perturbation (Figure 2-2a) (Villeneuve and Garcia-Reyero 2011). Non-animal-based approaches to inform chemical hazard and risk assessment, also called new approach methodologies (NAMs), have gained a foothold in the 15 years since the seminal NRC report was published (ECHA, 2016).

Increasing societal pressure to both eliminate whole-animal testing for new chemicals and to evaluate the immense number of data-poor legacy chemicals has resulted in legislation that supports NAMs, first in the European Union and followed by the US and Canada. Key pieces of legislation are the *Registration, Evaluation, Authorisation, and Restriction of Chemicals* (REACH) in the European Union (managed by the European Chemicals Agency (ECHA)), the *Canadian Environmental Protection Act* (CEPA) in Canada, and the *Toxic Substances Control Act* (TSCA) in the United States (van der Vegt *et al.*, 2021). REACH, first issued in 2007, was amended in 2017 to require that all toxicity testing use alternative methods unless animal-based tests are the only option (ECHA, 2020). In the US, an amendment to TSCA in 2016 included a directive to reduce reliance on animal testing (US Congress, 2016), and a memorandum issued in 2019 by the US EPA Director made this more explicit by committing to reduce funding for mammal studies by 30% by 2025 and to eliminate them by 2035, replacing them with NAMs (Wheeler, 2019). In 2021 in Canada, Bill C-28 was introduced to amend the preamble to CEPA to recognize the "importance of promoting the development and timely incorporation of scientifically justified alternative methods and strategies in the testing and assessment of substances" (Bill C-28, 2021). Thus, in addition to the general motivation to improve chemical management, researchers and government agencies are now under significant regulatory pressure to develop NAMs for toxicity testing.

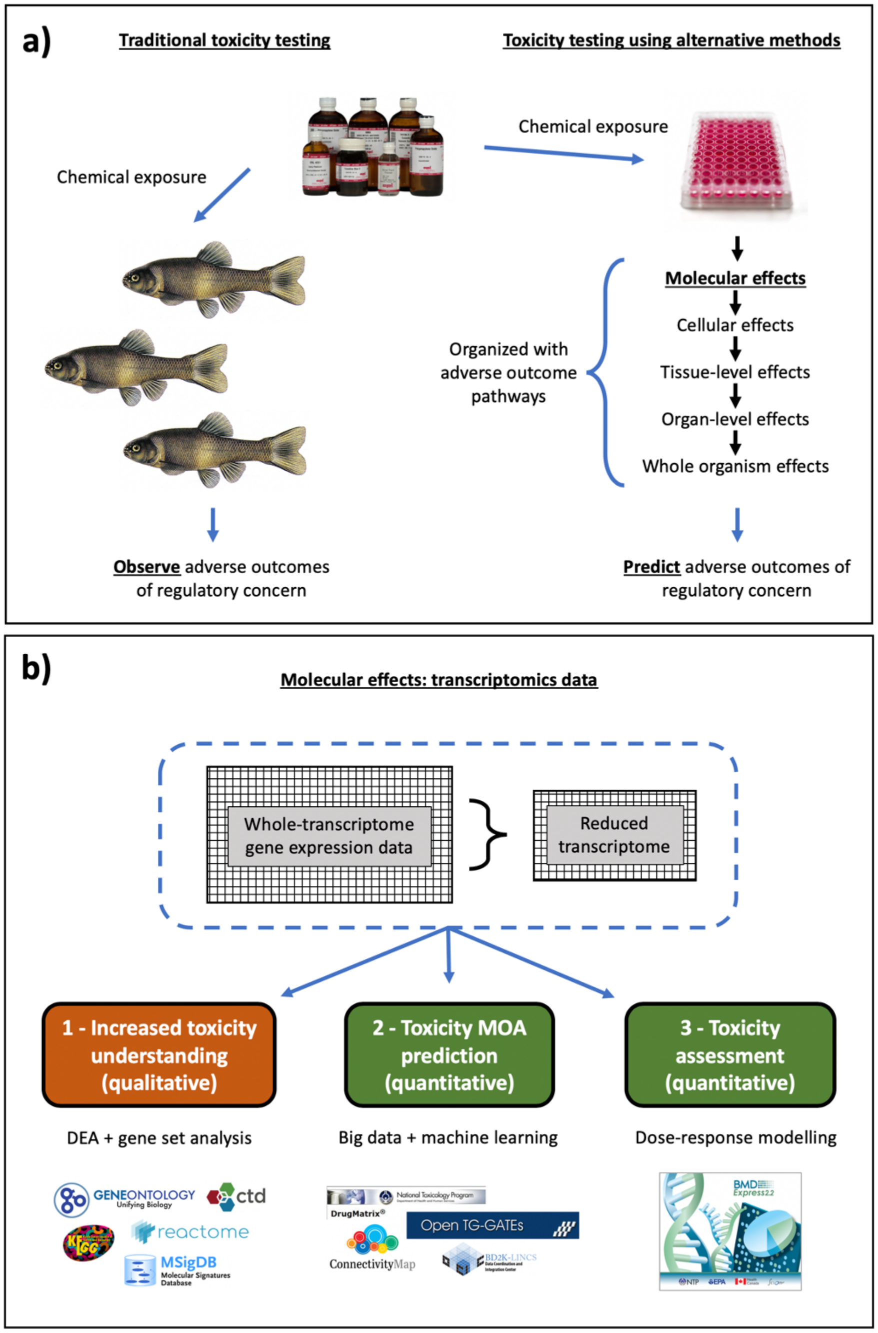

a)
Traditional toxicity testing
Toxicity testing using alternative methods
Chemical exposure
Chemical exposure
Molecular effects
Cellular effects
Tissue-level effects
Organ-level effects
Whole organism effects
Organized with adverse outcome pathways
Observe adverse outcomes of regulatory concern
Predict adverse outcomes of regulatory concern
b)
Molecular effects: transcriptomics data
Whole-transcriptome gene expression data
Reduced transcriptome
1 - Increased toxicity understanding (qualitative)
2 - Toxicity MOA prediction (quantitative)
3 - Toxicity assessment (quantitative)
DEA + gene set analysis
Big data + machine learning
Dose-response modelling
GENEONTOLOGY Unifying Biology
ctd
KEGG
reactome
MSigDB Molecular Signatures Database
National Toxicology Program
DrugMatrix
Open TG-GATEs
ConnectivityMap
BD2K-LINCS
BMDExpress2.2

Figure 2-2: a) Comparison of traditional and alternative toxicity testing methods. Traditional toxicity testing relies on exposing whole organisms to chemicals and then observing apical outcomes. Alternative methods aim to predict apical outcomes from measurements at the molecular level. b) Main objectives of whole transcriptome toxicogenomics data analysis. Gene expression is one of the molecular level effects that can be measured within the alternative toxicity testing framework – others include protein and metabolite levels. These methods can be applied to the whole transcriptome, or to a reduced set of genes that adequately cover the biological space of interest.
Note – MOA: mode of action; DEA: differential expression analysis.

2.1.2 Existing analytical methods for toxicogenomics data

Molecular endpoints collected after *in vitro* exposures promise to play a key part in the elimination of live animal testing (Andersen and Krewski 2008). The long-term objective is to predict apical outcomes in whole organisms from molecular outcomes measured *in vitro* (Burden *et al*. 2015; Dix *et al*. 2007). Thus, current research efforts aim to link molecular-level effects to downstream tissue and organism-level effects through organizing frameworks such as adverse outcome pathways (AOPs) (Villeneuve *et al*. 2014). These efforts rely heavily on the use of ‘omics technologies, which enable the comprehensive measurement of molecular profiles within a biological assay. ‘Omics data are broken down into more specific categories, for example transcriptomics for gene expression, metabolomics for metabolites, and proteomics for proteins. Toxicogenomics refers to the use of ‘omics technologies within the context of toxicology (Aardema and MacGregor 2003; Brockmeier *et al*. 2017).

The majority of toxicogenomics research has been conducted with transcriptomics data, partly because of early promises that gene expression profiling would revolutionize toxicity testing (Fields and Zacherewski, 2001). Transcriptomics data are collected using technologies such as microarrays and RNA sequencing that measure the expression of tens of thousands of genes in a single experiment. Most genes are not included in known toxicity pathways, for example there were only 173 genes in the AOP wiki as of 2018 (Wang *et al.*, 2018; Davis *et al.*, 2020). Thus, toxicogenomics research includes both efforts to expand resources such as the AOP wiki with focused mechanistic studies, and to develop statistical methods that link gene expression to apical outcomes or specific chemical exposures without focusing on mechanistic explanations

(Alexander-Dann *et al*. 2018). Existing statistical methods can be broadly sorted into three categories based their outcomes (Figure 2-2b):

1 - Increased toxicity understanding: Traditional bioinformatics methods can be applied to toxicogenomics data to investigate the effects of a toxic exposure. These methods use differential expression analysis (DEA) to identify genes perturbed by the experimental conditions, followed by functional analysis to identify gene sets that are overrepresented in the list of differentially expressed genes (DEGs). The resulting lists of DEGs and enriched pathways are difficult to compare quantitatively across exposures, requiring qualitative interpretation by expert toxicologists to increase understanding of toxicity mechanisms (Ankley *et al*. 2016; Cote *et al*. 2016).

2 - Toxicity MOA prediction: A recent review by Alexander-Dann *et al*. summarized existing methods for predicting compound-specific MOAs from gene expression data (Alexander-Dann *et al*. 2018). Briefly, the reviewed methods apply network and gene signature-based machine learning techniques to transcriptomics data to classify compounds based on their MOA (Alexander-Dann *et al*. 2018). This requires very large databases that contain gene expression data collected after exposure to hundreds or thousands of different compounds and corresponding phenotypic responses, such as DrugMatrix (Lea *et al*. 2016) and TG-GATEs (Igarashi *et al.* 2014). These databases are typically derived from mammalian organisms, focused on pharmaceutical exposures, and generated by large consortia.

3 - Toxicity assessment: Methods for applying dose-response modelling to gene expression data to quantitatively assess toxicity have been under development since 2003 (Table 2-1). These methods are of great interest because they can be used to compute transcriptomic versions of dose-response statistics such as BMDs and PODs. These statistics are familiar to regulators and already integrated into the risk assessment process. The most popular method for transcriptomic dose-response modelling in Table 2-2 is implemented in BMDExpress (Yang *et al*. 2007), and is the officially recommended method by the EPA, Health Canada, and Environment and Climate Change Canada (Farmahin *et al*. 2017). One advantage of this method is that it summarizes BMDs at the pathway level, which makes the results easier to interpret (Dean *et al*. 2017) and it has been shown to improve the prediction of apical outcomes (Farmahin *et al*. 2017; Thomas *et al.* 2012). The disadvantage is that by only analyzing genes in existing pathways, the coverage of biological space is limited, especially for genomes from ecological species that typically have sparse or no functional annotation.

Table 2-1: Methods for transcriptomic dose-response analysis. These papers were found by searching "transcriptomic*" AND "dose-response" in Web of Science, yielding 94 results. From this list, the papers with a primary focus of describing a new method or software for transcriptomic dose-response modeling were read and summarized.
Note – EC: effect concentration; NOTEL: no observable transcriptional effect level; 3P/4P: 3 or 4 parameters; POD: point of departure; GSEA: gene set enrichment analysis; BMD$_t$s: transcriptional benchmark dose.

Study description	**Software**	**Reference**
Developed software (BMDExpress) for transcriptomic dose-response analysis by adapting methods in the EPA BMDS software, a tool for dose-response analysis of apical endpoints. Considers linear, 2nd degree and 3rd degree polynomial, and power models for curve fitting.	Java application	(Yang *et al.*, 2007; Phillips *et al.*, 2019)
Developed a method to calculate an EC based on best fit, considering linear, Hill, exponential, Gauss-probit, and log-Gauss-probit models. Range of models chosen to cover both monotonic and biphasic dose-response curves. Adapted from Smetanová *et al.*	R package and web-based tool	(Larras *et al.*, 2018)

Developed a method to calculate the NOTEL based on a novel algorithm that uses a one-class classifier. The model is trained on controls in the TG-GATEs database and uses anomaly detection to determine whether an exposure of interest has significant transcriptional activity.	R script	(Quercioli *et al.*, 2018)
Measured the S1500+ reduced transcriptome using targeted RNA sequencing for multiple doses of four chemicals. Used best fit of constant, 3P-Hill, 4P-Hill, and Gain-Loss models to calculate the POD.	R script	(House *et al.*, 2017)
Integrates previous R packages for clustering, differential expression analysis, and curve fitting into one R package. Restricted to monotonic curves and does not include any pathway analysis.	R package	(Otava *et al.*, 2017)
GSEA was first performed to identify key pathways perturbed by a chemical exposure. BMD$_t$s were then calculated using the significantly enriched gene sets.	R script (used BMDExpress)	(Dean *et al.*, 2017)
Tested eleven different methods of identifying which genes should be used to calculate the BMD$_t$ for a specific chemical. Methods were evaluated based on concordance between resulting BMD$_t$s and BMDs from apical outcomes.	R script (used BMDExpress)	(Farmahin *et al.*, 2017)
Describes BMDExpress Data Viewer, a web-based tool for enhanced visualization of BMDExpress result files.	Web-based tool	(Kuo *et al.*, 2016)
ECs computed based on gene expression by choosing best fit from linear, linear-log, exponential, Michaelis-Menten, Hill, Weibull I, Weibull II, Gaussian-log, and Gaussian models.	R script	(Smetanová *et al.*, 2015)
Developed the ToxResponse Modeler to calculate transcriptomic POD from best fit of exponential, linear, Gaussian, quadratic, and sigmoidal models. Uses particle swarm optimization and iterative curve fitting to determine the best fit.	Java application	(Burgoon and Zacharewski *et al.*, 2008)
Described the dose-response analysis methodology implemented in BMDExpress. Demonstrated that summarizing BMD$_t$s at the pathway level provides mechanistic insight into dose-response statistics computed using apical outcomes.	R script	(Thomas *et al.*, 2007)
Genes are selected that increase monotonically according to either dose or time, based on order-restricted inference methodology.	R script	(Peddada *et al.*, 2003)

Overall, methods for the quantitative analysis of whole transcriptome toxicogenomics data are focused on summarizing the complex data as a smaller number of endpoints so that it is easier to compare gene expression profiles from different exposures. In doing so, these statistical methods attempt to achieve two main objectives. The first is to analyze and summarize any significant changes in gene expression following an exposure, which requires adequate coverage of the entire transcriptome. The second is to produce results that have an intuitive functional interpretation, so that the changes in gene expression can be related to mechanisms of toxicity or apical outcomes. Since only part of the genome is functionally annotated, there is an inherent trade-off between these two objectives. Analytical methods that use more of the transcriptome are more difficult to interpret because they include a higher number of genes with unknown function, while methods that focus on known toxicity pathways have limited transcriptome coverage.

2.1.3 Toxicogenomics for ecological risk assessment

Compared to chemical risk assessments that are done to protect human health, ecological risk assessments (ERAs) are more complex because they aim protect the health of ecosystems, which includes the hundreds to thousands of diverse taxa that live in them. Historically, this has been done by performing chemical exposures on a handful of model organisms that represent different taxonomic groups (Hope, 2006). Mammalian model organisms, for example mice and rats, are supplemented with non-mammalian species such as fathead minnow, zebrafish, chicken, Japanese quail, African clawed frog, roundworms, house flies, and green algae (LaLone *et al.*, 2021). These model organisms have been frequently used because of their compatibility with laboratory research, however in many cases they may not be representative of the specific ecosystems that ERAs aim to protect. For example, zebrafish, a tropical species, may not be representative of fish species found within the sub-arctic and arctic ecosystems of northern Canada. The need to extrapolate across species adds additional complexities to the methods described in sections 2.1.1 and 2.1.2. However, given the daunting magnitude of data required, it is even more important to harness NAMs that decrease cost, time, and live animal use for ERAs.

Toxicogenomics data do promise advantages that are specific to ERAs (Van Aggelen *et al.*, 2010; Basu *et al.*, 2019). One example is using ‘omics data to improve cross-species extrapolation (Celander *et al.*, 2010). Genetic backgrounds and activities of key toxicity pathways can be compared across species to develop a more nuanced understanding of the “taxonomic domain of applicability” of model organisms (LaLone, 2021). This will help to design better testing approaches by highlighting limitations of existing model organisms for specific ecosystems. Thus, in addition to the reasons outlined in section 2.1.1 and 2.1.2, there is great interest in developing toxicogenomics methods and resources for ecological species.

Developing toxicogenomics methods for ecological species is more difficult than for popular mammalian organisms due to the lack of genomes, functional annotations, and large, high-quality datasets (Primmer *et al*. 2013; Pruitt *et al*. 2007). Table 2-2 summarizes genome assembly and functional annotation resources for five vertebrate species frequently used as model organisms in ‘omics studies and six ecological species that are included in the EcoToxChip project, which is one of the largest toxicogenomics resources emerging from the ecological side (Basu *et al*. 2019). The genome annotation reports for non-mammalian species have undergone fewer releases (if they exist at all) and the actual genes have many fewer functional annotations (GO and KEGG). Additionally, there are no databases of comparable size and quality to TG-GATEs, DrugMatrix, and LINCS Connectivity Map for non-mammalian and/or ecological vertebrate species (Wang *et al.* 2016). Thus, existing machine learning methods that require large datasets for model training and testing, or that rely heavily on known toxicity pathways and gene sets, currently have limited relevance for ERAs.

Table 2-2: Summary of genome annotation resources for selected vertebrate species. In NCBI, there are genome assemblies (genomic DNA – accession ID starts with GCA) and genome annotations (location of features on the genome – accession ID starts with GCF). For species with no annotation report, the genome assembly ID is reported.

Species	**NCBI genome accession**	**# protein coding genes**	**# GO terms[1]**	**# KEGG pathways[2]**
Homo sapiens (human)	GCF_000001405 (release 38)	20 203	517 279	330
Mus musculus (mouse)	GCF_000001635 (release 24)	22 493	404 888	326
Rattus norvegicus (rat)	GCF_000001895 (release 5)	23 347	441 304	326
Danio rerio (zebrafish)	GCF_000002035 (release 6)	26 522	224 788	167
Gallus gallus (chicken)	GCF_000002315 (release 5)	17 477	139 687	168
Xenopus laevis (African clawed frog)	GCF_001663975 (release 1)	31 434	11 137	167
Oncorhynchus mykiss (rainbow trout)	GCF_002163495 (release 1)	42 884	2636	-
Coturnix japonica (Japanese quail)	GCF_001577835 (release 2)	16 057	458	168
Pimephales promelas (fathead minnow)	GCA_000700825 (release 1)	No NCBI annotation report	20	-
Phalacrocorax auritus (double-crested cormorant)	GCA_002173455 (release 1)	No NCBI annotation report	-	-
Lithobates pipiens (leopard frog)	Genome never sequenced	No NCBI annotation report	-	-

[1] - http://amigo.geneontology.org/amigo/search/annotation
[2] - https://www.genome.jp/kegg/pathway.html

When high-quality reference genomes do not exist, researchers must either A) align their raw reads to a reference genome from a different species, B) assemble their own genome, or C) assemble a *de novo* transcriptome. Option A is typically not preferred unless there is an extremely closely related species that has a reference genome. Option B is an immense amount of work, for example the ongoing effort by the US EPA to produce a high-quality reference genome for fathead minnow has taken a team of researchers more than 5 years to complete (Burns *et al.*, 2015; Martinson *et al.*, 2021). Thus, most researchers in the environmental life sciences currently use option C to analyze sequencing data for species that do not have a reference genome. *De novo* transcriptome assembly involves piecing together putative transcript sequences directly from RNA-seq data itself, and then aligning all reads to the assembled transcripts to quantify their expression across samples. This process requires extensive computational resources, multiple command-line software, advanced programming skills, and days to weeks of runtime (Martin *et al.*, 2011; Volshall *et al.*, 2018). The results are difficult to analyze – *de novo* transcriptomes often contain several hundred thousand transcripts (compared to 15-40k for a high-quality genome), the majority of which remain uncharacterized even after functional annotation with software like Blast2GO (Conesa and Gotz, 2008).

Thus, the lack of high-quality genome resources, pathway libraries, and accessible software solutions are immediate barriers to the widespread collection, processing, and interpretation of toxicogenomics data for ecologically relevant species.

2.1.4 Outlook for toxicogenomics data use in chemical risk assessment

Researchers have been investigating and promoting toxicogenomics data since at least 1999 (Nuwaysir *et al.*, 1999), and the field remains an area of active research. However, how much of this effort has been translated into actual use of toxicogenomics data by regulatory agencies in decision-making processes? In 2018, Health Canada published a report that evaluated their use of toxicogenomics data (Cheung *et al.*, 2018). They found that while some regulatory bureaus used toxicogenomics data within a weight-of-evidence approach to support MOA characterization, toxicogenomics are not currently well-established for decision-making. This is consistent with other findings that while many advances have been made in toxicogenomics, thus far, practical applications have not lived up to early projections (Cote *et al.*, 2016; Leung, 2017).

In a recent study, Pain *et al.* analyzed 56 publications and reports on the adoption of toxicogenomics for chemical risk assessment to identify drivers and obstacles related to this goal (Pain *et al.*, 2020). They found that toxicogenomics development is motivated by the drivers of *superior scientific understanding*, *new applications*, and *reduced cost and increased efficiency*. On the other hand, practical applications have been held back by the obstacles of *insufficient validation*, *complexity of interpretation*, and *lack of standardization*. Through further analysis, they derived the key insight that proponents of toxicogenomics data tend to have an "innovation-centric" perspective. Research often focuses on novelty (in both 'omics types and statistical methods) while largely ignoring the very practical needs of regulators. Pain *et al.* concludes with a call for people in the field of toxicogenomics to take on an "adopter-centric" perspective by focusing on the consistent calls by regulators for simple applications of toxicogenomics data that are consistent with existing decision-making processes (Pain *et al.*, 2020).

One tool that could help make toxicogenomics research more 'adopter-centric' is design-thinking. Design thinking is a framework for developing innovative solutions for complex and open-ended problems that has gained popularity in engineering, business and management, information technology, and education circles (Dorst, 2011; Jensen and Steinert, 2016). At its core, design-thinking focuses the design process on the people who have the problem that is trying to be solved. While there are many slightly different versions of design-thinking (Tschimmel, 2012), most contain these five main phases:

1. *Empathize*: understand users through activities such as interviews, shadowing, literature reviews, and the construction of user journeys.
2. *Define*: synthesize all observations from the empathize phase into the core problem that needs to be solved.
3. *Ideate*: generate potential solutions to the problem definition from the define phase. The goal here is to brainstorm many solutions, hopefully including innovative ideas.
4. *Prototype*: select the best few proposed solutions from the ideate phase and build a rapid prototype.
5. *Test*: evaluate solutions from the ideate phase by testing out prototypes with real users.

Another key characteristic of design-thinking is that it should be highly iterative (Tschimmel, 2012). After gaining insight into the strengths and weaknesses of solutions in the test phase, problem solvers may improve the prototype and then re-test, gradually improving the solution over time. They may go back to the ideate phase if the prototyped solutions were insufficient, or even all the way back to the empathize and define phases if testing revealed that the problem was not well-understood. The adoption of toxicogenomics requires bringing together diverse stakeholders, complicated technologies, and disparate fields of research (NASEM, 2017). Design-thinking, with its focus on user perspectives and iteration, could be well-suited for organizing the innovation of new solutions in this space. The need for involving actual end-users and other stakeholders during NAMs development has been recognized by initiatives such as the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) working groups, as laid out in their strategic roadmap for evaluating NAMs (ICCVAM, 2018).

2.2 REFERENCES

Aardema MJ, MacGregor JT. **2003**. Toxicology and genetic toxicology in the new era of "toxicogenomics": Impact of "-omics" technologies. In: *Toxicogenomics:Springer*, 171-193.

Alexander-Dann B, Pruteanu LL, Oerton E, Sharma N, Berindan-Neagoe I, Módos D, *et al*. **2018**. Developments in toxicogenomics: Understanding and predicting compound-induced toxicity from gene expression data. *Molecular Omics* 14:218-236.

Andersen ME, Krewski D. **2008**. Toxicity testing in the 21st century: Bringing the vision to life. *Toxicological Sciences* 107:324-330.

Ankley G, Daston GP, Degitz SJ, Denslow ND, Hoke RA, Kennedy SW, *et al.* **2006**. Toxicogenomics in Regulatory Ecotoxicology. *Environmental Science and Technology* 40(13):4055–4065

Ankley G, Escher BI, Hartung T, Shah I. **2016**. Pathway-based approaches for environmental monitoring and risk assessment. *Environmental Science and Technology* 50:10295-10296.

Balbus JM, Environmental Defense. **2005**. *Toxicogenomics: Harnessing the Power of New Technology*. New York, NY: Environmental Defense.

Basu N, Crump D, Head J, Hickey G, Hogan N, Maguire S, *et al.* **2019**. EcoToxChip: A next-generation toxicogenomics tool for chemical prioritization and environmental management. *Environmental Toxicology and Chemistry* 38:279.

Bill C-28, **2021**. *An Act to amend the Canadian Environmental Protection Act, 1999, to make related amendments to the Food and Drugs Act and to repeal the Perfluorooctane Sulfonate Virtual Elimination Act.* 2nd session, 43rd parliament, 2021.

Bourdon-Lacombe JA, Moffat ID, Deveau M, Husain M, Auerbach S, Krewski D, *et al.* **2015**. Technical guide for applications of gene expression profiling in human health risk assessment of environmental chemicals. *Regulatory Toxicology and Pharmacology* 72:292-309.

Brockmeier EK, Hodges G, Hutchinson TH, Butler E, Hecker M, Tollefsen KE, *et al*. **2017**. The role of omics in the application of adverse outcome pathways for chemical risk assessment. *Toxicological Sciences* 158:252-262.

Burden N, Mahony C, Müller BP, Terry C, Westmoreland C, Kimber I. **2015**. Aligning the 3Rs with new paradigms in the safety assessment of chemicals. *Toxicology* 330:62-66.

Burgoon LD, Zacharewski TR. **2008**. Automated quantitative dose-response modeling and point of departure determination for large toxicogenomic and high-throughput screening data sets. *Toxicological Sciences* 104:412-418.

Burns, F. R., Cogburn, A. L., Ankley, G. T., Villeneuve, D. L., Waits, E., Chang, Y. J., *et al.* **2016**. Sequencing and *de novo* draft assemblies of a fathead minnow (*Pimephales promelas*) reference genome. *Environmental Toxicology and Chemistry*, 35(1), 212-217.

Celander MC, Goldstone JV, Denslow ND, Iguchi T, Kille P, Meyerhoff RD, *et al.* **2011**. Species extrapolation for the 21st century. *Environmental Toxicology and Chemistry* 30:52-63.

Cheung, C., Jones-McLean, E., Yauk, C., Barton-Maclaren, T., Boucher, S., Bourdon-Lacombe, J., *et al.* **2018**. Evaluation of the Use of Toxicogenomics in Risk Assessment at Health Canada. *Health Canada,* Ottawa, Canada.

Cizmas L, Sharma VK, Gray CM, McDonald TJ. **2015**. Pharmaceuticals and personal care products in waters: Occurrence, toxicity, and risk. *Environmental Chemistry Letters* 13:381-394.

Conesa A, Götz S. **2008**. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 619832.

Cote I, Andersen ME, Ankley GT, Barone S, Birnbaum LS, Boekelheide K, *et al.* **2016**. The next generation of risk assessment multi-year study: Highlights of findings, applications to risk assessment, and future directions. *Environmental Health Perspectives* 124:1671-1682.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., Mattingly, C. J. **2021**. Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Research*, 49(D1), D1138-D1143.

Dean JL, Zhao QJ, Lambert JC, Hawkins BS, Thomas RS, Wesselkamper SC. **2017**. Application of gene set enrichment analysis for identification of chemically induced, biologically relevant transcriptomic networks and potential utilization in human health risk assessment. *Toxicological Sciences* 157:85-99.

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. **2007**. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences* 95:5-12.

Dorst, K. **2011**. The core of 'design thinking' and its application. *Design Studies*, 32(6), 521-532.

ECETOC. **2007**. *Workshop on the Application of Omic Technologies in Toxicology and Ecotoxicology: Case Studies and Risk Assessment*, 6-7 December 2007, Malaga. Brussels, Belgium: ECETOC.

ECHA. **2020**. *The use of alternatives to testing animals for the REACH regulation*. https://echa.europa.eu/documents/10162/0/alternatives_test_animals_2020_en.pdf/db66b8a3-00af-6856-ef96-5ccc5ae11026?t=1605088405285 Accessed 11 January 2022.

ECHA. **2016**. *New Approach Methodologies in Regulatory Science.* https://echa.europa.eu/documents/10162/21838212/scientific_ws_proceedings_en.pdf/a2087434-0407-4705-9057-95d9c2c2cc57 Accessed 11 January 2022.

Embry MR, Bachman AN, Bell DR, Boobis AR, Cohen SM, Dellarco M, *et al.* **2014**. Risk assessment in the 21st century: Roadmap and matrix. *Critical Reviews in Toxicology* 44:6-16.

Farmahin R, Williams A, Kuo B, Chepelev NL, Thomas RS, Barton-Maclaren TS, *et al.* **2017**. Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. *Archives of Toxicology* 91:2045-2065.

Fielden MR, Zacharewski TR. **2001**. Challenges and limitations of gene expression profiling in mechanistic and predictive toxicology. *Toxicological Sciences* 60(1):6–10.

Fent K, Sumpter JP. **2011**. Progress and promises in toxicogenomics in aquatic toxicology: is technical innovation driving scientific innovation? *Aquatic Toxicology* 105:25–39

Forbes VE, Calow P, Sibly RM. **2001**. Are current species extrapolation models a good basis for ecological risk assessment? *Environmental Toxicology and Chemistry* 20:442-447.

Guigueno MF, Fernie KJ. **2017**. Birds and flame retardants: A review of the toxic effects on birds of historical and novel flame retardants. *Environmental Research* 154:398-424.

Hope, B. **2006**. An examination of ecological risk assessment and management practices. *Environment International* 32(8):983-985.

House JS, Grimm FA, Jima DD, Zhou Y-H, Rusyn I, Wright FA. **2017**. A pipeline for high-throughput concentration response modeling of gene expression for toxicogenomics. *Frontiers in Genetics* 8:168.

Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM). **2018**. A strategic roadmap for establishing new approaches to evaluate the safety of chemicals and medical products in the United States. https://dx.doi.org/10.22427/NTP-ICCVAM-ROADMAP2018

Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, *et al.* **2014**. Open TG-GATEs: A large-scale toxicogenomics database. *Nucleic Acids Research* 43:D921-D927.

Jensen, M. B., Lozano, F., & Steinert, M. **2016**. The origins of design thinking and the relevance in software innovations. In *International Conference on Product-Focused Software Process Improvement* (pp. 675-678). Springer, Cham.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. **2016**. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 45:D353-D361.

Krewski D, Acosta D, Jr., Andersen M, Anderson H, Bailar JC, 3rd, Boekelheide K, *et al.* **2007**. Toxicity testing in the 21st century: A vision and a strategy. *Journal of Toxicology and Environmental Health, Part B* 13:51-138.

Kuo B, Francina Webster A, Thomas RS, Yauk CL. **2016**. BMDExpress data viewer-a visualization tool to analyze BMDExpress datasets. *Journal of Applied Toxicology* 36:1048-1059.

LaLone, C. A., Basu, N., Browne, P., Edwards, S. W., Embry, M., Sewell, F., Hodges, G. **2021**. International Consortium to Advance Cross-Species Extrapolation of the Effects of Chemicals in Regulatory Toxicology. *Environmental Toxicology and Chemistry*, *40*(12), 3226-3233.

Landrigan PJ, Fuller R, Acosta NJ, Adeyi O, Arnold R, Baldé AB, *et al.* **2018**. The Lancet commission on pollution and health. *The Lancet* 10119:462-512.

Larras F, Billoir E, Baillard V, Siberchicot A, Scholz S, Wubet T, *et al.* **2018**. DRomics: A turnkey tool to support the use of the dose–response framework for omics data in ecological risk assessment. *Environmental Science and Technology* 52:14461-14468.

Lea IA, Gong H, Paleja A, Rashid A, Fostel J. **2016**. CEBS: A comprehensive annotated database of toxicological data. *Nucleic Acids Research* 45:D964-D971.

Leung, K. M. **2018**. Joining the dots between omics and environmental management. *Integrated Environmental Assessment and Management*, 14(2), 169-173.

Li HH, Hyduke DR, Chen R, Heard P, Yauk CL, Aubrecht J, *et al.* **2015**. Development of a toxicogenomics signature for genotoxicity using a dose-optimization and informatics strategy in human cells. *Environmental and Molecular Mutagenesis* 56:505-519.

Martin JA, Wang Z. **2011**. Next-generation transcriptome assembly. *Nature Reviews Genetics* 12:671–682.

Martinson, J. W., Bencic, D. C., Toth, G. P., Kostich, M. S., Flick, R. W., See, M. J., *et al.* **2021**. *De Novo* Assembly of the Nearly Complete Fathead Minnow Reference Genome Reveals a Repetitive But Compact Genome. *Environmental Toxicology and Chemistry*. Online only.

NASEM (National Academies of Sciences, Engineering, and Medicine). **2017**. *Using 21st Century Science to Improve Risk-Related Evaluations*. Washington, DC: NASEM.

Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C., Afshari, C. A. **1999**. Microarrays and toxicology: the advent of toxicogenomics. *Molecular Carcinogenesis*, 24(3), 153-159.

Otava M, Sengupta R, Shkedy Z, Lin D, Pramana S, Verbeke T, *et al.* **2017**. Isogenegui: Multiple approaches for dose-response analysis of microarray data using R. *The R Journal* 9:14-26.

Pain, G., Hickey, G., Mondou, M., Crump, D., Hecker, M., Basu, N., Maguire, S. **2020**. Drivers of and obstacles to the adoption of toxicogenomics for chemical risk assessment: insights from social science perspectives. *Environmental Health Perspectives*, 128(10), 105002.

Pastoor TP, Bachman AN, Bell DR, Cohen SM, Dellarco M, Dewhurst IC, *et al.* **2014**. A 21st century roadmap for human health risk assessment. *Critical Reviews in Toxicology* 44:1-5.

Peddada SD, Lobenhofer EK, Li L, Afshari CA, Weinberg CR, Umbach DM. **2003**. Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference. *Bioinformatics* 19:834-841.

Primmer C, Papakostas S, Leder E, Davis M, Ragan M. **2013**. Annotated genes and non-annotated genomes: Cross-species use of gene ontology in ecology and evolution research. *Molecular Ecology* 22:3216-3241.

Pruitt KD, Tatusova T, Maglott DR. **2007**. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35:D61-D65.

Quercioli D, Roli A, Morandi E, Perdichizzi S, Polacchini L, Rotondo F, *et al.* **2018**. The use of omics-based approaches in regulatory toxicology: An alternative approach to assess the no observed transcriptional effect level. *Microchemical Journal* 136:143-148.

Smetanová S, Riedl J, Zitzkat D, Altenburger R, Busch W. **2015**. High-throughput concentration–response analysis for omics datasets. *Environmental Toxicology and Chemistry* 34:2167-2180.

Thomas RS, Allen BC, Nong A, Yang L, Bermudez E, Clewell III HJ, *et al.* **2007**. A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicological Sciences* 98:240-248.

Thomas RS, Clewell III HJ, Allen BC, Yang L, Healy E, Andersen ME. **2012**. Integrating pathway-based transcriptomic data into quantitative chemical risk assessment: A five chemical case study. *Mutation Research* 746:135-143.

Thomas, R. S., Bahadori, T., Buckley, T. J., Cowden, J., Deisenroth, C., Dionisio, K. L., et al. **2019**. The next generation blueprint of computational toxicology at the US Environmental Protection Agency. *Toxicological Sciences*, 169(2), 317-332.

Tschimmel, K. **2012**. Design Thinking as an effective Toolkit for Innovation. *ISPIM Conference Proceedings* (pp 1).

US Congress. **2016**. *Frank R. Lautenberg Chemical Safety for the 21st Century Act*. 130 STAT. 448. 114th United States Congress.

Van Aggelen, G., Ankley, G. T., Baldwin, W. S., Bearden, D. W., Benson, W. H., Chipman, J. K., *et al.* **2010**. Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. *Environmental Health Perspectives*, 118(1), 1-5.

van der Vegt, R. G., Maguire, S., Crump, D., Hecker, M., Basu, N., Hickey, G. M. **2021**. Chemical risk governance: Exploring stakeholder participation in Canada, the USA, and the EU. *Ambio* 1-13.

Villeneuve DL, Garcia-Reyero N. **2011**. Vision & strategy: Predictive ecotoxicology in the 21st century. *Environmental Toxicology and Chemistry* 30:1-8.

Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, *et al.* **2014**. Adverse outcome pathway (AOP) development I: Strategies and principles. *Toxicological Sciences* 142:312-320.

Voshall A, Moriyama EN. **2018**. Next-generation transcriptome assembly: strategies and performance analysis. In *Bioinformatics in the era of post genomics and Big Data* (ed. Abdurakhmonov IY), pp. 15–36. IntechOpen, London.

Wang R-L, Biales AD, Garcia-Reyero N, Perkins EJ, Villeneuve DL, Ankley GT, *et al.* **2016**. Fish connectivity mapping: Linking chemical stressors by their mechanisms of action-driven transcriptomic profiles. *BMC Genomics* 17:84.

Wheeler, A. **2019**. Directive to Prioritize Efforts to Reduce Animal Testing [Memorandum]. *US EPA*. https://www.epa.gov/sites/default/files/2019-09/documents/image2019-09-09-231249.pdf

WHO. **2012**. *State of the Science of Endocrine Disrupting Chemicals – 2012*. https://www.unep.org/resources/publication/state-science-endocrine-disputing-chemicals-ipcp-2012. Accessed January 21, 2022.

Yang L, Allen BC, Thomas RS. **2007**. BMDExpress: A software tool for the benchmark dose analyses of genomic data. *BMC Genomics* 8:387.

CONNECTING PARAGRAPH

Chapters 1 and 2 outlined barriers to using toxicogenomics data for decision-making in environmental risk assessments (ERAs). One of these is the lack of genomics resources for ecological species (section 2.1.3), and the inaccessibility of many bioinformatics workflows to people who do not have advanced programming skills and statistical training.

The lack of genomics resources mainly relates to a lack of annotated reference genomes. Chapter 3 presents a computing ecosystem for comprehensive analysis of raw RNA-seq data from any species regardless of whether that species has a reference genome. It is comprised of three tools: EcoOmicsAnalyst for raw data processing, ExpressAnalyst for statistical and functional analysis, and EcoOmicsDB for exploration of protein ortholog evidence. The core of the ecosystem is EcoOmicsAnalyst, which has workflows for both organisms with reference transcriptomes (Kallisto) and for those without (Seq2Fun). In addition to a robust computing environment with a web interface, chapter 3 presents version 2.0 of the Seq2Fun algorithm. Compared to version 1.0 (Liu *et al.,* 2021), version 2.0 has improved reads mapping efficiency, an expanded protein ortholog database, and expanded functional annotation of the orthologs. Together, the three tools allow users to go from raw RNA-seq data from any species to functional insights in less than 24 hours, all on a standard laptop computer and without writing a single line of code. This greatly increases the accessibility and value of toxicogenomics data from ecologically relevant species, especially for researchers with limited prior experience with next-generation sequencing technologies.

This chapter is being prepared as manuscript to be submitted to the journal *Nature Methods* (aim for end of March 2022). As such, formatting requirements for this journal require the Results and Discussion section to be presented before the Methods section. There are two co-first authors, the candidate and Dr. Peng Liu, formerly a postdoctoral fellow at McGill University and now a bioinformatics biologist at Environment and Climate Change Canada in Ottawa, Canada. The manuscript is additionally co-authored by Orcun Hacariz, Elena Legrand, Guangyan Zhou, Benjamin de Jourdan, Jessica Head, Niladri Basu, and Jianguo Xia. Supplemental information is provided in the thesis appendix.

CHAPTER 3. A COMPUTATIONAL ECOSYSTEM FOR RNA-SEQ DATA FROM NON-MODEL ORGANISMS

Peng Liu[1], Jessica Ewald[1*], Orcun Hacariz[1], Elena Legrand[1], Guangyan Zhou[1], Benjamin de Jourdan[2], Jessica Head[1], Niladri Basu[1], Jianguo Xia[1]*

[1] Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Canada

[2] Huntsman Marine Science Centre, St. Andrews, Canada

* These authors contributed equally to the work.

3.1 ABSTRACT

RNA-sequencing data is increasingly being collected from non-model organisms by researchers in the environmental life sciences. However, many of these researchers have limited bioinformatics experience and therefore do not have the computing resources, programming skills, or domain knowledge to extract value from these data. Here, we present an entirely web-based computing ecosystem for processing, analyzing, and interpreting RNA-seq data from any eukaryotic species. This is achieved by supporting both Kallisto-based pipelines for species with a reference transcriptome, and Seq2Fun-based pipelines for species without one. The ecosystem includes EcoOmicsAnalyst (www.ecoomicsanalyst.ca) for raw data processing, ExpressAnalyst (www.expressanalyst.ca) for statistical analysis, and EcoOmicsDB (www.ecoomicsdb.ca) for interpretation of Seq2Fun ortholog results. In this paper, we describe major improvements to the Seq2Fun algorithm and present two case studies to demonstrate 1) how Seq2Fun quantification

compares to Kallisto quantification with reference genomes, and 2) how the presented tools can be leveraged to produce high-quality functional insights for non-model organisms.

3.2 Introduction

As costs of RNA-sequencing have plummeted over the last decade, many researchers in the environmental life sciences have added a transcriptomics component to their research program (Wachi *et al*., 2017). Many are not computational scientists, thus lack of computing infrastructure, bioinformatics expertise, and programming skills present significant barriers to extracting value from the raw RNA-seq data (Ekblom and Galindo, 2011). Environmental life sciences researchers often study non-model organisms such as endangered species, under-studied taxonomic groups, or representatives from specific ecosystems. This makes RNA-seq data analysis even more challenging, since many of these species do not have high-quality, functionally annotated reference genomes (Sundaram *et al.,* 2017; Ekblom and Galindo, 2011). Assembling and annotating genomes is an immense amount of work that is beyond the scope of most research labs. For example, the ongoing effort by the US EPA to produce a high-quality reference genome for fathead minnow has taken a large team of researchers more than 5 years to complete (Burns *et al*., 2015; Martinson *et al.,* 2021).

Instead, many researchers use *de novo* transcriptome assembly to analyze RNA-seq data from non-model organisms (Wachi *et al*., 2017; Ekblom and Galindo, 2011). This process involves piecing together putative transcript sequences directly from the RNA-seq data itself, annotating these transcripts using BLAST-based algorithms, and then aligning all reads to the assembled transcripts to quantify their relative expression across samples. Analyses that include *de novo* transcriptome assembly require extensive computational resources, multiple command-line software, advanced programming skills, and days to weeks of runtime (Martin *et al.,* 2011; Volshall *et al*., 2018). Finally, the results that they produce are difficult to analyze because *de novo*-assembled transcriptomes often contain several hundred thousand transcripts (compared to 15-40k for a well-annotated reference genome), the majority of which remain uncharacterized even after functional annotation with software like Trinotate and Blast2GO (Conesa and Gotz, 2008). Moreover, *de novo* assembled transcripts are prone to false positives, which poses another layer of difficulties for data analysis and interpretation (Freedman *et al*., 2020).

For these reasons, our group previously developed the Seq2Fun algorithm for aligning raw RNA-seq reads from any species to a large collection of protein ortholog groups derived from the genomes of hundreds of species (www.seq2fun.ca). Seq2Fun allows researchers to avoid the difficult steps of *de novo* assembly and transcript annotation, while still producing counts tables that can be analyzed using conventional transcriptomics pipelines. We previously showed that Seq2Fun outperforms *de novo* transcriptomes in terms of accuracy, precision, computing time, RAM usage, and functional annotation of the results in analyses with zebrafish, mouse, chicken, and roundworm RNA-seq data (Liu *et al.,* 2021). However, Seq2Fun was originally published only as command line software, which may not be accessible to researchers who are not trained as computational scientists, and there was limited support for downstream analysis of the results. In addition, the initial protein ortholog databases were based on the KEGG Orthology database (Kanehisa *et al*., 2016), which limited transcriptomic coverage, resolution, and functional annotation.

The objective of this study was to develop an entirely web-based computing ecosystem for RNA-seq data from any eukaryotic non-model organisms, regardless of whether they have a reference genome. There are three main tools: EcoOmicsAnalyst for raw data processing, ExpressAnalyst for comprehensive downstream statistical analysis and visual analytics, and EcoOmicsDB for accessing ortholog evidence and information. With this ecosystem, researchers can extract functional insight from raw FASTQ files from any species without writing a single line of code, all on a regular laptop computer, and generally in less than 48 hours of runtime, the majority of which is unsupervised upload and processing time. In addition, we also present Seq2Fun version 2.0, which includes major improvements to the algorithm's protein ortholog databases and memory usage.

3.3 Results

EcoOmicsAnalyst (www.ecoomicsanalyst.ca) is the main tool in the computing ecosystem. It is supplemented by ExpressAnalyst (www.expressanalyst.ca) and EcoOmicsDB (www.ecoomicsdb.ca) for downstream analysis of the results (Figure 3-1). EcoOmicsAnalyst contains two main workflows, one based on Kallisto (organisms with a reference transcriptome) (Bray *et al*., 2016) and one based on Seq2Fun (organisms without a reference transcriptome) (Liu *et al*., 2021). ExpressAnalyst, a popular tool for comprehensive transcriptomics profiling and network visual analytics that was previously named NetworkAnalyst (Zhou *et al*., 2019), includes ID annotation and pathway library files for counts tables generated by EcoOmicsAnalyst. EcoOmicsDB contains detailed information for each Seq2Fun ortholog for targeted analysis of key features that emerge from statistical analysis in ExpressAnalyst.

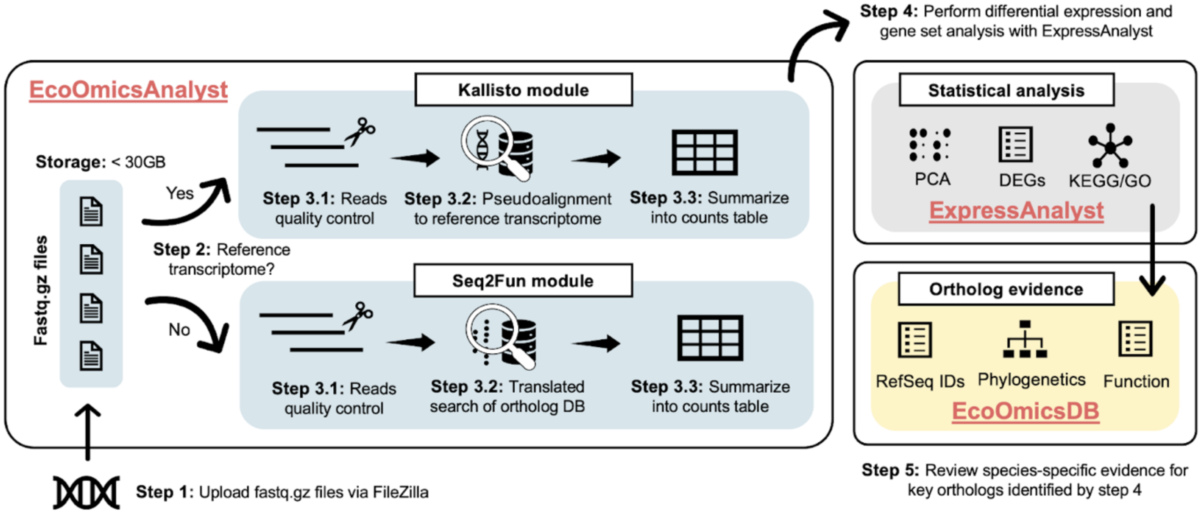

Step 4: Perform differential expression and gene set analysis with ExpressAnalyst
EcoOmicsAnalyst
Kallisto module
Storage: < 30GB
Yes
Step 3.1: Reads quality control
Step 3.2: Pseudoalignment to reference transcriptome
Step 3.3: Summarize into counts table
Fastq.gz files
Step 2: Reference transcriptome?
No
Seq2Fun module
Step 3.1: Reads quality control
Step 3.2: Translated search of ortholog DB
Step 3.3: Summarize into counts table
Statistical analysis
PCA
DEGs
KEGG/GO
ExpressAnalyst
Ortholog evidence
RefSeq IDs
Phylogenetics
Function
EcoOmicsDB
Step 5: Review species-specific evidence for key orthologs identified by step 4
Step 1: Upload fastq.gz files via FileZilla

Figure 3-1: Overview of the computational ecosystem and steps for a typical workflow.

Users can upload up to 30GB of compressed FASTQ files to their EOA account, and raw files are kept on the server for 30 days (alignment typically takes 6-9 hours). Processing results (abundance tables, annotation files, etc.) are accessible after raw files have been deleted.

3.3.1 Expanded ortholog databases in Seq2Fun version 2.0

Seq2Fun version 1.0 mapped reads to the KEGG ortholog (KO) database (Kanehisa *et al.*, 2016). While the quantification was highly accurate and functional analysis was consistent with analysis results that used reference transcriptomes (Liu *et al.*, 2021), the KO-based system has three main limitations. The first is limited and biased coverage of transcriptomes in the ortholog database. Not all protein-coding genes are annotated with KOs, for example the human genome has 19,648 protein-coding genes and only 14,964 (76.16%) are annotated with KOs (Kanehisa *et al.*, 2017). Coverage is even lower for non-mammalian species, for example the zebrafish genome has 26,584 protein-coding genes and only 16,322 (61.40%) are annotated with KOs. This lower coverage is not evenly distributed across the transcriptome. There are fewer KEGG pathways defined for non-mammalian species and some whole biological processes specific to non-mammalian species are missing, for example egg yolk formation in oviparous species. Vitellogenin, a frequently measured biomarker that is a precursor to egg yolk (Hansen *et al.*, 1998), is not found in the KO system. The second limitation is transcript resolution. KO groups are at a higher level than individual genes, often grouping many genes from one species together. While this is inevitable to some extent during ortholog definition, increased resolution would be an asset. Finally, the third limitation is lack of functional annotation with gene sets beyond KEGG pathways.

For Seq2Fun version 2.0, a custom ortholog database was created by using the OrthoFinders algorithm to group 12,828,537 protein-coding sequences from 687 species (Emms and Kelly, 2019), which took more than ten days on a server with 54-threads and 504GB RAM. The new database addresses the limitations described above: it includes 100% of protein-coding genes from each constituent genome, the sequence-similarity parameters for ortholog definition were chosen to produce more ortholog groups at a higher resolution than the KO system (Figure S1), and species-specific functional annotations were compiled to produce both KEGG and GO term gene sets for Seq2Fun ortholog IDs (Kanehisa *et al.*, 2017; Gene Ontology Consortium, 2019). Of the ~13 million protein-coding sequences, 5,871,017 were annotated with KEGG pathways and 1,567,627 were annotated with GO terms. The 687 species were organized into twelve phylogenetic groups (compared to six groups in version 1.0), based on the NCBI taxonomy database and which were used to define smaller sub-group ortholog databases (Table 3-1) (Schoch *et al.*, 2020).

Table 3-1: Expanded ortholog databases for Seq2Fun version 2.0.

Group	Species	Proteins	Ortholog	Ortholog (v1)
Eukaryotes	691	12 828 537	576 442	NA
Vertebrates	212	4 573 967	74 321	30 392
Mammals	94	1 909 225	51 570	19 323
Birds	31	482 205	24 076	15 537
Reptiles	20	382 462	24 823	15 884
Amphibians	3	75 261	22 211	15 199
Fishes	61	1 655 763	49 668	21 427
Arthropods	119	1 709 887	114 560	NA
Nematodes	6	103 321	38 733	NA
Invertebrate	158	2 502 377	198 800	NA
Plants	127	3 925 179	147 787	NA
Fungi	138	1 278 312	142 573	NA
Protists	52	655 135	138 844	NA

The distribution of the number of species represented by each ortholog group in the eukaryotes database is shown in Figure 3-2A. The majority of ortholog groups contain transcripts from only one species (416 413 out of 576 442, 72%). Some databases contain more orthologs relative to the number of species represented in them than others (Figure 3-2B). For example, there are more vertebrate species (n = 212) than invertebrate (n = 158), however there are more than twice as many distinct orthologs for invertebrate (n = 198, 800) than for vertebrate (n = 74,321). This is likely due to the variable taxonomic diversity represented in the different databases.

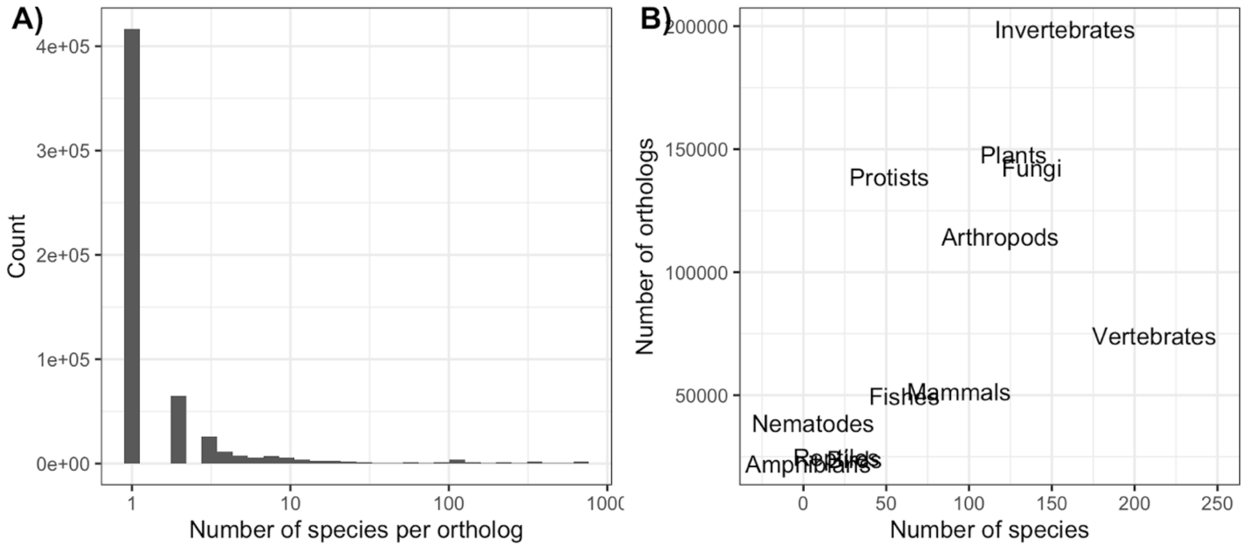


Figure 3-2: A) Histogram of the number of genes mapping to each ortholog in the "eukaryotes" database. B) Scatterplot of the number of species and the number of orthologs in a database.

3.3.2 EcoOmicsDB

EcoOmicsDB (www.ecoomicsdb.ca) contains the Entrez IDs, symbols, descriptions, functional annotations, and Seq2Fun IDs of all ~13 million genes in the ortholog database. The database can be queried by Seq2Fun ID, which retrieves details and generates graphics on all genes and species in that ortholog group. Grouping over 13 million sequences from 687 species and assigning each group a single symbol, description, and functional annotation is a complex task. In some cases, two distinct ortholog groups can be given the same symbol, for example when a gene has mutated between (ortholog) or within (paralog) species, but the symbol has remained the same. Another problematic case is when transcripts with different symbols and descriptions are included in the same ortholog group. This happens occasionally because each reference genome has its own unique history of development, and over time, different conventions have been accepted for naming and describing the same protein in different species. EcoOmicsDB does not resolve these conflicts but increases the transparency of Seq2Fun ortholog groups so that users can investigate the evidence behind specific cases of interest. For convenience, the table of differential expression results in ExpressAnalyst contain links to the EcoOmicsDB profiles for all Seq2Fun ortholog IDs. Further, graphics on the coverage of different species sub-groups provides valuable insights into the taxonomic domain of Seq2Fun ortholog groups.

3.3.3 Case Studies

Several datasets were used to develop, build, test, and update EcoOmicsAnalyst, ExpressAnalyst, and EcoOmicsDB. Here, we showcase three of them in two case studies to demonstrate how these tools simplify and improve transcriptomics data analysis for non-model organisms. Case study #1 uses data from two species that have reference genomes to enable comparisons between Kallisto and Seq2Fun. Case study #2 analyzes data from multiple species that do not have a reference genome to demonstrate strengths of Seq2Fun compared to de novo transcriptomes.

Case Study 1: Kallisto vs. Seq2Fun

In case study #1, RNA-seq profiles from two species with reference genomes (zebrafish and American lobster) were processed with both Seq2Fun and Kallisto using EcoOmicsAnalyst. The resulting counts tables were uploaded to ExpressAnalyst to perform DEA and GSA with KEGG pathways and Gene Ontology terms. Note that American lobster does not have any publicly available pathway libraries for the current version of the reference genome, so only DEA was done for the American lobster Kallisto pipeline.

Reference species #1: Zebrafish

The zebrafish RNA-seq data were collected as part of a previously published study that investigated the toxicity of perfluorooctane sulfonate (PFOS) compared to sodium p-perfluorous nonenoxybenzene sulfonate (OBS), a popular PFOS alternative (Huang *et al*., 2021). Three groups of zebrafish embryos were exposed to 20 mg/L PFOS, 20 mg/L OBS, and 30 mg/L OBS, in addition to a fourth unexposed control group, starting at six hours post fertilization. RNA-seq profiles were measured in whole embryos at four days post fertilization. Raw FASTQ files were downloaded from NCBI's Gene Expression Omnibus (GEO) at accession GSE164074. The original study found that while immune-related genes and pathways were impacted by both OBS and PFOS exposure, a higher number were dysregulated to a higher degree by PFOS. This supports their overall conclusion that while PFOS and OBS have a similar mechanism, PFOS is more toxic, likely because it is more bioaccumulative (Tu *et al*., 2019).

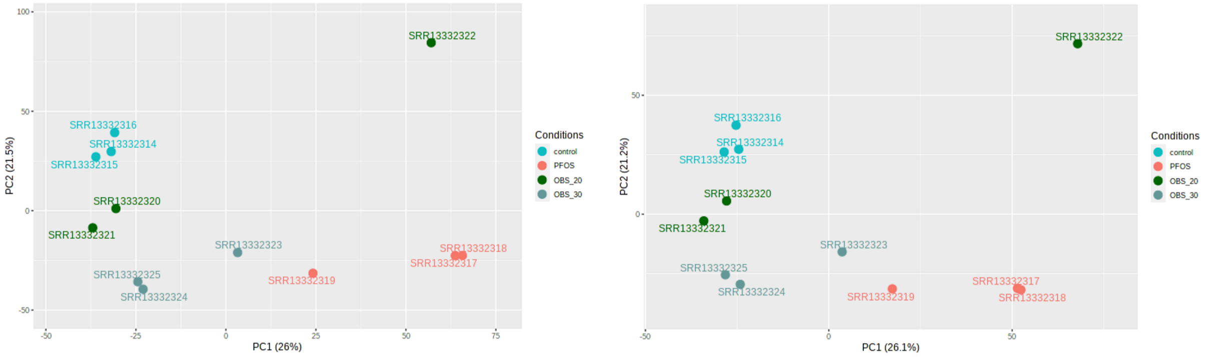


Figure 3-3: PCA of normalized zebrafish counts tables for a) Kallisto and b) Seq2Fun. This plot was generated by ExpressAnalyst.

Visual inspection of PCA plots of the normalized counts (Figure 3-3) shows that both Seq2Fun and Kallisto captured the same variability structure among the samples, even though Kallisto mapped a higher percentage of reads (78% vs. 53%) to a higher number of features (52 518 vs. 17 186) (Table 2). The higher number of reads is expected since Kallisto is quantifying multiple transcript isoforms for each coding transcript, as well as non-coding transcripts, while Seq2Fun is quantifying orthologs that do not distinguish isoforms and only include protein-coding sequences. The number of DEGs for each exposed vs. control group has the same pattern between Kallisto and Seq2Fun (PFOS > OBS30 > OBS20), with a slightly higher number in each contrast for Kallisto (Table 3-2).

Table 3-2: Results from zebrafish DEA and GSA

	Kallisto			**Seq2Fun**		
# Features quantified	41 345			17 186		
Reads mapping	77.49%			53.12%		
Gene mapping	73.57%			35.12%		
	OBS20	**OBS30**	**PFOS**	**OBS20**	**OBS30**	**PFOS**
DEGs	111	395	571	97	320	503
KEGG	6	1	1	1	10	23
GO BP	3	4	5	0	5	3
GO MF	8	9	5	0	8	3
GO CC	1	2	1	2	2	2

Directly comparing the DEGs across Kallisto and Seq2Fun is challenging because of a lack of mapping between zebrafish transcripts and ortholog groups, however a cursory examination of the results shows that the top DEGs are very similar across the two software. For example, in the PFOS vs. CTRL results, the symbols "mmp9" (Entrez = 406397; Seq2Fun = s2f_6058) and "ahsg2" (Entrez = 567406; Seq2Fun = s2f_12793) are in the top five DEGs for each software, and the quantified log2FC values are nearly the same (mmp9 = -3.81, -3.80 and ahsg2 = 6.84, 7.16 for Kallisto and Seq2Fun respectively). The number of significant pathways is more variable between the Kallisto and Seq2Fun results, however the biological themes of the top enriched pathways are quite similar for given contrasts. For example, for both software the PFOS vs. CTRL pathways are largely related to cellular signaling and the OBS30 vs. CTRL pathways are mostly related to lipid metabolism, lipid transport, and steroid biosynthesis (Tables S1 and S2).

Differences in the specific pathways flagged as enriched across the two software are expected because the Seq2Fun ortholog groups are more densely annotated than most transcriptomes. Overrepresentation analysis is very dependent on the total number of measured genes and on the number of genes in individual pathways. Thus, even if the lists of DEGs are very similar across two analyses, the specific pathways that get flagged as enriched can vary if there are substantial differences in annotation densities, as in the case for the Kallisto and Seq2Fun analyses. For example, in the analysis for case study #1, there were 41 345 genes quantified by Kallisto that were annotated by KEGG pathways in 13 212 cases, whereas for Seq2Fun, there were 17 186 quantified genes annotated by KEGG pathways in 32 922 cases. Thus, the ratio of KEGG annotations per gene is 0.32 for the Kallisto results and 1.92 for the Seq2Fun results. This explains many of the differences in the specific pathways that were flagged as enriched between the two analysis pipelines.

Reference species #2: Lobster

The American lobster RNA-seq profiles were collected as part of a study that investigated the biological impacts of exposure to heavy crude oil. One group was exposed to a water accommodated fraction (WAF) of oil at 72% concentration (WAF_72; n = 7), another to 0.39 mg/L of the polycyclic aromatic hydrocarbon (PAH) 1-methynaphthalene (positive control; n =

6), and a third group was unexposed (control; n = 6). These are new, unpublished data collected by the authors of this paper.

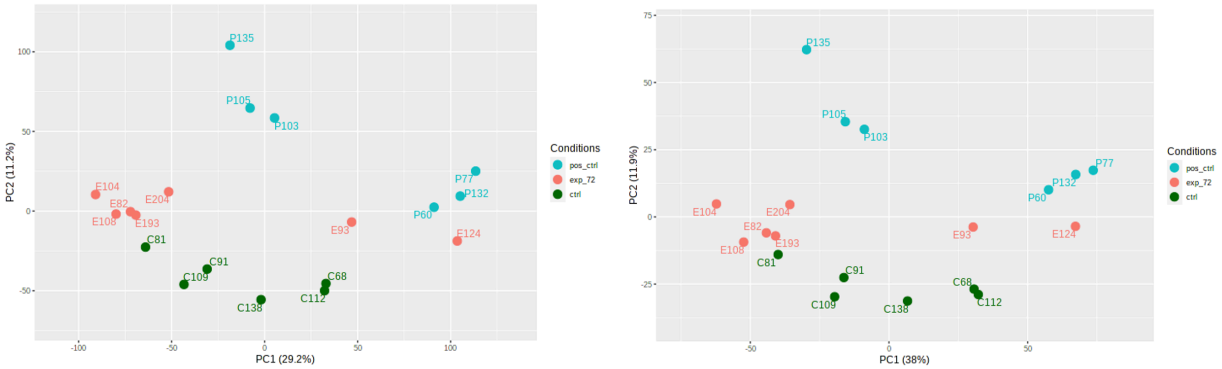


Figure 3-4: PCA of normalized lobster counts tables for a) Kallisto and b) Seq2Fun. This plot was generated by ExpressAnalyst.

The patterns with the lobster data are very similar to those with the zebrafish data. PCA plots of the normalized counts data show the same variance structure for both quantification methods (Figure 3-4), even though more features were quantified by Kallisto and Kallisto had a higher reads and gene mapping rate (Table 3-3). The pattern of DEGs for each contrast is also similar between Kallisto and Seq2Fun, with the positive control group having many more DEGs than the 72% WAF group. While the number of DEGs for each group is higher for Kallisto, the ratio between the two contrasts is very similar across both quantification methods.

Table 3-3: Results from lobster DEA and GSA.

	Kallisto	**Seq2Fun**
# Features quantified	36 925	10 332
Reads mapping	71.71%	29.13%
Gene mapping	74.20%	13.86%

	Pos. ctrl	*WAF 72%*	*Pos. ctrl*	*WAF 72%*
DEGs	908	50	287	14
KEGG	NA	NA	3	0
GO BP	NA	NA	1	0
GO MF	NA	NA	4	0
GO CC	NA	NA	0	0

There were lower reads and gene mapping rates for the lobster Seq2Fun results than for the zebrafish Seq2Fun results (reads: 53% vs. 29%; genes: 35% vs. 14%). This is likely explained by the taxonomic diversity of each database, with "invertebrates" covering both a higher number and a greater diversity of species than "fishes". As more genomes are published and added to the Seq2Fun databases, we expect that species coverage, reads mapping rates, the number of ortholog groups, and the number of genes in each ortholog group in each database will increase. However, the number of genes within an individual species' genome will stay the same, thus we expect genes mapping rates to decrease and needs to be properly considered when performing gene set analysis. It can also be addressed by making more narrow species groups, like the "fishes" and "birds" databases that fall within the "vertebrate" category.

Overall, case study #1 shows that Kallisto and Seq2Fun give very similar results in terms of relationships between samples, as shown by PCA plots and the relative number of DEGs for different experimental groups. While the number of pathways highlighted by overrepresentation analysis is less consistent, the functional interpretation is quite similar. The lobster analysis shows one advantage of Seq2Fun for recently published genomes - no publicly available gene set libraries exist for the American lobster transcriptome, however we are still able to easily perform functional analysis of the Seq2Fun results (Table S3).

Case Study 2: Comparative transcriptomics in salamander species

In case study #2, Seq2Fun was used to analyze RNA-seq profiles from three ambystomatid salamander species, one with a reference genome (*Ambystomatidae mexicanum*, abbreviation MEX) and two without (*Ambystomatidae andersoni*, abbreviation AND; *Ambystomatidae maculatum*, abbreviation MAC). The data were originally collected as part of a comparative study of transcriptional responses to limb regeneration (Dwaraka *et al*., 2019). The upper arm was amputated from larvae from each of the three species, and tissue samples were taken at the time of amputation (time0), and 24 hours after amputation (time24). Three replicates from each species and time group were sequenced, resulting in 3 reps * 2 time points * 3 species = 18 RNA-seq samples. Raw FASTQ files were downloaded from NCBI GEO at accession GSE116777. For this case study, FASTQ files were re-processed in 9 hours with the Seq2Fun module (vertebrates database) in EcoOmicsAnalyst.

The primary source of variability in the normalized counts matrix was species differences, as shown by the separation according to species along PC1 (Figure 3-5). AND and MEX samples fall closer to each other than to MAC, which makes sense given that AND is more closely related to MEX (estimated divergence time = 4.27 million years) than MAC is (estimated divergence time = 21.47 million years) (Hedges *et al*., 2015). The second largest source of variability was time since amputation, shown by the separation of samples with time0 and time24 annotations along PC2 (Figure 5).

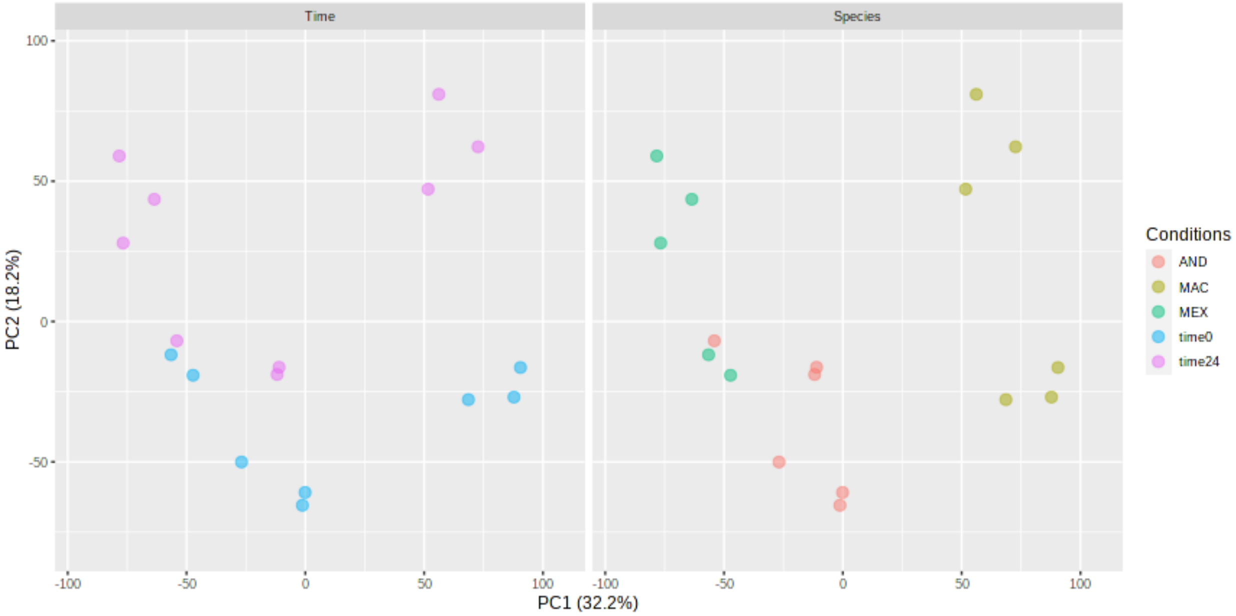


Figure 3-5: PCA of normalized counts tables for the salamander data, with samples annotated by A) time since amputation, and B) species. This plot was generated by ExpressAnalyst.

Differential expression analysis was performed in ExpressAnalyst to identify genes that were significantly different between time0 and time24 across species by analyzing all samples together and considering species as a blocking factor. This takes variability associated with species into account when calculating the p-value for differences between time0 and time24. Using the same statistical thresholds as the original publication (FDR adj. p-value < 0.1, no log2FC cut-off), there were 1566 DEGs.

The "Interactive Volcano Plot" tool in ExpressAnalyst was used to perform overrepresentation pathway analysis (ORA) separately on the up and down-regulated genes with KEGG, Gene Ontology Biological Process (GO BP), Molecular Function (GO MF), and Cellular Component (GO CC) gene sets. Overall, there were 24 significantly enriched pathways in the list of up-regulated genes and 48 in the list of down-regulated genes (Tables S4 and S5). The up-regulated pathways were mainly related to immune response, cell proliferation (many cancer-related pathways), and programmed cell death. The down-regulated pathways were mainly related to muscle tissue and cellular metabolism. This is very consistent with the functional analysis results reported by the original publication (Dwaraka *et al*., 2019), and makes sense when interpreted in the context of limb regeneration.

The top five DEGs (Table 3-4) were queried in EcoOmicsDB to demonstrate the type of information that can be retrieved. EcoOmicsDB shows that each of these genes is supported by lots of evidence (>190 genes and > 170 species for each). Three of the genes (s2f_3904, s2f_939, and s2f_2248) have many more genes than species. Visual examination of the main EcoOmicsDB table output shows that each species contributes multiple genes with very similar descriptions, for example the s2f_939 table contains descriptions for metalloproteinase inhibitors 1, 2, and 4, metalloproteinase-like genes, and multiple metalloproteinase isoform numbers. Taken together, there is ample evidence that these key differentially expressed genes are robust and represent real proteins, however it should be noted that the TIMP4 and MMP2 results could each represent a small group of highly similar genes within the salamander genomes. For these ortholog groups, each species is contributing, on average, two or three distinct genes as defined by the Entrez ID system.

Table 3-4: Top 5 DEGs from case study #2.

Seq2Fun ID	Symbol	Adj. p-value	log2FC	# Genes in EODB	# Species in EODB
s2f_11083	SERPINE1	3.2E-28	5.2	191	174
s2f_9201	PLEK2	1.3E-25	4.4	222	207
s2f_3904	TNC	2.6E-24	4.4	305	212
s2f_939	TIMP4	3.8E-22	5.1	650	212

s2f_2248	MMP2	3.1E-21	3.6	467	212

The original study quantified the RNA-seq data using a reference transcriptome from MEX and *de novo* transcriptomes from AND and MAC, and then performed DEA for each species. They then identified DEGs that were shared across species by searching for sequence similarities among the differentially expressed contigs from AND and MAC and the differentially expressed transcript sequences from MEX using the BLAST algorithm. Ultimately, they found 405 transcripts that were significantly impacted in all three species. We quantified all RNA-seq samples with Seq2Fun, even though there was a reference transcriptome for one species. This greatly simplified the downstream analysis. Since all quantified samples shared the same set of Seq2Fun IDs, the data could be integrated across species and analyzed in a single DEA without performing any ortholog mapping. It also improved the statistical power, with $n = 18$ instead of $n = 6$, likely explaining our 1566 DEGs versus their 405.

Finally, we note that the use of salamanders makes this a particularly strong case for Seq2Fun. Amphibians can have notoriously large genomes (Liedtke *et al*., 2018), estimated to range from 14 to 120 GB across salamander species (for reference, the human genome is 3.2 GB) (Nowoshilow *et al.*, 2018). Performing *de novo* assembly of two salamander genomes would be extremely computationally intensive. The full analysis for case study #2, including raw reads processing, statistical analysis, and figure preparation, took less than 24 hours, was completed without the command line or R, and was all done on a laptop computer.

3.4 DISCUSSION

Overall, the web-based computing ecosystem presented in this paper represents a valuable resource for researchers collecting RNA-sequencing data from non-model organisms. By removing barriers related to computing resources, programming skills, and knowledge of bioinformatics databases, we believe that the combination of EcoOmicsAnalyst, ExpressAnalyst, and EcoOmicsDB will make transcriptomics data processing and analysis more accessible to the environmental life sciences community.

Resolving conflicting ortholog annotations is unavoidable in RNA-seq analysis of non-model organisms: even when researchers choose to analyze their data with a *de novo* transcriptome, they still must annotate this de novo transcriptome by drawing on functional information from other species (Conesa and Gotz, 2008). Seq2Fun addresses ortholog grouping and annotation at the beginning of the raw data processing pipeline, while analyses with *de novo* transcriptomes do this at the end. Since Seq2Fun comprehensively addresses ortholog grouping and annotation across many species, it is extremely well suited for cross-species comparisons of transcriptomics data. There is already great interest in this, for example recent efforts to use 'omics data for cross-species extrapolation in the field of environmental toxicology (LaLone *et al*., 2021).

The current Seq2Fun algorithm (version 2.0) is based on a translated nucleotide to peptide search, which is extremely efficient at overcoming large evolutionary distances between query and target organisms in the database. However, this approach cannot quantify non-coding genes or distinguish gene isoforms. In the future, we will develop a tiered search strategy by first conducting DNA-to-DNA search for each read, then a DNA-to-protein (translated) search if the read fails. This novel strategy will not only enable Seq2Fun to quantify the whole transcriptome including non-coding sequences but will also have a higher resolution of features without compromising the ability to overcome long evolutionary distance.

We developed EcoOmicsDB for the limited initial goal of increasing transparency of the Seq2Fun ID system. However, with continued development and engagement of the environmental life sciences community, EcoOmicsDB could become the authoritative resource for transcript identification and functional annotation in non-model organisms. The genomes of non-model organisms have been much less studied compared to mammalian model organisms and there are still many uncharacterized IDs in published reference genomes. We envision a system in which functional information learned from individual transcriptomics studies is added to the Seq2Fun ortholog profiles in EcoOmicsDB. Over time, knowledge from such studies can be pooled across species to gain insights into uncharacterized proteins. This will not only improve the annotation and functional analysis of the Seq2Fun results but would also benefit the whole research community that works on orthologs and non-model organism transcriptomics.

3.5 METHODS

3.5.1 ORTHOLOG GROUP DEFINITION AND ANNOTATION

The core algorithm of Seq2Fun is the translated search of RNA-seq reads against the protein database. To create a comprehensive database, all protein-coding genes (*n* = 13,057,389) from 687 organisms which cover all major phylums of eukaryotes were downloaded from KEGG using KEGGREST (version 1.34.0). Protein FASTA files for each species were submitted to OrthoFinder (version 2.5.4) for classification of genes into ortholog groups. OrthoFinder (parameters: t = 56, a = 25) was run on a server with 56 threads and 504 GB's RAM and it took about ~10 days to finish ortholog grouping for all the organisms. Information from each gene was collapsed to generate a single gene symbol, description, KEGG pathway, and GO term annotation for each ortholog group. Phylogenetic groups of organisms were retrieved from KEGGREST to create sub-group databases, which is based on the NCBI taxonomy system.

3.5.2 Other updates to Seq2Fun algorithm

Seq2Fun version 2.0 also includes improvements to the memory usage by optimizing both the number of reads (*n* = 1000) in each pack and number of packs (*n* = 80) allowed in RAM to reduce the RAM consumption. In addition, more memory was allocated to heaps instead of stacks to further reduce the memory usage. Version 2.0 also includes a new function called SeqTract to retrieve mapped reads based on the list of genes. This subset of reads can be used for targeted gene assembly, which could be useful for primer design, isoform analysis, and phylogenetic analysis of specific genes.

3.5.3 Methods for case studies #1 and #2

Case study #1

Zebrafish RNA-seq data were obtained from NCBI's Sequence Read Archive (SRA; https://trace.ncbi.nlm.nih.gov/Traces/sra/) at sample accessions SRR13332314 - SRR13332325. Files were downloaded and converted from SRA to FASTQ format using the NCBI SRA ToolKit (version 2.11.3) before being uploaded to EcoOmicsAnalyst. Lobster RNA-seq data were collected by the authors of this paper. Briefly, stage I lobster larvae were exposed for 24hrs to four doses of WAF (10%, 19%, 37% and 72%) a positive control (1-methylnaphthalene, at 0.3

mg/L corresponding to the estimated concentration of EC20) or a negative control (0.22 µm filtered seawater). Due to the lack of effects reported on survival, molting and respiration after WAF exposures, only the highest dose of WAF (72% WAF) was considered for transcriptomics analysis. RNA extraction of whole larvae exposed to 72% WAF (*n* = 7), methylnaphthalene (*n* = 6) and filtered seawater (*n* = 6) was performed using Trizol. The transcriptomes were sequenced using Novaseq Illumina at 28M reads per library. FASTQ files were downloaded from the Genome Quebec portal and uploaded to EcoOmicsAnalyst.

For the Kallisto workflow, zebrafish samples were aligned to the GRCz11 *Danio rerio* genome assembly (accession: GCA_000002035.4) and lobster samples were aligned to the GMGI_Hamer_2.0 *Homarus americanus* genome assembly (accession: GCA_018991925.1). The minimum quality score parameter was set to 25. For the Seq2Fun workflow, zebrafish and lobster samples were aligned to the "Fishes" and "Invertebrates" databases respectively, with the parameters "maximum number of mismatches" = 2, "minimum matching length" = 19, and "minimum matching BLOSUM62 score" = 80.

Each of the four counts matrices (zebrafish: Kallisto and Seq2Fun; lobster: Kallisto and Seq2Fun) were analyzed with ExpressAnalyst. For upload, the appropriate organism and IDs were selected (Zebrafish - Ensembl, American lobster - RefSeq, Fishes - Seq2Fun ID, Invertebrates - Seq2Fun ID), data type was set to RNA-seq, and the gene summary method was set to "sum". Data were filtered to remove those with low abundance and low variation by setting the variance filter to 15 and the abundance filter to 4. Data were normalized using the "Log2 counts per million" option, which uses the *limma voom* R package (Law *et al.*, 2014).

Differential analysis was performed with the edgeR method, which uses the *edgeR* R package (Robinson *et al*., 2009). Each treatment group was compared to the control group using the "specific comparison" option. Genes were defined as differentially expressed if the adjusted p-value (FDR method) was less than 0.05 and the abs(log2FC) was greater than 1.5. For each contrast, the list of DEGs was analyzed for enriched KEGG, GO BP, GO MF, and GO CC gene sets using the "ORA Networks" tool. A pathway was defined as significantly enriched if the adjusted p-value (FDR method) was less than 0.05 and there were at least five DEGs in the gene set.

Case study #2

Salamander RNA-seq data were obtained from NCBI's SRA at sample accessions SRR7499348-SRR7499365, excluding sample SRR7499350 which was identified as an outlier by QA/QC performed by the original publication (Dwaraka *et al.,* 2019). Files were downloaded and converted from SRA to FASTQ format using the NCBI SRA ToolKit (version 2.11.3) before being uploaded to EcoOmicsAnalyst. Samples were aligned to the "Vertebrates" database using Seq2Fun. The count matrix was analyzed with ExpressAnalyst using the same methods as case study #1, until the differential expression step where case study #2 used a two-factor analysis. Time was set as the primary factor, and species as the secondary factor with the secondary factor defined as a "blocking factor". Then, a "specific comparison" was performed between 'time0' and 'time24'. Following the original publication, genes were considered differentially expressed if the adjusted p-value was less than 0.1. The list of DEGs was split into up and down-regulated genes, and each list was analyzed for enriched KEGG, GO BP, GO MF, and GO CC gene sets using the “Interactive Volcano Plot” tool. A pathway was defined as significantly enriched if the adj. p-value (FDR method) was less than 0.05 and there were at least five DEGs in the gene set.

3.5.4 IMPLEMENTATION OF WEB TOOLS

EcoOmicsAnalyst was implemented based on the PrimeFaces component library (version 11) (www.primefaces.org) and R (version 4.1.2). FASTQ file upload is handled with proFTP (version 1.3.6), user accounts are stored in a MariaDB instance (version 10.5.12), and job management is done with Slurm (version 20.11.2). All visualizations in EcoOmicsAnalyst were prepared with the R package *ggplot2*. At the time of publication, the versions for reads quality check and quantification were: fastp (version 0.21.1), Kallisto (version 0.46.1), and Seq2Fun (version 2.0.2).

ExpressAnalyst was also implemented based on the PrimeFaces component library (version 11) and R (version 4.1.2). Visual analytics tools are heatmap, volcano plot, gene set network, 3D PCA, and chord diagram, which are implemented using sigma.js (URL) and CanvasXpress (URL). For more details, see the latest publication on NetworkAnalyst, the original source of many ExpressAnalyst features (Zhou *et al*., 2019).

EcoOmicsDB was implemented based on the PrimeNG component library (version 13), R (version 4.1.2), and SQLite.

3.6 Acknowledgements

This project was funded by Genome Canada and Génome Québec. We thank early users who provided valuable feedback, including Yeonseon Jeon, Jonathan Sangiovanni, and Krittika Mittal. We also thank Joseph O'Brien and Zhiqiang Pang for technical assistance.

3.7 References

Bray, N. L., Pimentel, H., Melsted, P., Pachter, L. **2016**. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525-527.

Burns, F. R., Cogburn, A. L., Ankley, G. T., Villeneuve, D. L., Waits, E., Chang, Y. J., *et al*. **2016**. Sequencing and de novo draft assemblies of a fathead minnow (*Pimephales promelas*) reference genome. *Environmental Toxicology and Chemistry*, 35(1), 212-217.

Conesa A, Götz S. **2008**. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, 619832.

Dwaraka, V. B., Smith, J. J., Woodcock, M., Voss, S. R. **2019**. Comparative transcriptomics of limb regeneration: Identification of conserved expression changes among three species of Ambystoma. *Genomics*, 111(6), 1216-1225.

Ekblom, R., Galindo, J. **2011**. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), 1-15.

Emms, D. M., & Kelly, S. **2019.** OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 1-14.

Freedman, A. H., Clamp, M., Sackton, T. B. **2021**. Error, noise and bias in de novo transcriptome assemblies. *Molecular Ecology Resources*, 21(1), 18-29.

Gene Ontology Consortium. **2019**. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1), D330-D338.

Hansen, P. D., Dizer, H., Hock, B., Marx, A., Sherry, J., McMaster, M., Blaise, C. **1998**. Vitellogenin – a biomarker for endocrine disruptors. *TrAC Trends in Analytical Chemistry*, 17(7), 448-451.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., Kumar, S. **2015**. Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, 32(4), 835-845.

Huang, J., Wang, Q., Liu, S., Zhang, M., Liu, Y., Sun, L., Wu, Y., Tu, W. **2021**. Crosstalk between histological alterations, oxidative stress and immune aberrations of the emerging PFOS alternative OBS in developing zebrafish. *Science of the Total Environment*, 774, 145443.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. **2016**. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1), D457-D462.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K. **2017**. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1), D353-D361.

LaLone, C. A., Basu, N., Browne, P., Edwards, S. W., Embry, M., Sewell, F., Hodges, G. **2021**. International Consortium to Advance Cross-Species Extrapolation of the Effects of Chemicals in Regulatory Toxicology. *Environmental Toxicology and Chemistry*, 40(12), 3226-3233.

Law, C. W., Chen, Y., Shi, W., Smyth, G. K. **2014**. *voom*: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), 1-17.

Liedtke, H. C., Gower, D. J., Wilkinson, M., Gomez-Mestre, I. **2018**. Macroevolutionary shift in the size of amphibian genomes and the role of life history and climate. *Nature Ecology and Evolution*, 2(11), 1792-1799.

Liu, P., Ewald, J., Galvez, J. H., Head, J., Crump, D., Bourque, G., Basu, N., Xia, J. **2021**. Ultrafast functional profiling of RNA-seq data for nonmodel organisms. *Genome Research*, 31(4), 713-720.

Martin JA, Wang Z. **2011**. Next-generation transcriptome assembly. *Nature Reviews Genetics* 12:671–682.

Martinson, J. W., Bencic, D. C., Toth, G. P., Kostich, M. S., Flick, R. W., See, M. J., *et al*. **2021**. De Novo Assembly of the Nearly Complete Fathead Minnow Reference Genome Reveals a Repetitive But Compact Genome. *Environmental Toxicology and Chemistry*. Online only.

Nowoshilow, S., Schloissnig, S., Fei, J. F., Dahl, A., Pang, A. W., Pippel, M., *et al*. **2018**. The axolotl genome and the evolution of key tissue formation regulators. *Nature*, 554(7690), 50-55.

Robinson, M. D., McCarthy, D. J., Smyth, G. K. **2010**. *edgeR*: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., *et al*. **2020**. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*. Online only.

Sundaram, A., Tengs, T., Grimholt, U. **2017**. Issues with RNA-seq analysis in non-model organisms: a salmonid example. *Developmental and Comparative Immunology*, 75, 38-47.

Tu, W., Martinez, R., Navarro-Martin, L., Kostyniuk, D. J., Hum, C., Huang, J., *et al*. **2019**. Bioconcentration and metabolic effects of emerging PFOS alternatives in developing zebrafish. *Environmental Science and Technology*, 53(22), 13427-13439.

Voshall A, Moriyama EN. **2018**. Next-generation transcriptome assembly: strategies and performance analysis. In: *Bioinformatics in the era of post genomics and Big Data*, pp. 15–36. IntechOpen, London.

Wachi, N., Matsubayashi, K. W., Maeto, K. **2018**. Application of next-generation sequencing to the study of non-model insects. *Entomological Science*, 21(1), 3-11.

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E., Basu, N., Xia, J. **2019**. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research*, 47(W1), W234-W241.

CONNECTING PARAGRAPH

Chapter 3 presented web-based tools for the comprehensive analysis of raw RNA-sequencing data from any species. A typical analysis performed by these tools produce lists of impacted genes and pathways, which can contain hundreds to thousands of items. These types of results are overwhelming to interpret for people who do not have extensive bioinformatics experience.

Chapter 4 presents EcoToxModules, which are custom gene sets for organizing and analyzing transcriptomics data in the context of environmental toxicology. EcoToxModules aim to provide a high-level summary to help make sense of the extreme detail produced by typical ‘omics pipelines. Chapter 4 defines two types of EcoToxModules, one using a completely data-driven approach (statistical modules) and the other using a database of known molecular pathways (functional modules). Both sets were used to summarize changes in gene expression in a large microarray dataset measured in fathead minnow liver tissue. The overall conclusion is that while their performance in capturing major trends in the data is comparable, the functional modules were far easier to interpret. In addition, the functional modules were designed to be easily extended to new species with minimal additional work. Since the publication of this chapter, the functional EcoToxModules have been defined in terms of the protein orthologs from EcoOmicsAnalyst and incorporated into ExpressAnalyst.

This chapter was published in the journal *Environmental Science and Technology* in February of 2020 (volume 54 (7), pages 4376-4387). The candidate is the sole first author. It is additionally co-authored by Othman Soufan, Doug Crump, Markus Hecker, Jianguo Xia, and Niladri Basu, the candidate’s supervisor. Supplemental information is provided in the thesis appendix.

CHAPTER 4. ECOTOXMODULES: CUSTOM GENE SETS TO ORGANIZE AND ANALYZE TOXICOGENOMICS DATA FROM ECOLOGICAL SPECIES

Jessica D. Ewald[1], Othman Soufan[1], Doug Crump[2], Markus Hecker[3], Jianguo Xia[1], Niladri Basu[1]

[1] Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Canada

[2] Ecotoxicology and Wildlife Health Division, Environment and Climate Change Canada, National Wildlife Research Centre, Ottawa, Canada

[3] School of the Environment & Sustainability and Toxicology Centre, University of Saskatchewan, Saskatoon, Canada

4.1 Abstract

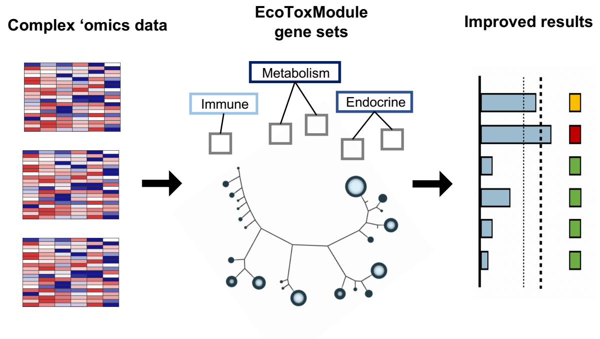


Figure 4-1: Graphical abstract published in manuscript.

Traditional results from toxicogenomics studies are complex lists of significantly impacted genes or gene sets, which are challenging to synthesize down to actionable results with a clear interpretation. Here we defined two sets of 21 custom gene sets, called the functional and statistical EcoToxModules, in fathead minnow (*Pimephales promelas*) to 1) re-cast pre-defined molecular pathways into a toxicological framework, and 2) provide a data-driven, unsupervised grouping of genes impacted by exposure to environmental contaminants. The functional EcoToxModules were identified by re-organizing KEGG pathways into biological processes that are more relevant to ecotoxicology based on input from expert scientists and regulators. The statistical EcoToxModules were identified using co-expression analysis of publicly available microarray data (n = 303 profiles) measured in livers of fathead minnows after exposure to 38 different conditions. Potential applications of the EcoToxModules were demonstrated with two case studies that represent exposure to a pure chemical and to environmental wastewater samples. In comparisons to differential expression and gene set analysis, we found that EcoToxModule responses were consistent with these traditional results. Additionally, they were easier to visualize and quantitatively compare across different conditions, which facilitated drawing conclusions about the relative toxicity of the exposures within each case study.

4.2 Introduction

Whole-transcriptome profiling based on microarray or RNAseq represents a powerful approach to comprehensively measure changes in gene expression. This type of data can provide mechanistic insights into the effects of exposures of organisms to environmental contaminants (Alexander-Dann *et al.*, 2018), particularly when using alternative toxicity testing strategies such as those called for by the amended United States Toxic Control Substances Act and the European Union REACH regulations (USEPA, 2015; Van der Jagt *et al.*, 2004). However, transcriptomics data are complex and challenging to interpret for users who may not have experience with the advanced bioinformatics required to synthesize these data into understandable, actionable and trusted results (Martyniuk *et al.*, 2018; Vachon *et al.*, 2017; Zaunbrecher *et al.*, 2017). As discussed in multiple papers published by the OpenTox Association, there is a need for open, computable, and standardized vocabularies for analyzing biological outcomes that can help to integrate measurements, including transcriptomics data, across levels of biological organization and within a toxicological framework (Hardy *et al.*, 2012; Tcheremenskaia *et al.*, 2012).

One strategy for translating gene-level signatures into higher-level endpoints that are more meaningful, reproducible, and relevant to toxicology is to analyze whole-transcriptome data at the cellular pathway or biological process level (Dean *et al.*, 2017; Farmahin *et al.*, 2017). Many toxicogenomics studies employ gene set analysis with traditional pathway libraries developed for broader biological research, such as KEGG, the Gene Ontology (GO), or Reactome (Alexander-Dann *et al.*, 2018; Bourdon-Lacombe *et al.*, 2015). However, these traditional methods have some limitations for analyzing toxicogenomics data. First, many of the pathways in these libraries have little relevance to toxicology because they were designed for more general biological and biomedical applications using the human and mouse genome, posing challenges when trying to extrapolate to ecotoxicologically relevant species (Perkins *et al.*, 2013). Also, there are hundreds to thousands of pathways in each library, so the results are still relatively complex and difficult to interpret and quantitatively compare across different exposures. Thus, there is a need for a toxicology-focused organizational scheme for existing molecular pathways to help make functional analysis of toxicogenomics data more relevant to ecotoxicological sciences, regulatory activities, and environmental management and monitoring.

Another strategy for organizing individual genes into larger gene sets is to group them based on statistical similarity within large, data-driven networks using unsupervised machine learning methods (Kitano, 2002; Ma *et al.*, 2012; Pavlopoulos *et al.*, 2011). This type of analysis infers interactions between genes by computing the pairwise correlation or mutual information of gene expression values across many samples. Clustering algorithms are then used to detect groups of co-expressed genes within the computed network, with the rationale that co-expressed genes are likely to be associated with the same biological process (Langfelder *et al.*, 2007; Van Dam *et al.*, 2017). Popular software for co-expression network analysis include ARACNE, DiffCor, and Weighted Gene Co-expression Network Analysis (WGCNA) (Margolin *et al.*, 2006; Fukushima, 2013; Langfelder *et al.*, 2008). Since these computationally generated gene sets do not depend on transcriptome annotation, and ecological species typically lack high-quality annotation compared to popular mammalian models, this makes them especially well-suited for analyzing ‘omics data in ecotoxicology studies (LaLone *et al.*, 2013; Pagé-Lariviére *et al.*, 2019). However, there is a paucity of existing studies that used co-expression network analysis to analyze toxicogenomics data from ecological species (Perkins *et al.*, 2011; Williams *et al.*, 2011).

In this era of "big data" and predictive toxicology, translating complex 'omics results into knowledge that can help inform decision-making is challenged by a lack of toxicologically-focused vocabularies and organizational schemes for interpreting these data (Hardy *et al.*, 2012). Therefore, the objective of this study was to help improve the interpretation and comparability of toxicogenomics results from ecological species by developing two custom gene set collections, called EcoToxModules, for analyzing whole-transcriptome data. In this particular case, EcoToxModule gene sets were developed for fathead minnow because of its popularity as a model species in aquatic toxicity testing, and because of the substantial amount of publicly available toxicogenomics data from this species (Ankley *et al.*, 2006; Wang *et al.*, 2016). However, the method for defining EcoToxModule gene sets is generic and can be applied to any other species with a sequenced transcriptome. The EcoToxModules were defined using two complementary approaches – one based on *a priori* knowledge of biological pathways (functional EcoToxModules) (Kanehisa *et al.*, 2016), and one based on statistical co-expression analysis of toxicogenomics data from fathead minnow (statistical EcoToxModules) (Sutherland *et al.*, 2016). This study was part of the EcoToxChip project (Basu *et al.*, 2019).

4.3 Materials and Methods

Overall the methods are separated into three phases (Figure 4-2). In phase I, fathead minnow microarray data were downloaded from NCBI's Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/), a database that consists of public gene expression data collected using either microarray or RNA sequencing technologies. The probe sequences were re-annotated by aligning them to the fathead minnow genome (NCBI assembly accession: GCA_000700825.1). In phase II, the statistical and functional EcoToxModules were defined and compared to each other based on their size, overlapping genes, and responsiveness to chemical exposure. In phase III, the data from phase I were re-analyzed to investigate how EcoToxModule-based analysis compares to traditional differential expression and gene set analysis. Two subsets of the fathead minnow microarray data were highlighted as case studies to demonstrate how EcoToxModules can be applied to analyze toxicogenomics data.

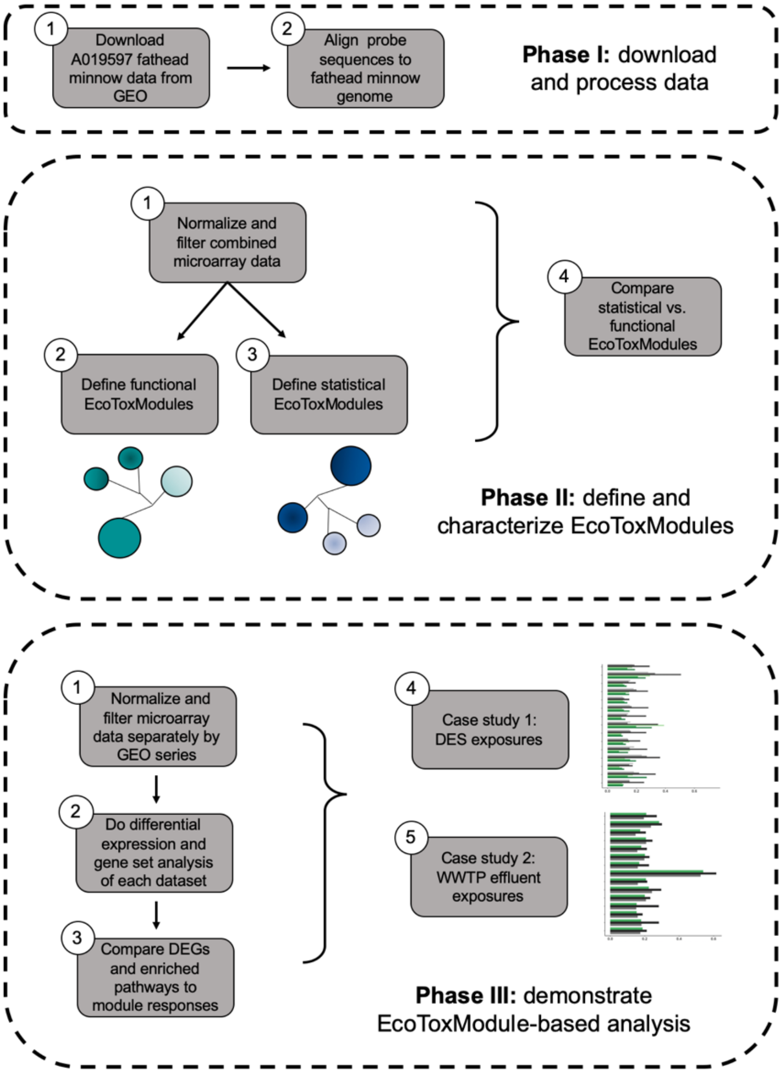

1
Download A019597 fathead minnow data from GEO
2
Align probe sequences to fathead minnow genome
Phase I: download and process data
1
Normalize and filter combined microarray data
2
Define functional EcoToxModules
3
Define statistical EcoToxModules
4
Compare statistical vs. functional EcoToxModules
Phase II: define and characterize EcoToxModules
1
Normalize and filter microarray data separately by GEO series
2
Do differential expression and gene set analysis of each dataset
3
Compare DEGs and enriched pathways to module responses
4
Case study 1: DES exposures
5
Case study 2: WWTP effluent exposures
Phase III: demonstrate EcoToxModule-based analysis

Figure 4-2: A visual representation of the different phases of the analysis workflow.

4.3.1 PHASE I: DATA PREPARATION AND PROCESSING

This study used pre-existing data from the Agilent-019597 (GPL10259) one-color microarray platform, due to the relatively high number of toxicogenomics studies that have been conducted using this platform (Wang *et al.*, 2016). The data (n = 303 microarray profiles) contain gene expression measurements in liver tissue from fathead minnows of both sexes and multiple life stages. We chose to focus on liver tissue because it is a common target tissue for many toxic chemicals and thus of prime interest in ecotoxicological studies (Wang *et al.*, 20016). They are from 11 distinct data series, available in NCBI's GEO DataSets database, and were generated by exposure studies that examined the toxicity of environmental water samples from 10 locations or varying doses of 6 chemicals (Tables 4-1 and S1) (Adedeji *et al.*, 2012; Gust *et al.*, 2011; Loughery *et al.*, 2018; Martinović-Weigelt *et al.*, 2014; Rodriguez-Jorquera *et al.*, 2015; Vidal-Dorsch *et al.*, 2013). These data are a subset of a larger dataset compiled in a previous study on fish connectivity mapping (Wang *et al.*, 2016). The Agilent-019597 microarray was developed using expressed sequence tag data before the fathead minnow genome was sequenced (Adedeji *et al.*, 2012; Larkin *et al.*, 2007). We aligned the probe sequences to the reference fathead minnow genome (GCA_000700825.1) using BLASTn so that they could be annotated with KEGG pathways in downstream steps (Burns *et al.*, 2016). Probes that were either significantly aligned (e-value < 10^{-6}) to multiple distinct locations on the genome or had no significant alignments were removed from downstream analyses.

Table 4-1: Overview of publicly available microarray data analyzed in Phases II and III.

GEO accession	Sample size (*n*)	Exposure	Sex	Life stage
GSE21100	60	Cyclonite (RDX)	female	adult
GSE21102	24	Cyclonite (RDX)	female	juvenile
GSE23521	18	Environmental sample (2 streams)	male	adult
GSE29350	38	Effluent (2 WWTPs); Estrogen (E2)	male	adult
GSE36465	12	Diethylstilbestrol (DES)	female	adult
GSE36466	11	Diethylstilbestrol (DES)	female	adult
GSE37550	15	Environmental sample (2 streams); Effluent (1 WWTP)	male	adult
GSE44839	20	Phenanthrene (PHN)	female	adult
GSE44840	28	Phenanthrene (PHN)	female	adult
GSE49098	62	Effluent (3 WWTPs)	male	adult
GSE54506	15	Perfluorooctanesulfonic acid (PFOS); Perfluorinated compounds (PFCs)	male	adult

4.3.2 Phase II: EcoToxModule definition and characterization

Defining functional EcoToxModules

The objective of the functional EcoToxModules was to group functionally annotated genes using a hierarchical organization of physiological categories and biological processes that resonate with researchers and decision-makers within the field of toxicology. The specific biological processes represented by the functional EcoToxModules were identified using a mixed methods approach. First, as part of the EcoToxChip project (www.ecotoxchip.ca), a panel of five principal investigators with expertise in ecotoxicology identified biological processes and pathways relevant to the fields of ecotoxicological sciences and applied environmental risk assessment, monitoring, and management by focusing on known toxicity mechanisms that affect survival, growth, and reproduction outcomes (Basu *et al.*, 2019). The team identified an initial list by consulting the Comparative Toxicogenomics Database (ctdbase.org) and the toxicogenomics literature (Davis *et al.*, 2017), and then this list was deliberated upon and refined by members of the larger project team including stakeholders from government and industry. This exercise resulted in the identification of 21 biological processes, each corresponding to a functional EcoToxModule, sorted into five larger physiological categories (signaling, metabolism, immune, endocrine, and cellular processes). Of the 330 KEGG pathways that exist for vertebrates, 172 were manually selected for inclusion in one of the functional EcoToxModules based on their relevance to the 21 biological processes and aided by the BRITE functional hierarchies devised by KEGG to organize individual pathways (Kanehisa *et al.*, 2007).

KEGG orthologues (KOs) are used to define groups of orthologous genes within the KEGG GENES database, and they form the basis of the reference pathway maps in the KEGG PATHWAY database (Kanehisa, 2008; Kanehisa *et al.*, 2000). Fathead minnow coding sequences were annotated with KOs by uploading the fathead minnow transcriptome to the KEGG automatic annotation server (KAAS) (Moriya *et al.*, 2007) and selecting all vertebrate species in the KEGG Genome database for sequence similarity computations. Microarray probes were assigned to functional EcoToxModules based on their alignment to fathead minnow transcript sequences. A complete list of the 21 modules and all of their KOs are available in the supporting information (SI) (Table S2).

This proposed functional EcoToxModule hierarchy was discussed with internal core EcoToxChip project members, wider project partners, and potential end-users, including scientists and regulators from Environment and Climate Change Canada, Health Canada, the

U.S. Environmental Protection Agency, the U.S. Army Core of Engineers, Shell, and Qiagen. Since 2017, we have consulted the broader environmental toxicology and risk assessment community by presenting the EcoToxModule hierarchy at six scientific conferences. Moving forward, we expect that the modules and genes within the EcoToxModule hierarchy will continue to evolve as particular users create customized and fit for purpose variations.

Defining the statistical EcoToxModules

We chose to use WGCNA to conduct the co-expression network analysis used to define the statistical EcoToxModules because it is well-documented and has been successfully used to analyze toxicogenomics data from mammalian species (Maertens *et al.*, 2018; Sutherland *et al.*, 2018). Prior to the co-expression analysis, the microarray data were first processed, background corrected and normalized with the *limma* R package using functions developed for the Agilent platform and parameters for single-channel microarrays (Langfelder *et al.*, 2008; Ritchie *et al.*, 2015). A standard WGCNA workflow was used to detect clusters of co-expressed genes within the normalized microarray dataset (Langfelder *et al.*, 2007; Zhao *et al.*, 2010). A detailed description of the processing, normalization, and WGCNA workflow are included in the SI. The detected clusters were annotated with enriched KEGG pathways and Gene Ontology (GO) biological process, molecular function, and cellular component terms using hypergeometric tests (adjusted p-value < 0.05, false discovery rate [FDR] method). The human homologue annotations in the GEO Platform file (GPL10259) were used for enrichment analysis, following previous studies using this microarray (Perkins *et al.*, 2011; Martinović-Weigelt *et al.*, 2014).

To compare quantitative characteristics of the functional and statistical EcoToxModules, we computed different statistics measuring gene set similarity and module responses. Gene set similarity was measured by calculating the number of overlapping genes and the Jaccard index between pairs of functional and statistical EcoToxModules. Two different module responses were calculated, the fold change and the eigengene. Here, the module fold change was defined as the average of the absolute $\log_2$FC of the EcoToxModule probes. Following a previous study that conducted co-expression network analysis based on the TG-GATES database (Sutherland *et al.*, 2016; Sutherland *et al.*, 2018; Igarashi *et al.*, 2014), the eigengene was defined as the first principal component computed from the $\log_2$FCs of the EcoToxModule probes.

4.3.3 Phase III: Comparison of EcoToxModule-based analysis to traditional bioinformatics

The objective of Phase III was to investigate how EcoToxModule-based analysis compares to traditional differential expression and gene set analyses. The data from each study within the fathead minnow microarray dataset were re-analyzed using traditional differential expression and gene set analyses, and these results were compared to EcoToxModule summaries of the data. Prior to differential expression analysis, the data were normalized separately for each study using the same methods as in Phase I (Ritchie *et al.*, 2015; Edwards, 2003; Silver *et al.*, 2008). Differentially expressed genes between each exposure and associated controls were selected using FDR adj. p-value (< 0.05) and abs(log2FC) (> 1.5). The genes were analyzed for significantly enriched GO biological process (GO BP) and KEGG pathway gene sets (FDR adj. p-value < 0.05) using hypergeometric tests with the *HTSanalyzeR* R package (GO Consortium, 2014; Wang *et al.*, 2011).

Two subsets of the data were highlighted as case studies to demonstrate how the EcoToxModules can be used for analyzing whole-transcriptome data in more detail. The case studies reflect one pure chemical exposure (diethylstilbestrol (DES); GSE36465 and GSE36466; case study 1) (Adedeji *et al.*, 2012), and one environmental water sample exposure (effluent from a wastewater treatment plant (WWTP) in Ely, Minnesota, USA; GSE49098; case study 2) (Martinović-Weigelt *et al.*, 2014). These case studies were chosen to demonstrate what EcoToxModule results could look like for whole-transcriptome data collected during chemical risk assessment (case study 1) and environmental monitoring (case study 2), two important regulatory activities in environmental toxicology.

The purpose of the DES exposure study was to collect information on the *in vivo* effects of exposure to environmentally relevant concentrations of DES, a synthetic estrogen, in fathead minnow. A complete description of the experimental design and methods are described in Adedeji *et al* (Adedeji *et al.*, 2012). Briefly, the authors exposed sexually mature fathead minnows to varying concentrations of DES (control, 1, 10, and 100 ng/L, *n* = 6 or 8 replicate tanks per condition with 3 individual fish of each sex per tank) for four days, and then allowed the fish to recover without DES exposure for four additional days. Half of the fish were sampled to measure biological endpoints at the end of the DES exposure, and half were sampled at the end of the recovery period. Gene expression measurements were made in the liver of female fathead minnow using the Agilent-019597 microarray for a subset of both the DES exposure (*n* = 3 individuals per condition) and recovery (*n* = 2 or 3 individuals per condition) groups.

One major objective of the WWTP study was to assess effluent-impacted field sites with ‘omics technologies. A complete description of the experimental design and methods are described in Martinovic-Weigelt *et al*. (Martinović-Weigelt *et al.*, 2012). The authors exposed sexually mature fathead minnows to water collected upstream of, downstream of, and at the effluent discharge location of three different WWTPs, as well as an external control exposed to filtered Lake Superior water (*n* = 3 replicate tanks per condition with 4 individual fish of each sex per tank). Gene expression measurements were made in the liver of male fathead minnow using the Agilent-019597 microarray (*n* = 6-7 individuals per condition). In this case study, we re-analyzed the data from the WWTP field site in Ely, Minnesota, USA.

4.4 Results and Discussion

4.4.1 Re-annotation of microarray probe sequences

Of the 15, 208 probe sequences on the A019597 microarray, *n* = 4, 317 could be significantly and uniquely aligned to coding transcripts in the fathead minnow genome. The remaining probes were aligned to multiple locations (*n* = 2, 463), aligned to locations predicted to be non-coding (*n* = 6, 897), or had no significant alignments to the genome (*n* = 1, 531). While the percentage of probes aligned to predicted non-coding sequences is high (45%), these probes have distinct expression patterns compared to the probes aligned to predicted coding sequences (Figure S3). Additionally, a previous study that similarly aligned the highly used Affymetrix Mouse 430 2.0 microarray probe sequences to the mouse genome found that 67 089 out of 496 468 probes (14%) aligned to predicted non-coding regions (Liao *et al.*, 2011). Only probes that could be aligned to coding sequences were included in the EcoToxModules.

4.4.2 General comparison of the statistical and functional EcoToxModules

A total of 1761 unique probes were identified across the two sets of 21 modules, with 962 in the functional EcoToxModules, 1097 in the statistical EcoToxModules, and 298 in both (Table S3, Table S4).

The number of probes per functional EcoToxModule ranged from 13 to 352 (median = 64). Each functional EcoToxModule was further categorized into one of five more general biological categories (Table S2, Table S5). Since KOs can be part of more than one KEGG pathway (Kanehisa, 2008), some probes were in multiple functional EcoToxModules. In all pairwise comparisons of functional EcoToxModules, the Jaccard index was < 0.3, indicating a small number of overlapping genes (Figure S4).

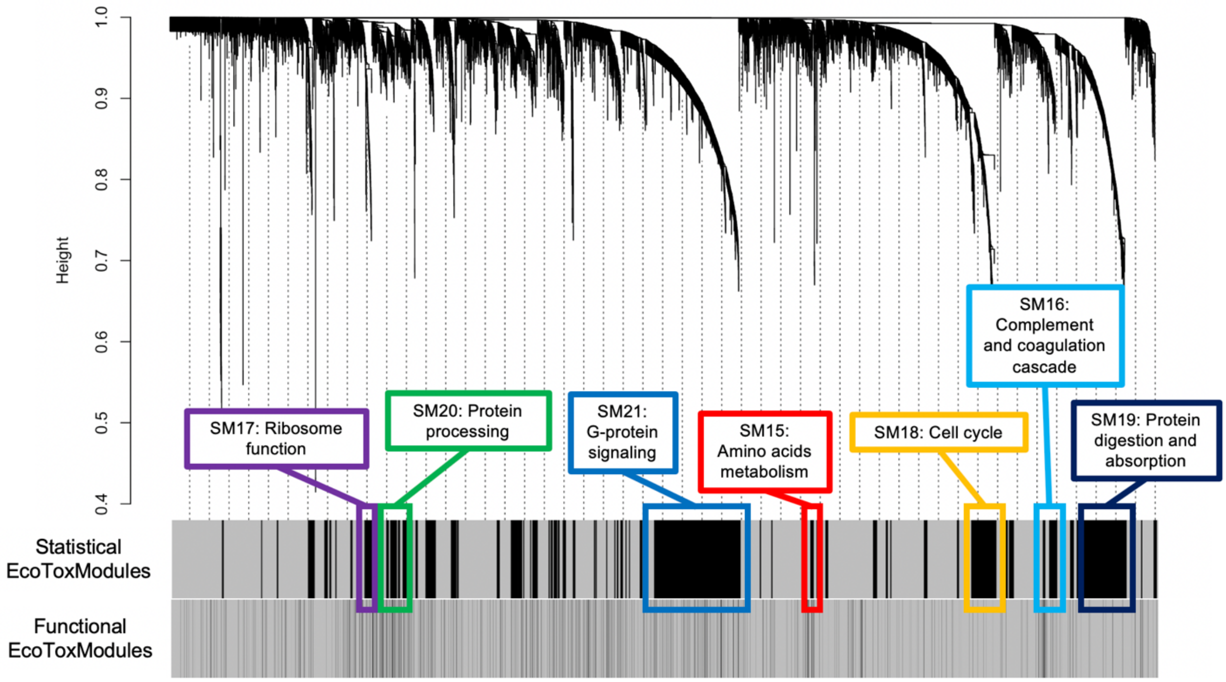


Figure 4-3: Co-expression network dendrogram showing the locations of the EcoToxModule probes. Each leaf in the dendrogram corresponds to one probe. Probes were hierarchically

clustered based on their pairwise correlation scores. The first band displays the locations of the clusters detected in the dendrogram, including both predicted coding and non-coding probes. The statistical EcoToxModules consist of the coding probes within the detected clusters. Clusters with a consistent biological theme in their gene set analysis results are highlighted. The second band displays the locations of the probes from the functional EcoToxModules.

The number of probes per statistical EcoToxModule ranged from 7 to 408 (median = 23). Seven of the statistical EcoToxModules were significantly enriched with multiple KEGG pathways and GO terms (Table S2, Table S6, Figure 4-3). Since biological annotations were inconsistent, the statistical EcoToxModules were named with alphanumeric IDs.

One major strength of the functional EcoToxModules is that each one has clear biological annotations, whereas biological annotation could only be found for 7/21 statistical EcoToxModules (Table S6). Previous studies have also noted that a limitation of co-expression analysis is the difficulty of establishing biological relevance for each detected cluster (Perkins *et al.*, 2011; Sutherland *et al.*, 2018; Abdul Hameed *et al.*, 2016). One strength of the statistical EcoToxModules is that they were better explained by their eigengenes (mean $R^2 = 0.44$, standard deviation = 0.15) compared to the functional EcoToxModules (mean $R^2 = 0.11$, standard deviation = 0.039), indicating that there is less variability in probe-level responses within the same module for a given exposure (Figure S5). This was expected since principal component analysis is more effective at capturing most of the variation in the first few principal components if the predictors are highly correlated to each other.

The functional and statistical EcoToxModules overlapped for all but two of the statistical modules; however, the number of shared genes was very low for nearly all pairs of statistical and functional EcoToxModules (Figure S6). Regardless, even though there was minimal overlap, the eigengenes produced similar global patterns of response across all of the exposure conditions for both the functional and statistical EcoToxModules (Figure 4-4A vs. 4-4B). In both heatmaps, the DES, DES recovery, E2, and WWTP4 – WWTP6 exposures cluster together, and the DES and DES recovery conditions displayed the highest transcriptional activity. This supports previous claims that module-based approaches of analyzing whole-transcriptome data are capable of capturing stable, reproducible trends within the complex data as compared to gene-level or pathway-level analysis (Sutherland *et al.*, 2018; Abdul Hameed *et al.*, 2016; Soufan *et al.*, 2019). By focusing on a higher level of biological organization, organizing the data with both the statistical and functional EcoToxModules can be an effective method for visualizing and understanding the "big picture" across many different exposures.

Overall, the EcoToxModule responses were representative of major trends in the differential expression analysis results. The sums of the 21 EcoToxModule fold changes for each exposure were positively correlated with the number of DEGs for both the statistical ($\rho = 0.82$) and functional ($\rho = 0.76$) EcoToxModules. This indicates that exposures with higher numbers of DEGs also had larger EcoToxModule fold changes. Additionally, the most perturbed functional EcoToxModules were biologically consistent with the significant gene set analysis results (Table S8). It was difficult to make a similar comparison for the statistical EcoToxModules since only one third of these modules had clear biological annotations.

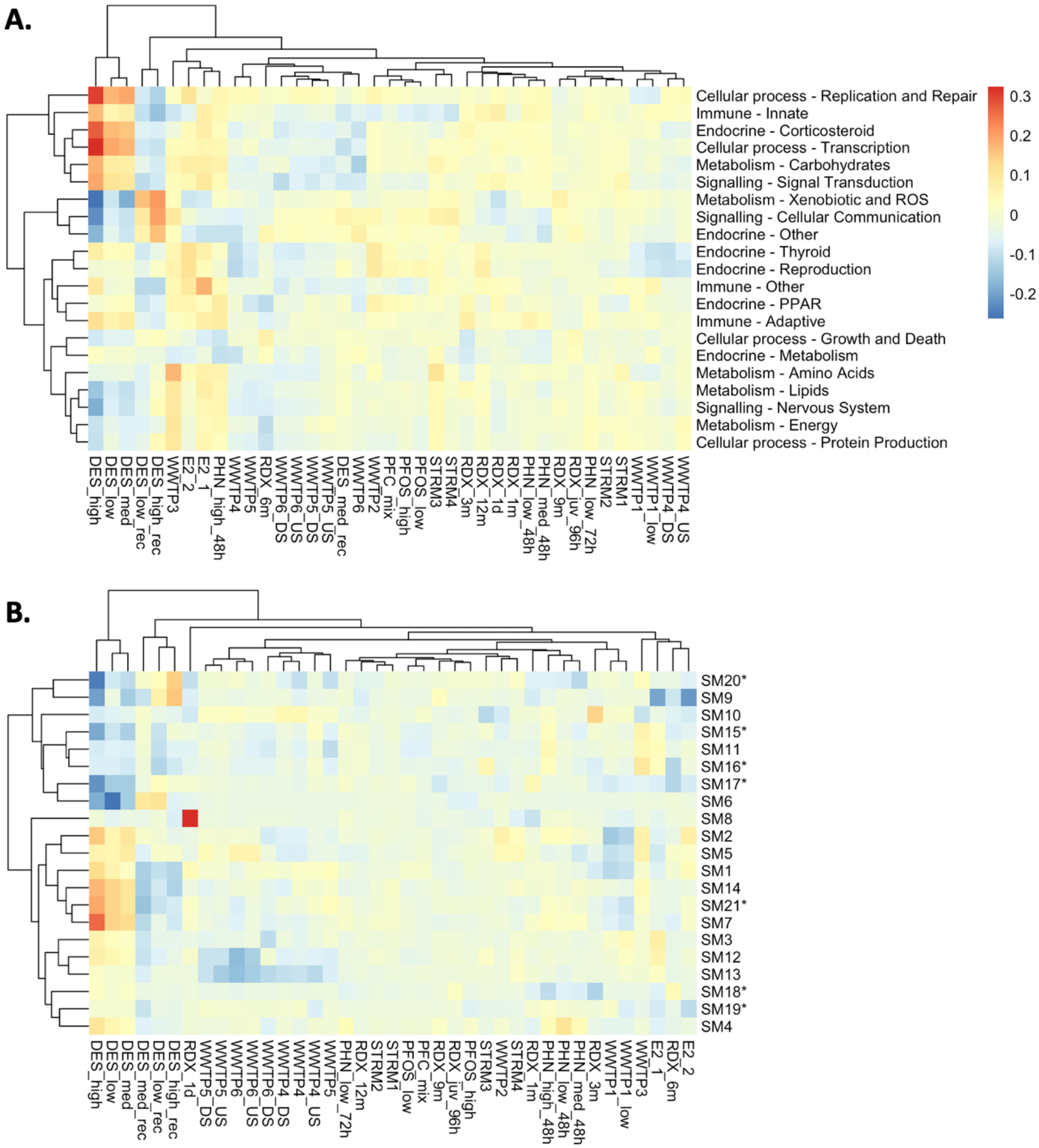


Figure 4-4: Heatmaps of EcoToxModule eigengene expression for all exposure conditions. Rows correspond to EcoToxModules, columns correspond to the exposure conditions, and squares correspond to the eigengene expression. The eigengene is calculated as the 1st principal component of log2FC of module probes for A: functional EcoToxModules, and B: statistical EcoToxModules. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. P-val < 0.05, FDR).

Case studies: overview

Subsets of the data were highlighted as case studies to demonstrate in more detail how the EcoToxModules can be helpful for organizing, analyzing, and visualizing whole transcriptome toxicogenomics data. The two case studies were chosen to represent both exposures to 1) varying doses of a pure chemical (Adedeji *et al.*, 2012), and 2) complex environmental water samples (Martinović-Weigelt *et al.*, 2014). These exposures generated many DEGs, so comparisons could be drawn between the traditional bioinformatics and EcoToxModule-based results.

4.4.3 Case study 1: DES exposures (GSE36465 and GSE36466)

The differential expression analysis results showed that there was a concentration-dependent increase in the number of DEGs for the DES exposures, with *n* = 39, 80, and 336 for the 1, 10, and 100 ng/L treatment groups, respectively (Table S7). The recovery exposures showed less consistent patterns, with *n* = 87, 4, and 200 DEGs for the 1, 10, and 100 ng/L recovery treatment groups, respectively. Out of the 21 modules in each set, there were dose-dependent perturbations for 14 functional and 12 statistical EcoToxModules, and the 100 ng/L (high) treatment had the greatest perturbation in 19 functional and 16 statistical EcoToxModules (Figure 4-5). The recovery treatments had less severe perturbations compared to the exposure treatments for 19 functional and 14 statistical EcoToxModules. Thus, both sets of EcoToxModule responses reproduced the traditional differential expression results, although the functional EcoToxModule responses were more consistent compared to the statistical EcoToxModules.

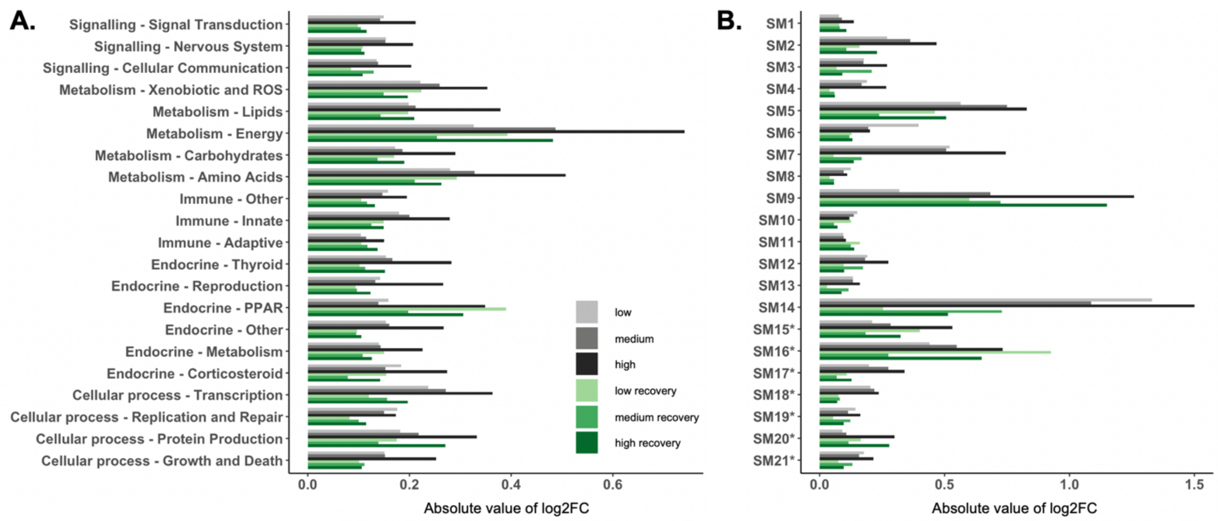


Figure 4-5: EcoToxModule fold changes for case study 1 (DES exposures). EcoToxModule fold changes are calculated as the mean of the abs(log2FC) of all probes in the module for a given exposure condition. Grey bars correspond to the low (1 ng/L), medium (10 ng/L), and high (100 ng/L) exposures to DES and green bars correspond to the same low, medium, and high exposures after a 4 day recovery period, for the A) functional EcoToxModules and B) statistical EcoToxModules. Data are from the GSE36465 and GSE36466 data series. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR).

The traditional approach of gene set analysis produced significant results for the DES 10 and 100 ng/L exposures, and the 100 ng/L recovery treatments (Table S8). The significant gene sets were generally related to cellular energy production and protein translation. For example, they included the KEGG pathways "Oxidative phosphorylation" and "Ribosome biogenesis", and the GO BP terms "tRNA aminoacylation for protein translation", "electron transport chain", and "rRNA processing" (Table S8). However, the list of significant results also included some gene sets that are clearly irrelevant to a toxicology study, such as the KEGG pathways "Huntington's disease" and "Systemic lupus erythematosus". Results such as these that include a mix of toxicologically relevant and irrelevant terms are difficult to synthesize into overall conclusions that make sense to a scientist with a general ecotoxicology background, and even more so for regulatory decision-makers who likely have less familiarity with molecular biology.

In contrast, the functional EcoToxModule responses provided simplified, toxicologically relevant summaries of these traditional results. The most and second-most perturbed modules were “Metabolism – Energy” and “Metabolism – Amino acids” for all six treatments, which was consistent with the overall biological theme represented by the significant gene sets (Table S8). Additionally, each functional EcoToxModule is associated with a quantitative measure of response, which makes it easier to compare results across different exposures and helps individuals to draw conclusions from the complex ‘omics data.

4.4.4 Case study 2: WWTP effluent exposures (GSE49098)

The differential expression analysis results showed that there was a similar number of DEGs for the upstream and downstream exposures (*n* = 84 and 79), and a higher number for the effluent exposure (*n* = 146) (Table S7). This was reproduced by the functional EcoToxModule responses as the effluent exposures caused the most severe perturbations for all 21 modules, while the upstream and downstream exposures caused perturbations that were both less severe and similar to each other (Figure 3-6a). The statistical EcoToxModule responses were less consistent with the differential expression results, with only 11 modules showing the most severe perturbation for the effluent exposures, and less similarity between the upstream and downstream exposures (Figure 3-6b).

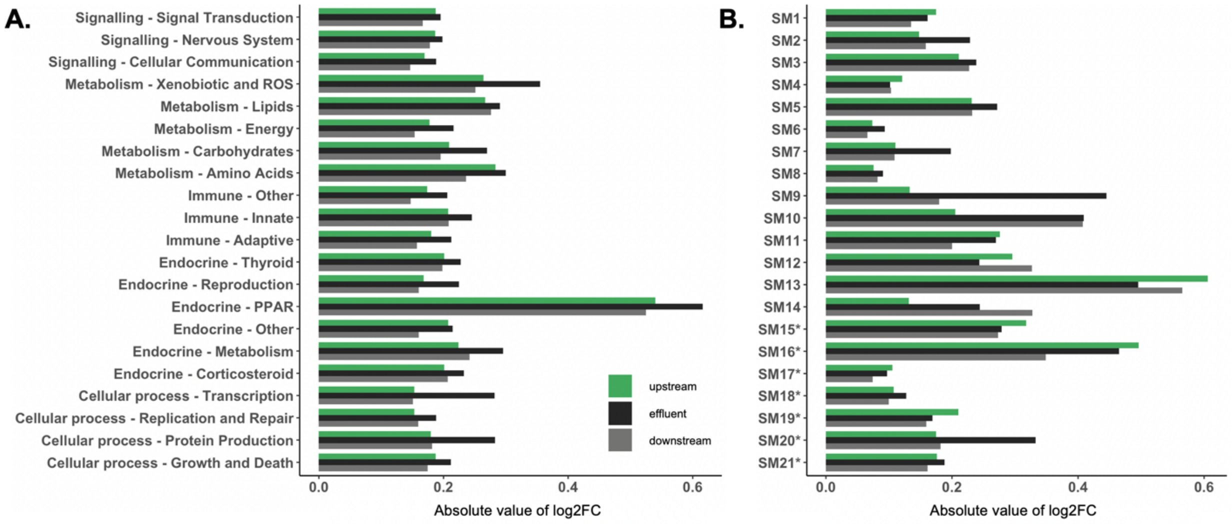


Figure 4-6: EcoToxModule fold changes for case study 2 (WWTP exposures). EcoToxModule fold changes are calculated as the mean of the abs(log2FC) of all probes in the module for a given exposure condition. Bars correspond to exposure to water samples collected upstream of (green), downstream of (grey), and at the effluent discharge point (black) of the WWTP in Ely, Minnesota, USA. Data are from the GSE49098 data series. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR).

There were significant gene set analysis results for all three exposures, with enriched KEGG pathways and GO BP terms related primarily to xenobiotic metabolism, oxidation-reduction, lipid metabolism, and PPAR signalling. Once again, the results also included biomedical terms such as the KEGG pathway "Leishmaniasis" and the GO BP term "defense response to virus" that are not linked to ecotoxicological outcomes (Table S8). The functional EcoToxModule responses were consistent with and are easier to interpret than the results from the traditional gene set analysis. The three most perturbed functional EcoToxModules were either "Endocrine – PPAR", "Metabolism – Lipids", "Metabolism – Xenobiotic and ROS", or "Metabolism – Amino acids" for all exposures.

The original publications for both case studies used the number of DEGs to compare the relative transcriptomic response across exposures, PCA plots to visualize similarity between the exposures, and lists of enriched pathways and DEGs of interest in either the text or a table to describe the main biological effects (Adedeji *et al.*, 2012; Martinović-Weigelt *et al.*, 2014). The lists of enriched pathways had an inconsistent vocabulary that ranged across different levels of detail and included biomedical terms that are unfamiliar and irrelevant to ecotoxicological sciences and regulatory activities related to ecological management and monitoring (Adedeji *et al.*, 2012; Martinović-Weigelt *et al.*, 2014). The functional EcoToxModule bar plots combine all of these main findings into one intuitive figure for each study (Figure 4-5a and 4-6a). For example, the bar plot for the DES case study clearly shows the dose-dependent response, the difference in response magnitude between the exposure and recovery samples, and the main biological processes affected by the perturbed genes and pathways (Figure 4-5a). We propose that EcoToxModule bar plots (Figure 4-5 and 4-6) be used to gain an overview of the whole-transcriptomic effects of chemical exposure, which can then help guide the interpretation of more detailed traditional differential expression and gene set analysis results.

4.4.5 STATISTICAL ECOTOXMODULES CAN BE USED TO IDENTIFY POTENTIAL BIOMARKERS

The statistical EcoToxModules were sparsely annotated because they include many probes with unknown function and, as the case studies showed, had more variable patterns of relative responses across different conditions. While their lack of functional annotation makes them less effective at summarizing traditional analysis results compared to the functional EcoToxModules, it also makes them more suitable for discovering novel biomarkers of chemical exposure.

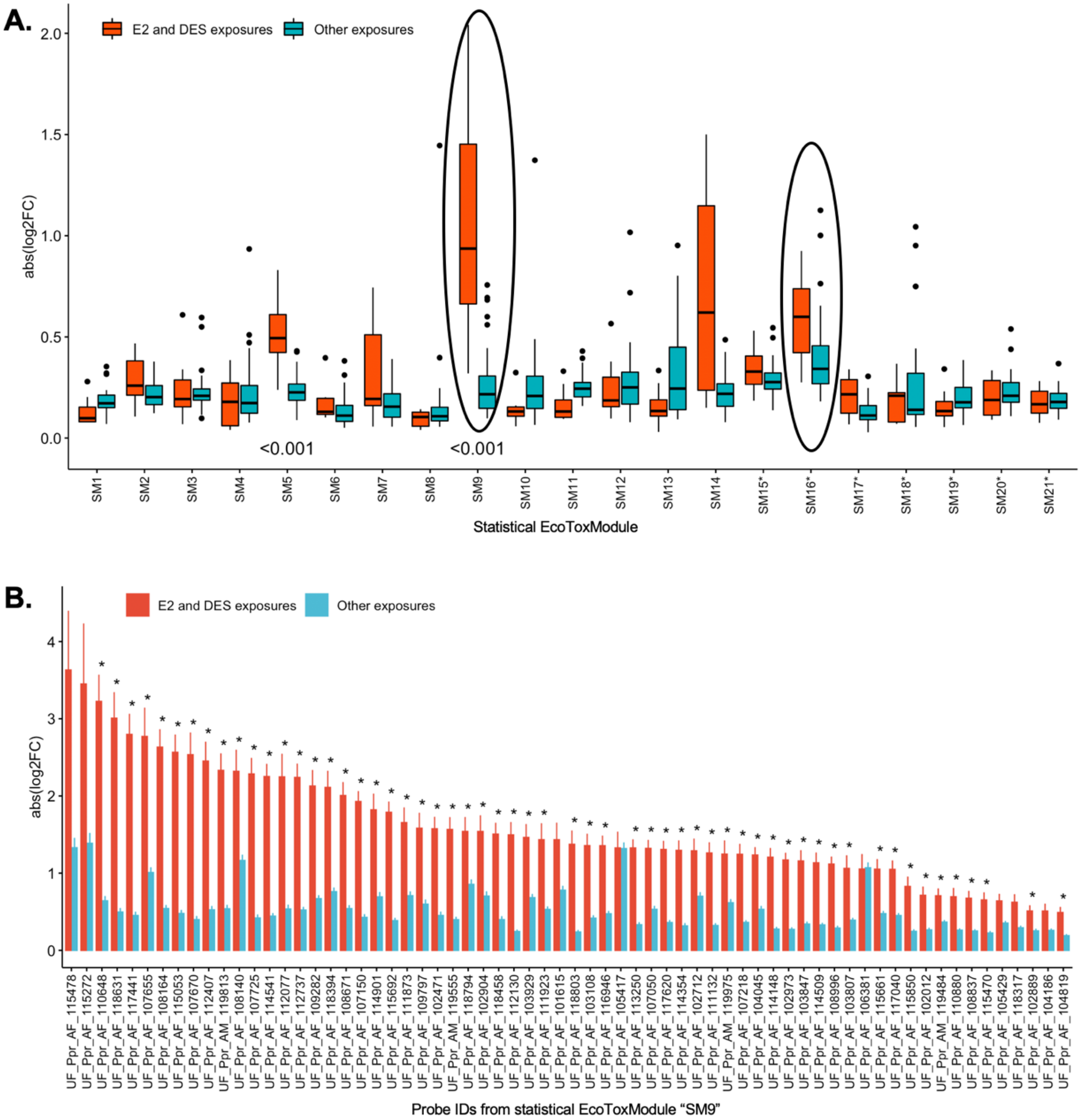


Figure 4-7: Statistical EcoToxModules that are potential exposure biomarkers. Exposure conditions (n = 38) were divided based on whether they had high estrogenic potency (n = 8, DES and E2 exposures, red fill) or medium-low estrogenic potency (n = 30, all other exposures, blue fill). Fold change, calculated as the mean of abs(log2FC), for each A) statistical EcoToxModules and B) probe within the SM9 EcoToxModule. Statistically significant differences (adj. p-val < 0.05) between the two groups were assessed using the Wilcoxon test for each module and probe. There are 66 probes in SM9, 23 of which correspond to coding sequences. P-values were adjusted using the Bonferroni method. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR).

For example, the SM16 module was in the top five most perturbed statistical EcoToxModules for 30 out of 38 exposure conditions, and thus stood out as being extremely responsive to the chemical exposures represented in the combined microarray dataset (Figure 4-7a). This module was enriched with five GO BP terms and five KEGG pathways, all of which were related to the innate immune system, and especially to complement activation (Table S6). Links between exposure to endocrine disrupting chemicals (EDCs) and modulation of the innate immune system in teleost fish have been established by many previous studies (Rehberger *et al.*, 2017; Torrealba *et al.*, 2018), including several that observed the perturbation of pathways related to complement activation with whole-transcriptome profiling (Rehberger *et al.*, 2017; Shelley *et al.*, 2012a; Shelley *et al.*, 2012b; Wenger *et al.*, 2011). One review found that parameters relating to complement activation showed significant responses 95% of the time (n = 19 exposures) for EDCs and 67% of the time (n = 144 exposures) for all exposures (Rehberger *et al.*, 2017). Thus, the SM16 EcoToxModule could be a potential source of biomarkers for exposure to a range of xenobiotics, and EDCs in particular, in the fathead minnow liver.

The SM9 module is another statistical EcoToxModule of interest because it appears to be especially responsive to estrogenic exposure (Figure 4-7). The exposure conditions in the microarray dataset (n = 38 total) were grouped based on whether they had high estrogenic potency (E2 and DES conditions, n = 8), or medium to low estrogenic potency (all other conditions, n = 30). A statistically significant difference was observed between the mean SM9 EcoToxModule fold changes of the high and medium-low estrogenic exposures (adjusted p-value < 0.01, Bonferroni correction). This is partially due to the presence of transcripts for vitellogenin (*vtg3*) and a vitellogenin precursor in the module (probe IDs: UF_Ppr_AF_115478 and UF_Ppr_AF_115272), which are known biomarkers for estrogen exposure (Flouriot *et al.*, 1995; Hansen *et al.*, 1998). However, statistically significant differences (adj. p-value < 0.05, Bonferroni correction) were observed for 58 out of 66 EcoToxModule probes (Figure 4-7b), suggesting that the separation between SM9 EcoToxModule responses is only partially driven by changes in the abundance of *vtg3* transcripts. This EcoToxModule provides candidates for alternative estrogen biomarkers that, if measured together, could be more robust than measurements of single genes such as *vtg* or *esr1*. For example, the DES exposures used female fathead minnow and only found a significant change in *vtg* expression for the highest dose (Adedeji *et al.*, 2012), yet the SM9 EcoToxModule was still significantly perturbed in a dose-dependent manner for all of the exposure and recovery conditions (Figure 4-5b).

4.4.6 ECOTOXMODULES AS TOOLS TO HELP IMPROVE THE INTERPRETATION OF TOXICOGENOMICS DATA

The interpretation of complex ‘omics data is challenged by a lack of toxicologically-focused vocabularies and organizational schemes that can handle this “big data” (Hardy *et al.*, 2012; Tcheremenskaia *et al.*, 2012). The functional and statistical EcoToxModules attempt to fill this gap using two distinct yet mutually reinforcing approaches. The main advantage of the functional EcoToxModules is that they can help to identify the main biological processes affected by different exposures because they provide a concise summary and organization of known pathways that are relevant to toxicology. However, they are limited to including genes with functional annotation, which is less extensive in ecological species compared to popular mammalian models. By providing a data-driven organization of the active transcriptional space that is unbiased by functional annotation, the main strength of the statistical EcoToxModules is that they can help to uncover novel toxicity mechanisms and biomarkers within the complex data.

While the EcoToxModules address the call from groups like the OpenTox Association for more practical methods to organize and describe complex biological data (Hardy *et al.*, 2012; Tcheremenskaia *et al.*, 2012), and thus offer exciting new possibilities for synthesizing toxicogenomics data, they also have limitations. For example, one limitation of the statistical method is that users must use their own judgment to decide whether module fold changes are meaningful. In the future, we plan to use RNAseq data collected by the EcoToxChip project (Basu *et al.*, 2019) to quantify the baseline variability of module responses within control samples to estimate module-specific thresholds that, when surpassed, suggest true biological responses. It was not possible to do this in the present study because of the variation in experimental design, including control choice, across the original studies (Table S1). Other limitations relate to uncertainty about how this new approach to analyzing toxicogenomics data will be received by potential users across academia, government, and industry. While this study is a start, further research that seeks to understand user experiences across a range of case studies within different decision-making contexts is needed to guide the continued development and validation of new methods for analyzing toxicogenomics data.

Despite these limitations, both the statistical and functional EcoToxModules distill toxicogenomics data down to simple, quantitative results that represent major trends within the whole transcriptome. In doing so, they improve the comparability and interpretability of toxicogenomics data in several ways, compared to traditional methods. Since there is a small number of modules and they have quantitative responses, they are easier to visualize using tools such as heatmaps and bar plots. Also, they focus on toxicologically relevant genes and pathways, which makes their responses easier to interpret in the context of ecotoxicological sciences and regulatory decision-making activities. Applied together, the two sets of EcoToxModules are complementary and powerful new approach methods for exploiting existing knowledge of molecular pathways to improve understanding of and find meaning in complex toxicogenomics data from ecological species.

4.5 Supporting Information

- Detailed methodology for co-expression analysis
- Additional results from mapping microarray probes to the fathead minnow genome
- Additional statistical characteristics of the EcoToxModules
- Tables summarizing sample meta-data, probes in each EcoToxModule, and gene set analysis results

4.6 ACKNOWLEDGEMENTS

JE was funded by NSERC CGS M and PGS D fellowships, as well as by a Graduate Excellence Award from McGill University. We acknowledge the support of all team members especially key project managers (Pascal Poulin, Anita Masse, Jenny Eng, Emily Boulanger) and our various trainees. We thank the financial sponsors of the larger EcoToxChip project (Genome Canada, Génome Québec, Genome Prairie, the Government of Canada, Environment and Climate Change Canada, Ministère de l'Économie, de la Science et de l'Innovation du Québec, the University of Saskatchewan, and McGill University) as well as core project partners (U.S. Environmental Protection Agency, U.S. Army Corps of Engineers, Qiagen, SGS AXYS, and Shell USA).

4.7 REFERENCES

Abdul Hameed, M. D. M.; Ippolito, D. L.; Stallings, J. D.; Wallqvist, A. **2016**. Mining kidney toxicogenomic data by using gene co-expression modules. *BMC Genomics* 17, (1), 790.

Adedeji, O. B.; Durhan, E. J.; Garcia-Reyero, N. l.; Kahl, M. D.; Jensen, K. M.; LaLone, C. A.; Makynen, E. A.; Perkins, E. J.; Thomas, L.; Villeneuve, D. L. **2012**. Short-term study investigating the estrogenic potency of diethylstilbesterol in the fathead minnow (Pimephales promelas). *Environmental Science and Technology* 46, (14), 7826-7835.

Alexander-Dann, B.; Pruteanu, L. L.; Oerton, E.; Sharma, N.; Berindan-Neagoe, I.; Módos, D.; Bender, A. **2018**. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. *Molecular Omics* 14, (4), 218-236.

Ankley, G. T.; Villeneuve, D. L. **2006**. The fathead minnow in aquatic toxicology: past, present and future. *Aquatic Toxicology* 78, (1), 91-102.

Basu, N.; Crump, D.; Head, J.; Hickey, G.; Hogan, N.; Maguire, S.; Xia, J.; Hecker, M. **2019**. EcoToxChip: A next-generation toxicogenomics tool for chemical prioritization and environmental management. *Environmental Toxicology and Chemistry* 38, (2), 279.

Bourdon-Lacombe, J. A.; Moffat, I. D.; Deveau, M.; Husain, M.; Auerbach, S.; Krewski, D.; Thomas, R. S.; Bushel, P. R.; Williams, A.; Yauk, C. L. **2015**. Technical guide for applications of gene expression profiling in human health risk assessment of environmental chemicals. *Regulatory Toxicology and Pharmacology* 72, (2), 292-309.

Burns, F. R.; Cogburn, A. L.; Ankley, G. T.; Villeneuve, D. L.; Waits, E.; Chang, Y. J.; Llaca, V.; Deschamps, S. D.; Jackson, R. E.; Hoke, R. A. **2016**. Sequencing and de novo draft assemblies of a fathead minnow (*Pimephales promelas*) reference genome. *Environmental Toxicology and Chemistry* 35, (1), 212-217.

Consortium, G. O. **2014**. Gene ontology consortium: going forward. *Nucleic Acids Research* 43, (D1), D1049-D1056.

Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; King, B. L.; McMorran, R.; Wiegers, J.; Wiegers, T. C.; Mattingly, C. J. **2017**. The comparative toxicogenomics database: update 2017. *Nucleic acids research* 45, (D1), D972-D978.

Dean, J. L.; Zhao, Q. J.; Lambert, J. C.; Hawkins, B. S.; Thomas, R. S.; Wesselkamper, S. C. **2017**. Application of gene set enrichment analysis for identification of chemically induced, biologically relevant transcriptomic networks and potential utilization in human health risk assessment. *Toxicological Sciences* 157, (1), 85-99.

Edwards, D. **2003**. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics* 19, (7), 825-833.

Farmahin, R.; Williams, A.; Kuo, B.; Chepelev, N. L.; Thomas, R. S.; Barton-Maclaren, T. S.; Curran, I. H.; Nong, A.; Wade, M. G.; Yauk, C. L. **2017**. Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. *Archives of Toxicology* 91, (5), 2045-2065.

Flouriot, G.; Pakdel, F.; Ducouret, B.; Valotaire, Y. **1995**. Influence of xenobiotics on rainbow trout liver estrogen receptor and vitellogenin gene expression. *Journal of Molecular Endocrinology* 15, (2), 143-151.

Fukushima, A. **2013**. DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene* 518, (1), 209-214.

Guillouzo, A.; Guguen-Guillouzo, C. **2008**. Evolving concepts in liver tissue modeling and implications for in vitro toxicology. *Expert Opinion on Drug Metabolism and Toxicology* 4, (10), 1279-1294.

Gust, K. A.; Brasfield, S. M.; Stanley, J. K.; Wilbanks, M. S.; Chappell, P.; Perkins, E. J.; Lotufo, G. R.; Lance, R. F. **2011**. Genomic investigation of year-long and multigenerational exposures of fathead minnow to the munitions compound RDX. *Environmental Toxicology and Chemistry* 30, (8), 1852-1864.

Hansen, P.-D.; Dizer, H.; Hock, B.; Marx, A.; Sherry, J.; McMaster, M.; Blaise, C. **1998**. Vitellogenin - a biomarker for endocrine disruptors. *TrAC Trends in Analytical Chemistry* 17, (7), 448-451.

Hardy, B.; Apic, G.; Carthew, P.; Clark, D.; Cook, D.; Dix, I.; Escher, S.; Hastings, J.; Heard, D. J.; Jeliazkova, N. **2012**. Toxicology ontology perspectives. *Alternatives to Animal Experimentation* 29, (2), 139-156.

Igarashi, Y.; Nakatsu, N.; Yamashita, T.; Ono, A.; Ohno, Y.; Urushidani, T.; Yamada, H. **2014**. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Research* 43, (D1), D921-D927.

Kanehisa, M.; Goto, S. **2000**. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28, (1), 27-30.

Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T. **2007**. KEGG for linking genomes to life and the environment. *Nucleic acids research* 36, (suppl_1), D480-D484.

Kanehisa, M. **2008**. In *The KEGG database*, 'In Silico'Simulation of Biological Processes: Novartis Foundation Symposium 247, 2008; Wiley Online Library: pp 91-103.

Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. **2016**. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 45, (D1), D353-D361.

Kitano, H. **2002**. Systems biology: a brief overview. *Science* 295, (5560), 1662-1664.

LaLone, C. A.; Villeneuve, D. L.; Burgoon, L. D.; Russom, C. L.; Helgen, H. W.; Berninger, J. P.; Tietge, J. E.; Severson, M. N.; Cavallin, J. E.; Ankley, G. T. **2013**. Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk of chemicals with known modes of action. *Aquatic Toxicology* 144, 141-154.

Langfelder, P.; Zhang, B.; Horvath, S. **2007**. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, (5), 719-720.

Langfelder, P.; Horvath, S. **2008**. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, (1), 559.

Larkin, P.; Villeneuve, D. L.; Knoebl, I.; Miracle, A. L.; Carter, B. J.; Liu, L.; Denslow, N. D.; Ankley, G. T. **2007**. Development and validation of a 2,000-gene microarray for the fathead minnow (*Pimephales promelas*). *Environmental Toxicology and Chemistry* 26, (7), 1497-1506.

Liao, Q.; Liu, C.; Yuan, X.; Kang, S.; Miao, R.; Xiao, H.; Zhao, G.; Luo, H.; Bu, D.; Zhao, H. **2011**. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Research* 39, (9), 3864-3878.

Loughery, J. R.; Kidd, K. A.; Mercer, A.; Martyniuk, C. J. **2018.** Part A: Temporal and dose-dependent transcriptional responses in the liver of fathead minnows following short term exposure to the polycyclic aromatic hydrocarbon phenanthrene. *Aquatic Toxicology* 199, 90-102.

Ma, X.; Gao, L. **2012**. Biological network analysis: insights into structure and functions. *Briefings in Functional Genomics* 11, (6), 434-442.

Maertens, A.; Tran, V.; Kleensang, A.; Hartung, T. **2018**. Weighted gene correlation network analysis (WGCNA) reveals novel transcription factors associated with Bisphenol A dose-response. *Frontiers in Genetics* 9, 508.

Margolin, A. A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla Favera, R.; Califano, A. **2006**. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7, (1), S7.

Martinović-Weigelt, D.; Mehinto, A. C.; Ankley, G. T.; Denslow, N. D.; Barber, L. B.; Lee, K. E.; King, R. J.; Schoenfuss, H. L.; Schroeder, A. L.; Villeneuve, D. L. **2014**. Transcriptomic effects-based monitoring for endocrine active chemicals: Assessing relative contribution of treated wastewater to downstream pollution. *Environmental Science and Technology* 48, (4), 2385-2394.

Martyniuk, C. J. **2018**. Are we closer to the vision? A proposed framework for incorporating omics into environmental assessments. *Environmental Toxicology and Pharmacology* 59, 87-93.

Moriya, Y.; Itoh, M.; Okuda, S.; Yoshizawa, A. C.; Kanehisa, M. **2007**. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* 35, W182-W185.

Pagé-Larivière, F.; Crump, D.; O'Brien, J. M. **2019**. Transcriptomic points-of-departure from short-term exposure studies are protective of chronic effects for fish exposed to estrogenic chemicals. *Toxicology and Applied Pharmacology* 114634.

Pavlopoulos, G. A.; Secrier, M.; Moschopoulos, C. N.; Soldatos, T. G.; Kossida, S.; Aerts, J.; Schneider, R.; Bagos, P. G. **2011**. Using graph theory to analyze biological networks. *BioData Mining* 4, (1), 10.

Perkins, E. J.; Chipman, J. K.; Edwards, S.; Habib, T.; Falciani, F.; Taylor, R.; Van Aggelen, G.; Vulpe, C.; Antczak, P.; Loguinov, A. **2011**. Reverse engineering adverse outcome pathways. *Environmental Toxicology and Chemistry* 30, (1), 22-38.

Perkins, E. J.; Ankley, G. T.; Crofton, K. M.; Garcia-Reyero, N.; LaLone, C. A.; Johnson, M. S.; Tietge, J. E.; Villeneuve, D. L. **2013**. Current perspectives on the use of alternative species in human health and ecological hazard assessments. *Environmental Health Perspectives* 121, (9), 1002-1010.

Rehberger, K.; Werner, I.; Hitzfeld, B.; Segner, H.; Baumann, L. **2017**.
20 Years of fish immunotoxicology–what we know and where we are. *Critical Reviews in Toxicology* **2017,** *47*, (6), 516-542.

Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. **2015**. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43, (7), e47-e47.

Rodriguez-Jorquera, I. A.; Kroll, K. J.; Toor, G. S.; Denslow, N. D. **2015**. Transcriptional and physiological response of fathead minnows (Pimephales promelas) exposed to urban waters entering into wildlife protected areas. *Environmental Pollution* 199, 155-165.

Shelley, L. K.; Ross, P. S.; Kennedy, C. J. **2012**. The effects of an in vitro exposure to 17β-estradiol and nonylphenol on rainbow trout (Oncorhynchus mykiss) peripheral blood leukocytes. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* 155, (3), 440-446.

Shelley, L. K.; Ross, P. S.; Kennedy, C. J. **2012**. Immunotoxic and cytotoxic effects of atrazine, permethrin and piperonyl butoxide to rainbow trout following in vitro exposure. *Fish & Shellfish Immunology* 33, (2), 455-458.

Silver, J. D.; Ritchie, M. E.; Smyth, G. K. **2008**. Microarray background correction: maximum likelihood estimation for the normal–exponential convolution. *Biostatistics* 10, (2), 352-363.

Soufan, O.; Ewald, J.; Viau, C.; Crump, D.; Hecker, M.; Basu, N.; Xia, J. **2019**. T1000: a reduced gene set prioritized for toxicogenomic studies. *PeerJ* 7, e7975.

Sutherland, J. J.; Jolly, R. A.; Goldstein, K. M.; Stevens, J. L. **2016**. Assessing concordance of drug-induced transcriptional response in rodent liver and cultured hepatocytes. *PLoS Computational Biology* 12, (3), e1004847.

Sutherland, J.; Webster, Y.; Willy, J.; Searfoss, G.; Goldstein, K.; Irizarry, A.; Hall, D.; Stevens, J. **2018**. Toxicogenomic module associations with pathogenesis: a network-based approach to understanding drug toxicity. *The Pharmacogenomics Journal* 18, (3), 377.

Tcheremenskaia, O.; Benigni, R.; Nikolova, I.; Jeliazkova, N.; Escher, S. E.; Batke, M.; Baier, T.; Poroikov, V.; Lagunin, A.; Rautenberg, M. **2012**. OpenTox predictive toxicology framework: toxicological ontology and semantic media wiki-based OpenToxipedia. *Journal of Biomedical Semantics* 3, (1), S7.

Torrealba, D.; More-Bayona, J. A.; Wakaruk, J.; Barreda, D. R. **2018**. Innate immunity as a bioindicator of health for teleosts exposed to nanoparticles. *Frontiers in Immunology* 9, 3074.

USEPA **2018**. *Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program.*; Washington, DC.

Vachon, J.; Campagna, C.; Rodriguez, M. J.; Sirard, M.-A.; Levallois, P. **2017**. Barriers to the use of toxicogenomics data in human health risk assessment: a survey of Canadian risk assessors. *Regulatory Toxicology and Pharmacology* 85, 119-123.

Van Dam, S.; Võsa, U.; van der Graaf, A.; Franke, L.; de Magalhães, J. P. **2017**. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics* 19, (4), 575-592.

Van der Jagt, K.; Munn, S.; Tørsløv, J.; de Bruijn, J. **2004**. *Alternative approaches can reduce the use of test animals under REACH*; pp 1-25.

Vidal-Dorsch, D. E.; Colli-Dula, R. C.; Bay, S. M.; Greenstein, D. J.; Wiborg, L.; Petschauer, D.; Denslow, N. D. **2013**. Gene expression of fathead minnows (Pimephales promelas) exposed to two types of treated municipal wastewater effluents. *Environmental Science and Technology* 47, (19), 11268-11277.

Wang, X.; Terfve, C.; Rose, J. C.; Markowetz, F. **2011**. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics* 27, (6), 879-880.

Wang, R.-L.; Biales, A. D.; Garcia-Reyero, N.; Perkins, E. J.; Villeneuve, D. L.; Ankley, G. T.; Bencic, D. C. **2016**. Fish connectivity mapping: linking chemical stressors by their mechanisms of action-driven transcriptomic profiles. *BMC Genomics* 17, (1), 84.

Wenger, M.; Sattler, U.; Goldschmidt-Clermont, E.; Segner, H. **2011**. 17Beta-estradiol affects the response of complement components and survival of rainbow trout (Oncorhynchus mykiss) challenged by bacterial infection. *Fish & Shellfish Immunology* 31, (1), 90-97.

Williams, T. D.; Turan, N.; Diab, A. M.; Wu, H.; Mackenzie, C.; Bartie, K. L.; Hrydziuszko, O.; Lyons, B. P.; Stentiford, G. D.; Herbert, J. M. **2011**. Towards a system level understanding of non-model organisms sampled from the environment: a network biology approach. *PLoS Computational Biology* 7, (8), e1002126.

Zaunbrecher, V.; Beryt, E.; Parodi, D.; Telesca, D.; Doherty, J.; Malloy, T.; Allard, P. **2017**. Has toxicity testing moved into the 21st Century? a survey and analysis of perceptions in the field of toxicology. *Environmental Health Perspectives* 125, (8), 087024.

Zhao, W.; Langfelder, P.; Fuller, T.; Dong, J.; Li, A.; Hovarth, S. **2010**. Weighted gene coexpression network analysis: state of the art. *Journal of Biopharmaceutical Statistics* 20, (2), 281-300.

CONNECTING PARAGRAPH

Chapter 4 presented the EcoToxModules, which in part aim to improve communication of complex toxicogenomics results to a broad audience. Transcriptomic dose-response modeling, the subject of chapter 4, can also be thought of as a tool for communicating toxicogenomics results to a non-bioinformatics audience. Dose-response analysis is a cornerstone of chemical risk assessment, and as such, statistics such as benchmark doses (BMD) and points of departure (POD), as well as the theory behind them, are already familiar to both toxicology scientists and regulators. Applying dose-response methods to transcriptomics data translates it into a familiar framework, while maintaining the rich mechanistic detail that toxicogenomics data captures. However, existing software requires users to process and normalize their transcriptomics data before performing curve fitting, can be very slow for large datasets, and must be locally installed. Thus, there is a need for more user-friendly and accessible options.

Chapter 5 is on FastBMD, a web-based tool for rapid transcriptomic dose-response modeling that goes from raw counts tables to results visualization. It was accepted for publication in *Bioinformatics* in February of 2021 (volume 37 (7), pages 1035-1036). The candidate is the sole first author, and additional co-authors are Othman Soufan, Jianguo Xia, and Niladri Basu. While the format of a *Bioinformatics Application Note* requires a short word count, there is considerable additional methodological details and case study results within the supplementary materials (provided in the thesis appendix), and in the FAQ and Tutorial pages of the web-tool (www.fastbmd.ca).

CHAPTER 5. FASTBMD: AN ONLINE TOOL FOR RAPID BENCHMARK DOSE-RESPONSE ANALYSIS OF TRANSCRIPTOMICS DATA

Jessica D. Ewald[1], *Othman Soufan*[1], *Jianguo Xia*[1], *Niladri Basu*[1]

[1] Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Canada

5.1 ABSTRACT

Motivation: Transcriptomics dose-response analysis is a promising new approach method for toxicity testing. While international regulatory agencies have spent substantial effort establishing a standardized statistical approach, existing software that follows this approach is computationally inefficient and must be locally installed.

Results: FastBMD is a web-based tool that implements standardized methods for transcriptomics benchmark dose-response analysis in R. It is >60 times faster than the current leading software, supports transcriptomics data from 13 species, and offers a comprehensive analytical pipeline that goes from processing and normalization of raw gene expression values to interactive exploration of pathway-level benchmark dose results.

Availability: FastBMD is freely available at www.fastbmd.ca

Contact: jeff.xia@mcgill.ca or niladri.basu@mcgill.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

5.2 Introduction

Dose-response analysis (DRA) is a classic method in toxicology used to identify the dose of a chemical that causes a predetermined change in a physiological response; traditionally, this would be an apical outcome such as mortality. This identified dose, called the benchmark dose (BMD), is a key component for developing regulatory standards that aim to protect human and ecological health from adverse effects associated with exposure to chemicals.

Benchmark doses derived from transcriptomics data from short-term exposures are similar to BMDs derived from apical outcomes from long-term exposures (Thomas *et al.*, 2013). Thus, there has recently been a concerted effort by regulatory agencies, including the US National Toxicology Program (NTP) and Health Canada, to develop standardized methods for DRA with transcriptomics data (Farmahin *et al.*, 2017; US NTP, 2018, Table S1). The leading tool, BMDExpress, is a Java-based software that follows these standardized transcriptomics DRA methods (Phillips *et al.*, 2019). However, it is computationally intensive, requires local installation, and is not easily integrated with the wealth of bioinformatics resources available in R.

Here, we present FastBMD, a computationally efficient implementation of the NTP's approach to transcriptomics DRA (US NTP, 2018). FastBMD is a freely available web-based software (www.fastbmd.ca) that supports BMD analysis at the gene, pathway, and transcriptomic levels. FastBMD currently supports microarray and RNA sequencing transcriptomics data from 13 species (Table S2), and also includes an annotation-free pipeline for gene and transcriptomic-level BMD analysis for non-model organisms.

5.3 IMPLEMENTATION

Previous work demonstrated that core non-linear curve fitting algorithms can be efficiently implemented in R (Larras *et al.*, 2018; Ritz *et al.*, 2015). This provided the foundation for FastBMD as an R-based DRA software that follows the NTP approach (US NTP, 2018).

Prior to curve fitting, a matrix of expression values is annotated (Table S2), summarized at the gene level, filtered according to abundance and variance thresholds, and normalized. Genes are filtered according to fold-change and p-value thresholds to eliminate those that are unlikely to have dose-dependent behaviour prior to the computationally intensive curve fitting step. For DRA, model parameters are found for up to 10 statistical models using base R non-linear search functions (Larras *et al.*, 2018; Ritz *et al.*, 2015), and filtered with a lack-of-fit p-value threshold (US NTP, 2018). The best-fit model is selected for each gene. The gene-level BMD (geneBMD) and its upper and lower 95% confidence interval are calculated using the likelihood-profiling method based on the mean and standard deviation of control expression values (US NTP, 2018; Ritz *et al.*, 2015). GeneBMDs are filtered to remove any that occur above the highest measured dose, or that have a wide confidence interval (US NTP, 2018). The transcriptomic-level (omicBMD) is calculated using the distribution of geneBMDs (Pagé-Larivière *et al.*, 2019). Pathway-level BMDs are calculated as the bootstrapped median of geneBMDs within each pathway. Finally, the gene, pathway, and transcriptomic-level results are integrated with interactive plots and tables (Fig. 5-1 A). Detailed explanations of each component of the analysis can be found in the "FAQ" and "Resources" tabs on www.fastbmd.ca.

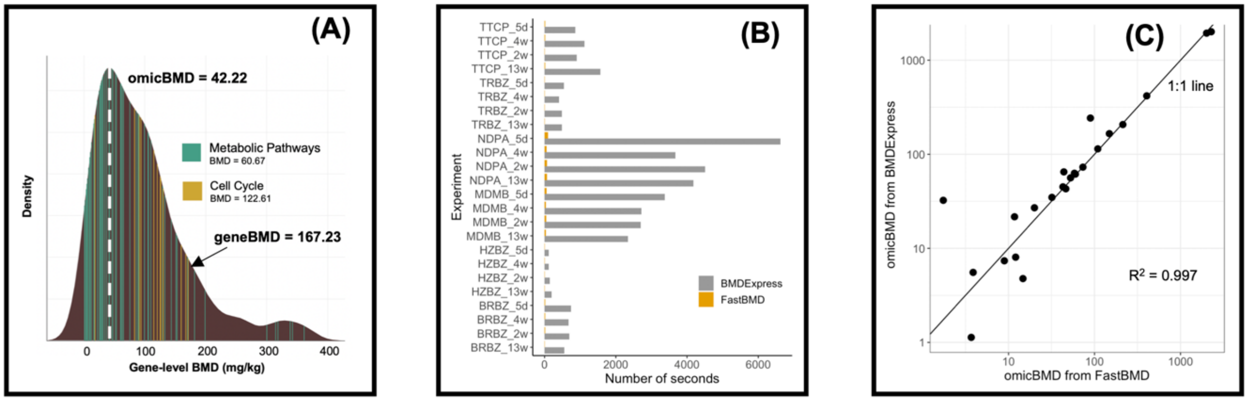


Figure 5-1: FastBMD performance. A) Snapshot of interactive gene, pathway, and transcriptomic BMD visualization in FastBMD for the TRBZ (13 weeks) dataset. B) Comparison of elapsed time during curve-fitting between FastBMD (yellow) and BMDExpress (grey) for 24 datasets. C) OmicBMDs from FastBMD and BMDExpress for 24 datasets.

FastBMD was implemented based on the PrimeFaces (v8.0) component library (http://primefaces.org/) and R (version 3.6.2). The interactive features were developed using Plotly.js (https://plot.ly). The system is hosted on a Google Cloud instance (12 virtual CPUs, 70GB RAM), except for curve fitting which is computed in parallel through a microservice hosted on a dedicated server (24 cores, 128GB RAM) using SpringBoot.

5.4 Materials and Methods

The performance of FastBMD was compared to that of BMDExpress 2 with microarray data measured in adult rats from 24 separate dose-response experiments (Table S3) (Thomas *et al.*, 2013). Data were quantile normalized in R, and then analyzed using both FastBMD and BMDExpress on a Macbook Pro with 4 cores and 8 GB RAM. Genes that did not have a fold change of at least two for any dose group compared to the control were filtered out. All statistical models other than 3º and 4º polynomials were fit to the expression of each remaining gene. More methodological details are given in the supplementary materials.

5.5 Results and Discussion

FastBMD was over 60 times faster than BMDExpress, taking 0.18 hours to compute geneBMDs for all 24 experiments compared to 11.03 hours for BMDExpress (Fig. 5-1B). The omicBMD^{mode} from each experiment were extremely similar with an even distribution around the 1:1 line and an R^2 of 0.997 (Fig. 5-1C). Sources of variation in the results between the two software are discussed in the supplementary materials. The increased efficiency can be mainly explained by the curve fitting algorithm. In FastBMD, parameters are roughly estimated using the input data, and then fed into a modern non-linear parameter search algorithm that quickly either converges or fails. In contrast, BMDExpress searches the parameter space until either a solution is found, or the user-specified timeout period is surpassed.

FastBMD is designed to be a flexible tool that can accommodate diverse transcriptomics data. It addresses reproducibility by allowing users to download results after each analytical step and generate summary reports. Since it is implemented in R, FastBMD is uniquely positioned to leverage existing statistical packages to rapidly implement future updates as the regulatory and scientific communities continue to refine the recommended approach. In the future, we plan to incorporate knowledgebases of environmental chemical concentrations, expanded gene-set libraries, and baseline gene expression levels in common target tissues to provide additional context for the BMD results.

5.6 Acknowledgements

This work was supported by Canada's Natural Sciences and Engineering Research Council (Discovery Grants to JX and NB; Canada Graduate Scholarship to JE), and the Canada Research Chairs Program (to JX and NB). We also thank Jason O'Brien, Florence Pagé-Larivière, Doug Crump, and other members of the EcoToxChip project (funded by Genome Canada) for their advice.

Conflict of Interest: none declared.

5.7 References

Farmahin, R., Williams, A., Kuo, B., Chepelev, N. L., Thomas, R. S., Barton-Maclaren, T. S., *et al*. **2017**. Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. *Archives of Toxicology*, 91(5), 2045-2065.

Larras, F., Billoir, E., Baillard, V., Siberchicot, A., Scholz, S., Wubet, T., *et al*. **2018**. DRomics: a Turnkey Tool to support the use of the dose–response framework for omics data in ecological risk assessment. *Environmental Science and Technology*, 52(24), 14461-14468.

Pagé-Larivière, F., Crump, D., O'Brien, J.M. **2019**. Transcriptomic points-of-departure from short-term exposure studies are protective of chronic effects for fish exposed to estrogenic chemicals. *Toxicology and Applied Pharmacology* 378:114634.

Phillips, J. R., Svoboda, D. L., Tandon, A., Patel, S., Sedykh, A., Mav, D., *et al*. **2019**. BMDExpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10), 1780-1782.

US NTP. **2018**. NTP Research Report on National Toxicology Program Approach to Genomic Dose-Response Modeling: Research Report 5.

Ritz, C., Baty, F., Streibig, J. C., Gerhard, D. **2015**. Dose-response analysis using R. *PloS One*, 10(12), e0146021.

Thomas, R. S., Wesselkamper, S. C., Wang, N. C. Y., Zhao, Q. J., Petersen, D. D., Lambert, J. C., *et al*. **2013**. Temporal concordance between apical and transcriptional points of departure for chemical risk assessment. *Toxicological Sciences*, 134(1), 180-194.

CONNECTING PARAGRAPH

Chapter 6 is on EcoToxXplorer, a web-based tool for analyzing transcriptomics data within the context of environmental toxicology. At the time of publication, EcoToxXplorer was split into three modules: 1: EcoToxChip Analysis, 2: Raw RNA-seq Processing, and 3: RNA-seq Analysis. Since publication of the Chapter 6 manuscript, the raw RNA-seq processing module has been expanded into EcoOmicsAnalyst and the RNA-seq analysis module has been incorporated into ExpressAnalyst, a more general platform for RNA-seq analysis. The EcoToxChip analysis module, at the time of publication and now, forms the core of EcoToxXplorer.

The selection of genes to include on the EcoToxChips was significantly informed by the functional EcoToxModules (Chapter 4), and the EcoToxModules are incorporated into visualizations and gene set analysis throughout the EcoToxChip analysis pipeline. In the spirit of design-thinking, significant updates were made to the EcoToxModules as we learned more about how they were used as they were incorporated into EcoToxXplorer. Since the publication of the Chapter 6 manuscript, a dose-response pipeline was added to the EcoToxChip module in EcoToxXplorer that uses the core algorithm from Chapter 5.

This chapter was published as a focus article in the journal *Environmental Toxicology and Chemistry* in November of 2021. The paper has two co-first authors: the candidate, and Othman Soufan, formerly a post-doc at McGill University and now an assistant professor at St. Francis Xavier University in Nova Scotia, Canada. It is additionally co-authored by Guangyan Zhou, Orcun Hacariz, Emily Boulanger, Alper James Alcaraz, Gordon Hickey, Steve Maguire, Guillaume Pain, Natacha Hogan, Markus Hecker, Doug Crump, Jessica Head, Niladri Basu (the candidate’s supervisor), and Jianguo Xia. Supplemental information is provided in the thesis appendix.

CHAPTER 6. ECOTOXXPLORER: LEVERAGING DESIGN THINKING TO DEVELOP A STANDARDIZED WEB-BASED TRANSCRIPTOMICS ANALYTICS PLATFORM FOR DIVERSE USERS

Othman Soufan[1,2,], Jessica Ewald[1,*], Guangyan Zhou[1], Orcun Hacariz[1], Emily Boulanger[1], Alper James Alcaraz[3], Gordon Hickey[1], Steve Maguire[4,5], Guillaume Pain[6], Natacha Hogan[3], Markus Hecker[3], Doug Crump[7], Jessica Head[1], Niladri Basu[1,*], Jianguo Xia[1,*]*

[1] Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Canada

[2] Department of Computer Science, Saint Francis Xavier University, Antigonish, Nova Scotia, Canada

[3] School of the Environment & Sustainability and Toxicology Centre, University of Saskatchewan, Saskatoon, Canada

[4] The University of Sydney Business School & Sydney Nano Institute, University of Sydney, Sydney, Australia

[5] Department of Chemistry, Faculty of Science, McGill University, Montreal, Quebec, Canada

[6] Faculty of Administrative Sciences, Université Laval, Quebec City, Quebec, Canada

[7] Ecotoxicology and Wildlife Health Division, Environment and Climate Change Canada, National Wildlife Research Centre, Ottawa, Canada

* These authors contributed equally to this work.

6.1 TRANSCRIPTOMICS DATA ARE TRANSFORMATIONAL YET CHALLENGING

The generation and use of transcriptomics data across the life sciences have risen sharply in recent years driven largely by advances in biotechnology and computational biology. Within the field of environmental toxicology, the data being generated from these efforts are providing important insights into stressor-induced perturbations at the molecular-level and helping increase understanding of causal linkages to connect such molecular perturbations with adverse outcomes at the whole organism level (Villeneuve *et al*., 2014). Despite impressive advances in these areas, the scope and pace of adoption of transcriptomics approaches in the practices of chemical risk assessment and environmental management have generally not met the expectations of their proponents (Mondou *et al*. 2021; Pain *et al*. 2021).

A major challenge with transcriptomics data is that they can be complex and difficult for users to distill and synthesize into clear and actionable insights. Transcriptomics technologies can generate a tremendous amount of data, and accordingly the handling and analysis of these data require powerful computers and comprehensive databases along with bioinformatics and programming know-how. Even studies of a few dozen genes can prove difficult for many users in terms of data management, analysis and interpretation. These challenges are compounded for ecological species, which have far fewer and less developed knowledgebases and user-friendly software tools compared to common model organisms. Further, tools that do exist are generally designed for ‘omics specialists rather than novice users.

Even though regulatory science communities are exhibiting increased awareness and acceptance of transcriptomics data (Mondou *et al*., 2020), activities in this field are mostly performed by research institutions and individuals who have deep ‘omics expertise and capabilities but tend to focus on fundamental biological questions (Health Canada 2019; Pain *et al*., 2021). For example, results from an online survey of 29 Canadian risk assessors revealed that the application of toxicogenomics data is marginal; 85% of respondents indicated that they never or rarely used such data (Vachon *et al*., 2017). Thus, while toxicogenomics data are now plentiful, their analysis and interpretation remain concentrated in the hands of a small group of domain experts. To help realize the potential for toxicogenomics data to transform chemical risk assessment, new tools are urgently needed to help novice users translate transcriptomics data efficiently and effectively into meaningful information and actionable knowledge.

6.2 DEMOCRATIZING ‘OMICS DATA VIA WEB-BASED VISUAL ANALYTICS

Over the years, authors of this paper have developed and published a series of web-based tools for metabolomics, transcriptomics, microbiomics, and multi-omics integration to help users translate complex ‘omics data into biological insights and actionable knowledge (Table 6-1). These tools have been gaining popularity among diverse user communities since their initial releases. As an example, over the last 12 months, the MetaboAnalyst platform has processed >3.8 million data analysis jobs submitted from >120,000 users worldwide. We have purposefully developed these tools to help users, including novice ones, overcome difficulties associated with the sheer size and complexity of big data as elaborated upon below.

In terms of the sheer size of big data, raw datafiles are typically gigabytes in size and usually require the use of a high-performance computing center and command-line programming scripts, and thus are generally not accessible to the novice user. Such barriers are now being overcome by innovations in computing technologies. Specifically, cloud computing services offer access to scalable high-performance computing power without the need for on-site hardware resources. When these cloud-based services are coupled with powerful, modern web browsers, efficient and easy-to-use software platforms can be developed. Such innovations have helped spur the development of computational platforms (e.g., Galaxy, KNIME, along with our tools listed in Table 6-1), that in recent years have started to make ‘omics data analysis more accessible, reproducible, and standardized for users.

Table 6-1: A list of web-based tools for ‘omics data analytics that team members have developed. The user statistics are based on Google Analytics (retrieved in May 2021). The metabolomics and microbiome data analysis tools are species agnostic and applicable to both model and non-model species. The transcriptomics data analysis tools are species dependent, though NetworkAnalyst, FastBMD, and Seq2Fun can support generic species based on KEGG orthologs as well as custom annotations. OmicsAnalyst can integrate different ‘omics data in a species independent manner.

Tool	Purpose	First released	Current version	Users/ year	Jobs / year
MetaboAnalyst.ca	Analyze metabolomics data	2009	5.0	135,000	3,800,000
NetworkAnalyst.ca	Analyze transcriptomics data	2014	3.0	43,500	540,000
MicrobiomeAnalyst.ca	Analyze microbiomics data	2017	1.0	24,000	100,000
FastBMD.ca	Calculate transcriptomics benchmark doses	2020	1.0	insufficient data	
Seq2Fun.ca	Functionally quantify RNAseq data in non-model organisms	2020	1.0		
EcoToxXplorer.ca (current study focus)	Analyze transcriptomics (EcoToxChip) data in an ecotoxicological context	2021	1.0		
OmicsAnalyst.ca	Integrate multi-omics data	2021	1.0		

In terms of the complexity of big data, advanced data management skills, statistical analysis tools, and visualization methods, along with a keen sense of the scientific context, need to be brought together to deal with the challenge of extracting biological meaning from datasets that tend to be large and complicated. Concerning data management and statistical analysis, most environmental researchers tend to rely on simple spreadsheets and these do not work well (nor scale well) for big data. Many environmental researchers are also used to visualizing each measured variable individually to decide which statistical methods to apply in each case. This approach is not feasible for ‘omics datasets, where instead the general approach is to uniformly apply relatively simple and well-established statistical models, and then to visually assess the results afterwards to ensure that nothing has gone drastically wrong. Many statistical methods commonly used for “small data” do not work well for big data due to the high dimensionality and heterogeneity characteristics of most ‘omics datasets, and so it is important to use well-established bioinformatics methods that have been proven robust under these conditions.

The scientific community now accepts that there has been an over-reliance on p-values obtained from routine statistical analyses, and that they should only be treated as one input amongst others into an overall evidence base. Rather than base conclusions primarily on a strict cut-off (i.e. p-value < 0.05), researchers are advised to compare and contrast the results from different analyses, while also evaluating scientific practices and contextual evidence (e.g., study design, biological plausibility) to obtain a comprehensive and deeper understanding of their data. Visual analytics has emerged as a promising solution to help extract even more biological meaning from any given dataset by combining human (i.e., pattern discovery through visualization) and computational (i.e., pattern validation through analytics) capabilities. In doing so, visualization tools such as interactive statistics and graphical plots that link to functional knowledgebases help make complex data more accessible and usable, while also engaging and empowering users to explore data in a user-friendly and iterative manner. This exemplifies the open data concept in data science and contrasts with the perception of many bioinformatics pipelines as closed (black-box approach) to end users.

One challenge associated with the open data concept concerns reproducibility where different algorithms, parameters, and their combinations may generate vastly different results. A lack of standardization has been flagged as a key obstacle for the regulatory uptake of toxicogenomics data (Pain *et al*., 2020), thus necessitating the need for data analysis tools that are built on rigid and predefined workflows. While it is necessary and technically feasible to build out such data analysis tools, there is also a need to achieve some balance between flexibility and standardization given that scientific discoveries sometimes arise from open ended exploration of large datasets. This balanced approach underpins the wide acceptance of the web-based tools developed by some of the authors (Table 6-1), and several features are included in these tools to ensure reproducibility. For example, each analysis session automatically records the procedures and parameters used, and these are captured in a detailed PDF report that documents each step (for general users) along with the underlying R command history (for those familiar with programming). In addition, the FAQ and tutorials pages provide guidance on the most appropriate analysis parameters.

6.3 OBJECTIVE

Actors in the fields of chemical risk assessment and environmental management express tremendous interest in harnessing the power of transcriptomics data, yet the lack of standardized and validated bioinformatics tools to help organize, analyze, visualize, and interpret the data remain key barriers in doing so (Health Canada 2019; Pain *et al*., 2021). Motivated by our team's experiences outlined above, the objective of this project was to design EcoToxXplorer as a next-generation bioinformatics tool that is high performance (i.e., scalable for large data or user traffic), intuitive to use (i.e., enables complex analytics via a simple interface), and universally accessible (i.e., web-based and user-friendly design) to handle transcriptomics data for the purpose of chemical risk assessment and environmental management.

Transcriptomics data in ecotoxicology is highly variable and can range from information on a few well-curated genes to the entire transcriptome, as well as from species with no (or limited) genomic understanding to those with fully characterized genomes. EcoToxXplorer was initially designed to handle data from these diverse situations though in recent updates we have focused the tool to handle data specifically from EcoToxChips. Foremost behind this decision is that the development of EcoToxXplorer falls under the EcoToxChip project which aims to develop, test, validate, and commercialize quantitative PCR arrays (EcoToxChips) and a data evaluation tool (EcoToxXplorer) for the characterization, prioritization, and management of environmental chemicals and complex mixtures of regulatory concern (Basu *et al*. 2019). Briefly, EcoToxChips consist of ~370 evidence-based gene targets carefully selected from a combination of input from domain experts (e.g., review of ecotoxicogenomics literature including AOPs), bioinformatics analyses of relevant datasets (e.g., *de novo* experiments performed through EcoToxChip project), and consultation with regulatory stakeholders to include genes they are familiar with and trust. EcoToxChips are manufactured through a commercial partnership with Qiagen (Frederick, Maryland, USA) in an ISO-compliant facility and available through a global distribution network. In terms of EcoToxXplorer's ability to handle other transcriptomics needs, for processing raw RNAseq data, EcoToxXplorer's home page has a 'Raw RNAseq Processing' module button that is detailed in Supporting Information 1. For the analysis of the entire transcriptome, the home page has a 'RNAseq Expression Profiling' module button that will route users to NetworkAnalyst, where we have expanded the number of annotation resources and analysis pipelines for ecological species. For the study of organisms without a reference transcriptome (or one that is poorly annotated), we recently developed Seq2Fun available through the EcoOmicsAnalyst portal (www.ecoomicsanalyst.ca).

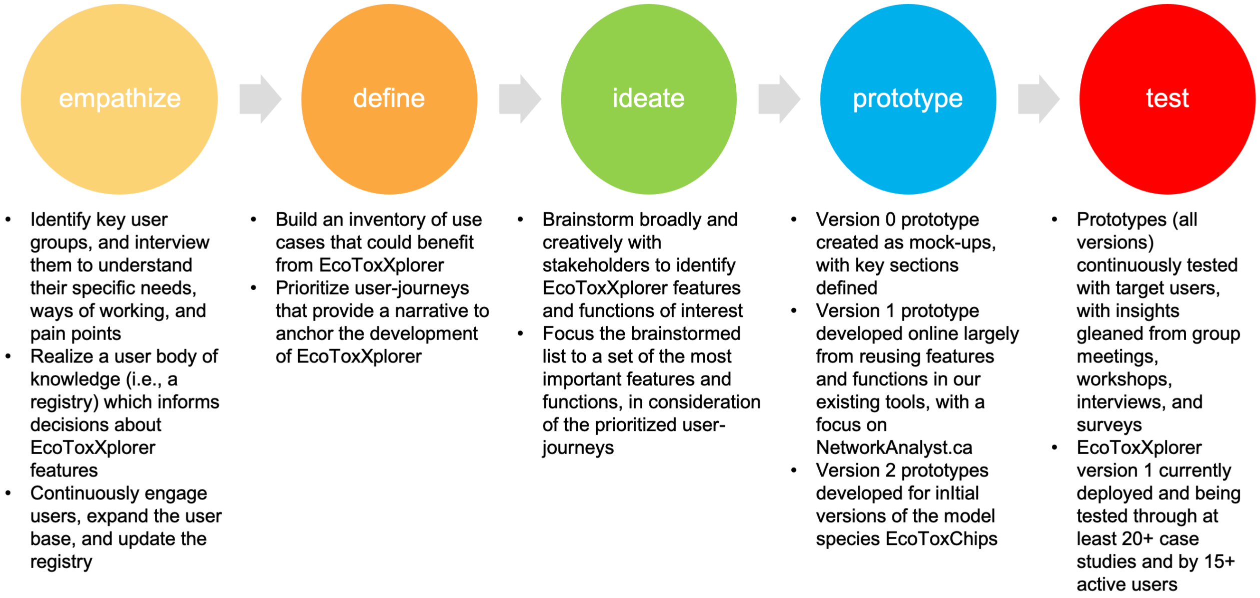
6.4 DESIGN THINKING

To ensure that end-user input and feedback were integrated systematically into every aspect of EcoToxXplorer (from overarching design, to key features, to user experience), the team leveraged a design thinking approach. Grown from roots in the fields of architecture and engineering in the 1950's and 60's, design thinking is an approach to problem solving that employs tools and methods typically used by designers. Designers strive to solve the problems of existing and prospective users, placing users' needs and inputs at the core of design thinking and its five components or "modes" – empathize, define, ideate, prototype, and test – as represented in Figure 1:

- "Empathizing" entails learning about users' values, behaviors, and experiences;
- "Defining" translates empathizing into an actionable problem statement;
- "Ideating" generates and explores design alternatives;
- "Prototyping" turns ideas into interactive models; and,
- "Testing" helps to gather feedback, refine models, and continue to learn about users.

By fostering collaborative, iterative, and meaningful engagements with the user community, design thinking can contribute to lowering knowledge barriers faced by nonexpert users, a necessary condition for the adoption of complex innovations such as the use of transcriptomics data for regulatory decision-making (Pain *et al*., 2021).

Accordingly, the development of EcoToxXplorer followed a design thinking process, which was in fact, built into all activities of the EcoToxChip project (Figure 6-1). Notably, over 100 project partners representing 67 distinct institutions informed the project with experience and knowledge on a range of topics pertinent to toxicity testing (e.g., animal species, new approach methods, 'omics measures), user communities (e.g., government scientists and regulators, industry, contract research organizations, academics, non-governmental organizations), and decision-making goals (e.g., chemical read across, screening and prioritization, environmental monitoring, basic research, etc.). These, and other, diverse perspectives have helped shape the design of EcoToxXplorer.



- Identify key user groups, and interview them to understand their specific needs, ways of working, and pain points
- Realize a user body of knowledge (i.e., a registry) which informs decisions about EcoToxXplorer features
- Continuously engage users, expand the user base, and update the registry

- Build an inventory of use cases that could benefit from EcoToxXplorer
- Prioritize user-journeys that provide a narrative to anchor the development of EcoToxXplorer

- Brainstorm broadly and creatively with stakeholders to identify EcoToxXplorer features and functions of interest
- Focus the brainstormed list to a set of the most important features and functions, in consideration of the prioritized user-journeys

- Version 0 prototype created as mock-ups, with key sections defined
- Version 1 prototype developed online largely from reusing features and functions in our existing tools, with a focus on NetworkAnalyst.ca
- Version 2 prototypes developed for inItial versions of the model species EcoToxChips

- Prototypes (all versions) continuously tested with target users, with insights gleaned from group meetings, workshops, interviews, and surveys
- EcoToxXplorer version 1 currently deployed and being tested through at least 20+ case studies and by 15+ active users

Figure 6-1: Design thinking components and activities therein that underpinned the development and evaluation of EcoToxXplorer.

Activities to map out both EcoToxXplorer's macro- and micro-level features were needed (Supporting Information 2). At the macro-level, the 'recipient system' within which EcoToxXplorer was to operate needed to be understood to help ensure that the tool could be integrated into existing routines and decision-making processes of users. Therefore, mapping out this system warranted an understanding of diverse aspects, ranging from the design of standard chemical exposure studies and the types of genomic data that EcoToxXplorer would handle, to how it would interface with transcriptomics data from EcoToxChips and other sources, to the process by which users would engage with EcoToxXplorer including their work routines and associated 'pain points'. At the micro-level, EcoToxXplorer needed to be designed by identifying, developing, testing, and optimizing key features of interest to users and their particular needs. Key design thinking activities we undertook are outlined in Figure 6-1, and identified macro and micro-level design features are summarized in Table 6-2.

Table 6-2: Key features identified through design thinking activities. These design features are described in more detail in the following sections.

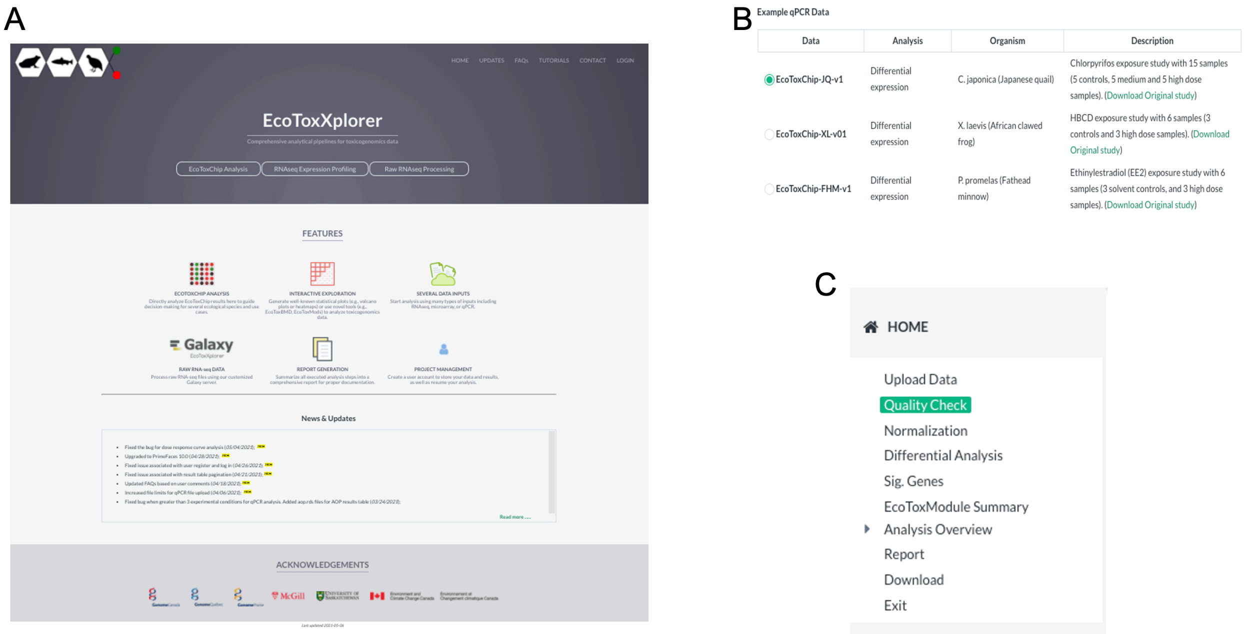
Macro-level features	Micro-level features
Give analysis pipeline a toxicology context	• Curate genes included on the EcoToxChips based on input from domain experts and regulatory stakeholders • Integrate toxicology-specific knowledgebases (EcoToxModules; AOPWiki) • Provide option for dose-response analysis
Help users navigate the complex results	• Edit gene symbols and names to be more familiar • Offer interactive visual analytics • Connect the big picture (EcoToxModules) to the granular details (statistics/plots for individual genes) • Generate comprehensive reports
Standardize analysis across species	• Lock down analysis pipeline regardless of species • Provide comparable biological coverage across species (i.e., common core genes) • Harmonize knowledgebases across species
Ensure novice users can comfortably use the tool	• Identify and flag low-quality data • Provide established statistical methods • Choose appropriate parameter defaults • Ensure design pipeline is "error proof" by not allowing incorrect combinations of parameters • Offer clear documentation, tutorials, and case studies

6.5 OVERVIEW OF ECOTOXXPLORER

EcoToxXplorer version 1 (available at www.EcoToxXplorer.ca) was built between 2017 and 2020 and has been modelled after successful cloud-based tools developed by team members (Table 6-1), most notably NetworkAnalyst.ca, which currently processes over 500,000 transcriptomic jobs per year. EcoToxXplorer has been developed with careful attention to employing best practices and standards to help realize a user-friendly tool (Helmy *et al*., 2016), specifically designed for the analysis of ecotoxicogenomics data. We note that side-by-side comparisons of results derived from EcoToxXplorer with alternate data analysis methods (e.g., programming in R, spreadsheets in Microsoft Excel) were regularly conducted by different project team members on diverse datasets to ensure that there was 100% concordance between workflows.

EcoToxXplorer was implemented based on the PrimeFaces (v10.0) component library and R (version 4.0.2). Visual analytic tools were developed using several JavaScript visualization components, including CanvasXpress, E-Charts, and Plotly.js. Interactive (2D) network visualization was realized through Sigma.js. The system is hosted on a Google Cloud n1-highmem-8 instance (64 GB RAM and eight virtual CPUs with 2.6 GHz each).

The web interface has evolved since its inception (Supporting Information 3), and the current version has links to pages that provide detailed FAQs, tutorials, updates, contact information, and project management login information (Figure 6-2A). Through the home page, users can access three main modules: 1) EcoToxChip Analysis; 2) RNAseq Expression Profiling; and 3) Raw RNAseq Processing. Once a specific module has been selected, users are then directed to that module's upload page within which there are example datasets to help users navigate the tools (Figure 6-2B). The 'RNAseq Expression Profiling' module routes users to NetworkAnalyst, where we have integrated annotation resources and analysis pipelines for ecological species that were curated as part of the EcoToxChip project. The 'Raw RNAseq Processing' module button routes users to https://galaxy.ecotoxxplorer.ca as detailed in Supporting Information 1. Within each module, navigation panels highlight the user's current execution step as well as the remaining ones (Figure 6-2C). Each module also includes options to generate and download figures, input and export data files, and track the history in R. There is no cost in using EcoToxXplorer, and users can register to set up a project management account. Registered users may store up to 10 projects, which may be used for up to one year to keep track of selected parameters, re-perform analyses, and save time uploading files.



ECOTOXCHIP ANALYSIS
Directly analyze EcoToxChip results here to guide decision-making for several ecological species and use cases.

INTERACTIVE EXPLORATION
Generate well-known statistical plots (e.g., volcano plots or heatmaps) or use novel tools (e.g., EcoToxBMD, EcoToxMods) to analyze toxicogenomics data.

SEVERAL DATA INPUTS
Start analysis using many types of inputs including RNAseq, microarray, or qPCR.

RAW RNA-seq DATA
Process raw RNA-seq files using our customized Galaxy server.

REPORT GENERATION
Summarize all executed analysis steps into a comprehensive report for proper documentation.

PROJECT MANAGEMENT
Create a user account to store your data and results, as well as resume your analysis.

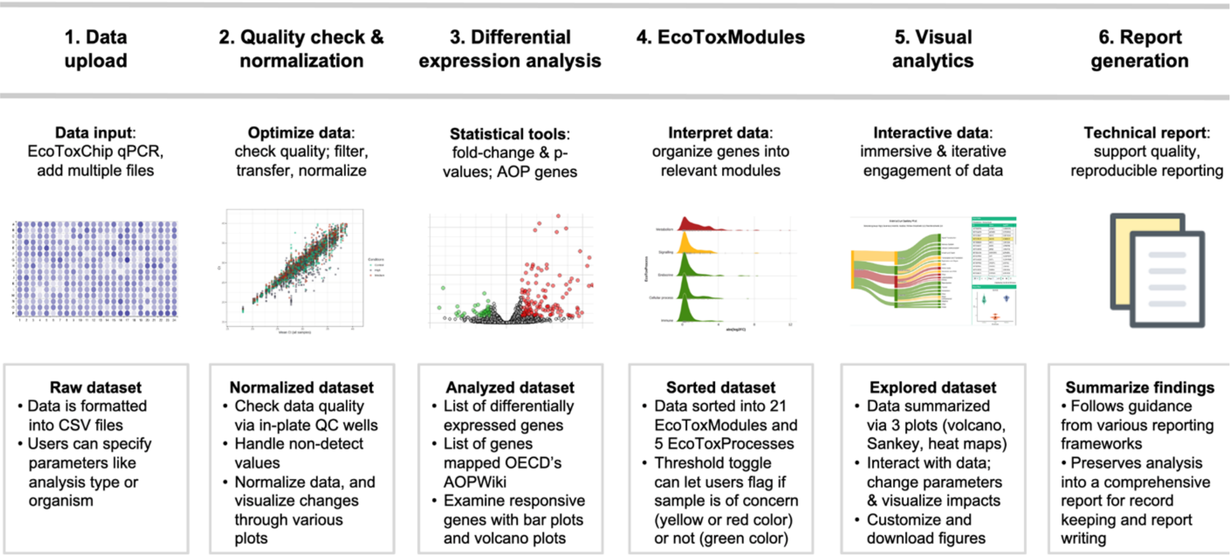
- Fixed the bug for dose response curve analysis (*05/04/2021*); NEW
- Upgraded to PrimeFaces 10.0 (*04/28/2021*); NEW
- Fixed issue associated with user register and log in (*04/26/2021*); NEW
- Fixed issue associated with result table pagination (*04/21/2021*); NEW
- Updated FAQs based on user comments (*04/18/2021*); NEW
- Increased file limits for qPCR file upload (*04/06/2021*); NEW
- Fixed bug when greater than 3 experimental conditions for qPCR analysis. Added aop.rds files for AOP results table (*03/24/2021*);

Example qPCR Data

Data	Analysis	Organism	Description
● EcoToxChip-JQ-v1	Differential expression	C. japonica (Japanese quail)	Chlorpyrifos exposure study with 15 samples (5 controls, 5 medium and 5 high dose samples). (Download Original study)
○ EcoToxChip-XL-v01	Differential expression	X. laevis (African clawed frog)	HBCD exposure study with 6 samples (3 controls and 3 high dose samples). (Download Original study)
○ EcoToxChip-FHM-v1	Differential expression	P. promelas (Fathead minnow)	Ethinylestradiol (EE2) exposure study with 6 samples (3 solvent controls, and 3 high dose samples). (Download Original study)

Figure 6-2: Screenshots of EcoToxXplorer's interface design. A) Home page with links to FAQs, tutorials, user accounts and other resources, along with entry points into the three main modules: 1) EcoToxChip Analysis; 2) RNAseq Expression Profiling; and 3) Raw RNAseq Processing. B) For the EcoToxChip Analysis module, three example datasets are provided to help the user learn how to navigate EcoToxXplorer. C) A navigation panel is provided to inform the user of their current processing step.

Note, in this paper we focus mainly on the 'EcoToxChip Analysis module' because developing the EcoToxChip data analysis pipeline was the primary motivation for developing EcoToxXplorer (Basu *et al*., 2019). The 'Raw RNAseq Processing' module is detailed in Supporting Information 1, and the 'RNAseq Expression Profiling' module is detailed in the NetworkAnalyst tool.



1. Data upload	2. Quality check & normalization	3. Differential expression analysis	4. EcoToxModules	5. Visual analytics	6. Report generation
Data input: EcoToxChip qPCR, add multiple files	**Optimize data**: check quality; filter, transfer, normalize	**Statistical tools**: fold-change & p-values; AOP genes	**Interpret data**: organize genes into relevant modules	**Interactive data**: immersive & iterative engagement of data	**Technical report**: support quality, reproducible reporting
Raw dataset • Data is formatted into CSV files • Users can specify parameters like analysis type or organism	**Normalized dataset** • Check data quality via in-plate QC wells • Handle non-detect values • Normalize data, and visualize changes through various plots	**Analyzed dataset** • List of differentially expressed genes • List of genes mapped OECD's AOPWiki • Examine responsive genes with bar plots and volcano plots	**Sorted dataset** • Data sorted into 21 EcoToxModules and 5 EcoToxProcesses • Threshold toggle can let users flag if sample is of concern (yellow or red color) or not (green color)	**Explored dataset** • Data summarized via 3 plots (volcano, Sankey, heat maps) • Interact with data; change parameters & visualize impacts • Customize and download figures	**Summarize findings** • Follows guidance from various reporting frameworks • Preserves analysis into a comprehensive report for record keeping and report writing

Figure 6-3: Data processing workflow of EcoToxXplorer. All major steps are listed with key aspects summarized.

6.6 Navigating EcoToxChip Analysis

The general workflow for using EcoToxXplorer to analyze EcoToxChip data is shown in Figure 6-3 and includes six key steps: 1) data upload, 2) quality check and normalization, 3) differential expression analysis, 4) EcoToxModules, 5) visual analytics, and 6) report generation.

Step 1: Data upload

EcoToxXplorer was designed to handle gene expression data arising from EcoToxChips that were run in a wide range of study conditions from investigation of single samples to concentration-response analysis of multiple treatments. The raw data generated from EcoToxChips consists of a table of gene names and corresponding Ct values (cycle threshold; point at which the fluorescent signal crosses a threshold in the qPCR instrument). Users can upload files with a maximum size of 50 MB (though a typical file is less than 10 KB), and data can be uploaded as a group of CSV files.

EcoToxChips are currently designed to handle 384- and 96-well microplate formats for six different species, including three laboratory model species that represent the most important vertebrate groups in ecological risk assessment (fish: fathead minnow; bird: Japanese quail; amphibian: *Xenopus laevis*), as well as three native species that are representative of Canadian ecosystems and actively monitored by current Environment and Climate Change Canada programs (fish: rainbow trout; bird: double-crested cormorant; amphibian: Northern leopard frog). Species and version-specific data annotation are built into EcoToxXplorer and linked to the data at the upload step and are necessary to enable downstream functional analyses (Supporting Information 4).

Step 2: Quality check and normalization

EcoToxXplorer allows users to assess the distribution of EcoToxChip data through an interactive diagram that provides an overview of Ct values averaged across all plates as well as values from individual plates (see '1. Data Upload' in Figure 6-3). In addition, EcoToxChips contain in-plate proprietary controls developed by Qiagen, which can be interrogated through features built into EcoToxXplorer. There is a single Genomic DNA Contamination (GDC) control well that contains primers that only amplify genomic DNA and not cDNA. Amplification with a Ct value <35 in considered to indicate that genomic DNA is present in the sample. There are also three Reverse Transcription Control (RTC) wells that detect an artificial external RNA control sequence spiked into the cDNA synthesis reaction. The RTC well can be used to assess the efficiency of the reverse transcription reaction when analyzed in conjunction with the three Positive PCR Control (PPC) wells. The latter contain pre-dispensed external DNA templates of known copy numbers, which produce a defined cycle threshold (Ct) value under ideal PCR conditions. A difference in Ct values between the average of the RTC wells and the average of the PPC wells of ≤ 5, signifies that the reverse transcription reaction was efficient. In addition to these controls, each EcoToxChip plate has five reference (housekeeping) genes that are used to normalize the data. The Ct values of each of these genes are compared across groups using ANOVAs. If there are statistically significant differences between housekeeping genes across treatment groups, or if the RTC or PPC controls exceed certain thresholds, then EcoToxXplorer will return an "inquiry" comment. This indicates that the user should look more closely at the raw data to determine the source of the issue and take further action. For each of these quality check factors, criteria for success are defined along with visual diagnostic plots to help support decision-making. Users are given the option of removing individual samples through the "Data Editor" if they fail multiple QA/QC criteria.

Once the overall quality of the data has been verified, users can next focus on outlier detection and normalization. In terms of outliers, options are provided to filter particular wells, to set a Ct cut-off, and to impute non-detect values. These can be tested iteratively to evaluate their impact on the final dataset. In terms of normalization, EcoToxXplorer provides two widely-used options; Delta Ct normalization and quantile normalization. In the 'Data Normalization' page, diagnostic plots (i.e., scatterplot, box plot, PCA plot, and density plot) are available to visualize the effects of data normalization. This empowers users to deeply investigate their data, make decisions on how to best deal with genes that do not amplify or are not detected, and which normalization steps should be taken. As an example, we illustrate pre-processing effects on density and PCA plots before and after normalization (Supporting Information 5).

Step 3: Differential expression analysis

Users can generate differentially expressed gene (DEG) lists, by selecting a particular statistical approach (parametric or non-parametric) and significance thresholds of interest (i.e., p-value and fold change). The initial result will be a simple tally of the number of significant genes that are up- and down-regulated. This is followed by a detailed table of computed statistics (i.e., fold changes and p-values) for each gene, along with an option to view the gene expression data as bar charts. When available, each Gene ID is hyperlinked to its corresponding page on NCBI from which a user can obtain more information. The adjusted p-value is calculated using the False Discovery Approach to account for multiple testing. To facilitate navigation and visual assessment of the result table, genes that are significantly up-regulated are highlighted in light red, and those that are significantly down-regulated are highlighted in light green. Given that the Adverse Outcome Pathway (AOP) concept is of interest to our community (Villeneuve *et al*., 2014), many genes selected for inclusion on EcoToxChips are also ones that can be mapped to AOP(s) and their underlying Key Events (KE). To draw attention to these specific genes, we developed a separate tab which allows users to view genes according to a given AOP or KE. The AOP and KE names hyperlink directly to the corresponding entry in the AOPWiki (aopwiki.org) led by the OECD's Extended Advisory Group for Molecular Screening and Toxicogenomics (EAGMST) group. Finally, a static volcano plot is provided that summarizes all the gene expression values while highlighting those that are statistically significant and meet a certain fold change cut-off.

Step 4: Functional analysis based on EcoToxModules

The ecological species in EcoToxXplorer include some with official genomes that have been annotated with gene names, gene ontology terms, and KEGG pathways (e.g. Japanese quail) and others with no existing resources (e.g. Double-crested cormorant). We intentionally designed the analytical pipeline to accommodate this range of pre-existing resources by focusing the functional interpretation on our previously developed EcoToxModules (Ewald *et al*., 2020), custom toxicology-focused gene sets that can be defined for transcript sequences from any species. Specifically, these functional EcoToxModules were identified by re-organizing 172 KEGG pathways into modules that are more relevant to ecotoxicology based on input from expert scientists and regulators. In doing so, we identified 21 biological modules (which we term EcoToxModules), which are further organized into five larger physiological process (signaling, metabolism, immune, endocrine, cellular processes; which we term EcoToxProcesses). Since there are only 377 genes on the EcoToxChips, most individual KEGG pathways have none or only a handful of representative genes, making them unsuitable for overrepresentation pathway analysis. The larger EcoToxModule gene sets do not suffer from this problem, and thus all pathway analysis embedded in the downstream visual analytics tools are conducted with the EcoToxModule and EcoToxProcess gene sets.

The EcoToxModule data are presented in both tabular and graphical formats, with the corresponding data summarizing gene expression values for all genes in a given module and not just those that are differentially expressed. A threshold toggle is provided to the user to flag whether the sample is “yellow” or “red”, which can aid in screening, prioritizing and visualizing the results. Currently, these threshold values are uniformly applied and arbitrary (similar to the use of a logFC threshold in differential expression analysis), though in the future we aim to derive data-driven thresholds that are more predictive.

Step 5: Visual analytics

A key feature of EcoToxXplorer is the development and support for visualization tools (volcano plots, Sankey plots, and heat maps) to allow users to interact with their results to gain further insights beyond the standard analyses described in earlier sections. In terms of the volcano plot, a static version was provided earlier in the workflow (see 'Differential gene expression step' in Figure 3), whereas in the Visual Analytics section, we offer a more interactive form in which users can click on a particular gene to learn more (e.g., a pop-up violin plot showing the expression values of a particular gene across treatment groups), select a certain area of the figure to explore deeper into a subset of genes, or sort the displayed genes according to various groupings (e.g., a specific EcoToxModule). The Sankey plot allows users to view and explore the EcoToxProcess and EcoToxModule data in a more interactive manner. For example, the width of the arrows is proportional to the number of genes a given EcoToxProcess or EcoToxModule contains. A user can click on a given arrow to view which genes are found in a given EcoToxProcess or EcoToxModule, along with some key summary statistics for those genes. Additionally, users can move the arrows to create their preferred figure and then download a high-quality version. The final visual analytic method, which complements the other two methods, is an interactive heat map allowing users to intuitively view the complete gene expression patterns across different EcoToxChips. Users can apply different clustering algorithms and then select regions of interest (i.e. those with clear patterns of change with regard to doses or chemical exposures) for more detailed examinations and functional analyses.

Step 6: Report generation

To help support the quality, reproducibility, and reporting of results from a given analysis, we developed a report-generating feature into EcoToxXplorer that drew inspiration from: 1) the on-going Transcriptomics Reporting Framework (TRF) project being led by the OECD Extended Advisory Group on Molecular Screening and Toxicogenomics (EAGMST), 2) a SETAC Pellston workshop with ideas on how to improve the usability of ecotoxicology data in regulatory decision making (Hanson *et al*. 2017), and 3) guidelines from the Minimum Information About a Microarray Experiment effort (MIAME). Thus, when users finish their analysis workflow, they are prompted to click the "Report" link from which an options page appears to allow users to select which sections they would like included. The default report keeps all sections, including an Introduction section (in which the user can input key information on study design) along with summaries of all tasks performed in terms of data upload, quality check, normalization, as well as key tables and figures from the various analyses performed. Each report is date and time stamped. This report is a key tool to help users properly document their analysis and permit reproducible research practices.

In addition to the Analysis Report, the "Download" link in the navigation panel allows users to download the processed numeric data, high-resolution images (PNG format), and the R command history. For anonymous or guest users, all raw and processed data files are stored on the server in a private folder for 72 hours before being deleted; for registered users, the results can be saved for up to one year.

Tutorials and Example Datasets

In order to help users navigate EcoToxXplorer, a number of supportive resources have been developed and offered on the website. First, tutorials are accessible through the 'Tutorials' link on the home page's header, within which there are files with step-by-step screenshots on how to navigate EcoToxXplorer as well as a video tutorial (https://youtu.be/BonFzvue4tg). Second, EcoToxXplorer currently has over 30 FAQs to complement the information found in the tutorials. 'Question mark' icons are also found on many pages that users can click on to learn more. Third, on the Contact page, feedback can be provided through an online form. Finally, there are three example datasets built into EcoToxXplorer that can help users navigate the workflow (Japanese quail exposed to chlorpyrifos, *Xenopus laevis* exposed to hexabromocyclododecane, fathead minnow exposed to ethinylestradiol). These can be accessed via the 'Try Example' button on the EcoToxChip Analysis page.

6.7 CONCLUDING REMARKS

There is palpable evidence that toxicity testing is undergoing a transformation. Alternatives to conventional testing approaches, which are often based on expensive and time-consuming animal studies that tend to focus on apical measures (e.g., survival, growth, development), are being sought, developed, and validated. These alternative (or new approach) methods are built on advances in biotechnology, focus on comprehensive understanding of molecular and systems biology, and favour non-animal testing strategies (i.e. *in silico*, *in vitro* and embryo-based approaches) that incorporate the "3Rs" principle (reduce, refine, replace). These alternative methods also generate big and complicated data, and thus there is a need to make the analyses standardized and the resulting information more accessible, understandable, and useable to the entire user community, not just to bioinformaticians.

In this Focus article, we detail EcoToxXplorer, a next-generation bioinformatics tool that is high performance (i.e., scalable for large data or user traffic), intuitive to use (i.e., enables complex analytics via a simple interface), universally accessible (i.e. web-based, user-friendly), and built to handle transcriptomics data for the purpose of chemical risk assessment and environmental management. In doing so, we also demonstrate how a design thinking approach can bring together innovators of new approach methods (i.e., EcoToxXplorer programmers and developers, along with toxicologists and social scientists) with adopters (i.e., users from academia, government, and industry), and ultimately co-produce a tool that is standardized and trusted to help organize, analyze, visualize, and interpret transcriptomics data for the purpose of toxicity testing. Moving ahead, we intend to update the tool every few years (as we have done with other web-based tools, Table 6-1) based on user feedback and their experiences as well as with evolving needs of the field.

6.8 REFERENCES

Basu N, Crump D, Head J, Hickey G, Hogan N, Maguire S, Xia J, Hecker M. **2019**. EcoToxChip: A next-generation toxicogenomics tool for chemical prioritization and environmental management. *Environmental Toxicology and Chemistry* 38(2):279-288.

Ewald JD, Soufan O, Crump D, Hecker M, Xia J, Basu N. **2020**. EcoToxModules: Custom gene sets to organize and analyze toxicogenomics data from ecological species. *Environmental Science and Technology* 54(7):4376-4387.

Hanson ML, Wolff BA, Green JW, Kivi M, Panter GH, Warne MS, Ågerstrand M, Sumpter JP. **2017**. How we can make ecotoxicology more valuable to environmental protection. *Science of the Total Environment* 578:228-235.

Health Canada. **2019**. Evaluation of the use of toxicogenomics in risk assessment at Health Canada: An exploratory document on current Health Canada practices for the use of toxicogenomics in risk assessment. *Health Canada Cat.* # H129-92/2018E-PDF.

Helmy M, Crits-Christoph A, Bader GD. **2016**. Ten simple rules for developing public biological databases. *PLoS Computational Biology* 12(11):e1005128.

Mondou, M., Hickey, G.M., Rahman, H.M.T., Maguire, S., Pain, G., Crump, D., Hecker, M. and Basu, N. **2020**. Factors affecting the perception of New Approach Methodologies (NAMs) in the ecotoxicology community. *Integrated Environmental Assessment and Management* 16:269-281.

Mondou, M., Maguire, S., Pain, G., Crump, D., Hecker, M., Basu, N., Hickey, G. **2021**. http://dx.doi.org/10.1002/ieam.4244 Envisioning an international validation process for New Approach Methodologies in chemical hazard and risk assessment. *Environmental Advances* 4: 100061.

Pain G, Hickey G, Mondou M, Crump D, Hecker M, Basu N, Maguire S. **2020**. Drivers of and Obstacles to the adoption of toxicogenomics for chemical risk assessment: Insights from social science perspectives. *Environmental Health Perspectives* 128(10):105002.

Vachon J, Campagna C, Rodriguez MJ, Sirard MA, Levallois P. **2017**. Barriers to the use of toxicogenomics data in human health risk assessment: A survey of Canadian risk assessors. *Regulatory Toxicology and Pharmacology* 85:119-123.

Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson T, LaLone CA, Landesmann B, Lettieri T, Munn S, Nepelska M, Ottinger MA, Vergauwen L, Whelan M. **2014**. Adverse Outcome Pathway (AOP) development I: Strategies and principles. *Toxicological Sciences* 142:312-320.

Chapter 7. GENERAL DISCUSSION

7.1 Toxicogenomics to support decision-making

My understanding of what "designing toxicogenomics to support decision-making in environmental toxicology" (title of this thesis) means has changed greatly since I started my doctoral work in 2017. At the outset, I interpreted it as working towards a machine learning algorithm that could universally predict toxicity from gene expression data and that would one day be capable of replacing some existing regulatory systems. Five years later, I understand that the way toxicity is defined is too context-dependent for simple and universal replacement. Chemical management is also situational. Within an individual country or province, different classes of substances are regulated by different agencies, each with its own legislation and mandates. Decisions are not a simple yes or no given by an individual person. Instead, they are characterized by large teams of people, integrating multiple data streams over periods of months to years, while trying to maintain public trust and balance the (often conflicting) needs of many stakeholders.

Over time, my interpretation of "toxicogenomics to support decision-making" has changed to making toxicogenomics data analysis and interpretation as accessible to as many people as possible. Put another way, my objective has become to empower rather than to replace. This is reflected in the breakdown of my work: for every hour spent on developing statistical methods, ten were spent on software infrastructure, user interface and user experience (UI/UX) considerations, documentation, and tutorials. Toxicogenomics is a rich source of information that has the power to transform chemical risk assessment, however, to do so, it must be usable and understandable by a broad group of people.

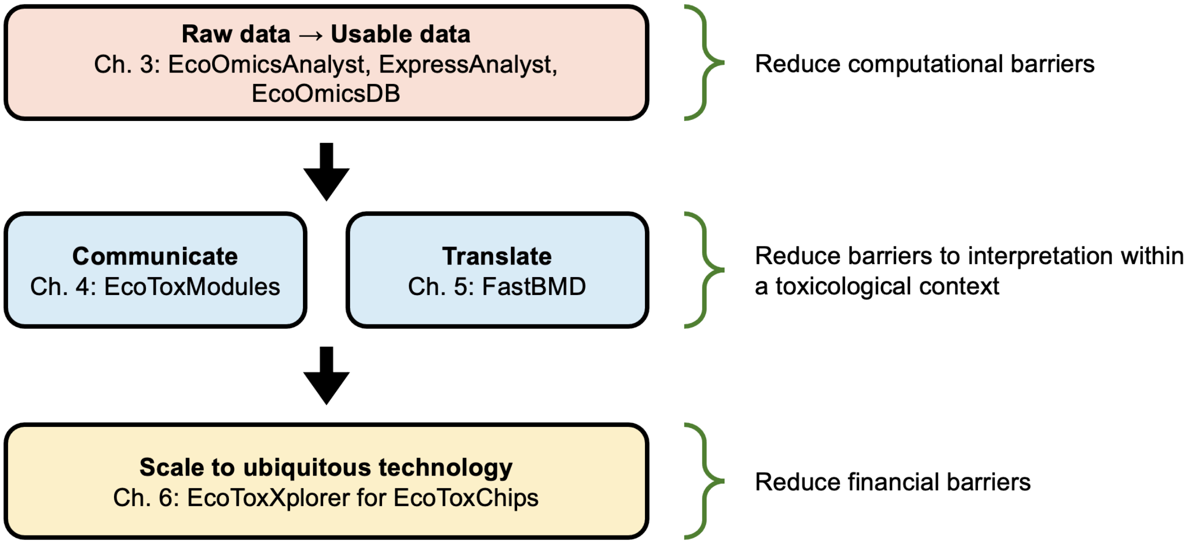

Raw data → Usable data
Ch. 3: EcoOmicsAnalyst, ExpressAnalyst, EcoOmicsDB
Reduce computational barriers
Communicate
Ch. 4: EcoToxModules
Translate
Ch. 5: FastBMD
Reduce barriers to interpretation within a toxicological context
Scale to ubiquitous technology
Ch. 6: EcoToxXplorer for EcoToxChips
Reduce financial barriers

Figure 7-1: Overview of barriers to toxicogenomics that were addressed by thesis chapters 3 – 6.

Chapters 3 – 6 presented various software and tools that contribute towards the goal of making toxicogenomics more accessible (Figure 7-1). Chapter 3 addressed raw RNA-seq data processing for non-model organisms, which is a large computational barrier (both in terms of computing resources and programming expertise) to many researchers working with ecologically relevant species. Chapters 4 and 5 addressed difficulties in interpreting and communicating transcriptomics data within an environmental toxicology context, through custom gene sets aimed at giving a big picture overview of the complicated data (Chapter 4) and dose-response methods that translate toxicogenomics results into statistics that are familiar to risk assessors (Chapter 5). Chapter 6 presented an analytical tool for EcoToxChips, which make measuring a panel of 384 genes more financially viable for labs that already have access to qPCR technology. EcoToxXplorer incorporates statistical methods from chapters 4 and 5 and was the original home of the raw data processing pipelines that became chapter 3.

7.2 DESIGN-THINKING IN CHAPTERS 3-6

Each of chapters 3 – 6 presented snap shots in time of statistical methods and software that have been and will continue to be continuously updated and improved. This follows the design-thinking framework, with its focus on cycles of prototyping and testing. This section highlights some of the groundwork and follow-up to methods and software presented in these chapters.

Chapter 3 is the first description of novel statistical methods and software in this thesis, but chronologically it is the most recent. However, ExpressAnalyst contains mainly features that were split from the NetworkAnalyst software, most recently published in a paper that I was a co-author of (Zhou *et al.*, 2019). EcoOmicsAnalyst is the continuation of the raw data processing pipelines that were initially hosted by EcoToxXplorer (chapter 5) (Soufan and Ewald *et al.*, 2021). Chapter 3 also describes major updates to the Seq2Fun algorithm, compared to version 1.0 published in a paper that I was also a co-author of (Liu *et al.*, 2021). The updates to Seq2Fun were motivated by various case studies with researchers at McGill University, University of Saskatchewan, Huntsman Marine Science Centre, US Geological Survey, and Environment and Climate Change Canada, including studies on the endangered copper redhorse, double crested cormorant, lobster, sturgeon, and rainbow trout.

Chapter 4 introduces the EcoToxModule gene sets and was published in 2020 (Ewald *et al.*, 2020). Since then, feedback from researchers within McGill University resulted in an update to the KEGG pathways included in the functional modules, which included merging the "Cellular Process – Transcription" and "Cellular Process – Protein Production" module into "Cellular Process – Transcription and Translation" and adding a "Metabolism – Other" module with more metabolism pathways. The updated EcoToxModules were used by the McGill University researchers to compare the transcriptomics response to chlorpyrifos exposures in Japanese quail and double-crested cormorant embryos (Desforges *et al.*, 2021). A second update was made in 2021 to manually add EcoToxModule annotations to all genes included in the EcoToxChips for Japanese quail, African clawed frog, and fathead minnow, and this version was incorporated into EcoToxXplorer (chapter 6) (Soufan and Ewald *et al.*, 2021). These additional annotations were also incorporated into the files for analyzing transcriptomics data. The most recent version of the EcoToxModule gene sets is available on the "Resources" tab of www.ecotoxxplorer.ca.

Since its publication in 2020, FastBMD (chapter 5) has been updated to include fathead minnow, rainbow trout, and Seq2Fun ortholog ID and pathway files, as requested and published by users. (Alcaraz *et al.*, 2021; Alcaraz *et al.*, 2022) In addition, the dose-response algorithm has been incorporated into the EcoToxChip pipeline in EcoToxXplorer, which involved developing custom normalization methods to enable non-linear curve fitting for qPCR data. Evaluation and publication of the qPCR pipeline are still underway.

7.3 ADDITIONAL OBSTACLES FACING TOXICOGENOMICS FOR REGULATORY DECISION-MAKING

In addition to "complexity of data analysis and interpretation", Pain *et al.*, identified "insufficient validation" and "lack of standardization" as obstacles to the adoption of toxicogenomics for chemical risk assessment (Pain *et al*, 2020). The OECD defines validation as demonstrating the reliability and relevance of a given method to a specific use case (Environment Directorate, 2005). In the context of toxicogenomics, establishing reliability has mainly meant demonstrating reproducibility of results between laboratories, platforms, and time points (Eskes and Whelan, 2016), while establishing relevance has meant demonstrating that use of ‘omics data leads to superior biological understanding (Pettit *et al*, 2010). Achieving widespread validation is further complicated by the fact that each government agency and use case may require different criteria (Malloy *et al.*, 2017). Addressing concerns about validation is largely beyond the scope of this thesis, especially because these activities are typically driven by institutions such as the OECD and involve coordination of multiple laboratories and government agencies (Mondou *et al.*, 2021). However, the software presented in this thesis could make such validation efforts easier by making uniform analytical workflows accessible to many different people in different locations. This could also contribute towards improving the reproducibility of toxicogenomics approaches.

The lack of standardization for toxicogenomics data is being addressed by efforts such as the transcriptomics reporting framework (Gant *et al.*, 2017), an international and multi-institution collaboration to establish parameters and methodology that should be reported by all ‘omics studies that are conducted within a regulatory context. A similar effort is underway for qPCR data, and we have been in contact with people involved in this effort to ensure that each analysis parameter in the reporting framework is exposed by EcoToxXplorer workflows so that users can report them. In the future, the plan is to incorporate the reporting framework parameters into the PDF report that can be generated at the end of the EcoToxChip analytical pipeline.

7.4 Strengths and limitations of this thesis

7.4.1 Limitations

This thesis focused on developing toxicogenomics methods and software for ecological species. One limitation is the types of species that were included in the definition of "ecological species", which in this thesis were limited to non-mammalian vertebrate species (ie. representatives of birds, fishes, and amphibians). Chapter 3, with its scope of enabling raw RNA-seq reads processing and analysis for any eukaryotic species, did cover more taxonomic categories. However, even this chapter was biased towards vertebrate species as these taxonomic categories were better represented by both reference transcriptomes for Kallisto, and in transcriptomes included in the ortholog databases for Seq2Fun. This reflects the bias towards vertebrate species in published whole genomes (Liu *et al*, 2021). Chapters 4 – 6 used exclusively vertebrate species. Important plant (ie. *Arabidopsis*) and invertebrate species (ie. *Daphnia*) are used in toxicity testing (Jonczyk and Gilron, 2005; Cobbett, 2003), however these were largely beyond the scope of this thesis. This is partly because this research was connected to the EcoToxChip project, which is based on vertebrate representatives of fish, bird, and frog species as part of an effort to address ethical concerns related to animal toxicity testing according to the "3Rs" principles of reduction, replacement, and refinement (Basu *et al.*, 2019).

The focus on using toxicogenomics data from ecologically relevant species also limited the scope of statistical methods that could be investigated and developed, due to the lack of large, uniform datasets that are well-suited for predictive modeling. There are large transcriptomics datasets available for mammalian species (both *in vivo* and *in vitro*) that include exposures to hundreds or thousands of compounds such as TG-GATES, DrugMatrix, and the LINCS L1000 datasets (Igarashi *et al.*, 2015; Svoboda *et al.*, 2019; Subramanian *et al.*, 2017). These datasets have enabled the development of statistical models for predicting adverse outcomes from gene expression data (Mohsen *et al.*, 2021; Ganter *et al.*, 2006; Gusenleitner *et al.*, 2014). The largest dataset from ecological species in this thesis was the fathead minnow dataset used to develop the statistical EcoToxModules in chapter 4 (sample size n = 303 microarray profiles). To achieve this size, samples were combined from 11 studies that used different exposure methods, creating a heterogeneous dataset that was not well suited for predictive modeling. However, since the beginning of this thesis, the EcoToxChip RNA-seq dataset has been collected (n > 700 samples across six species, eight chemicals, and multiple life stages), which I have begun using for preliminary cross-species analysis after processing all samples with Seq2Fun. There are also initiatives such as the US EPA's TARGET EcoTox Challenge to promote the creation of medium-to-high throughput toxicogenomics tools for ecological species. Hopefully, activity like this will spur research on predictive toxicogenomics models for ecologically relevant species in the coming years.

7.4.2 STRENGTHS

Overall, this thesis presents user-friendly software that make toxicogenomics analysis accessible to researchers with no previous training in bioinformatics. In my experience, it typically takes someone who already has previous training in both college-level biology and statistics three to four months of focused learning (the equivalent of one graduate-level bioinformatics course) before they can confidently perform basic differential expression and pathway analysis in R. This is consistent with concerns that increasing capacity for toxicogenomics is expensive and time consuming (Balbus and Environmental Defense, 2005; ECETOC 2007). With the software in this thesis, multiple researchers that I have worked with have been able to perform this standard analysis, as well as more advanced toxicogenomics analysis, interpretation, and visualization tasks, after one to two weeks of exploring the example datasets, tutorials, and FAQs provided online. Empowering more toxicologists to directly perform toxicogenomics analysis will help with the effort to integrate big molecular data into the activities of environmental toxicology research programs and government agencies worldwide (Thomas *et al*., 2019).

One strength of this thesis was the number of people who were involved across all the chapters, both those on the co-author lists and those listed in the acknowledgement sections. Working with and alongside the EcoToxChip project (www.ecotoxchip.ca) gave me access to a group of ~160 people that work in academia, government, and industry (Basu *et al.*, 2019). Previous sections of this thesis have discussed the design thinking framework, and it was the involvement of this large group of people that made applying this framework possible. Working with such a diversity of expertise and qualifications continually stretched my understanding of toxicity testing, and what needs to happen to create change in this field. This learning is mainly reflected in choices to include simpler and better understood statistical methods in the software developed for chapters 3 – 6, rather than the most novel and cutting edge possible. This focus is supported by the conclusions and recommendations of Pain *et al.* 2020:

> *"We observed that an innovation-centric perspective appears to have dominated discourse about the adoption of toxicogenomics in chemical risk assessment during the period 1998–2017, with proponents extolling the tools' putative superior and novel functionality but overlooking the tools' and data sets' understandability, ease of use, and fit with users' routines. We recommend that more attention be placed on ensuring the simplicity and compatibility of toxicogenomics tools and data, as well as creating opportunities for potential adopters to experiment with them directly*

> *(trialability) and vicariously (observability). We also conclude that the innovation-centric perspective would be usefully balanced with an adopter-centric one that highlights the importance of skill development and organizational learning in the adopting system."*

This quote highlights an additional unexpected strength of this thesis. While the methods and software were created mainly for researchers to use to analyze their data, over time they have also shown their usefulness as educational tools. Since each of the software are web-based and have built in example datasets, it is extremely easy to invite people to pull up the websites and start interacting directly with toxicogenomics data and workflows. Throughout my PhD, this has happened many times, both formally during conference presentations, workshops, and courses, and informally during conference networking sessions or after meetings. I am also aware of several professors who have incorporated FastBMD (chapter 5) into their undergraduate toxicology courses.

CHAPTER 8. GENERAL CONCLUSION

This thesis used a design-thinking framework to design new statistical methods and corresponding software for analyzing and visualizing toxicogenomics data to support decision-making in the context of environmental toxicology. By focusing on the needs of end-users, a focus was put on making the software easy to understand and use. Together, the software described in this thesis provide a powerful toolkit for analyzing transcriptomics data from ecological species for toxicogenomics applications. Specifically, there are tools to support data analysis for raw data processing (EcoOmicsAnalyst), filtering, normalization, and differential analysis (ExpressAnalyst, EcoToxXplorer, and FastBMD), dose-response modeling (FastBMD and EcoToxXplorer), visual analytics and figure preparation (ExpressAnalyst, FastBMD, and EcoToxXplorer), and various knowledgebases focused on toxicology and ecological species (EcoOmicsDB, EcoToxXplorer – EcoToxModules and AOPwiki genes). These methods and software reduce barriers to toxicogenomics analysis for people who do not have computing resources, programming skills, advanced statistical training, and knowledge of bioinformatics databases.

MASTER REFERENCE LIST

This section lists references from the front-matter, general introduction, connecting paragraphs, general discussion, and general conclusion sections.

Alcaraz, A. J. G., Mikulasek, K., Potesil, D., Park, B., Shekh, K., Ewald, J., *et al*. **2021**. Assessing the toxicity of 17α-ethinylestradiol in rainbow trout using a 4-day transcriptomics benchmark dose (BMD) embryo assay. *Environmental Science & Technology*, 55(15), 10608-10618.

Alcaraz, A. J. G., Baraniuk, S., Mikulášek, K., Park, B., Lane, T., Burbridge, C., *et al*. **2022**. Comparative analysis of transcriptomic points-of-departure (tPODs) and apical responses in embryo-larval fathead minnows exposed to fluoxetine. *Environmental Pollution*, 295, 118667.

Basu, N., Crump, D., Head, J., Hickey, G., Hogan, N., Maguire, S., *et al*. **2019**. EcoToxChip: a next-generation toxicogenomics tool for chemical prioritization and environmental management. *Environmental Toxicology and Chemistry.*

Balbus JM, Environmental Defense. **2005**. *Toxicogenomics: Harnessing the Power of New Technology*. New York, NY: Environmental Defense.

Cobbett, C. **2003**. Heavy metals and plants: Model systems and hyperaccumulators. *New Phytologist*, 289-293.

Desforges, J. P., Legrand, E., Boulager, E., Liu, P., Xia, J., Butler, H., *et al*. **2021**. Using Transcriptomics and Metabolomics to Understand Species Differences in Sensitivity to Chlorpyrifos in Japanese Quail and Double-Crested Cormorant Embryos. *Environmental Toxicology and Chemistry*, *40*(11), 3019-3033.

ECETOC. **2007**. *Workshop on the Application of Omic Technologies in Toxicology and Ecotoxicology: Case Studies and Risk Assessment*, 6-7 December 2007, Malaga. Brussels, Belgium: ECETOC.

Environment Directorate, OECD **2005**. Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment.

Eskes, C., Whelan, M. (Eds.). **2016**. *Validation of alternative methods for toxicity testing* (Vol. 856). Springer.

Ewald, J. D., Soufan, O., Crump, D., Hecker, M., Xia, J., Basu, N. **2020**. EcoToxModules: custom gene sets to organize and analyze toxicogenomics data from ecological species. *Environmental Science and Technology*, 54(7), 4376-4387.

Gant, T. W., Sauer, U. G., Zhang, S. D., Chorley, B. N., Hackermüller, J., Perdichizzi, S., *et al*. **2017**. A generic transcriptomics reporting framework (TRF) for 'omics data processing and analysis. *Regulatory Toxicology and Pharmacology*, 91, S36-S45.

Ganter, B., Snyder, R. D., Halbert, D. N., Lee, M. D. **2006**. Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix® database. *Pharmacogenomics.* 1025-1044.

Gusenleitner, D., Auerbach, S. S., Melia, T., Gomez, H. F., Sherr, D. H., Monti, S. **2014**. Genomic models of short-term exposure accurately predict long-term chemical carcinogenicity and identify putative mechanisms of action. *PloS One*, 9(7), e102579.

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., Yamada, H. **2015**. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Research*, *43*(D1), D921-D927.

Jonczyk, E., Gilron, G. **2005**. Acute and chronic toxicity testing with Daphnia sp. In *Small-scale freshwater toxicity investigations* (pp. 337-393). Springer, Dordrecht.

Liu, P., Ewald, J., Galvez, J. H., Head, J., Crump, D., Bourque, G., *et al*. **2021**. Ultrafast functional profiling of RNA-seq data for nonmodel organisms. *Genome Research*, *31*(4), 713-720.

Malloy, T., Zaunbrecher, V., Beryt, E., Judson, R., Tice, R., Allard, P., *et al*. **2017**. Advancing alternatives analysis: the role of predictive toxicology in selecting safer chemical products and processes. *Integrated Environmental Assessment and Management*, 13(5), 915-925.

Mohsen, A., Tripathi, L. P., Mizuguchi, K. **2021**. Deep Learning Prediction of Adverse Drug Reactions in Drug Discovery Using Open TG–GATEs and FAERS Databases. *Frontiers in Drug Discovery*, 3.

Mondou, M., Maguire, S., Pain, G., Crump, D., Hecker, M., Basu, N., Hickey, G. M. **2021**. Envisioning an international validation process for New Approach Methodologies in chemical hazard and risk assessment. *Environmental Advances*, 4, 100061.

Pain, G., Hickey, G., Mondou, M., Crump, D., Hecker, M., Basu, N., Maguire, S. **2020**. Drivers of and obstacles to the adoption of toxicogenomics for chemical risk assessment: insights from social science perspectives. *Environmental Health Perspectives*, 128(10), 105002.

Pettit, S., Des Etages, S. A., Mylecraine, L., Snyder, R., Fostel, J., Dunn, R. T., *et al*. **2010**. Current and future applications of toxicogenomics: Results summary of a survey from the HESI Genomics State of Science Subcommittee. *Environmental Health Perspectives*, 118(7), 992-997.

Soufan, O., Ewald, J., Zhou, G., Hacariz, O., Boulanger, E., Alcaraz, A. J., *et al*. **2022**. EcoToxXplorer: Leveraging Design Thinking to Develop a Standardized Web-Based Transcriptomics Analytics Platform for Diverse Users. *Environmental Toxicology and Chemistry*, 41(1), 21-29.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X. **2017**. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6), 1437-1452.

Svoboda D.L., Saddler T., Auerbach S.S. **2019**. An Overview of National Toxicology Program's Toxicogenomic Applications: DrugMatrix and ToxFX. *Advances in Computational Toxicology. Challenges and Advances in Computational Chemistry and Physics*, vol 30. Springer, Cham.

Thomas, R. S., Bahadori, T., Buckley, T. J., Cowden, J., Deisenroth, C., Dionisio, K. L., *et al.* **2019**. The next generation blueprint of computational toxicology at the US Environmental Protection Agency. *Toxicological Sciences*, 169(2), 317-332.

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E., Basu, N., Xia, J. **2019**. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research*, 47(W1), W234-W241.

Appendix: Supplementary information for chapters 3-6

SI for Chapter 3

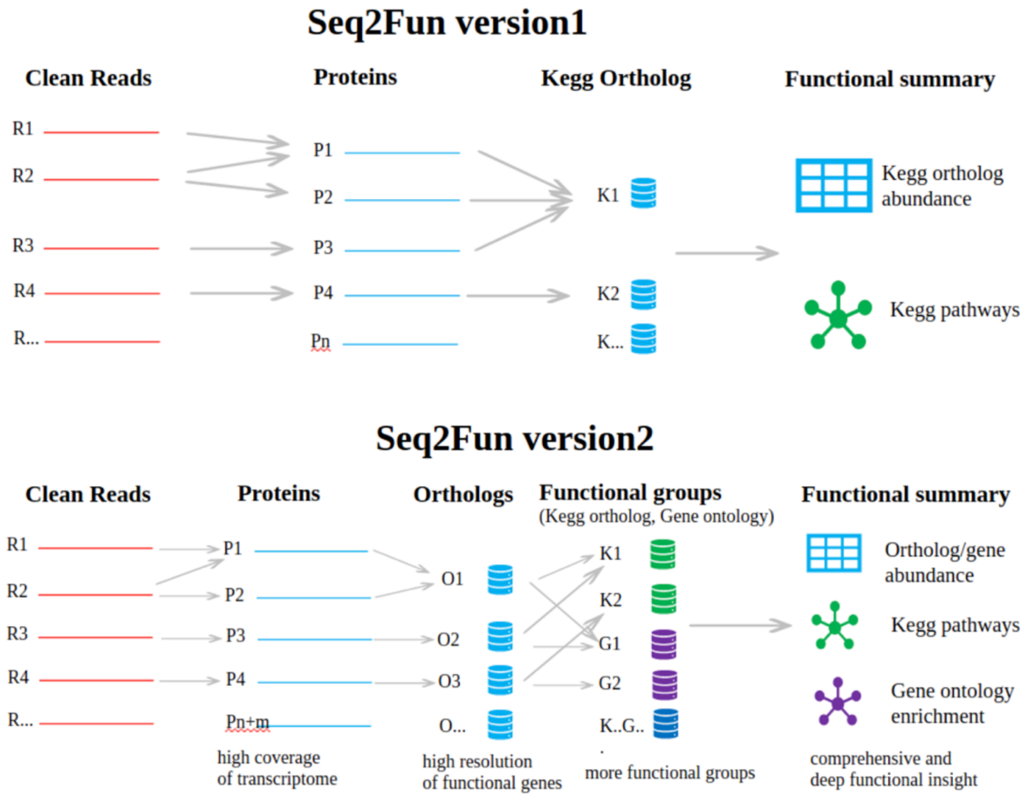


Figure S1. Difference in ortholog mapping systems between Seq2Fun version 1.0 and version 2.0.

Table S1: Significantly enriched pathways for case study #1 - zebrafish, Kallisto dataset.

Pathway	Total	Hits	FDR	Library
OBS20				
Proteolysis	457	9	0.00283	GO BP
Lipid metabolic process	163	6	0.00283	GO BP
Xenobiotic metabolic process	4	2	0.00733	GO BP
Extracellular region	574	10	0.019	GO CC
Protein methyltransferase activity	534	13	0.000414	GO MF
Aldo_keto reductase (NADP) activity	22	4	0.000551	GO MF
Phosphatase inhibitor activity	100	5	0.0122	GO MF
Lipid transporter activity	56	4	0.0122	GO MF
Cytochrome_c oxidase activity	112	5	0.0133	GO MF
Unfolded protein binding	135	5	0.0262	GO MF
Calmodulin binding	298	7	0.0268	GO MF
Growth factor binding	91	4	0.0391	GO MF
Metabolism of xenobiotics by cytochrome P450	37	5	1.21E-05	KEGG
Metabolic pathways	1500	15	9.81E-05	KEGG
Drug metabolism - cytochrome P450	34	4	0.000172	KEGG
Glutathione metabolism	59	4	0.0012	KEGG
Drug metabolism - other enzymes	66	4	0.0015	KEGG
alpha-Linolenic acid metabolism	18	2	0.0388	KEGG
OBS30				
Proteolysis	457	29	1.83E-08	GO BP
Lipid metabolic process	163	13	0.000232	GO BP
Lipid transport	60	7	0.004	GO BP
Immune response	181	10	0.0418	GO BP
Extracellular region	574	32	5.96E-08	GO CC
Extracellular space	382	17	0.00982	GO CC
Unfolded protein binding	135	18	1.69E-09	GO MF
Calmodulin binding	298	23	7.64E-08	GO MF
Growth factor binding	91	13	2.36E-07	GO MF
Protein methyltransferase activity	534	23	0.00139	GO MF
Lipid transporter activity	56	7	0.00193	GO MF
Receptor binding	13	4	0.00217	GO MF

Enzyme inhibitor activity	88	8	0.00387	GO MF
Phosphatase inhibitor activity	100	8	0.00832	GO MF
Aldo_keto reductase (NADP) activity	22	4	0.0132	GO MF
Steroid biosynthesis	19	4	0.016	KEGG
PFOS				
G_protein coupled receptor signaling pathway	732	36	0.0138	GO BP
Transmission of nerve impulse	11	4	0.0156	GO BP
Response to wounding	12	4	0.0156	GO BP
Ion transport	424	23	0.0242	GO BP
Response to stress	16	4	0.0319	GO BP
Extracellular region	574	41	3.63E-07	GO CC
Enzyme inhibitor activity	88	16	3.75E-08	GO MF
Protein serine/threonine/tyrosine kinase activity	673	35	0.000666	GO MF
Transmembrane signaling receptor activity	178	14	0.00592	GO MF
Cysteine_type peptidase activity	11	4	0.00592	GO MF
Amino acid transmembrane transporter activity	18	4	0.0387	GO MF
Neuroactive ligand-receptor interaction	499	38	1.16E-11	KEGG

Table S2: Significantly enriched pathways for case study #1 - zebrafish, Seq2Fun dataset.

Pathway	Total	Hits	FDR	Library
OBS20				
Extracellular region	772	16	0.000222	GO CC
Extracellular space	253	9	0.000835	GO CC
Staphylococcus aureus infection	77	4	0.0331	KEGG
OBS30				
Proteolysis	553	24	0.00109	GO BP
Lipoprotein metabolic process	16	5	0.00237	GO BP
Lipid transport	61	7	0.0163	GO BP
Immune response	117	9	0.0203	GO BP
Signal transduction	1080	31	0.0234	GO BP
Extracellular region	772	55	8.20E-21	GO CC
Extracellular space	253	20	3.16E-07	GO CC
Hormone activity	90	13	1.08E-06	GO MF
Iron ion binding	99	11	0.000192	GO MF
Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	66	8	0.00225	GO MF
Peptidase activity	325	17	0.00225	GO MF
Serine_type peptidase activity	89	9	0.00225	GO MF
Serine_type endopeptidase activity	126	10	0.00499	GO MF
Heme binding	92	8	0.0162	GO MF
Oxidoreductase activity	472	18	0.0459	GO MF
Cholesterol metabolism	90	10	0.000158	KEGG
Cytokine-cytokine receptor interaction	319	16	0.00151	KEGG
Fat digestion and absorption	65	7	0.00385	KEGG
Staphylococcus aureus infection	77	7	0.00869	KEGG
IL-17 signaling pathway	127	8	0.0264	KEGG
Vitamin digestion and absorption	47	5	0.0264	KEGG
Malaria	73	6	0.0264	KEGG
Steroid biosynthesis	28	4	0.0264	KEGG
Neuroactive ligand-receptor interaction	604	19	0.0264	KEGG
African trypanosomiasis	58	5	0.0473	KEGG
PFOS				

Signal transduction	1080	73	5.64E-17	GO BP
G protein_coupled receptor signaling pathway	626	41	6.23E-08	GO BP
Immune response	117	12	0.00517	GO BP
Extracellular region	772	65	1.70E-17	GO CC
Extracellular space	253	20	0.000601	GO CC
G protein_coupled receptor activity	566	38	1.37E-07	GO MF
Hormone activity	90	15	3.15E-07	GO MF
Cytokine activity	62	10	0.000277	GO MF
Cytokine-cytokine receptor interaction	319	32	1.75E-13	KEGG
Neuroactive ligand-receptor interaction	604	39	9.44E-11	KEGG
IL-17 signaling pathway	127	16	6.80E-08	KEGG
TNF signaling pathway	183	16	1.03E-05	KEGG
Rheumatoid arthritis	152	13	0.000175	KEGG
Toll-like receptor signaling pathway	156	13	0.000195	KEGG
Malaria	73	8	0.00204	KEGG
JAK-STAT signaling pathway	199	13	0.00204	KEGG
Th17 cell differentiation	176	12	0.00238	KEGG
Pertussis	125	10	0.0024	KEGG
Lipid and atherosclerosis	352	17	0.00459	KEGG
C-type lectin receptor signaling pathway	203	12	0.00663	KEGG
NF-kappa B signaling pathway	174	11	0.00663	KEGG
Viral protein interaction with cytokine and cytokine receptor	104	8	0.0124	KEGG
Chagas disease	194	11	0.0146	KEGG
Inflammatory bowel disease	84	7	0.0154	KEGG
Relaxin signaling pathway	231	12	0.0159	KEGG
Primary immunodeficiency	47	5	0.0282	KEGG
Estrogen signaling pathway	226	11	0.04	KEGG
Th1 and Th2 cell differentiation	135	8	0.0473	KEGG
Hepatitis B	271	12	0.0473	KEGG
Leishmaniasis	108	7	0.0473	KEGG
Non-alcoholic fatty liver disease	272	12	0.0473	KEGG

Table S3: Significantly enriched pathways for case study #1 - lobster, Seq2Fun dataset for the POSl vs. CTRL contrast. There were no significantly enriched pathways for the WAF_72 vs. CTRL contrast.

Pathway	Total	Hits	FDR	Library
Metabolic process	102	7	0.000372	GO BP
Oxidoreductase activity	479	13	0.00226	GO MF
Hydrolase activity	1050	19	0.00226	GO MF
Hydrolase activity, acting on glycosyl bonds	85	6	0.00434	GO MF
Diacylglycerol O_acyltransferase activity	3	2	0.0445	GO MF
Metabolic pathways	2360	35	5.93E-05	KEGG
Biosynthesis of secondary metabolites	727	15	0.00758	KEGG
Carbohydrate digestion and absorption	34	4	0.00799	KEGG

Table S4: Significantly enriched pathways in the list of up-regulated genes from case study #2.

Pathway	Total	Hits	FDR	Library
Osteoclast differentiation	162	22	1.04E-05	KEGG
MicroRNAs in cancer	215	23	0.000207	KEGG
TNF signaling pathway	148	18	0.000342	KEGG
Transcriptional misregulation in cancer	324	26	0.00362	KEGG
Lipid and atherosclerosis	307	25	0.00362	KEGG
MAPK signaling pathway	334	26	0.00455	KEGG
Small cell lung cancer	129	14	0.00685	KEGG
Adipocytokine signaling pathway	92	11	0.0138	KEGG
Parathyroid hormone synthesis, secretion and action	125	13	0.0142	KEGG
Cholesterol metabolism	66	9	0.0149	KEGG
IL-17 signaling pathway	115	12	0.0179	KEGG
Bladder cancer	56	8	0.0179	KEGG
Toxoplasmosis	150	14	0.0179	KEGG
Plasma membrane	975	63	0.022	GO CC
Protein tyrosine kinase activity	63	12	0.0264	GO MF
NF-kappa B signaling pathway	140	13	0.0272	KEGG
Growth hormone synthesis, secretion and action	143	13	0.0309	KEGG
Phospholipase D signaling pathway	200	16	0.0325	KEGG
Ferroptosis	50	7	0.0337	KEGG
Pathways in cancer	703	39	0.0349	KEGG
Biosynthesis of nucleotide sugars	52	7	0.0363	KEGG
Human T-cell leukemia virus 1 infection	308	21	0.0363	KEGG
Glutathione metabolism	67	8	0.0363	KEGG
Amino sugar and nucleotide sugar metabolism	71	8	0.0481	KEGG

Table S5: Significantly enriched pathways in the list of down-regulated genes from case study #2.

Pathway	Total	Hits	FDR	Library
Mitochondrial inner membrane	148	39	7.05E-19	GO CC
Oxidative phosphorylation	188	37	1.12E-17	KEGG
Diabetic cardiomyopathy	314	47	1.12E-17	KEGG
Cardiac muscle contraction	99	28	1.12E-17	KEGG
Mitochondrion	443	64	3.77E-17	GO CC
Thermogenesis	281	43	1.59E-16	KEGG
Non-alcoholic fatty liver disease	227	35	1.93E-13	KEGG
Prion disease	368	44	5.45E-13	KEGG
Chemical carcinogenesis - reactive oxygen species	299	38	5.31E-12	KEGG
Citrate cycle (TCA cycle)	50	15	5.98E-10	KEGG
Biosynthesis of secondary metabolites	599	51	1.24E-09	KEGG
Carbon metabolism	191	27	1.24E-09	KEGG
Hypertrophic cardiomyopathy	93	18	1.20E-08	KEGG
Dilated cardiomyopathy	105	18	8.67E-08	KEGG
Retrograde endocannabinoid signaling	158	22	8.67E-08	KEGG
Amyotrophic lateral sclerosis	484	40	3.16E-07	KEGG
Respirasome	39	13	4.53E-07	GO CC
Electron transport chain	31	12	3.45E-06	GO BP
Z disc	8	6	2.07E-05	GO CC
Tricarboxylic acid cycle	31	11	2.60E-05	GO BP
Calcium ion binding	443	44	3.44E-05	GO MF
Sarcolemma	14	7	6.68E-05	GO CC
Glycolysis / Gluconeogenesis	108	14	0.000118	KEGG
Pyruvate metabolism	67	11	0.000118	KEGG
Arginine and proline metabolism	58	10	0.000191	KEGG
Sarcoplasmic reticulum	7	5	0.000264	GO CC
Arrhythmogenic right ventricular cardiomyopathy	80	11	0.000615	KEGG
Cysteine and methionine metabolism	73	10	0.00138	KEGG
Mitochondrial matrix	38	9	0.00138	GO CC
Catalytic activity	384	36	0.00147	GO MF
ECM-receptor interaction	156	15	0.0017	KEGG

Adrenergic signaling in cardiomyocytes	175	16	0.00176	KEGG
Mitochondrial respiratory chain complex IV	11	5	0.00339	GO CC
Central carbon metabolism in cancer	98	11	0.00342	KEGG
Calcium signaling pathway	276	20	0.00543	KEGG
Basement membrane	21	6	0.0086	GO CC
Oxidoreductase activity	451	37	0.014	GO MF
Pyridoxal phosphate binding	51	10	0.014	GO MF
Propanoate metabolism	39	6	0.0158	KEGG
Oxidoreductase activity, acting on NAD(P)H	11	5	0.016	GO MF
Biosynthesis of amino acids	141	12	0.0191	KEGG
beta-Alanine metabolism	28	5	0.0193	KEGG
Glycolytic process	31	8	0.0255	GO BP
Muscle organ development	11	5	0.0348	GO BP
Valine, leucine and isoleucine degradation	48	6	0.0396	KEGG
Starch and sucrose metabolism	48	6	0.0396	KEGG
cGMP-PKG signaling pathway	199	14	0.0409	KEGG
Oxidative phosphorylation	12	5	0.0462	GO BP

SI FOR CHAPTER 4

In addition to the textual supplementary tables provided below, there is an Excel spreadsheet with supplementary tables S1-S9 in the *Supporting Information* section of the published manuscript. It can be found at https://pubs.acs.org/doi/10.1021/acs.est.9b06607?goto=supporting-info.

WGCNA workflow used to define the statistical EcoToxModules

Prior to computing pairwise co-expression between genes, the microarray data were first processed, background corrected and normalized with the limma R package using functions developed for the Agilent platform and parameters for single-channel microarrays (Langfelder *et al.*, 2008; Ritchie *et al.*, 2015). The "normexp" method (offset = 16) was used for background correction (Edwards, 2003; Silver *et al.*, 2008), and the "quantile" method was used for between-array normalization. Probes were ranked by their variance and the lowest 15% across all datasets were removed. Previous work has shown that removing up to 50% of genes based on variance can decrease the false positive rate during differential expression analysis (Bourgon *et al.*, 2010). Since we combined data from multiple studies, we chose 15% with the rationale that a threshold as high as 50% may filter out many genes that were significantly perturbed by only a few of the exposure conditions. We did not attempt to account for potential batch effects because it has been shown that Pearson correlation is a valid method of computing connections between microarray probes using the WGCNA workflow, even in the presence of significant batch effects (Li *et al.*, 2018).

The matrix of normalized gene expression values was transformed into a similarity matrix by calculating the absolute value of the pairwise Pearson correlation of the probes. Next, a soft threshold parameter (β) was chosen such that the adjacency function transforms the similarity matrix to one that approximates scale-free topology (Yip *et al.*, 2007). For each β = [2, 4, … 20], the similarity matrix was raised to the power β, and the mean connectivity and a scale-free topology fit index were calculated. Since connectivity decreases as β increases, the lowest β parameter that resulted in an $R^2 > 0.8$ was selected (β = 8, Figure S1). The similarity matrix was further transformed using the topological overlap measure (TOM), which uses information from neighboring nodes to reduce the effect of random noise on inferred connections between genes (Yip *et al.*, 2007).

Hierarchical clustering and the dynamic tree cut algorithm were used to detect clusters in the co-expression network (Langfelder *et al.*, 2008). First, the dissimilarity matrix, computed as 1 – TOM similarity matrix, was hierarchically clustered into a dendrogram using the average linkage distance measure. The dynamic tree cut algorithm was then used to detect clusters within the dendrogram. The algorithm has parameters that can be changed to adjust the cluster detection sensitivity. A low sensitivity results in a smaller number of larger clusters, while high sensitivity results in a larger number of smaller clusters (Langfelder *et al.*, 2008). Since clusters with a small number of genes (< 10) are difficult to annotate using gene set analysis, larger clusters are generally preferable; however, directly detecting clusters with a low sensitivity can miss important groups of co-expressed genes. For this reason, initial cluster detection was done with high sensitivity (deepSplit parameter = 3), and then clusters were merged if their dissimilarity score was less than 0.3 (Langfelder *et al.*, 2008) (Figure S2).

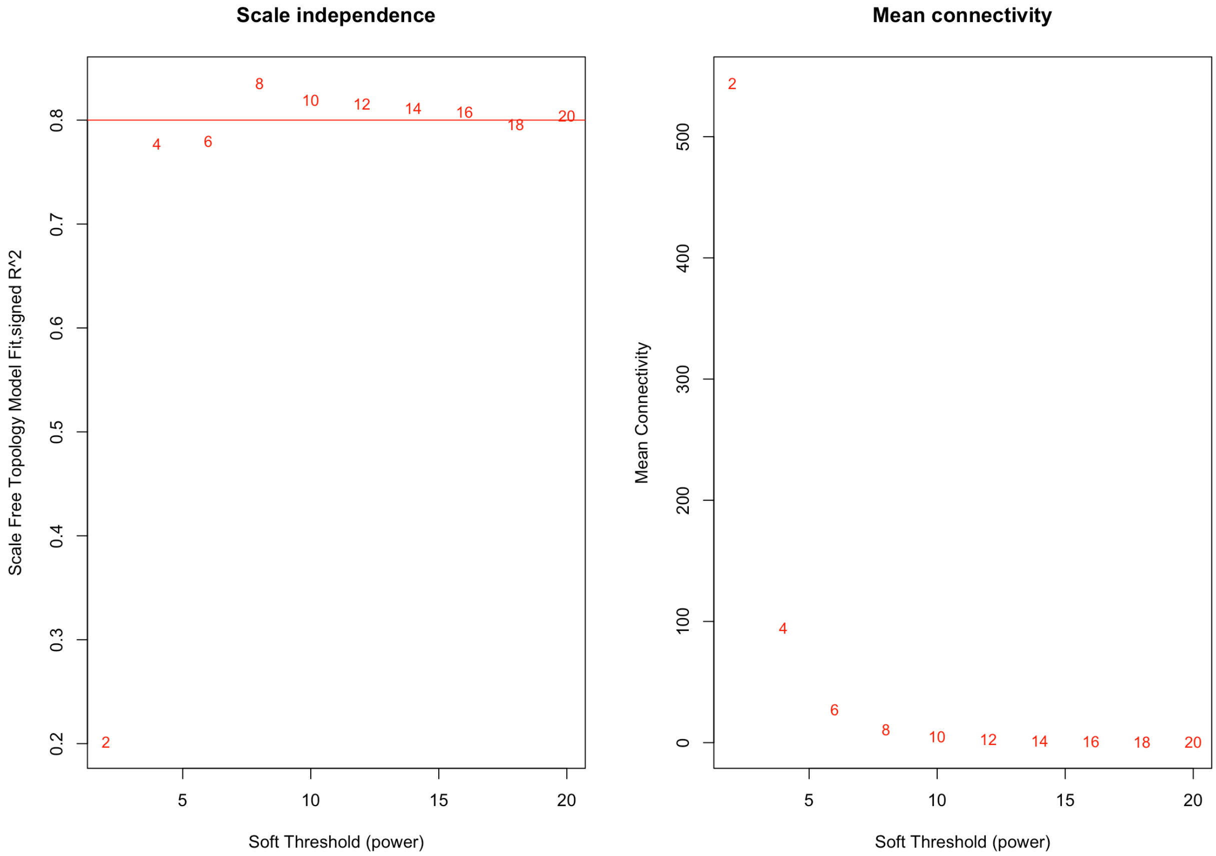


Figure S1: Elbow plots used to select the soft threshold parameter. The first plot displays the R^2 of fitting the scale free model to the adjacency network for different values of the soft threshold (β). The second plot displays the mean node connectivity (calculated as the mean sum of edges from each node) for different values of the soft threshold. The objective here is to use the two plots to maximize the scale free model fit and the mean connectivity simultaneously. Here, selecting β = 6 or 8 would be acceptable. We chose β = 8.

Supplemental information: merged clusters

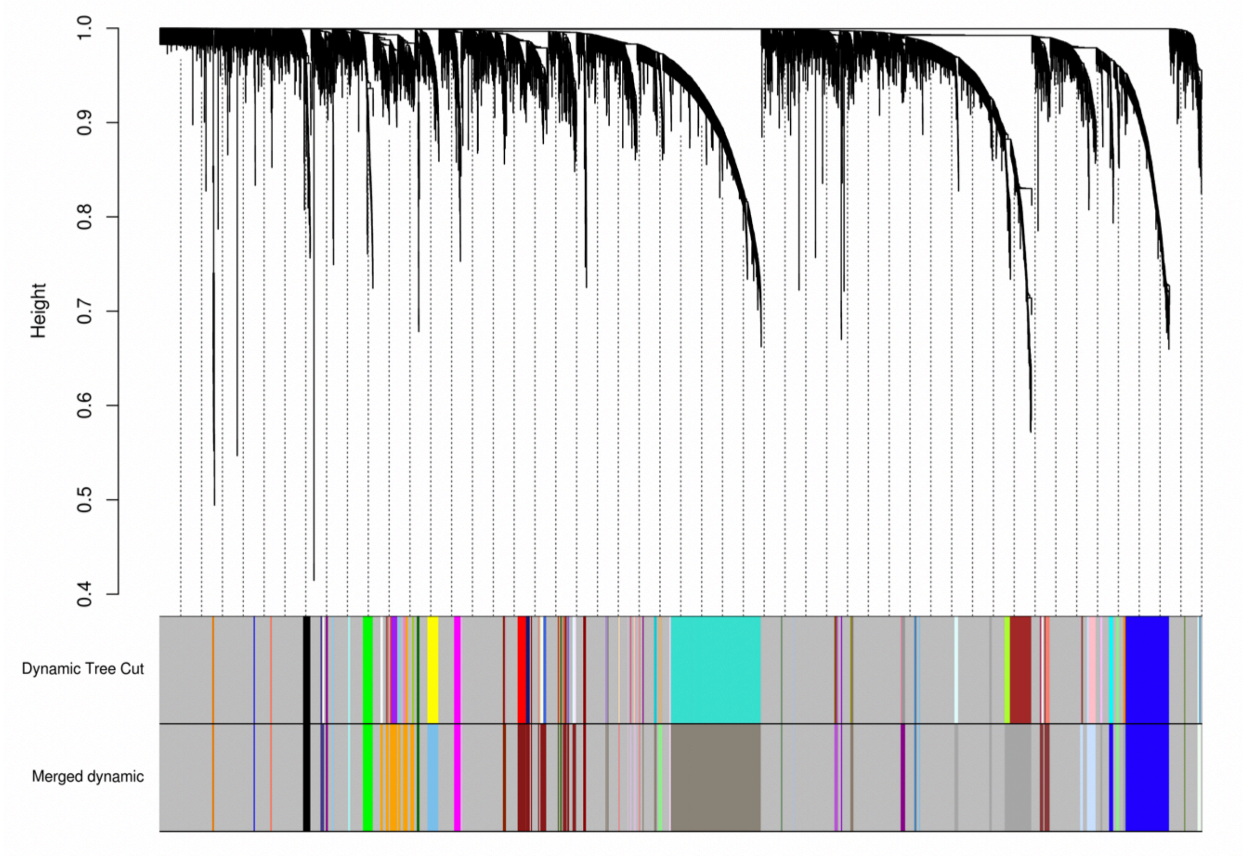


Figure S2: Dendrogram showing the location of the original and merged clusters. Each leaf in the dendrogram corresponds to one probe. Probes were hierarchically clustered based on their pairwise correlation scores. The first band shows the locations of the clusters detected with a deepSplit parameter of 3. Each cluster is given a distinct colour. The second band shows the locations of the clusters after being merged. Clusters were merged if the dissimilarity score between their eigengenes (computed using pairwise eigengene correlation) was < 0.3.

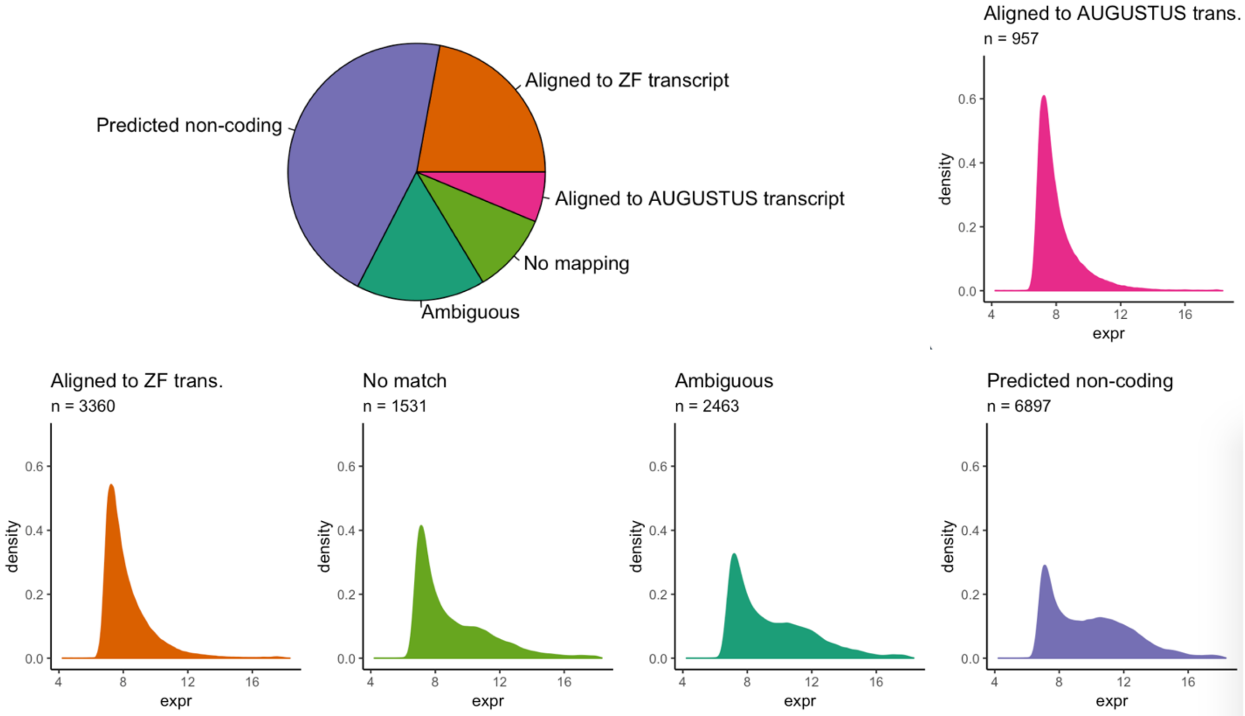


Figure S3: Results of the microarray probe re-annotation. Pie chart shows the proportions of microarray probes mapped to different locations on the fathead minnow reference genome using BLASTn. "Aligned to ZF transcript" were significantly (e-value < 10^{-6}) mapped to a transcript homologous to a coding transcript from the zebrafish genome, "aligned to AUGUSTUS transcript" were significantly mapped to a transcript predicted as a coding gene by the AUGUSTUS software and not mapped to a ZF transcript, "Predicted non-coding" were mapped to location with no coding transcript prediction, "no mapping" were not significantly mapped to any location, and "ambiguous" were significantly mapped to multiple locations. Density plots show the distribution of probe expression values across all n = 303 microarray samples for each re-annotation category.

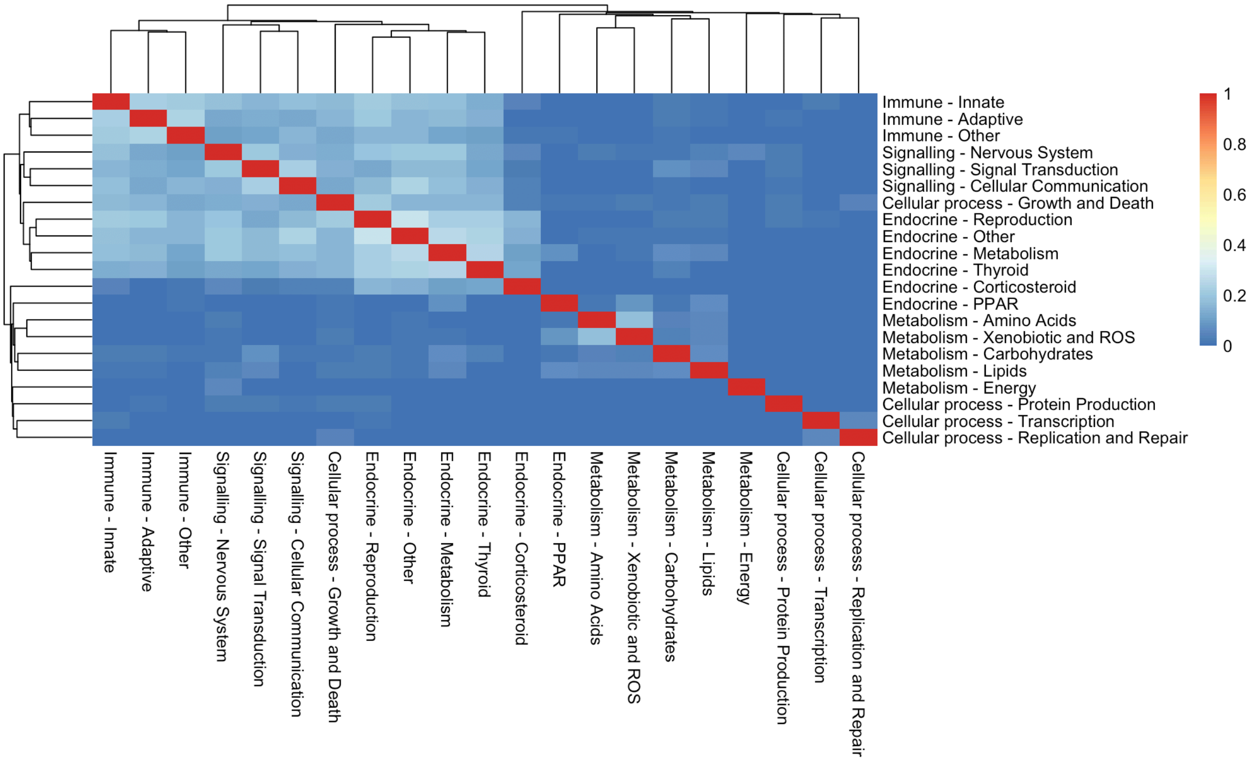


Figure S4: Pairwise overlap coefficient between functional EcoToxModules. Each square in the heatmap displays the Jaccard index between a pair of functional EcoToxModules. The Jaccard index is computed as the number of intersecting genes divided by the size of the union of the two gene sets.

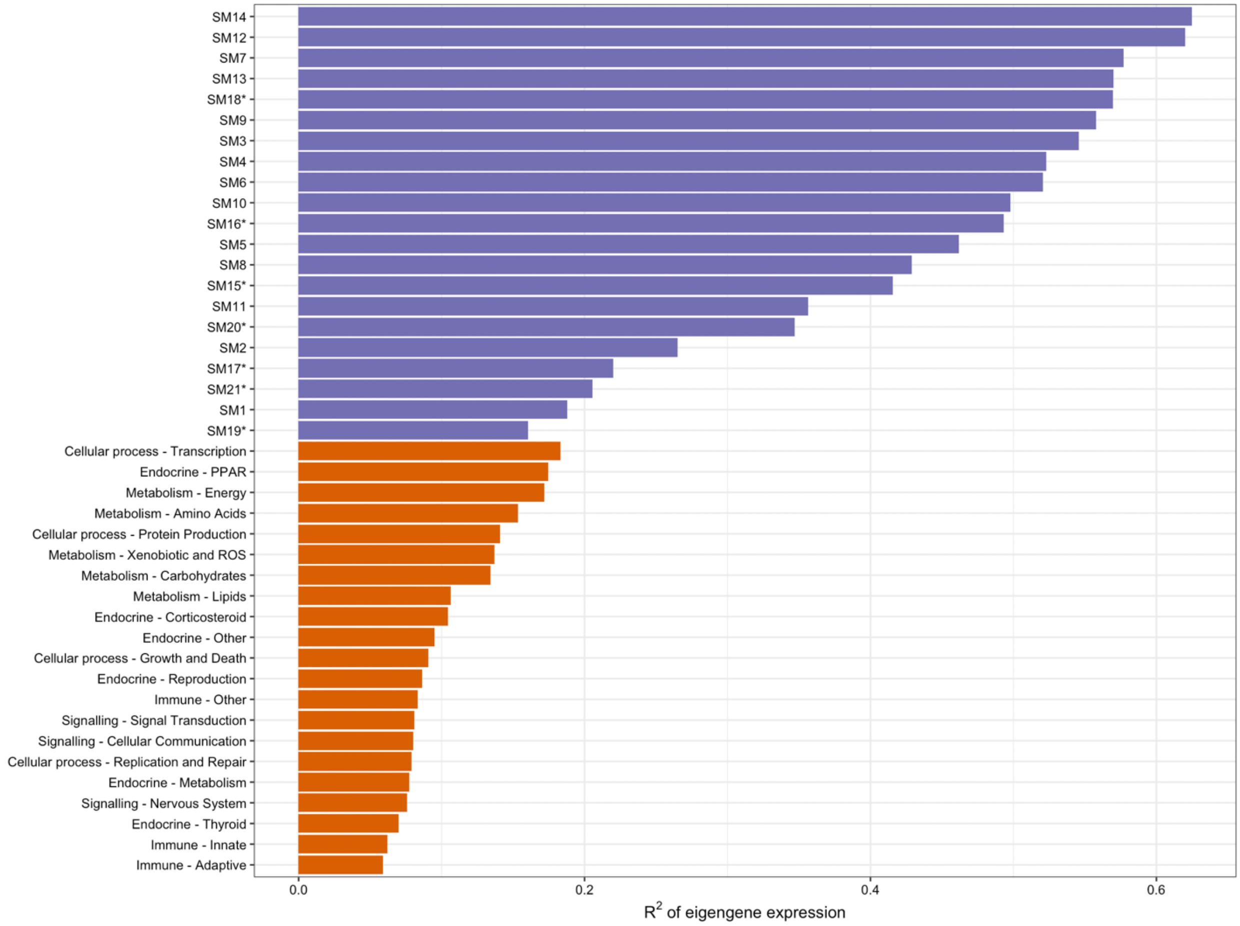


Figure S5: R^2 of statistical EcoToxModule eigengene expression. The eigengene expression is calculated as the 1st principal component of the probe fold changes with a module. Purple bars correspond to statistical EcoToxModules and orange corresponds to functional EcoToxModules. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR).

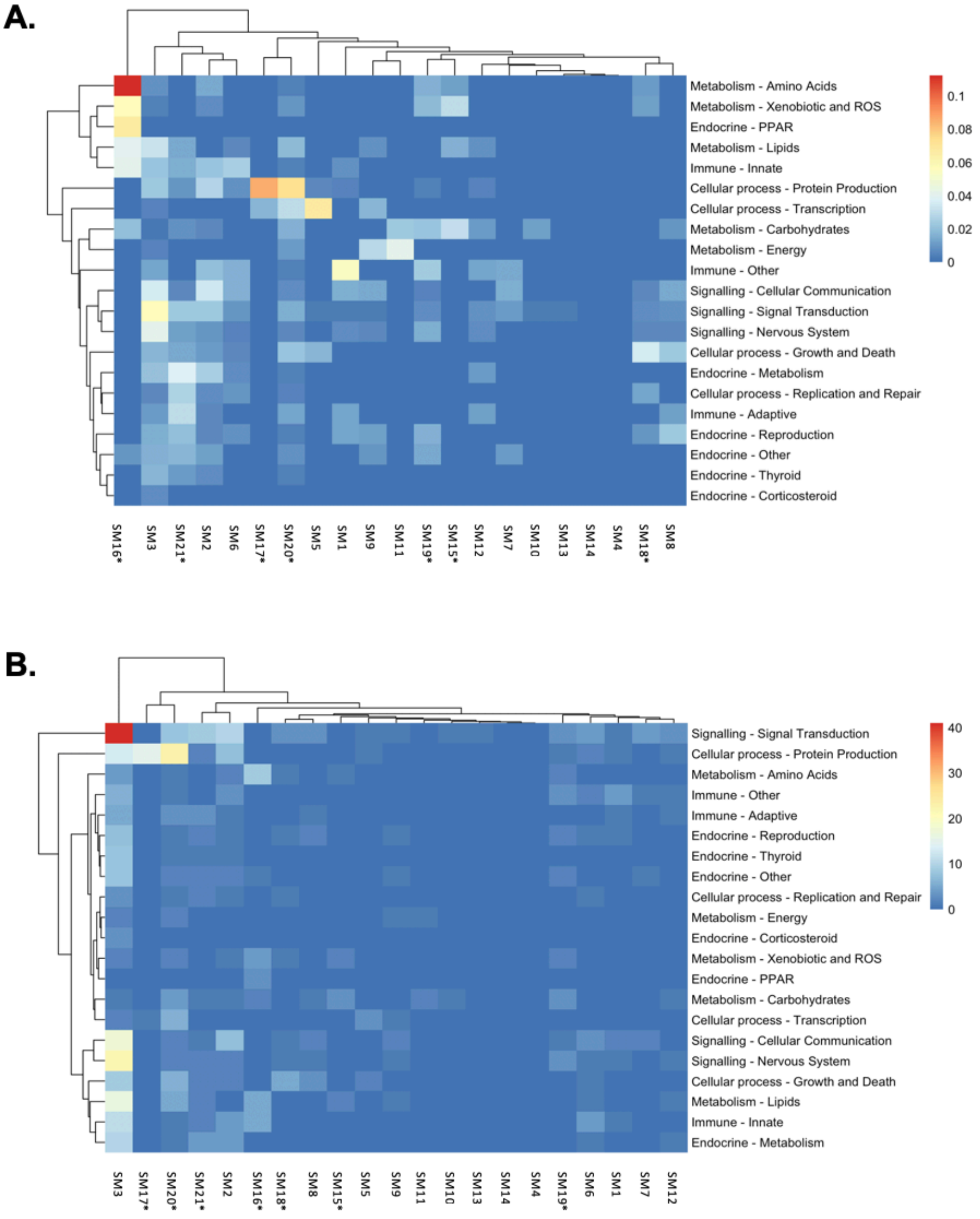


Figure S6: Similarity of statistical and functional EcoToxModules. The first heatmap displays the Jaccard index between pairs of one statistical and one functional EcoToxModule. The Jaccard index is computed as the number of intersecting genes divided by the size of the union of the two gene sets. The second heatmap displays the number of overlapping genes between pairs of one statistical and one functional EcoToxModule. Statistical EcoToxModules with an "*" have at least one significantly enriched KEGG pathway (adj. p-val < 0.05, FDR).

References

Bourgon, R.; Gentleman, R.; Huber, W., **2010**. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*. 107, (21), 9546-9551.

Edwards, D., **2003**. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics* 19, (7), 825-833.

Langfelder, P.; Zhang, B.; Horvath, S., **2007**. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, (5), 719-720.

Langfelder, P.; Horvath, S., **2008**. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, (1), 559.

Li, J.; Zhou, D.; Qiu, W.; Shi, Y.; Yang, J.-J.; Chen, S.; Wang, Q.; Pan, H., **2018**. Application of weighted gene co-expression network analysis for data from paired design. *Scientific Reports* 8, (1), 622.

Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K., **2015**. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43, (7), e47-e47.

Silver, J. D.; Ritchie, M. E.; Smyth, G. K., **2008**. Microarray background correction: maximum likelihood estimation for the normal–exponential convolution. *Biostatistics* 10, (2), 352-363.

Yip, A. M.; Horvath, S., **2007**. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 8, (1), 22.

SI for Chapter 5

SM Table 1: Comparison of FastBMD to popular dose-response software. The platform type impacts accessibility and flexibility, with web platforms being the most accessible and R being the most flexible. Speed is estimated relative to the other software, with one check representing the slowest and three being the fastest. See Figure 1 for a more detailed speed comparison with BMDExpress 2. "Batch analysis" refers to the ability to upload and process multiple transcriptomics dose-response matrices at one time.

Software	**FastBMD**	**BMDExpress 2**	**BMDS 3**	**PROAST**	**DROmics**
Platform	Web	Locally installed**	Locally installed***	R and Web	R and Web
Speed	✓✓✓	✓✓	✓*	R: ✓✓✓* Web: ✓*	R: ✓✓✓ Web: ✓
Number of continuous models	10	10	10	47	7
Interactive data visualization	Yes	Yes	No	No	No
Designed for transcriptomics	Yes	Yes	No	No	Yes
Supports batch analysis	No	Yes	No	No	No
Supports gene set analysis	Yes	Yes	NA	NA	No
Follows NTP approach	Yes	Yes	NA	NA	No
Citation	NA	(Phillips *et al.*, 2019)	(Davis *et al.*, 2011)	(Hardy *et al.*, 2017)	(Larras *et al.*, 2018)

*: This software is not purposefully designed for transcriptomics data, thus additional code must be written by the user to automate curve fitting for 100s – 1000s of genes. The time it would take to develop this additional code is not considered in the speed ranking.

**: BMDExpress 2.0 is compatible with Windows, Mac OSX, and Linux operating systems.

***: BMDS 3.0 is compatible with Windows operating systems. Microsoft Excel must also be installed.

SM Table 2: Species and annotation ID types supported by FastBMD. Specific supported microarray platforms can be found by visiting www.fastbmd.ca, choosing the organism of interest from the "Specify Organism" dropdown menu, and then consulting the "ID type" dropdown menu. Ens. G = Ensembl Gene; Ens. P = Ensembl Protein; Ens. T = Ensembl Transcript; OGS = Official Gene Symbol; ORF = Open Reading Frame.

Species	**Entrez**	**Refseq**	**Ens. G**	**Ens. P**	**Ens. T**	**OGS**	**Genbank**	**Uniprot**	**String**	**Other**
A. thaliana (Arabidopsis)	X	X	X	X	X	X	X	X		Tair
B. taurus (cow)	X	X	X	X	X	X	X	X		
C. elegans (roundworm)	X	X	X	X	X	X	X	X	X	Wormbase
C. japonica (Japanese quail)	X	X	X	X	X	X				
D. melanogaster (fruit fly)	X	X	X	X	X	X	X	X	X	Flybase
D. rerio (zebrafish)	X	X	X	X	X	X	X	X	X	1 microarray
G. gallus (chicken)	X	X	X	X	X	X				
H. sapiens (human)	X	X	X	X	X	X	X	X		22 microarrays
M. musculus (mouse)	X	X	X	X	X	X	X	X	X	19 microarrays
R. norvegicus (rat)	X	X	X	X	X	X	X	X		4 microarrays
S. cerevisae (yeast)	X	X	X	X	X	X		X	X	ORF identifiers
S. scrofa (pig)	X	X	X	X	X	X	X	X		
X. laevis (African clawed frog)	X	X								

Comparison between FastBMD and BMDExpress 2

Datasets

The performance of FastBMD was compared to that of BMDExpress 2 (Phillips *et al.*, 2019) by analyzing 24 previously published microarray dose-response datasets (Thomas *et al.*, 2013) with both software. The transcriptomic data were measured in adult rats (*Rattus norvegicus*) that were exposed to five doses of six chemicals for four different exposure durations (five days, two weeks, four weeks, thirteen weeks) with Affymetrix HT Rat230+ PM microarrays. This resulted in 24 distinct datasets, as described in SM Table 3. Data were downloaded from NCBI's Gene Expression Omnibus (accession = GSE45892). More details on the exposures can be found in the original publication (Thomas *et al.*, 2013).

SM Table 3: Description of dose-response experiments used to test FastBMD performance. Exposures that generated the previously published datasets used to compare the performance of FastBMD to BMDExpress. For dose units: mkd = mg per kg per day; ppm = parts per million.

CHEMICAL	DOSES	TISSUE
1,2,4-Tribromobenzene (TRBZ)	0, 2.5, 5, 10, 25, 75 mkd	Liver
Bromobenzene (BRBZ)	0, 25, 100, 200, 300, 400 mkd	Liver
2,3,4,6-Tetrachlorophenol (TTCP)	0, 10, 25, 50, 100, 200 mkd	Liver
4,4'-Methylenebis(*N,N*-dimethyl) benzenamine (MDMB)	0, 50, 200, 375, 500, 750 ppm	Thyroid
N-Nitrosodiphenylamine (NDPA)	0, 250, 1000, 2000, 3000, 4000 ppm	Bladder
Hydrazobenzene (HZBZ)	0, 5, 20, 80, 200, 300 ppm	Liver

Statistical analysis

Prior to dose-response analysis, each dataset was quantile normalized in R using the 'limma' R package (Ritchie *et al.,* 2015). FastBMD was run offline on the same computer and using the same computational resources as BMDExpress so that the elapsed time for each analysis could be directly compared between the two software. Within both software, each of the 24 datasets were filtered to remove any probe that did not have a fold-change of greater than two for any dose group. For the BMD analysis, all models except for the higher order polynomials (Exp2, Exp3, Exp4, Exp5, Linear, Poly2, Hill, and Power models) were fit to the expression values of each probe. These model fits were then used to calculate the geneBMDs and their 95% upper (geneBMD$_u$) and lower (geneBMD$_l$) confidence intervals.

For BMD analysis in BMDExpress, the following parameters were selected:

- Maximum Iterations: 250
- Confidence Level: 0.95
- Constant Variance: TRUE
- BMR Type: Standard Deviation
- BMR Factor: 1 SD
- Restrict Power: >= 1
- BMDL and BMDU: Compute but ignore non-convergence in best model selection
- Best Poly Model Test: Lowest AIC
- P-value Cutoff: 0.1
- Number of threads: 8
- Model Execution Timeout (secs): 30

For BMD analysis in FastBMD, the following parameters were selected:

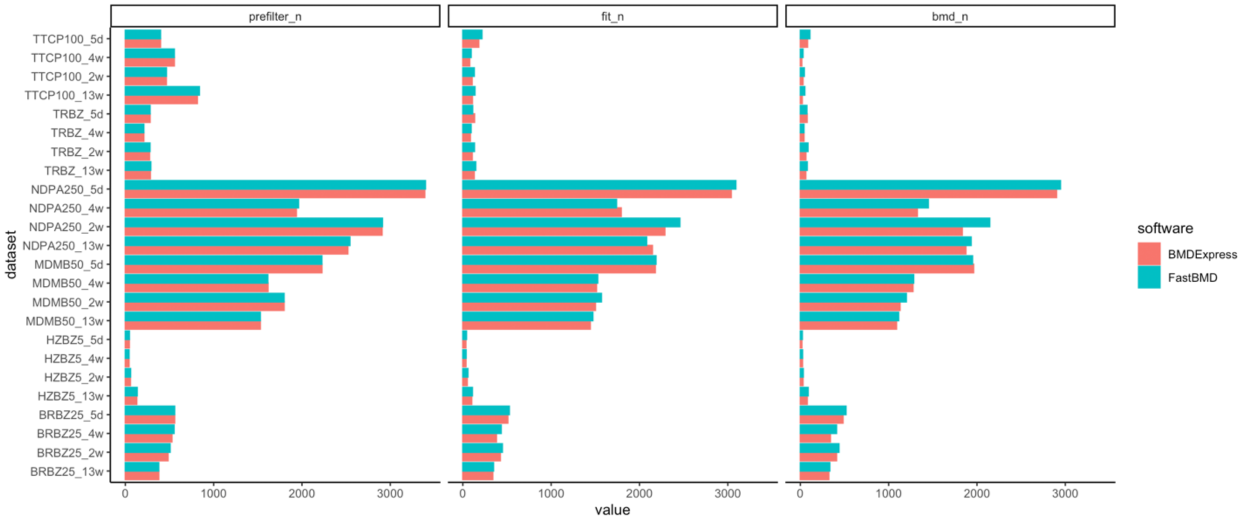
- Lack-of-fit p-value: 0.1
- BMR Factor: 1 SD
- Control expression: Evaluate model at 0

After gene-level BMD analysis, model fits and geneBMD results were downloaded from both software and imported into R. These results were filtered to remove any probes where the geneBMD was greater than the highest dose, or where the geneBMD$_u$/geneBMD$_l$ was greater than 40. The remaining geneBMDs were used to compute an omicBMD for each dose-response experiment using the "mode" method. Curve fitting results were compared between FastBMD and BMDExpress based on the time to compute model fits and geneBMDs, number of probes that passed each filtering step, omicBMD values, geneBMD values, distribution of best fit models, and model fit quality. Model fit quality was assessed using Akaike's Information Criterion (AIC), which is a modified measure of a model's prediction error that applies increasingly large penalties for more complex models. Thus, a model that has a smaller prediction error, but many parameters, may have a worse (higher) AIC score than a simpler model that has a slightly larger prediction error.

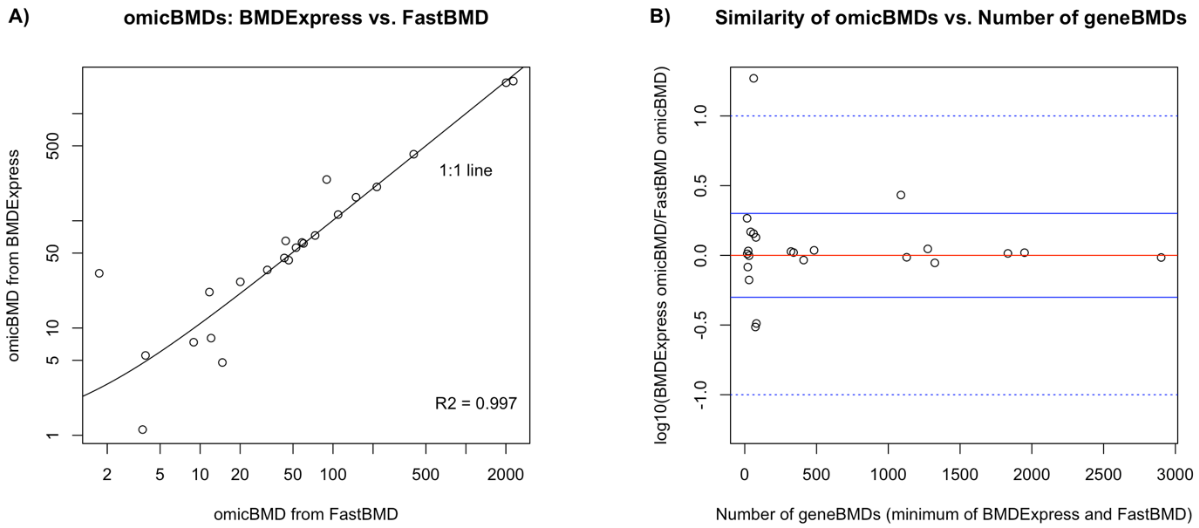
Comparing AIC scores across BMDExpress and FastBMD models was challenging as the AIC values returned by BMDExpress and FastBMD were different even when the model had the exact same coefficients. Thus, AIC values had to be re-computed in R using the same method for both software, however there is no straightforward method to manually create model fit objects in R with previously defined coefficients. To overcome this, we re-ran "*nls*", a non-linear model fitting algorithm in base R, for each model fit. Using the previously found coefficients as starting values and restricting the number of iterations forced "*nls*" to converge on the coefficients from FastBMD and BMDExpress, returning model fit objects that could be used to compute the AIC.

Results and Discussion

Some datasets exhibited a much stronger dose-dependent response to chemical exposure than others, resulting in different numbers of probes that passed the fold change, curve fitting, and BMD filters across the 24 datasets (SM Figure 1). The number of geneBMDs computed for each dataset ranged from a minimum of 16 (TTCP – 4 weeks, BMDExpress) to a maximum of 2945 (NDPA – 5 days, FastBMD). In general, the omicBMDs from both software were similar to each other, with an R^2 of 0.997 (SM Figure 2a). OmicBMDs were more variable between the two software when there were smaller numbers of geneBMDs (Figure 2b). This makes sense as an omicBMD that is computed from 10s of geneBMDs likely has more associated uncertainty than one that is computed from 100s or 1000s of geneBMDs.

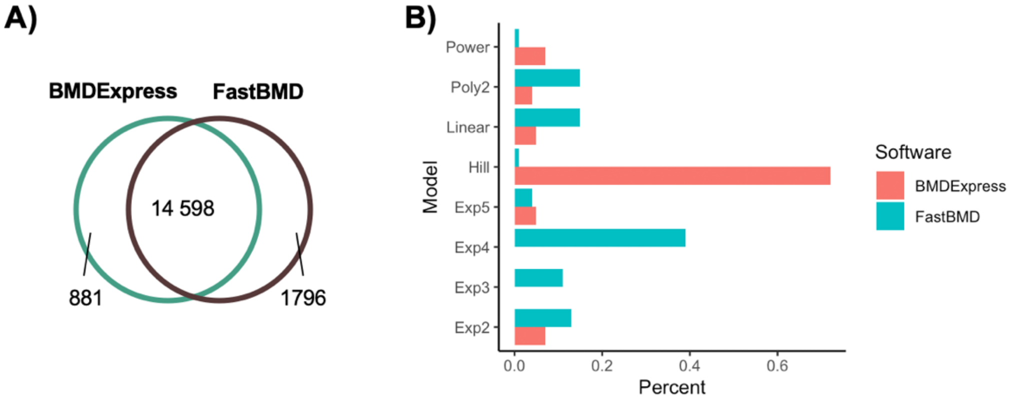


SM Figure 1: Number of probes that passed each filter. Bars show the number of probes that passed each of the filtering steps. (blue = FastBMD; pink = BMDExpress; prefilter_n = after fold-change filter; fit_n = after lack-of-fit p-value filter; bmd_n = after BMD quality filters).



SM Figure 2: Comparison of the omicBMDs from FastBMD and BMDExpress. A) omicBMDs computed with the "mode" method using both FastBMD and BMDExpress for the datasets summarized in SM Table 2. B) log10 of the ratio of BMDExpress:FastBMD omicBMDs plotted against the minimum number of geneBMDs used to compute the omicBMD from both software. Given that the y-axis is on a log10 scale, the solid red line indicates a fold-change of 1 (no change), the solid blue line is a fold-change of 2, and the dotted blue line is a fold-change of 10.

Across all 24 experiments, BMDExpress and FastBMD found 15 479 and 16 394 geneBMDs respectively. The majority of the fits were for the same probes (*n* = 14 598), however BMDExpress and FastBMD both found model fits that passed all of the quality criteria for some probes (BMDExpress: *n* = 881; FastBMD: *n* = 1796) that the other software did not find (SM Figure 3a). For the models fit to the unique probes in particular, there was a different distribution of best-fitting model types, with BMDExpress tending to find more Hill and Power model fits and FastBMD finding more Exponential and Polynomial fits (SM Figure 3b). This shows that while neither curve fitting algorithm is the best choice in all scenarios, the FastBMD algorithm does find high-quality model fits more often than the BMDExpress algorithm does.

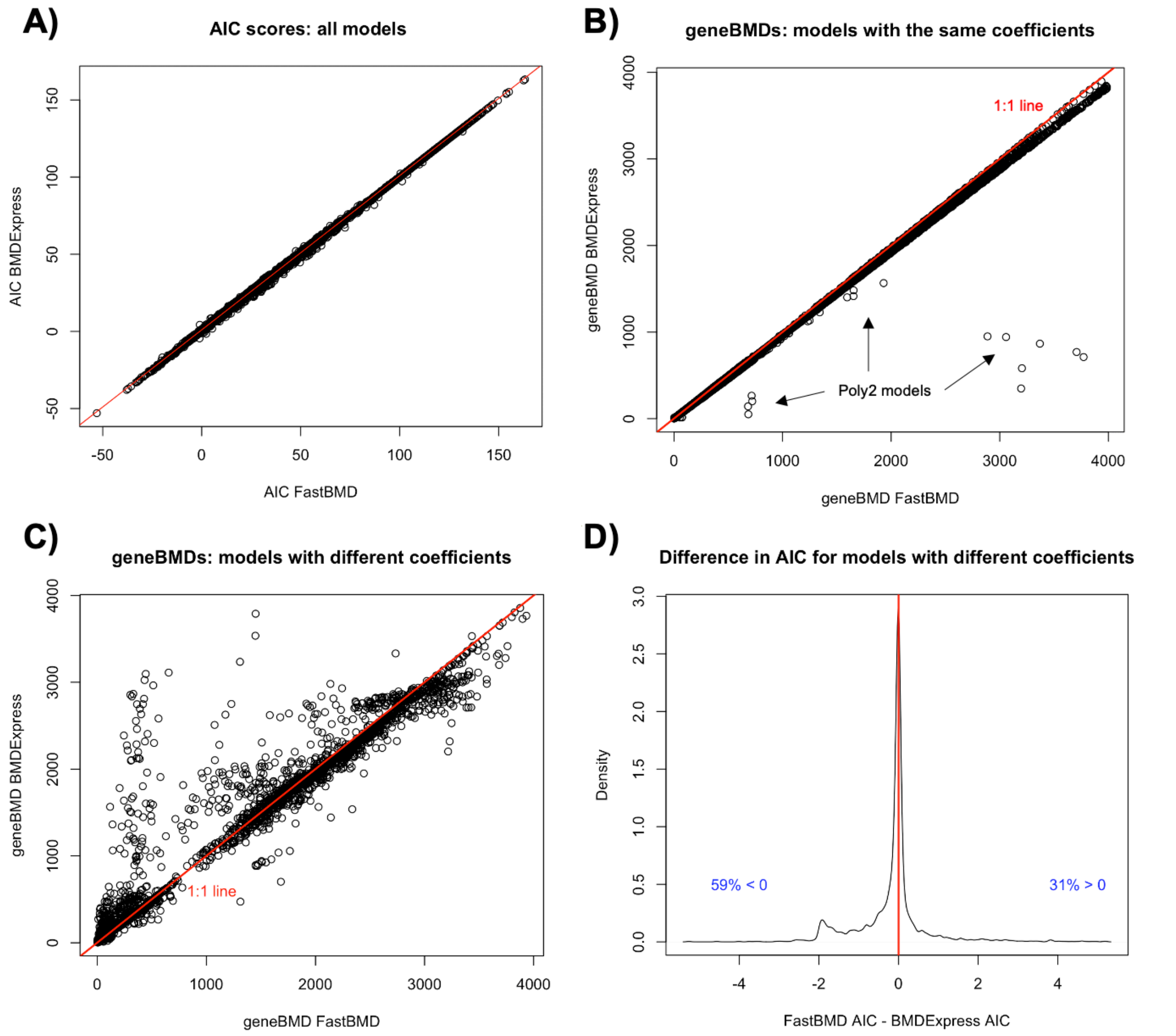


SM Figure 3: Overlapped and unique probes with geneBMDs. A) The overlap of probes with geneBMDs that passed all quality filters from BMDExpress and FastBMD. B) Percent of best fit models of each type that had a geneBMD from one software but not the other (*n* = 881 for BMDExpress; *n* = 1796 for FastBMD).

The AIC scores were re-computed for the models that were fit to the same probes by both software. Overall, AIC scores were able to be re-computed for both the FastBMD and BMDExpress model fits for 13 224 out of the 14 598 shared probes (90.6%). AIC scores were unable to be computed for all genes since forcing 'nls' to rerun with certain restrictions caused errors for some data, while loosening the restrictions had the effect of allowing 'nls' to find new model parameters. Errors were thrown by both FastBMD and BMDExpress models, so we are working under the assumption that the re-computed AIC values for 90.6% of probes are representative of the full set of models.

The re-computed AIC values were very similar across all probes (SM Figure 4a). Out of the 13 224 probes with AIC values for both software, 10 080 had the exact same individual parameters and model types. They had nearly identical BMDs, except for 18 of the Poly2 fits (SM Figure 4b). This can be explained by how each software computes the benchmark response (BMR). The benchmark response (BMR) is found by evaluating the fitted model at zero, and then either adding or subtracting the standard deviation of the residuals depending on whether the adverse direction is positive or negative. In FastBMD, the adverse direction is determined by fitting a linear model to the expression values for each probe and seeing if the slope coefficient is positive or negative. Since Poly2 curves can change direction, it's possible for the overall slope to be in the opposite direction as when the curve first surpasses one standard deviation of the residuals. Thus, on rare occasions (1.4% of Poly2 fits; 18/1282 occurrences), BMDExpress and FastBMD can return substantially different gene-level BMDs from the same Poly2 fits.

Probes with different models (*n* = 3144) produced geneBMDs that were less consistent across the two software (SM Figure 4c). The vast majority of these probes (99%) had different best-fitting model types, which then returned less similar geneBMDs for the same probe expression values. To determine which software tended to find models with higher quality fits, the differences between FastBMD and BMDExpress AIC scores were visualized with a density plot (SM Figure 4d). The differences were negative 59% of the time, indicating that when the software found different models, FastBMD found higher quality fits more often than BMDExpress, although neither algorithm performed the best in all scenarios.



SM Figure 4: Probe-level results from FastBMD and BMDExpress. A) AIC scores for models returned by FastBMD and BMDExpress (*n* = 13 224). B) geneBMD values for models with the same type and coefficients from FastBMD and BMDExpress that also have AIC scores (*n* = 10 080). Upon investigation, all noticeable outliers are from Poly2 model fits (*n* = 18 Poly2 outliers). C) geneBMD vaues for models with different coefficients and/or types (*n* = 3144). D) Distribution of differences between FastBMD and BMDExpress AIC values from models with different coefficients and/or types (*n* = 3144). Negative and positive differences indicate that the FastBMD and BMDExpress models had a higher quality fit, respectively.

The majority of the variation in geneBMDs appears to be from scenarios where BMDExpress and FastBMD find different best-fitting model types (SM Figure 4c), even though these model fits still had very similar AIC values (SM Figure 4a). It is concerning that two models with almost the same quality fit can produce substantially different BMDs; this may be an inherent limitation to using parametric models for BMD analysis. When there were differences in the best-fitting model type, FastBMD tended to find higher quality model fits more often than BMDExpress (SM Figure 3a, SM Figure 4d), although BMDExpress does perform slightly better at computing BMDs from Poly2 fits.

BMDExpress and FastBMD both implement the modeling philosophy that is outlined in the NTP approach to transcriptomic dose-response modeling. Even though the NTP recommendations are quite detailed, there are still practical design choices that must be made when implementing the statistical workflow, for example which non-linear parameter search algorithm to choose or how to break ties between model fit AIC scores. For many of these decisions, there isn't a clear choice that performs best in all scenarios. However, despite small implementation differences like these, FastBMD and BMDExpress returned the exact same best-fit model for 76% of fitted models. When considering both the number of times that both software did not find any model and when the software returned the same model across all 24 datasets, FastBMD and BMDExpress produced the same results for >99% of the probes.

References

Phillips JR, Svoboda DL, Tandon A, Patel S, Sedykh A, Mav D, Kuo B, Yauk CL, Yang L, Thomas RS. **2019**. BMDExpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics* 35:1780-1782.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. **2015**. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43:e47-e47.

Thomas RS, Wesselkamper SC, Wang NCY, Zhao QJ, Petersen DD, Lambert JC, Cote I, Yang L, Healy E, Black MB, Clewell HJ, Allen BC, Andersen ME. **2013**. Temporal concordance between apical and transcriptional points of departure for chemical risk assessment. *Toxicological Sciences* 134.1: 180-194.

Allen DJ, Gift JS, and Zhao QJ. **2011**. "Introduction to benchmark dose methods and US EPA's benchmark dose software (BMDS) version 2.1.1." *Toxicology and Applied Pharmacology* 254.2: 181-191.

Larras F, Billoir E, Baillard V, Siberchicot A, Scholz S, Wubet T, Tarkka M, Schmitt-Jansen M, Delignette-Muller M. **2018**. DRomics: A turnkey tool to support the use of the dose–response framework for Omics data in ecological risk assessment. *Environmental Science & Technology* 52.24: 14461-14468.

EFSA Scientific Committees, Hardy A, Benford D, Halldorsson D, Jeger MJ, Knutsen KH, More S, Mortensen A, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Silano V, Solecki R, Turck D, Aerts M, Bodin L, Davis A, Edler L, Gundert-Remy U, Sand S, Slob W, Bottex B, Abrahantes JC, Marques DC, Kass G, Schlatter JR. **2017**. Update: use of the benchmark dose approach in risk assessment. *EFSA Journal*.

SI FOR CHAPTER 6

Supporting Information 1. Raw RNAseq processing instructions.

For ‘raw RNAseq processing’, users can click the appropriate button on the main page, which will permit the analysis of their raw data through the Galaxy for Raw Data section of EcoToxXplorer [https://galaxy.ecotoxxplorer.ca; which uses our in-house customized Galaxy server (Afgan *et al*. 2018)]. After completing registration, users can upload their data (fastq.gz, which can be seen under the shared data section), and then choose the ultra-fast pseudoalignment-based method (Kallisto) or the classical spliced aligner (HISAT2) to map raw RNAseq data against different species. In the Kallisto workflow, the raw data are subjected to Trim Galore (Galaxy version 0.4.3.1) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove adapter sequences. After this step, trimmed data are mapped to a transcriptome index and read counts for each sample are quickly estimated through the pseudoalignment approach of Kallisto (Bray *et al*., 2016). In the HISAT2 workflow, the raw data are subjected to Trim Galore as above-mentioned and trimmed data are mapped to the defined genome with the related annotation file (GTF or GFF) by HISAT2 (Kim *et al*., 2015) (Galaxy version 2.1.0). Read counts for each sample are obtained using HTSeq with the intersection-strict mode (Galaxy version 0.9.1) (Anders, Pyl, & Huber, 2015). At the end of both workflows, the read counts for all the processed samples are provided in a single tab file. Further details are described in a tutorial in the Galaxy for Raw Data section (https://galaxy.ecotoxxplorer.ca), which also allows users to check the data quality with FASTQC (Galaxy version 0.72) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

References Cited:

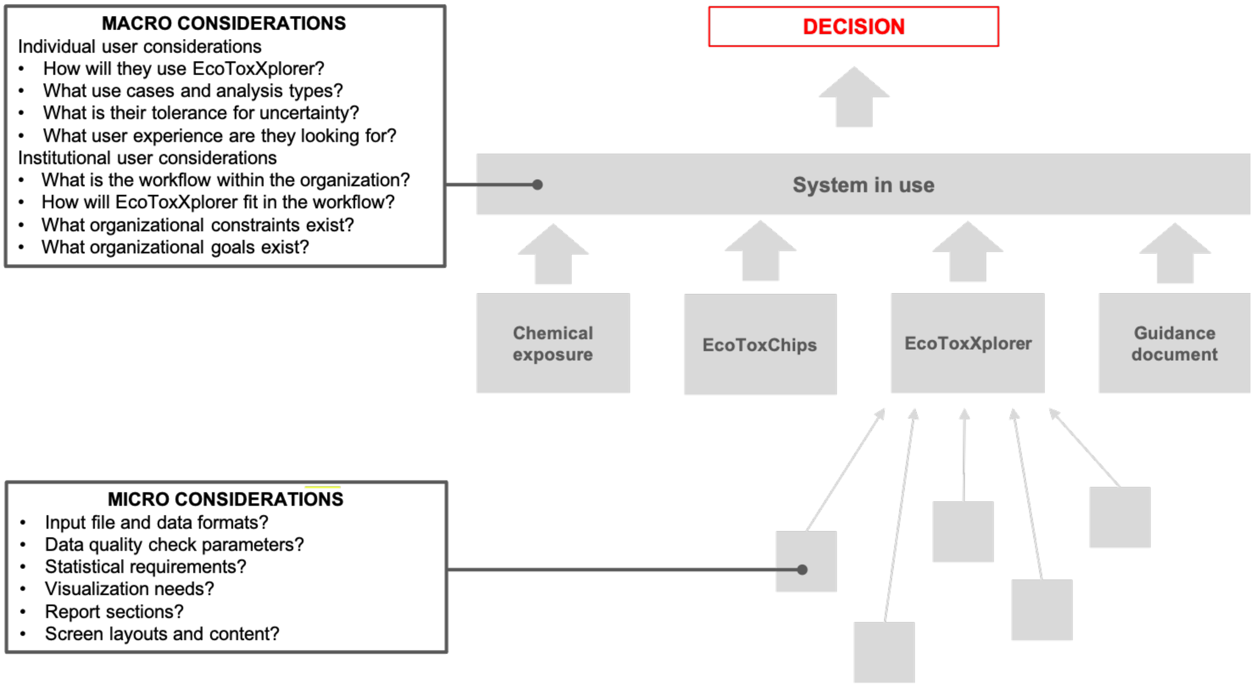
Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. **2018**. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*. 46(W1):W537-W544. doi: 10.1093/nar/gky379.

Anders S, Pyl PT, Huber W. **2015**. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31(2):166-9. doi: 10.1093/bioinformatics/btu638.

Bray, N., Pimentel, H., Melsted, P. *et al*. **2016**. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34, 525–527. https://doi.org/10.1038/nbt.3519

Kim, D., Langmead, B., Salzberg, S. **2015**. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 12, 357–360. https://doi.org/10.1038/nmeth.3317

Supporting Information 2. Representative examples of micro and macro features of EcoToxXplorer.ca identified through design thinking activities.



MACRO CONSIDERATIONS

Individual user considerations

- How will they use EcoToxXplorer?
- What use cases and analysis types?
- What is their tolerance for uncertainty?
- What user experience are they looking for?

Institutional user considerations

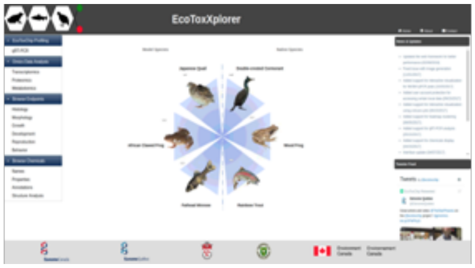
- What is the workflow within the organization?
- How will EcoToxXplorer fit in the workflow?
- What organizational constraints exist?
- What organizational goals exist?

MICRO CONSIDERATIONS

- Input file and data formats?
- Data quality check parameters?
- Statistical requirements?
- Visualization needs?
- Report sections?
- Screen layouts and content?

<u>Supporting Information 3.</u> Evolution of the EcoToxXplorer homepage.

Circa 2017:



EcoToxXplorer

Japanese Quail

Double-crested Cormorant

African Clawed Frog

Wood Frog

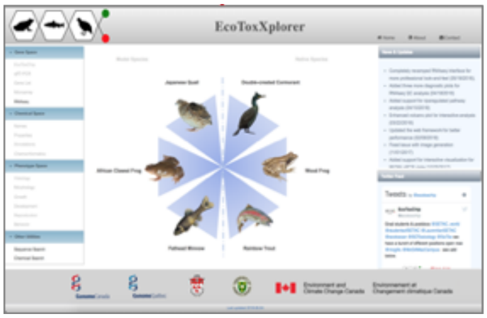
Fathead Minnow

Rainbow Trout

Environment Canada

Environnement Canada

Circa 2018:



EcoToxXplorer

Japanese Quail

Double-crested Cormorant

African Clawed Frog

Wood Frog

Fathead Minnow

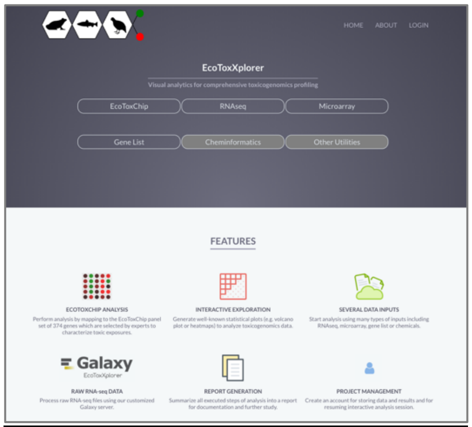
Rainbow Trout

Tweets

Environment and Climate Change Canada

Environnement et Changement climatique Canada

Circa 2019:



HOME ABOUT LOGIN

EcoToxXplorer

Visual analytics for comprehensive toxicogenomics profiling

EcoToxChip | RNAseq | Microarray

Gene List | Cheminformatics | Other Utilities

FEATURES

ECOTOXCHIP ANALYSIS
Perform analysis by mapping to the EcoToxChip panel set of 374 genes which are selected by experts to characterize toxic exposures.

INTERACTIVE EXPLORATION
Generate well-known statistical plots (e.g. volcano plot or heatmaps) to analyze toxicogenomics data.

SEVERAL DATA INPUTS
Start analysis using many types of inputs including RNAseq, microarray, gene list or chemicals.

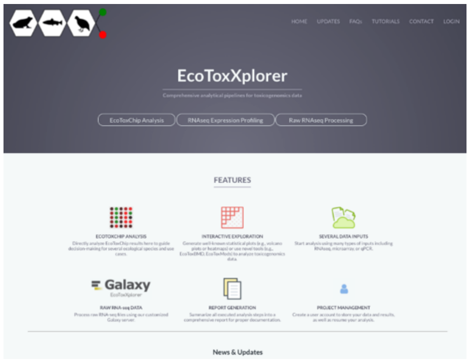
Galaxy EcoToxXplorer

RAW RNA-seq DATA
Process raw RNA-seq files using our customized Galaxy server.

REPORT GENERATION
Summarize all executed steps of analysis into a report for documentation and further study.

PROJECT MANAGEMENT
Create an account for storing data and results and for resuming interactive analysis session.

Current (since 2020):



HOME UPDATES FAQs TUTORIALS CONTACT LOGIN

EcoToxXplorer

Comprehensive analytical pipelines for toxicogenomics data

EcoToxChip Analysis | RNAseq Expression Profiling | Raw RNAseq Processing

FEATURES

ECOTOXCHIP ANALYSIS
Directly analyze EcoToxChip results here to guide decision-making for several ecological species and use cases.

INTERACTIVE EXPLORATION
Generate well-known statistical plots (e.g., volcano plots or heatmaps) or use novel tools (e.g., EcoToxBMD, EcoToxMods) to analyze toxicogenomics data.

SEVERAL DATA INPUTS
Start analysis using many types of inputs including RNAseq, microarray, or qPCR.

Galaxy EcoToxXplorer

RAW RNA-seq DATA
Process raw RNA-seq files using our customized Galaxy server.

REPORT GENERATION
Summarize all executed analysis steps into a comprehensive report for proper documentation.

PROJECT MANAGEMENT
Create a user account to store your data and results, as well as resume your analysis.

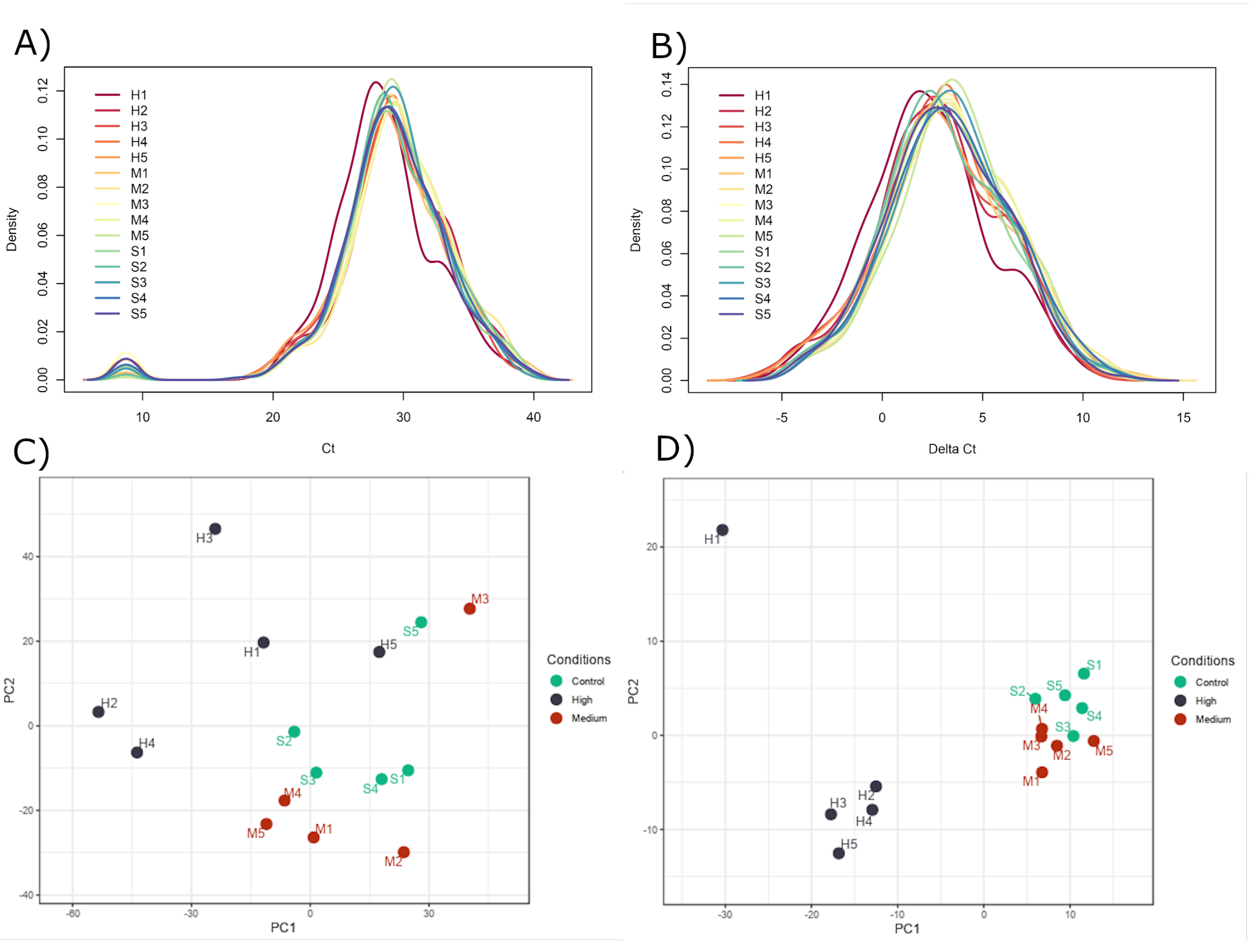
News & Updates

Supporting Information 4. Summary of current species annotation libraries in EcoToxXplorer.

Note, these will be updated over time with additional species and information to be added.

Species	Data type	ID type	Functional Annotation
Coturnix japonica (JQ; Japanese quail)	qPCR, EcoToxChip (v0.1, v1), RNA-seq, microarray	Ensembl gene ID, Entrez gene ID, Official gene symbol	KEGG (pathways, module, process), EcoTox (module, process), Gene ontologies
Pimephales promelas (FHM; Fathead minnow)	qPCR, EcoToxChip (v0.1, v1), RNA-seq, microarray	Ensembl gene ID, Entrez gene ID, Official gene symbol, RefSeq gene ID, Ensemble Transcript ID, Ensemble Protein ID, Uniprot protein ID, Agilent-019597 FHM Denslow (8x15K)	KEGG (pathways, module, process), EcoTox (module, process), Gene ontologies
Xenopus laevis (XL; African clawed frog)	qPCR, EcoToxChip (v0.1), RNA-seq, microarray	Ensembl gene ID, Entrez gene ID, Official gene symbol	KEGG (pathways, module, process), EcoTox (module, process), Gene ontologies
Oncorhynchus mykiss (RT; Rainbow trout)	RNA-seq, microarray	Ensembl gene ID, Entrez gene ID, Official gene symbol	KEGG (pathways), Gene ontologies

Supporting Information 5. Effects of normalization on density (raw plot A to normalized plot B) and PCA (raw plot C to normalized plot D).


A)
B)
C)
D)
Density
Ct
Delta Ct
H1
H2
H3
H4
H5
M1
M2
M3
M4
M5
S1
S2
S3
S4
S5
PC1
PC2
Conditions
Control
High
Medium