Masthead Logo

University of Iowa
Iowa Research Online

Theses and Dissertations

Spring 2018

Exploring and analyzing omics using bioinformatics tools and techniques

Mrutyunjaya Parida
University of Iowa

Copyright © 2018 Mrutyunjaya Parida

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/6244

Recommended Citation
Parida, Mrutyunjaya. "Exploring and analyzing omics using bioinformatics tools and techniques." PhD (Doctor of Philosophy) thesis, University of Iowa, 2018.
https://doi.org/10.17077/etd.854wv5hc

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Bioinformatics Commons

EXPLORING AND ANALYZING OMICS USING BIOINFORMATICS TOOLS AND TECHNIQUES

by

Mrutyunjaya Parida

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in
Informatics (Health Informatics)
in the Graduate College of
The University of Iowa

May 2018

Thesis Supervisor: Associate Professor J. Robert Manak

Copyright by

MRUTYUNJAYA PARIDA

2018

All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Mrutyunjaya Parida

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Informatics (Health Informatics) at the May 2018 graduation.

Thesis Committee:

J. Robert Manak, Thesis Supervisor

Josep Comeron

Prakash Nadkarni

Benjamin W. Darbro

Nick Street

ACKNOWLEDGEMENTS

The work in this thesis is a result of collaborative effort from my committee members, collaborators, lab members and most importantly my research advisor, Professor J Robert Manak. When I first interviewed with John in August 2011, I was moved by his passion towards science. He showed great excitement and pride when describing the ongoing projects in his lab. It was his remarkable intelligence and scientific expertise that made me join his lab. I am utterly grateful that he accepted to train me as a Ph.D. student. John taught me about multitasking, time management, presentation skills, logical thinking, fundamentals of biology and genomics, sequencing data analysis, visual inspection and verification of data analysis, and above all doing good science. John has always given me time from his busy schedule to help me with my project related questions and concerns. His comments and insights during our weekly lab meetings have extremely benefited me. I deeply appreciate and support his motto “put science before you”. He challenged my mind by training me on a variety of projects and have always trusted my ability to work on them. Additionally, he imparted me with his wisdom and rigorous but smart scientific work culture for which I am truly thankful. I strongly believe it has made me scientifically skilled, strong and ethical and will continue to benefit me in my future endeavors.

I am sincerely thankful to all my committee members, Professor Josep Comeron, Dr. Prakash Nadkarni, Dr. Benjamin W. Darbro, and Professor Nick Street, for their support, guidance and training that helped me to complete my Ph.D. study. I learnt a variety of skills from them such as probability and statistics from Dr. Comeron, clinical informatics from Dr. Darbro, heavy metal biology from Dr. NadkarNi^{2+} and data mining from Dr. Street. Additionally, I convey my gratitude to all our current and previous lab members namely, Juan Santana, Lisa Lansdon, Xiaojing Hong, Anthony Lilienthal, and Salleh Ehaideb. They have helped me in my projects by taking time off from theirs, made valuable suggestions to my projects and presentations, encouraged my work and always had my back, especially Juan Santana and Lisa Lansdon. I am very honored for this opportunity to train at the University of Iowa, under a marvelous advisor, work under an exceptional committee, and be surrounded by intelligent and good-hearted lab mates who have been nothing but kind, giving and have enriched my scientific knowledge base.

My Ph.D. is dedicated to my father, mother, sister and nephew who have always believed in my abilities to achieve my goals. I am proud to represent them here at the UIOWA and become the first Ph.D. in my immediate and associative family.

ABSTRACT

During the Human Genome Project the first hundred billion bases were sequenced in four years, however, the second hundred billion bases were sequenced in four months (NHGRI, 2013). As efforts were made to improve every aspect of sequencing in this project, cost became inversely proportional to the speed (NHGRI, 2013). Human Genome Project ended in April 2003 but research in faster and cheaper ways to sequence the DNA is active to date (NHGRI, 2013). On the one hand, these advancements have allowed the convenient and unbiased generation and interrogation of a variety of omics datasets; on the other hand, they have substantially contributed towards the ever-increasing size of biological data. Therefore, informatics techniques are indispensable tools in the field of biology and medicine due to their ability to efficiently store and probe large datasets. Bioinformatics is a specialized domain under informatics that focusses on biological data storage, organization and analysis (NHGRI, 2013). Here, I have applied informatics approaches such as database designing and web development in the context of biological datasets or bioinformatics, to create a novel web-based resource that allows users to explore the comprehensive transcriptome of common aquatic tunicate named *Oikopleura dioica* (*O .dioica*), and access their associated annotations across key developmental time points, conveniently. This unique resource will substantially contribute towards studies on development, evolution and genetics of chordates using *O. dioica* as a model.

Mendelian or single-gene disorders such as cystic fibrosis, sickle-cell anemia, Huntington’s disease, and Rett’s syndrome run across generations in families (Chial, 2008). Allelic variations associated with Mendelian disorders primarily reside in the

protein-coding regions of the genome, collectively called an exome (Stenson et al., 2009). Therefore, sequencing of exome rather than whole genome is an efficient and practical approach to discover etiologic variants in our genome (Bamshad et al., 2011). Renal agenesis (RA) is a severe form of congenital anomalies of the kidney and urinary tract (CAKUT) where children are born with one (unilateral renal agenesis) or no kidneys (bilateral renal agenesis) (Brophy et al., 2017; Yalavarthy & Parikh, 2003). In this study, we have applied exome-sequencing technique to selective human patients in a renal agenesis (RA) pedigree that followed a Mendelian mode of disease transmission. Exome sequencing and molecular techniques combined with my bioinformatics analysis has led to the discovery of a novel RA gene called *GREB1L* (Brophy et al., 2017). In this study, we have successfully demonstrated the validation of exome sequencing and bioinformatics techniques to narrow down disease-associated mutations in human genome. Additionally, the results from this study has substantially contributed towards understanding the molecular basis of CAKUT. Discovery of novel etiologic variants will enhance our understanding of human diseases and development.

High-throughput sequencing technique called RNA-Seq has revolutionized the field of transcriptome analysis (Z. Wang, Gerstein, & Snyder, 2009). Concisely, a library of cDNA is prepared from a RNA sample using an enzyme called reverse transcriptase (Nottingham et al., 2016). Next, the cDNA is fragmented, sequenced using a sequencing platform of choice and mapped to a reference genome, assembled transcriptome, or assembled *de novo* to generate a transcriptome (Grabherr et al., 2011; Nottingham et al., 2016). Mapping allows detection of high-resolution transcript boundaries, quantification of transcript expression and identification of novel transcripts in the genome. We have applied RNA-Seq to analyze the gene expression patterns in water flea otherwise known as *D. pulex* to work out the genetic details underlying heavy metal induced stress (unpublished) and predator induced phenotypic plasticity (PIPP) (Rozenberg et al., 2015), independently. My bioinformatics analysis of the RNA-Seq data has facilitated the discovery of key biological processes participating in metal induced stress response and predator induced defense mechanisms in *D. pulex*. These studies are great additions to the field of ecotoxicogenomics, phenotypic plasticity and have aided us in gaining mechanistic insight into the impact of toxicant and predator exposure on *D. pulex* at a bimolecular level.

PUBLIC ABSTRACT

Cells are the fundamental structural and functional components of all living beings (USDOEGRP, 2008). A cell’s function is governed by the genetic code of its DNA (USDOEGRP, 2008). A genome of an organism consists of a comprehensive set of its DNA molecules (USDOEGRP, 2008). Each constituent DNA in the genome is called a chromosome (USDOEGRP, 2008). A chromosome contains genes that holds the genetic code to make proteins (USDOEGRP, 2008). The transfer of genetic information from genes to proteins involves an intermediate biomolecule called RNA (Janet, Szostak, & Bell, 2008). A gene is first copied into RNA that is later translated into proteins. A complete collection of all protein coding RNAs or transcripts of an organism makes its transcriptome. Proteins perform all the work required for maintaining the structure and function of our tissues and organs (GHR, 2018) . A cell’s proteome comprises of all the proteins in a cell (USDOEGRP, 2008). Concisely, central dogma of molecular biology as explained by Francis Crick says, protein is made from RNA and RNA is made from DNA (Crick, 1970). Ability to read and analyze the genetic data in these biomolecules will provide us the systematic view into the genetic constitution of an organism and will enlighten us with the molecule basis behind their observable traits or phenotype. Application of this information comprises of understanding human development and diseases, genetic differences and similarities between different organisms and species, and the impact of genetics and environment on their evolution.

Persistent advancements in genomics such as invention of high-throughput sequencing methodologies has facilitated the framework to access and analyze various

omics (genome, proteome and transcriptome) data on humans and other organisms in a rapid, unbiased and cost effective manner. This has led to the enrichment of our scientific knowledgebase and attributed towards the explosion of biological data. The ever-increasing size of this data requires methods with ability to accurately and efficiently probe them and mine meaningful patterns. Computational approaches have been indispensable in storing, managing and analyzing large datasets. Bioinformatics is a collection of computational approaches that specialize in probing and managing biological datasets. In chapter 2 of this thesis, I have focused on describing the design and development of a genome browser tool called OikoBase that allows convenient access to integrated omics data on a water tunicate called *Oikopleura dioica* (*O. dioica*). *O. dioica* is a model organism used in multiple fields of research such as developmental biology, genome evolution and phenotypic plasticity in chordates. This browser is a novel computational web resource that allows investigation of 18,020 genes and their expression patterns across 18 key life cycle stages in *O. dioica.* Additionally, it also offers homology search capability to find similar genes from other organisms in *O. dioica*, which will immensely contribute towards a more comprehensive functional characterization of their genes.

Exomes are the protein coding regions of the genome. Mendelian disorders are single gene disorders that are known to pass from one generation to another in a family (Chial, 2008). Research in human disease modelling of Mendelian disorders have established exomes as the genomic region to harbor disease causing mutations (Stenson et al., 2009). Identification of these etiologic mutations will further our knowledge on their genetic role in human development and diseases. Targeted sequencing of exome instead of the whole genome has been considered as a cost-efficient and practical strategy to investigate all mutations in these genomic regions (Bamshad et al., 2011). We have applied this sequencing technique to explore the exomes from selected human patients in a renal agenesis (RA) pedigree that followed a Mendelian disease transmission pattern. Congenital anomalies of the kidney and urinary tract (CAKUT) represents a wide range of birth defects in children (Rodriguez, 2014). RA is a severe form of CAKUT where children are born with one (unilateral renal agenesis (URA)) or no kidneys (bilateral renal agenesis (BRA)) (Brophy et al., 2017; Yalavarthy & Parikh, 2003). Due to the sheer amount of mutation data generated from exomes, it is extremely hard to identify the disease causing variants manually, with efficiency and accuracy. In chapter 3 of this thesis, I have explained various computational approaches that I applied, especially under section “Materials and methods” and subsection “Exome sequencing analysis, Iowa”, that enabled the discovery of a novel RA gene called *GREB1L*, which when mutated results in URA and BRA phenotypes. This knowledge is a significant addition towards resolving the genetic basis behind CAKUT. Identification of novel disease-associated genes will empower our perception on human disorders and development.

RNA-Seq is a high-throughput sequencing technique that has allowed unbiased analysis of whole transcriptome. We have applied this technique to explore gene expression patterns in *D. pulex*, otherwise known as common water flea, when exposed to predators (Rozenberg et al., 2015) and heavy metals (unpublished). *D. pulex* can be easily maintained in a laboratory and their genome is highly sensitive to chemicals released from their predators called kairomones (Rozenberg et al., 2015), and environmental toxicants. Additionally, the compact size of their genome allows us to conveniently model them for ecotoxicological studies (ecotoxicology: characterizing genetic interactions between chemical stressors and biological organisms) (J. K. Colbourne et al., 2011). We have bioinformatically discovered key biological processes in this organism explaining their defense response against predators and genetic response against environmental toxicants using RNA-Seq. In chapter 4 and 5 of this thesis, under the "Materials and methods" section, I have explained the bioinformatics strategies employed to analyze the transcriptome of *D. pulex* under heavy metal and predator induced stress, respectively. Similar studies will boost our understanding of the correlation between genetic responses as a solution to the current environmental challenges. Molecular response patterns at an organism level can be extrapolated to understand changes at their population levels (Schirmer, Fischer, Madureira, & Pillai, 2010). This will allow us to perform the following long standing goals in ecotoxicology:

a) Predicting the influence of environmental changes on the evolution and survival of an organism.

b) Identify, resolve and deter harmful impact of pollutants on our environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTAL FIGURES

CHAPTER 1. INTRODUCTION

An ocean of biological data is accumulating at a faster rate due to advancements in high-throughput and reasonably inexpensive DNA sequencing technologies(Zou, Ma, Yu, & Zhang, 2015). GenBank database contains over 200 million independent nucleotide sequences from at least 370,000 species (Benson et al., 2017; GenBank, 2018). Uniprot Knowledgebase (UniprotKB) is the major resource for publicly available protein sequences and their aligned annotations ("UniProt: the universal protein knowledgebase," 2017). It currently holds over 60 million unique amino acid or protein sequences from at least 5,631 independent proteomes ("UniProt: the universal protein knowledgebase," 2017). Furthermore, data from whole genome sequencing, gene expression studies, structural and functional characterization of proteins, and protein-protein interaction are contributing towards the ever-increasing volume and diversity of biological data (Luscombe, Greenbaum, & Gerstein, 2001). Informatics approaches can efficiently probe enormous amount of data and mine meaningful patterns from them (Luscombe et al., 2001). For example, the Interpro database and InterproScan algorithm (Jones et al., 2014), have been instrumental in the analysis and annotation of sequences in UniprotKB (Robert D. Finn et al., 2017). InterproScan algorithm outputs protein domain predictions and their associated annotations for each input amino acid sequence by allowing a comprehensive scan against the “signatures” or predictive models such as profile hidden Markov models (HMMs) and position-specific scoring matrices (PSSM), of 14 member databases such as CATH-Gene3D (Lam et al., 2016), SMART (Letunic, Doerks, & Bork, 2015), HAMAP (Pedruzzi et al., 2015), SUPERFAMILY (Oates et al., 2015), PANTHER (Mi, Poudel, Muruganujan, Casagrande, & Thomas, 2016), Pfam (R. D. Finn et al., 2016), TIGRFAMs (Haft et al., 2013), PIRSF (C. H. Wu et al., 2004), ProDom (Bru et al., 2005), PRINTS (Attwood et al., 2012), PROSITE patterns (Sigrist et al., 2013), Structure–Function Linkage Database (SFLD) (Akiva et al., 2014), PROSITE Profiles (Sigrist et al., 2013), and the Conserved Domains Database (CDD) (Marchler-Bauer et al., 2015) (Robert D. Finn et al., 2017). The unique biological focus of each member database collectively empowers the prediction process by offering dimensionality and reducing false positive predictions. The Interpro “signatures” and InterproScan algorithm have successfully predicted annotations for at least 56 million unique protein sequences in UniprotKB (Robert D. Finn et al., 2017). The InterproScan web based application, processes at least 40 million query sequences per month from users (Robert D. Finn et al., 2017). Additionally, informatics approaches have made significant contributions towards improving data storage, organization, sharing, and retrieval. Therefore, they are invaluable to the current and future research in biology and medicine (Luscombe et al., 2001).

Bioinformatics is a branch of informatics that specializes in gathering and analyzing complex omics data as follows:

- Genomics (study of genes and genomes (whole DNA of an organism); genes are DNA sequences that hold instructions to make protein molecules (NHGRI, 2015))

- Transcriptomics (a comprehensive study of RNA such as mRNA; biomolecules produced from genes to be translated into amino acid sequences or proteins (Nature, 2018), and other non-coding RNAs such as tRNA and rRNA; biomolecules that participate in protein synthesis, and snRNA; biomolecules that process RNA prior to their translation into proteins (Education, 2018)).
- Proteomics (large scale study of proteins; proteins are biomolecules that perform all the functions required for life (Scitable, 2014c)).
- Metabolomics (study of metabolites; metabolites are small molecules that represent the end products of metabolism (Jordan et al., 2009). These biomolecules explain details concerning cellular physiology in a comprehensive manner) (Horgan & Kenny, 2011; Manzoni et al., 2018).

A principal task for bioinformaticians involves creating and maintaining public databases that can help biologists to share, retrieve, visualize and integrate biological data and their aligned annotations from multiple sources (Zou et al., 2015). Genome browsers are graphical user interfaces (GUI) that allow researchers to connect with a database, graphically visualize omics data and download their associated annotations with convenience. During my training as a bioinformatician in Dr. Manak’s lab, I followed a three-pronged approach as follows:

- First part of my training involved developing a novel database and genome browser for *Oikopleura dioica* (*O. dioica*) called OikoBase (Gemma. Danks & Parida, 2013).

- Second part of my training required me to bioinformatically analyze selective exomes from a renal agenesis (RA) pedigree that led to the discovery of a novel RA associated gene (Brophy et al., 2017).
- Third part of my training included independently processing and analyzing RNA-Seq data to study acute heavy metal exposure (unpublished) and predator-induced phenotypic plasticity (Rozenberg et al., 2015) in *D. pulex*.

A traditional DNA microarray or DNA chip allows interrogation of thousands of known genes in a genome simultaneously (Scitable, 2014a). A typical DNA chip is a glass microscopic slide that holds thousands of spots at defined positions. The glass is usually coated with an organo-functional alkoxysilane coating that facilitates a link between the glass and organic biomolecules such as DNA fragments or probes, of flexible lengths (Conzone & Pantano, 2004). Each spot on the coated glass slide contains radioactively labeled probes that exhibit fluorescence when they bind to complementary sequence of a known gene in a sample (Conzone & Pantano, 2004). Tiling arrays are whole genome microarray platforms that differ from traditional DNA microarrays by probing contiguous regions of a genome (Choudhuri, 2014). They are primarily used to perform functional characterization of genomic sequences with no prior known function (Choudhuri, 2014). Probe designs such as partially overlapping probes, and spaced apart probes can be applied to control the resolution power of tiling arrays (Choudhuri, 2014).

O. dioica is a model organism used in cross-disciplinary research ranging from chordate genetics and development to molecular ecology (G. Danks et al., 2013). It has a miniature genome of 70 megabases (Mb) in size that contains 18,020 predicted genes and can be easily maintained in the laboratory (G. Danks et al., 2013; Denoeud et al., 2010; Seo et al., 2001). The compact genome of *O. dioica* has allowed generation of ultra-high resolution tiling array data using the NimbleGen platform. The generated data captures snapshots of *O. dioica*'s gene expression profile at key life cycle stages such as early development (oocyte, 2-8 cell embryos, 1 hour post fertilization (1HPF), tailbud and hatched), organogenesis and metamorphosis (early tadpole, tailshift), somatic growth (Day 1-3), and reproductive organ growth and differentiation (Day 4 and 5) (G. Danks et al., 2013). In addition to this dataset, gene expression data on testis and ovaries were also generated to describe reproductive organ development in adult animals (G. Danks et al., 2013). Importantly, *O. dioica* ceases to develop at Day 3 under high animal density (restrictive conditions); therefore, gene expression data on three supplementary (Day 2-4 dense) stages covering the developmental arrest were also added to this dataset (Chapter 2, subsection named "Developmental transcriptome of *O. dioica*", under section "Biological source materials and data generation" of this thesis explains these life cycle stages in detail) (G. Danks et al., 2013). This dataset serves as a rich resource for studying chordate evolution and development at a biomolecular level in *O. dioica* (G. Danks et al., 2013). OikoBase is a unique web resource that allows researchers in these fields to query, retrieve, and graphically visualize crucial omics data across multiple developmental time points of this organism. I am the webmaster of the *O. dioica* genome browser project that includes 1260 genomic sequence scaffolds as well as genes, transcripts and coding sequence tracks (G. Danks et al., 2013; Gemma. Danks & Parida, 2013). My tasks included as follows:

- Designing and developing the OikoBase (Gemma. Danks & Parida, 2013).
- Taking user requests via email and suitably implementing them in OikoBase to accommodate integration and retrieval of diverse omics data.
- Maintaining the database and performing necessary checks on the operation of crucial features and associated links for each web page in OikoBase.

Similar model organism genome browsers such as wFleaBase (John K. Colbourne, Singan, & Gilbert, 2005), FlyBase (Gramates et al., 2017), and WormBase (Harris et al., 2010) are excellent resources when exploring omics data on closely related species. A crucial feature that separates OikoBase from these databases is it allows users to explore *O. dioica* transcriptome at an ultra-high resolution and conveniently download genome-wide gene expression data for a comprehensive set of developmental stages.

The second part of my training concerned bioinformatically analyzing selective human exomes from a renal agenesis (RA) pedigree that followed disease transmission in a Mendelian fashion and prioritize candidate disease-associated damaging mutations. The Mendelian fashion of disease transmission involves a single gene that is likely causing the disease across generations in a family (Genetic, District of, & Department of, 2010). The majority of known allelic variants associated with Mendelian disorders reside in the protein-coding regions of the genome or exome (Stenson et al., 2009). These etiologic variants such as missense single nucleotide substitutions or insertions/deletions

(INDELS) are rare, protein altering and deleterious by nature (Bamshad et al., 2011). Therefore, targeted sequencing of the exome instead of the whole genome is an efficient and cost-effective way to explore approximately all variations existing in these regions (Bamshad et al., 2011). Furthermore, the whole exome sequencing (WES) approach has helped detect disease-causing variants in multiple studies involving human (Mao et al., 2016; Micheal et al., 2017; S. B. Ng et al., 2010; Sarah B. Ng et al., 2009; Salih et al., 2017) and mouse (Fairfield et al., 2011; Hilton et al., 2011).

Abnormalities in kidney and/or urinary tract morphogenesis can cause a broad range of structural defects collectively known as Congenital Anomalies of the Kidney and Urinary Tract (CAKUT) (Vivante et al., 2017). It is a common cause of morbidity and mortality in children (Sanna-Cherchi et al., 2009). The most severe form of CAKUT is RA, or the complete absence of renal tissue at birth (Yalavarthy & Parikh, 2003). Unilateral renal agenesis (URA) is the absence of one kidney that can cause proteinuria, hypertension and early renal failure but is typically compatible with life (Schreuder, Langemeijer, Bökenkamp, Delemarre-Van De Waal, & Van Wijk, 2008). However, bilateral renal agenesis (BRA) is the absence of both kidneys, which always results in death at birth (Potter, 1946, 1965). URA occurs at a relatively higher frequency, than BRA (1/1000 (Woolf & Hillman, 2007) versus; 1/3000-1/5000 (Loendersloot, Verjaal, & Leschot, 1978; McPherson, 2007), respectively). It is important to note these occurrence estimations are affected by under-reporting (Norwood & Chevalier, 2003; Yalavarthy & Parikh, 2003), as renal ultrasounds are not a standard recommendation during newborn screening.

We used exome-sequencing strategy to identify causal variants in two independent RA pedigrees of which I analyzed the Iowa pedigree only (Brophy et al., 2017). Bioinformatics strategies aided in maintaining the quality of sequencing reads and adequate read depth at targeted exome sites for confident variant calling against the human reference genome (version-glk_v37) (Brophy et al., 2017). These raw variant calls were then bioinformatically filtered to generate a refined list of candidate genes that carried damaging mutations potentially associated with the disease (Brophy et al., 2017). Bioinformatics tools and techniques are extremely effective in narrowing down highly likely disease-causing mutations from a substantial amount of variant calls. Researchers can then experimentally confirm pathogenicity of these candidate mutations. Identifying novel disease-causing mutations will contribute towards comprehending the genetic basis of Mendelian diseases and their impact on human development (Z. Wang, Liu, Yang, & Gelernter, 2013). I worked on the following tasks:

- Maintaining the quality of raw sequencing reads and mapping them to the human reference genome (version-glk_v37).
- Processing the sequence alignments to make confident variant calls.
- Filtering the raw variant calls and generating a list of candidate genes carrying potential disease-associated mutations.

The candidate gene list generated from my analysis led to the identification of a novel RA gene called *GREB1L* (Brophy et al., 2017).

Finally, I independently generated and analyzed the whole transcriptome expression data to study heavy metal-induced stress and predator-induced defenses, in

Daphnia pulex (*D.pulex*) as a part of my three-pronged training approach. Genomic methodologies such as whole genome microarrays and next-generation sequencing techniques have advanced transcriptome studies in ecotoxicology (Bouetard et al., 2012; Schirmer et al., 2010). Currently, RNA-Seq is the technique of choice that allows genome-wide gene expression analysis due to the following reasons:

- RNA-Seq allows quantification of gene expression values at a wider dynamic range.
- It has low or close to negligible background signal.
- It allows detection of known and novel transcript boundaries up to single base pair resolution when compared to traditional array hybridization techniques such as DNA microarrays (McGettigan, 2013; Z. Wang et al., 2009).

RNA-Seq is a high-throughput sequencing technique that sequences a cDNA library generated from RNAs in tissues or cells (Qi, Liu, & Rong, 2011). These sequences either are mapped to a pre-existing reference genome or assembled *de novo* to generate a transcriptome map (Grabherr et al., 2011). This map defines high-resolution transcript boundaries and displays the height of their expression based on mapped read depth. We have applied this technique to obtain a comprehensive transcriptome profile of the microcrustacean and key aquatic indicator *D. pulex* when exposed to common environmental heavy metals such as cadmium (Cd), nickel (Ni) and zinc (Zn) (Cempel & Nikel, 2006; Eisler et al., 1985; Lambert, Leven, & Green, 2000). *D. pulex* whole transcriptome expression data will aid us to highlight key biological processes and interaction between different processes as a response to acute heavy metal exposure. A comparison between our gene expression data and the regulation of genes involved in known heavy metal-induced pathways such as oxidative stress, hypoxia, lipid peroxidation, antioxidant response and calcium and iron homeostasis (Ercal, Gurer-Orhan, & Aykin-Burns, 2001; Stohs & Bagchi, 1995), will enhance our understanding of these biological processes in this organism and further validate *D. pulex* as a convenient model for ecotoxicological studies. I have applied bioinformatics tools and techniques to perform the following tasks:

- Quality control and mapping of raw sequencing reads to the *D. pulex* reference transcriptome (Ye et al., 2017).
- Quantifying transcript expression for all the known transcripts in heavy metal treated (experimental) and untreated (control) *D. pulex* samples.
- Independently annotating transcripts with their associated biological processes.
- Highlighting the biological processes responding to the heavy metal stress.
- Bioinformatically identifying binding sites for key stress associated regulatory elements potentially driving the heavy metal stress-induced pathways.

D. pulex exhibits considerable morphological changes under predator risk that makes it a suitable ecological genomic model to study predator induced phenotypic plasticity (PIPP). Phenotypic plasticity is defined as an organism’s ability to generate multiple phenotypes from one genotype (Rozenberg et al., 2015). We applied RNA-Seq to detect unbiased genome-wide gene expression in *D. pulex* and identify regulation of gene families associated with predator-induced defenses. The mapping of gene regulation to observed morphological alterations would further our understanding of molecular basis underlying PIPP in *D. pulex*. In future, similar studies will enhance our understanding of the various solutions generated by the genetic code of an organism to solve specific environmental problems (Lewontin, 2001). I have applied bioinformatics tools and techniques to perform the following tasks:

- Quality control, processing and mapping the raw sequencing reads to the *D. pulex* draft genome (J. K. Colbourne et al., 2011).
- Computing genome-wide gene expression data comparing predator induced (experimental) vs non-induced (control) *D. pulex* samples using the “*Daphnia pulex* Genes 2010 beta 3” gene annotations from wFleabase (Genome Informatics Lab, 2005).

Bioinformatics comprises of tools and techniques to mine and create biologically relevant insights from genetic data. In the chapters below, I have described various bioinformatics methods that have aided in the investigation of crucial biological questions, and led to the discovery of new knowledge. The results obtained from my analysis have significantly contributed towards expanding our knowledgebase in fields such as human diseases, ecotoxicogenomics, phenotypic plasticity and chordate developmental biology.

CHAPTER 2. OIKOBASE: A GENOMICS AND DEVELOPMENTAL TRANSCRIPTOMICS RESOURCE FOR THE UROCHORDATE *OIKOPLEURA DIOICA.*[2]

ABSTRACT

We report the development of OikoBase (Gemma. Danks & Parida, 2013), a tiling array-based genome browser resource for *Oikopleura dioica*, a metazoan belonging to the urochordates, the closest extant group to vertebrates. OikoBase facilitates retrieval and mining of a variety of useful genomics information. First, it includes a genome browser which interrogates 1260 genomic sequence scaffolds and features gene, transcript and CDS annotation tracks. Second, we annotated gene models with gene ontology (GO) terms and InterPro domains which are directly accessible in the browser with links to their entries in the GO (Carbon et al., 2009) and InterPro (Robert D. Finn et al., 2017) databases, and we provide transcript and peptide links for sequence downloads. Third, we introduce the transcriptomics of a comprehensive set of developmental stages of *O. dioica* at high resolution and provide downloadable gene expression data for all developmental stages. Fourth, we incorporate a Basic Local Alignment Search Tool (BLAST) (Altschul, Gish, Miller, Myers, & Lipman, 1990) to identify homologs of genes and proteins. Finally, we include a tutorial that describes how to use OikoBase as well as a link to detailed methods, explaining the data generation and analysis pipeline. OikoBase

[2] This chapter has been published in Danks, G., Campsteijn, C., **Parida, M.**, Butcher, S., Doddapaneni, H., Fu, B., . . . Manak, J. R. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica. Nucleic Acids Res, 41*(Database issue), D845-853. doi:10.1093/nar/gks1159

will provide a valuable resource for research in chordate development, genome evolution and plasticity and the molecular ecology of this important marine planktonic organism.

INTRODUCTION

The urochordate (or tunicate) appendicularian, *Oikopleura dioica*, is a coastal marine planktonic chordate with a pan-global distribution. As an abundant component of mesozooplankton communities (Gorsky & Fenaux, 1998), appendicularians are noted for their ability to rapidly expand population size in response to algal blooms. In a defining feature, urochordates are partially or completely enclosed in extracellular, cellulose-based 'tunics' or 'houses' involved in filter-feeding, making them the only animals known to synthesize cellulose (Sagane et al., 2010). Appendicularians frequently resynthesize their house during their short life cycle, and discarded houses (marine snow) play significant, and sometimes dominant, roles in oceanic vertical carbon flux (Alldredge, 2005; Robison, Reisenbichler, & Sherlock, 2005), thus impacting global carbon cycles. The urochordates are the closest extant relatives to the vertebrates (Delsuc, Brinkmann, Chourrout, & Philippe, 2006) and have a simplified chordate body plan with a notochord, dorsal neural tube, gill slits and endostyle. *Oikopleura dioica* remains transparent throughout its short (less than 1 week) chordate life cycle, exhibits high fecundity [more than 300 eggs per female (Troedsson, Bouquet, Aksnes, & Thompson, 2002)] and can be cultured in the laboratory for hundreds of generations (Bouquet et al., 2009). It is the only known dioecious urochordate species. Compared with ascidians, *O. dioica* undergoes a morphologically simple metamorphosis, in which the tail is not resorbed but only shifts from a posterior directed orientation in the tadpole to become orthogonal to the trunk at metamorphosis. Development is very rapid, with metamorphosis completed after 12–14 h at 15°C. Throughout the life cycle, the tail retains the notochord as its axial structure in order to function in pumping water through the filters of the house for feeding. The animal is generally found from 0 to 200 m depth and tolerates a wide range of temperatures and salinities (Fenaux, Bone, & Deibel, 1998). The 70-Mb compact, sequenced genome containing more than 18 000 predicted genes ranks among the smallest known metazoan genomes (Denoeud et al., 2010; Seo et al., 2001), with both gene regulatory and intronic regions highly reduced in size. This rapidly evolving lineage exhibits profound alteration of deeply conserved features of metazoan genome architecture, thus offering interesting perspectives in the study of genome plasticity (Denoeud et al., 2010) and developmental gene regulation.

Here we present a tiling array-based developmental transcriptome genome browser resource, OikoBase ((Gemma. Danks & Parida, 2013); based on the popular GBrowse software), at ultra-high resolution in terms of both developmental chronology and the underlying genome sequence features. The unbiased transcriptome datasets described here comprise 12 key developmental stages encompassing the entire life span of *O. dioica*, in addition to testis, ovary and somatic body-specific sample sets. Finally, three transcriptome datasets generated from animals under growth/developmental arrest are included. The compact genome of *O. dioica*, as well as its short lifespan, permits a level of comprehensiveness and resolution that rivals that of well-established model organisms. A total of 62 Mb of the non-repeat genome (93% of the total genome) was tiled on a single microarray (2.1 million features, with a 2-fold genome coverage) using isothermal oligonucleotide probes with a median size of 53 and 24 base pairs (bp) overlaps between adjacent probes, permitting reliable detection of short introns (47-bp peak size in the *Oikopleura* genome). This platform and its ancillary features, which will be used to explore both the transcriptome as well as global chromatin structure, will serve as a rich resource in understanding the developmental and evolutionary biology of this ecologically important organism in the context of its transcriptional response to changing experimental and environmental inputs.

BIOLOGICAL SOURCE MATERIALS AND DATA GENERATION

ANIMAL CULTURE AND COLLECTION

Oikopleura dioica were maintained in culture at 15°C (Bouquet et al., 2009). Oocytes were collected from mature females, and developmental stages up to and including metamorphosis were generated by *in vitro* fertilization (Schulmeister, Schmid, & Thompson, 2007), with embryos left to develop in artificial sea water (Red Sea, final salinity 30.4–30.5 g/l) at room temperature to the desired stage. Day1–Day5 animals were placed in artificial seawater, chased from their houses, left for 30 min to empty their gut, anesthetized in cold ethyl 3-aminobenzoate methanesulfonate salt (MS-222, 0.125 mg/ml; Sigma), and then collected. Ovary and testes extracts were collected as previously described (Campsteijn, Øvrebø, Karlsen, & Thompson, 2012). To obtain the trunk samples, Day5 animals were washed in cold phosphate-buffered saline and anesthetized in cold MS-222, transferred to cold Buffer A [10 mM Tris (pH 7.5), 360 mM sucrose, 75 mM NaCl, 10 mM ethylenediaminetetraacetic acid, 10 mM ethyleneglycol bis[aminoethylether]-tetraacetic acid, 3 mM dithiothreitol, 1 mM phenylmethylsulfonyl fluoride and 1:500 RNase OUT (Invitrogen)] and gonads were punctured and removed. Remaining trunks were washed once in cold Buffer A, pooled and processed for RNA isolation. For collection of the Day2–Day4 dense samples, the normal dilution of animal spawn at Day1 was omitted (Bouquet et al., 2009), yielding a culture with ~5-fold higher animal density. In partial compensation for elevated animal numbers while suppressing algal overgrowth, the food concentration was doubled at each developmental stage. The elevated animal density and feeding regime were maintained throughout the experiment, with animal collection performed at time points as indicated.

GENE ANNOTATION

A very detailed reporting of gene annotation protocols is provided in the supplementary information of Denoeud et al. (2010). A summary version is presented here. Semi-HMM-based Nucleic Acid Parser (SNAP) ab initio gene prediction (Korf, 2004) was used to create a clean set of *O. dioica* genes. This set was used to train gene prediction algorithms and optimize their parameters. SNAP was launched using the Caenorhabditis elegans configuration file, and only models with all introns confirmed by at least one *O. dioica* cDNA were retained. Models that contained at least one exon that overlapped a cDNA intron were rejected. Three hundred models were randomly selected to create the *O. dioica* clean training set.

Exofish (Crollius et al., 2000) comparisons were done with Biofacet (www.gene-it.com). When ecores (Evolutionarily COnserved REgions) were contiguous in the two genomes, they were included in the same ecotig (contig of ecores). Exofish comparisons were performed between *O. dioica* and four organisms: Tetraodon nigroviridis, Strongulocentrus purpuratus, Ciona savignyi and Ciona intestinalis. High scoring segment pairs were filtered by length and percent identity.

The Uniprot (Apweiler et al., 2012) database was then used to detect genes conserved between *O. dioica* and other species. Since GeneWise (Birney, Clamp, & Durbin, 2004) is computationally intensive, sequences in the Uniprot database were first aligned with the *O. dioica* genome assembly using BLAST-Like Alignment Tool (BLAT)(Kent, 2002). Each significant match was chosen for a GeneWise alignment. The default GeneWise gene parameter file was modified to account for unusual splice sites of *O. dioica* genes. Geneid (Parra, Blanco, & Guigó, 2000) and SNAP ab initio gene prediction software were then trained on the 300 genes from the training set.

Expressed Sequence Tags (ESTs) generated from three full-length-enriched cDNA libraries were prepared from a cultured outbred population [large pools of unfertilized eggs, embryos at mixed stages from 1 to 3 h post-fertilization (pf), tadpoles 6–10 h pf and Day 4 animals] and 180 000 cDNA clones were end sequenced. After assembly of 5′- and 3′-sequences, a total of 177 439 sequences were aligned to the *O. dioica* genome using the following pipeline: after masking of polyA tails and spliced leaders, the sequences were aligned with BLAT on the genome assembly and matches with scores within 99% of the best score were extended by 5 kb on each end and realigned with the cDNA clones using Exonerate (Slater & Birney, 2005), allowing for non-canonical splice sites, with the following parameters; –model est2genome –minintron 25 –maxintron 15 000 –gapextend -8 –dnahspdropoff 12 –intronpenalty -23.

An National Center for Biotechnology Information (NCBI) (Sayers et al., 2012) collection of ~1,500,000 public tunicate ESTs was then aligned with the *O. dioica* genome assembly using BLAT alignments refined with Est2Genome (Mott, 1997). Significant matches were chosen for alignment with Est2Genome. BLAT alignments used default parameters between translated genomic and translated ESTs.

The above resources were combined to automatically build *O. dioica* gene models using GAZE (Howe, Chothia, & Durbin, 2002). Individual predictions from each program (Geneid, SNAP, Exofish, GeneWise, Est2genome and Exonerate) were broken into segments (coding, intron and intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start and transcript stop). Exons predicted by ab initio software, Exofish, GeneWise, Est2Genome and Exonerate were used as coding segments. Introns predicted by GeneWise and Exonerate were used as intron segments. Intergenic segments were created from the span of each mRNA, with a negative score (forcing GAZE not to split genes). Predicted repeats were used as intron and intergenic segments, to avoid prediction of protein-coding genes in such regions.

The whole genome was scanned to find signals (splice sites, start and stop codons). In order to annotate genes containing non-canonical splice sites, all G* (GT, GA, GC and GG) donor sites were authorized. GAZE used segments predicting exon boundaries only if GAZE chose the same boundaries. Each segment or signal from a given program was accorded a value reflecting our confidence in the data, and these values were used as scores for the arcs of GAZE automaton. All signals were assigned a fixed score, but segment scores were context sensitive: coding segment scores were linked to the percentage identity (%ID) of the alignment; intronic segment scores were linked to the %ID of the flanking exons. A weight was assigned to each resource to further reflect reliability and accuracy in predicting gene models. This weight acted as a multiplier for the score of each information source, before processing by GAZE. When applied to the entire assembled sequence, GAZE predicted 17 113 gene models on the reference assembly. The final proteome of 18 020 gene models was obtained by adding 907 gene models from an allelic assembly that were not present in the reference assembly.

RNA ISOLATION AND CDNA SYNTHESIS

RNA isolation was performed using RNeasy (Qiagen) and RNA quality was assessed using a Bioanalyzer (Agilent). Purified total RNA was briefly treated (15 min at room temperature) with 0.5 units of DNase I (AMP grade; Invitrogen) per microgram of RNA to remove residual DNA, followed by ethanol precipitation and purification of RNA. Double-stranded (ds) cDNA was synthesized using SuperScript™ Double-Stranded cDNA Synthesis Kit (Invitrogen) according to Roche NimbleGen protocols, followed by phenol:chloroform:isoamylalcohol (25:24:1) as well as chloroform:isoamylalcohol (24:1) back extraction and ds cDNA was recovered by ethanol precipitation. Typically, 4–10 µg of input total RNA yielded 3–9 µg of ds cDNA.

TILING ARRAY HYBRIDIZATION AND PROCESSING

A tiled genomic microarray was designed to interrogate the entire non-repeat genome of *O. dioica* based on genome sequence [GENBANK/European Molecular Biology Laboratory accession numbers are CABV01000001-CABV01005917, CABW01000001-CABW01006678, FN653015-FN654274, FN654275-FN658470, FP700189-FP710243, FP710258-FP791398 and FP791400-FP884219; (Denoeud et al., 2010)] assembled in collaboration between the Sars International Centre for Marine Molecular Biology (www.sars.no) and Genoscope (www.genoscope.cns.fr/secure-nda/*Oikopleura*). For each developmental stage (or stressor), cDNA was labeled and amplified using the Roche NimbleGen gene expression protocol (http://www.nimblegen.com/products/lit/05434505001_NG_Expression_UGuide_v6p0.pdf) with the following modifications: 15 µg of the labeled cDNA was hybridized to each array, and hybridization was performed on three separate arrays (three technical replicates; four for the oocyte sample). The arrays were washed and processed as per the manufacturer's recommendations, and the arrays were scanned on Molecular Devices GenePix Axon scanners 4000B or 4400 A at 5 or 2.5 µm resolution, respectively. The scanned array data were converted into pair and GFF files. These were processed to generate normalized, background subtracted probe intensities that comprise the transcriptomics tracks in the browser (together with log2-transformed data tracks). Raw and processed probe intensities for all samples have been deposited in NCBI's Gene Expression Omnibus (Barrett et al., 2011) and are accessible through GEO Series accession number GSE39568.

We define a transcriptionally active region (TAR) as any stretch of consecutive positive probes in a particular sample. To compare TARs between samples, we constructed a set of ‘superTARs’ that represent maximal continuous regions in which transcription occurs in any one or more samples. Interestingly, we identified 37,941 superTARs (covering 5.4 Mb), not overlapping annotated genes, suggesting that these TARs represent either unannotated exons of known genes or novel transcripts.

GENE ONTOLOGY AND INTERPRO DOMAIN ANNOTATION

We used Blast2GO (Götz et al., 2008) to annotate all predicted gene model protein sequences with gene ontology (GO) terms and protein names using the non-redundant protein sequences (nr) database at an E-value cut-off of 1×10^{-3} and default weighting parameters in the GO term annotation step (see Blast2GO documentation for further details). This gave us a set of 9,667 gene models with GO annotations. We also used Blast2GO to annotate each protein with InterPro (Hunter et al., 2012) domains using InterProScan. The resulting GO terms and InterPro domains associated with each gene model provide researchers with valuable information on the putative functions of these *Oikopleura* proteins and allow further analyses of overrepresented GO terms on genes of interest. To facilitate the investigation of individual genes, we provide users of the browser with GO and InterPro domain identifiers and name codes associated with each gene (Figure 2.2). Furthermore, we provide links for each to the relevant entry in the GO/InterPro databases. The full sets of GO annotations and InterPro domain annotations are also available for download at OikoBase using the Downloads link.

DEVELOPMENTAL TRANSCRIPTOME OF *O. DIOICA*

To assemble a complete developmental transcriptome for *O. dioica*, we subdivided organismal development into 12 segments. We began with unfertilized oocytes (‘Oocyte’) that are arrested in Metaphase I of meiosis. Oocytes were collected from females that were allowed to spawn naturally by rupture of the gonad wall. The next stage consisted of two to eight cell embryos that encompass the maternal to zygotic transition in the transcriptional regulatory control of development. This was followed by a sample at 1 h pf, a stage at which blastomere fates for most tissue types have already been determined (Stach, Winter, Bouquet, Chourrout, & Schnabel, 2008). Continued rapid cell proliferation within determined blastomere lineages was captured by the ‘tailbud’ and ‘hatched’ stages, the latter stage occurring when the animal emerges from the protective chorion. The ‘early tadpole’ stage is a period of very active organogenesis. The ‘tailshift’ stage represents the completion of metamorphosis and initiation of filter-feeding. This was then followed by samples isolated from stages in which the bulk of animal somatic growth occurs (‘Day1’, ‘Day2’ and ‘Day3’). From the fourth day of development onwards, nutritional resources are increasingly allocated to growth and differentiation of the male and female reproductive organs (‘Day4’ and ‘Day5’). To specifically interrogate development of the reproductive organs, we collected samples derived from dissected and isolated testes or ovaries of Day5.5 animals, as well as a complementary animal sample consisting of Day5.5 animals with their gonads removed (‘trunk’). Importantly, we observed that culturing *O. dioica* at higher animal density (‘restrictive conditions’) causes it to cease development and somatic growth just prior to germline differentiation at Day3 (Ganot, Kallesøe, & Thompson, 2007). Therefore, we included three time points covering the entry into this developmental arrest (‘Day2 dense’, ‘Day3 dense’ and ‘Day4 dense’). Material derived from these 18 samples was processed and hybridized to tiling arrays.

We used our processed tiling array data to calculate gene expression levels, at each stage, for each *Oikopleura* gene model covered by our array (see the ‘Methods’ section of OikoBase for details and qPCR validation) and include a matrix of these values in OikoBase. This gives researchers access to the developmental expression profiles of genes of interest as well as their expression during developmental stasis and in the male and female gonads and the trunk of adult animals. These data are provided in a format that facilitates genome-wide analyses.

In total, out of the 16,749 annotated genes that are covered by our tiling array, we find that 13,081 are expressed at some point during *O. dioica* development. The remaining 3,668 genes (1,161 of which have zero positive probes; the others have <50% positive CDS probes in any stage) could represent silent pseudogenes, genes expressed during very short time points not represented by our time course or genes having environment-specific roles such as during stress responses. Gene expression traffic is very dense in *O. dioica*, and an example of the complex interlaced developmental transcriptome profile of a short 12-kb region of the genome is shown in Figure 2.1. Note that the tiling array data are of sufficient resolution to identify a distal novel 5′-exon of the right-most gene in this interval (see legend of Figure 2.1 for details). The array data also accurately portray gene expression profiles associated with developmental life history characteristics of the animal. For example, secretion of the first *O. dioica* pre-house rudiment is detected in the tailshift stage and inflation of this rudiment into a functional house coincides with the beginning of active filter-feeding. Tissue-specifically, the houses are produced from a specialized oikoplastic epithelium (Ganot & Thompson, 2002; Spada et al., 2001; Thompson, Kallesøe, & Spada, 2001) covering the trunk of the animal. Culturing *O. dioica* at higher animal density and reduced relative food availability causes animal growth and development to cease at the Day3 stage. Despite this, animals continue feeding and replace their house regularly upon house degeneration and filter clogging, demonstrating continued house production even in the absence of apparent animal and cellular growth (Day2 dense to Day4 dense). The full expression profile of a structural component of the house, oikosin 19 (Hosp, Sagane, Danks, & Thompson, 2012) faithfully recapitulates all of these features (Supplementary Figure 2.S1).

DATA MINING OF THE *OIKOPLEURA* GENOME

NAVIGATING OIKOBASE

The OikoBase tiling array-based genome browser contains various functions to query the *O. dioica* genome and the high-resolution developmental transcriptome. Entering the GBrowse link on the Facepage allows queries based on genomic coordinates or annotated gene model, gene transcript or protein identifiers (GSOIDG, GSOIDT and GSOIDP codes, respectively). For each region of interest, unique gene identifiers, CDS, transcript and GC-content can be visualized. Processed transcriptomics tracks for any, up to all, of the 18 samples can be activated and customized in linear or log2 scales. Please note that log2-transformed data appear ‘smoothened’ and less choppy, which may be preferable for some users who wish to view larger scale dynamic expression changes. Selecting the untransformed data files should be preferable for viewers who wish to identify more subtle changes in gene expression.

In order to create an annotation database that allows for mining of gene and protein information, we have built a pipeline that facilitates identification of both GO terms as well as InterPro terms on a gene-by-gene basis. Figure 2.2 illustrates these features in a composite genome browser shot of a 50-kb segment of scaffold one, containing several gene models. The operating mode for this pipeline is described in the Figure 2.2 legend. Options to directly download transcript (GSOIDT code) and protein (GSOIDP code) sequences of interest are also provided, in addition to links that access GO and InterPro domain information. These features should prove very useful in rapidly obtaining additional information for genes expressed in developmental patterns of interest identified in the transcriptomics portion of the OikoBase browser.

BLAST PIPELINE

We have incorporated a BLAST pipeline that allows researchers to find relevant *Oikopleura* homologs to their genes of interest from other organisms. Entering via the BLAST option on the OikoBase Facepage, the complete *O. dioica* reference genome as well as ESTs and gene model-based transcript and proteome resources can be directly queried. An example of this workflow using the human cytoplasmic actin protein sequence as a query is developed in Figure 2.3. This function permits rapid assessment of expression profiles of putative homologs and additional annotation as well as corresponding nucleotide and protein sequences can be retrieved (Figure 2.2) for any desired subsequent analyses. Finally, following the Gene Expression Matrix (GEM) link on the Facepage, it is possible to access the entire set of processed expression values for each annotated gene model across developmental time in an HTML-sortable table. Selecting a desired gene model identifier (GSOIDG code) allows retrieval of the developmental expression profile of that gene model.

CONCLUSIONS AND PERSPECTIVES

Here we have presented a tiling array-based genome browser resource for the research community, which provides unbiased whole-genome transcription data across the key events in development of the urochordate *O. dioica*, as well as providing somatic-, testes- and ovary-specific transcriptional outputs. In addition to the comprehensive chronological resolution of the developmental transcriptome profile, the spatial resolution of the data is such that the small (47-bp peak size) introns that predominate in the compact *O. dioica* genome are reliably detected. These data are useful in improving the gene model annotation of the *Oikopleura* genome and in identifying transcribed unannotated DNA, which likely represent new transcript isoforms of known genes as well as uncharacterized novel genes. The rapid evolution of *O. dioica* has included lineage-specific expansion of gene families including central developmental genes (Denoeud et al., 2010). The analysis of retention and diversification of expression patterns among paralogs should therefore provide insights into mechanisms of sub- and neofunctionalization. For example, a preliminary transcriptomics analysis shows that genes containing the most overrepresented domains (Chavali, Morais, Gough, & Babu,

2011) have highly clustered expression profiles, correlating well with the scaffolding of specific house substructures, a process that requires coordination of cell growth and positioning in addition to tuning of metabolic output within cell fields and between cell fields. The generation of multiple gene paralogs would allow for modulation of cell-specific functional output, as has previously been shown for paralogs of the house structural components, ‘oikosins’ (Hosp et al., 2012; Spada et al., 2001; Thompson et al., 2001). Numerous regulators of mitotic cell division have also been amplified in *O. dioica* (Campsteijn et al., 2012), and our data indicate that expression of these paralogs is frequently anti-correlated, with novel variants expressed in non-mitotic tissues, such as endocycling cells and differentiating testes. Therefore, combining the developmental transcriptomics with DNA sequence divergence of these paralogs could yield novel insights into the evolutionary constraints on mitotic determinants as well as the genetic adaptations required to accommodate variant cell cycles.

Complementing the existing Genoscope *Oikopleura* database, OikoBase now provides transcriptomics data across the animal’s entire life cycle. OikoBase also annotates Genoscope gene models with predicted functions including GO terms, InterPro domains and informative gene names, giving researchers readily accessible, biologically relevant data on these genes through a browser interface. We have also implemented a BLAST function that allows researchers to search for homologs of their gene of interest. As a systemic library of temporal and spatial *O. dioica in situ* images becomes available to significantly complement the temporal transcriptome profiles, such a library will also be incorporated into OikoBase. Genoscope provides sequence resources for a large and expanding number of species, whereas OikoBase is a more dedicated and focused database that will remain an active and improving urochordate resource. Studies are underway to provide more detailed analysis of developmental programs and we plan to layer into the OikoBase resource further transcriptomics interrogations of animal responses to environmental stressors (both tiling array and RNA-seq data), developmental CAGE data for accurate annotation of transcription start sites, and global analysis of histone modifications (regulating chromatin structure and transcriptional states in specific developmental contexts) through ChIP-chip and ChIP-seq.

IMPLEMENTATION

Geneic Model Organism Database (GMOD; http://www.gmod.org) tools, MySQL version 5.1.47 (x86_64), and GBrowse 2.07 were implemented to create the OikoBase genome browser. Perl scripts were used to manage and load the appropriate databases. All transcriptome array data were converted to the Genetic Feature Format Version 3 (GFF3). We also installed a BLAST feature accessible from the Facepage to query sequences against the *Oikopleura* genome. The Facepage was created using JavaScript, Jquery, CSS and HTML5 scripts.

ACKNOWLEDGEMENTS

The authors thank Jean-Marie Bouquet and the staff of Appendic Park, Sars Centre, for their efforts in producing a continuous supply of *Oikopleura* and David Osborne for identifying the phenomenon of density induced developmental stasis. The

authors also thank Willem Haagmans, Roche NimbleGen, for initial consultations on the design of the custom *Oikopleura* tiling array.

EPILOGUE

This chapter was published in Nucleic Acids Res, volume 41(Database issue), page number D845-853, 2013. Gemma Danks is the first author in this manuscript and I am the third author. I have cited the web resource and manuscript as Gemma. Danks and Parida (2013) and G. Danks et al. (2013) in this chapter respectively. I cited the web resource independently of the manuscript due to my significant contribution in designing, building and maintaining it.

OikoBase is a unique web resource that offers researchers an effortless and convinient access to the genome-wide developmental profile of *O. dioica*'s life cycle. Before OikoBase there was no comprehensive and powerful knowledgebase on *O. dioica* that integrated various omics data in one resource such as transcript and protein sequences, ESTs (ESTs are fragments of mRNA characterizing expressed regions of a genome that are used in gene identification and validating predicted gene models (Parkinson & Blaxter, 2009)), gene expression data for 18,020 predicted genes across 18 developmental time points, and homology search against *O. dioica* assemblies using the BLAST feature.

My role in this project involved the following tasks:

- Creating and designing the web pages such as the About, Tutorial, People, GBrowse, BLAST, Methods, GEM, Downloads, Related Links and Updates, and their associated links.

- Formatting the GBrowse page to display gene expression track for each stage.
- Uploaded the tiling array expression data for all 18 time points using custom Perl scripts to the background database.
- Adding a bubble feature on each transcript in the GBrowse page to easily access their associated information such as Gene ontology (GO) terms, InterPro terms, and transcript and peptide sequences.
- Adding a button feature to each column header in the GEM page using jQuery JavaScript library that allows sorting of the entire GEM page in an ascending or descending order.
- Installing and linking the BLAST feature to the BLAST page and creating a dropdown menu that allows users to choose an assembly of sequences such as *O. dioica* complete assembly, ESTs, peptide, and transcript reference assembly, to search for similar sequences based on their query sequence or homologs in *O. dioica*.
- Linking the BLAST output homologs to the Gbrowse page using custom PHP, Python and sed scripts.
- Creating a Downloads page for downloading raw data files such as the whole genome FASTA sequences, published article, GEM dataset, all the GO and InterPro terms associated with each *O. dioica* transcript, and the *O. dioica* gene, transcript, peptide, and EST sequences.

My contributions on the interface design and database develpoment of OikoBase has made it a user friendly resource for biologists. This web resource will continue to serve as an indispensable tool for advancing research in chordate evolution and genomics, and molecular ecology of this crucial model organism.

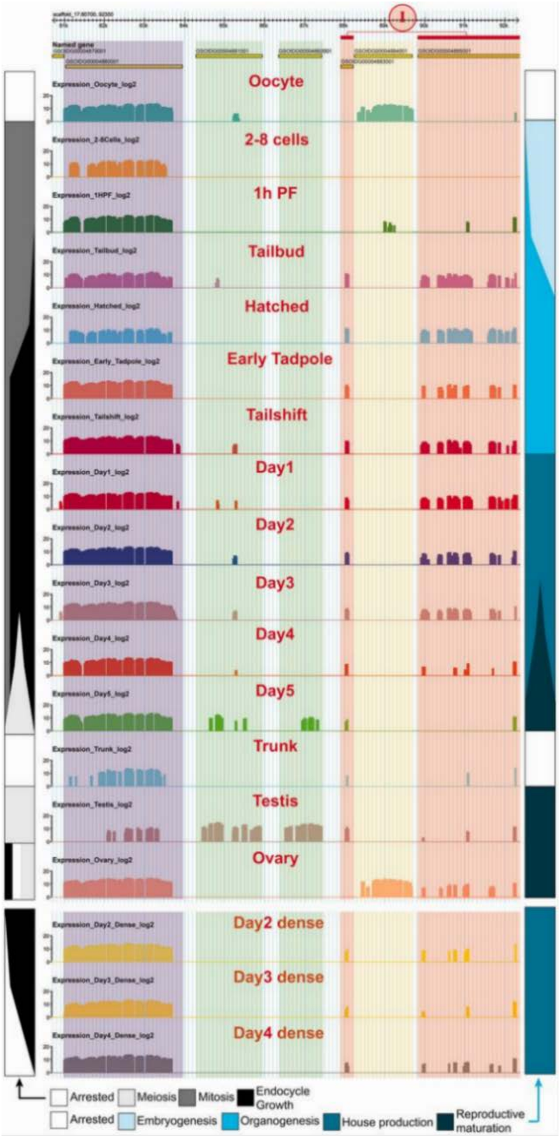
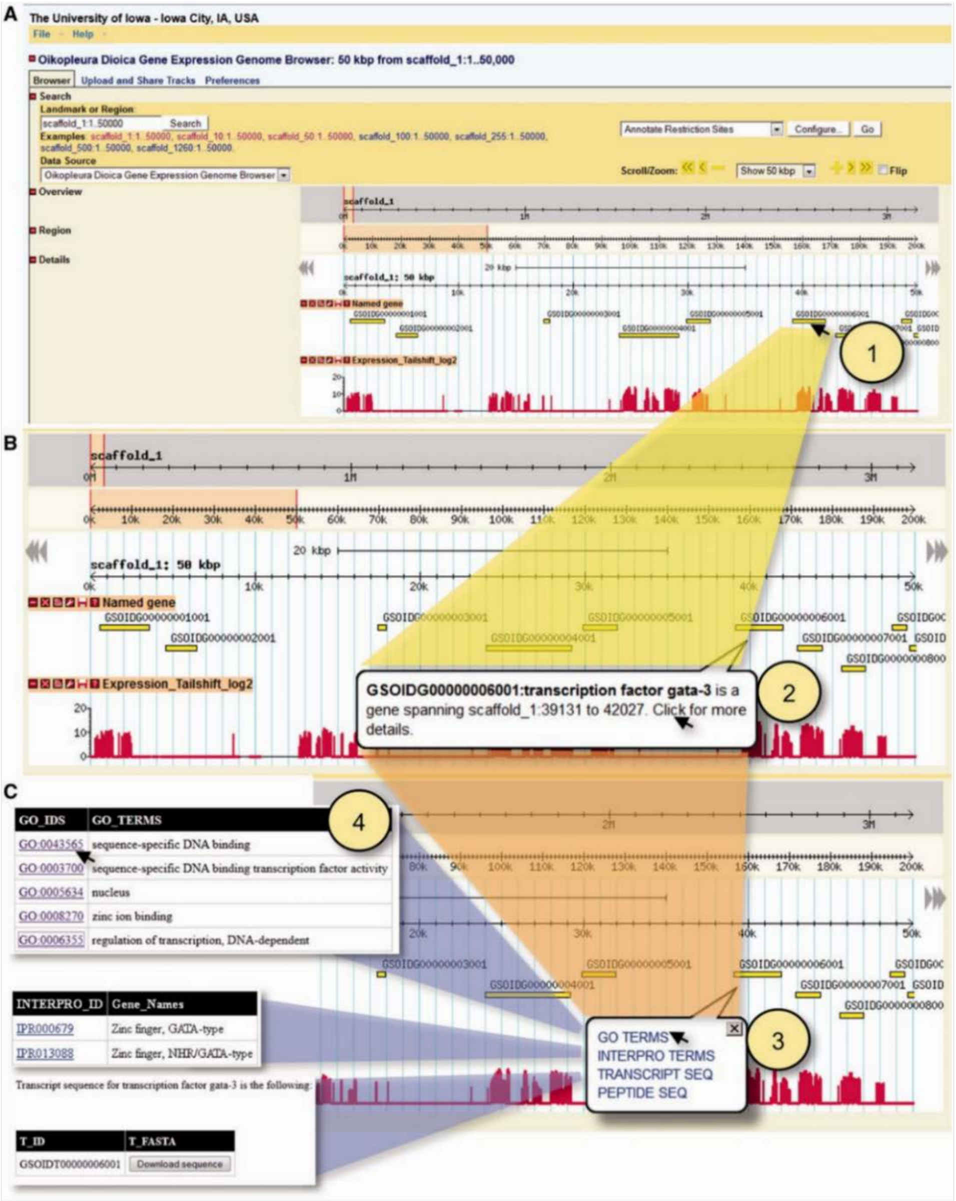


Figure 2.1 Complex developmental gene expression traffic in the compact *O. dioica* genome as visualized in the OikoBase genome browser. In addition to the normal browser output, the developmental stages are indicated in larger font in red type for clarity. Flanking the browser expression data, vertical timelines are included to illustrate key cell cycle (grey-scale, left-hand side) and organismal (blue-scale, right-hand side) processes and transitions covered by our transcriptomics analysis. The illustrated region includes testes-specific gene models (GSOIDG00004881001 and GSOIDG00004882001, green shading) that are also more weakly expressed in Day5 animals, a sub-portion of which contain testes at an earlier developmental stage. Immediately to the left of these is a ubiquitously expressed gene model (GSOIDG00004880001, purple shading). Notably, gene models GSOIDG00004883001 and GSOIDG00004885001 (pink shading) are expressed predominantly from tailbud to Day3, and RACE analysis indicated that they are part of the same gene (unpublished data; the red 'I' indicates their connectedness), expressed from a promoter left of GSOIDG00004883001. However, the short 5′-exon (GSOIDG00004883001) is not expressed in the ovary at which point GSOIDG00004885001 expression is driven by an internal bidirectional promoter also driving GSOIDG0004884001 (yellow shading), as confirmed by RACE and qRT–PCR analysis (unpublished data). This high diversity of developmentally regulated expression occurs in a region measuring <15 kb and indicates that our transcriptome analysis at high temporal and spatial resolution can improve gene model annotation. The Y-axis scale is exponential (log2).

Source: Danks, G., Campsteijn, C., **Parida, M.**, Butcher, S., Doddapaneni, H., Fu, B., . . . Manak, J. R. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res*, *41*(Database issue), D845-853. doi:10.1093/nar/gks1159



The University of Iowa - Iowa City, IA, USA

Oikopleura Dioica Gene Expression Genome Browser: 50 kbp from scaffold_1:1..50,000

Landmark or Region: scaffold_1:1..50000

Examples: scaffold_1:1..50000, scaffold_10:1..50000, scaffold_50:1..50000, scaffold_100:1..50000, scaffold_255:1..50000, scaffold_500:1..50000, scaffold_1260:1..50000.

Data Source: Oikopleura Dioica Gene Expression Genome Browser

Annotate Restriction Sites

Scroll/Zoom: Show 50 kbp

GSOIDG0000006001:transcription factor gata-3 is a gene spanning scaffold_1:39131 to 42027. Click for more details.

GO_IDS	GO_TERMS
GO:0043565	sequence-specific DNA binding
GO:0003700	sequence-specific DNA binding transcription factor activity
GO:0005634	nucleus
GO:0008270	zinc ion binding
GO:0006355	regulation of transcription, DNA-dependent

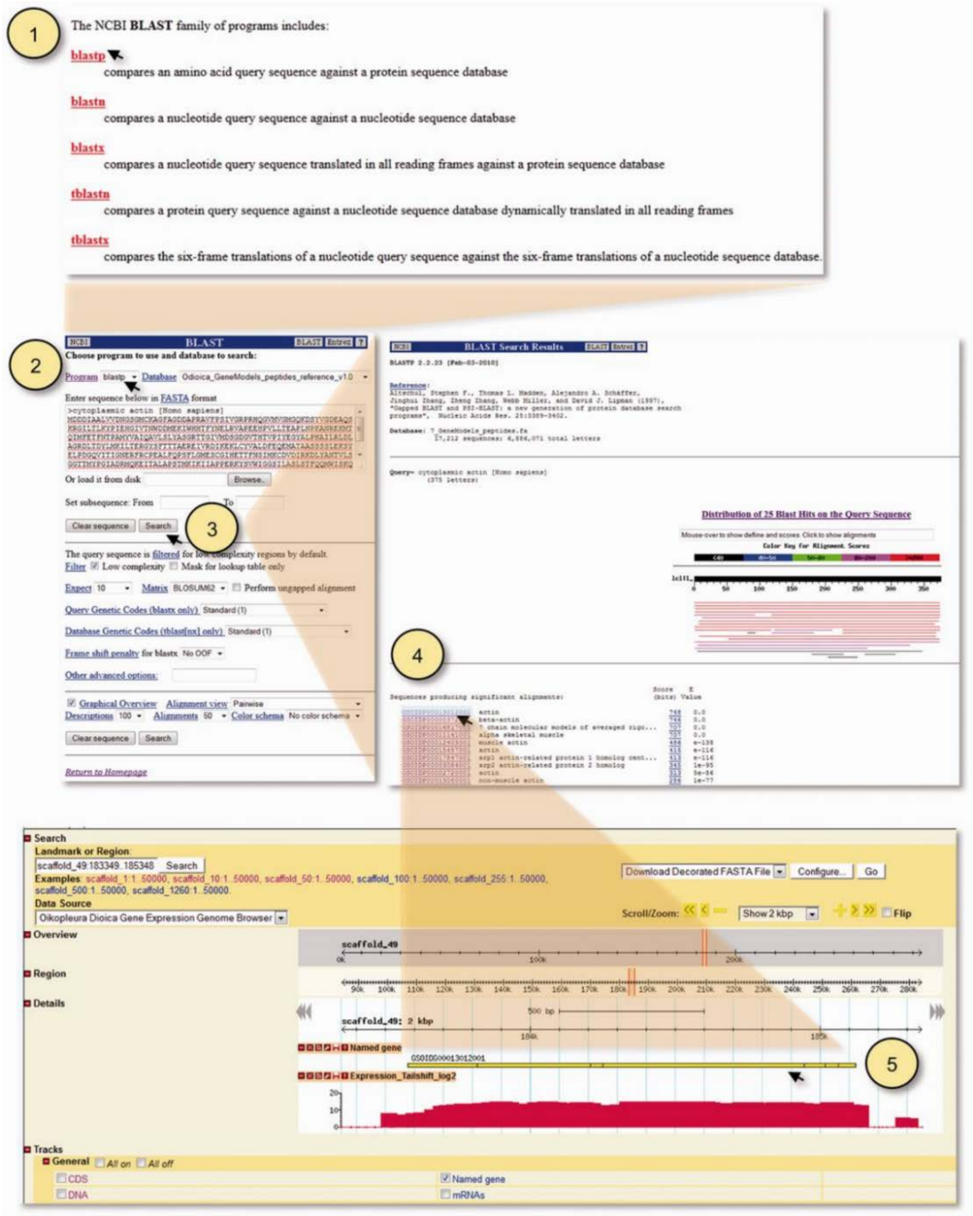
INTERPRO_ID	Gene_Names
IPR000679	Zinc finger, GATA-type
IPR013088	Zinc finger, NHR/GATA-type

Transcript sequence for transcription factor gata-3 is the following:

T_ID	T_FASTA
GSOIDT0000006001	Download sequence

Figure 2.2. **Data mining in OikoBase.** (**A**) Browser window of a 50-kb segment of scaffold 1 showing several genes, with the tailshift expression track selected as an example of a time point. (**B**) By placing the pointer over a gene model (1) (here, the sixth gene from the left) a balloon appears that specifies the gene name, gata-3 (assigned by Blast2GO) and its genomic coordinates (2). (**C**) By clicking on the gene model, a different balloon (3) is opened in its place which allows links to both the GO website (http://www.geneontology.org/) (4) and the InterPro website (http://www.ebi.ac.uk/interpro/). By clicking on either link, the browser is redirected to the relevant GO or InterPro terms associated with this gene which catalogue information regarding biological processes, molecular functions and cellular components as well as evolutionarily conserved protein domains. Options to directly download nucleotide (TRANSCRIPT SEQ) and protein (PEPTIDE SEQ) are also provided in this balloon.

Source: Danks, G., Campsteijn, C., **Parida, M.**, Butcher, S., Doddapaneni, H., Fu, B., . . . Manak, J. R. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res*, *41*(Database issue), D845-853. doi:10.1093/nar/gks1159



The NCBI **BLAST** family of programs includes:

blastp
compares an amino acid query sequence against a protein sequence database

blastn
compares a nucleotide query sequence against a nucleotide sequence database

blastx
compares a nucleotide query sequence translated in all reading frames against a protein sequence database

tblastn
compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames

tblastx
compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Choose program to use and database to search:

Program: blastp; Database: Odioica_GeneModels_peptides_reference_v1.0

Enter sequence below in FASTA format

```
>cytoplasmic actin [Homo sapiens]
MDDDIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQS
KRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMT
QIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDL
AGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSY
ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS
GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQ
```

Or load it from disk

Set subsequence: From To

The query sequence is filtered for low complexity regions by default.

Filter ☑ Low complexity ☐ Mask for lookup table only

Expect: 10; Matrix: BLOSUM62; ☐ Perform ungapped alignment

Query Genetic Codes (blastx only): Standard (1)

Database Genetic Codes (tblast[nx] only): Standard (1)

Frame shift penalty for blastx: No OOF

Other advanced options:

☑ Graphical Overview; Alignment view: Pairwise

Descriptions: 100; Alignments: 50; Color schema: No color schema

```
BLASTP 2.2.23 [Feb-03-2010]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Database: 7_GeneModels_peptides.fa
          17,212 sequences; 4,886,071 total letters

Query= cytoplasmic actin [Homo sapiens]
         (375 letters)
```

Distribution of 25 Blast Hits on the Query Sequence

Mouse-over to show define and scores. Click to show alignments

```
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

[illegible]  actin                                                  748   0.0
[illegible]  beta-actin                                             744   0.0
[illegible]  7 chain molecular models of averaged rigo...           707   0.0
[illegible]  alpha skeletal muscle                                  707   0.0
[illegible]  muscle actin                                           484   e-138
[illegible]  actin                                                  415   e-116
[illegible]  arp1 actin-related protein 1 homolog cent...           413   e-116
[illegible]  arp2 actin-related protein 2 homolog                   345   1e-95
[illegible]  actin                                                  313   5e-86
[illegible]  non-muscle actin                                       286   1e-77
```

Search

Landmark or Region: scaffold_49:183349..185348

Examples: scaffold_1:1..50000, scaffold_10:1..50000, scaffold_50:1..50000, scaffold_100:1..50000, scaffold_255:1..50000, scaffold_500:1..50000, scaffold_1260:1..50000.

Data Source: Oikopleura Dioica Gene Expression Genome Browser

Scroll/Zoom: Show 2 kbp ☐ Flip

Tracks

General ☐ All on ☐ All off

☐ CDS ☑ Named gene

☐ DNA ☐ mRNAs

Figure 2.3. **Obtaining developmental expression profiles for *O. dioica* homologs of genes-of-interest from other organisms.** To provide an example of how to use this pipeline, the human cytoplasmic actin protein has been selected as the query sequence. First, the 'Programs available for the BLAST search' page (1) is accessed by clicking the Facepage BLAST link. The desired Program should then be matched up with the appropriate Database using the dropdown menus (2). Note that selecting 'blastp' (as in this example) or 'blastx' will automatically pair these programs with the '*O. dioica* peptides reference' database. After executing the search (3), the desired putative *O. dioica* homologs are displayed on the BLAST search results page where unique GSOID identifiers and corresponding gene names are provided (4). By clicking on the desired live link GSOIDP identifier the user is taken to the location of the putative homolog (5). Subsequently, any desired additional annotation or transcript tracks can be activated in the browser. Here, the expression of an *O. dioica* cytoplasmic actin homolog is shown at the tailshift stage with the view zoomed to 2 kb.

Source: Danks, G., Campsteijn, C., **Parida, M.**, Butcher, S., Doddapaneni, H., Fu, B., . . . Manak, J. R. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res*, *41*(Database issue), D845-853. doi:10.1093/nar/gks1159

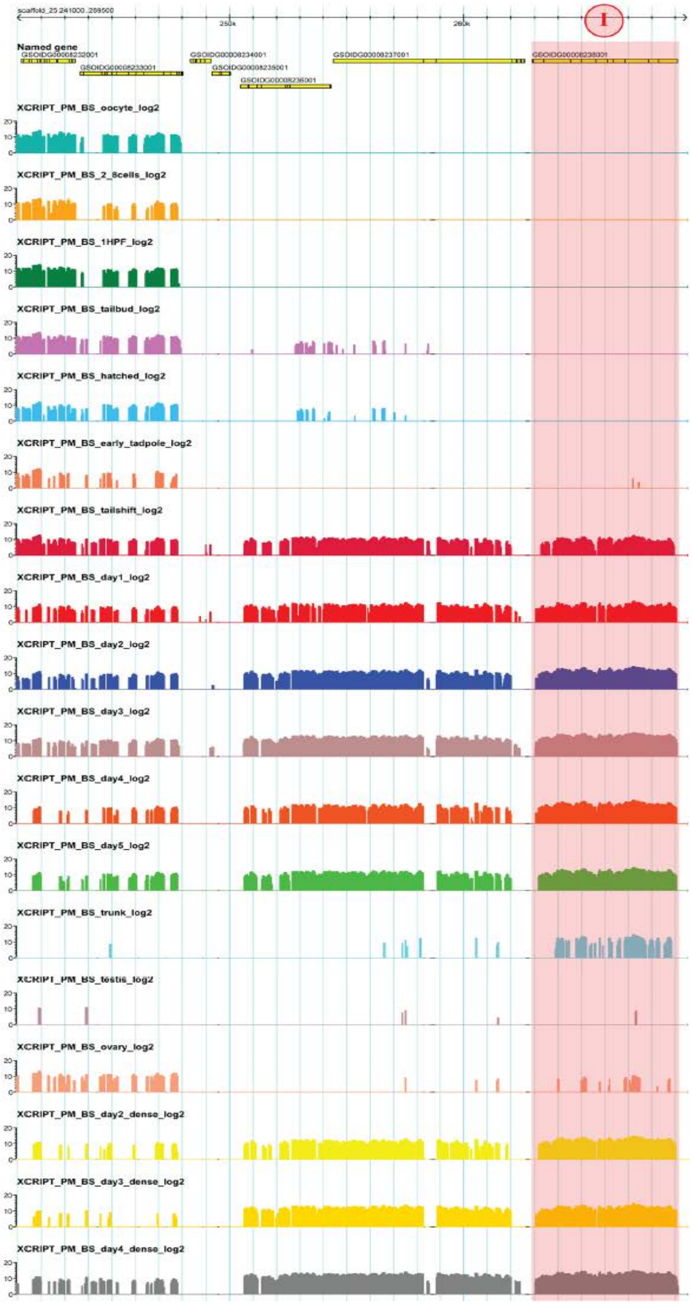


Figure 2.S1. **Characteristic developmental expression profile of a gene, Oikosin 19, involved in repetitive production of the filter-feeding house.** Oikosin 19 contains both EGF and lectin domains and the encoding gene (GSOIDG0008238001; gene model on the right, shaded in pink [I]) is expressed from a specialized oikoplastic epithelium located on the somatic trunk, from tailshift onwards, a time period during which the animal actively filter feeds. The gene is not expressed in the sexual organs but is expressed during developmental stasis conditions (day2-4 dense) when houses continue to be produced.

Source: Danks, G., Campsteijn, C., **Parida, M.**, Butcher, S., Doddapaneni, H., Fu, B., . . . Manak, J. R. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res*, *41*(Database issue), D845-853. doi:10.1093/nar/gks1159

CHAPTER 3. A GENE IMPLICATED IN ACTIVATION OF RETINOIC ACID RECEPTOR TARGETS IS A NOVEL RENAL AGENESIS GENE IN HUMANS.[3]

ABSTRACT

Renal agenesis (RA) is one of the more extreme examples of congenital anomalies of the kidney and urinary tract (CAKUT). Bilateral renal agenesis is almost invariably fatal at birth, and unilateral renal agenesis can lead to future health issues including endstage renal disease. Genetic investigations have identified several gene variants that cause RA, including *EYA1, LHX1, and WT1*. However, whereas compound null mutations of genes encoding a and g retinoic acid receptors (RARs) cause RA in mice, to date there have been no reports of variants in RAR genes causing RA in humans. In this study, we carried out whole exome sequence analysis of two families showing inheritance of an RA phenotype, and in both identified a single candidate gene, *GREB1L*. Analysis of a zebrafish *greb1l* loss-of-function mutant revealed defects in the pronephric kidney just prior to death, and F0 CRISPR/Cas9 mutagenesis of *Greb1l* in the mouse revealed kidney agenesis phenotypes, implicating *Greb1l* in this disorder. *GREB1L* resides in a chromatin complex with RAR members, and our data implicate *GREB1L* as a coactivator for RARs. This study is the first to associate a component of the RAR pathway with renal agenesis in humans.

[3] This chapter has been published in Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

INTRODUCTION

CONGENITAL anomalies of the kidney and urinary tract (CAKUT) are one of the more common sets of birth defects noted in children and represent a significant cause of morbidity and mortality (Sanna-Cherchi et al., 2009), including end-stage renal disease (ESRD) (USRDS, 1999). Renal agenesis (RA) is defined as the complete absence of renal tissue at birth, which can be separated into unilateral and bilateral renal agenesis (Yalavarthy & Parikh, 2003), and represents the most severe form of CAKUT. While unilateral renal agenesis (URA) can lead to proteinuria, hypertension, and early renal failure, it is generally compatible with life (Schreuder et al., 2008). Bilateral Renal Agenesis (BRA), in contrast, is almost invariably fatal at birth (Potter, 1946, 1965). It is estimated that BRA occurs at a frequency of 1/3000–1/5000 births, while URA occurs more frequently (up to 1/1000 births), although estimating the incidence is hampered by underreporting (Norwood & Chevalier, 2003; Yalavarthy & Parikh, 2003). In humans, genetic etiologies for RA were first identified 30 years ago, when it was shown that relatives of a person with a nonsyndromic RA had an increased risk (from 4 to 9%) of having RA themselves (Carter, Evans, & Pescia, 1979; Roodhooft, Birnholz, & Holmes, 1984). At least 70 different clinical conditions or syndromes exist where RA has been identified as a component (Kerecuk, Schreuder, & Woolf, 2008; Sanna-Cherchi et al., 2007), including: branchio-oto-renal (Brophy et al., 2013); hypoparathyroidism, deafness, and renal dysplasia (Van Esch et al., 2000); Townes–Brocks (Kohlhase, Wischermann, Reichenbach, Froster, & Engel, 1998); and Fraser (Vrontou et al., 2003) syndromes. Additionally, variants identified in several genes (including *EYA1, SIX1 and SIX2,* *FRAS1, GATA3, WNT4, RET, FGF20, UPK3A*, and *ITGA8*) have been implicated in human nonsyndromic RA (Barak et al., 2012; Humbert et al., 2014; Jenkins et al., 2005; Sanna-Cherchi et al., 2007; Skinner, Safford, Reeves, Jackson, & Freemerman, 2008; Toka, Toka, Hariri, & Nguyen, 2010).

In mice, variants in a variety of genes have been identified that cause RA, and several of these genes are involved in regulating developmental processes such as nephric duct formation (*Pax2, Lim1*) or ureter budding (*GDNF, Ret, GFR alpha1*) (Uetani & Bouchard, 2009). Additionally, a large number of the RA-associated genes are required for the proper expression of *GDNF* and *Ret/GFR alpha1,* including *Eya1, Six2, Fras1, Gata3, and Emx2* (Uetani & Bouchard, 2009). However, most of the genes identified in monogenic mutant animal models have not yet been correlated with the equivalent human disease. More recently, *ESRRG* (Estrogen Related Receptor Gamma), a gene encoding an estrogen receptor-related nuclear hormone receptor, was implicated in RA based on chromosomal breakpoint analysis in cases affected by RA (Harewood et al., 2010), although targeted inactivation in mice only revealed agenesis of the renal papilla (Berry et al., 2011). Additionally, glomerulonephritis was observed in mice lacking the *estrogen receptor alpha* gene (Shim, Kis, Warner, & Gustafsson, 2004).

Various studies have shown that retinoic acid signaling plays a key role in kidney development. Retinoic acid [which binds a nuclear receptor highly homologous to steroid hormone receptors (Petkovich, Brand, Krust, & Chambon, 1987)], can expand the pronephric region of the kidney in animal cap assays as well as promote expression of many markers of the intermediate mesoderm and its derivatives in mouse embryonic stem cells (D. Kim & Dressler, 2005; Osafune, Nishinakamura, & Komazaki, 2002). Moreover, retinoic acid can promote ureteric bud outgrowth in the developing metanephros (Rosselot et al., 2010), which is thought to work by regulating *Ret* expression in the bud (Batourina et al., 2005). Furthermore, studies in *Xenopus* and zebrafish showed that several genes required for specification and development of the pronephros (*pax8, lhx1, wt1, pteg*) are under the control of retinoic acid signaling (Bollig et al., 2009; Carroll & Vize, 1999; Cartry et al., 2006; S. J. Lee, Kim, Choi, & Han, 2010; Perner, Englert, & Bollig, 2007), and compound null mutations of genes encoding α and γ retinoic acid receptors (RARs) cause a renal agenesis phenotype (Mendelsohn et al., 1994). Finally, a recent report showed that mutation of the *Nuclear Receptor Interacting Protein 1 (NRIP1)* gene, encoding a transcriptional cofactor of retinoic acid receptors, caused a range of CAKUT, including renal hypo/dysplasia and vesicoureteral reflux (VUR) and/or ectopia (Vivante et al., 2017).

Here we describe identification of a novel renal agenesis locus, *GREB1L*, through exome sequence analysis of cases chosen from two independent RA pedigrees, and show that (1) zebrafish *greb1l* is required for proper specification of the pronephros and (2) F0 CRISPR mouse *Greb1l* mutants present with kidney agenesis phenotypes, confirming a role for *GREB1L* in this disorder. *GREB1L* was initially identified as a paralog of *GREB1*, and *GREB1* expression was upregulated upon estrogen treatment of a human breast carcinoma cell line and shown to be highly correlated with both estrogen receptor (ER) and androgen receptor (AR) expression in breast/prostate cancer cell lines and primary tumors (Ghosh, Thompson, & Weigel, 2000; Mohammed et al., 2013; Rae et al., 2005). Notably, *GREB1* (which acts as a coactivator of the ER) resides in a chromatin complex with both *GREB1L* and Retinoic Acid Receptor components. *GREB1L*, on the other hand, is upregulated in a well-established cell line model of retinoic acid signaling (Laursen, Wong, & Gudas, 2012), and mutation of retinoic acid targets expressed in the developing pronephros are associated with RA in mice or humans (Bouchard, Souabni, Mandler, Neubüser, & Busslinger, 2002; Brophy, Ostrom, Lang, & Dressler, 2001; Kreidberg et al., 1993; Meeus et al., 2004; Shawlot & Behringer, 1995; M. Torres, Gómez-Pardo, Dressler, & Gruss, 1995; Trueba et al., 2005). Taken together, these data strongly implicate *GREB1L* as a coactivator for RARs that, when reduced in dose, causes kidney agenesis phenotypes.

MATERIALS AND METHODS

CASE ASCERTAINMENT, IOWA

The Iowa case ascertainment has been carried out on joint projects and replication efforts throughout the world. In 2005, the Brophy laboratory established an Internal Review Board (IRB) approved website for collecting RA samples (IRB # 200711705) (www.kidneygenes.com). Participants and their physicians are made aware of this study through our website, the National Center for Biotechnology Information (NCBI) web resource www.genetests.com, and our work with the National Potter’s Syndrome Support Group. From our website, the participant or their physician downloads appropriate consent forms and other paperwork. This has resulted in a worldwide data and sample collection including in-depth phenotypic, clinical, and genetic material. Dr. Michael Schneider while at the University of Southern Illinois originally brought the proband of the family included in this study to our attention. Adult family members voluntarily filled out a health questionnaire that collected their personal health history as well as that of their extended family history in an anonymous manner. Through this method, additional potentially affected family members and their immediate relatives were identified. Members who were enrolled were asked, at their own discretion, to reach out to additional family members to inquire about participating. Members who were willing to participate contacted us and were enrolled through their local medical provider.

CASE ASCERTAINMENT, DENMARK

The Danish cases were ascertained as part of a project on prenatally diagnosed kidney anomalies. Data on pre- and postnatal findings in the families were collected as well as DNA. The second affected fetus from family 2 was analyzed by our in-house-designed kidney-gene-targeted panel including 108 genes associated with kidney disease. This analysis did not reveal any disease-causing variants. The family was therefore selected for novel kidney-gene discovery using whole exome sequencing.

CASE PHENOTYPES AND SAMPLES, IOWA

Kidney ultrasound, MRI, and intravenous pyelogram examination revealed URA as well as hypertrophy of the left kidney for II-5, and URA for II-7 (Figure 3.1A). Additionally, kidney ultrasound revealed URA for III-3 and III-4. *In utero* kidney ultrasound revealed BRA for one family member (III-6), and *in utero* kidney ultrasound as well as MRI revealed BRA for another (III-8). BRA was suspected in two family members (II-1, II-6) but not confirmed. Four confirmed affected family members with either URA or BRA (II-5, II-7, III-4, III-8) were selected for whole exome sequencing. The Institutional Review Board of the Carver College of Medicine, University of Iowa, approved this study, and all participants provided written consent in addition to DNA samples after being properly counseled regarding the potential of incidental findings from whole exome sequencing.

CASE PHENOTYPES AND SAMPLES, DENMARK

Two pregnancies with affected fetuses presenting with BRA (indicated in Figure 3.1B) were terminated following parental request and approval by the regional abortion committee. Following the second termination, the parents had a renal ultrasound, which showed left-sided URA in the mother. The father had normal kidneys. Subsequently, the mother’s parents and brother had renal ultrasound examinations, which showed normal kidneys. Prenatal ultrasound examinations of the three live-born healthy brothers were unremarkable. The second affected fetus, the affected mother, and the unaffected father as well as the unaffected maternal grandparents were selected for exome sequencing analysis. Blood samples were obtained from the adult family members. Subsequently, buccal smear samples were obtained from the live-born brothers.

The Danish National Committee of Ethics approved the whole exome sequencing study, and written informed consent was obtained for all four adult family members included in the study after being properly counseled regarding the potential of incidental findings from whole exome sequencing.

EXOME SEQUENCING ANALYSIS, IOWA

Genomic DNA was obtained from either lymphocytes isolated from whole blood samples or from tissue samples obtained during autopsy using standard laboratory methods. The genomic DNA samples from four affected individuals (II-5, II-7, III-4, III-8; Figure 3.1A) were prepared for WES (whole exome sequencing) using the Illumina paired-end sample prep kit (Illumina, San Diego, CA) and captured using the Nimblegen SeqCap EZ Human Exome Library v2.0 kit (Roche NimbleGen Inc, Madison, WI) following the manufacturer's instructions. Captured samples were sequenced using Illumina HiSeq100-bp paired-end sequencing (Duke Center for Genomic and Computational Biology, Ontario Institute for Cancer Research). Next, quality control was performed using FastQC software (Andrews, 2011). Reads with average quality scores <20 were trimmed using the Burrows Wheeler Aligner (BWA) (H. Li & Durbin, 2009) and reads <35 bp were not used for the downstream analysis. Reads were mapped to the human reference genome (version-glk_v37) using BWA. Mapping statistics of the aligned reads and coverage of exome target regions were analyzed using Qualimap software (http://qualimap.bioinfo.cipf.es/) (García-Alcalde et al., 2012) and BEDtools (Quinlan & Hall, 2010) (Table 3.1).

Local realignment and base quality score recalibration was performed using Genome Analysis Toolkit (GATK) (http://www.broadinstitute.org/gatk/) (McKenna et al., 2010), and fixing mate information and marking duplicates was performed using Picard tools (http://picard.sourceforge.net). Finally, Unified Genotyper was used to call genetic variants in standard Variant Call Format. Variants were annotated using SnpEff

(Cingolani et al., 2012) software, the University of California, Santa Cruz, human reference genome assembly hg19, and dbSNP 137 (Sherry et al., 2001). Additionally, minor allele frequencies (MAF) for all variants were generated using two databases, the 1000 Genomes Project and the National Heart Lung Blood Institute Exome Sequencing Project (NHLBI-ESP) using the wANNOVAR web server (http://wannovar.usc.edu/) (Chang & Wang, 2012). We then applied GATK's best practices of variant quality and coverage thresholds to account for false positive variant calls. A genotype filter was applied to exclude variants with diverse genotypes across all samples. Assuming that variants involved in causing Mendelian disorders would be rare in nature, we excluded variants that had an MAF ≥1% in 1000 Genomes and NHLBI-ESP. Moreover, we also excluded those variants that had an MAF tag of >5% in their dbSNP 137 annotations.

Lastly, we checked the effects of amino acid substitution on protein structure using the database of human nonsynonymous SNPs and function predictions (dbNSFP v2.0) (https://sites.google.com/site/jpopgen/dbNSFP) (X. Liu, Jian, & Boerwinkle, 2013). We focused our analysis on both Polymorphism Phenotyping version 2 (PolyPhen-2) (http://genetics.bwh.harvard.edu/pph2/) and Separating Intolerant from Tolerant (SIFT) (http://sift.jcvi.org/). The CONsensus DELeteriousness (CONDEL) program was then used to generate the weighted average of the normalized scores from PolyPhen-2 and SIFT (http://bg.upf.edu/fannsdb/) (González-Pérez & López-Bigas, 2011). The deleterious variants based on CONDEL predictions were retained in our final list for downstream analysis. Directed Sanger sequencing (carried out by the Iowa Institute of Human Genetics, Genomics Division, University of Iowa Carver College of Medicine) along with a TaqMan Allelic Discrimination Assay (Applied Biosystems) was then used to determine which of these variants showed the predicted segregation pattern for an etiologic variant.

EXOME SEQUENCING ANALYSIS, DENMARK

Genomic DNA was extracted from cultured fetal fibroblasts, formalin-fixed paraffin-embedded fetal tissue, whole blood samples, and buccal smear samples using standard laboratory methods. DNA from two affected and three unaffected family members (Figure 3.1B) was prepared for WES using the KAPA HTP Library Preparation Kit (KAPA Biosystems Inc, Wilmington, MA) and captured using the SeqCap EZ MedExome Kit (Roche NimbleGen Inc, Madison, WI) according to the manufacturer's instructions. Next, Illumina NextSequation 500 sequencing was used to generate paired-end reads (carried out by the Department of Molecular Medicine, Aarhus University Hospital, Denmark).

Reads were aligned to the human reference genome (GRCh37/hg19) and variants were called and annotated in coding exons ± 10 bp using Biomedical Genomics Workbench version 2.0 (CLC bio-Qiagen, Aarhus, Denmark). Standard settings on QIAGEN's Ingenuity Variant Analysis (www.qiagen.com/ingenuity) software were used for data analysis. False positive variant calls were removed based on default coverage and quality thresholds. Variants with an MAF $\geq$0.1% in the 1000 Genomes Project, the National Heart Lung Blood Institute Exome Sequencing Project (NHLBI-ESP), the Allele Frequency Community, and the Exome Aggregation Consortium (ExAC) were excluded. Variants predicted deleterious and listed in the Human Gene Mutation

Database were retained. A filter was applied to retain variants present in heterozygous form in the affected mother and the second affected fetus. Finally, a filter was applied to retrain variants present in heterozygous form only in the affected family members but not present in unaffected family members. The variants were confirmed by direct sequencing using BigDye Terminator v1.1 Cycle Sequencing Kit according to the description of the manufacturer (Applied Biosystems, Life Technology) and analyzed using ABI 3500xl Genetic Analyzer (Applied Biosystems, Foster City, CA). Additionally, Sanger sequencing tested in the female fetus and in the live-born brothers the presence of the variant. Primer sequences and PCR details are available upon request.

ZEBRAFISH ANALYSES

UNIVERSITY OF IOWA:

Zebrafish embryos and adults were reared as described previously (Westerfield & Zfin, 2000), in the University of Iowa Zebrafish Facility, which is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International, following procedures approved by the University of Iowa's Institutional Animal Care and Use Committee (IACUC). Embryos were staged by hours or days post fertilization (hpf or dpf) at 28.5° (Kimmel, Ballard, Kimmel, Ullmann, & Schilling, 1995). Homozygous *greb1l* mutant embryos were generated from heterozygous adults of the sa1260 allele obtained from the Zebrafish International Resource Center, Eugene, OR.

To inhibit *grebl1* expression we ordered an antisense morpholino oligonucleotide (MO) targeting the exon 3–intron 3 junction (sequence: 5′-TATTGGAACACCAACCTAAAAGTGC-3′) (Gene Tools, Philomath, OR). To test efficacy of the MO, we harvested RNA (separately) from embryos injected with the control MO or with the *greb1l* MO, generated first-strand complementary DNA (cDNA), and carried out PCR with primers flanking the splice junction on both cDNA templates. The band of expected size was found in both templates, but in the *greb1l* MO template an additional larger band was present. Sequence from both products confirmed the smaller band corresponded to correctly spliced RNA and the larger band to RNA in which the third intron was unspliced. To mutate the *grebl1* gene with CRISPR/Cas9 we used the website https://chopchop.rc.fas.harvard.edu/index.php to identify a high-scoring guide RNA target site. An oligo specific to exon 17 of the zebrafish grebl1 gene was selected. The target site (GGTCCACACAAAAATGG) was synthesized (Integrated DNA Technologies; IDT) with the T7 promoter sequence on the 5′ end and a 20-bp overlap at the 3′ end complementary to the generic Cas9-binding scaffold oligo. The guide sequence oligo was then annealed to the generic 119-bp Cas9-binding scaffold oligo as described in the cloning-free method of generating single-stranded guide RNA (sgRNA) (Talbot & Amacher, 2014). Once annealed, this product provides a DNA template complete with T7 promoter for *in vitro* synthesis of a sgRNA. We co-injected 1- to 2-cell-stage embryos with sgRNA (200–400 pg per embryo) and/or Cas9 protein (IDT) at 2 ng.

For *in situ* hybridization, 592 bp of *greb1l* cDNA was amplified from 24 hpf wild-type zebrafish first-strand cDNA using the following primers: forward, 5′-GTCAAGCAGGAAAAGATCTGC-3′; reverse, 5′-GGAACGATCGGTAATGTCTT-3′. The cDNA was engineered into the StrataClone vector (Agilent Technologies, Santa Clara, CA) and a DIG-labeled, antisense RNA probe was generated by *in vitro* transcription (Roche Diagnostics, Indianapolis, IN). Whole-mount *in situ* hybridization was carried out following procedures described previously (Thisse & Thisse, 2008). For immunohistochemistry, a monoclonal anti-ATPase, [Na(+) K(+)] α-1 subunit antibody (a6F, Developmental Studies Hybridoma Bank at the University of Iowa), was used at a 1:100 dilution. Following primary antibody incubation for 48 hr, the embryos were blocked and then incubated with an Alexa-488 conjugated goat-anti-mouse secondary antibody for 48 hr.

MAYO CLINIC:

Zebrafish procedures were approved by the Mayo Clinic's IACUC. Embryos developed at 28.5° with 0.003% 1-phenyl-2-thiourea (Sigma-Aldrich) added at 24 hpf to prevent pigmentation for facilitating cyst visualization. Embryos were anesthetized using 0.02% tricaine (Aquatic Habitats) before observation by microscopy. Embryos were examined for cysts at 2 and 3 dpf using a Zeiss Lumar stereo fluorescence microscope and Zen software.

GENERATION OF CRISPR MUTAGENIZED F0 EMBRYOS:

A guide sequence, GTTTATATGAGGCATGTTGA, targeting the orthologous region in mouse to the L1793R mutation was synthesized as an Ultramer (IDT) with the guide sequence embedded between the T7 promoter and portion of stem loop as described in Bassett, Tibbit, Ponting, and Liu (2013). The resulting DNA template was column purified (QIAGEN) prior to *in vitro* transcription reaction, column purified (Zymogen), and quantified via Nanodrop before microinjection. A single-stranded donor oligonucleotide (ssODN) was designed to introduce the desired T > G point mutation to create L1793R missense mutation and also included a silent C > T substitution to ablate the PAM sequence. The ssODN was synthesized as a 125-bp Ultramer (IDT) with the introduced base pair changes underlined: ATCCTGCCCCTTCAGTACGTCTGCGCCCCTGACAGTGAACACACACTCCTGGCAGCCCCTGCACAGTTCCTCCTGGAGAAGTTTCGTCAACATGCCTCATATAAACTCTTCCCTAAAGCCATCCA. One-cell embryos were obtained from superovulated C57BL/6NJ (B6NJ; JAX stock number 5304) female donors crossed to B6NJ males. For microinjection, reagents were injected into the pronucleus at the following concentrations: 30 ng/µl Cas9 mRNA (Trilink); 15 ng/µl sgRNA; and 20 ng/µl ssODN. Embryos were collected at E15.5 and processed for microCT as described in Dickinson et al. (2016). PCR genotyping was performed on tail tip DNA using primers flanking the region of interest: Greb1l-GT-F TGACAGGCACATCTCCCATG and Greb1l-GT-R TCCAAGTCATCAAGGCAGGC that generate a 433-bp product. Individual genotypes were first assessed using Sanger sequencing and subsequently confirmed by T/A cloning and sequencing of at least eight independent clones of tail tip DNA for each putative mutant.

DATA AVAILABILITY

Supplementary data such as Figure 3.S1, 3.S2 and 3.S3, compare**s** sequences between human and zebrafish *GREB1L* protein**s**, compares kidney phenotypes in zebrafish for *greb1l* morpholino knockdown and CRISPR-Cas9 deletion and show**s** the sequences of mutagenized alleles recovered from CRISPR F0 mouse embryos. This study was approved by the University of Iowa under IRB#200711705 as well as by the Danish

National Committee of Ethics. WES data is available in the Sequence Read Archive database (accession number SRP112780).

RESULTS

EXOME SEQUENCE ANALYSIS OF TWO PEDIGREES REVEALS *GREB1L* AS AN RA GENE

Four URA/BRA family members (II-5, II-7, III-4, III-8) from pedigree 1, which suggests autosomal dominant inheritance of the RA phenotype (Figure 3.1A), were selected for WES analysis. Across the four samples, we achieved an average targeted exome coverage of 172× with a mean mapping quality of 45.30 for calling high-quality variants (Table 3.1). We focused on identifying variants shared by all four cases, and this revealed heterozygosity for novel missense variants in three genes (*LHX9* c.1127 C > T, *GYLTL1B* c.442 T > A, *GREB1L* c.5378 T > G) as well as heterozygosity for a novel stop-loss variant (*CLEC9A* c.724 T > C) in *CLEC9A* (no novel variants showed homozygosity or compound heterozygosity shared by all four affected family members). PolyPhen-2, SIFT, and CONDEL analyses predicted all three missense variants to be damaging. Directed Sanger sequencing along with a TaqMan Allelic Discrimination Assay (Applied Biosystems) revealed that only the *GREB1L* variant showed the predicted segregation pattern for an etiologic variant: six affected family members (II-5, II-7, III-3, III-4, III-6, and III-8) all harbored the variant, while seven unaffected family members (I-3, I-5, II-3, II-4, III-1, III-2, III-5) lacked the variant; female II-2 was hypothesized to be a carrier of the *GREB1L* variant exhibiting incomplete penetrance, and presence of the variant was confirmed (Figure 3.1A). This missense variant, which was absent from the

ExAC database, changes a conserved leucine to arginine in the highly conserved c-terminus of the protein (see Figure 3.2 and Figure 3.S1).

Two URA/BRA affected (II-2, III-2) and three unaffected family members (I-1, I-2, II-1) from pedigree 2 (Figure 3.1B) were selected for WES analysis. We achieved a mean target region coverage of 119× and mapping quality of 61 for calling high-quality variants (Table 3.S2). Ingenuity Variant Analysis revealed two variants that were present only in affected family members, i.e., *GREB1L* c.5608 + 1delG and *FAM21C* c.1837G > C. The *FAM21C* missense variant was predicted to be tolerated and benign by SIFT and PolyPhen-2, respectively, with the base at that position being weakly conserved. The *GREB1L* variant is novel and deletes one of two G residues located at the splice donor site of the last intron (the transcribed wild-type RNA sequence reads AAAG at the 3′ end of the exon followed by GUAA at the 5′ end of the intron). Both G residues represent highly conserved nucleotides involved in splicing and therefore there are two potential effects of a single G at the splice donor site: first, a novel splice site could be created, shifted by 1 bp, resulting in the protein sequence as depicted in Figure 3.2 lter the c-terminus of the protein; and second, splicing efficiency could be diminished, causing partial intronic read through of nonspliced messenger RNA prior to encountering a stop codon during translation. Importantly, in either scenario, the highly conserved c-terminus of the protein encoded by the last exon would be deleted.

Sanger sequencing identified the *GREB1L* variant in the other affected fetus as well as the two eldest healthy live-born brothers. Additionally, based on the exome sequencing data, no disease-associated copy number variants were identified.

GREB1L IS A HAPLOINSUFFICIENCY LOCUS

Large-scale microarray and sequencing studies have helped elucidate numerous haploinsufficient or variation-intolerant regions within the genome (Petrovski, Wang, Heinzen, Allen, & Goldstein, 2013; Ruderfer et al., 2016; Zarrei, MacDonald, Merico, & Scherer, 2015), and interrogation of the Database of Genomic Variants (DGV; http://dgv.tcag.ca) finds only two deletion events in control populations that involve coding portions of *GREB1L*. This is very significant given that as of early 2017, the DGV had identified over six million-sample level CNVs from control populations of over 70 studies. Further supporting this claim is the absence of any *GREB1L* coding deletion CNVs within the DECIPHER database and only one within the ClinGen resource (https://decipher.sanger.ac.uk/; https://www.clinicalgenome.org/). Finally, there are no deletions involving *GREB1L* noted in the CNV calls from the ExAC database (http://exac.broadinstitute.org/) and the gene itself is predicted to be haploinsufficient (%HI 9.33 reported by DECIPHER) (N. Huang, Lee, Marcotte, & Hurles, 2010). Given that most of the copy number variable regions of the genome are pericentromeric, the lack of such variation within the *GREB1L* gene is also significant due to its proximity to the centromere of chromosome 18 (Iafrate et al., 2004; Redon et al., 2006; Sebat et al., 2004; Zarrei et al., 2015). Taken together, these data identify *GREB1L* as a likely haploinsufficiency locus.

GREB1L IS EXPRESSED IN THE DEVELOPING KIDNEY

To determine whether *GREB1L* is expressed during genitourinary development, we accessed gene expression microarray data cataloged in GUDMAP (Genitourinary

Development Molecular Anatomy Project) (Harding, Armit, Armstrong, Brennan, Cheng, Haggarty, Houghton, Lloyd-MacGilp, Pi, Roochun, Sharghi, Tindal, McMahon, Gottesman, Little, Georgas, Aronow, Steven Potter, et al., 2011; McMahon et al., 2008). These results revealed that *GREB1L* is expressed primarily in the early proximal tubule as well as metanephric mesenchyme, and also in the ureteric bud (Georgas et al. 2009). However, most kidney expression studies have been performed during morphogenesis of the metanephric kidney, and thus information on factors associated with early pronephric specification is lacking. Since early events in kidney development are strongly conserved between zebrafish and mammals (Drummond, 2005; Drummond & Davidson, 2016a), we thus turned to the zebrafish model to explore *greb1l* expression patterns and function during development.

PRONEPHRIC KIDNEY DEVELOPMENT IS ALTERED IN ZEBRAFISH *GREB1L* LOSS-OF-FUNCTION MUTANTS AND *GREB1L*-DEPLETED EMBRYOS

The single ortholog of *GREB1L* in the zebrafish genome, *greb1l*, is predicted to encode a protein that is 61% identical and 73% similar to the human ortholog (Figure 3.S1). Whole-mount *in situ* hybridization on wild-type embryos at 90% epiboly (8.5 hpf) and early somitogenesis stages (11.5 hpf) revealed expression of *greb1l* in the mesoderm, including the intermediate mesoderm, the origin of the pronephros (Figure 3.3, A and B).

An N-ethyl-N-nitrosourea (ENU)-induced T to A substitution in the *grebl1* gene, the sa1260 allele which introduces a stop codon at amino acid 1915 of the 1942 residue full-length protein, was isolated by large-scale screening (Kettleborough et al., 2013). Heterozygotes for the sa1260 allele are morphologically normal and are fertile.

Homozygotes for this allele (hereafter, *greb1l* mutants) were readily recognized at 3 dpf by periorbital edema and pericardial effusions, consistent with a defect in ion and fluid homeostasis (16 embryos in a clutch of 109 embryos showed this phenotype and were all found to be homozygous mutants by sequencing) (Figure 3.4B). At 2 dpf, the developing kidney is readily discernible in normal living embryos. In most of the clutch, presumed to be wild types or heterozygous mutants, the pronephric tubule appeared normal (Figure 3.4C). By contrast, in *greb1l* mutants, tubules were dilated and kinked (n = 16) (Figure 3.4D). At 3 dpf, dilation was more pronounced, and the majority of the mutants had obvious cysts (15 of 16 mutants, Figure 3.4F). In embryos processed to reveal immunoreactivity of anti-Na/K ATPase antibody (a6f), an early marker of pronephric mesoderm, dilation of the proximal straight tubule in mutants in comparison to nonmutant siblings was evident (Figure 3.5, A and B). In all wild types examined, there was a characteristic hairpin turn between the proximal convoluted tubule and the neck region of the pronephros (n = 12, shown at 6 dpf, Figure 3.5C). In all *greb1l* mutants examined, this region of the pronephros is serpentine (n = 14, Figure 3.5D). *greb1l* mutants died between 10 and 12 dpf.

Because the *greb1l* sa1260 mutant allele came from a chemical mutagenesis screen, it is conceivable that there are other mutations cosegregating with the *greb1l* mutation. To confirm that the abnormal kidney phenotype in sa1260 mutants results from the mutation in *greb1l* we reduced *greb1l* expression by injecting wild-type embryos with antisense MO targeting an early splice junction, or with control MO. We used RT-PCR and sequencing to confirm the MO was effective at inhibiting splicing of *greb1l* (see

Materials and Methods). Additionally, we employed CRISPR technology by injecting wild-type embryos with Cas9 protein and a guide RNA targeting an evolutionarily conserved exon of *greb1l*; this is expected to yield mosaic embryos in which a variable frequency of cells experience a biallelic mutation in *greb1l* (Talbot & Amacher, 2014). In both cases we fixed injected embryos at 4 dpf and processed them to reveal anti-Na/K ATPase immunoreactivity. The large majority of embryos injected with *greb1l* MO exhibited the abnormal morphology of the proximal kidney seen in sa1260 mutants (Figure 3.S2). Moreover, ~30% of F0 embryos injected with *greb1l* gRNA with Cas9 protein exhibited the proximal kidney defects, suggesting mosaic, biallelic mutation of *greb1l* is sufficient to yield this phenotype; notably, the efficiency of phenotypic penetrance in CRISPR/Cas9-injected F0 embryos is comparable to that seen by other groups targeting other genes (Jao, Wente, & Chen, 2013). The convergent phenotype of *greb1l* mutants, embryos injected with MO targeting *greb1l*, and CRISPR/Cas9 reagents targeting *greb1l* strongly support a requirement for *Greb1l* in kidney morphogenesis in zebrafish and are consistent with a role for *GREB1L* in morphogenesis of the human kidney.

F0 CRISPR MOUSE *GREB1L* MUTANTS DISPLAY KIDNEY AGENESIS PHENOTYPES

The single ortholog of mouse *GREB1L*, *Greb1l*, is predicted to encode a protein that is 90% identical and 94% similar to the human ortholog. To test whether *Greb1l* plays a similar role in a mammalian model, we generated F0 mutant embryos in mice using CRISPR/Cas9, targeting mouse exon 31 and an ssODN designed to introduce the orthologous L1793R Iowa mutation (Figure 3.6A). Our F0 approach has previously been shown to faithfully recapitulate human disease phenotypes despite the mosaicism intrinsic to the genome editing process, and that we can establish a clear and robust genotype–phenotype relationship (Guimier et al., 2015). Our microinjections produced 56 F0 embryos that were subsequently analyzed at E15.5. Of these, we identified 9 (16%) embryos showing evidence of CRISPR/Cas9 mutagenesis with 3 (33%) phenotypically affected mutants that displayed a range of gross phenotypes including exencephaly and craniofacial dysmorphology including unilateral and bilateral cleft lip (Figure 3.6, B–D). For the mutant embryos, we took advantage of our high-throughput microCT imaging platform established for the Knockout Mouse Phenotyping Program (KOMP2) (Dickinson et al., 2016) to examine developmental kidney defects. In two of the affected embryos, we observed unilateral agenesis, and bilateral agenesis in the third (Figure 3.6, E–G). Notably, in each case of unilateral agenesis, the contralateral kidney also appears abnormal or incompletely developed. To determine the nature of the CRISPR-induced mutations, we cloned and sequenced the mutations of all embryos showing evidence of CRISPR activity and confirmed that all three embryos with RA phenotypes harbored mutations in *Greb1l*. Two affected embryos carried knock-in alleles harboring the L1793R mutation along with an accompanying in-frame deletion removing a conserved glutamine residue adjacent to L1793 (Figure 3.S3). The other affected embryo was homozygous for a 2-bp insertion resulting in a frameshift mutation and stop codon 33 amino acids downstream. The remaining six mutagenized embryos showed evidence of nonhomologous end joining (NHEJ)-induced indels but also contained wild-type alleles suggesting only partial impairment of *Greb1l* function. In summary, these phenotypes are consistent with an essential role for *Greb1l* in kidney development and suggest additional roles for *Greb1l* during embryonic development.

DISCUSSION

GREB1L CODING VARIANTS ARE ASSOCIATED WITH RA IN TWO FAMILIES

The Iowa pedigree structure (Figure 3.1A) is consistent with autosomal dominant inheritance of the RA phenotype, and the *GREB1L* missense variant was the only variant to cosegregate with the phenotype in all cases tested (six) but none of the unaffected family members tested (seven) except the female carrier (II-2, Figure 3.1A) exhibiting incomplete penetrance. The odds of such a segregation pattern occurring by chance is 1 in over 16,000. The variant was called as damaging by SIFT/PolyPhen-2/CONDEL and alters a conserved residue in a highly conserved domain in the c-terminus of the protein.

In the Danish family, both the *GREB1L* frameshift variant as well as a missense variant of *FAM21C* had arisen *de novo* in the affected mother and was found in one affected fetus. The *FAM21C* variant was predicted to be tolerated/benign, while the *GREB1L* variant causes a profound alteration of the c-terminus (notably, the same region affected by the Iowa variant). The *GREB1L* variant was also found in the second affected fetus (and thus all three affected family members), along with two unaffected brothers. Collectively, these data reveal that the *GREB1L* variants identified in both the Iowa and Danish families are likely the etiologic variants causing the RA phenotype with incomplete penetrance.

GREB1L EXPRESSION PATTERN SUPPORTS ITS ROLE IN KIDNEY MORPHOGENESIS

Transcriptome analysis shows that *GREB1L* is expressed in a variety of tissues (www.genecards.org), with particularly high expression in brain, kidney, and ovary (GEO accession number GSM35549, profile GDS3052; Hildner et al. (2008)). *GREB1L* is also robustly expressed in the early proximal tubule and metanephric mesenchyme of the metanephric kidney, with lower expression levels seen in the ureteric bud (Georgas et al., 2009). Since these expression studies were performed on the developing metanephric kidney but not on earlier morphogenetic events, we performed *in situ* hybridization of *greb1l* in the developing zebrafish embryo and found it to be expressed in the portion of the intermediate mesoderm that gives rise to the pronephros (Figure 3.3, A and B). Since both *GREB1L* human variants (as well as the variants in zebrafish and mouse *Greb1l*) alter the c-terminus of the protein (Figure 3.2 and Figures 3.S1 and 3.S3), it is possible that this region may be associated with a kidney-specific function and that its alteration could produce a dominant effect. Both variants are located in one of the most highly conserved domains of the protein when comparing the *GREB1* and *GREB1L* paralogs, or the *GREB1L* human and zebrafish orthologs (amino acid identity 61%). The Iowa missense variant alters a residue residing in a stretch of 24/27 (89%) conserved amino acids across paralogs as well as between human and zebrafish *GREB1L*, while the Danish frameshift variant deletes a region of 47/54 (87%) conserved amino acids across paralogs and 50/54 (93%) across orthologs. However, given that premature stop codons produced by both the Danish and mouse variants would probably lead to nonsense-mediated decay, it is more likely that these variants are loss-of-function, which would effectively reduce

GREB1L gene dosage to half. Consistent with this idea, we have found that *GREB1L* is likely to be a haploinsufficiency gene. This might also explain the range of observed phenotypes (two kidneys, URA, BRA) observed in the pedigrees, with the reduced gene dosage effectively creating a “teeter-totter” scenario of stochastic developmental decisions that either result in the morphogenesis of a mature kidney, or no kidney at all.

GREB1L IS A LIKELY COFACTOR FOR STEROID HORMONE/RARS

Although UniProt predicts *GREB1L* to be a single-pass membrane protein due to a predicted membrane-associated domain, we do not favor this hypothesis. Its paralog, *GREB1* (54% identical and 67% similar to *GREB1L*), was shown to be a nuclear chromatin-bound ER coactivator that is (1) upregulated after estrogen treatment and (2) essential for ER-mediated transcription (Mohammed et al., 2013; Rae et al., 2005). Importantly, *GREB1L* and retinoic acid receptor members are part of the ER/GREB1 chromatin complex (Mohammed et al., 2013). Consistent with *GREB1L* playing a similar coactivator role, but in concert with RARs, retinoic acid treatment of F9 embryonal carcinoma stem cells (a well-established model for retinoic acid signaling) was shown to robustly upregulate *GREB1L* (Laursen et al., 2012), which would then be predicted to bind and activate RARs. Intriguingly, RNAi knockdown of *GREB1* in cell lines was shown to block estrogen-induced growth (Rae et al., 2005), produce a G0/1 arrest with increased G1 DNA content (Kittler et al., 2007), and decrease cell viability after treatment with Paclitaxel (Sinnott et al., 2014; Whitehurst et al., 2007), suggesting that GREB proteins might be playing a role in mediating cell growth.

ZEBRAFISH *GREBL1* IS REQUIRED FOR PROPER MORPHOGENESIS OF THE PRONEPHROS

Whole-mount ISH of zebrafish embryos revealed widespread labeling of the mesoderm during early somitogenesis, an area that includes the intermediate mesoderm that gives rise to the pronephros and also expresses the pronephric markers *wt1a, wt1b, pax2a, pax8*, and *lhx1a* at similar stages of development (Bollig et al., 2006; Drummond & Davidson, 2016a; Perner et al., 2007). It is worth noting that these genes are the orthologs of the genes that pattern the mammalian pronephros (see below), demonstrating the evolutionary conservation of pronephros specification and relevance of the zebrafish model.

Zebrafish *greb1l* mutants had abnormal pronephric morphology and evidence of altered function, including presence of cysts and dilated tubules evident by 2 dpf when the pronephros begins filtering (Drummond et al., 1998). Later, the mutants developed edema and disrupted proximal tubule convolution, phenotypes that could result from/contribute to defects in fluid and ion transport (Vasilyev, Liu, Hellman, Pathak, & Drummond, 2012; Vasilyev et al., 2009) thus contributing to death of the mutants. In particular, loss of fluid flow leads to fluid accumulation and organ distension including pronephric cysts and tubule dilation (Kramer-Zucker, Wiessner, Jensen, & Drummond, 2005). Since *greb1l* zebrafish mutants die just prior to the time when the mesonephros can be reliably detected (Diep et al., 2015), we were unable to assess development of the mesonephros, the mature kidney of the zebrafish. Nonetheless, these data point toward a role of *greb1*l in controlling early pronephros specification/morphogenesis and ultimately function.

F0 CRISPR *GREB1L* MOUSE MUTANTS PRESENT WITH URA AND BRA PHENOTYPES

The high efficiency of Cas9 coupled with the short gestation period of the mouse provides a significant opportunity to functionally validate discoveries uncovered from WES efforts of human cohorts. Here, we demonstrate this powerful combination using CRISPR/Cas9-mediated genome editing to model a novel point mutation in *GREB1L* directly in F0 mouse embryos, thereby removing the traditional constraints of establishing animal lines and performing timed matings. The mutagenized embryos all displayed a spectrum of kidney abnormalities ranging from URA to BRA, confirming the causative etiology of the human mutations in RA. Additionally, several craniofacial abnormalities were observed in the affected embryos, highlighting a critical and more widespread role for *Greb1l* during embryonic development. While two of the three mutants harbored nonnull allelic combinations, the third was homozygous for a frameshift mutation, consistent with the highly conserved nature of the mutated residue and c-terminal domain of the *GREB1L* protein. These findings along with current advances in genome editing hold great promise for the future of performing rapid and precise modeling of human developmental disorders in a mammalian system.

GREB1L MAY MEDIATE PROLIFERATION AND INDUCTIVE EVENTS IN EARLY KIDNEY DEVELOPMENT

In addition to connections between estrogen/estrogen-related nuclear steroid hormone receptors and kidney morphogenesis (Berry et al., 2011; Harewood et al., 2010; Shim et al., 2004), retinoic acid also plays key roles in genitourinary system development, including promoting early pronephric kidney morphogenesis (Bollig et al., 2009; Carroll & Vize, 1999; Cartry et al., 2006; D. Kim & Dressler, 2005; S. J. Lee et al., 2010; Osafune et al., 2002; Perner et al., 2007) as well as metanephros development (Vilar, Gilbert, Moreau, & Merlet-Bénichou, 1996), and absence of α and γ RARs results in murine renal agenesis (Mendelsohn et al., 1994). The hypothesis that *GREB1L* may be promoting kidney development through activation of RARs is particularly attractive, since several vertebrate genes required for pronephros specification and development in fish, frogs, and/or mice (*Pax2, Pax8, Lhx1, Wt1*, and *Pteg*; mouse abbreviations used for clarity) are under the control of retinoic acid signaling (Bollig et al., 2009; Carroll & Vize, 1999; Cartry et al., 2006; S. J. Lee et al., 2010; Perner et al., 2007), to determine rostral/caudal and multiciliated/transporting epithelial cell fate (Cheng & Wingert, 2015; Y. Li, Cheng, Verdun, & Wingert, 2014; Marra & Wingert, 2016; Wingert et al., 2007), and mouse or human studies have associated several of these early expressed genes (*Pax2, Pax8, Lhx1*, and *Wt1*) with RA phenotypes (Bouchard et al., 2002; Brophy et al., 2001; Kreidberg et al., 1993; Meeus et al., 2004; Shawlot & Behringer, 1995; M. Torres et al., 1995; Trueba et al., 2005). Collectively, these data suggest a mechanism whereby retinoic acid signaling activates *GREB1L* expression, which in turn allows interaction of *GREB1L* and RARs, both of which are required for robust activation of the pronephros patterning genes. Failure to properly activate GREB1L expression, or expression of the retinoic acid-responsive *PAX2/8, LHX1*, and *WT1* targets, could then lead to RA phenotypes.

The pronephros is formed from intermediate mesoderm (where *greb1l* expression is observed in the zebrafish), and although it is considered a rudimentary structure that will be temporarily replaced by the mesonephros, studies have demonstrated that the pronephric duct is essential for promoting both mesonephric as well as metanephric (adult) kidney formation via key inductive signaling events (Carroll & Vize, 1999; Natarajan, Jeyachandran, Subramaniyan, Thanigachalam, & Rajagopalan, 2013; Saxén & Sariola, 1987; Vize, Seufert, Carroll, & Wallingford, 1997). Early on, the pronephric duct signals nearby intermediate mesoderm to form mesonephric tubules and these allow drainage into the mesonephric duct, the most caudal portion of the original pronephric duct. Later on, the mesonephric duct forms the ureteric bud, and mutual induction between the metanephric mesenchyme and the ureteric bud promotes mature kidney development (Clarke et al., 2006; Costantini, 2010; Piscione & Rosenblum, 2002). Of particular note, studies on mutants of the retinoic acid-responsive pronephros specification gene *Pax2* revealed that homozygous mutant embryos were able to form both a pronephros and a mesonephros, but failed to induce the mature metanephric kidney (Bouchard, 2004; Brophy et al., 2001; M. Torres et al., 1995). These studies underscore the importance of proper pronephros specification for mature kidney development, and our zebrafish results suggest that *GREB1L* might be functioning in early pronephric development to ensure that the proper downstream developmental decisions are made.

Although the RA-associated *Wt1, Pax2,* and *Lhx1* genes are expressed in the early pronephros, they are also necessary for proper metanephric mesenchyme induction and development (Clarke et al., 2006; Donovan et al., 1999; Shawlot & Behringer, 1995). Remarkably, in mice *Greb1l* is also expressed at high levels in the metanephric mesenchyme, suggesting that similar to pronephros development, a second retinoic acid signaling event involving *Greb1l* and retinoic acid signaling targets is employed. Additionally, *Greb1l* is expressed in ureteric buds, albeit at lower levels, and retinoic acid signaling has been shown to be required for proper expression of Ret, itself a gene associated with RA. It is thus conceivable that the agenesis phenotype seen in *Greb1l* mutants may also be due to alterations in specification of either the metanephric mesenchyme or ureteric bud. Further studies are needed to establish which mechanisms underlie the agenesis phenotypes.

ACKNOWLEDGMENTS

The authors thank Senuri Jayatilleka for initial work on the *greb1l* morpholino and zebrafish *in situ* assays, and Caleb Heffner for assistance with mouse embryo processing. This work was funded in part by the National Institutes of Health grant R01 DE021071 (J.R.M.) as well as by the National Institutes of Health grant RC4 DK090937 (P.D.B. and J.R.M.), a National Institutes of Health grant UM1 OD023222 (S.A.M.), and a Maria Dorthea and Holger From, Haderslevs Foundation grant (M.R.).

EPILOGUE

This chapter was published in GENETICS, volume 207 (Investigation), issue number 1, page number 215-228, 2017. This was one of the highlighted articles in GENETICS for the September, 2017 issue (http://www.genetics.org/content/207/1/NP). I am one of the first co-authors in this manuscript. My analysis was significant in maintaining the quality of our data, confidence in our varaint calls and generation of a small candidate variant list, that led to the identification of *GREB1L* as an etiologic variant in RA. I believe this analysis can be used in other exome sequencing studies focusing on identifying Mendelian disease associated mutations.

Discovery of *GREB1L* is a novel and crucial addition to the field of understanding the underlying genetic basis for RA. Deleterious mutations in this key gene causes children to be born with one kidney or no kidneys (Brophy et al., 2017).

The subsection called “Exome sequencing analysis, Iowa” under “Materials and methods” section of this chapter, describes majority of my work for this manuscript. Here I describe my role after the exome seqeuncing data was generated for four affected individuals in the Iowa RA pedigree (II-5, II-7, III-4, III-8; Figure 3.1A) as follows:

- Checked the quality of the raw sequencing data using FastQC (Andrews, 2011). This step is extremely important as it verifies the data quality prior to mapping them. Higher data quality allows making confident variant calls against the reference genome.
- Adjusted the quality of the sequences and mapped them to the human reference genome (version-glk_v37) using BWA algorithm (H. Li & Durbin, 2009). Additionally, I performed mapping quality and exome target region coverage analysis using Qualimap algorithm (García-Alcalde et al., 2012) to ensure depth of coverage with high quality sequence alignments to make confident variant calls.
- Applied standard exome alignment-processing steps suggested by GATK (McKenna et al., 2010) before calling variants.

- Called the variants using Unified Genotyper algorithm under GATK (McKenna et al., 2010) against the human reference genome (version-glk_v37).
- Annotated the raw variant calls using SnpEff software, human reference assembly (version-hg19) and dbSNP database (Sherry et al. (2001), version-137).
- Computed MAF for each variant call in our samples using databases such as the 1000 Genomes Project and the NHLBI-ESP in the wANNOVAR web server (Chang & Wang, 2012) and dbSNP database (Sherry et al. (2001), version-137). The MAF frequency is a percentage that shows the prevalence of our variants in these databases. Assuming, the disease causing variants underlying Mendelian disorders would be rare we decided to remove variant calls with MAF ≥1%. dbSNP tagged variant calls with MAF >5%, therefore, I excluded them from my further analysis.
- Applied additional filters such as variant quality and depth of coverage to exclude false positive variant calls using GATK best practices guidelines (GATK_BP).
- Retained variant calls that have identical genotypes across all our samples.
- Checked the impact of each variant on their protein structure and function using algorithms such as PolyPhen-2 (Adzhubei et al., 2010) and SIFT (Kumar, Henikoff, & Ng, 2009) in the database of human nonsynonymous SNPs and function predictions (dbNSFP v2.0) (X. Liu et al., 2013). Certain times a variant can be called as benign by PolyPhen-2 and damaging by SIFT and vice-versa. Therefore, I applied a third algorithm called CONsensus DELeteriousness (CONDEL) (González-Pérez & López-Bigas, 2011) that generates a weighted

average of the normalized scores from PolyPhen-2 and SIFT and uses that score to classify a variant as benign or damaging. The damaging variants predicted by CONDEL were retained in my final list of genetic variants for further analysis.

The candidate gene list of variants were searched against the GUDMAP database (Harding, Armit, Armstrong, Brennan, Cheng, Haggarty, Houghton, Lloyd-MacGilp, Pi, Roochun, Sharghi, Tindal, McMahon, Gottesman, Little, Georgas, Aronow, Potter, et al., 2011; McMahon et al., 2008). GUDMAP is a database of gene expression patterns associated with organs of the developing genitourinary tract and kidney. This search confirmed the expression of a gene in my candidate gene list called, *GREB1L,* in the developing kidney, supporting its involvement in RA. Finally, the *GREB1L* variant was validated to be present in six affected family members (II-5, II-7, III-3, III-4, III-6 and III-8; Figure 3.1A) and absent in seven unaffected family members (I-3, I-5, II-3, II-4, III-1, III-2, III-5; Figure 3.1A) of our RA pedigree, confirming the predicted segregation pattern of an etiologic variant. One unaffected family member II-2 in our pedigree also exhibited the *GREB1L* variant, suggesting incomplete penetrance of the RA phenotype. This knowledge has and will continue to contribute towards follow up studies on *GREB1L*'s role in kidney development and CAKUT. Additionally, sqeuncing of exomes and bionformatically analyzing them to identify new disease causing variants will empower our knowledge on the genetic basis underlying human diseases and development.

Individual	Total # reads	# mapped reads / % of total reads	# mapped reads in target region / % of total reads	Mean coverage in target region	Mean mapping quality in target region
II.4	138,952,013	121,216,737 / 87.24%	100,480,821 / 72.31%	131.08	44.53
II.6	151,277,545	149,022,896 / 98.51%	123,983,267 / 81.96%	160.78	45.87
III.4	179,371,347	163,104,270 / 90.93%	138,330,046 / 77.12%	177.45	45.07
III.7	201,991,695	198,300,809 / 98.17%	166,643,662 / 82.50%	217.45	45.71
Average	**167,898,150**	**157,911,178 / 93.71%**	**132,359,449 / 78.47%**	**171.69**	**45.30**

Table 3.1. Whole exome mapping statistics for four cases (Iowa family). Across the samples, we achieved an average targeted exome coverage of 171.69X with a mean mapping quality 45.30 for calling high quality variants.

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

Individual	Total # reads	# mapped reads / % of total reads	# mapped reads in target region / % of total reads	Mean coverage in target region	Mean mapping quality in target region
Fetus	161,010,696	159,630,813 /99.14%	137,922,285 /85.66%	223.7	62.33
Mother	70,125,707	69,581,369 /99.22%	59,504,884 /84.85%	100.4	61.3
Father	69,880,361	69,349,566 /99.24%	59,425,709 /85.03%	98.4	61.56
Maternal Grandmother	64,728,968	64,216,353 /99.2%	54,985,396 /84.94%	92.2	61.44
Maternal Grandfather	85,311,447	84,643,305 /99.21%	72,529,359 /85%	120.5	61.44
Average	**167,898,150**	**157,911,178 / 93.71%**	**132,359,449 / 78.47%**	**127**	**62**

Table 3.2. Whole exome mapping statistics for two cases and three unaffected family members (Denmark family). Across the samples, we achieved an average targeted exome coverage of 127X with a mean mapping quality of 62 for calling variants of high quality.

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

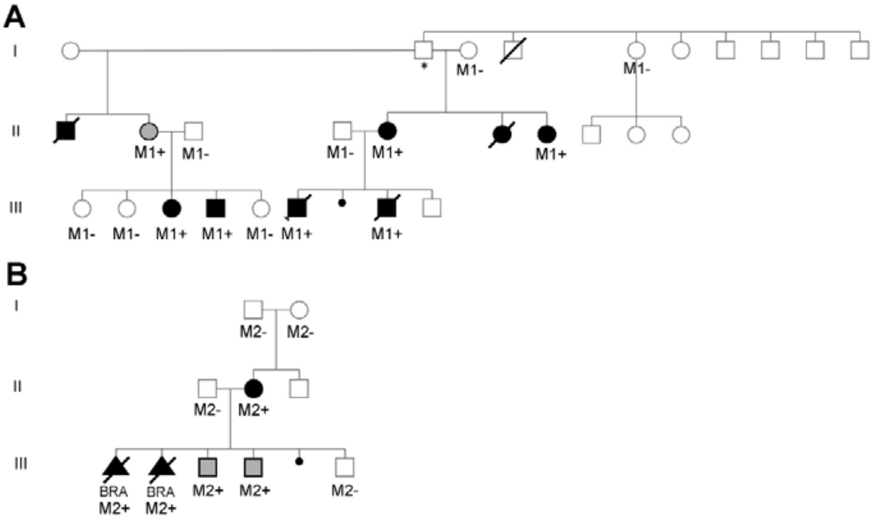
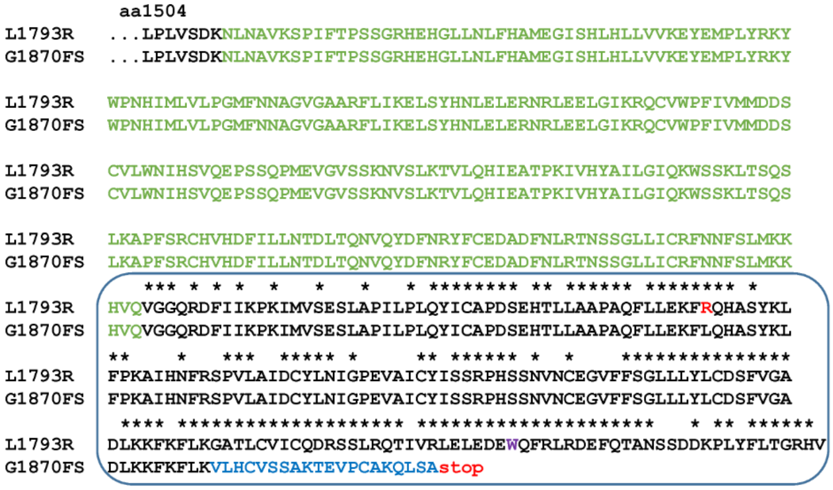


Figure 3.1. Iowa and Danish renal agenesis pedigrees. (A) Iowa pedigree showing dominant inheritance of the agenesis phenotype. (B) Danish pedigree showing transmission of the de novo *GREB1L* variant to both fetuses. M1 = Iowa variant, M2 = Danish variant, + = presence, - = absence, * = likely origin of the Iowa mutation. II-2 shaded female (Iowa), III-3,4 shaded males (Denmark) = family members with incomplete penetrance.

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125



```
           aa1504
L1793R    ...LPLVSDKNLNAVKSPIFTPSSGRHEHGLLNLFHAMEGISHLHLLVVKEYEMPLYRKY
G1870FS   ...LPLVSDKNLNAVKSPIFTPSSGRHEHGLLNLFHAMEGISHLHLLVVKEYEMPLYRKY

L1793R    WPNHIMLVLPGMFNNAGVGAARFLIKELSYHNLELERNRLEELGIKRQCVWPFIVMMDDS
G1870FS   WPNHIMLVLPGMFNNAGVGAARFLIKELSYHNLELERNRLEELGIKRQCVWPFIVMMDDS

L1793R    CVLWNIHSVQEPSSQPMEVGVSSKNVSLKTVLQHIEATPKIVHYAILGIQKWSSKLTSQS
G1870FS   CVLWNIHSVQEPSSQPMEVGVSSKNVSLKTVLQHIEATPKIVHYAILGIQKWSSKLTSQS

L1793R    LKAPFSRCHVHDFILLNTDLTQNVQYDFNRYFCEDADFNLRTNSSGLLICRFNNFSLMKK
G1870FS   LKAPFSRCHVHDFILLNTDLTQNVQYDFNRYFCEDADFNLRTNSSGLLICRFNNFSLMKK
             *** * * *  *   *   *   * ********* ** ****** ********* *
L1793R    HVQVGGQRDFIIKPKIMVSESLAPILPLQYICAPDSEHTLLAAPAQFLLEKFRQHASYKL
G1870FS   HVQVGGQRDFIIKPKIMVSESLAPILPLQYICAPDSEHTLLAAPAQFLLEKFLQHASYKL
          **    *   ***  ***** *     ** ****** *  *   ****************
L1793R    FPKAIHNFRSPVLAIDCYLNIGPEVAICYISSRPHSSNVNCEGVFFSGLLLYLCDSFVGA
G1870FS   FPKAIHNFRSPVLAIDCYLNIGPEVAICYISSRPHSSNVNCEGVFFSGLLLYLCDSFVGA
           **** ********************* *********************   * ** ******
L1793R    DLKKFKFLKGATLCVICQDRSSLRQTIVRLELEDEWQFRLRDEFQTANSSDDKPLYFLTGRHV
G1870FS   DLKKFKFLKVLHCVSSAKTEVPCAKQLSAstop
```

Figure 3.2 Comparison of proteins encoded by Iowa, Danish and zebrafish *GREB1L* mutants. The human, mouse, and fish variants (position indicated in human protein) encode proteins that are altered in the conserved c-terminus of the protein. L1793R = Iowa protein, G1870FS = Danish protein; predicted glycosyltransferase domain = green lettering (Iyer, Zhang, Burroughs, & Aravind, 2013). Iowa L to R mutation indicated in red lettering, which was recapitulated in two out of three mouse mutants along with deletion of the following Q residue (see Supplementary Figure 3.S3); Danish frameshift amino acids indicated in blue lettering; zebrafish W to STOP mutation indicated by purple W residue. Conserved c-terminus indicated by boxed region, with asterisks denoting amino acids conserved between GREB1 and GREB1L paralogs.

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

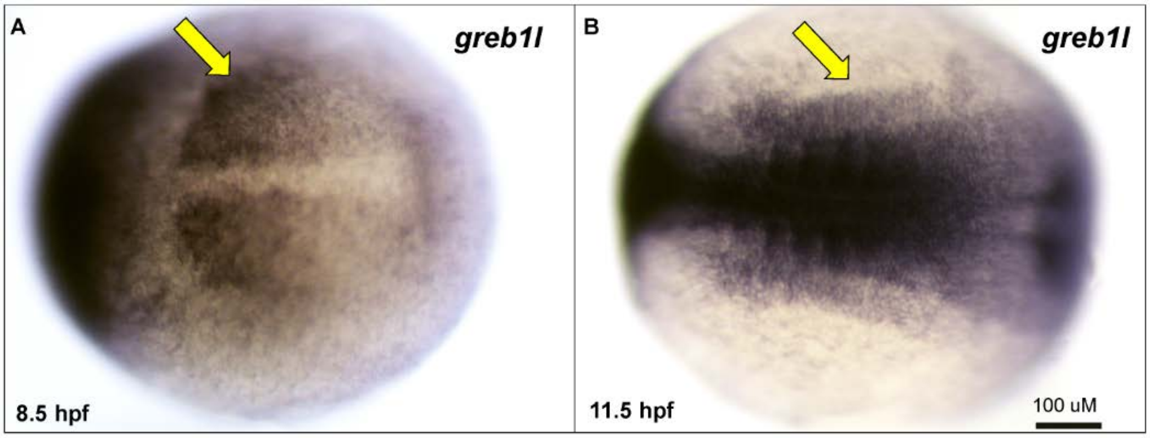


Figure 3.3. Endogenous expression of *greb1l* during zebrafish development. *In situ* hybridization of *greb1l* anti-sense probes on embryos fixed at indicated stages. Embryos are presented in a dorsal view with rostral to the left. Arrows indicate intermediate mesoderm signal.

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics*, *207*(1), 215-228. doi:10.1534/genetics.117.1125

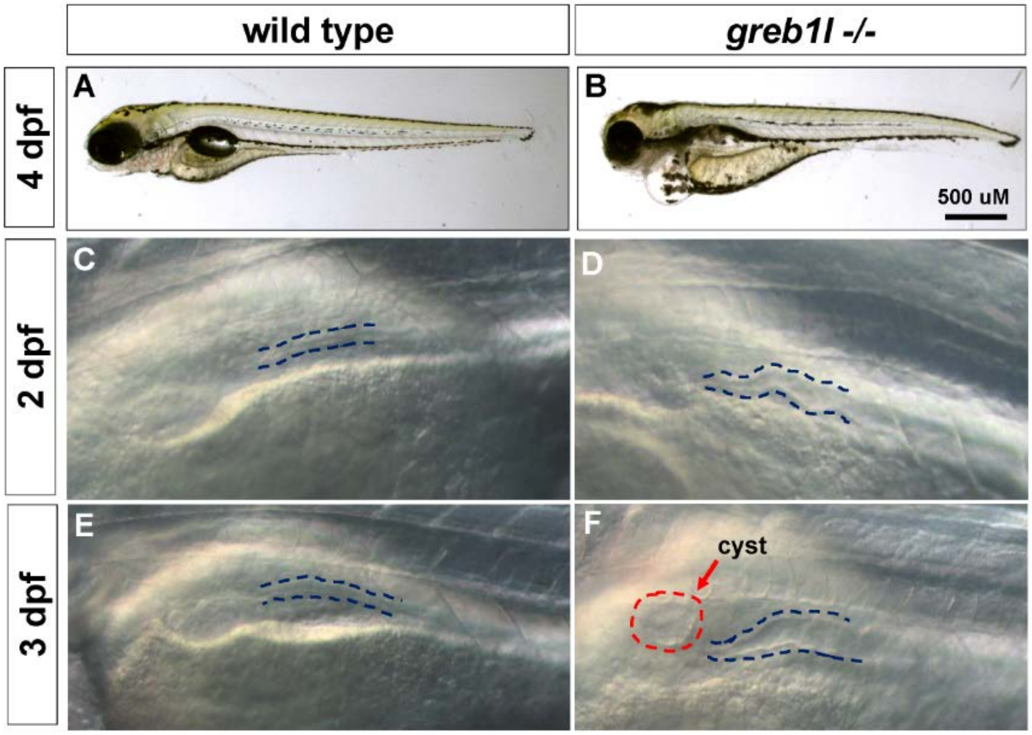


Figure 3.4. *greb1l* mutants have edema and abnormal pronephros development. (A-F) Lateral views of live larvae at the indicated stage and of the indicated genotype. (B) At 4 dpf mutants exhibit edema particularly around the heart and eyes, (D) 2 dpf mutant embryos have dilated and kinked tubules (F; hatched lines). At 3 dpf in mutant embryos the kidney remains dilated (hatched lines) and cysts are evident in most cases (red hatched circle).

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics*, *207*(1), 215-228. doi:10.1534/genetics.117.1125

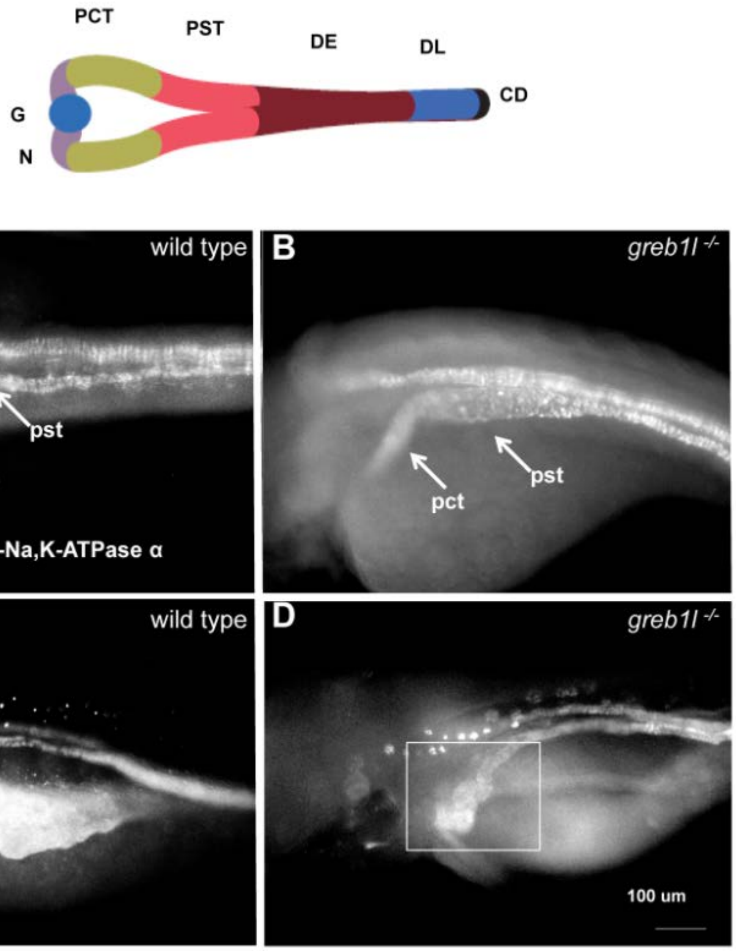


Figure 3.5. Renal morphology of zebrafish *greb1l* mutants. In all images, rostral is to the left and embryos are processed with anti-Na,K ATPase antibody (a6F) and a fluorescent secondary antibody. (A) and (B) dorsal views of representative embryos at 3 days post fertilization (dpf). Mutants present with swelling of the proximal convoluted tubule (PCT) and proximal straight tubule (PST). (C-D) are ventral-lateral views. Mutants have deformed junction between the PCT and neck. Top, schematic modified from Drummond and Davidson(Drummond & Davidson, 2016b).

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

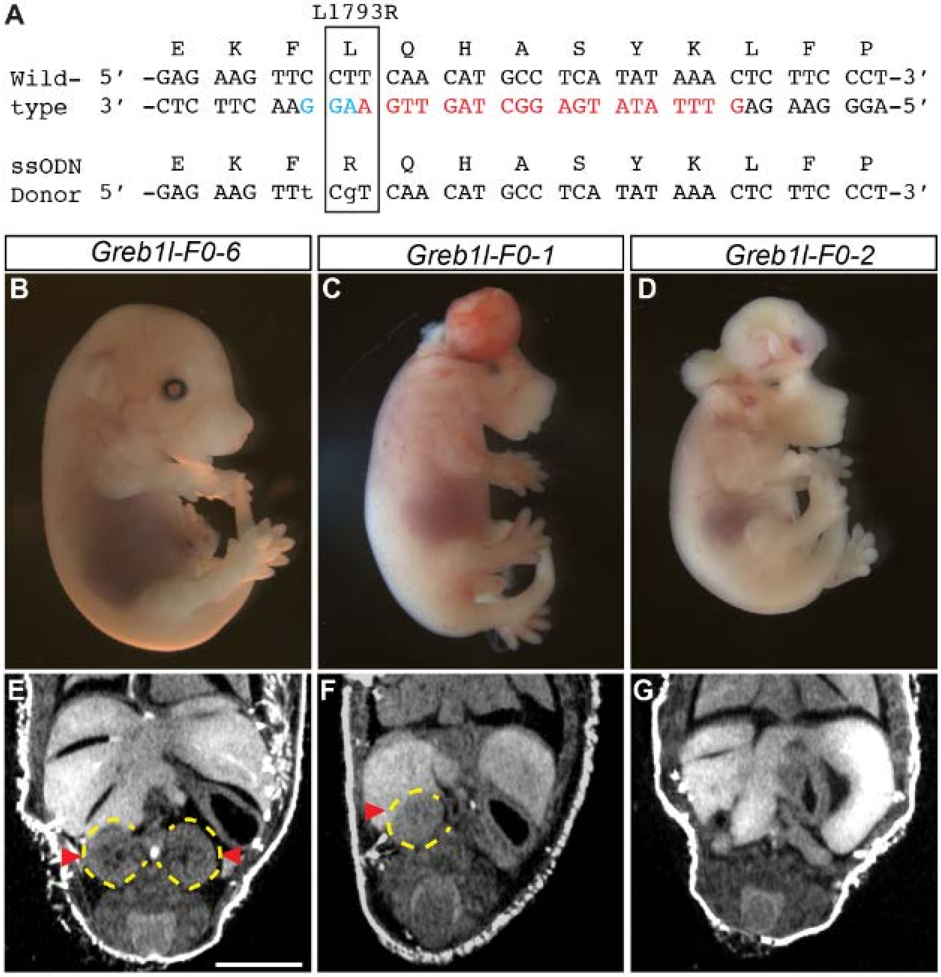
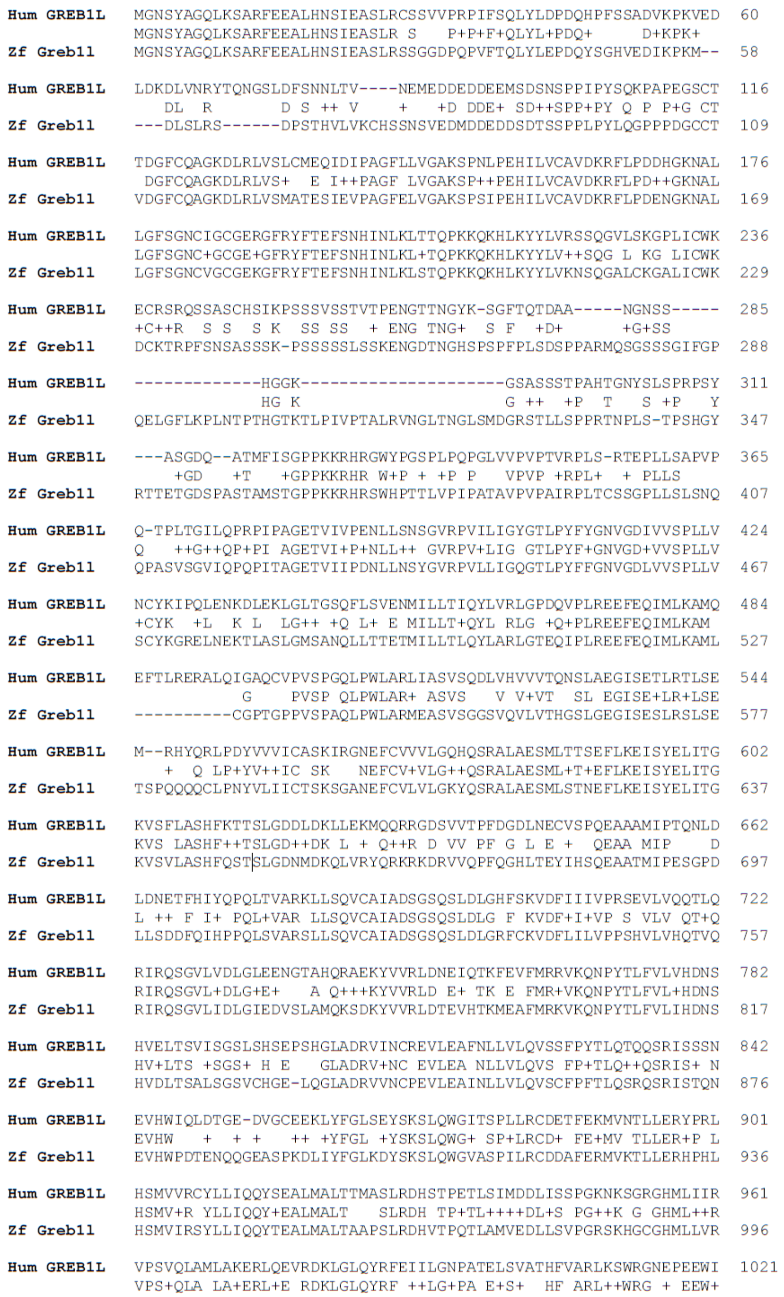

A
L1793R
E K F L Q H A S Y K L F P
Wild- 5′ -GAG AAG TTC CTT CAA CAT GCC TCA TAT AAA CTC TTC CCT-3′
type 3′ -CTC TTC AAG GAA GTT GAT CGG AGT ATA TTT GAG AAG GGA-5′
ssODN E K F R Q H A S Y K L F P
Donor 5′ -GAG AAG TTt CgT CAA CAT GCC TCA TAT AAA CTC TTC CCT-3′
Greb1l-F0-6
Greb1l-F0-1
Greb1l-F0-2
B
C
D
E
F
G

Figure 3.6. Analysis of *Greb1l* function in CRISPR/Cas9 mutagenized F0 mouse embryos. (A) CRISPR/Cas9 strategy for introducing the L1793R mutation using an ssODN donor template. The guide sequence is colored red and the adjacent PAM sequence (AGG) is indicated in turquoise. Point mutations engineered into the donor are shown as lowercase. (B-D) Whole-mount images highlighting the observed exencephaly in two mutants (C, D) carrying KI alleles as compared to wild-type (B). (E-F) Coronal sections of microCT data show unilateral and complete kidney agenesis in mutagenized F0 embryos. The position of kidneys is indicated by red arrowheads and yellow dotted circles. Scale bar is 2 mm.

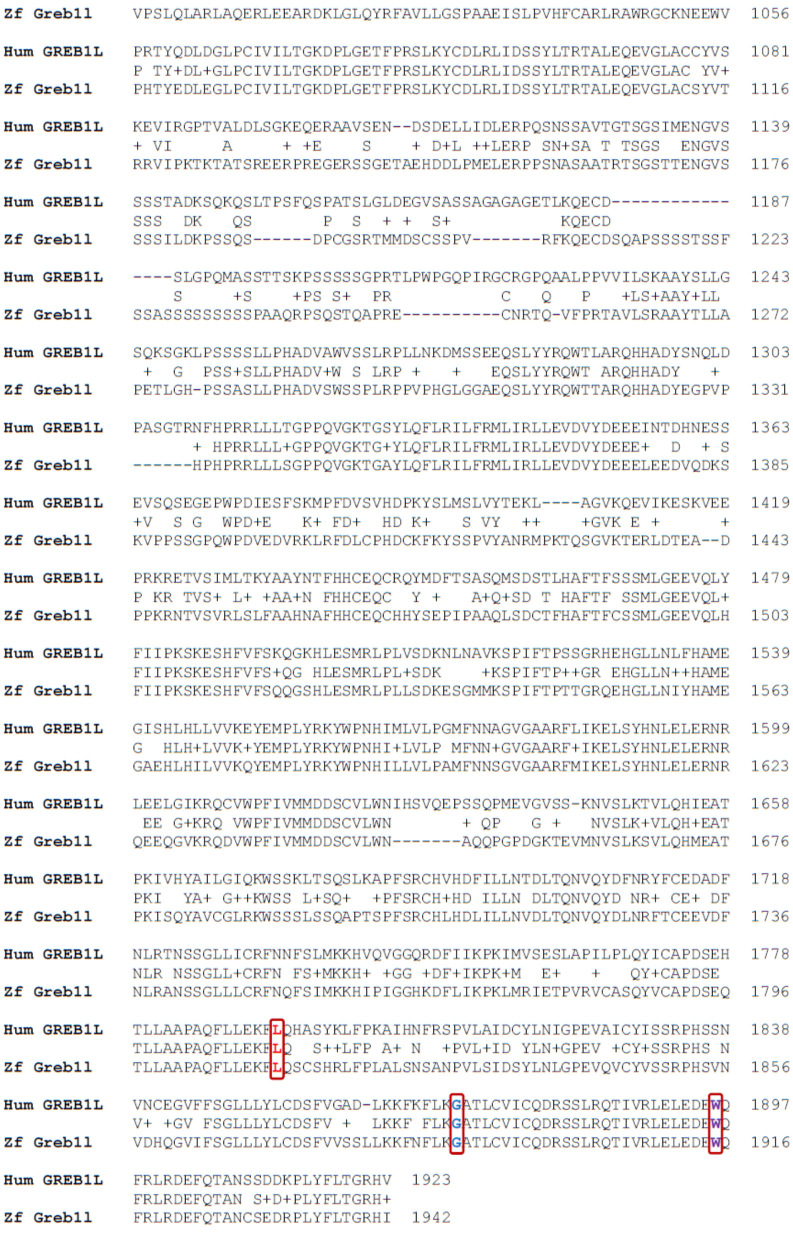
Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125



```
Hum GREB1L   MGNSYAGQLKSARFEEALHNSIEASLRCSSVVPRPIFSQLYLDPDQHPFSSADVKPKVED  60
             MGNSYAGQLKSARFEEALHNSIEASLR S   P+P+F+QLYL+PDQ+     D+KPK+
Zf Greb1l    MGNSYAGQLKSARFEEALHNSIEASLRSSGGDPQPVFTQLYLEPDQYSGHVEDIKPKM--  58

Hum GREB1L   LDKDLVNRYTQNGSLDFSNNLTV----NEMEDDEDDEEMSDSNSPPIPYSQKPAPEGSCT  116
                DL  R       D S ++ V    +   +D DDE+ SD++SPP+PY Q P P+G CT
Zf Greb1l    ---DLSLRS------DPSTHVLVKCHSSNSVEDMDDEDDSDTSSPPLPYLQGPPPDGCCT  109

Hum GREB1L   TDGFCQAGKDLRLVSLCMEQIDIPAGFLLVGAKSPNLPEHILVCAVDKRFLPDDHGKNAL  176
              DGFCQAGKDLRLVS+  E I++PAGF LVGAKSP++PEHILVCAVDKRFLPD++GKNAL
Zf Greb1l    VDGFCQAGKDLRLVSMATESIEVPAGFELVGAKSPSIPEHILVCAVDKRFLPDENGKNAL  169

Hum GREB1L   LGFSGNCIGCGERGFRYFTEFSNHINLKLTTQPKKQKHLKYYLVRSSQGVLSKGPLICWK  236
             LGFSGNC+GCGE+GFRYFTEFSNHINLKL+TQPKKQKHLKYYLV++SQG L KG LICWK
Zf Greb1l    LGFSGNCVGCGEKGFRYFTEFSNHINLKLSTQPKKQKHLKYYLVKNSQGALCKGALICWK  229

Hum GREB1L   ECRSRQSSASCHSIKPSSSVSSTVTPENGTTNGYK-SGFTQTDAA-----NGNSS-----  285
             +C++R  S S  S K  SS SS  + ENG TNG+  S F  +D+      +G+SS
Zf Greb1l    DCKTRPFSNSASSSK-PSSSSSLSSKENGDTNGHSPSPFPLSDSPPARMQSGSSSGIFGP  288

Hum GREB1L   -------------HGGK---------------------GSASSSTPAHTGNYSLSPRPSY  311
                          HG K                     G ++  +P  T   S +P   Y
Zf Greb1l    QELGFLKPLNTPTHGTKTLPIVPTALRVNGLTNGLSMDGRSTLLSPPRTNPLS-TPSHGY  347

Hum GREB1L   ---ASGDQ--ATMFISGPPKKRHRGWYPGSPLPQPGLVVPVPTVRPLS-RTEPLLSAPVP  365
                 +GD   +T   +GPPKKRHR W+P + +P P   VPVP +RPL+  + PLLS
Zf Greb1l    RTTETGDSPASTAMSTGPPKKRHRSWHPTTLVPIPATAVPVPAIRPLTCSSGPLLSLSNQ  407

Hum GREB1L   Q-TPLTGILQPRPIPAGETVIVPENLLSNSGVRPVILIGYGTLPYFYGNVGDIVVSPLLV  424
             Q   ++G++QP+PI AGETVI+P+NLL++ GVRPV+LIG GTLPYF+GNVGD+VVSPLLV
Zf Greb1l    QPASVSGVIQPQPITAGETVIIPDNLLNSYGVRPVLLIGQGTLPYFFGNVGDLVVSPLLV  467

Hum GREB1L   NCYKIPQLENKDLEKLGLTGSQFLSVENMILLTIQYLVRLGPDQVPLREEFEQIMLKAMQ  484
             +CYK  +L  K L  LG++ +Q L+ E MILLT+QYL RLG +Q+PLREEFEQIMLKAM
Zf Greb1l    SCYKGRELNEKTLASLGMSANQLLTTETMILLTLQYLARLGTEQIPLREEFEQIMLKAML  527

Hum GREB1L   EFTLRERALQIGAQCVPVSPGQLPWLARLIASVSQDLVHVVVTQNSLAEGISETLRTLSE  544
                        G    PVSP QLPWLAR+ ASVS   V V+VT  SL EGISE+LR+LSE
Zf Greb1l    ----------CGPTGPPVSPAQLPWLARMEASVSGGSVQVLVTHGSLGEGISESLRSLSE  577

Hum GREB1L   M--RHYQRLPDYVVVICASKIRGNEFCVVVLGQHQSRALAESMLTTSEFLKEISYELITG  602
                +  Q LP+YV++IC SK   NEFCV+VLG++QSRALAESML+T+EFLKEISYELITG
Zf Greb1l    TSPQQQQCLPNYVLIICTSKSGANEFCVLVLGKYQSRALAESMLSTNEFLKEISYELITG  637

Hum GREB1L   KVSFLASHFKTTSLGDDLDKLLEKMQQRRGDSVVTPFDGDLNECVSPQEAAAMIPTQNLD  662
             KVS LASHF++TSLGD++DK L + Q++R D VV PF G L E +  QEAA MIP    D
Zf Greb1l    KVSVLASHFQSTSLGDNMDKQLVRYQRKRKDRVVQPFQGHLTEYIHSQEAATMIPESGPD  697

Hum GREB1L   LDNETFHIYQPQLTVARKLLSQVCAIADSGSQSLDLGHFSKVDFIIIVPRSEVLVQQTLQ  722
             L ++ F I+ PQL+VAR LLSQVCAIADSGSQSLDLG F KVDF+I+VP S VLV QT+Q
Zf Greb1l    LLSDDFQIHPPQLSVARSLLSQVCAIADSGSQSLDLGRFCKVDFLILVPPSHVLVHQTVQ  757

Hum GREB1L   RIRQSGVLVDLGLEENGTAHQRAEKYVVRLDNEIQTKFEVFMRRVKQNPYTLFVLVHDNS  782
             RIRQSGVL+DLG+E+   A Q+++KYVVRLD E+ TK E FMR+VKQNPYTLFVL+HDNS
Zf Greb1l    RIRQSGVLIDLGIEDVSLAMQKSDKYVVRLDTEVHTKMEAFMRKVKQNPYTLFVLIHDNS  817

Hum GREB1L   HVELTSVISGSLSHSEPSHGLADRVINCREVLEAFNLLVLQVSSFPYTLQTQQSRISSSN  842
             HV+LTS +SGS+ H E   GLADRV+NC EVLEA NLLVLQVS FP+TLQ++QSRIS+ N
Zf Greb1l    HVDLTSALSGSVCHGE-LQGLADRVVNCPEVLEAINLLVLQVSCFPFTLQSRQSRISTQN  876

Hum GREB1L   EVHWIQLDTGE-DVGCEEKLYFGLSEYSKSLQWGITSPLLRCDETFEKMVNTLLERYPRL  901
             EVHW   +  + +   ++ +YFGL +YSKSLQWG+ SP+LRCD+ FE+MV TLLER+P L
Zf Greb1l    EVHWPDTENQQGEASPKDLIYFGLKDYSKSLQWGVASPILRCDDAFERMVKTLLERHPHL  936

Hum GREB1L   HSMVVRCYLLIQQYSEALMALTTMASLRDHSTPETLSIMDDLISSPGKNKSGRGHMLIIR  961
             HSMV+R YLLIQQY+EALMALT   SLRDH TP+TL++++DL+S PG++K G GHML++R
Zf Greb1l    HSMVIRSYLLIQQYTEALMALTAAPSLRDHVTPQTLAMVEDLLSVPGRSKHGCGHMLLVR  996

Hum GREB1L   VPSVQLAMLAKERLQEVRDKLGLQYRFEIILGNPATELSVATHFVARLKSWRGNEPEEWI  1021
             VPS+QLA LA+ERL+E RDKLGLQYRF ++LG+PA E+S+  HF ARL++WRG + EEW+
```

Figure 3.S1. Amino acid sequence comparison between orthologous human and zebrafish *GREB1L* proteins. Note the strong conservation at the c-termiNi^{2+} of the proteins, as well as the conservation at positions affected by the mutations. Circled red L residue = position of the Iowa mutation, circled blue G residue = position of the Danish frameshift mutation, circled purple W mutation = position of the stop codon mutation in the fish. See Supplementary Figure 3.S3 for mouse mutations.

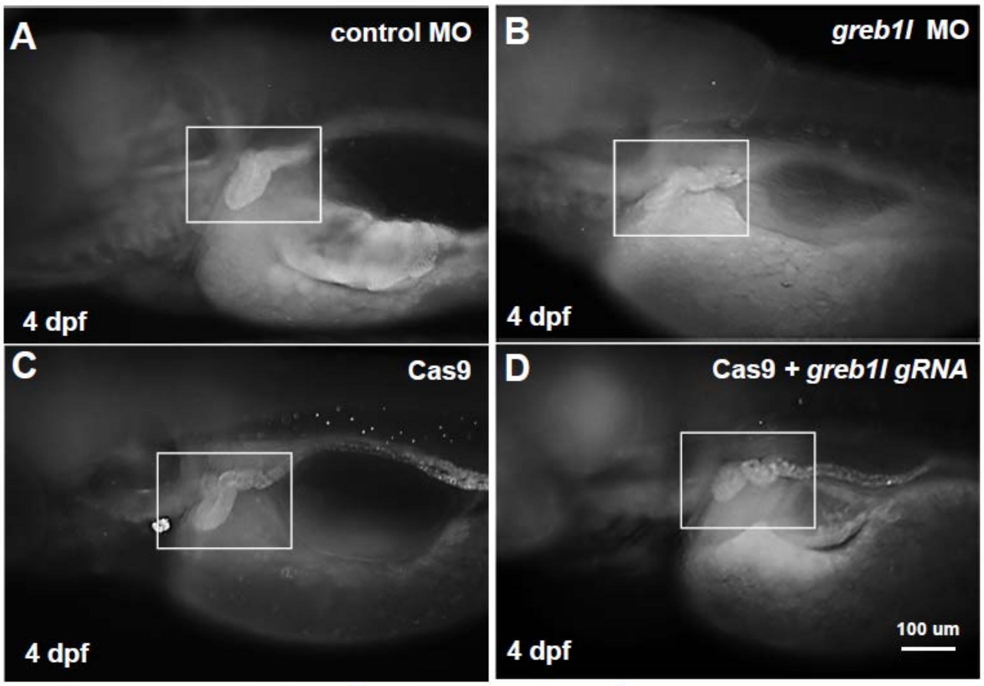
Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125



```
Zf Greb1l     VPSLQLARLAQERLEEARDKLGLQYRFAVLLGSPAAEISLPVHFCARLRAWRGCKNEEWV  1056

Hum GREB1L    PRTYQDLDGLPCIVILTGKDPLGETFPRSLKYCDLRLIDSSYLTRTALEQEVGLACCYVS  1081
              P TY+DL+GLPCIVILTGKDPLGETFPRSLKYCDLRLIDSSYLTRTALEQEVGLAC YV+
Zf Greb1l     PHTYEDLEGLPCIVILTGKDPLGETFPRSLKYCDLRLIDSSYLTRTALEQEVGLACSYVT  1116

Hum GREB1L    KEVIRGPTVALDLSGKEQERAAVSEN--DSDELLIDLERPQSNSSAVTGTSGSIMENGVS  1139
              + VI     A     + +E    S    + D+L ++LERP SN+SA T TSGS  ENGVS
Zf Greb1l     RRVIPKTKTATSREERPREGERSSGETAEHDDLPMELERPPSNASAATRTSGSTTENGVS  1176

Hum GREB1L    SSSTADKSQKQSLTPSFQSPATSLGLDEGVSASSAGAGAGETLKQECD------------  1187
              SSS  DK   QS       P  S  + +  S+           KQECD
Zf Greb1l     SSSILDKPSSQS------DPCGSRTMMDSCSSPV-------RFKQECDSQAPSSSSTSSF  1223

Hum GREB1L    ----SLGPQMASSTTSKPSSSSSGPRTLPWPGQPIRGCRGPQAALPPVVILSKAAYSLLG  1243
                  S     +S    +PS S+  PR           C   Q   P   +LS+AAY+LL
Zf Greb1l     SSASSSSSSSSSPAAQRPSQSTQAPRE----------CNRTQ-VFPRTAVLSRAAYTLLA  1272

Hum GREB1L    SQKSGKLPSSSSLLPHADVAWVSSLRPLLNKDMSSEEQSLYYRQWTLARQHHADYSNQLD  1303
               +  G  PSS+SLLPHADV+W S LRP +   +   EQSLYYRQWT ARQHHADY   +
Zf Greb1l     PETLGH-PSSASLLPHADVSWSSPLRPPVPHGLGGAEQSLYYRQWTTARQHHADYEGPVP  1331

Hum GREB1L    PASGTRNFHPRRLLLTGPPQVGKTGSYLQFLRILFRMLIRLLEVDVYDEEEINTDHNESS  1363
                    + HPRRLLL+GPPQVGKTG+YLQFLRILFRMLIRLLEVDVYDEEE+  D  + S
Zf Greb1l     ------HPHPRRLLLSGPPQVGKTGAYLQFLRILFRMLIRLLEVDVYDEEELEEDVQDKS  1385

Hum GREB1L    EVSQSEGEPWPDIESFSKMPFDVSVHDPKYSLMSLVYTEKL----AGVKQEVIKESKVEE  1419
              +V  S G  WPD+E   K+ FD+  HD K+   S VY  ++    +GVK E +      +
Zf Greb1l     KVPPSSGPQWPDVEDVRKLRFDLCPHDCKFKYSSPVYANRMPKTQSGVKTERLDTEA--D  1443

Hum GREB1L    PRKRETVSIMLTKYAAYNTFHHCEQCRQYMDFTSASQMSDSTLHAFTFSSSMLGEEVQLY  1479
              P KR TVS+ L+ +AA+N FHHCEQC  Y +   A+Q+SD T HAFTF SSMLGEEVQL+
Zf Greb1l     PPKRNTVSVRLSLFAAHNAFHHCEQCHHYSEPIPAAQLSDCTFHAFTFCSSMLGEEVQLH  1503

Hum GREB1L    FIIPKSKESHFVFSKQGKHLESMRLPLVSDKNLNAVKSPIFTPSSGRHEHGLLNLFHAME  1539
              FIIPKSKESHFVFS+QG HLESMRLPL+SDK     +KSPIFTP++GR EHGLLN++HAME
Zf Greb1l     FIIPKSKESHFVFSQQGSHLESMRLPLLSDKESGMMKSPIFTPTTGRQEHGLLNIYHAME  1563

Hum GREB1L    GISHLHLLVVKEYEMPLYRKYWPNHIMLVLPGMFNNAGVGAARFLIKELSYHNLELERNR  1599
              G  HLH+LVVK+YEMPLYRKYWPNHI+LVLP MFNN+GVGAARF+IKELSYHNLELERNR
Zf Greb1l     GAEHLHILVVKQYEMPLYRKYWPNHILLVLPAMFNNSGVGAARFMIKELSYHNLELERNR  1623

Hum GREB1L    LEELGIKRQCVWPFIVMMDDSCVLWNIHSVQEPSSQPMEVGVSS-KNVSLKTVLQHIEAT  1658
               EE G+KRQ VWPFIVMMDDSCVLWN       + QP   G +   NVSLK+VLQH+EAT
Zf Greb1l     QEEQGVKRQDVWPFIVMMDDSCVLWN-------AQQPGPDGKTEVMNVSLKSVLQHMEAT  1676

Hum GREB1L    PKIVHYAILGIQKWSSKLTSQSLKAPFSRCHVHDFILLNTDLTQNVQYDFNRYFCEDADF  1718
              PKI  YA+ G++KWSS L+SQ+  +PFSRCH+HD ILLN DLTQNVQYD NR+ CE+ DF
Zf Greb1l     PKISQYAVCGLRKWSSSLSSQAPTSPFSRCHLHDLILLNVDLTQNVQYDLNRFTCEEVDF  1736

Hum GREB1L    NLRTNSSGLLICRFNNFSLMKKHVQVGGQRDFIIKPKIMVSESLAPILPLQYICAPDSEH  1778
              NLR NSSGLL+CRFN FS+MKKH+ +GG +DF+IKPK+M  E+   +   QY+CAPDSE
Zf Greb1l     NLRANSSGLLLCRFNQFSIMKKHIPIGGHKDFLIKPKLMRIETPVRVCASQYVCAPDSEQ  1796

Hum GREB1L    TLLAAPAQFLLEKFLQHASYKLFPKAIHNFRSPVLAIDCYLNIGPEVAICYISSRPHSSN  1838
              TLLAAPAQFLLEKFLQ  S++LFP A+ N  +PVL+ID YLN+GPEV +CY+SSRPHS N
Zf Greb1l     TLLAAPAQFLLEKFLQSCSHRLFPLALSNSANPVLSIDSYLNLGPEVQVCYVSSRPHSVN  1856

Hum GREB1L    VNCEGVFFSGLLLYLCDSFVGAD-LKKFKFLKGATLCVICQDRSSLRQTIVRLELEDEWQ  1897
              V+ +GV FSGLLLYLCDSFV +  LKKF FLKGATLCVICQDRSSLRQTIVRLELEDEWQ
Zf Greb1l     VDHQGVIFSGLLLYLCDSFVVSSLLKKFNFLKGATLCVICQDRSSLRQTIVRLELEDEWQ  1916

Hum GREB1L    FRLRDEFQTANSSDDKPLYFLTGRHV  1923
              FRLRDEFQTAN S+D+PLYFLTGRH+
Zf Greb1l     FRLRDEFQTANCSEDRPLYFLTGRHI  1942
```

Figure 3.S1. —continued

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125



E	% abnormal morphology of PCT/Neck junction	Number of embryos (# of experiments)
Control MO	4	45 (3)
greb1l e3i3 MO	85	135 (3)
Cas9	6	33 (3)
Cas9 + *greb1l* gRNA	42	144 (3)

Figure 3.S2. Renal morphology of zebrafish in which greb1l expression is inhibited with morpholino or mutated with CRISPR/Cas9. All images are lateral views of zebrafish larvae at 4 days post fertilization, rostral to the left, fixed and processed to reveal anti-Na,K ATPase antibody (a6F) immunoreactivity. Wild type embryos were injected with (A) the "standard control" MO that is complementary to no sequence in the zebrafish genome (Gene Tools), (B) an MO targeting e3/i3 of greb1l, (C) Cas9 protein alone, or (D) Cas9 protein plus guide RNA targeting greb1l exon 17 (of 34). (A, C) In control embryos, the junction of the proximal convoluted tubule and the neck exhibited the hairpin structure characteristic of wild-type embryos. (B, D) In strongly-affected (B) greb1l MO-injected or (D) greb1l guide RNA injected embryos, the structure of this region was kinked. (E) Table depicting the number of fish observed and the percentage of them that displayed kidney deformities.

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

A

```
     E   K   F   L   Q   H   A   S   Y   K   L   F   P
WT- GAG AAG TTC CTT CAA CAT GCC TCA TAT AAA CTC TTC CCT
```

Greb1l-F0-1

```
     E   K   F   R   Q   H   A   S   Y   K   L   F   P
#1  GAG AAG TTT CGT CAA CAT GCC TCA TAT AAA CTC TTC CCT (4/8)
     E   K   F   L       H   A   S   Y   K   L   F   P
#2  GAG AAG TTC CTT CA- --T GCC TCA TAT AAA CTC TTC CCT (4/8)
```

Greb1l-F0-2

```
     E   K   F   R   Q   H   A   S   Y   K   L   F   P
#1  GAG AAG TTT CGT CAA CAT GCC TCA TAT AAA CTC TTC CCT (3/8)
     E   K   F   L       H   A   S   Y   K   L   F   P
#2  GAG AAG TTC CTT CA- --T GCC TCA TAT AAA CTC TTC CCT (5/8)
```

Greb1l-F0-14

```
    E   K   F   L   Q   H   M   P   H   I   N   S   S   R   *
#1 GAG AAG TTC CTT CAA CAc aTG CCT CAT ATA AAC TCT TCC…AGG TAG (3/3)
```

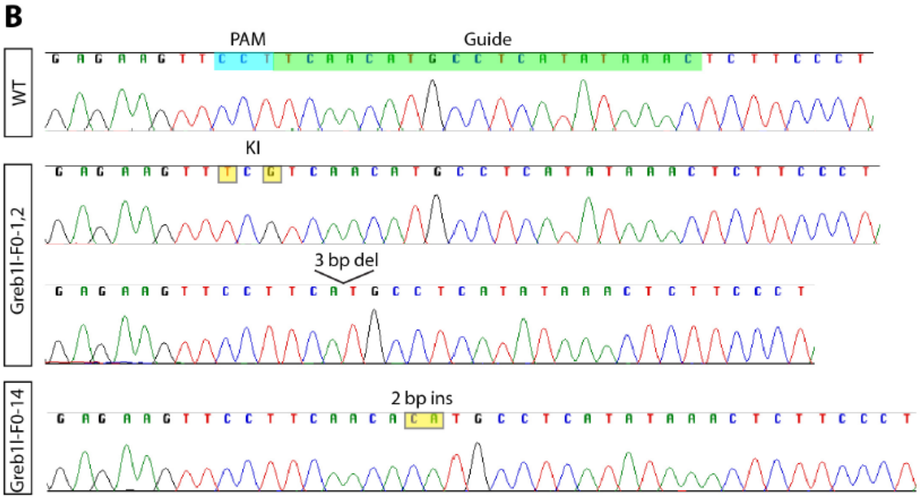


Figure 3.S3. Mutagenized alleles recovered from affected CRISPR F0 embryos. (A) Reference genomic sequence is shown at top with recovered mutant alleles shown below for each embryo displaying kidney phenotype. Notably, embryos #1 and 2 shared the same-in-frame deletion in addition to the desired KI allele. The number in parentheses indicated the number of clones identified with the specific mutation. (B) Chromatograms of clones reads corresponding to mutations described in (A).

Source: Brophy, P. D., Rasmussen, M., **Parida, M.**, Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

CHAPTER 4. BIOLOGICAL PROCESSES AND PATHWAYS ASSOCIATED WITH HEAVY METAL INDUCED STRESS IN *DAPHNIA PULEX*. [4]

INTRODUCTION

Ecotoxicology is the branch of toxicology that studies the distribution and impact of environmental pollutants on organisms and their population levels, and factors in species interactions to estimate the consensus effect of these pollutants on the ecosystem (Schirmer et al., 2010; Truhaut, 1977). In the recent years, exponential increase in heavy metal exposure due to natural causes such as volcanic eruptions, weathering of rocks, soil erosion and leaching of metals to ground water sources, and primarily through, anthropogenic activities such as industrial mining and agricultural operations, has become a worldwide public health concern (Tchounwou, Yedjou, Patlolla, & Sutton, 2012). Heavy metals are relatively denser than water (E Fergusson, 1991) and can be categorized into essential and non-essential metals (Gunther, Lindert, & Schaffner, 2012). In trace amounts, essential metals such as Ni^{2+} and copper (Cu) are part of proteins that perform and regulate vital biological processes such as carbohydrate metabolism, hemoglobin formation, and sperm production and motility (Gunther et al., 2012; Tchounwou et al., 2012; Yokoi, Uthus, & Nielsen, 2003). However, in excess concentrations they can induce genotoxicity (Skreb & Fischer, 1984) and ultimately lead to a wide-range of human diseases such as Wilson's disease due to Cu^{2+} accumulation (Tchounwou, Newsome, Williams, & Glass, 2008) and allergic contact dermatitis due to

[4] This chapter is unpublished.

Ni^{2+} exposure (F. Torres, das Gracas, Melo, & Tosti, 2009). On the other hand, non-essential heavy metals such as Cd^{2+}, mercury (Hg), lead (Pb), arsenic (As), and chromium (Cr), do not have a defined biological function. Our lack of homeostatic mechanisms after their absorption (Neathery & Miller, 1975), combined with their ability to accumulate and induce oxidative stress at low concentrations has been associated with organ failure, cytotoxicity, DNA damage and apoptosis (Patlolla, Barnes, Yedjou, Velma, & Tchounwou, 2009; Sutton, Tchounwou, Ninashvili, & Shen, 2002; Tchounwou et al., 2012; X. F. Wang, Xing, Shen, Zhu, & Xu, 2006; Yedjou & Tchounwou, 2006).

D. pulex, otherwise called the water flea, is a ubiquitous freshwater micro-crustacean (J. K. Colbourne et al., 2011). It is an important part of the aquatic food hierarchy placed above algae, bacteria and protozoans, and below fish (J. R. Shaw et al., 2007). Their sensitivity to environmental toxicants makes them an appropriate model organism to study the genetic basis behind toxicant induced stress (J. R. Shaw, Dempsey, Chen, Hamilton, & Folt, 2006; Joseph R. Shaw et al., 2008). Additionally, their compact genome of 200 Mb (J. K. Colbourne et al., 2011), easy maintenance in the laboratory and a recently published version of high quality transcriptome assembly with 23, 244 predicted transcript models (Ye et al., 2017), has furthered their appeal as a model organism to study ecotoxicology at a biomolecular level. Since, cellular biomolecules are the primary interactors of biotic and abiotic chemicals in an organism; studying ecotoxicology at a molecular level will provide us the following information in general:

- Identification of biomarker genes specific or common to toxicants.

- The role of biological pathways underlying the organism’s response against toxicants.
- Determining different modes of toxicity.
- The impact of toxicants on their physiology, population levels and evolution (Altshuler et al., 2011; H. J. Kim, Koedrith, & Seo, 2015; Schirmer et al., 2010).

The knowledge gathered from these studies will ultimately empower our evaluations on harmful effects of toxicants on humans and the environment, and aid in designing better preventive and counter measures (Eggen, Behra, Burkhardt-Holm, Escher, & Schweigert, 2004).

Daphnia is a key aquatic indicator of the inland water quality (Gunatilaka, Diehl, & Puzicha, 2001). Environmental protection agencies use *Daphnia* based toxicant responses from ecotoxicological studies to gauge the quality of industrial and municipal wastes, minimize their influx and impact on the environment (Joseph R. Shaw et al., 2007). Several transcriptomic studies have used *Daphnia* as an ecotoxicological model organism to characterize the molecular mechanisms behind toxicant induced genetic responses (Asselman, Shaw, Glaholt, Colbourne, & De Schamphelaere, 2013; De Schamphelaere et al., 2008; Orsini et al., 2017; Poynton et al., 2008; Poynton et al., 2007; Joseph R. Shaw et al., 2007; Soetaert, van der Ven, et al., 2007; Soetaert, Vandenbrouck, et al., 2007; Vandenbrouck, Soetaert, van der Ven, Blust, & De Coen, 2009). These studies comprise of the following:

- Transcriptome analysis by using cDNA microarray platforms primarily, and RNA-seq (Orsini et al., 2017).

- Studying genetic response of *Daphnia* at varying concentrations and time of exposure.

Since, gene expression patterns differ based on time of exposure, dosage and type of the toxicant, and homolog of a gene, it is crucial to explore these patterns at varying exposure periods and concentrations of diverse toxicants to obtain a comprehensive picture of *Daphnia* based genomic response (Asselman et al., 2013; Soetaert, Vandenbrouck, et al., 2007). Therefore, in this current study we are exploring the transcriptome of *D. pulex* to identify the underlying genetic pathways associated with heavy metal induced stress after 24 hours exposure using Cd^{2+}, Ni^{2+} and Zn^{2+}independently, at 20ug/L, 200ug/L, and 200ug/L concentrations that has not been studied before.

Genomic techniques such as microarrays and high-throughput next generation sequencing has facilitated research in ecotoxicogenomics (Bouetard et al., 2012; Schirmer et al., 2010). RNA-Seq is a high-throughput sequencing approach that allows an unbiased view of the transcriptome expression patterns both quantitatively and qualitatively over traditional array-based hybridization techniques (McGettigan, 2013; Z. Wang et al., 2009). In this study, we have applied RNA-Seq technique to quantify and profile the *D. pulex* transcriptome under heavy metal stress. Here, we have attempted to bioinformatically characterize the key biological processes and pathways responding to acute heavy metal stress using gene expression data, gene ontology data, published literature, and regulatory element binding sites in the promoter region of stress responsive genes.

Our results suggest, prominent classes of genes regulated after heavy metal exposure are participating in oxidative stress and xenobiotic metabolism, structural constituent of cuticle, and proteolysis processes. The transcript expression patterns exhibit significant regulation of gene families under all three phases of xenobiotic metabolism that has not been associated with *D. pulex* in the context of heavy metal stress before. Moreover, *in silico* detection of binding sites for known stress induced transcription factors such as aryl hydrocarbon receptor (AHR) (Dietrich, 2016) and NF-E2 related factor-1 (NRF1) (Ohtsuji et al., 2008; Schultz, Abdel-Mageed, & Mondal, 2010) in the promoter region of oxidative stress and xenobiotic metabolism genes imply the potential drivers behind the genetic response of *D. pulex* under toxicant stress. The data generated from this study is a crucial map to link acute stress response genetic pathways in *D. pulex* after heavy metal exposure.

MATERIALS AND METHODS

SAMPLE PREP

D. pulex RNA samples from treated (with Cd^{2+} at 20ug/L, Ni^{2+} at 200ug/L and Zn^{2+}at 200ug/L sub-lethal concentration (~LC01), independently) and untreated animals were obtained from Dr. Joseph R Shaw's lab using standard laboratory protocols (Joseph R. Shaw et al., 2007). Biological replicates (BR) are independent samples generated from independent experiments following the same treatment protocols (Consortium, 2011). Standard RNA-Seq experimental design requires at least three BRs per treatment group to overcome biological variability among transcript expression patterns (Consortium, 2011). Therefore, three BR per treated group and four BR for the untreated group were prepared.

Next, the TruSeq RNA library prep (v2 RevD) protocol was applied to process 500ng of total RNA from each sample. Finally, samples were sequenced in a paired-end manner to generate 100 bp sequencing reads, by first barcoding and then loading them onto a single flow cell lane of an Illumina HiSeq sequencer.

QUALITY CONTROL, PROCESSING AND MAPPING

Quality control of RNA-Seq data is a required step to avoid incorrect interpretation of the data (Sheng et al., 2017). The FastQC algorithm (Andrews, 2011), was used to check the quality of raw sequencing reads and detect data quality issues. A FASTA file of contaminants was downloaded from Tamir (2012), that included adapter and primer sequences used during Illumina high-throughput sequencing steps. Trimmomatic software (Bolger, Lohse, & Usadel, 2014) was used to trim sequencing read bases that matched to a contaminant sequence, maintain an average read quality score of at least 20 (error rate of 1 in 100, (Cliften, 2015)), and exclude reads shorter than 36 bp (Rozenberg et al., 2015). Trimmomatic categorizes reads that survived processing as a pair (paired) separately than reads where only one pair was retained (single). Therefore, we mapped paired and single reads to the *D. pulex* hybrid transcriptome (explained in the Results section) using TopHat v2.1.1 (Trapnell, Pachter, & Salzberg, 2009), independently. Finally, we merged the paired and unpaired alignments to produce one alignment file for each sample using samtools 0.1.18 (H. Li et al., 2009).

BR VALIDATION

Correlations are a standard measure to capture transcript expression pattern variability between BRs. eXpress-1.5.1 algorithm (Roberts & Pachter, 2013) was used to generate individual transcript expression data per sample. Next, samples were grouped based on their treatment and a pairwise Pearson correlation analysis was performed within samples of each group using a custom R script. Sample pairs with Pearson correlation of 0.97 or higher were retained for further analysis.

DIFFERENTIAL GENE EXPRESSION

Differential gene expression analysis allows us to quantify and detect transcript expression changes in treated over untreated samples. Cuffdiff v2.2.1 (Trapnell et al., 2013) was used to compare BRs 2 and 3 under each metal treatment to BRs 2 and 4 of the untreated category. Transcripts altered by at least one heavy metal exposure with a standard q-value cutoff of <0.05 were retained for our downstream analysis.

GENE ONTOLOGY AND PATHWAY TERM ASSIGNMENT

Gene ontology (GO) terms are annotations that characterize a gene product based on their involvement in a biological process (e.g. signal transduction, carbohydrate metabolism, etc.), molecular function (e.g. catalytic activity, proteolysis, etc.) and their location in a cellular or sub-cellular component (e.g. ribosome, nucleus, etc.) (Ashburner et al., 2000). InterProScan uses 14 member databases under the Interpro consortium (Robert D. Finn et al., 2017) as described in the Introduction chapter of this thesis, to informatically predict protein domains for each input protein sequence. These predictions also include associated GO term information for each protein domain. We used the InterProScan algorithm (Jones et al., 2014) to predict protein domains for gene products (proteins or amino acid sequences) in our hybrid transcriptome. Default threshold of e-value <0.001 (error rate of 1 in a 1000, (B. NCBI)) was applied to detect known protein domains in our sequences and their related GO and pathway terms. A copy of the 2016/17 InterPro consortium database was locally installed to improve the efficiency of the prediction process. InterProScan rejects amino acid sequences with multiple stop codons or '*', therefore, sequences with multiple stop codons were trimmed after their first stop codon to prevent run errors and under the assumption that this is the translation termination site of the transcript. Input sequence filtering and processing was done using a custom Python script. This resulted in 23,149 processed amino acid sequences. Only, one amino acid sequence was excluded under probable erroneous gene model assumption because it begun with a '*'. The default InterProScan output was in XML format, therefore a custom Python script was written to convert the output into a tabular format for easier access to the data.

Interpro also provides associated pathway terms for each predicted protein domain. The pathway databases available under the Interpro consortium include KEGG (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016) , MetaCyc (Caspi et al., 2016) and Reactome (Fabregat et al., 2016). These terms inform us of a protein's involvement in known biological pathways.

GO AND PATHWAY TERM ENRICHMENT ANALYSIS

To discover enriched biological processes underlying each heavy metal exposure in our samples we used the GOSeq tool in R package (Young, Wakefield, Smyth, & Oshlack, 2010). This package is designed for RNA-Seq data and accounts for biases associated with length and expression pattern of transcripts during the GO term enrichment analysis (Young et al., 2010). GO terms in the entire *D. pulex* hybrid transcriptome served as our background to test the statistical significance of each enriched GO term in the SDE transcript lists. A GO term with q-value <0.05 (to control false discovery) was called statistically significantly enriched (SSE). We divided the SDE transcript lists based on their regulation such as only up, only down, and up and down, to discover SSE GO terms specific to transcript regulation and metal treatment.

Similar strategy and tool was applied to detect significantly enriched pathways in our list of SDE transcripts.

TRANSCRIPTION FACTOR BINDING SITE IDENTIFICATION

Transcription factors (TFs) are proteins that regulate the transcription (process of making RNA from DNA) of a gene (Scitable, 2014b). They have distinctive DNA-binding domains that allow them to bind to specific non-coding genomic regions called promoters and enhancers that are present in the upstream or downstream sequence of a gene’s transcription start site (genomic location where transcription begins, typically at the 5’ end of a gene) (Scitable, 2014b). Enrichment of these TFs in the upstream region of genes involved in key stress associated pathways might aid us to identify the master regulators behind the metal stress induced genetic response in *D. pulex*.

Oxidative stress is one of the prominent class of genes altered against metal induced stress response in our dataset. Given metal induced ROS generation can cause lipid peroxidation, disrupt thiol group and calcium homeostasis, DNA damage, and oxidative stress (Valko, Morris, & Cronin, 2005), we decided to focus our analysis towards identifying stress associated regulatory element binding sites in the upstream region of our SDE transcripts. Position weight matrices (PWM) of known stress response TFs (Table 4.9) for all taxon in the JASPAR database, excluding Plantae and Fungi were downloaded (Mathelier et al., 2016) (Figure 4.2). These matrices contain frequencies of all 4 nucleotides (A, C, G, and T) at each position of a consensus sequence. Consensus sequences or motifs are short sequences with length typically ranging between 6 and 21 bp, and are widely used to represent specificity of TF binding (Stormo, 2000; Wasserman & Sandelin, 2004). To avoid redundancy in TF binding sites by similar TF PWMs we collapsed their binding sites under one representative TF. First, TF PWM profiles were grouped based on their similarity with other stress TFs in our list. Each stress TF PWM in Table 4.9, was matched against the JASPAR animal database (excluding Plantae and Fungi) using TOMTOM software (Gupta, Stamatoyannopoulos, Bailey, & Noble, 2007). Next, JASPAR matches for each input stress TF with an e-value <0.01 (error rate of 1 in a 100, (B. NCBI)), were overlapped with TFs from our list excluding itself to form a group of similar stress TF PWMs of variable widths and offsets (Table 4.14). This grouping was used for further analysis.

To decide on a threshold for upstream region length we collected intergenic regions from top 10 longest scaffolds in the *D. pulex* genome. Based on the average of the median length for these intergenic regions, we chose 800 bp as our search space to scan for TF binding sites (Table 4.3). Next, we collected up to 800 bp upstream region for all the transcripts in our hybrid transcriptome and accordingly trimmed them to avoid their overlap with neighboring transcripts on the same strand (Corà, Di Cunto, Provero, Silengo, & Caselle, 2004) (Figure 4.2). Upstream regions equal to or greater than 6 bp were retained to allow scan for the shortest TF in our list of 6 bp width (*Ahr::Arnt*).

According to Jayaram, Usvyat, and AC (2016), Find Individual Motif Occurrences (FIMO) (Grant, Bailey, & Noble, 2011) tool performs the best compared to other existing tools when searching for individual TF binding sites. Therefore, we used FIMO to detect the stress response TFs binding sites or motifs in the upstream sequences of SDE transcripts (Figure 4.2). Additionally, we applied a q-value filter of less than 0.2 from Peng, Hu, and Yang (2016) (Figure 4.2) to call a motif statistically significant. The statistical power of identifying biologically relevant motifs using FIMO depends entirely on the information content of the TF PWM and the background model (Grant et al., 2011). We prepared a background model for this analysis using the "fasta-get-markov" tool under the MEME suite (Bailey, Johnson, Grant, & Noble, 2015) and all the upstream sequences in our transcriptome (excluding the upstream sequences associated with the SDE transcripts). A short and non-distinctive motif of 6 bp width can exhibit a vast amount of good matches purely by random chance, however, identifying a motif of 10 bp width by random chance is extremely rare. To enrich our results with biologically relevant motifs (BRM), we matched the statistically significant motifs (q-value <0.2), against the animal JASPAR CORE 2016 database using the TOMTOM tool (Gupta et al.,

2007) (Figure 4.2). Motifs were considered biologically relevant, if they matched to the TF that was used to discover them or a TF within the same family (Peng et al., 2016), with an e-value<0.05 (error rate of 5 in a 100, (B. NCBI)) (Gupta et al., 2007) (Figure 4.2).

MOTIF BASED NETWORK, MODULARITY AND CONSENSUS CLUSTERING

To identify key transcriptional regulatory patterns in our dataset we build a motif based network using python's "Networkx" package (Hagberg, 2008), where the nodes and edges comprised of SDE transcripts containing at least one occurrence of a BRM and a pair of SDE transcripts sharing at least one BRM, in their upstream region, respectively. The edges were assigned a weight using Jaccard's index (total number of matching motifs between two transcripts /total number of unique motifs combining the two transcripts). Modularity is a metric that describes the strength of a network's partition into modules (Newman & Girvan, 2004). A high modularity would implicate denser connections within nodes of a module over their connections with the remaining network (Newman & Girvan, 2004). Since, we found shared motifs amongst transcripts in our dataset we decided to examine the possibility of groups of master stress regulators driving the expression of a set of transcripts under metal stress in our dataset. Therefore, we applied a popular clustering algorithm for this task called "Louvain modularity" using python's Louvain package. Additionally, it detects communities or groups in an unbiased fashion without requiring user input of k (desired number of clusters). The algorithm discovers communities in large networks using a two-step approach such as, first it optimizes modularity locally and detects little communities and next, it aggregates the nodes of little communities as one node, forms a new abstract network and reapplies the first step. It repeats these steps in iteration until the maximum modularity is reached and any further iteration will not improve this metric (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Given the stochastic nature of this algorithm we applied the method of consensus clustering to stabilize and improve the partitions of our motif based network (Lancichinetti & Fortunato, 2012) using a custom written python script. Consensus clustering of our network involved the following steps:

- Ran the algorithm on the same network 100 independent times and generated modularity each time.
- Due to the stochastic nature of this algorithm, it will output slightly different results each time. Therefore, we built a new network from the previous runs comprising of nodes from our original network and the edges between two nodes carrying a weight = total number of times these two nodes appeared in the same cluster out of 100 runs/100. Additionally, we compute the average modularity from the previous runs.
- Next, we repeated the first and second steps on the new network, iteratively, and compared the average modularity of the current iteration with the previous iteration until no significant improvement in average modularity could be made.

A partitioned network was built from our final consensus clustering iteration using a custom python script. Next, we computed some statistics such as total number of genes, up and down regulated transcripts, transcripts with a known GO domain and orphans, and GO and orphan transcripts with at least one shared FUCD appearing in the same cluster, per cluster using a custom python script. Additionally, we performed a hypergeometric test per cluster to determine the significance of enriched TFs based on their gene target frequency using (Graeber_lab, 2009). The obtained significance was adjusted for false discovery rate to identify significantly enriched TFs per cluster using the “p.adjust” function in R “stats” package (R Development Core Team, 2006). Next, the GO and orphan genes sharing the same FUCd^{2+}, for each individual cluster, were plotted along with their TF binding sites using RSAT feature map tool (Medina-Rivera et al., 2015) and a custom python script, to visually compare their motif profiles. Similar plot was drawn for the entire set of phase II xenobiotic metabolizing enzymes irrespective of their cluster assignments to inspect stress associated TF binding profiles. Finally, we compared the motif similarity distributions between GO and orphan transcripts sharing at least one FUCD domain and assigned to one cluster, and GO and orphan pairs sharing at least one FUCD domain but assigned to different clusters, based on their motif similarities (Figure 4.8).

RESULTS

D. PULEX HYBRID TRANSCRIPTOME

The *D. pulex* hybrid transcriptome is a combination of overlapping transcripts between the “Genes 2010 beta 3”   and the “JGI_2011_Frozen_Cat**”** transcript annotations published by wFleabase (Genome Informatics Lab, 2005) and Ye et al. (2017) respectively, and additional novel transcript models predicted by Ye et al. (2017). Transcript models were compared between the “JGI_2011_Frozen_Cat**”** and the “Genes 2010 beta 3” versions using custom Python scripts. 22,724 out of 23,210 total transcripts

(~97%) in the “JGI_2011_Frozen_Cat” version were entirely overlapping with 22,630 transcript models in the “Genes 2010 beta 3” version (Table 4.1). To further verify the common transcript models between the two releases, we mapped 100,000 paired end processed reads from one sample independently mapped to these transcriptome versions using TopHat v2.1.1. We found nearly 8% more mapped reads and 7% more uniquely mapped reads using the “Genes 2010 beta 3” version over the “JGI_2011_Frozen_Cat” version (Table 4.1). Based on these results, we created the *D. pulex* hybrid transcriptome involving 22,630 transcript models from the “Genes 2010 beta 3” version validated by RNA-Seq mapping and 520 novel transcripts from the “JGI_2011_Frozen_Cat” version. Since, “JGI_2011_Frozen_Cat” is at present the most accurate version of the *D. pulex* transcriptome, “Genes 2010 beta 3” transcripts that did not match entirely with any transcript model from this annotation were excluded from further analysis due to their potential invalidity.

MAPPING STATISTICS

On average, we found nearly 68.5% of total reads mapping unambiguously to known transcripts in the *D. pulex* hybrid transcriptome (Table 4.2). The RNA-Seq read alignment threshold for the human genome is between 70-90% (based on the quality of the data and the aligner used), but this percentage can be slightly lower if the reads are mapped to a transcriptome instead of genome due to the inability to align reads to novel genomic locations in the genome (Conesa et al., 2016). Additionally, the number of multi-mapping alignments significantly increase in transcriptome based mapping over whole genome, because sequencing reads can potentially map to overlapping exons between multiple transcript isoforms of a gene (Conesa et al., 2016).

BR VALIDATION ANALYSIS

For this analysis we preferred the eXpress algorithm to generate individual transcript expression data per BR because it claimed to be better at handling multi-mapping alignments, more efficient and accurate when quantifying transcript expression than other available tools (Roberts & Pachter, 2013). Since, multi-mapping alignments tend to increase in transcriptome based mapping (Conesa et al., 2016), we used this tool to quantify transcriptome expression in our dataset.

According to the ENCODE consortium a Pearson correlation coefficient between 0.92 and 0.98 is enough to consider compatible BRs for differential expression analysis. Pearson correlation coefficient is vulnerable to outliers. Additionally, transcript expression variability between BRs is commonly prevalent in lowly expressed transcripts (Consortium, 2011). To detect truly compatible BRs we examined their correlation scatterplots and chose a correlation coefficient of at least 0.97 or higher. According to the scatterplots in Figure 4.1, it is clear that BR 2 and 3 have an overall lower variability than their independent comparisons with BR 1, for each heavy metal treatment. We performed similar analysis for samples in our untreated cohort. We found BR 2 and 4 as the best correlating pair (Figure 4.1). Using a higher correlation coefficient allowed us to detect stable transcript expression patterns across the BRs of a sample and to counter the impact of outliers and highly variable BRs that might lead to incorrect interpretation of our analysis. Two BRs per treatment group and control group were retained for further analysis.

SIGNIFICANTLY DIFFERENTIALLY EXPRESSED GENES

We found 2,292 significantly differentially expressed (SDE) genes with a q-value <0.05, expressed in at least one heavy metal exposure (Table 4.10). Cd^{2+} exposure alone resulted in 2,090 SDE genes projecting a substantially higher level of toxicity inducing nearly 5-fold greater gene alterations than Ni^{2+} and Zn^{2+}(< 400 SDE genes under each exposure) (Table 4.10). 106 genes under Ni^{2+} and 82 genes under Zn^{2+}exposure shared the same gene regulation pattern as Cd^{2+} Additionally, a set of 46 and 96 SDE genes were found to be commonly up and down regulated against Cd^{2+}, Ni^{2+} and Zn^{2+}exposures, respectively and only 18 SDE genes showed variable regulation patterns across metals.

PROTEIN DOMAIN AND PATHWAY TERM PREDICTIONS

We predicted protein domains for nearly 67% of genes in our hybrid transcriptome (Table 4.4). 76% of the 67% had a GO term assigned to them and the remaining genes were orphans that contained functionally uncharacterized protein domains (FUCD) (Table 4.4). Out of the 2,292 SDE genes, ~57% had a GO term assigned to them and 13% were orphans (Table 4.4). The remaining genes did not have a predicted protein domain due to the following reasons:

a) Predicted domain had an e-value >0.001.

b) No known domain exists for the amino acid sequence suggesting the protein is specific to *D. pulex*.

Interestingly, some GO genes in addition to containing protein domains with an assigned GO term also carried FUCDs (Table 4.5). We found shared FUCDs between 452 and 200, SDE GO and orphan genes, respectively (Table 4.5).

ENRICHED GO AND PATHWAY TERMS

We found 38 significantly enriched (SE) GO terms with a q-value <0.05 (Table 4.11). Cd^{2+} was the dominant metal among the three heavy metals covering 100% of SE GO terms. Ni^{2+} and Zn^{2+}had 4 and 1 overlapping SE GO terms with Cd^{2+}, respectively. Based on Figure 4.3 below, nearly 60% (23/38), 75% (18/24), 65% (17/26), and 66% (25/38) of these GO categories contained genes from both up and down regulation patterns for Cd^{2+}, Ni, $Zn^{2+}$and all the metals combined, respectively. This pattern suggests a mixed transcript expression pattern amongst the highlighted biological process responding to heavy metal exposure.

Prominent SE categories of affected GO terms involved, oxidation-reduction process, response to oxidative stress, peroxidase activity, oxygen binding and transport, heme binding and transport, hemoglobin complex, carbohydrate metabolism, iron transport and homeostasis, cell cycle and DNA repair, chitin binding and metabolism, cuticle and reproduction , calcium ion and apoptosis, steroid biosynthesis. Additionally, we discovered xenobiotic metabolism pathway that was substantially affected due to heavy metal exposure that has not been reported previously at a comprehensive level in

D. pulex (Table 4.12). These categories are discussed in detail in the Discussion section of this chapter.

Nearly, 19% of our genes were assigned a pathway term suggesting their participation in a known biological pathway. 9% of the 19% were present in our SDE gene list (Table 4.6)

REGULATORY ELEMENT BINDING SITE ANALYSIS AND CONSENSUS CLUSTERING

We detected 4891 unique biologically relevant TF binding sites for 1622 out of 2292 SDE transcripts involving 17 out of 19 stress associated transcription factors. We did not find any relevant binding sites for hypoxia-inducible factor 1-alpha (*HIF1A*) and aryl hydrocarbon receptor nuclear translocator (*Arnt*) using our current methodology. We found forkhead box protein o (*Foxo1/FOXO3/6*), aryl hydrocarbon receptor and aryl hydrocarbon receptor nuclear translocator (*Ahr::Arnt*) heterodimer complex, activating transcription factor 3 (*Atf3*), metal transcription factor (*MTF1*), aryl hydrocarbon receptor nuclear translocator like (*Arntl*), and nuclear factor kappa-light-chain-enhancer of activated B cells (*NF-κB* (*RELA*)) targeting nearly 200 and higher number of SDE transcripts in our dataset (Table 4.7). The binding sites for all TFs were evenly distributed across up and down regulated target transcripts and did not implicate TF binding specificity towards a particular transcript expression pattern (Figure 4.9). However, nuclear respiratory factor 1 (*NRF1*) and *Arntl* binding sites were found relatively high among up and down regulated SDE transcripts (Table 4.7), respectively. Biologically relevant binding sites for TFs such as CCAAT/enhancer-binding protein (*CEBP A*,*B*), *Ahr::Arnt*, *Foxo1/FOXO3/6*, *Atf* TF family, *NRF1*, *MTF1*, *NF-κB* (*RELA*) were detected in the upstream region of known oxidative stress response genes such as copper-zinc cu-Zn^{2+}superoxide dismutase (SOD), carbonyl reductase (CBR3), glutathione s-transferase (GST), aldo-keto reductase (AKR), peroxinectin, sulfotransferase (SULT), Prostaglandin G/H synthase 2 (PTGS2 or COX2), chorion peroxidase, NADPH-cytochrome P450 oxidoreductase (POR), and GSH (Table 4.8).

As described in the materials and method, SDE genes were clustered based on their TF binding site profile using consensus clustering method. A maximum modularity of 0.72 was reached after 4 iterations of consensus clustering (Figure 4.7). This resulted in 5 clusters with cluster 1 being the most populated cluster comprising of 550 transcripts including 328 transcripts with a GO term, 69 transcripts with FUCDs, 153 transcripts with no known protein domain, 307 up regulated and 246 downregulated transcripts (Table 4.13). Cluster 1 and 2 contained majority of GO and orphan pairs with at least one common FUCD in the same cluster (Table 4.13). The up and down regulated transcripts were evenly distributed across all the clusters exhibiting no specificity between TF binding sites and expression pattern of transcripts (Table 4.13). Consensus clustering and hypergeometric test allowed us to detect significantly enriched TFs per cluster.

Hypergeometric analysis for detecting enriched TF target genes per cluster revealed the following results:

- *ATF* family TF targets were significantly enriched in cluster 1 (Table 4.15).
- *Ahr::Arnt* target genes was significantly enriched in cluster 2 (Table 4.15).
- *Foxo* family TFs targets were significantly enriched in cluster 3 (Table 4.15).

- *MTF1* and *Ddit3::Cebpa* targets were found significantly enriched in cluster 4 (Table 4.15).
- *NF-κB (RELA), Arntl, CEBPA/B, NRF1,* and *Pparg::Rxra* were significantly enriched in cluster 5 (Table 4.15).

The motif similarity distribution of GO and orphan genes sharing at least one FUCD and clustered as a pair was substantially different than those that did not clustered as a pair (Figure 4.8). The plot shows large amounts of GO and orphan FUCD pairs that clustered separately were due to absence of overlapping motifs between them (Figure 4.8).

Lastly, the feature map plot of TF binding sites between GO and orphan pairs sharing at least one FUCD and assigned to the same cluster revealed high similarity in motif profiles as illustrated for cluster 2 (Figure 4.6). All the GO and orphan pairs in this cluster have at least one *Ahr::Arnt* binding site in common (Figure 4.6) which agrees with our result of cluster 2 being significantly enriched with *Ahr::Arnt* targets (Table 4.15).

DISCUSSION

SIGNIFICANTLY DIFFERENTIALLY EXPRESSED TRANSCRIPTS

We found >5 fold differentially regulated transcripts in our Cd^{2+} treated samples over Ni^{2+} and Zn^{2+} treatments (Table 4.10). Garrett, Somji, Sens, Zhang, and Sens (2011) study of gene expression patterns under Cd^{2+} stress using a human renal epithelial cell culture model (HPT) and microarray analysis (GeneChip Human Genome, U133 Plus 2.0 arrays, Affymetrix), resulted in 1,848 SDE genes after 24 hour exposure to Cd^{2+} at 9, 27, and 45 μM concentrations and 923 SDE genes after 13 days exposure to Cd^{2+} at 4.5, 9, and 27 μM. There were only 387 overlapping genes between the two exposures. Their results suggested that during acute Cd^{2+} exposure cellular stress window is highly active and that might have influenced the large alterations in gene expression (Garrett et al., 2011). Cd^{2+} is known to induce oxidative stress in cells after acute exposure by elevating reactive oxygen species (ROS) production such as hydroxyl radical and superoxide ions, and depleting antioxidant levels such as glutathione (GSH) and protein-bound thiol groups, leading to lipid peroxidation and DNA damage (Bagchi et al., 1997; F. Liu & Jan, 2000; Manca, Ricard, Vantra, & Chevalier, 1994). Due to their higher affinity to thiol (-SH) groups than essential heavy metals such as Cu^{2+} and Zn^{2+}, they are more cytotoxic in nature (Buffle, Chalmers, Masson, & Midgley, 1991; Skreb & Fischer, 1984). Therefore, they can greatly influence cell physiology by interacting and inhibiting these groups in essential enzymes and proteins (Kinraide & Yermiyahu, 2007; Webb, 1979). Cd^{2+} is known to stimulate Metallothionein (MT) synthesis, a low molecular weight cysteine rich metal ion binding protein, and form Cd^{2+}-MT complexes due to their thiol group abundance. MT helps sequester Cd^{2+} as a detoxification mechanism (Yang & Shu, 2015). Kidneys are the major Cd^{2+}- sequestration organs where the half-life period of Cd^{2+} is up to 10 years or higher (Godt et al., 2006; Orlowski & Piotrowski, 2003). Due to small size of MT the Cd^{2+}-MT complex can filter through the glomerulus of the kidney and get reabsorbed via endocytosis into the epithelial cells of the proximal tubule (Yang & Shu, 2015). Therefore, the Cd^{2+}-MT complex degrades rapidly in the epithelial cells and Cd^{2+} binds to newly synthesized MT in the cytoplasm (Klaassen, Liu, & Choudhuri, 1999). At higher concentrations, Cd^{2+}-MT binding can saturate and contribute to cytotoxicity (Yang & Shu, 2015). Intracellular organelles such as lysosomes act as detoxification and degradation components in vertebrate and invertebrate cells (Longhurst, 1988; Sterling et al., 2007), by accumulating essential and non-essential heavy metals, and catabolizing exogenous and endogenous molecules (Dingle & Dean, 1976; Glaumann & Ballard, 1987) Glaumann and Ballard, 1987) respectively. In invertebrates, lysosomes play a key role in detoxification of non-essential heavy metals such as Pb^{2+}, Cd^{2+}, and Hg^{2+} by lowering their availability due to long-term sequestration through precipitation, and subsequent excretion (Sterling et al., 2007). However, an *in vitro* study of livers isolated from male Sprague Dawley rat analyzing the role of cathepsins (specific acidic proteases that degrade MT) in degrading MT- (Zn^{2+}or Cd^{2+}) complexes found, at lysosomal pH most Cd^{2+} is not easily released from Cd^{2+}-MT complex (McKim, Choudhuri, & Klaassen, 1992). This led researchers to conclude that the Cd^{2+}-MT complex has a higher half-life period in cells than Zn^{2+}-MT complexes (McKim et al., 1992). Therefore, longer half-life period, reabsorption, and stronger affinity to thiol groups can collectively describe lower excretion rate of Cd^{2+} and their ability to induce cytotoxicity and strong genetic response than Ni^{2+} and Zn^{2+}.

CADMIUM, CALCIUM AND APOPTOSIS

Calcium (Ca^{2+}) is an abundant element in our body that plays a key role in the proper functioning of biological processes such as nerve impulse transmission, regulating muscle contraction and relaxation, clotting of blood, signaling cascades and secretion of hormones (Choong, Liu, & Templeton, 2014; Metheny & Metheny, 2012). Intracellular

Ca^{2+} levels are tightly controlled by Ca^{2+} channel and exchanger proteins (Choong et al., 2014). Disturbance in calcium regulation can contribute to a variety of conditions such as hypocalcemia, hypercalcemia and osteoporosis (Metheny & Metheny, 2012). Cd^{2+} and Ca^{2+} are divalent cations with analogous physiochemical properties that allows their exchange in Ca^{2+} binding proteins. Cd^{2+} has been shown to displace Ca^{2+} ions in proteins such as calmodulin (CaM) (Chao, Suzuki, Zysk, & Cheung, 1984), sarcolemma (Langer & Nudd, 1983) and troponin C (Ellis, Strang, & Potter, 1984). One of the processes linked to increase Cd^{2+} uptake in cells is through voltage gated calcium channels (VDCC). Cd^{2+} inability to induce cellular toxicity has been demonstrated in HeLa cells lacking VDCCs (Gavazzo, Morelli, & Marchetti, 2005). We found all 5 SDE VDCC genes upregulated under Cd^{2+} exposure in our dataset. Cd^{2+} is known to cause increase in Ca^{2+} concentrations in specific cell types via induction of inositol trisphosphate (IP3) that releases stored intracellular Ca^{2+} (Choong et al., 2014) (Figure 4.5). Ca^{2+}-dependent phospholipase C (PLC) enzyme cleaves phosphatidylinositol 4, 5-bisphosphate (PIP_2) to generate signaling molecules such as IP3 and diacylglycerol (DAG) (Falkenburger, Jensen, Dickson, Suh, & Hille, 2010). Inhibition of PLC lowered Cd^{2+} induced Ca^{2+} concentration rise in *Xenopus* oocytes (Hague, Matifat, Louvet, Brûlé, & Collin, 2000). Additionally, phosphatidylinositol 4-kinase alpha (PI4KA) participates in the synthesis of inositol 1, 4, 5-trisphosphate (Gehrmann et al., 1999) and inositol 1, 4, 5-trisphosphate receptor type 1 is stimulated by inositol 1, 4, 5-trisphosphate to facilitate intracellular Ca^{2+} release from endoplasmic reticulum (ER) (Gerber et al., 2016). Diacylglycerol kinase delta (DAGKD) is an enzyme involved in phosphorylation of DAG into phosphatidic acid (PA) (Harada et al., 2008). DAG and PA are signaling molecules that have been shown to activate protein kinase C (PKC) (Jornayvaz & Shulman, 2012; Lang, Malviya, Hubsch, Kanfer, & Freysz, 1995). PKC is known to phosphorylate cellular biomolecules resulting in activation of NADPH-oxidase (NOX, ROS generator) and apoptosis pathways (Dekker et al., 2000; Reyland, 2007) (Figure 4.5). NOX produced ROS is a part of cellular defense mechanism against pathogens, however, the ROS leakage from phagosome into the cytosol has been associated oxidative stress (Morgan & Liu, 2011) (Figure 4.5). We found phospholipase C epsilon (PLCE), IPTR1, PI4KA, DAGKD, PKC delta, and NADPH-oxidase in our SDE gene list upregulated against Cd^{2+} exposure suggesting the role of Cd^{2+} in altering cellular structure and function by targeting Ca^{2+} regulation (Figure 4.5). Plasma membrane Ca^{2+} ATPases (PMCAs) are P-type ion pump enzymes that regulate the ionic gradients such as (Ca^{2+}, Na^{+}/K^{+} or H^{+}) across cell membranes (Møller, Juul, & le Maire, 1996; Palmgren & Nissen, 2011). Cd^{2+} can cause inhibition of Ca^{2+} efflux in erythrocytes via non-competitive inhibition of Ca^{2+}-ATPases (Akerman, Honkaniemi, Scott, & Andersson, 1985; Visser, Peters, & Theuvenet, 1993). However, after 24 hrs Cd^{2+} exposure in *D. pulex*, we found significantly high expression of Plasma membrane calcium-transporting ATPase 3 (PMCA3). In a study where zebrafish larvae were exposed to 0.08 μM Cd^{2+} in water containing 0.2 mM Ca^{2+} (low calcium environment), PMCA2 expression decreased significantly after 96 hours and increased significantly after 72 hours in water containing 2 mM Ca^{2+} (high calcium environment) (C. T. Liu, Chou, Lin, & Wu, 2012). This study showed a restoration of PMCA2 gene expression based on the Ca^{2+} to Cd^{2+} concentration ratio. The Cd^{2+} induced intracellular Ca^{2+} release in our case is possibly allowing the significantly high expression of PMCA.

Ca^{2+} /CaM -dependent protein kinases (CaMK) are effectors that are known to sustain calcium signaling cascades. Increased expression of CaMK-II are linked to cardiac hypertrophy (Maier, Bers, & Brown, 2007) and reoxygenation injury (Vila-Petroff et al., 2007) in humans. Signaling cascades downstream of CaMK-II include Mitogen activated protein kinases (MAPKs) that may induce cellular pathways such as pro-survival and apoptosis (Choong et al., 2014; Erickson, He, Grumbach, & Anderson, 2011). Cd^{2+} induced ROS and intracellular Ca^{2+} release can activate CaMK-II (Choong et al., 2014; Y. Liu & Templeton, 2007). We found CaMK-II and MAPK9 (GO terms include, cellular response to Cd^{2+} , cellular response to ROS and positive regulation of apoptotic process, (Uniprot, 2018)) significantly upregulated after Cd^{2+} exposure in our dataset suggesting an active apoptotic process. Protein phosphatase 2A (PP2A) is known to induce apoptosis via activating pro-apoptotic proteins and suppressing anti-apoptotic proteins. However, a subset of PP2A enzymes have been linked to anti-apoptotic process in *Drosophila* (Van Hoof & Goris, 2003). Additionally, L. Chen, Liu, and Huang (2008) showed Cd^{2+} induced ROS causing inhibition of PP2A and PP5 and neuronal cell death via MAPK based activation of c-Jun N-terminal kinase (JNK) and extracellular signal-regulated kinase 1/2 (Erk1/2). PP2A and PP5 can negatively regulate Erk1/2, JNK and p38 (S. Huang et al., 2004; Y. Liu, Shepherd, & Nelin, 2007; Morita et al., 2001; Van Kanegan, Adams, Wadzinski, & Strack, 2005) . In our dataset, we found PP2A significantly upregulated. MAP kinase-interacting serine/threonine-protein kinase 2

(MKNK2) encodes a Ca^{2+} /CaM dependent protein kinase and is a downstream target of MAPK (RefSeq, 2011). It has been shown to participate as a negative regulator of arsenic-trioxide induced apoptosis in leukemia cell lines (Dolniak et al., 2008). Bifunctional apoptosis regulator (BFAR) is another anti-apoptotic gene that has been shown to protect neurons from multiple apoptotic pathways (W. Roth et al., 2003). We found both of these genes significantly downregulated in our dataset against Cd^{2+} exposure. Eleawa et al. (2014) have shown downregulation of an anti-apoptotic regulator called Bcl-2 in rat testes after $CdCl_2$ exposure. Our data shows significant downregulation of Bcl-2 against Cd^{2+} exposure. This suggests a more positive regulation of apoptosis in *D. pulex* after 24 hr Cd^{2+} exposure.

LIPID PEROXIDATION AND RESPONSE TO OXIDATIVE STRESS

Cd^{2+} is known to induce oxidative stress and reduce the antioxidant substrate glutathione, modify antioxidant enzymes and cause structural damage to the cell through lipid peroxidation (Bagchi et al., 1997; Manca et al., 1994). Primary targets of lipid peroxidation involve cholesterol, membrane glycolipids, and phospholipids (Ayala, Munoz, & Arguelles, 2014) (Figure 4.5). Arachidonic acid (AA) is a polyunsaturated fatty acid that binds to plasma membrane phospholipids and is a precursor to metabolites such as prostaglandins, hydroepoxyeicosatrienoic acids (HETEs) and leukotrienes that are involved in inflammatory signaling (Elabdeen et al., 2013; Lyons, Tovar-y-Romo, Thakur, McArthur, & Haughey, 2015; D. Wang & Dubois, 2010) (Figure 4.5). Intracellular Ca^{2+} release has been associated with increase in arachidonic acid production in parathyroid cells (Almaden et al., 2002) (Figure 4.5). Phospholipase A2 is an enzyme that uses Ca^{2+} as a cofactor to hydrolyze membrane phospholipids and release lysophospholipids and AA (Cupillard, Koumanov, Mattei, Lazdunski, & Lambeau, 1997; Gunawardena, Govindaraghavan, & Münch, 2014; Pan et al., 2002) (Figure 4.5). Lysophospholipases are enzymes that are known to degrade cytotoxic lysophospholipids (Weller, Bach, & Austen, 1984). Prostaglandin G/H synthase 2 or Cyclooxygenase 2 (COX2) metabolizes arachidonic acid (AA) into prostaglandin H_2 and generate superoxide radicals during this process that contributes towards oxidative stress (Morgan & Liu, 2011; D. Wang & Dubois, 2010) (Figure 4.5). COX2 up regulation has been linked to inflammatory response, resistance to apoptosis and tumor progression, and cell adhesion (S. F. Kim, Huri, & Snyder, 2005). Prostaglandin D synthase further metabolizes PGH2 to generate prostanoids such as prostaglandin D_2, E_2, and $F_{2\alpha}$ ($PGD_{2,}$ PGE_2, and $PGF_{2\alpha}$ respectively), and thromboxane A_2 such as TXA_2. Additionally, PTGD2 has also been shown to enable xenobiotic and endogenous compound detoxification through glutathione conjugation in *Nilaparvata lugens* (Yamamoto et al., 2017). We found group IID secretory phospholipase A2 (PA2GD), group 10 secretory phospholipase A2 (PA2GX), eosinophil lysophospholipase (LPPL), COX2, and glutathione-requiring prostaglandin D synthase (PTGD2) significantly up regulated against Cd^{2+} in our dataset (Figure 4.5).

Nearly 30% (21/69) of genes involved in the oxidative stress response in *D. pulex* were found to be SDE. 66% of those 30% were up regulated against Cd^{2+} exposure only**.** These genes include COX2, peroxidasin, phospholipid hydroperoxide glutathione peroxidase (GPx), peroxinectin and chorion peroxidases. GPx is an antioxidant enzyme that protects cell membrane against lipid peroxidation and is also involved in detoxification of xenobiotics (Esworthy, Doan, Doroshow, & Chu, 1994; Sidhu, Sharma, Bhatia, Awasthi, & Nath, 1993). Peroxidasin gene encodes for a heme-containing peroxidase that aids in consolidation of the extracellular matrix, defense response and phagocytosis of cells undergoing apoptosis (Nelson et al., 1994). Peroxinectin is linked to cell adhesion function in the black tiger shrimp, *Penaeus monodon* (Sritunyalucksana, Wongsuebsantati, Johansson, & Soderhall, 2001). Cell adhesion and differentiation was one of the affected categories of genes after an acute 24 hr Cd^{2+} exposure in HPT cells in Garrett et al. (2011). Shrimp peroxinectin has 51% cDNA similarity to *Drosophila melanogaster* peroxinectin and peroxidasin sequences, especially to their peroxidase domains (Sritunyalucksana et al., 2001). Cd^{2+} is known to induce generation of ROS such superoxide ion, H_2O_2 and hydroxyl radicals (Stohs & Bagchi, 1995). Peroxidases primarily break down toxic hydrogen peroxides (H_2O_2) into water and oxygen (Flohe & Ursini, 2008). Chorion peroxidases are known to be expressed throughout the life of adult female insects to protect their eggs against oxidative stress by contributing towards the rigid, insoluble egg chorion (Tootle & Spradling, 2008; Tufail & Takeda, 2012). Glutathione s-transferases (GST) have also been known to play a major role in protecting cells against several xenobiotic agents (Salinas & Wong, 1999). GSTs conjugate glutathione to toxic electrophilic compounds produced during membrane oxidation processes such as 4-hydroxynonenal and cholesterol α-oxide (Danielson, Esterbauer, & Mannervik, 1987; Hubatsch, Ridderstrom, & Mannervik, 1998). Overall, the conjugation process makes the highly toxic compounds less reactive and water soluble (Veal, Toone, Jones, & Morgan, 2002). We found 8 GSTs in our SDE list and 7/8 were upregulated against Cd^{2+} exposure only (Figure 4.5). Additionally, the first rate-limiting enzyme in synthesizing glutathione, the Glutamate-cysteine ligase catalytic subunit (Siegmund et al., 2011), was also upregulated under Cd^{2+} exposure (Figure 4.5). Glutathione up regulation has be directly associated with Cd^{2+} chelation and detoxification and antioxidant response against oxidative stress (Delalande et al., 2010; Jozefczak, Remans, Vangronsveld, & Cuypers, 2012) (Figure 4.5).

XENOBIOTIC METABOLISM PATHWAY

Humans encounter xenobiotics via exposure to drugs, environmental pollutants, processed food, chemicals used in agriculture and cosmetic products. Generally, xenobiotics are lipophilic in nature and cause toxicity by accumulating in the body (Davies, 2007). Xenobiotic metabolizing enzymes (XME) eliminate most xenobiotics in discrete phases such as phase 1, 2 and 3 (Davies, 2007; Zhang et al., 2013). A nuclear receptor that regulates expression of phase I genes is aryl hydrocarbon receptor (Ahr) (Xu, Li, & Kong, 2005). As shown in chapter 4 “Arms Race Between Plants and Animals: Biotransformation System” of Steinberg (2012), xenobiotics enters the cell and binds to Ahr in the cytoplasm. Ahr is then translocated to the nucleus after activation by transport factors (Steinberg, 2012). The activated complex interacts with the xenobiotic response element (XRE) in the upstream of its target genes and allows their transcription (Steinberg, 2012). Transcription factors in cap-n-collar family such as NRF1 and NRF2 target antioxidant response elements (ARE) sequence and regulate the expression of oxidative stress response or phase II genes

(Biswas & Chan, 2010; Itoh et al., 1997). Additionally, Ahr can interact with both XRE and ARE sequences (Kohle & Bock, 2009). We found both Ahr and NRF1 genes significantly upregulated against Cd^{2+} exposure in *D. pulex*. Since, xenobiotics are mostly hydrophobic in nature; XMEs convert them into hydrophilic derivatives for easier elimination (Davies, 2007).

In phase I, xenobiotic compounds are modified by adding a reactive group such as hydroxyl radical (-OH), carboxylic acid (-COOH), sulfanyl group (-SH), or amine group (NH2) (Davies, 2007; Liska, 1998). Enzymes that participate in these reactions are cytochrome P450 (CypP450), flavin-containing monooxygenases (FMO), carboxylesterases (CES) and epoxide hydrolase (Davies, 2007; Liska, 1998; Zhang et al., 2013). We found CypP450 (4 transcripts) significantly upregulated against Cd^{2+} exposure in *D. pulex*. Among them were, CYP2J2 and CYP46A that are known to oxidize lipid metabolites such as AA and cholesterol into other pro-inflammatory and cytotoxic metabolites such as HETEs and 24S-hydroxycholesterol, respectively (C. Chen & Wang, 2013; Lutjohann & von Bergmann, 2003; NCBI, 2016; Yamanaka, Urano, Takabe, Saito, & Noguchi, 2014) (Figure 4.5). CypP450 perform oxidation of substrates by using O2 (due to the presence of a heme molecule in their active site) and H^+ from nicotinamide adenine dinucleotide phosphate (NADPH) supplied by NADPH-cytochrome P450 oxidoreductase (Davies, 2007). We found NADPH-CypP450 reductase (alias for NADPH-cytochrome P450 oxidoreductase) significantly upregulated in our dataset against Cd^{2+} treatment. In an uncoupled reaction CypP450 may consume more O_2 for metabolizing a substrate and produce ROS such as activated oxygen radical O_{2-}(Davies, 2007). Superoxide dismutase (SOD) can protect the cell from oxidative damage by degrading ROS into less reactive metabolites and hydrogen peroxide (Y. Li et al., 1995; Morgan & Liu, 2011). Our analysis shows SOD significantly upregulated against Cd^{2+} treatment. Inability to metabolize phase I intermediates can cause oxidative damage to DNA, RNA and proteins (Davies, 2007; Liska, 1998). Therefore, phase II enzymes participate in the antioxidant pathway through conjugation reactions that stabilizes the reactive intermediary metabolites from phase I and converts them into hydrophilic compounds for easier elimination via excretory mechanisms or phase III (Liska, 1998) (Figure 4.5).

Phase II enzymes include aldo-keto reductase (AKR), γ-glutamylcysteine synthetase (GCL), glutathione peroxidase (GPX), glutathione s-transferase (GST), heme oxygenase 1 (HO-1), menadione reductase (NMO), N-acetyltransferase (NAT), NADPH quinine oxidoreductase 1 (NQO-1), peroxiredoxin (PRX), sulfiredoxin (SRXN), sulfotransferase (SULT), thioredoxin (Trx), glucose-6-phosphate dehydrogenase, thioredoxin reductase (TrxR), carbonyl reductase [NADPH] 3 (CBR3), and UDP-glucuronosyltransferase (UGT) (Barski, Tipparaju, & Bhatnagar, 2008; Zhang et al., 2013). AKRs are involved in oxidation-reduction reactions during detoxification process and intermediary metabolism. Substrates of AKRs include both endogenous and exogenous molecules such as drugs, environmental pollutants, steroids, lipid peroxidation metabolites and glycosylation products (Barski et al., 2008). Lipids are the barrier between cell and the extracellular environment and crucial in cell signaling process (Larregle et al., 2008). Cd^{2+} causes free radical damage to the cell membrane by lipid peroxidation (Bagchi et al., 1997; Manca et al., 1994). Lipid peroxidation of inner mitochondrial membrane can disrupt mitochondrial membrane integrity (L. Muller, 1986). AKR upregulation likely suggests detoxification of genotoxic metabolites because of lipid peroxidation. GCL plays a crucial role in the biosynthesis and homeostasis of antioxidant glutathione (GSH) (Hibi et al., 2004). GST catalyzes the transfer of GSH to reactive electrophiles from phase I and prevents their interaction with cellular macromolecules (Davies, 2007). GPX protects cell from oxidative damage and lipid peroxide induced toxicity (Yant et al., 2003). Sulfotransferases (SULT) and UDP-glucuronosyltransferase (UGT) add a sulfonate and glucoronic acid moiety to increase the water solubility of a xenobiotic compound (Gamage et al., 2006; Steinberg, 2012) 2012). UGT have also been shown to conjugate with pro-inflammatory and unstable AA metabolites such as HETEs and leukotriene B4 (LTB4) and remove them from our bodies (Turgeon et al., 2003). NADPH participates in oxidation-reduction reactions such as protecting cells against ROS toxicity by acting as a coenzyme with glutathione reductase to convert glutathione disulfide (GSSG) (oxidized) reduction to GSH (reduced) (Deneke & Fanburg, 1989; Rush et al., 1985). Glucose-6-phosphate dehydrogenase (G6PD) helps to reduce $NADP^+$ (oxidized) to NADPH (reduced). CBR3 participates in conversion of endogenous molecules such as steroids and prostaglandins, and exogenous active carbonyl derivatives into their respective alcohols (Miura, Nishinaka, & Terada, 2008). We found AKR (4 transcripts), glutamate--cysteine ligase catalytic subunit (alias GCL, GSH), GPX4, GST (6 transcripts), SULT (8 transcripts), UGT (7 transcripts), G6PD and CBR3 significantly upregulated genes in our dataset against Cd^{2+} exposure (Figure 4.5).

Given phase II genes are involved in antioxidant defense and reduction of reactive and oxidized intermediates from phase I, we decided to look in their TF binding sites for potential similarities. Our clustering results showed phase II genes are scattered across different clusters, most likely due to presence of variable combinations of TF binding sites in their upstream regions (Figure 4.10). Nonetheless, *Ahr::Arnt* and *Atf* binding sites were prominently present among these genes (~64%, 18/28) and all phase II genes with *Atf* binding sites were grouped under the same cluster (cluster 1, ~36%, 10/28). These *Atf* target genes include Ahr, AKR, GST, PTGD2, SULT, and UDP. Our results, point towards a combinations of TFs likely causing a variety of antioxidant response based on specific xenobiotics and reactive endogenous substrates. Further research is required to verify the molecular mechanisms underlying the interplay between TFs and antioxidant defense response target genes.

Phase III is considered as antiporter activity that involves genes such as organic anion transporting polypeptide 2 (OATP2) and ATP binding cassette transporters (ABC) such as multidrug resistance associated protein (MRP), ABCG2, p-glycoprotein (P-gp) (Glavinas, Krajcsi, Cserepes, & Sarkadi, 2004; Zhang et al., 2013). These act as xenobiotic efflux pumps that participate in exporting intracellular signaling molecules, toxic metabolites and xenobiotic compounds (Chin, Pastan, & Gottesman, 1993; Toyoda et al., 2008; D. Wang & Dubois, 2010) (Figure 4.5). OATPs can transport large hydrophobic anions into the cell, MRPs can eject uncharged hydrophobic molecules and hydrophilic anions, and ABCG2 and P-gps can remove large positively charged molecules that are insoluble in water (Glavinas et al., 2004; M. Roth, Obaidat, &

Hagenbuch, 2012). We found Cystic fibrosis transmembrane conductance regulator (CFTR, member of MRP family) and Canalicular multispecific organic anion transporter 2 (MRP3), and ABCG2 significantly upregulated in our dataset against Cd^{2+} exposure. This suggests activation of all phases of xenobiotic metabolism pathway to protect cells from Cd^{2+} induced oxidative stress and toxic metabolites in *D. pulex*. The remaining genes in the phase I, II and III absent in our SDE list either did not change significantly or did not have a homologue in our dataset.

Clustering of TF binding site profiles revealed a set of phase I, II and III genes grouped under a single cluster most likely due to the presence of common *Atf* motifs among them (cluster 1, Figure 4.11). Our results show a strong enrichment of *Atf* binding sites among oxidative stress response and xenobiotic metabolism genes. This may suggest its major involvement in regulating heavy metal induced stress response transcription in *D. pulex*.

REPRODUCTION, CHITIN AND EXOSKELETON

Vitellogenin is the precursor from which egg yolk proteins are derived (Wiley & Wallace, 1981). We found 3 *vitellogenin-1* transcripts in our SDE gene list. Two out of the three transcripts are downregulated after exposure to Cd^{2+} This result is in agreement with downregulation of *vitellogenin* observed in microarray analysis after a 24 hr Cd^{2+} (18 ug/L) exposure in *Daphnia magna* (*D. magna*) by Poynton et al. (2007) and after *D. magna* exposure to Cd^{2+} for 96 hrs by Soetaert, Vandenbrouck, et al. (2007). Cervera, Maymo, Martinez-Pardo, and Garcera (2006) study also showed downregulation of *vitellogenin* in the large milk weed bug (*Oncopeltus fasciatus*) after Cd^{2+} exposure.

Moreover, Cd^{2+} exposure has also been shown to reduce reproduction in *D. magna* (Baillieul, Smolders, & Blust, 2005; Bodar, Van Leeuwen, Voogt, & Zandee, 1988). Therefore, the downregulation of *Vitellogenin* might suggest an impact of Cd^{2+} exposure on *D. pulex* reproduction. *Daphnids* release their neonates after shedding their exoskeleton (Poynton et al., 2007), and chitin is the major component of insect, crustacean, mollusk and fungi exoskeletons (Daraghmeh, Chowdhry, Leharne, Al Omari, & Badwan, 2011). Chitinases are enzymes that help in maintaining and reshaping the exoskeleton. It has been shown that the downregulation of chitinases correlates with chronic effects to *D. magna* reproduction after a dose dependent Zn^{2+}exposure (Poynton et al., 2007). However, in the same study chitinase activity did not change in Cd^{2+} exposed animals when compared to control animals. Conversely, chitinase-1, chitinase-2 and chitotriosidase were shown to be significantly upregulated after 48 hrs Cd^{2+} treatment in *D. pulex* (Joseph R. Shaw et al., 2007). In our dataset, we found 4/6 chitinase transcripts were downregulated and the remaining 2 upregulated in response to Cd^{2+} (1/2 was also upregulated under Zn^{2+}). These findings point towards influence of Cd^{2+} on ecdysis and molt regulation in *D. pulex*. Insect cuticle is an exoskeleton that contributes to the insect's shape and separates their living tissue from the environment by acting as a barrier (Andersen, 1979). Due to sequestration of heavy metals in the cuticle of insects such as grasshoppers and carabidae, the molting process has been linked to potential heavy metal detoxification (Lindqvist & Block, 1995). We found a total of 141 cuticle associated transcripts altered after independent Cd^{2+}, Ni^{2+} and Zn^{2+}exposure. Joseph R. Shaw et al. (2007) study showed upregulation of cuticle proteins in response to Cd^{2+}

Cd^{2+} alone with 38 upregulated and 83 downregulated genes was responsible for 85% (121/141) of genetic response in this biological process. Poynton et al. (2007) showed downregulation of a cuticle associated gene under Zn^{2+} exposure. In our analysis, 40 cuticle genes were affected by Zn^{2+} and 85% (34/40) of them were downregulated. Under the chitin binding and metabolic processes, we found 36 genes in our SDE gene list. 19/36 were upregulated and 17/36 were downregulated against Cd^{2+} exposure. In this list we found an enrichment of proteins associated with insect Peritrophic matrix (PM). The PM is the insect midgut that consists of chitin and proteins (Tellam, Wijffels, & Willadsen, 1999). Peritrophic matrix proteins (PMP) are known to protect insects against pathogens, support digestion, strength, elasticity and permeability of the PM (Jasrapuria et al., 2010; Tellam et al., 1999). Most PMPs have cysteine-rich domains that facilitates chitin binding (Tellam et al., 1999). Specific PMPs have been shown to regulate PM permeability to maintain insect fat body that is crucial for insect survival (Agrawal et al., 2014). We found a mix response pattern of both up and down regulated genes for PMPs against Cd^{2+} exposure. Our data suggests Cd^{2+} exposure is altering the regulation of PMPs and is likely affecting the PM structure and function.

CARBOHYDRATE METABOLISM

Carbohydrate metabolism was significantly downregulated in our GO enrichment analysis against Cd^{2+} exposure. We found 39 genes in total involved in this process overlapping with our SDE gene list. 84% (33/39) of these genes are downregulated in response to Cd^{2+}. Previous studies have shown suppression of digestion and feeding rates due to Cd^{2+} and Zn^{2+} exposures in *D. magna*, followed by a decrease in digestive enzyme expression (De Coen & Janssen, 1998; Guan & Wang, 2004) . Additionally, shrinking and paralysis of the digestive system has been shown because of chronic Cd^{2+} and Zn^{2+} exposure at sub-lethal dosages in *D. magna* (Griffiths, 1980). Therefore, an overall suppression of this process would likely suggest a decrease in digestive enzyme activity (Poynton et al., 2007). Salivary glands of humans and other animals to initiate the digestion of starch (Alpers, 2003) produce alpha amylase (ptyalin) and pancreatic amylase continues the process in the small intestine (M. E. Smith & Morton, 2011). 4 alpha amylase transcripts were downregulated and 1 was upregulated after Cd^{2+} exposure showing reduced metabolism of dietary carbohydrate under Cd^{2+} exposure in our dataset. Endoglucanase are cellulases that degrade cellulose into simple sugars (Yennamalli, Rader, Kenny, Wolt, & Sen, 2013). 7 genes under this family were downregulated after Cd^{2+} exposure. Other enzymes in our SDE downregulated gene list for carbohydrate metabolism include exoglucanase-1, 3-beta-glucanase, mannosidase (4 transcripts), mannanase (3 transcripts), and trehalases (2 transcripts) .This result is in agreement with Poynton et al. (2007).

OXYGEN, HEME BINDING AND TRANSPORT

Hemoglobin (Hb) is present in a wide range of animals but is narrowly represented by large invertebrate taxa including insecta and crustacea (Zheng, Xu, Qin, Wu, & Wei, 2017). Invertebrate Hb functions similarly to vertebrate Hb such that it facilitates transfer of oxygen to breathing tissues from the environment (Ha & Choi, 2008). The Hb transcripts in *Propsilocerus akamusi* showed higher expression compared to control samples after exposure to a sub-lethal dose of Cd^{2+} (2.4 mmol/L) at major Cd^{2+}

accumulation sites such as the epidermis (after 48 hrs exposure), gut (after 48, 72 hrs exposure) and malpighian tubules (after 48, 72 and 96 hrs exposure) (Zheng et al., 2017). Altered expression of Hb after Cd^{2+} exposure was postulated as a likely response to meet increasing oxygen demand during xenobiotic metabolism (Zheng et al., 2017). Joseph R. Shaw et al. (2007) have shown upregulation of 3 Hb transcripts in *D. pulex* after 48 hr exposure to sub-lethal dose of Cd^{2+} at 20 µg/L. We found a total of 11 Hb transcripts in our SDE list participating in oxygen binding and transfer. 81% (9/11) of them were all upregulated against Cd^{2+} treatment. Soetaert, van der Ven, et al. (2007) have shown downregulation of Hb transcripts in *D. magna* when exposed to Cd^{2+} at sub-lethal concentrations of 10, 50 and 100µg/L for 48 and 96 hrs. 2/11 of our Hb transcripts were downregulated with Cd^{2+} treatment. However, it is important to note that concentration and time of exposure of a toxicant in an ecotoxicological experiment can produce variable gene expression profiles for the same gene (H. J. Kim et al., 2015).

CELL CYCLE AND DNA REPAIR

The cell division process in all eukaryotic organisms is governed by cyclin dependent kinases (CDKs) and cyclin proteins (Sobkowiak & Deckert, 2004). Cyclins are proteins that regulate CDK activity during discrete phases of the cell cycle (Sobkowiak & Deckert, 2004). We found some cyclins were downregulated in our SDE gene list against Cd^{2+} exposure. Cyclin-A2 is known to regulate CDK2 during late S phase to transit the cell into G2 phase (Risal, Adhikari, & Liu, 2016). We found Cyclin-A2 was downregulated in our dataset after Cd^{2+} exposure. Cyclin-B is a key component for progression of cell cycle during G2-M transition (Risal et al., 2016). Our data showed two transcripts for mitotic-specific cyclin-b1 were downregulated under Cd^{2+} treatment. A study on Cd^{2+} impact on soybean cells showed a link between Cd^{2+} exposure and decrease in cyclin B1 mRNA resulting in cell cycle disruption during G2-M transition (Sobkowiak & Deckert, 2004). 2004). Mammalian polo-like kinases (Plk) are crucial regulators in the progression of cell cycle phases, mitosis, response to DNA damage and cytokinesis (Winkles & Alberts, 2005). Plk genes differentially regulate in cells under stress (Burns, Fei, Scata, Dicker, & El-Deiry, 2003). Plk2 mRNA expression increased in a human cell line when exposed to UV light, leading to cell cycle arrest (Burns et al., 2003), and Plk1 gene expression was suppressed in mammalian cells after exposure to DNA-damaging agents contributing to cell cycle arrest (Ando et al., 2004; Ree, Bratland, Nome, Stokke, & Fodstad, 2003) . We found Serine/threonine-protein kinase PLK1 was downregulated and Serine/threonine-protein kinase Plk2 was upregulated in our dataset against Cd^{2+} exposure, providing further evidence of cell cycle arrest after Cd^{2+} exposure. Cd^{2+} is known to cause DNA damage through oxidative stress (Badisa et al., 2007). Base excision repair (BER) enzymes remove damaged bases in DNA and suppress spontaneous mutagenesis (Nakamura et al., 2017; Sirbu & Cortez, 2013). Thymine DNA glycosylase (TDG) is shown to regulate DNA damage response in human fibroblast cells (Nakamura et al., 2017). Our data shows the upregulation of G/T mismatch-specific thymine DNA glycosylase against Cd^{2+} exposure. Another gene called Three prime repair exonuclease 2 (TREX2) was also upregulated after Cd^{2+} treatment. 3’ excision of DNA nucleotides is crucial in a variety of processes such as replication, repair and recombination of DNA (Mazur & Perrino, 2001). TREX1 and TREX2 are known mammalian genes involved in the DNA 3’ excision process (Mazur & Perrino, 2001).

METALLOTHIONEIN EXPRESSION

Metallothionein (MT) is important biomarker frequently associated with cellular protection against Cd^{2+} exposure (Bi, Lin, Millecchia, & Ma, 2006; Klaassen et al., 1999; Poynton et al., 2007; Joseph R. Shaw et al., 2007). They are highly conserved, cytosolic, non-enzymatic, low molecular weight (< 10 kDa), cysteine-rich (30-33% of the protein), and metal ion binding proteins (Amiard, Amiard-Triquet, Barka, Pellerin, & Rainbow, 2006; Y. Liu et al., 2014; Joseph R. Shaw et al., 2007). MTs participate in maintaining the homeostasis of essential metals in cells such as Cu^{2+} and Zn^{2+} (Amiard et al., 2006; Langston, Bebianno, & Burt, 1998; Viarengo & Nott, 1993).They scavenge free radicals and protect cells against oxidative stress (Ruttkay-Nedecky 2013, Sato and Bremner 1993). Additionally, MT play a crucial role in detoxification of non-essential metals heavy such as Cd^{2+}, silver (Ag^{2+})and mercury (Hg^{2+}) (Langston et al., 1998) (Amiard et al., 2006; Joseph R. Shaw et al., 2007). Joseph R. Shaw et al. (2007) and Asselman et al. (2013) have reported upregulation of MT in *D. pulex* when exposed to Cd^{2+} at 48 hrs for 20μg Cd^{2+}/L concentration and after 96 hrs for 0.5μg Cd^{2+}/L concentration respectively. Poynton et al. (2007) also observed significant MT upregulation in *D. magna* (same species as *D. pulex* with high amino acid sequence similarity to *D. pulex* MTs, Asselman et al. (2013)), when treated with Cd^{2+} at 18μg Cd^{2+}/L concentration for 24 hours. However, Soetaert, Vandenbrouck, et al. (2007) did not report any change in MT when *D. magna* was exposed to (10, 50, 100μg Cd^{2+}/L) for 48 and 96 hours. Asselman et al.

(2013) have described MT expression as time and homolog dependent. Soetaert, Vandenbrouck, et al. (2007) observed high correlation between gene expression changes and gradual increase in metal dose and exposure time. This implicates that MT expression is dependent on the organism, time of exposure and dosage of the heavy metal. Metal transcription factor I (MTF-1) is a crucial regulator of MT gene expression (Grzywacz et al., 2015). Nuclear Factor I (NFI) is a direct interactor of MTF-1 (Gunther et al., 2012). NFI interaction with MTF-1 has been indicated to either enhance or inhibit MT-I basal and heavy metal such Cd^{2+} and Zn^{2+} induced expression (Jacob, Majumder, & Ghoshal, 2002; LaRochelle et al., 2008; Majumder, Ghoshal, Gronostajski, & Jacob, 2001). Majumder et al. (2001) have showed that direct interaction between NFI and MT-I promoter is not required for NFI proteins to repress MT-I expression. NFI protein has four isoforms such as NFI-A,-B,-C and -X. Overexpression of all isoforms in HepG2 cells showed repression of MT-I. LaRochelle et al. (2008) described the discrepancy between their results and Majumder et al. (2001) was due to the difference in overexpression of NFI. However, we found NFI-C significantly upregulated against Cd^{2+} exposure in *D. pulex* without any tampering with the expression of NFI isoforms. Jacob et al. (2002) study has shown an inverse correlation between MT-I and GST in certain prostate cancer cell lines. Additionally, they also found expression of GCL correlating with GST in lymphosarcoma and hepatoma cells indicating a mechanism compensating for decrease in MT-I expression (Jacob et al., 2002). We found MT significantly downregulated and GST and GCL significantly upregulated in our dataset against Cd^{2+} exposure.

IRON TRANSPORT AND HOMEOSTASIS

Ferritins are iron (Fe^{2+}) storage and scavenging proteins (Poynton et al., 2007). Their regulation depends primarily on bioavailable Fe, iron response protein (IRP) and iron response element complex (IRE) (Arosio & Levi, 2002; Poynton et al., 2007). In *Xenopus laevis* cells, Ferritin expression increased after Cd^{2+} exposure (J. P. Muller, Vedel, Monnot, Touzet, & Wegnez, 1991). Poynton et al. (2007) showed upregulation of Ferritin gene in *D. magna* after independent Cu^{2+} (6 ug/L) and Cd^{2+} (18 ug/L) treatment for 24 hrs. This microarray analysis also showed two transcripts of Ferritin were downregulated when exposed to Cd^{2+} only (Poynton et al., 2007). Joseph R. Shaw et al. (2007) mentioned differential regulation for one Ferritin gene in *D. pulex* after Cd^{2+} treatment at 20ug/L for 48 hrs.

Activating transcription factor 1 (ATF1) is known to repress ferritin H gene transcription (Hailemariam, Iwasaki, Huang, Sakamoto, & Tsuji, 2010; Iwasaki, Hailemariam, & Tsuji, 2007). However, homeodomain-interacting protein kinase 2 (HIPK2) has been shown to counter the effect of ATF1 on ferritin H transcriptional repression (Hailemariam et al., 2010). Even though, HIPK2 is significantly upregulated under Cd^{2+} exposure, we found seven ferritin heavy (H) chain transcripts significantly downregulated in our dataset. We also found all seven ferritin H chain transcripts grouped under cluster 1 based on our TF binding site profile clustering. Cluster 1 is significantly enriched with ATF transcription factor gene targets indicating a potential role of ATF TFs in regulating ferritin H gene expression. Further research in this direction is needed to understand the underlying mechanism behind the downregulation of ferritins under 24 hr Cd^{2+} exposure in *D. pulex*.

CONCLUSION

In conclusion, we have successfully demonstrated validation of *D. pulex* as a model organism to study ecotoxicogenomics. RNA-Seq has enhanced the power of ecotoxicological science by allowing us to detect changes at the biomolecular level. This is the biggest gene expression dataset generated for heavy metal induced toxicity in *D. pulex*. This study has bioinformatically predicted oxidation-reduction as a significantly enriched biological process against Cd^{2+} induced toxicity that has not previously been shown at a comprehensive level in *D. pulex*.

Our TF factor binding site analysis pipeline is a fresh attempt at solving an existing problem of biologically relevant motif discovery. Using the binding site detection pipeline along with consensus clustering has allowed us to detect combinations of TFs potentially driving parts of the genetic response against heavy metal induced stress. *Ahr::Arnt* and *NRF1* bind to xenobiotic (XRE) and antioxidant (ARE) response elements in the upstream regions of genes involved in environmental toxicant response or xenobiotic metabolism pathway (Beischlag, Luis Morales, Hollingshead, & Perdew, 2008; Biswas & Chan, 2010; Kohle & Bock, 2009). The significant upregulation of *Ahr* and *NRF1* against Cd^{2+} exposure in our dataset and exhaustive literature search aided us to discover xenobiotic metabolism pathway that was not detected using conventional GO enrichment analysis. Additionally, foxo transcription factors have been linked to regulate cellular stress resistance, apoptosis, and metabolism (Martins, Lithgow, & Link, 2016).

We found highest number of SDE gene targets for foxo TFs compared to other stress associated TFs in our list (Table 4.7), suggesting a significant role of this TF in regulating the stress response of this organism to heavy metal exposure. GO and orphan pairs sharing FUCDs that did not cluster together mainly had no known stress motif in similar. This may suggest a possibility of common motifs between them that are associated with other biological processes or novel motifs associated with stress. Antioxidant defense response genes showed a variety of stress TF binding site profiles suggesting modularity in their regulatory patterns. However, a set of xenobiotic and oxidative stress response genes grouped under the same cluster implicates their role as a unit responding under the heavy metal and especially Cd^{2+} stress in *D. pulex*. Further research needs to be conducted to collect more evidence on the molecular mechanisms of *D. pulex* genomic response to heavy metal contaminants.

The analysis in this chapter was entirely performed by me, therefore, it does not contain an epilogue section.

ACKNOWLEDGEMENTS

The RNA-Sequencing of *D. pulex* sample was supported by funding from Dr. J Robert Manak lab. I am thankful to Dr. J Robert Manak and Dr. Josep Comeron for their significant contribution in the bioinformatics analysis of this study.

	Genes 2010 beta 3	JGI_2011_Frozen_Cat
Total genes	47712	23210
Common genes	22630	22724
Total mapped reads	152834/200000 (76.4%)	136272/200000 (68.1%)
Uniquely mapped reads	140547/200000 (70.2%)	127782/200000 (63.8%)

Table 4.1. Transcript version comparison statistics.

Samples	Mapped reads	Unique reads	Unique percentage
Control 1	32754569	26781719	0.674420451
Control 2	28635831	23286565	0.686202291
Control 3	23813276	20883728	0.695967303
Control 4	19956021	16313391	0.688898427
Cadmium 1	40385281	34897457	0.685324907
Cadmium 2	29029006	24494492	0.674518413
Cadmium 3	37326026	31047093	0.675283915
Nickel 1	42109420	35638244	0.673396224
Nickel 2	23221157	19394164	0.689756958
Nickel 3	35936381	29378732	0.688068568
Zinc 1	43267688	37650949	0.68621579
Zinc 2	35827270	29093188	0.6856887
Zinc 3	35631505	29480893	0.691686024

Table 4.2. Read mapping statistics.

Scaffold number	Median length of Intergenic regions
1	704 bp
3	871 bp
2	741 bp
4	781 bp
5	690 bp
6	761 bp
8	656 bp
7	1031 bp
9	891 bp
12	793 bp
Average	~ 800 bp

Table 4.3. Upstream region statistics.

	Genes with domain	GO genes	Orphans	SDE	SDE GO	SDE Orphan	SDE without domains
Y_2016/2017	15692	12001	3691	2292	1323	317	652

Table 4.4. InterProScan statistics.

Shared FUCD genes	GO with FUCD	Orphans with FUCD
Transcriptome	5719	2523
SDE	452	200

Table 4.5. Comparison between orphans and GO genes carrying the same FUCD

	Genes with pathway terms	SDE	SDE with pathway terms
Y_2017	4516	2292	410

Table 4.6. Pathway term statistics.

Transcription factor	Total number of gene targets	Up regulated gene targets	Down regulated gene targets	Mixed
Ahr::Arnt	649	361	291	3
Arntl	195	78	118	1
Atf1/3/ATF7	545	302	246	3
CEBPA/B	123	65	60	2
Foxo1/FOXO3/6	733	385	354	6
NF-κB (RELA)	197	97	102	2
MTF1	413	224	191	2
NRF1	100	64	37	1
Pparg::Rxra	110	59	52	1
Ddit3::Cebpa	14	10	4	0
ATF4	60	31	29	0
PPARG	3	1	2	0

Table 4.7. Stress associated transcription factor binding site on oxidative stress genes.
Note: discrepancy in the sum of up and down regulated genes is due to genes with mixed (up and down) regulation patterns across metals.

Oxidative stress genes	TF profile	Regulation
GSTA1	*ATF4;Ahr::Arnt*	Cd_Up
SODC	*RELA;Atf1/3; Foxo1/FOXO3/6*	Cd_Up
AKR1C4	*MTF1;Atf1; FOXO6*	Cd_Up
CBR3	*MTF1;NRF1*	Cd_Up
PGH2	*Atf1/3/ATF7;Ahr::Arnt;ATF4;NRF1*	Cd_Up
GSH	*CEBPA/B;MTF1; FOXO3/6;Ahr::Arnt*	Cd_Up
POR	*RELA;Ahr::Arnt;*	Cd_Up

Table 4.8. Stress associated genes TF binding profile.

TF	Involved in biological processes	Publication
Ahr::Arnt	Response to environmental contaminants	Beischlag et al. (2008); Dietrich (2016)
Arnt	Oxidative stress	Wells, Gu, and dela Paz (2009)
Arntl	Circadian rhythm and regulation of antioxidant response	J. Lee et al. (2013); Wilking, Ndiaye, Mukhtar, and Ahmad (2013)
NRF1	Antioxidant and xenobiotic response	Ohtsuji et al. (2008)
MTF1	Heavy metal exposure and oxidative stress	Gunther et al. (2012)
Atf1	Oxidative stress	Hailemariam et al. (2010)
Atf3	Oxidative stress	Okamoto, Iwamoto, and Maru (2006)
Atf4	Oxidative stress	Lange et al. (2008)
ATF7	Stress response	Maekawa et al. (2018)
NF-κB (REL, RELA)	ROS regulation, cell survival, inflammation, and immunity	Morgan and Liu (2011)
HIF1A	Hypoxia, ROS and oxidative stress	Pialoux et al. (2009)
Ddit3	Cellular stress response	Jauhiainen et al. (2012)
CEBP	Regulation of Oxidative stress response	Hsiao et al. (2014); Huggins et al. (2015); Manea, Todirita, Raicu, and Manea (2014)
PPARG	Energy homeostasis, inflammatory response, xenobiotic metabolism and oxidative stress	Omiecinski, Vanden Heuvel, Perdew, and Peters (2011); Polvani, Tarocchi, and Galli (2012)
FoxO	Cellular stress response and antioxidant defense	Klotz et al. (2015); Martins et al. (2016)

Table 4.9. Known stress associated TFs.

Sample	Upregulated genes	Downregulated genes	Total genes
Cd	1185	905	2090
Ni	186	186	372
Zn	109	288	397

Table 4.10. Significantly differentially expressed (SDE) gene statistics.

Metal	Number of GO terms affected
Cd	38
Ni	4
Zn	1

Table 4.11. SE GO terms statistics.

Biological processes	Supporting literature
Oxidation-reduction process, Response to oxidative stress, and Peroxidase activity	Poynton et al. (2007); Joseph R. Shaw et al. (2007)
Oxygen, heme binding and transport, and Hemoglobin complex	Joseph R. Shaw et al. (2007), Soetaert, van der Ven, et al. (2007)
Carbohydrate metabolism	Soetaert, van der Ven, et al. (2007), Poynton et al. (2007)
Iron (Fe) transport and homeostasis	Poynton et al. (2007); Joseph R. Shaw et al. (2007)
Cell cycle and DNA repair	Garrett et al. (2011)
Reproduction, chitin and exoskeleton (cuticle)	Poynton et al. (2007); Soetaert, van der Ven, et al. (2007)
Calcium ion (Ca2+) and cell death	Choong et al. (2014)
Xenobiotic metabolism pathway	Novel as a whole pathway
Steroid biosynthetic process	Bochud et al. (2018); Paksy, Varga, and Lazar (1992); Voogt, den Besten, Kusters, and Messing (1987)

Table 4.12. Biological processes affected by heavy metals, especially Cd^{2+}

Cluster number	Total number of gene targets	Up	Down	GO	Orphan	NO domain genes	GO+FUCD	Orphan+FUCD
1	550	307	246	328	69	153	42	27
2	396	219	180	234	51	111	39	21
3	279	154	129	150	46	83	19	11
4	247	126	122	131	31	85	7	10
5	150	68	84	82	29	39	8	7

Table 4.13. Motif based cluster statistics.
Note: discrepancy in the sum of up and down regulated genes is due to genes with mixed (up and down) regulation patterns across metals.

Similarity group	TF	Width	TF family
1	*CEBPB*	10	C/EBP-related
1	*CEBPA*	11	C/EBP-related
1	*CEBPG*	10	C/EBP-related
1	*CEBPE*	10	C/EBP-related
1	*CEBPD*	10	C/EBP-related
2	*FOXO3*	8	Forkhead box (FOX) factors
2	*Foxo1*	11	Forkhead box (FOX) factors
2	*FOXO6*	7	Forkhead box (FOX) factors
2	*FOXO4*	7	Forkhead box (FOX) factors
3	*ATF7*	14	Jun-related factors
3	*Atf1*	8	CREB-related factors
3	*Atf3*	8	Fos-related factors
4	*RELA*	10	NF-kappaB-related factors
4	*REL*	10	NF-kappaB-related factors
5	*Arntl*	10	PAS domain factors
5	*Arnt*	6	PAS domain factors
6	*ARNT::HIF1A*	8	PAS domain factors::PAS domain factors
7	*Pparg::Rxra*	15	Thyroid hormone receptor-related factors (NR1)::RXR-related receptors (NR2)
8	*PPARG*	20	Thyroid hormone receptor-related factors (NR1)
9	*Ahr::Arnt*	6	PAS domain factors::PAS domain factors
10	*MTF1*	14	More than 3 adjacent zinc finger factors
11	*Ddit3::Cebpa*	12	C/EBP-related::C/EBP-related
12	*ATF4*	13	ATF-4-related factors
13	*NRF1*	11	Jun-related factors

Table 4.14. Similarity based grouping of known stress associated TF using TOMTOM (e-value <0.01).
Note: The highlighted TF consensus of variable lengths and offsets within the same "Similarity group" were used for further analysis.

CLUSTER	TF	# of SDE TF targets	# of SDE TF targets in the cluster	# of SDE targets for this TF	# of SDE genes in the cluster	p-value	FDR adjusted p-value
CLUSTER_1	*Atf1/3_ATF7*	545	550	545	1622	0	0
CLUSTER_1	*ATF4*	34	550	60	1622	0.000191873	0.001151236
CLUSTER_1	*NRF1*	34	550	100	1622	0.5314166	1
CLUSTER_1	*Ddit3::Cebpa*	4	550	14	1622	0.7544307	1
CLUSTER_1	*Pparg::Rxra*	30	550	110	1622	0.9501108	1
CLUSTER_1	*PPARG*	1	550	3	1622	0.7115841	1
CLUSTER_1	*Ahr::Arnt*	210	550	649	1622	0.8711236	1
CLUSTER_1	*CEBPA/B*	36	550	123	1622	0.8916267	1
CLUSTER_1	*Foxo1/FOXO3/6*	193	550	733	1622	1	1
CLUSTER_1	*MTF1*	109	550	413	1622	0.9999409	1
CLUSTER_1	*RELA*	56	550	197	1622	0.966493	1
CLUSTER_1	*Arntl*	51	550	195	1622	0.9947937	1
CLUSTER_2	*Ahr::Arnt*	396	396	649	1622	1.22E-203	1.22E-202
CLUSTER_2	*NRF1*	28	396	100	1622	0.2267698	1
CLUSTER_2	*ATF4*	7	396	60	1622	0.9963103	1
CLUSTER_2	*Pparg::Rxra*	26	396	110	1622	0.6163758	1
CLUSTER_2	*PPARG*	1	396	3	1622	0.5684222	1
CLUSTER_2	*CEBPA/B*	20	396	123	1622	0.9914228	1
CLUSTER_2	*Foxo1/FOXO3/6*	134	396	733	1622	0.9999999	1
CLUSTER_2	*MTF1*	62	396	413	1622	1	1
CLUSTER_2	*RELA*	40	396	197	1622	0.9379738	1
CLUSTER_2	*Arntl*	38	396	195	1622	0.9660599	1

Table 4.15. Cluster-wise transcription factor enrichment analysis.
Note: Highlighted TFs are statistically significantly enriched in each cluster after hypergeometric test (p-value) and false discovery rate (FDR) multiple test correction (adjusted p-value). # = Number.

CLUSTER	**TF**	**# of SDE TF targets**	**# of SDE TF targets in the cluster**	**# of SDE targets for this TF**	**# of SDE genes in the cluster**	**p-value**	**FDR adjusted p-value**
CLUSTER_3	*Foxo1/FOXO3/6*	279	279	733	1622	1.53E-112	1.38E-111
CLUSTER_3	*Pparg::Rxra*	19	279	110	1622	0.5340316	0.96125688
CLUSTER_3	*PPARG*	1	279	3	1622	0.4325751	0.96125688
CLUSTER_3	*CEBPA/B*	23	279	123	1622	0.3616633	0.96125688
CLUSTER_3	*Arntl*	36	279	195	1622	0.340853	0.96125688
CLUSTER_3	*NRF1*	10	279	100	1622	0.9869584	0.991337
CLUSTER_3	*ATF4*	7	279	60	1622	0.9142887	0.991337
CLUSTER_3	*Ddit3::Cebpa*	1	279	14	1622	0.9296508	0.991337
CLUSTER_3	*RELA*	23	279	197	1622	0.991337	0.991337
CLUSTER_4	*MTF1*	242	247	413	1622	4.09E-166	4.09E-165
CLUSTER_4	*Ddit3::Cebpa*	9	247	14	1622	3.82E-05	0.000191056
CLUSTER_4	*Foxo1/FOXO3/6*	127	247	733	1622	0.01957799	0.065259967
CLUSTER_4	*NRF1*	7	247	100	1622	0.9968172	1
CLUSTER_4	*ATF4*	9	247	60	1622	0.5771939	1
CLUSTER_4	*Pparg::Rxra*	18	247	110	1622	0.4082018	1
CLUSTER_4	*Ahr::Arnt*	43	247	649	1622	1	1
CLUSTER_4	*CEBPA/B*	12	247	123	1622	0.9756154	1
CLUSTER_4	*RELA*	24	247	197	1622	0.9183718	1
CLUSTER_4	*Arntl*	21	247	195	1622	0.9781572	1
CLUSTER_5	*RELA*	54	150	197	1622	6.70E-16	4.02E-15
CLUSTER_5	*Arntl*	49	150	195	1622	1.13E-12	3.39E-12
CLUSTER_5	*CEBPA/B*	32	150	123	1622	1.01E-08	2.02E-08
CLUSTER_5	*NRF1*	21	150	100	1622	0.000168614	0.000252921
CLUSTER_5	*Pparg::Rxra*	17	150	110	1622	0.02044984	0.024539808
CLUSTER_5	*ATF4*	3	150	60	1622	0.9282335	0.9282335

Table 4.15. —continued

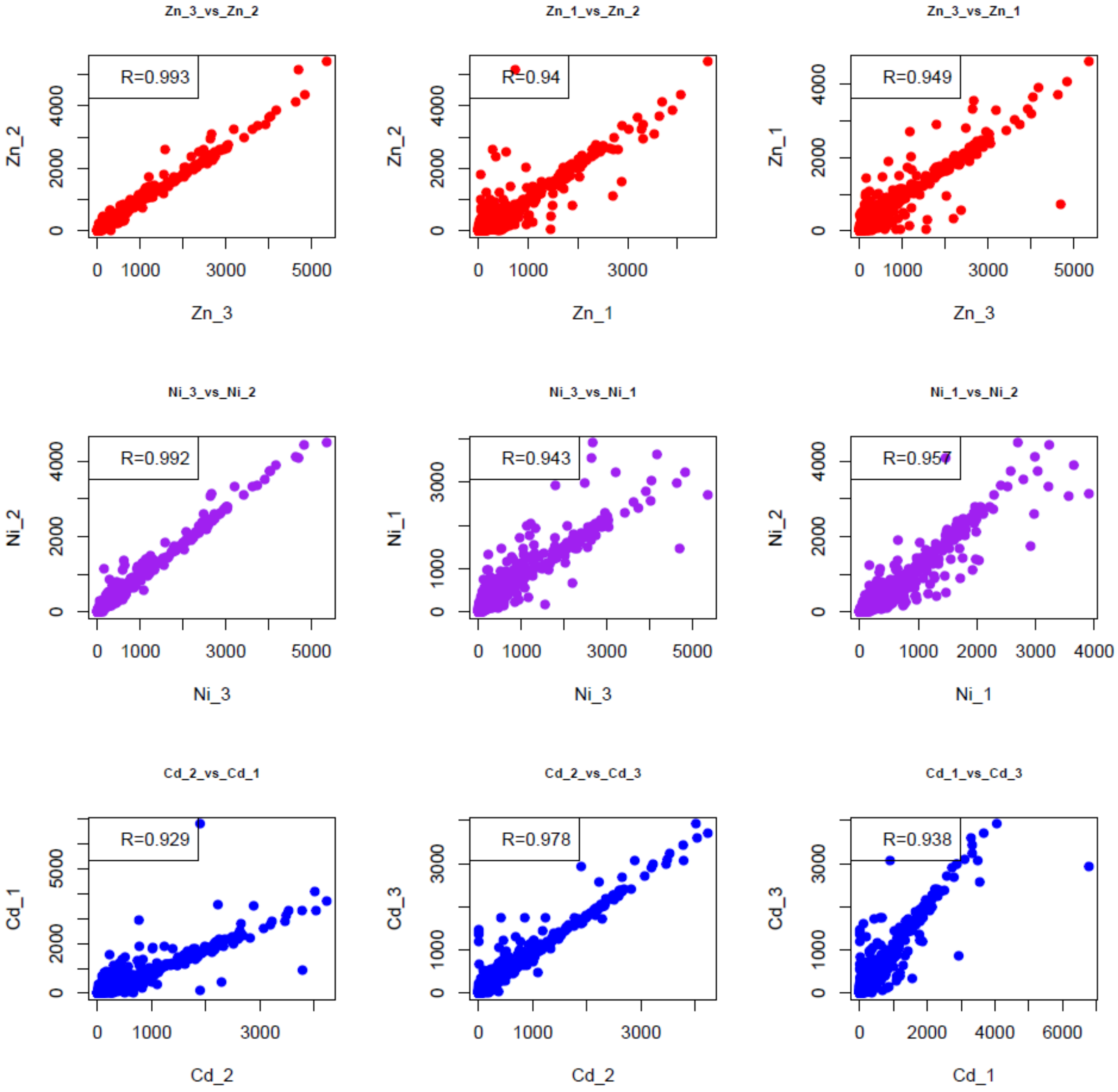


Figure 4.1. BR correlation scatter plots comparing a pair of samples under each treatment and their coefficient of Pearson correlation (R).

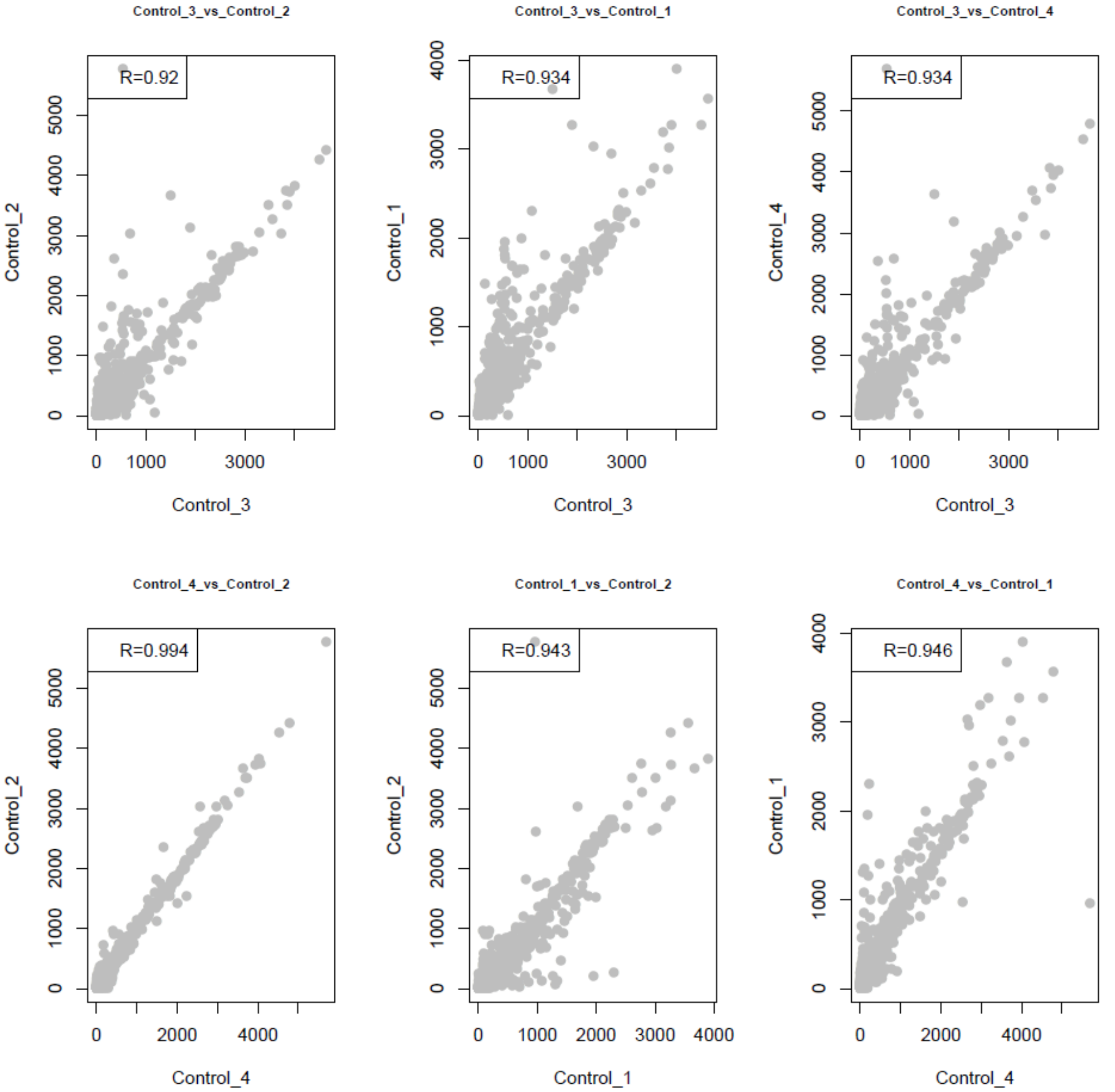


Figure 4.1. —continued

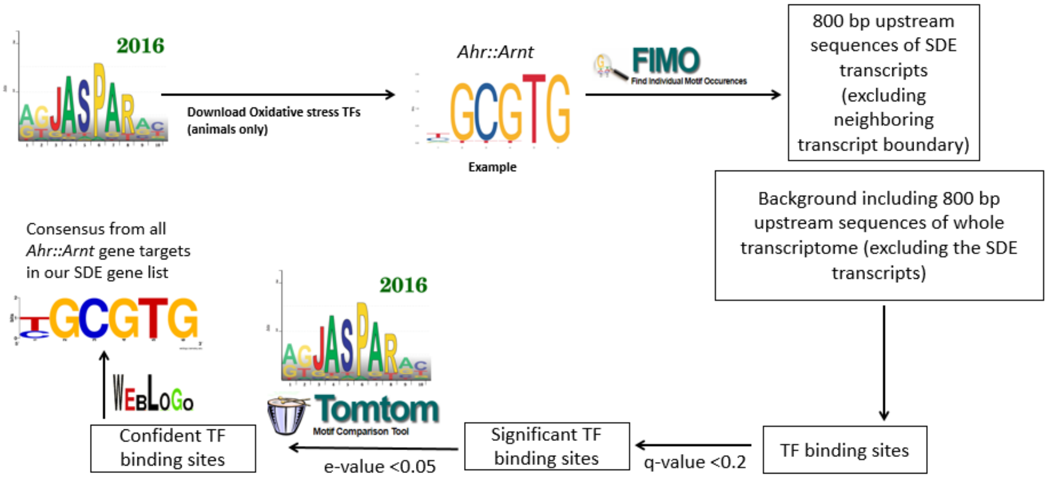

2016
JASPAR
Download Oxidative stress TFs (animals only)
Ahr::Arnt
GCGTG
Example
FIMO
Find Individual Motif Occurences
800 bp upstream sequences of SDE transcripts (excluding neighboring transcript boundary)
Background including 800 bp upstream sequences of whole transcriptome (excluding the SDE transcripts)
TF binding sites
q-value <0.2
Significant TF binding sites
e-value <0.05
Tomtom
Motif Comparison Tool
2016
JASPAR
Confident TF binding sites
WEBLOGO
Consensus from all Ahr::Arnt gene targets in our SDE gene list
GCGTG

Figure 4.2. Regulatory motif analysis pipeline.

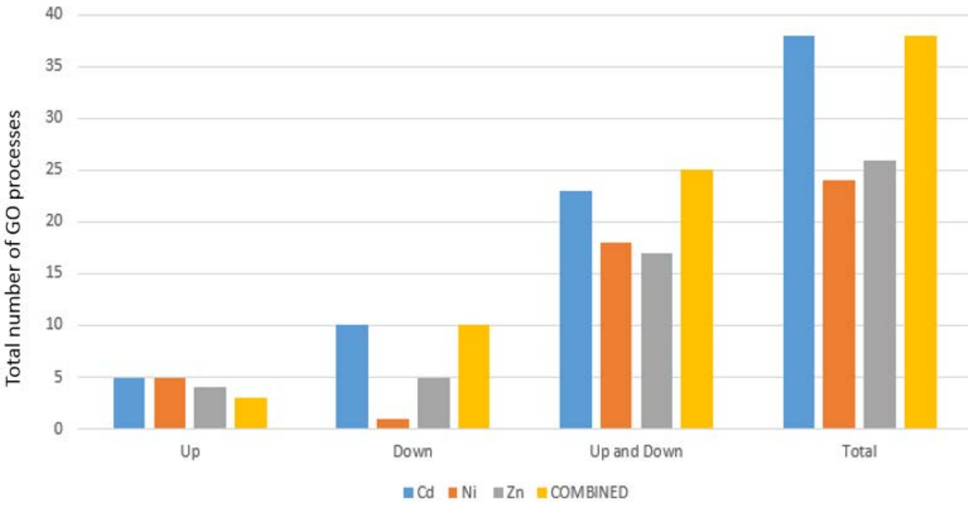


Figure 4.3. Enriched GO terms and metal induced gene regulation comparison.

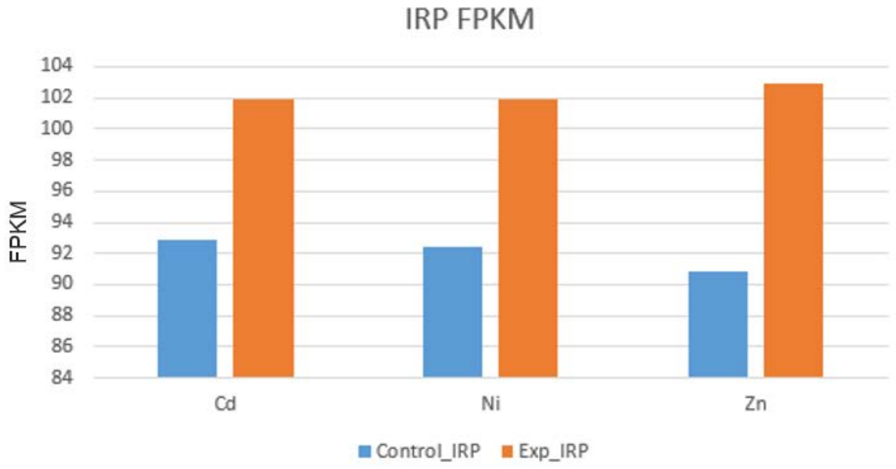


Figure 4.4. Expression of IRP across metals.

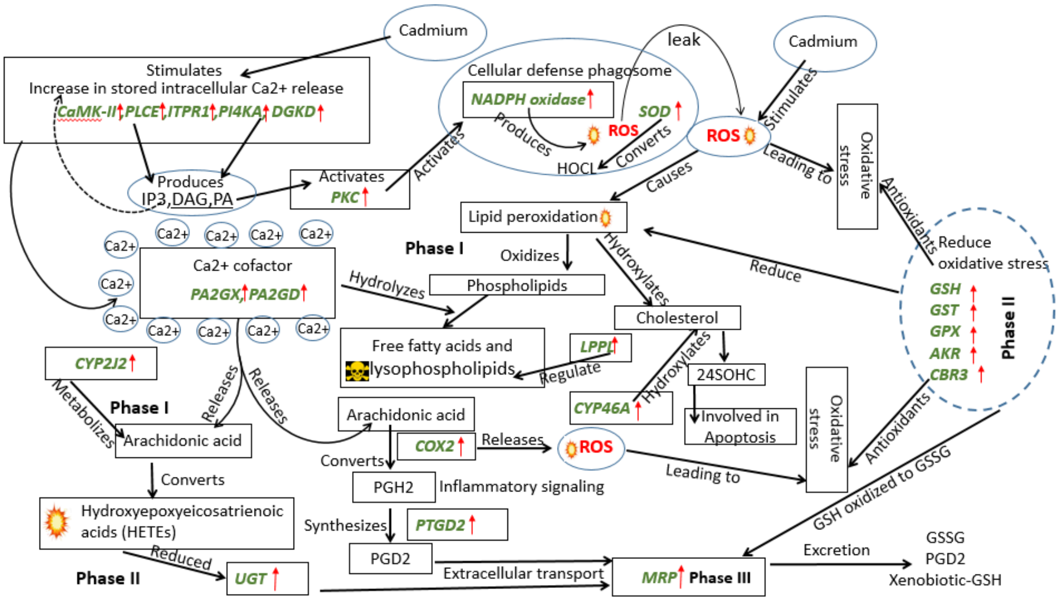


Figure 4.5. Oxidative stress and xenobiotic metabolism.
Note: Green color and italics are gene symbols and red arrow means SDE up regulation. Black color is for a substrate, or a reaction, or a biological process. ✹ stands for reactive intermediate and ☠ stands for toxic metabolites.

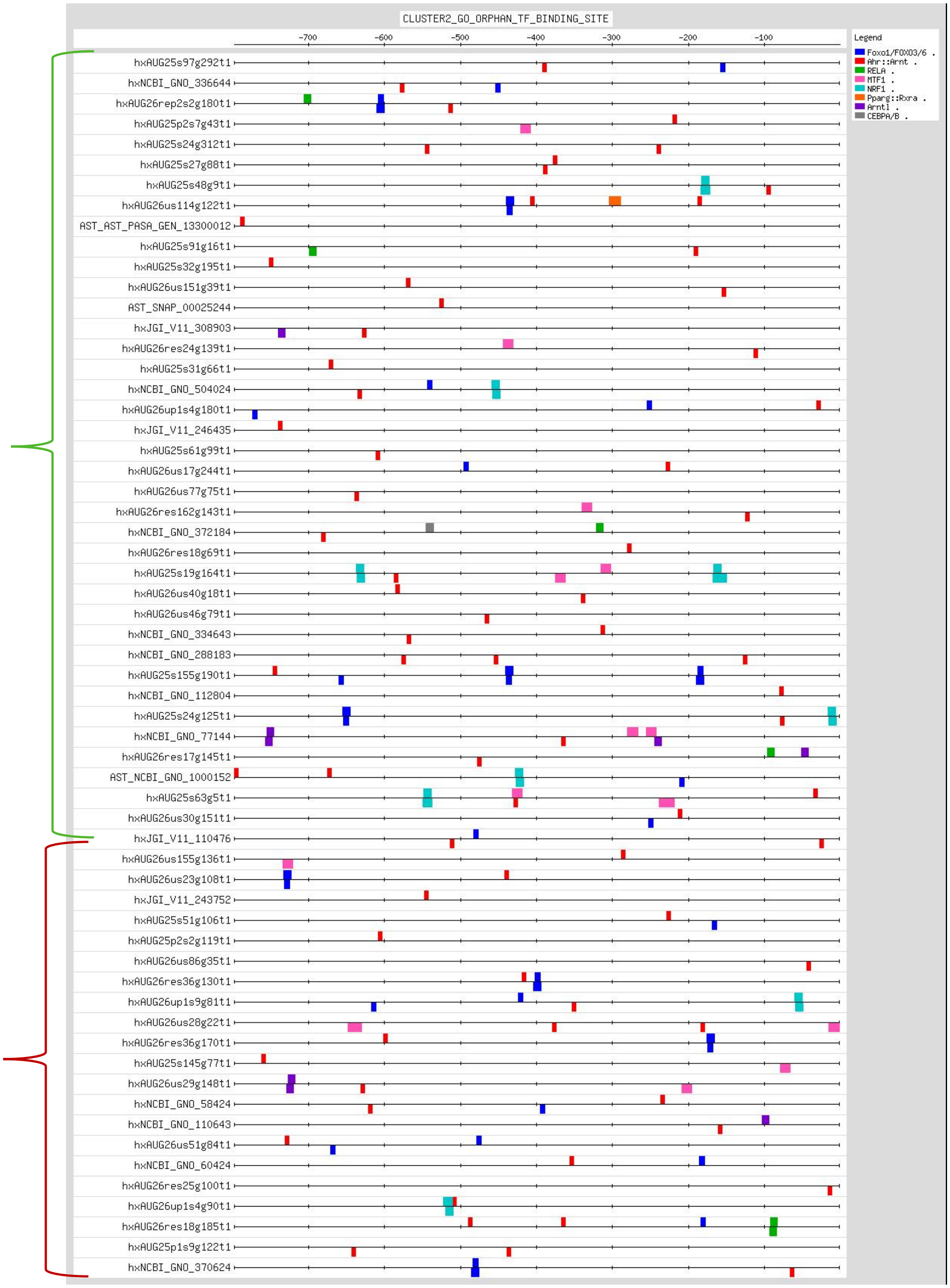


Figure 4.6 GO and Orphan genes sharing FUCDs and *Ahr::Arnt* binding sites, which are clustered together (cluster 2).

Note: GO genes = green color, orphan genes = maroon color and *Ahr::Arnt* binding sites= orange rectangles.

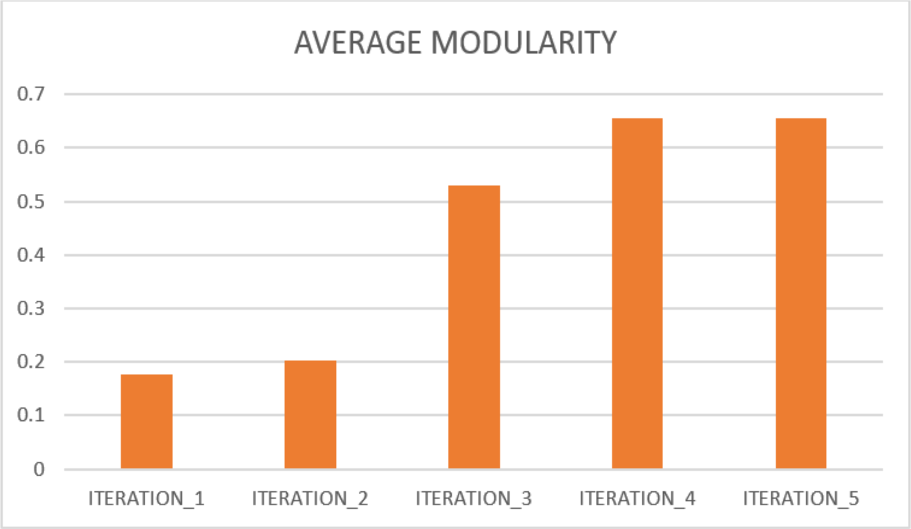


Figure 4.7 Iterative average modularity.

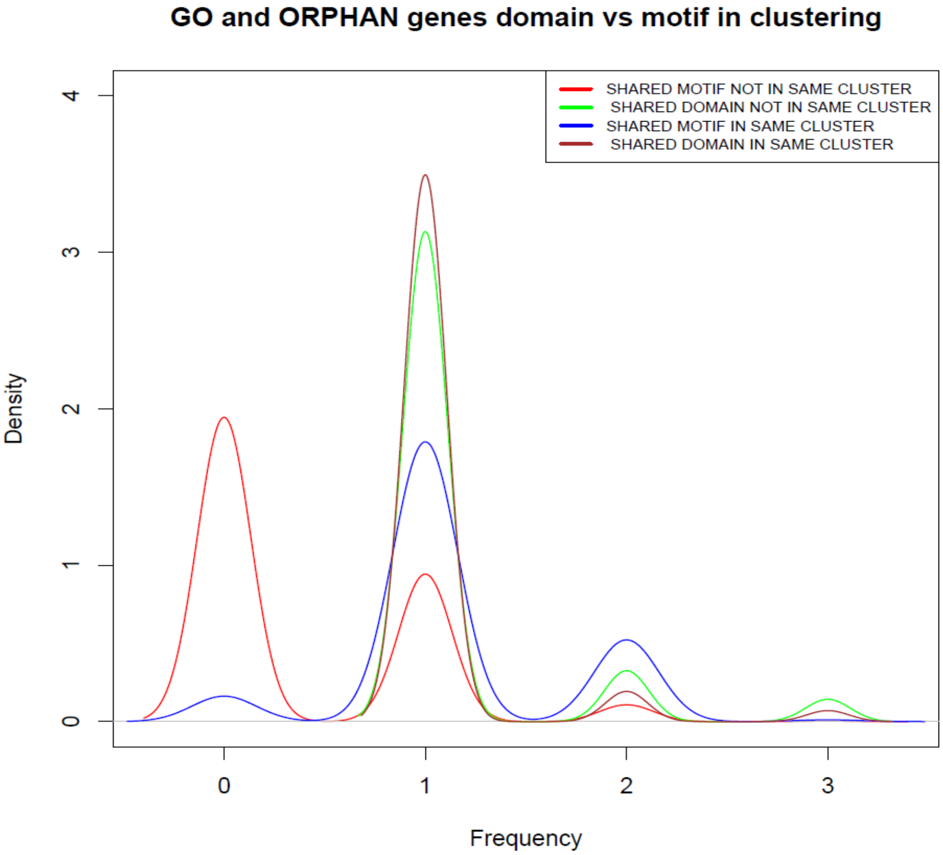


Figure 4.8 Distribution of GO and orphan motif similarity between clustered and non-clustered GO and orphan pairs sharing a FUCD domain.

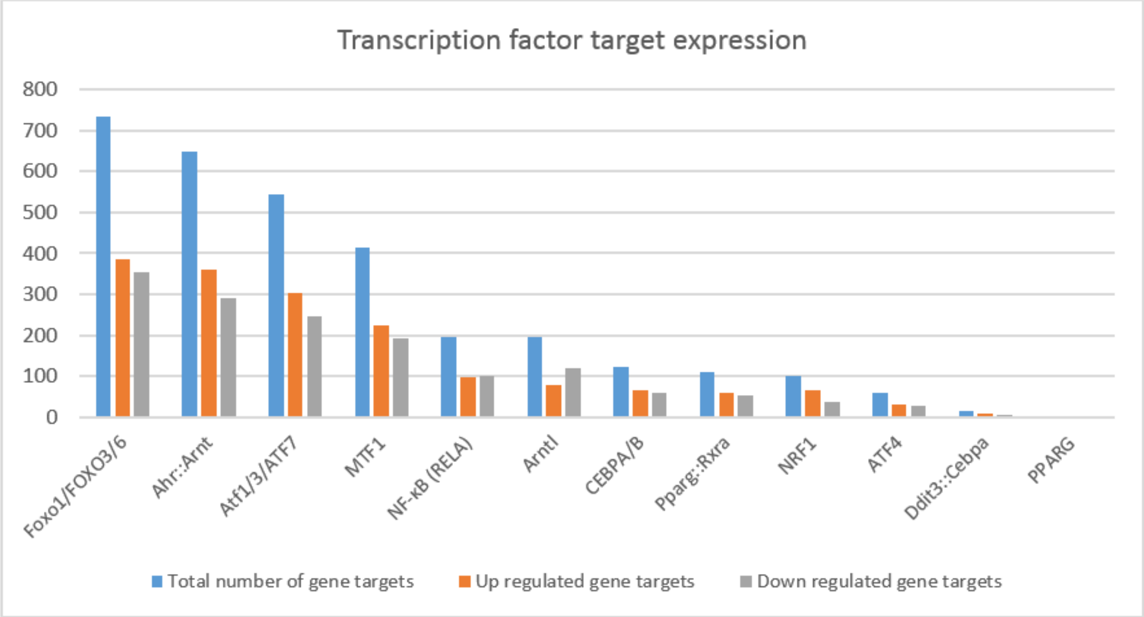


Figure 4.9 Expression pattern of TF target genes

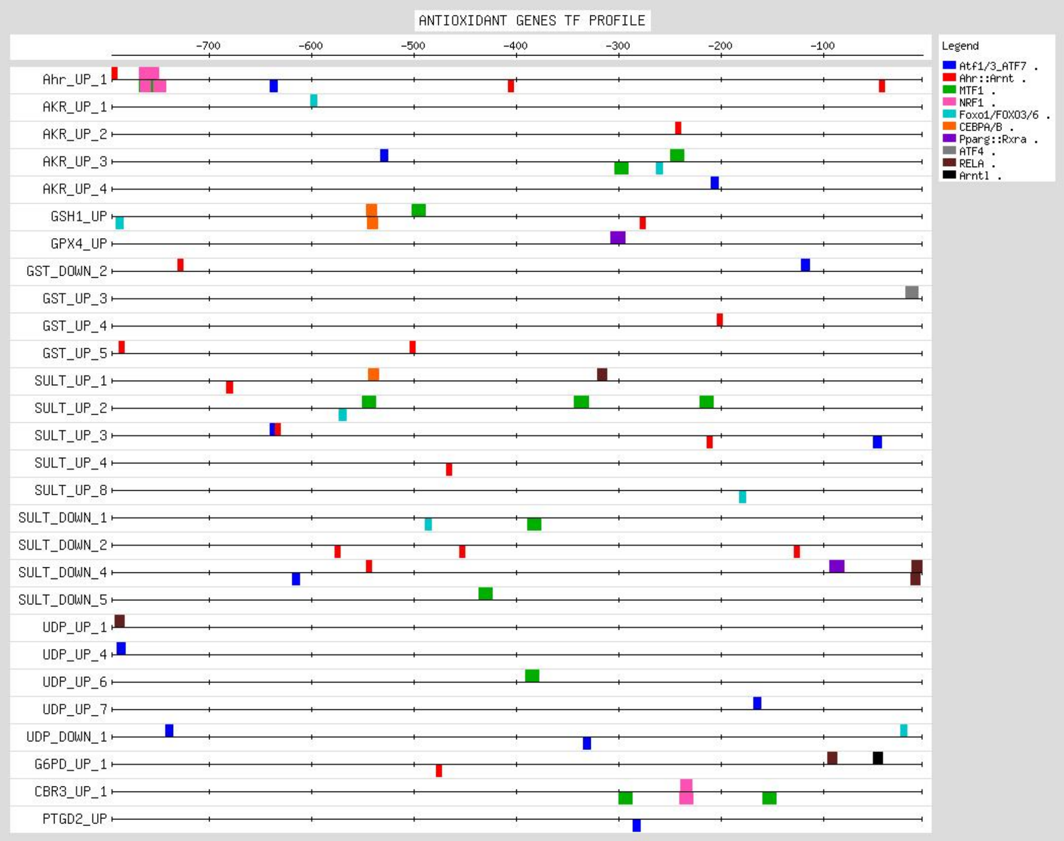


Figure 4.10 Antioxidant defense response genes TFBS profile.

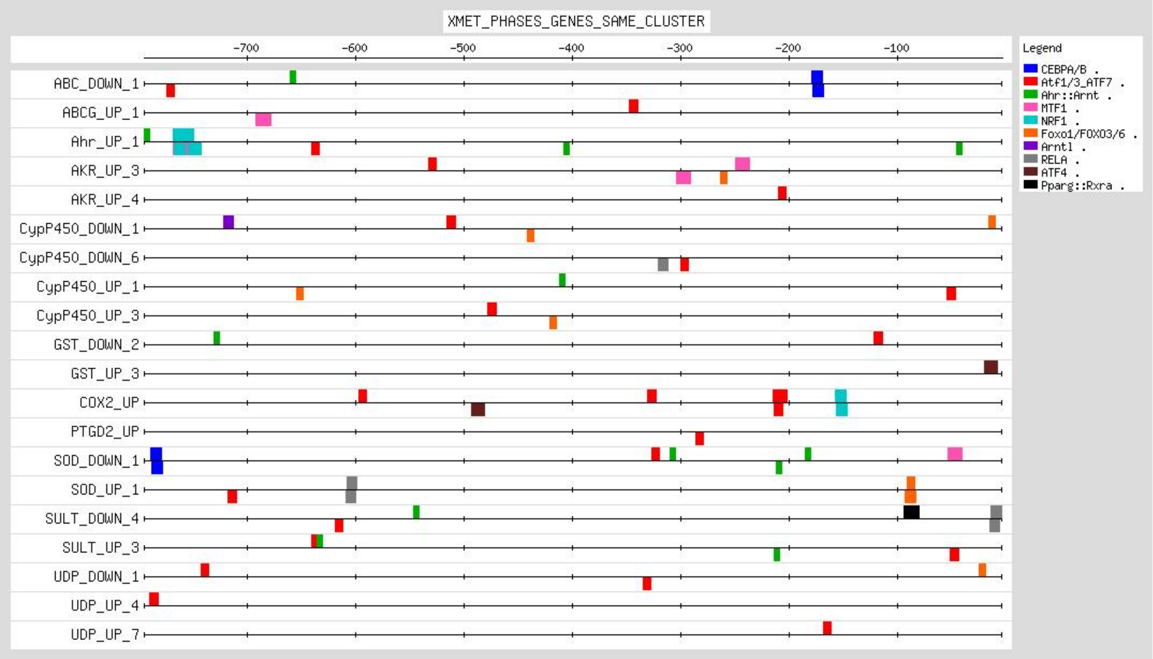


Figure 4.11 TFBS profile of Xenobiotic metabolism genes in the same cluster.

CHAPTER 5. TRANSCRIPTIONAL PROFILING OF PREDATOR-INDUCED PHENOTYPIC PLASTICITY IN *DAPHNIA PULEX*.[5]

ABSTRACT

BACKGROUND

Predator-induced defenses are a prominent example of phenotypic plasticity found from single-celled organisms to vertebrates. The water flea *Daphnia pulex* is a very convenient ecological genomic model for studying predator-induced defenses as it exhibits substantial morphological changes under predation risk. Most importantly, however, genetically identical clones can be transcriptionally profiled under both control and predation risk conditions and be compared due to the availability of the sequenced reference genome. Earlier gene expression analyses of candidate genes as well as a tiled genomic microarray expression experiment have provided insights into some genes involved in predator-induced phenotypic plasticity. Here we performed the first RNA-Seq analysis to identify genes that were differentially expressed in defended vs. undefended *D. pulex* specimens in order to explore the genetic mechanisms underlying predator-induced defenses at a qualitatively novel level.

RESULTS

We report 230 differentially expressed genes (158 up- and 72 down-regulated) identified in at least two of three different assembly approaches. Several of the

[5] This chapter has been published in Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x

differentially regulated genes belong to families of paralogous genes. The most prominent classes amongst the up-regulated genes include cuticle genes, zinc-metalloproteinases and vitellogenin genes. Furthermore, several genes from this group code for proteins recruited in chromatin-reorganization or regulation of the cell cycle (cyclins). Down-regulated gene classes include C-type lectins, proteins involved in lipogenesis, and other families, some of which encode proteins with no known molecular function.

CONCLUSIONS

The RNA-Seq transcriptome data presented in this study provide important insights into gene regulatory patterns underlying predator-induced defenses. In particular, we characterized different effector genes and gene families found to be regulated in *Daphnia* in response to the presence of an invertebrate predator. These effector genes are mostly in agreement with expectations based on observed phenotypic changes including morphological alterations, i.e., expression of proteins involved in formation of protective structures and in cuticle strengthening, as well as proteins required for resource re-allocation. Our findings identify key genetic pathways associated with anti-predator defenses.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (doi: 10.1186/s12983-015-0109-x) contains supplementary material, which is available to authorized users.

BACKGROUND

The common freshwater micro-crustacean *Daphnia* has become a model organism for many biological disciplines (Altshuler et al., 2011; J. K. Colbourne et al., 2011; Lampert, 2011; Miner, de Meester, Pfrender, Lampert, & Hairston Jr, 2012; Schaack, 2008; R. Tollrian & Leese, 2010). The extensive knowledge of its ecology (Lampert, 2006, 2011; Miner et al., 2012) and its biological responses to environmental changes (Altshuler et al., 2011; S. I. Dodson, 1989; R. Tollrian & Dodson, 1999) together with the availability of genomic resources (J. K. Colbourne et al., 2011) make the system highly attractive for evolutionary ecology research and provides the unique opportunity to study ecological traits with the aid of emerging molecular biological tools. One of the most intriguing ecological responses of *Daphnia* species to environmental changes is their ability to develop different phenotypes given the same genetic background, a phenomenon called phenotypic plasticity. Prominent examples of phenotypic plasticity include inducible defenses.

Inducible defenses are interpreted as adaptations to heterogeneous predation risks and are found in many organisms from protists to vertebrates (Brönmark, Pettersson, & Nilsson, 1999; Kuhlmann, Kusch, & Heckmann, 1999; Ralph Tollrian & Harvell, 1999). *Daphnia* evolved sensitivity against specific chemical compounds, which are unintentionally emitted by their predators. These so-called kairomones serve as signals which prompt the daphnid prey to develop individuals which are better defended. Previous work has shown that different predators, e.g. fish and the phantom midge *Chaoborus* spp., can induce different, sometimes opposite phenotypic reactions in the same species or clone (Boersma, Spaak, & De Meester, 1998; Lüning, 1995; Stibor & Luning, 1994; R. Tollrian & Dodson, 1999; Weiss, Kruppert, Laforsch, & Tollrian, 2012). This means that the genome must encode multiple developmental programs triggered by environmental conditions. Induced defenses in *Daphnia* include prominent morphological modifications: from tiny cuticular teeth to very elongated tail and head spines, helmets or even giant crests (Beaton & Hebert, 1997; Juračka, Laforsch, & Petrusek, 2011; Laforsch & Tollrian, 2004; Petrusek, Tollrian, Schwenk, Haas, & Laforsch, 2009), but also changes in life history and different behaviours, which ultimately all act as deterrents to encounter, capture and ingestion by the predator (S. I. Dodson, 1974, 1989; Hebert, 1978; Hülsmann, Vijverberg, Boersma, & Mooij, 2004; Latta Iv, Bakelar, Knapp, & Pfrender, 2007).

In the model species *Daphnia pulex*, kairomones from the phantom-midge larvae *Chaoborus* trigger production of neck-teeth, the most easily detectable trait, and overall hardening of the cuticle (Laforsch, Ngwa, Grill, & Tollrian, 2004). These external, cuticle-associated alterations effectively protect juveniles from predation (Repka, Walls, & Ketola, 1995; R. Tollrian, 1995b). At the same time, induced *D. pulex* females shift resources from reproduction to somatic growth, thereby reaching maturity at a larger size and producing less but larger offspring (Riessen, 1999, 2012; R. Tollrian, 1995a). Vertical migration is deemed to comprise the main behavioural reaction to the presence of the predator in *D. pulex*: thus, *Chaoborus*-induced specimens prefer shallower depths in comparison to control specimens (Boeing, Ramcharan, & Riessen, 2006; S. Dodson, 1988). *Chaoborus* is an ambush predator, such that *Daphnia* is also expected to reduce its swimming speed, although in the case of *D. pulex* this habit is displayed only by some clones (S. I. Dodson, 1996; S. I. Dodson, Hanazato, & Gorski, 1995; Weber & Van Noordwijk, 2002) (LCW, unpublished observations).

Instability of environmental conditions (periodicity of predation risk, different predators) and costs of defenses explain the inducible nature of the defensive morphs. This is also in line with the fact that the neck-teeth are present only in certain juvenile instars when the daphnids reach preferred prey size of their gape-limited predators (Riessen, 1992; Riessen & Sprules, 1990; R. Tollrian, 1995a; R. Tollrian & Dodson, 1999; Ralph Tollrian & Harvell, 1999).

Based on the experimental evidence we can make the following predictions regarding the underlying functional classes of effector genes that might contribute to *Daphnia*'s anti-predatory response (Figure 4.1):

1. the structural changes in the cuticle are expected to mirror changes in the amounts/types of cuticle-associated proteins;
2. life history modifications are expected to be controlled by several physiological changes affecting both somatic growth and reproduction;
3. one can predict that other metabolic functions should be down regulated under predation risk in order to allocate energy resources primarily to the above mentioned pathways;
4. all levels of the response must ultimately be controlled by cascades of receptors, humoral factors, signaling pathways and transcription factors.

Currently, the technique of choice suitable for addressing response patterns in gene expression at a genome-wide level with potentially unlimited depth of coverage is RNA-Seq (McGettigan, 2013; Ozsolak & Milos, 2011; Z. Wang et al., 2009). The availability of the *D. pulex* draft genome (J. K. Colbourne et al., 2011) greatly facilitates the power of such analyses in that RNA-Seq reads can be specifically mapped to a particular genomic location. Investigations of the genome-environment interactions in *Daphnia* are ongoing, with the results of the first analyses of differential gene expression patterns in ecological experiments recently becoming available (Asselman et al., 2012; De Coninck et al., 2014; Jansen et al., 2013; Pauwels, Stoks, & De Meester, 2005). A number of features have been discovered that point to an ecological responsiveness of the *D. pulex* genome; e.g. a large overall number of genes, organized in the many families of paralogous genes that in many cases do not show homologies to genes in other organisms, but show differential expression under different environmental conditions (Boucher, Ditlecadet, Dubé, & Dufresne, 2010; J. K. Colbourne et al., 2011; Latta, Weider, Colbourne, & Pfrender, 2012; Pẽalva-Arana, Lynch, & Robertson, 2009; Thomson, Baldwin, Wang, Kwon, & LeBlanc, 2009).

While preliminary analysis of the transcriptomic responses of *D. pulex* to the predator was performed earlier (J. K. Colbourne et al., 2011), gene expression was assessed with tiling microarrays and was restricted to the second juvenile instar, after the onset of neck-teeth production. Here we aimed at providing the first whole-genome analysis of gene expression changes involved in formation of predator-induced defenses in *D. pulex*. To accomplish this, we apply the most versatile technique to study transcriptomes, RNA-Seq and focus on the first juvenile instar, the developmental point at which the defense is expected to unfold.

MATERIAL AND METHODS

DAPHNIA CLONE USED IN THE EXPERIMENT

In contrast to the clone chosen for genome sequencing by the Daphnia Genomics Consortium ("TCO"), the clone we utilized for the experiment (designated as "R9") is known to show pronounced production of defensive morphs in the presence of the phantom midge larvae *Chaoborus* (J. K. Colbourne et al., 2011). It originates from Canada and according to mitochondrial markers belongs to the so-called Panarctic *Daphnia pulex* clade. The TCO clone in turn belongs to a group of populations united under the name "*Daphnia arenaria*" which is likely of hybrid origin with its mitochondrial genome coming from the same clade as R9, while nuclear markers point to closer relationships with North American *Daphnia pulicaria* (J. K. Colbourne et al., 1998; J. K. Colbourne et al., 2011; Vergilino, Markova, Ventura, Manca, & Dufresne, 2011).

EXPERIMENT

We utilized a simple experimental design: one pooled series of *Daphnia pulex* juveniles exposed to *Chaoborus* and a control set of specimens without predator induction. Fifty age-synchronized specimens of the *D. pulex* R9 clone, served as the founding generation for the experimental animals. For induction, the mothers bearing late embryos were exposed to the *Chaoborus* larvae contained in a net cage and fed with 100 juvenile daphnids. Progeny obtained from the induced and control mother specimens was collected at the first juvenile instar, and stored in RNA Later (Qiagen, Hilden, Germany) for 24 h. Subsequently, specimens were stored at -20 °C until RNA extraction. Three independent induction experiments were performed leading to 90 juveniles in each group (induction and control) in total, which were pooled together to level individual variation.

RNA EXTRACTION AND SEQUENCING

Total RNA was extracted with the TRIzol Reagent (Life Technologies, Carlsbad, CA, USA) according to the manufacturer's protocol with modifications. DNA was further depleted with DNase treatment (TurboDNase, Life Technologies, Carlsbad, CA, USA) and its absence was confirmed afterwards via PCR with primers spanning exon-intron boundaries in an α-tubulin gene. Quality and amount of the purified RNA in the samples was analyzed on the Experion System (Bio-Rad, Hercules, CA, USA) with the aid of the Experion RNA StdSens Analysis Kit according to the supplied manual. The samples were shipped to Otogenetics (Norcross, GA, USA) for library preparation and sequencing. cDNA was synthesized with random hexamers after rRNA depletion. Size-selected cDNA fragments (250–300 bp range) were sequenced on an Illumina HiSeq 2000 from both ends. Overall, two sequencing runs were performed yielding 10–20 million 100 bp read pairs per sample.

Quality of the reads was analyzed with FastQC v. 0.10.1 (Andrews, 2012)and necessary filtering steps were performed with trimmomatic v. 0.22 (Bolger et al., 2014). Since a considerable contamination of the data with adapter sequences and non-coding

RNAs (primarily, rRNA) was detected in the first experiment, we performed two rounds of trimmomatic treatment: removal of reads showing similarity to sequences of non-coding RNA (100 bp tiling fragments with 50 bp overlaps were used as queries) (ILLUMINACLIP:non_coding_rnas:4:40:12 MINLEN:81, for positive matches the whole pair was discarded) and adapter and quality trimming on the second step (ILLUMINACLIP:adapters:2:40:9 LEADING:15TRAILING:15MINLEN:36). After each step of contamination removal a quality control analysis using FastQC was performed to check the validity of our data processing steps.

The raw reads are available from the NCBI Sequence Read Archive (BioProject accession number: PRJNA287609).

ASSEMBLY

We employed two principal alternative approaches to assemble the transcriptome: mapping-oriented and *de novo* assembly. For mapping we took the latest set of scaffolds for the *Daphnia pulex* genome: the 06.09.2005 version with further filtering steps as available on http://genome.jgi-psf.org/Dappu1 (5,191 scaffolds with the total length of 197,261,574 bp). “*Daphnia pulex* Genes 2010 beta 3” annotations, provided by the wFleaBase (Genome Informatics Lab, 2005) were utilized for the reference-genes-guided steps. Those loci which were located on the filtered out scaffolds were excluded, and the final gene set comprised 41,561 transcripts in total. Intron length boundaries were estimated based on the official gene annotations.

Mapping of the reads was performed independently with TopHat v. 2.0.6 (Trapnell et al., 2009) and GSNAP v. 2012-12-20 (T. D. Wu & Nacu, 2010). In TopHat the following principal options were specified for mapping paired reads: --read-mismatches 7 --read-gap-length 2 --read-edit-dist 7 --mate-inner-dist 0 --mate-std-dev 100 --min-anchor-length 5 --min-intron-length 10 --max-intron-length 50000 --max-insertion-length 3 --max-deletion-length 3 --microexon-search --segment-mismatches 3 --segment-length 18 --no-coverage-search --min-segment-intron 10 --max-segment-intron 50000 --min-coverage-intron 10 --max-coverage-intron 20000 --b2-sensitive --report-secondary-alignments --max-multihits 10. Singleton reads decoupled during read filtering were mapped independently with analogous parameters, except that no new junctions were allowed (--no-novel-juncs), besides junctions, obtained on the first step, as recommended by the developers. Mapping options for GSNAP were as follows: --max-mismatches=0.07 --suboptimal-levels=2 --novelsplicing=1 --localsplicedist=50000 --pairmax-rna=50000 --sam-multiple-primaries with the paired reads and singletons analyzed together.

Transcripts for both mapping approaches were assembled with Cufflinks v. 2.0.2 (Roberts, Pimentel, Trapnell, & Pachter, 2011; Trapnell et al., 2010). The assembly was performed with the following general parameters: --multi-read-correct --upper-quartile-norm --max-intron-length 50000 --min-intron-length 10 --overlap-radius 10 and the bias correction option turned on. Two alternative assemblies were made: *de novo* assembly without predefined annotations and reference-guided assembly (the --GTF option) with cuffmerge and cuffcompare, respectively, used to compare the reconstructed transcripts to the reference.

Mapping-independent *de novo* assembly was done with Trinity r. 2012-10-05 (Grabherr et al., 2011). Input dataset for Trinity was prepared as follows. First, for the paired reads in each pair an attempt was made to merge the mates in one longer fragment with the requirement of at least 15 bases of overlap and allowance of no more than 2 % mismatches, with the fastq-join program from the ea-utils package v. 1.1.2 (http://code.google.com/p/ea-utils). Resulting merged fragments, remaining paired reads and singleton reads were treated as unpaired sequenced and were united together both from the control and treatment with subsequent exclusion of redundant identical sequences by fastx_collapser, a program from the FASTX-Toolkit v. 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit). Resulting dataset was processed by Trinity with defaults settings, with the exception of butterfly --max_diffs_same_path=5 --min_per_align_same_path=0.95 -SW options and minimal contig length of 60 nucleotides. To obtain expression levels for the Trinity contigs, individual reads were aligned back to them with the aid of Bowtie2 v. 2.0.4 (Langmead & Salzberg, 2012) with --end-to-end --sensitive --rdg 7,3 --rfg 7,3 settings separately for the control and the treatment. Resulting mappings were processed by cufflinks with the following parameters: --multi-read-correct --upper-quartile-norm --overlap-radius 10 and the whole contigs specified as predefined annotations, as well as input for bias correction. To define genomic positions of the Trinity contigs, we mapped them to the reference genome with blat (default settings for RNA and minimal identity 93 %). To detect presence and classes of overlaps of the Trinity contigs and the official annotations, cuffcompare from the cufflinks package was used.

Differential expression on the level of genes was assessed with cuffdiff with parameters compatible with those of cufflinks.

FUNCTIONAL ANNOTATIONS

Two different sources of information on gene functions were used:

- Official annotations as supplied with the “Genes 2010 beta 3” gene set.
- InterProScan (Quevillon et al., 2005) motif-based analysis for the whole official gene set was performed with the RunIprScan v. 1.0.0 client (http://michaelrthon.com/runiprscan/) in April 2013.

Accordingly, two sets of gene ontology (GO) assignments were used for GO-term enrichment analysis: from the official gene annotations and from the information provided by InterPro for individual domains and families.

Predictions made with RunIprScan were expanded to all members of the respective families of paralogous genes. These families were defined based on pairwise similarity of the amino-acid sequences of all genes longer than 16 amino-acids with the aid of blat, similar to the approach utilized for detection of potentially indistinguishable cDNA sequences outlined above. A pair of proteins were assigned to the same family if alignment segments with a minimum of 60 % identity covered at least 90 % of the longer protein sequence, which can be considered as a conservative and safe approach (compare with Addou, Rentzsch, Lee, and Orengo (2009)).

Term enrichment analyses for GO-terms, gene families and InterPro domain and family assignments were performed in GOseq v. 1.12.0 (Young et al., 2010) for lists of differentially expressed genes obtained with mapping-based approaches only. All genes we weighted by their length and the resulting probability weighting function was used for the over-representation analysis with default settings, independently for up- and down-regulated genes. Only genes with respective annotations (GO classes, gene family membership etc.) were involved in the calculation. Lists of over-represented terms were filtered by controlling false discovery rate at the level of 0.025 within tests.

ASSIGNING BIOLOGICAL FUNCTION TO GENES

The multitude of genes without identifiable orthologs in other animal genomes makes correct functional characterization of *Daphnia* genes challenging. In the current work, we overcame this obstacle in two ways. First, we used not only the gene function assignments as reported in the official gene prediction set, but also results from an independent domain-oriented analysis (InterProScan). The second strategy was to expand these functional designations to whole families of paralogous genes defined in a conservative way. This last approach inevitably leads to propagation of false positives, but nevertheless it provides a good starting point for gene function prediction in a situation when no other source of evidence is available. The *D. pulex* genome contains an extraordinarily large number of genes organized in families of paralogs (J. K. Colbourne et al., 2011). Given this multiplicity of gene families, even with sensitive mapping-oriented transcriptome assembly, there is always uncertainty about precise locations of the reads coming from genes which are very similar on the mRNA sequence level. This problem becomes even more evident when dealing with *D. pulex* clones different from the one chosen for genome sequencing, as was the case in this study. If not compensated, this can lead to incorrect estimations of differential expression as well as biased results with regard to enrichment analyses. As described in detail below, we decided to select only a single gene from a group if the similarity between their transcripts exceeded a given threshold, 90 % of a pairwise alignment covered by 96 % identical segments.

DETECTION OF POTENTIALLY INDISTINGUISHABLE CDNA SEQUENCES

Highly similar cDNA sequences were detected with blat (Kent, 2002): mRNA sequences from the reduced official gene set were used as the query and the target in the same run. blat alignment segments with 96 % identity for individual hits were merged and pairs with overall alignment length at least 90 % were detected as potentially indistinguishable. Individual groups of highly similar genes were created based on the resulting network of pairwise hits. For down-stream analyses only single representatives of the respective groups were utilized.

ANALYSIS BACKBONE

All of the necessary format conversions and file rearrangements were performed with the aid of SAMtools v. 0.1.18 (H. Li et al., 2009), standard Unix commands, MySQL queries and custom PHP, R and bash scripts. The data and the reference genome were visualized and manipulated on a local installation of the UCSC Genome Browser (James Kent et al., 2002) with the assembly 06.09.2005 of the *Daphnia pulex* genome (J. K. Colbourne et al., 2011) and associated annotations (raw files were downloaded from the wFleaBase (Genome Informatics Lab, 2005)). MySQL-database for the genome browser was further customized to incorporate the results of our functional annotations and information on the lists of differentially expressed genes (see above) and besides the standard interface was accessed directly with SQL commands and a local web-based search tool. Venn and pie diagrams were constructed with the aid of the VennEuler v. 1.1-0 R package (Wilkinson, 2011) and the SVGGraph library (http://www.goat1000.com/svggraph.php), respectively.

RESULTS

RNA-SEQ DATA QUALITY

After the stringent filtering steps, 8.0 and 8.6 million read pairs, as well as 6.7 and 7.5 million singleton reads were retained for the control and treatment libraries, respectively. 58.8–62.5 and 85.1–90.7 % of the reads were successfully mapped to the reference genome by TopHat and GSNAP. On average, 19,652–26,008 transcripts encompassed at least one mapped fragment, with the mean number of mapped fragments 441.4–605.6 per transcript. Analysis of the obtained assemblies yielded 288 and 364 differentially expressed (DE) genes for TopHat and GSNAP respectively (see Supplementary Figure 4.S1 and 4.S2). The discrepancies between the two mapping methods could not be attributed to the potential differences in read assignment in cases of very similar paralogs as only one such unambiguous case was detected (data not shown). *De novo* reconstruction of full-length transcripts was limited due to moderate read coverage. Therefore, the contigs obtained by the reference-independent assembler Trinity were used only as a supplement to the two mapping-oriented approaches.

LISTS OF DIFFERENTIALLY EXPRESSED GENES

Reference-independent and reference-guided TopHat and GSNAP assemblies yielded very similar lists of DE genes (data not shown), and all of the DE genes were previously annotated in the reference genome. For further analysis, the lists produced with these assembly strategies were merged producing two united lists for the two mapping programs (referred below as “TopHat” and “GSNAP” lists).

Overall, 435 genes were identified as differentially expressed by at least one approach. The two mapping approaches and the *de novo* assembly yielded similar lists of DE genes. 256 genes shown to be regulated by at least two approaches were considered the strongest candidates for differential expression (see Supplementary Figure 4.1 and 4.2). From the sets of DE genes we further excluded paralogous genes with nearly identical mRNA sequences, since the actual number of members being differentially expressed cannot be deduced for gene groups with high sequence similarities. A random representative of each group was retained.

The list of DE genes identified by at least two approaches is composed of a set of 230 genes: 158 up- and 72 down-regulated in the presence of the kairomone. Distribution of the log_2 fold changes in expression levels for these genes is shown in Figure 4.2. Several genes were shown to be regulated in an on/off manner (i.e. showing no expression in either control or treatment) by individual assembly approaches. After averaging over the assemblies this strictly binary regulation was retained for only one of them: *hxAUG26rep1s6g18t1*, a protein without assigned function in the official gene set, for which a mollusc metallothionein family 2 signature was identified by InterProScan (see Supplementary Figure 4.S1). Published results generated by qPCR (for predefined candidate genes) and tiled genomic microarrays to examine differentially expressed genes after treating *Daphnia* with *Chaoborus* kairomones (J. K. Colbourne et al., 2011; Miyakawa et al., 2010) were compared with our RNA-Seq dataset. All three lists show a low degree of overlap with the other two sets, but nevertheless the lists compiled from the results of the microarray and RNA-Seq analyses do share 31 genes with concordant patterns of expression, as well as four additional genes which show differential regulation in the opposite direction.

FUNCTIONAL ANNOTATIONS

To functionally characterize DE genes, we used two independent sources of information: official gene annotations and InterProScan domain predictions.

In many cases even single amino-acid differences precluded domain identification in some otherwise identical proteins. Thus, to increase the power of the enrichment analyses, functional assignments and gene ontologies were interpolated from hits to individual members of paralogs to whole families of paralogous genes. As the gene family assignments provided with the official gene annotations were too broad, we performed independent, more stringent analysis of similarity between protein sequences.

OVER-REPRESENTED INTERPRO TERMS

Table 4.1 shows significantly over-represented domains and families as identified by InterProScan for the gene sets obtained with the aid of the two mapping methods. Among the up-regulated genes, genes coding for presumed cuticle-associated proteins (with 26–32 of them tagged as “insect cuticle proteins”) are most prevalent. Less abundant are proteins with evidence of lipid transport domains (lipoproteins), vitellogenin domains, and vWF domains, followed by genes coding for globins (together with cruorins) and cyclins.

The list of down-regulated candidates is enriched for genes coding for proteins with lectin-C, CUB, fibrinogen, collagen, TNF-like and complement C1q domains as well as proteins assigned to the GNS1/SUR4 family of unknown molecular function.

GO-ENRICHMENT ANALYSIS

Two sets of GO-term assignments were used in the GO-enrichment analysis: one deriving from the official gene annotations and another obtained with domain and family annotations reported by InterProScan. Corresponding lists of over-represented ontologies are shown in Tables 4.2 and 4.3. The terms obtained with InterProScan tend to be more precise and some of them have no corresponding categories in the second list, such as “regulation of cell cycle”, as well as the most abundant term pointing to cuticle-associated proteins. Over-representation of yolk proteins (“nutrient reservoir activity”) is explicitly detected as such only with the official annotations. For the down-regulated genes only one term was detected as being significantly over-represented with both annotation sources: carbohydrate binding, with collagen-related terms being additionally represented only in the InterPro-based list.

CHROMATIN AND CELL CONTROL PROTEINS

Many of the up-regulated genes detected by RNA-Seq code for chromatin-associated or cell-cycle promoting proteins, although not all of the respective functions have been shown to be significantly over-represented (Figure 4.3). Among them are nucleosome assembling proteins such as CAF-1, and histones H3 and H2b. Another histone H2b variant, distinct at the mRNA level (84.5 % identity), but similar on the amino-acid level (with the exception of the N-terminus, overall 77.2 % identity) is down-regulated. Cyclins encoded by three up-regulated genes belong to the A (1 gene) and B (2 genes) types.

CHEMORECEPTORS AND HORMONES

Among the DE genes detected by at least two assembly approaches only one is annotated as a gustatory receptor in the official gene set, *hxAUG25p1s10g327t1* possessing a Scavenger receptor CD36 domain as reported in the wFleaBase (Genome Informatics Lab, 2005) and by InterProScan (Jones et al., 2014), with a fold change of 6.2 in our experiment. The only protein with identifiable humoral function in the RNA-Seq list of DE genes is an uncharacterized gene *hxAUG26res18g88t1* with insulin-like domains identified by InterProScan. Its expression showed a 7.0 fold increase in the kairomone-treated juveniles (see Supplementary Figure 4.S1).

OVER-REPRESENTED FAMILIES OF PARALOGS

In the list of genes shown to be differentially expressed by at least two approaches, 27 % of the up-regulated and 22 % of the down-regulated genes belong to families of paralogs. Some of these families are represented by several candidate genes in our list of DE genes and among them a considerable number of families is significantly over-represented (Table 4.4). The largest family among the up-regulated genes includes genes coding for products similar to a) pupal cuticle proteins, followed by b) Zinc-metalloproteinases, c) vitellogenin, d) a second family of cuticle-associated proteins, e) globin-cruorins, and f) other smaller families (Table 4.4). Products of gene families with significant down-regulation are characterized as a) C-type lectins, b) proteins of unknown molecular function with similarity to the C1q complement protein, c) proteins involved in elongation of very long chain fatty acids, and d) other less abundant families with nearly (Table 4.4).

DISCUSSION

GENE EXPRESSION PATTERNS IN THE PHYSIOLOGICAL CONTEXT

Our RNA-Seq results generally agree with the hypothesis proposed in the introduction (see also Figure 4.1):

1. The most abundant and significantly over-represented functional group of up-regulated genes is composed of genes coding for cuticle-associated proteins (Tables 4.1, 4.2 and 4.4), which directly echoes morphological observations: i.e.,

production of neck-teeth (the main defense mechanism of the juvenile *D. pulex*), and changes in cuticle ultrastructure (Laforsch et al., 2004).

In addition, we observed increased transcription of genes involved in chromatin restructuring and the cell cycle (cyclins). This is likely related to the increased proliferative activity recently reported in the region underlying neck-teeth in induced *D. pulex* juveniles (Naraki, Hiruta, & Tochinai, 2013).

2. A clear hint of resource re-allocation is suggested by the increase in expression of genes involved in lipid transport and metabolism, as well as globins. Six vitellogenin (precursor of the major yolk protein, vitellin (Kato, Tokishita, Ohta, & Yamagata, 2004; Zaffagnini, 1987) genes belong to this group as well. Production of yolk in daphnids starts as early as late juvenile stages, but the onset of vitellogenin mRNA synthesis takes place even earlier (J. Kim et al., 2011; Stibor, 2002; Tokishita et al., 2006). Vitellogenin is synthesized in fat bodies (Zaffagnini & Zeni, 1986); thus, the presence of residual maternal mRNA in our experiment can be excluded. In this respect, the increased expression of vitellogenin genes seems to point to one of the following factors or their combination: earlier onset of vitellogenesis, increased fecundity, or increased size of progeny. In a physiological study of vitellogenesis in *D. magna*, it was discovered that induction with the *Chaoborus* kairomone has no effect on the onset of yolk production, but decreases its rate (Stibor, 2002). Moreover, in a recent proteomics study of *D. magna* responses to another invertebrate predator, *Triops*, vitellogenin was shown to be among the proteins with decreased

production in induced specimens (Otte, Fröhlich, Arnold, & Laforsch, 2014). These observations were based on protein content measurements for a distantly related species producing no neck-teeth, likely explaining the clear contradiction with our results. However, for *D. pulex* it was shown that kairomone induction leads to production of bigger offspring (R. Tollrian, 1995a), which presumably requires a larger pool of yolk.

3. Among functional groups significantly over-represented for the down-regulated genes, we find a large number of genes coding for proteins with domains that play various cellular roles: lectins, and proteins with CUB, fibrinogen and TNF domains. Intriguingly, proteins containing these domains function in immune responses in other invertebrates (Fujita, Matsushita, & Endo, 2004; Hanington & Zhang, 2011; L. C. Smith, Azumi, & Nonaka, 1999). Although many details of molecular mechanisms of immune responses in Cladocera remain unknown (Auld, 2014; McTaggart, Conlon, Colbourne, Blaxter, & Little, 2009), decreased expression of these proteins may be causally connected to the observation that inducible defenses in *Daphnia* lead to decreased resistance to diseases (Yin, Laforsch, Lohr, & Wolinska, 2011).
4. Potential regulatory genes involved in metabolism of hormones and neurotransmitters or coding for chemoreceptors are nearly absent from our lists of differentially expressed genes. Among the up-regulated genes we found a gustatory receptor, which was designated as such in wFleaBase (Genome Informatics Lab, 2005), but was not reported in an extensive *in silico* study of

chemoreceptors in *D. pulex* (Pẽalva-Arana et al., 2009) (see Supplementary Figure 4.S1). We speculate that this receptor may be involved in perception of the kairomone, but this requires further experimental evidence.

No genes with identifiable roles in transcriptional regulation were discovered with our approach, which is likely related to the moderate sequencing coverage of the RNA-Seq.

COMPARISON TO PREVIOUS STUDIES, FUTURE PERSPECTIVES

A comparison of the lists of DE genes discovered in *Chaoborus*-induction experiments on juvenile *D. pulex* presented here with the results obtained from tiling microarrays shows that the two lists share a group of 31 genes showing concordant expression patterns. The overall discrepancy is nevertheless noticeable and may be attributed to differences in experimental set-up and/or water conditions and not to the differences in the platform, since comparative analyses generally reveal good correlation between microarray- and sequencing-based techniques (Babbitt et al., 2010; Nookaew et al., 2012; Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014). Although both experiments were performed on the same *D. pulex* clone, the stage chosen for the microarray experiment was more advanced in comparison to the animals utilized in the current study. This observation signifies the necessity of an experiment involving several developmental stages to sample genes involved in different steps of the predator perception, signal perception and neck-teeth production and maintenance.

It is clear that any results obtained for a single genetic clone should not be over-extrapolated. More experiments on different *D. pulex* clones are necessary to make firm conclusions about the intraspecific variation of the genetic mechanisms acting in the anti-predatory response. Moreover, the existence of neck-teeth producing *Daphnia* species not directly related to *D. pulex* (Juračka et al., 2011) calls for even broader sample of species to investigate the differences in trajectories ultimately leading to similar morphological features.

CONCLUSIONS

This study provides important insights into gene regulatory patterns underlying predator-induced defenses, utilizing for the first time unbiased whole-transcriptome RNA-Seq expression data. In particular, our study characterizes different effector genes, gene families underlying morphological, and life-history changes, which are largely in agreement with expectations based on observed phenotypic changes. Our data represent the largest dataset on the genetic basis of anti-predator defenses in *Daphnia* to date and add an important contribution to link a phenotypically plastic response directly with the underlying molecular genetic processes. A deeper understanding of these processes would be achieved with experiments on different genetic clones and at different developmental stages.

ACKNOWLEDGEMENTS

This work was supported in part by a grant of the Dinter-Foundation (Deutsches Stiftungszentrum) to FL and RT, startup funding of JRM and funding of RT.

EPILOGUE

This chapter was published in BMC Frontiers of Zoology, volume 12, number of pages 18, 2015. I am a co-first author in this manuscript. Andrey Rozenberg and I have both done parallel transcriptome analysis of this dataset. I focused primarily on quality control and mapping the sequencing data to the reference genome using Tophat (Trapnell et al., 2009). My analysis was essential in validating his gene expression data and allowed the discovery of key genes involved in the defense process, immune response, and resource re-allocation during PIPP in *D. pulex*. This was the first study to perform RNA-Seq analysis to detect gene expression patterns associated with PIPP in *D. pulex* (Rozenberg et al., 2015). The conclusions from this study are an important addition towards understanding the underlying genetic processes related to PIPP in *D. pulex*, at a whole genome level.

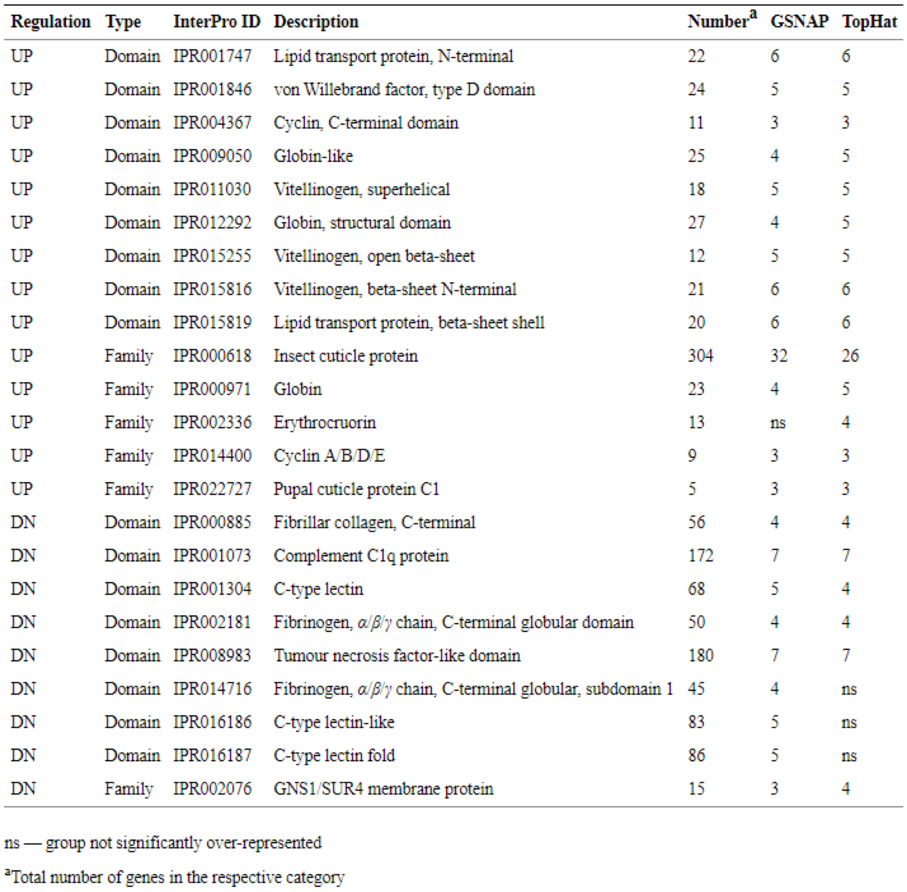
My role in this project involved the following tasks:

- Cleaning the adapter contaminants from the raw sequencing data to enrich the quality of our dataset. For all of our samples, I generated quality control data and plotted them using FastQC (Andrews, 2012) to visualize the quality of our dataset. I downloaded a list of known contaminants such as adapter and primer sequences used by multiple sequencing experiments from Tamir (2012). I used FastQC to search for matching contaminants in our dataset. My FastQC analysis showed an over-representation of adapter contaminants from my list in our sequencing data. Adapter and primer sequences aid in the sequencing of a DNA fragment but they are biologically irrelevant to our research goals. However, their presence can potentially contaminate the sequence alignment step to the reference genome. Therefore, I trimmed them from our sequence data using Trimmomatic (Bolger et al., 2014) and applying the parameters mentioned in the subsection “RNA extraction and sequencing” under the “Materials and methods”. I maintained a minimum sequence length of 36 bp post trimming and excluded sequences shorter than this length to allow unique alignment of a sequence to a genomic region.

- Theoretically, a library of DNA fragments should exhibit a diverse set of sequences. FastQC assigns an over-representation tag to a sequence if the same sequence represents at least 1% of the entire sample. Therefore, it is crucial to determine the annotation of these sequences to decide the biological relevance to our research goals. I did a nucleotide BLAST (Altschul et al., 1990) search against the non-redundant database for our sequences that were showing over-representation according to FastQC but did not match to any sequencing contaminants in my list. These sequences were primarily ribosomal RNA (rRNA). rRNA represent a substantial portion of the total RNA in a cell (Eun, 1996). They play a key role in translation of mRNA to protein (Eun, 1996). However, they are non-protein coding in nature and are not associated directly to our research goal. Therefore, I excluded sequences from our dataset using Trimmomatic and applying the parameters mentioned in the subsection “RNA extraction and sequencing” under the “Materials and methods”. The post-processing quality control plots showed good quality of our sequence data.

- Mapped our processed sequences to the *D. pulex* reference genome (version-06.09.2005, http://genome.jgi-psf.org/Dappu1) using the TopHat (version-2.0.6) algorithm (Trapnell et al., 2009). I used a small subset of sequences from one sample of our dataset to run a combination of parameters that aligned maximum amount of uniquely mapped sequences to the *D. pulex* genome. Next, I applied those parameters in Tophat and mapped all our samples to the reference genome.
- Used cuffdiff algorithm (Trapnell et al., 2013), to generated significantly differentially expressed (SDE) list of genes (q-value <0.05, to reduce false positives) by comparing the predator induced *Daphnia* samples (experimental) against the control samples. I selected standard parameters suggested by the cuffdiff manual.

I believe my quality control and alignment strategy can be applied in any research problem aiming towards gene expression pattern analysis using RNA-Seq, regardless of the model organism, to account for over-representative sequencing contaminants, non-coding RNA and improving parameter selection of the sequence mapping process.



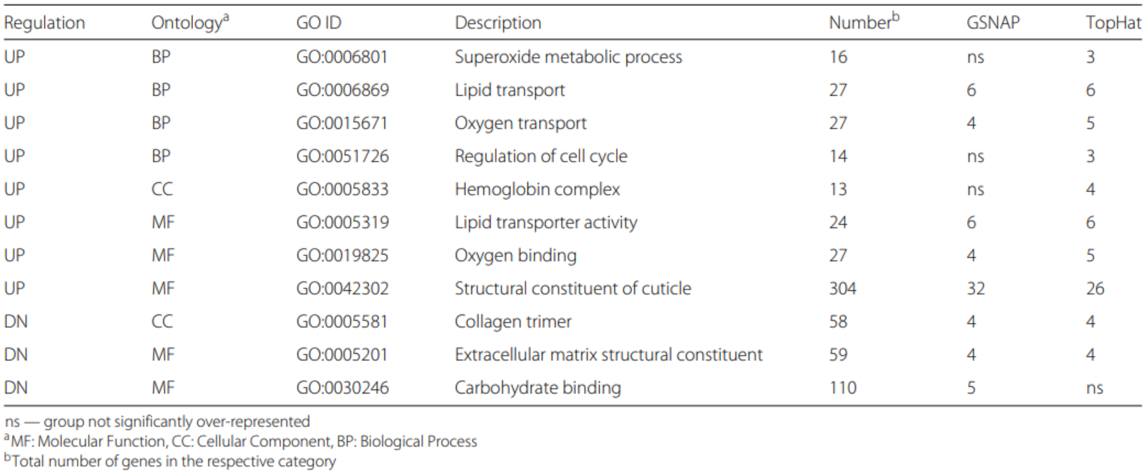
Regulation	Type	InterPro ID	Description	Number[a]	GSNAP	TopHat
UP	Domain	IPR001747	Lipid transport protein, N-terminal	22	6	6
UP	Domain	IPR001846	von Willebrand factor, type D domain	24	5	5
UP	Domain	IPR004367	Cyclin, C-terminal domain	11	3	3
UP	Domain	IPR009050	Globin-like	25	4	5
UP	Domain	IPR011030	Vitellinogen, superhelical	18	5	5
UP	Domain	IPR012292	Globin, structural domain	27	4	5
UP	Domain	IPR015255	Vitellinogen, open beta-sheet	12	5	5
UP	Domain	IPR015816	Vitellinogen, beta-sheet N-terminal	21	6	6
UP	Domain	IPR015819	Lipid transport protein, beta-sheet shell	20	6	6
UP	Family	IPR000618	Insect cuticle protein	304	32	26
UP	Family	IPR000971	Globin	23	4	5
UP	Family	IPR002336	Erythrocruorin	13	ns	4
UP	Family	IPR014400	Cyclin A/B/D/E	9	3	3
UP	Family	IPR022727	Pupal cuticle protein C1	5	3	3
DN	Domain	IPR000885	Fibrillar collagen, C-terminal	56	4	4
DN	Domain	IPR001073	Complement C1q protein	172	7	7
DN	Domain	IPR001304	C-type lectin	68	5	4
DN	Domain	IPR002181	Fibrinogen, $\alpha/\beta/\gamma$ chain, C-terminal globular domain	50	4	4
DN	Domain	IPR008983	Tumour necrosis factor-like domain	180	7	7
DN	Domain	IPR014716	Fibrinogen, $\alpha/\beta/\gamma$ chain, C-terminal globular, subdomain 1	45	4	ns
DN	Domain	IPR016186	C-type lectin-like	83	5	ns
DN	Domain	IPR016187	C-type lectin fold	86	5	ns
DN	Family	IPR002076	GNS1/SUR4 membrane protein	15	3	4

ns — group not significantly over-represented

[a]Total number of genes in the respective category

Table 5.1. Significantly over-represented InterPro domains and families among the differentially expressed genes. In total 46,928 annotations for 18,168 genes were available. The last two columns represent gene counts for significantly over-represented groups as revealed with the aid of the two mapping strategies.

Source: Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology*, *12*, 18. doi:10.1186/s12983-015-0109-x



Regulation	Ontology[a]	GO ID	Description	Number[b]	GSNAP	TopHat
UP	BP	GO:0006801	Superoxide metabolic process	16	ns	3
UP	BP	GO:0006869	Lipid transport	27	6	6
UP	BP	GO:0015671	Oxygen transport	27	4	5
UP	BP	GO:0051726	Regulation of cell cycle	14	ns	3
UP	CC	GO:0005833	Hemoglobin complex	13	ns	4
UP	MF	GO:0005319	Lipid transporter activity	24	6	6
UP	MF	GO:0019825	Oxygen binding	27	4	5
UP	MF	GO:0042302	Structural constituent of cuticle	304	32	26
DN	CC	GO:0005581	Collagen trimer	58	4	4
DN	MF	GO:0005201	Extracellular matrix structural constituent	59	4	4
DN	MF	GO:0030246	Carbohydrate binding	110	5	ns

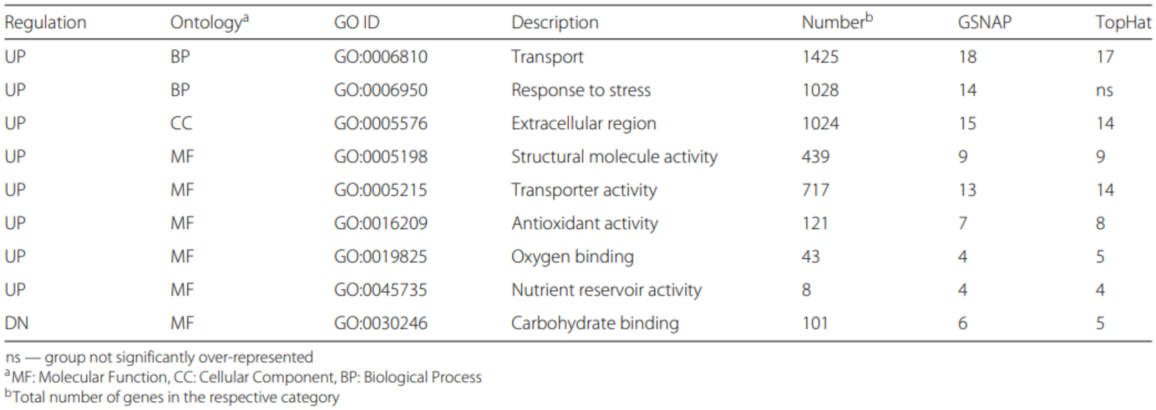
ns — group not significantly over-represented

[a] MF: Molecular Function, CC: Cellular Component, BP: Biological Process

[b] Total number of genes in the respective category

Table 5.2. Over-represented Gene Ontology terms based on the assignments from InterPro-annotations. In total 40,177 annotations for 14,503 genes were available.

Source: Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x



Regulation	Ontology[a]	GO ID	Description	Number[b]	GSNAP	TopHat
UP	BP	GO:0006810	Transport	1425	18	17
UP	BP	GO:0006950	Response to stress	1028	14	ns
UP	CC	GO:0005576	Extracellular region	1024	15	14
UP	MF	GO:0005198	Structural molecule activity	439	9	9
UP	MF	GO:0005215	Transporter activity	717	13	14
UP	MF	GO:0016209	Antioxidant activity	121	7	8
UP	MF	GO:0019825	Oxygen binding	43	4	5
UP	MF	GO:0045735	Nutrient reservoir activity	8	4	4
DN	MF	GO:0030246	Carbohydrate binding	101	6	5

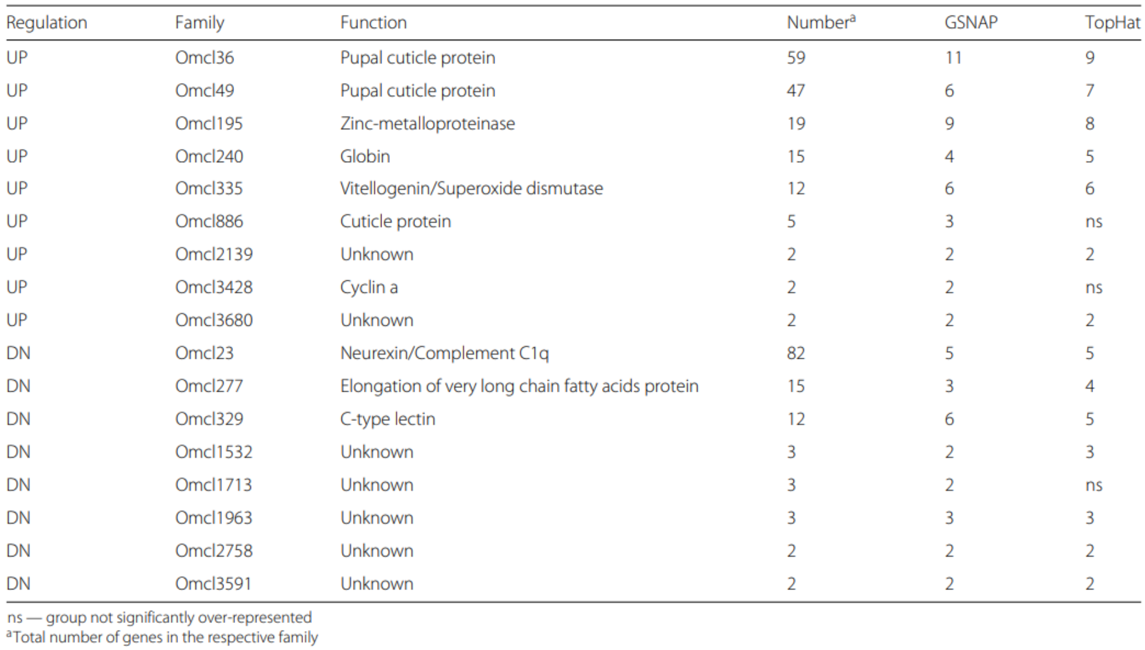
ns — group not significantly over-represented

[a] MF: Molecular Function, CC: Cellular Component, BP: Biological Process

[b] Total number of genes in the respective category

Table 5.3. Over-represented Gene Ontology terms based on the GO assignments in the wFleaBase. In total 87,517 annotations for 13,612 genes were available.

Source: Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x



Regulation	Family	Function	Number^{a}	GSNAP	TopHat
UP	Omcl36	Pupal cuticle protein	59	11	9
UP	Omcl49	Pupal cuticle protein	47	6	7
UP	Omcl195	Zinc-metalloproteinase	19	9	8
UP	Omcl240	Globin	15	4	5
UP	Omcl335	Vitellogenin/Superoxide dismutase	12	6	6
UP	Omcl886	Cuticle protein	5	3	ns
UP	Omcl2139	Unknown	2	2	2
UP	Omcl3428	Cyclin a	2	2	ns
UP	Omcl3680	Unknown	2	2	2
DN	Omcl23	Neurexin/Complement C1q	82	5	5
DN	Omcl277	Elongation of very long chain fatty acids protein	15	3	4
DN	Omcl329	C-type lectin	12	6	5
DN	Omcl1532	Unknown	3	2	3
DN	Omcl1713	Unknown	3	2	ns
DN	Omcl1963	Unknown	3	3	3
DN	Omcl2758	Unknown	2	2	2
DN	Omcl3591	Unknown	2	2	2

ns — group not significantly over-represented

[a] Total number of genes in the respective family

Table 5.4. Over-represented families of paralogous genes based on the wFleaBase annotations. In total 3,978 families encompassed 24,102 genes (genes per family: median — 3, 5–95 % interval — 2–18).

Source: Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x

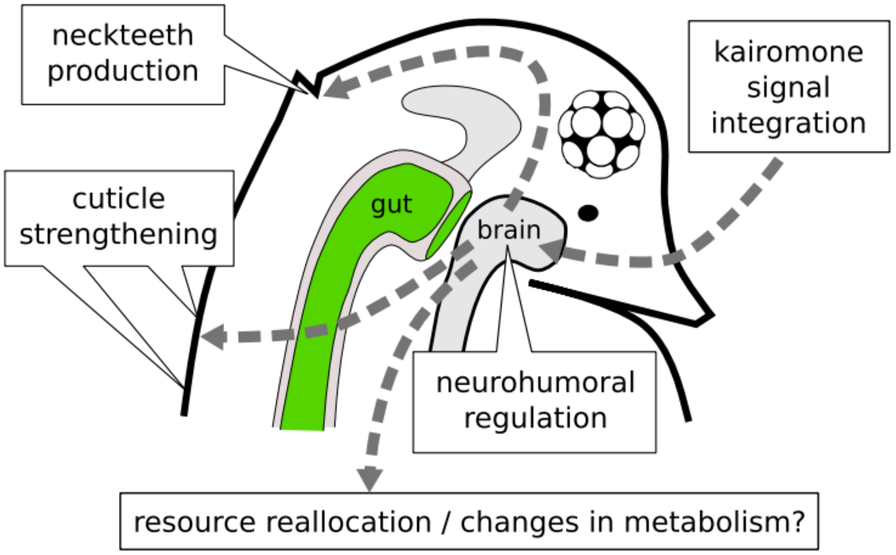

neckteeth production
kairomone signal integration
gut
cuticle strengthening
brain
neurohumoral regulation
resource reallocation / changes in metabolism?

Figure 5.1. Presumptive scheme of physiological and morphological changes during kairomone-induction in *D. pulex*

Source: Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x

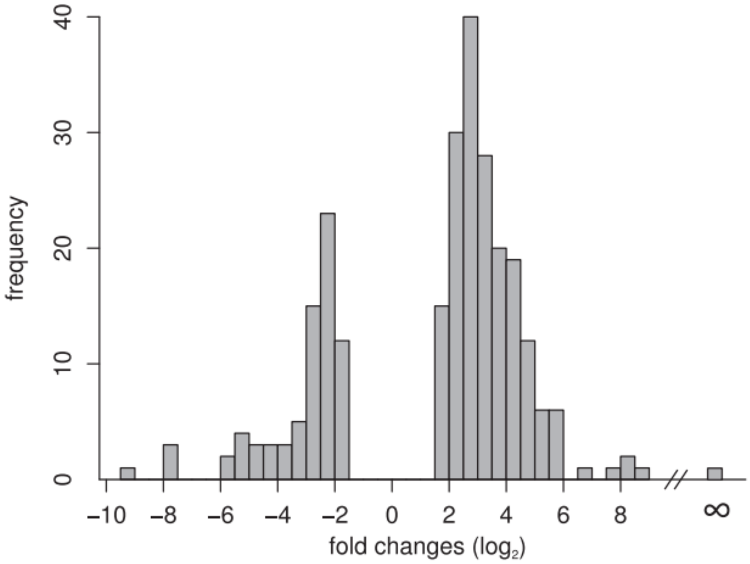


Figure 5.2. Presumptive scheme of physiological and morphological changes during kairomone-induction in *D. pulex*

Source: Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x

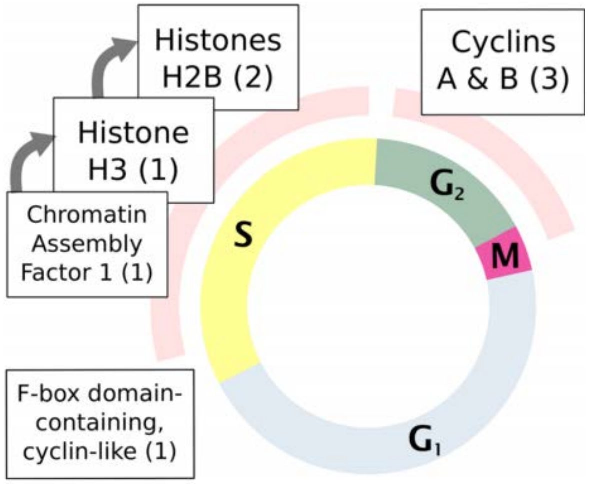


Figure 5.3. Differentially expressed genes involved in cell-cycle control and chromatin regulation with their relative timing as known for other animals. The internal circle represents cell-cycle. The boxes corresponding to different proteins are arranged according to their role in progressing respective cell-cycle stages. The role of the F-box domain-containing protein cannot be predicted with certainty. Numbers in parentheses represent numbers of differentially expressed genes belonging to respective families.

Source: Rozenberg, A., **Parida, M.**, Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x

CHAPTER 6. CONCLUSIONS

In conclusion, my training in bioinformatics has allowed me to contribute to the field of biology and human disease modelling as follows:

- Creating a unique web resource called OikoBase (Gemma. Danks & Parida, 2013) for the chordate genomics and developmental biology community using *O. dioica* as a model organism, that did not exist before. Multiple groups all over the world using *O. dioica* as a model animal to perform research in these fields have benefited from OikoBase. Following are the published journals that have cited OikoBase as a testament to its usefulness, Olsen et al. (2018), Navratilova et al. (2017), Omotezako, Onuma, and Nishida (2015), Omotezako, Matsuo, Onuma, and Nishida (2017), Henriet et al. (2015),   and Appels, Barrero, and Bellgard (2013).
- Spearheading the exome sequencing analysis of a RA pedigree that led to the identification of a novel RA associated gene called *GREB1L* (Brophy et al., 2017). Two independent high impact peer-reviewed journals have verified and cited our work exhibiting this gene's role in kidney and genital development in humans and mice (De Tomasi et al., 2017) and implicating its role in RA and hypodysplasia (RHD) (Sanna-Cherchi et al., 2017). This novel discovery has made a significant contribution towards understanding the genetic basis underlying kidney development and CAKUT in humans. Identification of novel

etiologic variants associated with other Mendelian disorders will expand our knowledgebase on human diseases and development.

- Processing, generating and analyzing transcriptome expression data to understand genetic pathways concerning heavy metal stress response in *D. pulex*. Highlighting key stress response biological processes and identifying important regulatory element binding sites that are potentially driving the gene regulation patterns associated with heavy metal induced oxidative and xenobiotic stress in this organism. This is the biggest gene expression dataset generated to study the underlying genetic basis of heavy metal response in *Daphnia* to date. This data has added to the validity of *D. pulex* as a model organism to do ecotoxicogenomics and contributed substantially towards mapping the links between acute heavy metal exposure and genetic response pathways. Comparing our results against similar studies that have used different exposure periods and dosages will paint a comprehensive atlas of gene expression patterns across time and under different concentrations of heavy metals in *Daphnia*. Knowledge derived from these patterns can be extrapolated to diagnose the exposure periods and dosage of similar heavy metals in other organisms based on their genetic profile. This work is unpublished.
- Processing and generating transcriptome level expression data to understand PIPP in *D. pulex*, which led to the discovery of key genes involved in defense mechanisms, immune response and resource re-allocation. A group comparing within and across generational gene expression patterns concerning PIPP in

> *Daphnia* has cited our work in their published manuscript (Hales et al., 2017). The data generated from this analysis has made a crucial addition towards validating *D. pulex* as a suitable animal model to explore and understand the underlying genetic processes associated with PIPP (Rozenberg et al., 2015).

Nobel laureate Paul nurse argues studying the art of processing data in biological systems will lead to the next “quantum leap” in biology (Hogeweg, 2011). Bioinformatics is constantly aiding medical and biological research in a substantial manner to solve contemporary problems and gain meaningful insights. With the ever increasing size of biological and clinical data the demand for efficient and faster computational approaches will keep increasing. As a bioinformatics scientist I will continue to serve as a principal component in the process of scientific discovery associated with human disease modelling and biology.

REFERENCES

Addou, S., Rentzsch, R., Lee, D., & Orengo, C. A. (2009). Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer. *Journal of Molecular Biology, 387*(2), 416-430. doi:10.1016/j.jmb.2008.12.045

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Agrawal, S., Kelkenberg, M., Begum, K., Steinfeld, L., Williams, C. E., Kramer, K. J., . . . Merzendorfer, H. (2014). Two essential peritrophic matrix proteins mediate matrix barrier functions in the insect midgut. *Insect Biochem Mol Biol, 49*, 24-34. doi:10.1016/j.ibmb.2014.03.009

Akerman, K. E., Honkaniemi, J., Scott, I. G., & Andersson, L. C. (1985). Interaction of Cd2+ with the calmodulin-activated (Ca2+ + Mg2+)-ATPase activity of human erythrocyte ghosts. *Biochim Biophys Acta, 845*(1), 48-53.

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, Custer, A. F., Hicks, M. A., . . . Babbitt, P. C. (2014). The Structure-Function Linkage Database. *Nucleic Acids Res, 42*(Database issue), D521-530. doi:10.1093/nar/gkt1130

Alldredge, A. L. (2005). The contribution of discarded appendicularian houses to the flux of particulate organic carbon from oceanic surface waters. *Response of Marine Ecosystems to Global Change: Ecological Impact of Appendicularians*, 309-326.

Almaden, Y., Canalejo, A., Ballesteros, E., Anon, G., Canadillas, S., & Rodriguez, M. (2002). Regulation of arachidonic acid production by intracellular calcium in parathyroid cells: effect of extracellular phosphate. *J Am Soc Nephrol, 13*(3), 693-698.

Alpers, D. H. (2003). CARBOHYDRATES | Digestion, Absorption, and Metabolism A2 - Caballero, Benjamin *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (pp. 881-887). Oxford: Academic Press.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Altshuler, I., Demiri, B., Xu, S., Constantin, A., Yan, N. D., & Cristescu, M. E. (2011). An integrated multi-disciplinary approach for studying multiple stressors in freshwater ecosystems: Daphnia as a model organism. *Integrative and Comparative Biology, 51*(4), 623-633. doi:10.1093/icb/icr103

Amiard, J. C., Amiard-Triquet, C., Barka, S., Pellerin, J., & Rainbow, P. S. (2006). Metallothioneins in aquatic invertebrates: Their role in metal detoxification and their use as biomarkers. *Aquatic Toxicology, 76*(2), 160-202. doi:https://doi.org/10.1016/j.aquatox.2005.08.015

Andersen, S. O. (1979). Biochemistry of Insect Cuticle. *Annual Review of Entomology, 24*(1), 29-59. doi:10.1146/annurev.en.24.010179.000333

Ando, K., Ozaki, T., Yamamoto, H., Furuya, K., Hosoda, M., Hayashi, S., . . . Nakagawara, A. (2004). Polo-like kinase 1 (Plk1) inhibits p53 function by physical interaction and phosphorylation. *J Biol Chem, 279*(24), 25549-25561. doi:10.1074/jbc.M314182200

Andrews, S. (2011). FastQC. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Andrews, S. (2012). FastQC. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Appels, R., Barrero, R., & Bellgard, M. (2013). Advances in biotechnology and informatics to link variation in the genome to phenotypes in plants and animals. *Funct Integr Genomics, 13*(1), 1-9. doi:10.1007/s10142-013-0319-2

Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., . . . Zhang, J. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research, 40*(D1), D71-D75. doi:10.1093/nar/gkr981

Arosio, P., & Levi, S. (2002). Ferritin, iron homeostasis, and oxidative damage. *Free Radic Biol Med, 33*(4), 457-463.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25*(1), 25-29. doi:10.1038/75556

Asselman, J., De Coninck, D. I. M., Glaholt, S., Colbourne, J. K., Janssen, C. R., Shaw, J. R., & De Schamphelaere, K. A. C. (2012). Identification of pathways, gene networks, and paralogous gene families in daphnia pulex responding to exposure to the toxic cyanobacterium Microcystis aeruginosa. *Environmental Science and Technology, 46*(15), 8448-8457. doi:10.1021/es301100j

Asselman, J., Shaw, J. R., Glaholt, S. P., Colbourne, J. K., & De Schamphelaere, K. A. (2013). Transcription patterns of genes encoding four metallothionein homologs in Daphnia pulex exposed to copper and cadmium are time- and homolog-dependent. *Aquatic Toxicology, 142-143*, 422-430. doi:10.1016/j.aquatox.2013.09.010

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., . . . Mitchell, A. L. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database (Oxford), 2012*, bas019. doi:10.1093/database/bas019

Auld, S. K. J. R. (2014). Physiology of the Cladocera. 219-233.

Ayala, A., Munoz, M. F., & Arguelles, S. (2014). Lipid peroxidation: production, metabolism, and signaling mechanisms of malondialdehyde and 4-hydroxy-2-nonenal. *Oxid Med Cell Longev, 2014*, 360438. doi:10.1155/2014/360438

Babbitt, C. C., Fedrigo, O., Pfefferle, A. D., Boyle, A. P., Horvath, J. E., Furey, T. S., & Wray, G. A. (2010). Both noncoding and protein-coding rnas contribute to gene expression evolution in the primate brain. *Genome Biology and Evolution, 2*(1), 67-79. doi:10.1093/gbe/evq002

Badisa, V. L., Latinwo, L. M., Odewumi, C. O., Ikediobi, C. O., Badisa, R. B., Ayuk-Takem, L. T., . . . West, J. (2007). Mechanism of DNA damage by cadmium and interplay of antioxidant enzymes and agents. *Environ Toxicol, 22*(2), 144-151. doi:10.1002/tox.20248

Bagchi, D., Vuchetich, P. J., Bagchi, M., Hassoun, E. A., Tran, M. X., Tang, L., & Stohs, S. J. (1997). Induction of oxidative stress by chronic administration of sodium dichromate [chromium VI] and cadmium chloride [cadmium II] to rats. *Free Radical Biology and Medicine, 22*(3), 471-478. doi:Doi 10.1016/S0891-5849(96)00352-8

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res, 43*(W1), W39-49. doi:10.1093/nar/gkv416

Baillieul, M., Smolders, R., & Blust, R. (2005). The effect of environmental stress on absolute and mass-specific scope for growth in Daphnia magna Strauss. *Comp Biochem Physiol C Toxicol Pharmacol, 140*(3-4), 364-373. doi:10.1016/j.cca.2005.03.007

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics, 12*, 745. doi:10.1038/nrg3031 https://www.nature.com/articles/nrg3031#supplementary-information

Barak, H., Huh, S. H., Chen, S., Jeanpierre, C., Martinovic, J., Parisot, M., . . . Kopan, R. (2012). FGF9 and FGF20 Maintain the Stemness of Nephron Progenitors in Mice and Man. *Developmental Cell, 22*(6), 1191-1207. doi:10.1016/j.devcel.2012.04.018

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., . . . Soboleva, A. (2011). NCBI GEO: Archive for functional genomics data sets-10 years on. *Nucleic Acids Research, 39*(SUPPL. 1), D1005-D1010. doi:10.1093/nar/gkq1184

Barski, O. A., Tipparaju, S. M., & Bhatnagar, A. (2008). The aldo-keto reductase superfamily and its role in drug metabolism and detoxification. *Drug Metab Rev, 40*(4), 553-624. doi:10.1080/03602530802431439

Bassett, A. R., Tibbit, C., Ponting, C. P., & Liu, J. L. (2013). Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System. *Cell Reports, 4*(1), 220-228. doi:10.1016/j.celrep.2013.06.020

Batourina, E., Tsai, S., Lambert, S., Sprenkle, P., Viana, R., Dutta, S., . . . Mendelsohn, C. L. (2005). Apoptosis induced by vitamin A signaling is crucial for connecting the ureters to the bladder. *Nature Genetics, 37*(10), 1082-1089. doi:10.1038/ng1645

Beaton, M. J., & Hebert, P. D. N. (1997). The cellular basis of divergent head morphologies in Daphnia. *Limnology and Oceanography, 42*(2), 346-356. doi:10.4319/lo.1997.42.2.0346

Beischlag, T. V., Luis Morales, J., Hollingshead, B. D., & Perdew, G. H. (2008). The aryl hydrocarbon receptor complex and the control of gene expression. *Crit Rev Eukaryot Gene Expr, 18*(3), 207-250.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research, 45*(Database issue), D37-D42. doi:10.1093/nar/gkw1070

Berry, R., Harewood, L., Pei, L., Fisher, M., Brownstein, D., Ross, A., . . . FitzPatrick, D. R. (2011). Esrrg functions in early branch generation of the ureteric bud and is essential for normal development of the renal papilla. *Human Molecular Genetics, 20*(5), 917-926. doi:10.1093/hmg/ddq530

Bi, Y., Lin, G. X., Millecchia, L., & Ma, Q. (2006). Superinduction of metallothionein I by inhibition of protein synthesis: role of a labile repressor in MTF-1 mediated gene transcription. *J Biochem Mol Toxicol, 20*(2), 57-68. doi:10.1002/jbt.20116

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Research, 14*(5), 988-995. doi:10.1101/gr.1865504

Biswas, M., & Chan, J. Y. (2010). Role of Nrf1 in antioxidant response element-mediated gene expression and beyond. *Toxicol Appl Pharmacol, 244*(1), 16-20. doi:10.1016/j.taap.2009.07.034

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics-Theory and Experiment*. doi:Artn P10008
10.1088/1742-5468/2008/10/P10008

Bochud, M., Jenny-Burri, J., Pruijm, M., Ponte, B., Guessous, I., Ehret, G., . . . Ackermann, D. (2018). Urinary Cadmium Excretion Is Associated With Increased Synthesis of Cortico- and Sex Steroids in a Population Study. *The Journal of Clinical Endocrinology & Metabolism, 103*(2), 748-758. doi:10.1210/jc.2017-01540

Bodar, C. W. M., Van Leeuwen, C. J., Voogt, P. A., & Zandee, D. I. (1988). Effect of cadmium on the reproduction strategy of Daphnia magna. *Aquatic Toxicology, 12*(4), 301-309. doi:https://doi.org/10.1016/0166-445X(88)90058-6

Boeing, W. J., Ramcharan, C. W., & Riessen, H. P. (2006). Multiple predator defence strategies in Daphnia pulex and their relation to native habitat. *Journal of Plankton Research, 28*(6), 571-584. doi:10.1093/plankt/fbi142

Boersma, M., Spaak, P., & De Meester, L. (1998). Predator-mediated plasticity in morphology, life history, and behavior of Daphnia: The uncoupling of responses. *American Naturalist, 152*(2), 237-248. doi:10.1086/286164

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bollig, F., Mehringer, R., Perner, B., Hartung, C., Schäfer, M., Schartl, M., . . . Englert, C. (2006). Identification and comparative expression analysis of a second wt1 gene in zebrafish. *Developmental Dynamics, 235*(2), 554-561. doi:10.1002/dvdy.20645

Bollig, F., Perner, B., Besenbeck, B., Köthe, S., Ebert, C., Taudien, S., & Englert, C. (2009). A highly conserved retinoic acid responsive element controls wt1a expression in the zebrafish pronephros. *Development, 136*(17), 2883-2892. doi:10.1242/dev.031773

Bouchard, M. (2004). Transcriptional control of kidney development. *Differentiation, 72*(7), 295-306. doi:10.1111/j.1432-0436.2004.07207001.x

Bouchard, M., Souabni, A., Mandler, M., Neubüser, A., & Busslinger, M. (2002). Nephric lineage specification by Pax2 and Pax8. *Genes and Development, 16*(22), 2958-2970. doi:10.1101/gad.240102

Boucher, P., Ditlecadet, D., Dubé, C., & Dufresne, F. (2010). Unusual duplication of the insulin-like receptor in the crustacean Daphnia pulex. *BMC Evolutionary Biology, 10*(1). doi:10.1186/1471-2148-10-305

Bouetard, A., Noirot, C., Besnard, A. L., Bouchez, O., Choisne, D., Robe, E., . . . Coutellec, M. A. (2012). Pyrosequencing-based transcriptomic resources in the pond snail Lymnaea stagnalis, with a focus on genes involved in molecular response to diquat-induced stress. *Ecotoxicology, 21*(8), 2222-2234. doi:10.1007/s10646-012-0977-1

Bouquet, J. M., Spriet, E., Troedsson, C., Otter, H., Chourrout, D., & Thompson, E. M. (2009). Culture optimization for the emergent zooplanktonic model organism Oikopleura dioica. *Journal of Plankton Research, 31*(4), 359-370. doi:10.1093/plankt/fbn132

Brönmark, C., Pettersson, L. B., & Nilsson, P. A. (1999). The ecology and evolution of inducible defenses. 203-217.

Brophy, P. D., Alasti, F., Darbro, B. W., Clarke, J., Nishimura, C., Cobb, B., . . . Manak, J. R. (2013). Genome-wide copy number variation analysis of a Branchio-oto-renal syndrome cohort identifies a recombination hotspot and implicates new candidate genes. *Human Genetics, 132*(12), 1339-1350. doi:10.1007/s00439-013-1338-8

Brophy, P. D., Ostrom, L., Lang, K. M., & Dressler, G. R. (2001). Regulation of ureteric bud outgrowth by Pax2-dependent activation of the glial derived neurotrophic factor gene. *Development, 128*(23), 4747-4756.

Brophy, P. D., Rasmussen, M., Parida, M., Bonde, G., Darbro, B. W., Hong, X., . . . Manak, J. R. (2017). A Gene Implicated in Activation of Retinoic Acid Receptor Targets Is a Novel Renal Agenesis Gene in Humans. *Genetics, 207*(1), 215-228. doi:10.1534/genetics.117.1125

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res, 33*(Database issue), D212-215. doi:10.1093/nar/gki034

Buffle, J., Chalmers, R. A., Masson, M. R., & Midgley, D. (1991). *Complexation reactions in aquatic systems : an analytical approach*. Chichester, West Sussex, England; New York: E. Horwood ; Halsted Press [distributor].

Burns, T. F., Fei, P., Scata, K. A., Dicker, D. T., & El-Deiry, W. S. (2003). Silencing of the novel p53 target gene Snk/Plk2 leads to mitotic catastrophe in paclitaxel (taxol)-exposed cells. *Mol Cell Biol, 23*(16), 5556-5571.

Campsteijn, C., Øvrebø, J. I., Karlsen, B. O., & Thompson, E. M. (2012). Expansion of cyclin D and CDK1 paralogs in Oikopleura dioica, a chordate employing diverse cell cycle variants. *Molecular Biology and Evolution, 29*(2), 487-502. doi:10.1093/molbev/msr136

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., . . . Web Presence Working, G. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics, 25*(2), 288-289. doi:10.1093/bioinformatics/btn615

Carroll, T. J., & Vize, P. D. (1999). Synergism between Pax-8 and lim-1 in embryonic kidney development. *Developmental Biology, 214*(1), 46-59. doi:10.1006/dbio.1999.9414

Carter, C. O., Evans, K., & Pescia, G. (1979). A family study of renal agenesis. *Journal of Medical Genetics, 16*(3), 176-188.

Cartry, J., Nichane, M., Ribes, V., Colas, A., Riou, J. F., Pieler, T., . . . Umbhauer, M. (2006). Retinoic acid signalling is required for specification of pronephric cell fate. *Developmental Biology, 299*(1), 35-51. doi:10.1016/j.ydbio.2006.06.047

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., . . . Karp, P. D. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research, 44*(D1), D471-D480. doi:10.1093/nar/gkv1164

Cempel, M., & Nikel, G. (2006). Nickel: A review of its sources and environmental toxicology. *Polish Journal of Environmental Studies*, *15*(3), 375-382.

Cervera, A., Maymo, A. C., Martinez-Pardo, R., & Garcera, M. D. (2006). Vitellogenin polypeptide levels in one susceptible and one cadmium-resistant strain of Oncopeltus fasciatus (Heteroptera: Lygaeidae), and its role in cadmium resistance. *J Insect Physiol, 52*(2), 158-168. doi:10.1016/j.jinsphys.2005.10.001

Chang, X., & Wang, K. (2012). Wannovar: Annotating genetic variants for personal genomes via the web. *Journal of Medical Genetics, 49*(7), 433-436. doi:10.1136/jmedgenet-2012-100918

Chao, S. H., Suzuki, Y., Zysk, J. R., & Cheung, W. Y. (1984). Activation of calmodulin by various metal cations as a function of ionic radius. *Mol Pharmacol, 26*(1), 75-82.

Chavali, S., Morais, D. A., Gough, J., & Babu, M. M. (2011). Evolution of eukaryotic genome architecture: Insights from the study of a rapidly evolving metazoan, Oikopleura dioica: Non-adaptive forces such as elevated mutation rates may influence the evolution of genome architecture. *BioEssays, 33*(8), 592-601. doi:10.1002/bies.201100034

Chen, C., & Wang, D. W. (2013). CYP epoxygenase derived EETs: from cardiovascular protection to human cancer therapy. *Curr Top Med Chem*, *13*(12), 1454-1469.

Chen, L., Liu, L., & Huang, S. (2008). Cadmium activates the mitogen-activated protein kinase (MAPK) pathway via induction of reactive oxygen species and inhibition of protein phosphatases 2A and 5. *Free Radic Biol Med, 45*(7), 1035-1044. doi:10.1016/j.freeradbiomed.2008.07.011

Cheng, C. N., & Wingert, R. A. (2015). Nephron proximal tubule patterning and corpuscles of Stannius formation are regulated by the sim1a transcription factor and retinoic acid in zebrafish. *Developmental Biology, 399*(1), 100-116. doi:10.1016/j.ydbio.2014.12.020

Chial, H. (2008). Rare Genetic Disorders: Learning About Genetic Disease Through Gene Mapping, SNPs, and Microarray Data *Scitable* (Vol. 1, pp. 192): Nature Education.

Chin, K. V., Pastan, I., & Gottesman, M. M. (1993). Function and regulation of the human multidrug resistance gene. *Adv Cancer Res, 60*, 157-180.

Choong, G., Liu, Y., & Templeton, D. M. (2014). Interplay of calcium and cadmium in mediating cadmium toxicity. *Chem Biol Interact, 211*, 54-65. doi:10.1016/j.cbi.2014.01.007

Choudhuri, S. (2014). Chapter 3 - Genomic Technologies* *Bioinformatics for Beginners* (pp. 55-72). Oxford: Academic Press.

Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly, 6*(2), 80-92. doi:10.4161/fly.19695

Clarke, J. C., Patel, S. R., Raymond Jr, R. M., Andrew, S., Robinson, B. G., Dressler, G. R., & Brophy, P. D. (2006). Regulation of c-Ret in the developing kidney is responsive to Pax2 gene dosage. *Human Molecular Genetics, 15*(23), 3420-3428. doi:10.1093/hmg/ddl418

Cliften, P. (2015). Chapter 7 - Base Calling, Read Mapping, and Coverage Analysis A2 - Kulkarni, Shashikant. In J. Pfeifer (Ed.), *Clinical Genomics* (pp. 91-107). Boston: Academic Press.

Colbourne, J. K., Crease, T. J., Weider, L. J., Hebert, P. D. N., Dufresne, F., & Hobæk, A. (1998). Phylogenetics and evolution of a circumarctic species complex (Cladocera: Daphnia pulex). *Biological Journal of the Linnean Society, 65*(3), 347-365. doi:10.1006/bijl.1998.0251

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., . . . Boore, J. L. (2011). The ecoresponsive genome of Daphnia pulex. *Science, 331*(6017), 555-561. doi:10.1126/science.1197761

Colbourne, J. K., Singan, V. R., & Gilbert, D. G. (2005). wFleaBase: the Daphnia genome database. *BMC Bioinformatics, 6*(1), 45. doi:10.1186/1471-2105-6-45

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., . . . Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol, 17*, 13. doi:10.1186/s13059-016-0881-8

Consortium, T. E. (2011). Standards, Guidelines and Best Practices for RNA-Seq. Retrieved from https://genome.ucsc.edu/encode/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

Conzone, S. D., & Pantano, C. G. (2004). Glass slides to DNA microarrays. *Materials Today, 7*(3), 20-26. doi:https://doi.org/10.1016/S1369-7021(04)00122-1

Corà, D., Di Cunto, F., Provero, P., Silengo, L., & Caselle, M. (2004). Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrep-resented upstream motifs. *BMC Bioinformatics, 5*(1), 57. doi:10.1186/1471-2105-5-57

Costantini, F. (2010). GDNF/Ret signaling and renal branching morphogenesis: From mesenchymal signals to epithelial cell behaviors. *Organogenesis, 6*(4), 252-262. doi:10.4161/org.6.4.12680

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*(5258), 561-563.

Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., . . . Weissenbach, J. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics, 25*(2), 235-240. doi:10.1038/76118

Cupillard, L., Koumanov, K., Mattei, M. G., Lazdunski, M., & Lambeau, G. (1997). Cloning, chromosomal mapping, and expression of a novel human secretory phospholipase A2. *J Biol Chem, 272*(25), 15745-15752.

Danielson, U. H., Esterbauer, H., & Mannervik, B. (1987). Structure-activity relationships of 4-hydroxyalkenals in the conjugation catalysed by mammalian glutathione transferases. *Biochem J, 247*(3), 707-713.

Danks, G., Campsteijn, C., Parida, M., Butcher, S., Doddapaneni, H., Fu, B., . . . Manak, J. R. (2013). OikoBase: a genomics and developmental transcriptomics resource for the urochordate Oikopleura dioica. *Nucleic Acids Res, 41*(Database issue), D845-853. doi:10.1093/nar/gks1159

Danks, G., & Parida, M. (2013). OikoBase. Retrieved from http://oikoarrays.biology.uiowa.edu/Oiko/

Daraghmeh, N. H., Chowdhry, B. Z., Leharne, S. A., Al Omari, M. M., & Badwan, A. A. (2011). Chapter 2 - Chitin. In H. G. Brittain (Ed.), *Profiles of Drug Substances, Excipients and Related Methodology* (Vol. 36, pp. 35-102): Academic Press.

Davies, J. E. (2007). The pharmacological basis of therapeutics. *Occupational and Environmental Medicine, 64*(8), e2-e2. doi:10.1136/oem.2007.033902

De Coen, W. M., & Janssen, C. R. (1998). The use of biomarkers in Daphnia magna toxicity testing - I. The digestive physiology of daphnids exposed to toxic stress. *Hydrobiologia, 367*, 199-209. doi:Doi 10.1023/A:1003240502946

De Coninck, D. I. M., Asselman, J., Glaholt, S., Janssen, C. R., Colbourne, J. K., Shaw, J. R., & De Schamphelaere, K. A. C. (2014). Genome-wide transcription profiles reveal genotype-dependent responses of biological pathways and gene-families in daphnia exposed to single and mixed stressors. *Environmental Science and Technology, 48*(6), 3513-3522. doi:10.1021/es4053363

De Schamphelaere, K. A. C., Vandenbrouck, T., Muyssen, B. T. A., Soetaert, A., Blust, R., De Coen, W., & Janssen, C. R. (2008). Integration of molecular with higher-level effects of dietary zinc exposure in Daphnia magna. *Comparative Biochemistry and Physiology D-Genomics & Proteomics, 3*(4), 307-314. doi:10.1016/j.cbd.2008.09.001

De Tomasi, L., David, P., Humbert, C., Silbermann, F., Arrondel, C., Tores, F., . . . Jeanpierre, C. (2017). Mutations in GREB1L Cause Bilateral Kidney Agenesis in Humans and Mice. *Am J Hum Genet, 101*(5), 803-814. doi:10.1016/j.ajhg.2017.09.026

Dekker, L. V., Leitges, M., Altschuler, G., Mistry, N., McDermott, A., Roes, J., & Segal, A. W. (2000). Protein kinase C-beta contributes to NADPH oxidase activation in neutrophils. *Biochem J, 347 Pt 1*, 285-289.

Delalande, O., Desvaux, H., Godat, E., Valleix, A., Junot, C., Labarre, J., & Boulard, Y. (2010). Cadmium-glutathione solution structures provide new insights into heavy metal detoxification. *FEBS J, 277*(24), 5086-5096. doi:10.1111/j.1742-4658.2010.07913.x

Delsuc, F., Brinkmann, H., Chourrout, D., & Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature, 439*(7079), 965-968. doi:10.1038/nature04336

Deneke, S. M., & Fanburg, B. L. (1989). Regulation of cellular glutathione. *Am J Physiol, 257*(4 Pt 1), L163-173. doi:10.1152/ajplung.1989.257.4.L163

Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J. M., Da Silva, C., Brinkmann, H., . . . Chourrout, D. (2010). Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science, 330*(6009), 1381-1385. doi:10.1126/science.1194167

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., . . . Murakami, A. (2016). High-throughput discovery of novel developmental phenotypes. *Nature, 537*(7621), 508-514. doi:10.1038/nature19356

Diep, C. Q., Peng, Z., Ukah, T. K., Kelly, P. M., Daigle, R. V., & Davidson, A. J. (2015). Development of the zebrafish mesonephros. *Genesis, 53*(3-4), 257-269. doi:10.1002/dvg.22846

Dietrich, C. (2016). Antioxidant Functions of the Aryl Hydrocarbon Receptor. *Stem Cells Int, 2016*, 7943495. doi:10.1155/2016/7943495

Dingle, J. T., & Dean, R. T. (1976). *Lysosomes in biology and pathology. 5 5*. Amsterdam; Oxford; New York: North-Holland publ. comp. ; Elsevier.

Dodson, S. (1988). The ecological role of chemical stimuli for the zooplankton: Predator‐avoidance behavior in Daphnia. *Limnology and Oceanography, 33*(6part2), 1431-1439. doi:10.4319/lo.1988.33.6part2.1431

Dodson, S. I. (1974). Adaptive change in plankton morphology in response to size‐selective predation: A new hypothesis of cyclomorphosis. *Limnology and Oceanography, 19*(5), 721-729. doi:10.4319/lo.1974.19.5.0721

Dodson, S. I. (1989). The ecological role of chemical stimuli for the zooplankton: predator-induced morphology in Daphnia. *Oecologia, 78*(3), 361-367. doi:10.1007/BF00379110

Dodson, S. I. (1996). Optimal swimming behavior of zooplankton. *Zooplankton: Sensory Ecology and Physiology*, 365-374.

Dodson, S. I., Hanazato, T., & Gorski, P. R. (1995). Behavioral responses of Daphnia pulex exposed to carbaryl and Chaoborus kairomone. *Environmental Toxicology and Chemistry, 14*(1), 43-50. doi:10.1002/etc.5620140106

Dolniak, B., Katsoulidis, E., Carayol, N., Altman, J. K., Redig, A. J., Tallman, M. S., . . . Platanias, L. C. (2008). Regulation of arsenic trioxide-induced cellular responses by Mnk1 and Mnk2. *J Biol Chem, 283*(18), 12034-12042. doi:10.1074/jbc.M708816200

Donovan, M. J., Natoli, T. A., Sainio, K., Amstutz, A., Jaenisch, R., Sariola, H., & Kreidberg, J. A. (1999). Initial differentiation of the metanephric mesenchyme is independent of WT1 and the ureteric bud. *Developmental Genetics, 24*(3-4), 252-262. doi:10.1002/(SICI)1520-6408(1999)24:3/4&lt;252::AID-DVG8&gt;3.0.CO;2-K

Drummond, I. A. (2005). Kidney development and disease in the zebrafish. *Journal of the American Society of Nephrology, 16*(2), 299-304. doi:10.1681/ASN.2004090754

Drummond, I. A., & Davidson, A. J. (2016a) Zebrafish kidney development. *Vol. 134. Methods in Cell Biology* (pp. 391-429).

Drummond, I. A., & Davidson, A. J. (2016b). Zebrafish kidney development. *Methods Cell Biol, 134*, 391-429. doi:10.1016/bs.mcb.2016.03.041

Drummond, I. A., Majumdar, A., Hentschel, H., Elger, M., Solnica-Krezel, L., Schier, A. F., . . . Fishman, M. C. (1998). Early development of the zebrafish pronephros and analysis of mutations affecting pronephric function. *Development, 125*(23), 4655-4667.

E Fergusson, J. (1991). *The Heavy Elements—Chemistry, environmental impact and health effects* (Vol. 69).

Education, P. (2018). Concept 3: Different Genes for Different RNAs. Retrieved from http://www.phschool.com/science/biology_place/biocoach/transcription/difgns.html

Eggen, R. I., Behra, R., Burkhardt-Holm, P., Escher, B. I., & Schweigert, N. (2004). Challenges in ecotoxicology. *Environ Sci Technol, 38*(3), 58A-64A.

Eisler, R., Fish, U. S., Wildlife, S., Fish, U. S., Wildlife, Fish, U. S., . . . Development. (1985). Cadmium hazards to fish, wildlife, and invertebrates : a synoptic review.

Elabdeen, H. R., Mustafa, M., Szklenar, M., Ruhl, R., Ali, R., & Bolstad, A. I. (2013). Ratio of pro-resolving and pro-inflammatory lipid mediator precursors as potential markers for aggressive periodontitis. *PLoS ONE, 8*(8), e70838. doi:10.1371/journal.pone.0070838

Eleawa, S. M., Alkhateeb, M. A., Alhashem, F. H., Bin-Jaliah, I., Sakr, H. F., Elrefaey, H. M., . . . Khalil, M. A. (2014). Resveratrol reverses cadmium chloride-induced testicular damage and subfertility by downregulating p53 and Bax and upregulating gonadotropins and Bcl-2 gene expression. *J Reprod Dev, 60*(2), 115-127.

Ellis, P. D., Strang, P., & Potter, J. D. (1984). Cadmium-substituted skeletal troponin C. Cadmium-113 NMR spectroscopy and metal binding investigations. *J Biol Chem, 259*(16), 10348-10356.

Ercal, N., Gurer-Orhan, H., & Aykin-Burns, N. (2001). Toxic metals and oxidative stress part I: mechanisms involved in metal-induced oxidative damage. *Curr Top Med Chem, 1*(6), 529-539.

Erickson, J. R., He, B. J., Grumbach, I. M., & Anderson, M. E. (2011). CaMKII in the cardiovascular system: sensing redox states. *Physiol Rev, 91*(3), 889-915. doi:10.1152/physrev.00018.2010

Esworthy, R. S., Doan, K., Doroshow, J. H., & Chu, F. F. (1994). Cloning and sequencing of the cDNA encoding a human testis phospholipid hydroperoxide glutathione peroxidase. *Gene, 144*(2), 317-318.

Eun, H.-M. (1996). 1 - Enzymes and Nucleic Acids: General Principles *Enzymology Primer for Recombinant DNA Technology* (pp. 1-108). San Diego: Academic Press.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., . . . D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Research, 44*(D1), D481-D487. doi:10.1093/nar/gkv1351

Fairfield, H., Gilbert, G. J., Barter, M., Corrigan, R. R., Curtain, M., Ding, Y., . . . Reinholdt, L. G. (2011). Mutation discovery in mice by whole exome sequencing. *Genome Biology, 12*(9), R86. doi:10.1186/gb-2011-12-9-r86

Falkenburger, B. H., Jensen, J. B., Dickson, E. J., Suh, B. C., & Hille, B. (2010). Phosphoinositides: lipid regulators of membrane proteins. *J Physiol, 588*(Pt 17), 3179-3185. doi:10.1113/jphysiol.2010.192153

Fenaux, R., Bone, Q., & Deibel, D. (1998). Appendicularian distribution and zoogeography. *The Biology of Pelagic Tunicates*, 251-264.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., . . . Mitchell, A. L. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research, 45*(Database issue), D190-D199. doi:10.1093/nar/gkw1107

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., . . . Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res, 44*(D1), D279-285. doi:10.1093/nar/gkv1344

Flohe, L., & Ursini, F. (2008). Peroxidase: a term of many meanings. *Antioxid Redox Signal, 10*(9), 1485-1490. doi:10.1089/ars.2008.2059

Fujita, T., Matsushita, M., & Endo, Y. (2004). The lectin-complement pathway - Its role in innate immunity and evolution. *Immunological Reviews, 198*, 185-202. doi:10.1111/j.0105-2896.2004.0123.x

Gamage, N., Barnett, A., Hempel, N., Duggleby, R. G., Windmill, K. F., Martin, J. L., & McManus, M. E. (2006). Human sulfotransferases and their role in chemical metabolism. *Toxicol Sci, 90*(1), 5-22. doi:10.1093/toxsci/kfj061

Ganot, P., Kallesøe, T., & Thompson, E. M. (2007). The cytoskeleton organizes germ nuclei with divergent fates and asynchronous cycles in a common cytoplasm during oogenesis in the chordate Oikopleura. *Developmental Biology, 302*(2), 577-590. doi:10.1016/j.ydbio.2006.10.022

Ganot, P., & Thompson, E. M. (2002). Patterning through differential endoreduplication in epithelial organogenesis of the chordate, Oikopleura dioica. *Developmental Biology, 252*(1), 59-71. doi:10.1006/dbio.2002.0834

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., . . . Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics, 28*(20), 2678-2679. doi:10.1093/bioinformatics/bts503

Garrett, S. H., Somji, S., Sens, M. A., Zhang, K., & Sens, D. A. (2011). Microarray Analysis of Gene Expression Patterns in Human Proximal Tubule Cells Over a Short and Long Time Course of Cadmium Exposure. *Journal of Toxicology and Environmental Health-Part a-Current Issues, 74*(1), 24-42. doi:Pii 930339229 10.1080/15287394.2010.514230

GATK_BP. GATK best practices. Retrieved from https://software.broadinstitute.org/gatk/best-practices/

Gavazzo, P., Morelli, E., & Marchetti, C. (2005). Susceptibility of insulinoma cells to cadmium and modulation by L-type calcium channels. *Biometals, 18*(2), 131-142.

Gehrmann, T., Gulkan, H., Suer, S., Herberg, F. W., Balla, A., Vereb, G., . . . Heilmeyer, L. M., Jr. (1999). Functional expression and characterisation of a new human phosphatidylinositol 4-kinase PI4K230. *Biochim Biophys Acta, 1437*(3), 341-356.

GenBank. (2018). GenBank and WGS Statistics. Retrieved from https://www.ncbi.nlm.nih.gov/genbank/statistics/

Genetic, A., District of, C., & Department of, H. (2010). Understanding genetics : a District of Columbia guide for patients and health professionals.

Genome Informatics Lab, I. U. B. D. (2005). wFleaBase. Retrieved from http://wfleabase.org

Georgas, K., Rumballe, B., Valerius, M. T., Chiu, H. S., Thiagarajan, R. D., Lesieur, E., . . . Little, M. H. (2009). Analysis of early nephron patterning reveals a role for distal RV proliferation in fusion to the ureteric tip via a cap mesenchyme-derived connecting segment. *Developmental Biology, 332*(2), 273-286. doi:10.1016/j.ydbio.2009.05.578

Gerber, S., Alzayady, K. J., Burglen, L., Bremond-Gignac, D., Marchesin, V., Roche, O., . . . Fares Taie, L. (2016). Recessive and Dominant De Novo ITPR1 Mutations Cause Gillespie Syndrome. *Am J Hum Genet, 98*(5), 971-980. doi:10.1016/j.ajhg.2016.03.004

Ghosh, M. G., Thompson, D. A., & Weigel, R. J. (2000). PDZK1 and GREB1 are estrogen-regulated genes expressed in hormone-responsive breast cancer. *Cancer Research, 60*(22), 6367-6375.

GHR. (2018). What are proteins and what do they do? Retrieved from https://ghr.nlm.nih.gov/primer/howgeneswork/protein

Glaumann, H., & Ballard, F. J. (1987). *Lysosomes : their role in protein breakdown*. London; Orlando: Academic Press.

Glavinas, H., Krajcsi, P., Cserepes, J., & Sarkadi, B. (2004). The role of ABC transporters in drug resistance, metabolism and toxicity. *Curr Drug Deliv, 1*(1), 27-42.

Godt, J., Scheidig, F., Grosse-Siestrup, C., Esche, V., Brandenburg, P., Reich, A., & Groneberg, D. A. (2006). The toxicity of cadmium and resulting hazards for human health. *J Occup Med Toxicol, 1*, 22. doi:10.1186/1745-6673-1-22

González-Pérez, A., & López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics, 88*(4), 440-449. doi:10.1016/j.ajhg.2011.03.004

Gorsky, G., & Fenaux, R. (1998). The role of Appendicularia in marine food webs. *The Biology of Pelagic Tunicates*, 161-169.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., . . . Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research, 36*(10), 3420-3435. doi:10.1093/nar/gkn176

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol, 29*(7), 644-652. doi:10.1038/nbt.1883

Graeber_lab. (2009). hypergeometric.   Retrieved from http://systems.crump.ucla.edu/hypergeometric/index.php

Gramates, L. S., Marygold, S. J., dos Santos, G., Urbano, J.-M., Antonazzo, G., Matthews, B. B., . . . the FlyBase Consortium. (2017). FlyBase at 25: looking to the future. *Nucleic Acids Research, 45*(Database issue), D663-D671. doi:10.1093/nar/gkw1016

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics, 27*(7), 1017-1018. doi:10.1093/bioinformatics/btr064

Griffiths, P. R. E. (1980). Morphological and ultrastructural effects of sublethal cadmium poisoning on Daphnia. *Environmental Research, 22*(2), 277-284. doi:https://doi.org/10.1016/0013-9351(80)90140-1

Grzywacz, A., Gdula-Argasinska, J., Muszynska, B., Tyszka-Czochara, M., Librowski, T., & Opoka, W. (2015). Metal responsive transcription factor 1 (MTF-1) regulates zinc dependent cellular processes at the molecular level. *Acta Biochim Pol, 62*(3), 491-498. doi:10.18388/abp.2015_1038

Guan, R., & Wang, W. X. (2004). Cd and Zn uptake kinetics in Daphnia magna in relation to Cd exposure history. *Environ Sci Technol, 38*(22), 6051-6058.

Guimier, A., Gabriel, G. C., Bajolle, F., Tsang, M., Liu, H., Noll, A., . . . Gordon, C. T. (2015). MMP21 is mutated in human heterotaxy and is required for normal left-right asymmetry in vertebrates. *Nature Genetics, 47*(11), 1260-1263. doi:10.1038/ng.3376

Gunatilaka, A., Diehl, P., & Puzicha, H. (2001). The Evaluation of "Dynamic Daphnia Test" after a Decade of Use: Benefits and Constraints. *Environmental science research., 56*, 29-58.

Gunawardena, D., Govindaraghavan, S., & Münch, G. (2014). Chapter 30 - Anti-Inflammatory Properties of Cinnamon Polyphenols and their Monomeric Precursors *Polyphenols in Human Health and Disease* (pp. 409-425). San Diego: Academic Press.

Gunther, V., Lindert, U., & Schaffner, W. (2012). The taste of heavy metals: gene regulation by MTF-1. *Biochim Biophys Acta, 1823*(9), 1416-1425. doi:10.1016/j.bbamcr.2012.01.005

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology, 8*(2). doi:ARTN R24 10.1186/gb-2007-8-2-r24

Ha, M. H., & Choi, J. (2008). Effects of environmental contaminants on hemoglobin of larvae of aquatic midge, Chironomus riparius (Diptera: Chironomidae): a potential biomarker for ecotoxicity monitoring. *Chemosphere, 71*(10), 1928-1936. doi:10.1016/j.chemosphere.2008.01.018

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res, 41*(Database issue), D387-395. doi:10.1093/nar/gks1234

Hagberg, A. A., Schult, D.A., and Swart, P.J. (2008). *Exploring Network Structure, Dynamics, and Function using NetworkX*. https://networkx.github.io/documentation/networkx-1.9/index.html

Hague, F., Matifat, F., Louvet, L., Brûlé, G., & Collin, T. (2000). *The carcinogen Cd2+ activates InsP3-mediated Ca2+ release through a specific metal ions receptor in Xenopus oocyte* (Vol. 12).

Hailemariam, K., Iwasaki, K., Huang, B. W., Sakamoto, K., & Tsuji, Y. (2010). Transcriptional regulation of ferritin and antioxidant genes by HIPK2 under genotoxic stress. *J Cell Sci, 123*(Pt 22), 3863-3871. doi:10.1242/jcs.073627

Hales, N. R., Schield, D. R., Andrew, A. L., Card, D. C., Walsh, M. R., & Castoe, T. A. (2017). Contrasting gene expression programs correspond with predator-induced phenotypic plasticity within and across generations in Daphnia. *Molecular Ecology, 26*(19), 5003-5015. doi:10.1111/mec.14213

Hanington, P. C., & Zhang, S. M. (2011). The primary role of fibrinogen-related proteins in invertebrates is defense, not coagulation. *Journal of Innate Immunity, 3*(1), 17-27. doi:10.1159/000321882

Harada, B. T., Knight, M. J., Imai, S., Qiao, F., Ramachander, R., Sawaya, M. R., . . . Bowie, J. U. (2008). Regulation of enzyme localization by polymerization: polymer formation by the SAM domain of diacylglycerol kinase delta1. *Structure, 16*(3), 380-387. doi:10.1016/j.str.2007.12.017

Harding, S. D., Armit, C., Armstrong, J., Brennan, J., Cheng, Y., Haggarty, B., . . . Davidson, D. (2011). The GUDMAP database--an online resource for genitourinary research. *Development, 138*(13), 2845-2853. doi:10.1242/dev.063594

Harding, S. D., Armit, C., Armstrong, J., Brennan, J., Cheng, Y., Haggarty, B., . . . Davidson, D. (2011). The GUDMAP database - an online resource for genitourinary research. *Development, 138*(13), 2845-2853. doi:10.1242/dev.063594

Harewood, L., Liu, M., Keeling, J., Howatson, A., Whiteford, M., Branney, P., . . . FitzPatrick, D. R. (2010). Bilateral renal agenesis/Hypoplasia/Dysplasia (BRAHD): Postmortem analysis of 45 cases with breakpoint mapping of two De novo translocations. *PLoS ONE, 5*(8). doi:10.1371/journal.pone.0012375

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., . . . Sternberg, P. W. (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research, 38*(Database issue), D463-D467. doi:10.1093/nar/gkp952

Hebert, P. D. N. (1978). The adaptive significance of cyclomorphosis in Daphnia: more possibilities. *Freshwater Biology, 8*(4), 313-320. doi:10.1111/j.1365-2427.1978.tb01452.x

Henriet, S., Sumic, S., Doufoundou-Guilengui, C., Jensen, M. F., Grandmougin, C., Fal, K., . . . Chourrout, D. (2015). Embryonic expression of endogenous retroviral RNAs in somatic tissues adjacent to the Oikopleura germline. *Nucleic Acids Res, 43*(7), 3701-3711. doi:10.1093/nar/gkv169

Hibi, T., Nii, H., Nakatsu, T., Kimura, A., Kato, H., Hiratake, J., & Oda, J. (2004). Crystal structure of gamma-glutamylcysteine synthetase: insights into the mechanism of catalysis by a key enzyme for glutathione homeostasis. *Proc Natl Acad Sci U S A, 101*(42), 15052-15057. doi:10.1073/pnas.0403277101

Hildner, K., Edelson, B. T., Purtha, W. E., Diamond, M., Matsushita, H., Kohyama, M., . . . Murphy, K. M. (2008). Batf3 deficiency reveals a critical role for CD8α+dendritic cells in cytotoxic T cell immunity. *Science, 322*(5904), 1097-1100. doi:10.1126/science.1164206

Hilton, J. M., Lewis, M. A., Grati, M. h., Ingham, N., Pearson, S., Laskowski, R. A., . . . Steel, K. P. (2011). Exome sequencing identifies a missense mutation in Isl1associated with low penetrance otitis media in dearisch mice. *Genome Biology, 12*(9), R90. doi:10.1186/gb-2011-12-9-r90

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Comput Biol, 7*(3), e1002021. doi:10.1371/journal.pcbi.1002021

Horgan, R. P., & Kenny, L. C. (2011). ‘Omic’ technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist, 13*(3), 189-195. doi:10.1576/toag.13.3.189.27672

Hosp, J., Sagane, Y., Danks, G., & Thompson, E. M. (2012). The evolving proteome of a complex extracellular matrix, the Oikopleura house. *PLoS ONE, 7*(7). doi:10.1371/journal.pone.0040172

Howe, K. L., Chothia, T., & Durbin, R. (2002). GAZE: A genetic framework for the integration of gene-prediction data by dynamic programming. *Genome Research, 12*(9), 1418-1427. doi:10.1101/gr.149502

Hsiao, T. H., Lin, C. J., Chung, Y. S., Lee, G. H., Kao, T. T., Chang, W. N., . . . Fu, T. F. (2014). Ethanol-induced upregulation of 10-formyltetrahydrofolate dehydrogenase helps relieve ethanol-induced oxidative stress. *Mol Cell Biol, 34*(3), 498-509. doi:10.1128/MCB.01427-13

Huang, N., Lee, I., Marcotte, E. M., & Hurles, M. E. (2010). Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics, 6*(10), 1-11. doi:10.1371/journal.pgen.1001154

Huang, S., Shu, L., Easton, J., Harwood, F. C., Germain, G. S., Ichijo, H., & Houghton, P. J. (2004). Inhibition of mammalian target of rapamycin activates apoptosis signal-regulating kinase 1 signaling by suppressing protein phosphatase 5 activity. *J Biol Chem, 279*(35), 36490-36496. doi:10.1074/jbc.M401208200

Hubatsch, I., Ridderstrom, M., & Mannervik, B. (1998). Human glutathione transferase A4-4: an alpha class enzyme with high catalytic efficiency in the conjugation of 4-hydroxynonenal and other genotoxic products of lipid peroxidation. *Biochem J, 330 (Pt 1)*, 175-179.

Huggins, C. J., Mayekar, M. K., Martin, N., Saylor, K. L., Gonit, M., Jailwala, P., . . . Johnson, P. F. (2015). C/EBPgamma Is a Critical Regulator of Cellular Stress Response Networks through Heterodimerization with ATF4. *Mol Cell Biol, 36*(5), 693-713. doi:10.1128/MCB.00911-15

Hülsmann, S., Vijverberg, J., Boersma, M., & Mooij, W. M. (2004). Effects of infochemicals released by gape-limited fish on life history traits of Daphnia: A maladaptive response? *Journal of Plankton Research, 26*(5), 535-543. doi:10.1093/plankt/fbh054

Humbert, C., Silbermann, F., Morar, B., Parisot, M., Zarhrate, M., Masson, C., . . . Jeanpierre, C. (2014). Integrin alpha 8 recessive mutations are responsible for bilateral renal agenesis in humans. *American Journal of Human Genetics, 94*(2), 288-294. doi:10.1016/j.ajhg.2013.12.017

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., . . . Yong, S. Y. (2012). InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Research, 40*(D1), D306-D312. doi:10.1093/nar/gkr948

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., . . . Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics, 36*(9), 949-951. doi:10.1038/ng1416

Itoh, K., Chiba, T., Takahashi, S., Ishii, T., Igarashi, K., Katoh, Y., . . . Nabeshima, Y. (1997). An Nrf2/small Maf heterodimer mediates the induction of phase II detoxifying enzyme genes through antioxidant response elements. *Biochem Biophys Res Commun, 236*(2), 313-322.

Iwasaki, K., Hailemariam, K., & Tsuji, Y. (2007). PIAS3 interacts with ATF1 and regulates the human ferritin H gene through an antioxidant-responsive element. *J Biol Chem, 282*(31), 22335-22343. doi:10.1074/jbc.M701477200

Iyer, L. M., Zhang, D., Burroughs, A. M., & Aravind, L. (2013). Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res, 41*(16), 7635-7655. doi:10.1093/nar/gkt573

Jacob, S. T., Majumder, S., & Ghoshal, K. (2002). Suppression of metallothionein-I/II expression and its probable molecular mechanisms. *Environ Health Perspect, 110 Suppl 5*, 827-830.

James Kent, W., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research, 12*(6), 996-1006. doi:10.1101/gr.229102. Article published online before print in May 2002

Janet, I., Szostak, J., & Bell, T. (2008). What is RNA?   Retrieved from http://exploringorigins.org/rna.html

Jansen, M., Vergauwen, L., Vandenbrouck, T., Knapen, D., Dom, N., Spanier, K. I., . . . De Meester, L. (2013). Gene expression profiling of three different stressors in the water flea Daphnia magna. *Ecotoxicology, 22*(5), 900-914. doi:10.1007/s10646-013-1072-y

Jao, L. E., Wente, S. R., & Chen, W. (2013). Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proceedings of the National Academy of Sciences of the United States of America, 110*(34), 13904-13909. doi:10.1073/pnas.1308335110

Jasrapuria, S., Arakane, Y., Osman, G., Kramer, K. J., Beeman, R. W., & Muthukrishnan, S. (2010). Genes encoding proteins with peritrophin A-type chitin-binding domains in Tribolium castaneum are grouped into three distinct families based on phylogeny, expression and function. *Insect Biochem Mol Biol, 40*(3), 214-227. doi:10.1016/j.ibmb.2010.01.011

Jauhiainen, A., Thomsen, C., Strombom, L., Grundevik, P., Andersson, C., Danielsson, A., . . . Aman, P. (2012). Distinct cytoplasmic and nuclear functions of the stress induced protein DDIT3/CHOP/GADD153. *PLoS ONE, 7*(4), e33208. doi:10.1371/journal.pone.0033208

Jayaram, N., Usvyat, D., & AC, R. M. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*. doi:10.1186/s12859-016-1298-9

Jenkins, D., Bitner-Glindzicz, M., Malcolm, S., Hu, C. C. A., Allison, J., Winyard, P. J. D., . . . Woolf, A. S. (2005). De novo Uroplakin IIIa heterozygous mutations cause human renal adysplasia leading to severe kidney failure. *Journal of the American Society of Nephrology, 16*(7), 2141-2149. doi:10.1681/ASN.2004090776

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240. doi:10.1093/bioinformatics/btu031

Jordan, K. W., Nordenstam, J., Lauwers, G. Y., Rothenberger, D. A., Alavi, K., Garwood, M., & Cheng, L. L. (2009). Metabolomic characterization of human rectal adenocarcinoma with intact tissue magnetic resonance spectroscopy. *Dis Colon Rectum, 52*(3), 520-525. doi:10.1007/DCR.0b013e31819c9a2c

Jornayvaz, F. R., & Shulman, G. I. (2012). Diacylglycerol activation of protein kinase Cepsilon and hepatic insulin resistance. *Cell Metab, 15*(5), 574-584. doi:10.1016/j.cmet.2012.03.005

Jozefczak, M., Remans, T., Vangronsveld, J., & Cuypers, A. (2012). Glutathione is a key player in metal-induced oxidative stress defenses. *International Journal of Molecular Sciences, 13*(3), 3145-3175. doi:10.3390/ijms13033145

Juračka, P. J., Laforsch, C., & Petrusek, A. (2011). Neckteeth formation in two species of the Daphnia curvirostris complex (Crustacea: Cladocera). *Journal of Limnology, 70*(2), 359-368. doi:10.3274/JL11-70-2-20

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. doi:10.1093/nar/gkv1070

Kato, Y., Tokishita, S. I., Ohta, T., & Yamagata, H. (2004). A vitellogenin chain containing a superoxide dismutase-like domain is the major component of yolk proteins in cladoceran crustacean Daphnia magna. *Gene, 334*(1-2), 157-165. doi:10.1016/j.gene.2004.03.030

Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research, 12*(4), 656-664. doi:10.1101/gr.229202. Article published online before March 2002

Kerecuk, L., Schreuder, M. F., & Woolf, A. S. (2008). Renal tract malformations: Perspectives for nephrologists. *Nature Clinical Practice Nephrology, 4*(6), 312-325. doi:10.1038/ncpneph0807

Kettleborough, R. N. W., Busch-Nentwich, E. M., Harvey, S. A., Dooley, C. M., De Bruijn, E., Van Eeden, F., . . . Stemple, D. L. (2013). A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature, 496*(7446), 494-497. doi:10.1038/nature11992

Kim, D., & Dressler, G. R. (2005). Nephrogenic factors promote differentiation of mouse embryonic stem cells into renal epithelia. *Journal of the American Society of Nephrology, 16*(12), 3527-3534. doi:10.1681/ASN.2005050544

Kim, H. J., Koedrith, P., & Seo, Y. R. (2015). Ecotoxicogenomic approaches for understanding molecular mechanisms of environmental chemical toxicity using aquatic invertebrate, Daphnia model organism. *International Journal of Molecular Sciences, 16*(6), 12261-12287. doi:10.3390/ijms160612261

Kim, J., Kim, Y., Lee, S., Kwak, K., Chung, W. J., & Choi, K. (2011). Determination of mRNA expression of DMRT93B, vitellogenin, and cuticle 12 in Daphnia magna and their biomarker potential for endocrine disruption. *Ecotoxicology, 20*(8), 1741-1748. doi:10.1007/s10646-011-0707-0

Kim, S. F., Huri, D. A., & Snyder, S. H. (2005). Inducible nitric oxide synthase binds, S-nitrosylates, and activates cyclooxygenase-2. *Science, 310*(5756), 1966-1970. doi:10.1126/science.1119407

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., & Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics, 203*(3), 253-310. doi:10.1002/aja.1002030302

Kinraide, T. B., & Yermiyahu, U. (2007). A scale of metal ion binding strengths correlating with ionic charge, Pauling electronegativity, toxicity, and other physiological effects. *J Inorg Biochem, 101*(9), 1201-1213. doi:10.1016/j.jinorgbio.2007.06.003

Kittler, R., Pelletier, L., Heninger, A. K., Slabicki, M., Theis, M., Miroslaw, L., . . . Buchholz, F. (2007). Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nature Cell Biology, 9*(12), 1401-1412. doi:10.1038/ncb1659

Klaassen, C. D., Liu, J., & Choudhuri, S. (1999). Metallothionein: an intracellular protein to protect against cadmium toxicity. *Annu Rev Pharmacol Toxicol, 39*, 267-294. doi:10.1146/annurev.pharmtox.39.1.267

Klotz, L. O., Sanchez-Ramos, C., Prieto-Arroyo, I., Urbanek, P., Steinbrenner, H., & Monsalve, M. (2015). Redox regulation of FoxO transcription factors. *Redox Biol, 6*, 51-72. doi:10.1016/j.redox.2015.06.019

Kohle, C., & Bock, K. W. (2009). Coordinate regulation of human drug-metabolizing enzymes, and conjugate transporters by the Ah receptor, pregnane X receptor and constitutive androstane receptor. *Biochem Pharmacol, 77*(4), 689-699. doi:10.1016/j.bcp.2008.05.020

Kohlhase, J., Wischermann, A., Reichenbach, H., Froster, U., & Engel, W. (1998). Mutations in the SALL1 putative transcription factor gene cause Townes- Brocks syndrome. *Nature Genetics, 18*(1), 81-83. doi:10.1038/ng0198-81

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics, 5*. doi:10.1186/1471-2105-5-59

Kramer-Zucker, A. G., Wiessner, S., Jensen, A. M., & Drummond, I. A. (2005). Organization of the pronephric filtration apparatus in zebrafish requires Nephrin, Podocin and the FERM domain protein Mosaic eyes. *Developmental Biology, 285*(2), 316-329. doi:10.1016/j.ydbio.2005.06.038

Kreidberg, J. A., Sariola, H., Loring, J. M., Maeda, M., Pelletier, J., Housman, D., & Jaenisch, R. (1993). WT-1 is required for early kidney development. *Cell, 74*(4), 679-691. doi:10.1016/0092-8674(93)90515-R

Kuhlmann, H. W., Kusch, J., & Heckmann, K. (1999). Predator-induced defenses in ciliated protozoa. *The Ecology and Evolution of Inducible Defenses*, 142-159.

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi:10.1038/nprot.2009.86

Laforsch, C., Ngwa, W., Grill, W., & Tollrian, R. (2004). An acoustic microscopy technique reveals hidden morphological defenses in Daphnia. *Proceedings of the National Academy of Sciences of the United States of America, 101*(45), 15911-15914. doi:10.1073/pnas.0404860101

Laforsch, C., & Tollrian, R. (2004). Embryological aspects of inducible morphological defenses in Daphnia. *Journal of Morphology, 262*(3), 701-707. doi:10.1002/jmor.10270

Lam, S. D., Dawson, N. L., Das, S., Sillitoe, I., Ashford, P., Lee, D., . . . Lees, J. G. (2016). Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res, 44*(D1), D404-409. doi:10.1093/nar/gkv1231

Lambert, M., Leven, B. A., & Green, R. M. (2000). New methods of cleaning up heavy metal in soils and water; Environmental science and technology briefs for citizens**;** Manhattan, KS: Kansas State University; .

Lampert, W. (2006). Daphnia: Model herbivore, predator and prey. *Polish Journal of Ecology, 54*(4), 607-620.

Lampert, W. (2011). Daphnia: development of a model organism in ecology and evolution, in: Kinne, O. *Daphnia: Development of a Model Organism in Ecology and Evolution*.

Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. *Sci Rep, 2*, 336. doi:10.1038/srep00336

Lang, D., Malviya, A. N., Hubsch, A., Kanfer, J. N., & Freysz, L. (1995). Phosphatidic acid activation of protein kinase C in LA-N-1 neuroblastoma cells. *Neurosci Lett, 201*(3), 199-202.

Lange, P. S., Chavez, J. C., Pinto, J. T., Coppola, G., Sun, C. W., Townes, T. M., . . . Ratan, R. R. (2008). ATF4 is an oxidative stress-inducible, prodeath transcription factor in neurons in vitro and in vivo. *J Exp Med, 205*(5), 1227-1242. doi:10.1084/jem.20071460

Langer, G. A., & Nudd, L. M. (1983). Effects of cations, phospholipases, and neuraminidase on calcium binding to "gas-dissected" membranes from cultured cardiac cells. *Circ Res, 53*(4), 482-490.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Langston, W. J., Bebianno, M. J., & Burt, G. R. (1998). Metal handling strategies in molluscs. 219-283.

LaRochelle, O., Labbe, S., Harrisson, J. F., Simard, C., Tremblay, V., St-Gelais, G., . . . Seguin, C. (2008). Nuclear factor-1 and metal transcription factor-1 synergistically activate the mouse metallothionein-1 gene in response to metal ions. *J Biol Chem, 283*(13), 8190-8201. doi:10.1074/jbc.M800640200

Larregle, E. V., Varas, S. M., Oliveros, L. B., Martinez, L. D., Anton, R., Marchevsky, E., & Gimenez, M. S. (2008). Lipid metabolism in liver of rat exposed to cadmium. *Food Chem Toxicol, 46*(5), 1786-1792. doi:10.1016/j.fct.2008.01.018

Latta Iv, L. C., Bakelar, J. W., Knapp, R. A., & Pfrender, M. E. (2007). Rapid evolution in response to introduced predators II: The contribution of adaptive plasticity. *BMC Evolutionary Biology, 7*. doi:10.1186/1471-2148-7-21

Latta, L. C., Weider, L. J., Colbourne, J. K., & Pfrender, M. E. (2012). The evolution of salinity tolerance in Daphnia: A functional genomics approach. *Ecology Letters, 15*(8), 794-802. doi:10.1111/j.1461-0248.2012.01799.x

Laursen, K. B., Wong, P. M., & Gudas, L. J. (2012). Epigenetic regulation by RARα maintains ligand-independent transcriptional activity. *Nucleic Acids Research, 40*(1), 102-115. doi:10.1093/nar/gkr637

Lee, J., Moulik, M., Fang, Z., Saha, P., Zou, F., Xu, Y., . . . Yechoor, V. K. (2013). Bmal1 and beta-cell clock are required for adaptation to circadian disruption, and their loss of function leads to oxidative stress-induced beta-cell failure in mice. *Mol Cell Biol, 33*(11), 2327-2338. doi:10.1128/MCB.01421-12

Lee, S. J., Kim, S., Choi, S. C., & Han, J. K. (2010). XPteg (Xenopus proximal tubules-expressed gene) is essential for pronephric mesoderm specification and tubulogenesis. *Mechanisms of Development, 127*(1-2), 49-61. doi:10.1016/j.mod.2009.11.001

Letunic, I., Doerks, T., & Bork, P. (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res, 43*(Database issue), D257-260. doi:10.1093/nar/gku949

Lewontin, R. C. (2001). *The triple helix : gene, organism, and environment*. Cambridge, Mass.: Harvard University Press.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, Y., Cheng, C. N., Verdun, V. A., & Wingert, R. A. (2014). Zebrafish nephrogenesis is regulated by interactions between retinoic acid, mecom, and notch signaling. *Developmental Biology, 386*(1), 111-122. doi:10.1016/j.ydbio.2013.11.021

Li, Y., Huang, T. T., Carlson, E. J., Melov, S., Ursell, P. C., Olson, J. L., . . . Epstein, C. J. (1995). Dilated cardiomyopathy and neonatal lethality in mutant mice lacking manganese superoxide dismutase. *Nat Genet, 11*(4), 376-381. doi:10.1038/ng1295-376

Lindqvist, L., & Block, M. (1995). Excretion of cadmium during moulting and metamorphosis in Tenebrio molitor (Coleoptera; Tenebrionidae). *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology, 111*(2), 325-328. doi:https://doi.org/10.1016/0742-8413(95)00057-U

Liska, D. J. (1998). The detoxification enzyme systems. *Altern Med Rev, 3*(3), 187-198.

Liu, C. T., Chou, M. Y., Lin, C. H., & Wu, S. M. (2012). Effects of ambient cadmium with calcium on mRNA expressions of calcium uptake related transporters in zebrafish (Danio rerio) larvae. *Fish Physiol Biochem, 38*(4), 977-988. doi:10.1007/s10695-011-9583-z

Liu, F., & Jan, K. Y. (2000). DNA damage in arsenite- and cadmium-treated bovine aortic endothelial cells. *Free Radical Biology and Medicine, 28*(1), 55-63. doi:Doi 10.1016/S0891-5849(99)00196-3

Liu, X., Jian, X., & Boerwinkle, E. (2013). dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation, 34*(9), E2393-E2402. doi:10.1002/humu.22376

Liu, Y., Shepherd, E. G., & Nelin, L. D. (2007). MAPK phosphatases--regulating the immune response. *Nat Rev Immunol, 7*(3), 202-212. doi:10.1038/nri2035

Liu, Y., & Templeton, D. M. (2007). Cadmium activates CaMK-II and initiates CaMK-II-dependent apoptosis in mesangial cells. *FEBS Lett, 581*(7), 1481-1486. doi:10.1016/j.febslet.2007.03.003

Liu, Y., Wu, H., Kou, L., Liu, X., Zhang, J., Guo, Y., & Ma, E. (2014). Two metallothionein genes in Oxya chinensis: molecular characteristics, expression patterns and roles in heavy metal stress. *PLoS ONE, 9*(11), e112759. doi:10.1371/journal.pone.0112759

Loendersloot, E. W., Verjaal, M., & Leschot, N. J. (1978). Bilateral Renal Agenesis (Potters Syndrome) in 2 Consecutive Infants. *European Journal of Obstetrics Gynecology and Reproductive Biology, 8*(3), 137-142. doi:Doi 10.1016/0028-2243(78)90063-1

Longhurst, A. R. (1988). *Analysis of Marine Ecosystems*. London; Toronto: Academic Press.

Lüning, J. (1995). Life-history responses to Chuoborus of spined and unspined Daphniu pulex. *Journal of Plankton Research, 17*(1), 71-84. doi:10.1093/plankt/17.1.71

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? An introduction and overview. *IMIA Yearbook*, 83-99.

Lutjohann, D., & von Bergmann, K. (2003). 24S-hydroxycholesterol: a marker of brain cholesterol metabolism. *Pharmacopsychiatry, 36 Suppl 2*, S102-106. doi:10.1055/s-2003-43053

Lyons, J. L., Tovar-y-Romo, L. B., Thakur, K. T., McArthur, J. C., & Haughey, N. J. (2015). Chapter 28 - Pathobiology of CNS Human Immunodeficiency Virus Infection A2 - Zigmond, Michael J. In L. P. Rowland & J. T. Coyle (Eds.), *Neurobiology of Brain Disorders* (pp. 444-466). San Diego: Academic Press.

Maekawa, T., Liu, B., Nakai, D., Yoshida, K., Nakamura, K. I., Yasukawa, M., . . . Ishii, S. (2018). ATF7 mediates TNF-alpha-induced telomere shortening. *Nucleic Acids Res*. doi:10.1093/nar/gky155

Maier, L. S., Bers, D. M., & Brown, J. H. (2007). Calmodulin and Ca2+/calmodulin kinases in the heart - physiology and pathophysiology. *Cardiovasc Res, 73*(4), 629-630. doi:10.1016/j.cardiores.2007.01.005

Majumder, S., Ghoshal, K., Gronostajski, R. M., & Jacob, S. T. (2001). Downregulation of constitutive and heavy metal-induced metallothionein-I expression by nuclear factor I. *Gene Expr, 9*(4-5), 203-215.

Manca, D., Ricard, A. C., Vantra, H., & Chevalier, G. (1994). Relation between Lipid-Peroxidation and Inflammation in the Pulmonary Toxicity of Cadmium. *Archives of Toxicology, 68*(6), 364-369. doi:DOI 10.1007/s002040050083

Manea, S. A., Todirita, A., Raicu, M., & Manea, A. (2014). C/EBP transcription factors regulate NADPH oxidase in human aortic smooth muscle cells. *J Cell Mol Med, 18*(7), 1467-1477. doi:10.1111/jcmm.12289

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform, 19*(2), 286-302. doi:10.1093/bib/bbw114

Mao, C.-y., Yang, J., Zhang, S.-y., Luo, H.-y., Song, B., Liu, Y.-t., . . . Xu, Y.-m. (2016). Exome capture sequencing identifies a novel CCM1 mutation in a Chinese family with multiple cerebral cavernous malformations. *International Journal of Neuroscience, 126*(12), 1071-1076. doi:10.3109/00207454.2015.1118628

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., . . . Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res, 43*(Database issue), D222-226. doi:10.1093/nar/gku1221

Marra, A. N., & Wingert, R. A. (2016). Epithelial cell fate in the nephron tubule is mediated by the ETS transcription factors etv5a and etv4 during zebrafish kidney development. *Developmental Biology, 411*(2), 231-245. doi:10.1016/j.ydbio.2016.01.035

Martins, R., Lithgow, G. J., & Link, W. (2016). Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. *Aging Cell, 15*(2), 196-207. doi:10.1111/acel.12427

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., . . . Wasserman, W. W. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research, 44*(D1), D110-D115. doi:10.1093/nar/gkv1176

Mazur, D. J., & Perrino, F. W. (2001). Excision of 3' termini by the Trex1 and TREX2 3'-->5' exonucleases. Characterization of the recombinant proteins. *J Biol Chem, 276*(20), 17022-17029. doi:10.1074/jbc.M100623200

McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology, 17*(1), 4-11. doi:10.1016/j.cbpa.2012.12.008

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

McKim, J. M., Jr., Choudhuri, S., & Klaassen, C. D. (1992). In vitro degradation of apo-, zinc-, and cadmium-metallothionein by cathepsins B, C, and D. *Toxicol Appl Pharmacol, 116*(1), 117-124.

McMahon, A. P., Aronow, B. J., Davidson, D. R., Davies, J. A., Gaido, K. W., Grimmond, S., . . . Zhang, P. (2008). GUDMAP: The genitourinary developmental molecular anatomy project. *Journal of the American Society of Nephrology, 19*(4), 667-671. doi:10.1681/ASN.2007101078

McPherson, E. (2007). Renal anomalies in families of individuals with congenital solitary kidney. *Genet Med, 9*(5), 298-302. doi:10.1097GIM.0b013e3180544516

McTaggart, S. J., Conlon, C., Colbourne, J. K., Blaxter, M. L., & Little, T. J. (2009). The components of the Daphnia pulex immune system as revealed by complete genome sequencing. *BMC Genomics, 10*. doi:10.1186/1471-2164-10-175

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., . . . van Helden, J. (2015). RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res, 43*(W1), W50-56. doi:10.1093/nar/gkv362

Meeus, L., Gilbert, B., Rydlewski, C., Parma, J., Roussie, A. L., Abramowicz, M., . . . Vassart, G. (2004). Characterization of a novel loss of function mutation of PAX8 in a familial case of congenital hypothyroidism with in-place, normal-sized thyroid. *Journal of Clinical Endocrinology and Metabolism, 89*(9), 4285-4291. doi:10.1210/jc.2004-0166

Mendelsohn, C., Lohnes, D., Decimo, D., Lufkin, T., LeMeur, M., Chambon, P., & Mark, M. (1994). Function of the retinoic acid receptors (RARs) during development. (II) Multiple abnormalities at various stages of organogenesis in RAR double mutants. *Development, 120*(10), 2749-2771.

Metheny, N. M., & Metheny, N. M. (2012). *Fluid and electrolyte balance : nursing considerations*. Sudbury, MA: Jones & Bartlett Learning.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res, 44*(D1), D336-342. doi:10.1093/nar/gkv1194

Micheal, S., Hogewind, B. F., Khan, M. I., Siddiqui, S. N., Zafar, S. N., Akhtar, F., . . . den Hollander, A. I. (2017). Variants in the PRPF8 Gene are Associated with Glaucoma. *Mol Neurobiol*. doi:10.1007/s12035-017-0673-5

Miner, B. E., de Meester, L., Pfrender, M. E., Lampert, W., & Hairston Jr, N. G. (2012). Linking genes to communities and ecosystems: Daphnia as an ecogenomic model. *Proceedings of the Royal Society B: Biological Sciences, 279*(1735), 1873-1882. doi:10.1098/rspb.2011.2404

Miura, T., Nishinaka, T., & Terada, T. (2008). Different functions between human monomeric carbonyl reductase 3 and carbonyl reductase 1. *Mol Cell Biochem, 315*(1-2), 113-121. doi:10.1007/s11010-008-9794-5

Miyakawa, H., Imai, M., Sugimoto, N., Ishikawa, Y., Ishikawa, A., Ishigaki, H., . . . Miura, T. (2010). Gene up-regulation in response to predator kairomones in the water flea, Daphnia pulex. *BMC Developmental Biology, 10*. doi:10.1186/1471-213X-10-45

Mohammed, H., D'Santos, C., Serandour, A. A., Ali, H. R., Brown, G. D., Atkins, A., . . . Carroll, J. S. (2013). Endogenous Purification Reveals GREB1 as a Key Estrogen Receptor Regulatory Factor. *Cell Reports, 3*(2), 342-349. doi:10.1016/j.celrep.2013.01.010

Møller, J. V., Juul, B., & le Maire, M. (1996). Structural organization, ion transport, and energy transduction of P-type ATPases. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes, 1286*(1), 1-51. doi:https://doi.org/10.1016/0304-4157(95)00017-8

Morgan, M. J., & Liu, Z. G. (2011). Crosstalk of reactive oxygen species and NF-kappaB signaling. *Cell Res, 21*(1), 103-115. doi:10.1038/cr.2010.178

Morita, K., Saitoh, M., Tobiume, K., Matsuura, H., Enomoto, S., Nishitoh, H., & Ichijo, H. (2001). Negative feedback regulation of ASK1 by protein phosphatase 5 (PP5) in response to oxidative stress. *Embo Journal, 20*(21), 6028-6036. doi:10.1093/emboj/20.21.6028

Mott, R. (1997). Estgenome: a program to align spliced dna sequences to unspliced genomic dna. *Computer Applications in the Biosciences, 13*(4), 477-478.

Muller, J. P., Vedel, M., Monnot, M. J., Touzet, N., & Wegnez, M. (1991). Molecular cloning and expression of ferritin mRNA in heavy metal-poisoned Xenopus laevis cells. *DNA Cell Biol, 10*(8), 571-579. doi:10.1089/dna.1991.10.571

Muller, L. (1986). Consequences of cadmium toxicity in rat hepatocytes: mitochondrial dysfunction and lipid peroxidation. *Toxicology, 40*(3), 285-295.

Nakamura, T., Murakami, K., Tada, H., Uehara, Y., Nogami, J., Maehara, K., . . . Sugasawa, K. (2017). Thymine DNA glycosylase modulates DNA damage response and gene expression by base excision repair-dependent and independent mechanisms. *Genes Cells, 22*(4), 392-405. doi:10.1111/gtc.12481

Naraki, Y., Hiruta, C., & Tochinai, S. (2013). Identification of the precise kairomone-sensitive period and histological characterization of necktooth formation in predator-induced polyphenism in Daphnia pulex. *Zoological Science, 30*(8), 619-625. doi:10.2108/zsj.30.619

Natarajan, G., Jeyachandran, D., Subramaniyan, B., Thanigachalam, D., & Rajagopalan, A. (2013). Congenital anomalies of kidney and hand: A review. *Clinical Kidney Journal, 6*(2), 144-149. doi:10.1093/ckj/sfs186

Nature. (2018). Transcriptomics. Retrieved from https://www.nature.com/subjects/transcriptomics

Navratilova, P., Danks, G. B., Long, A., Butcher, S., Manak, J. R., & Thompson, E. M. (2017). Sex-specific chromatin landscapes in an ultra-compact chordate genome. *Epigenetics Chromatin, 10*, 3. doi:10.1186/s13072-016-0110-4

NCBI. (2016). RefSeq. Retrieved from https://www.ncbi.nlm.nih.gov/gene?cmd=Retrieve&dopt=full_report&list_uids=1573

NCBI, B. E-value NCBI. Retrieved from https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect

Neathery, M. W., & Miller, W. J. (1975). Metabolism and Toxicity of Cadmium, Mercury, and Lead in Animals - Review. *Journal of Dairy Science, 58*(12), 1767-1781. doi:DOI 10.3168/jds.S0022-0302(75)84785-0

Nelson, R. E., Fessler, L. I., Takagi, Y., Blumberg, B., Keene, D. R., Olson, P. F., . . . Fessler, J. H. (1994). Peroxidasin - a Novel Enzyme-Matrix Protein of Drosophila Development. *Embo Journal, 13*(15), 3438-3447.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys, 69*(2 Pt 2), 026113. doi:10.1103/PhysRevE.69.026113

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., . . . Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet, 42*(1), 30-35. doi:10.1038/ng.499

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., . . . Shendure, J. (2009). Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. *Nature, 461*(7261), 272-276. doi:10.1038/nature08250

NHGRI. (2013). Bioinformatics: Introduction. Retrieved from https://www.genome.gov/25020000/online-education-kit-bioinformatics-introduction/

NHGRI. (2015). A Brief Guide to Genomics. Retrieved from https://www.genome.gov/18016863/a-brief-guide-to-genomics/

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in Saccharomyces cerevisiae. *Nucleic Acids Research, 40*(20), 10084-10097. doi:10.1093/nar/gks804

Norwood, V. F., & Chevalier, R. L. (2003). *Renal Developmental Disorders of the Fetus and Newborn*.

Nottingham, R. M., Wu, D. C., Qin, Y., Yao, J., Hunicke-Smith, S., & Lambowitz, A. M. (2016). RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *RNA, 22*(4), 597-613. doi:10.1261/rna.055558.115

Oates, M. E., Stahlhacke, J., Vavoulis, D. V., Smithers, B., Rackham, O. J., Sardar, A. J., . . . Gough, J. (2015). The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Res, 43*(Database issue), D227-233. doi:10.1093/nar/gku1041

Ohtsuji, M., Katsuoka, F., Kobayashi, A., Aburatani, H., Hayes, J. D., & Yamamoto, M. (2008). Nrf1 and Nrf2 play distinct roles in activation of antioxidant response element-dependent genes. *J Biol Chem, 283*(48), 33554-33562. doi:10.1074/jbc.M804597200

Okamoto, A., Iwamoto, Y., & Maru, Y. (2006). Oxidative stress-responsive transcription factor ATF3 potentially mediates diabetic angiopathy. *Mol Cell Biol, 26*(3), 1087-1097. doi:10.1128/MCB.26.3.1087-1097.2006

Olsen, L. C., Kourtesis, I., Busengdal, H., Jensen, M. F., Hausen, H., & Chourrout, D. (2018). Evidence for a centrosome-attracting body like structure in germ-soma segregation during early development, in the urochordate Oikopleura dioica. *BMC Dev Biol, 18*(1), 4. doi:10.1186/s12861-018-0165-5

Omiecinski, C. J., Vanden Heuvel, J. P., Perdew, G. H., & Peters, J. M. (2011). Xenobiotic metabolism, disposition, and regulation by receptors: from biochemical phenomenon to predictors of major toxicities. *Toxicol Sci, 120 Suppl 1,* S49-75. doi:10.1093/toxsci/kfq338

Omotezako, T., Matsuo, M., Onuma, T. A., & Nishida, H. (2017). DNA interference-mediated screening of maternal factors in the chordate Oikopleura dioica. *Sci Rep, 7,* 44226. doi:10.1038/srep44226

Omotezako, T., Onuma, T. A., & Nishida, H. (2015). DNA interference: DNA-induced gene silencing in the appendicularian Oikopleura dioica. *Proc Biol Sci, 282*(1807), 20150435. doi:10.1098/rspb.2015.0435

Orlowski, C., & Piotrowski, J. K. (2003). Biological levels of cadmium and zinc in the small intestine of non-occupationally exposed human subjects. *Hum Exp Toxicol, 22*(2), 57-63. doi:10.1191/0960327103ht326oa

Orsini, L., Brown, J. B., Shams Solari, O., Li, D., He, S., Podicheti, R., . . . De Meester, L. (2017). Early transcriptional response pathways in Daphnia magna are coordinated in networks of crustacean-specific genes. *Mol Ecol*. doi:10.1111/mec.14261

Osafune, K., Nishinakamura, R., & Komazaki, S. (2002). In vitro induction of the pronephric duct in Xenopus explants. *Development Growth and Differentiation, 44*(2), 161-167. doi:10.1046/j.1440-169x.2002.00631.x

Otte, K. A., Fröhlich, T., Arnold, G. J., & Laforsch, C. (2014). Proteomic analysis of Daphnia magna hints at molecular pathways involved in defensive plastic responses. *BMC Genomics, 15*(1). doi:10.1186/1471-2164-15-306

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87-98. doi:10.1038/nrg2934

Paksy, K., Varga, B., & Lazar, P. (1992). Cadmium interferes with steroid biosynthesis in rat granulosa and luteal cells in vitro. *Biometals, 5*(4), 245-250.

Palmgren, M. G., & Nissen, P. (2011). P-type ATPases. *Annu Rev Biophys, 40*, 243-266. doi:10.1146/annurev.biophys.093008.131331

Pan, Y. H., Yu, B. Z., Singer, A. G., Ghomashchi, F., Lambeau, G., Gelb, M. H., . . . Bahnson, B. J. (2002). Crystal structure of human group X secreted phospholipase A2. Electrostatically neutral interfacial surface targets zwitterionic membranes. *J Biol Chem, 277*(32), 29086-29093. doi:10.1074/jbc.M202531200

Parkinson, J., & Blaxter, M. (2009). Expressed Sequence Tags: An Overview. In J. Parkinson (Ed.), *Expressed Sequence Tags (ESTs): Generation and Analysis* (pp. 1-12). Totowa, NJ: Humana Press.

Parra, G., Blanco, E., & Guigó, R. (2000). GeneId in Drosophila. *Genome Research, 10*(4), 511-515. doi:10.1101/gr.10.4.511

Patlolla, A. K., Barnes, C., Yedjou, C., Velma, V. R., & Tchounwou, P. B. (2009). Oxidative stress, DNA damage, and antioxidant enzyme activity induced by hexavalent chromium in Sprague-Dawley rats. *Environ Toxicol, 24*(1), 66-73. doi:10.1002/tox.20395

Pauwels, K., Stoks, R., & De Meester, L. (2005). Coping with predator stress: Interclonal differences in induction of heat-shock proteins in the water flea Daphnia magna. *Journal of Evolutionary Biology, 18*(4), 867-872. doi:10.1111/j.1420-9101.2005.00890.x

Pẽalva-Arana, D. C., Lynch, M., & Robertson, H. M. (2009). The chemoreceptor genes of the waterflea Daphnia pulex: Many Grs but no Ors. *BMC Evolutionary Biology, 9*(1). doi:10.1186/1471-2148-9-79

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., de Castro, E., . . . Bridge, A. (2015). HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res, 43*(Database issue), D1064-1070. doi:10.1093/nar/gku1002

Peng, F. Y., Hu, Z., & Yang, R.-C. (2016). Bioinformatic prediction of transcription factor binding sites at promoter regions of genes for photoperiod and vernalization responses in model and temperate cereal plants. *BMC Genomics, 17*(1), 573. doi:10.1186/s12864-016-2916-7

Perner, B., Englert, C., & Bollig, F. (2007). The Wilms tumor genes wt1a and wt1b control different steps during formation of the zebrafish pronephros. *Developmental Biology, 309*(1), 87-96. doi:10.1016/j.ydbio.2007.06.022

Petkovich, M., Brand, N. J., Krust, A., & Chambon, P. (1987). A human retinoic acid receptor which belongs to the family of nuclear receptors. *Nature, 330*(6147), 444-450.

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics, 9*(8). doi:10.1371/journal.pgen.1003709

Petrusek, A., Tollrian, R., Schwenk, K., Haas, A., & Laforsch, C. (2009). A "crown of thorns" is an inducible defense that protects Daphnia against an ancient predator. *Proceedings of the National Academy of Sciences of the United States of America, 106*(7), 2248-2252. doi:10.1073/pnas.0808075106

Pialoux, V., Mounier, R., Brown, A. D., Steinback, C. D., Rawling, J. M., & Poulin, M. J. (2009). Relationship between oxidative stress and HIF-1 alpha mRNA during sustained hypoxia in humans. *Free Radic Biol Med, 46*(2), 321-326. doi:10.1016/j.freeradbiomed.2008.10.047

Piscione, T. D., & Rosenblum, N. D. (2002). The molecular control of renal branching morphogenesis: Current knowledge and emerging insights. *Differentiation, 70*(6), 227-246. doi:10.1046/j.1432-0436.2002.700602.x

Polvani, S., Tarocchi, M., & Galli, A. (2012). PPARgamma and Oxidative Stress: Con(beta) Catenating NRF2 and FOXO. *PPAR Res, 2012*, 641087. doi:10.1155/2012/641087

Potter, E. L. (1946). Facial characteristics of infants with bilateral renal agenesis. *American journal of obstetrics and gynecology, 51*, 885-888.

Potter, E. L. (1965). Bilateral absence of ureters and kidneys: A report of 50 cases. *Obstetrics and Gynecology, 25*(1), 3-12.

Poynton, H. C., Loguinov, A. V., Varshavsky, J. R., Chan, S., Perkins, E. J., & Vulpe, C. D. (2008). Gene expression profiling in Daphnia magna part I: concentration-dependent profiles provide support for the No Observed Transcriptional Effect Level. *Environ Sci Technol, 42*(16), 6250-6256.

Poynton, H. C., Varshavsky, J. R., Chang, B., Cavigiolio, G., Chan, S., Holman, P. S., . . . Vulpe, C. D. (2007). Daphnia magna ecotoxicogenomics provides mechanistic insights into metal toxicity. *Environmental Science & Technology, 41*(3), 1044-1050. doi:10.1021/es0615573

Qi, Y. X., Liu, Y. B., & Rong, W. H. (2011). [RNA-Seq and its applications: a new technology for transcriptomics]. *Yi Chuan, 33*(11), 1191-1202.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: Protein domains identifier. *Nucleic Acids Research, 33*(SUPPL. 2), W116-W120. doi:10.1093/nar/gki442

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

R Development Core Team, R. (2006). *A language and environment for statistical computing* (Vol. 1).

Rae, J. M., Johnson, M. D., Scheys, J. O., Cordero, K. E., Larios, J. M., & Lippman, M. E. (2005). GREB1 is a critical regulator of hormone dependent breast cancer growth. *Breast Cancer Research and Treatment, 92*(2), 141-149. doi:10.1007/s10549-005-1483-4

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., . . . Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454. doi:10.1038/nature05329

Ree, A. H., Bratland, A., Nome, R. V., Stokke, T., & Fodstad, O. (2003). Repression of mRNA for the PLK cell cycle gene after DNA damage requires BRCA1. *Oncogene, 22*(55), 8952-8955. doi:10.1038/sj.onc.1207000

RefSeq. (2011). MKNK2. Retrieved from https://www.ncbi.nlm.nih.gov/gene/2872

Repka, S., Walls, M., & Ketola, M. (1995). Neck spine protects Daphnia pulex from predation by Chaoborus, but individuals with longer tail spine are at a greater risk. *Journal of Plankton Research, 17*(2), 393-403. doi:10.1093/plankt/17.2.393

Reyland, M. E. (2007). Protein Kinase C and Apoptosis. In R. Srivastava (Ed.), *Apoptosis, Cell Signaling, and Human Diseases: Molecular Mechanisms, Volume 2* (pp. 31-55). Totowa, NJ: Humana Press.

Riessen, H. P. (1992). Cost-benefit model for the induction of an antipredator defense. *American Naturalist, 140*(2), 349-362. doi:10.1086/285416

Riessen, H. P. (1999). Chaoborus predation and delayed reproduction in Daphnia: A demographic modeling approach. *Evolutionary Ecology, 13*(4), 339-363. doi:10.1023/A:1006715120109

Riessen, H. P. (2012). Costs of predator-induced morphological defences in Daphnia. *Freshwater Biology, 57*(7), 1422-1433. doi:10.1111/j.1365-2427.2012.02805.x

Riessen, H. P., & Sprules, W. G. (1990). Demographic costs of antipredator defenses in Daphnia pulex. *Ecology, 71*(4), 1536-1546. doi:10.2307/1938290

Risal, S., Adhikari, D., & Liu, K. (2016). Animal Models for Studying the In Vivo Functions of Cell Cycle CDKs. *Methods Mol Biol, 1336*, 155-166. doi:10.1007/978-1-4939-2926-9_13

Roberts, A., & Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods, 10*(1), 71-U99. doi:10.1038/Nmeth.2251

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-seq. *Bioinformatics, 27*(17), 2325-2329. doi:10.1093/bioinformatics/btr355

Robison, B. H., Reisenbichler, K. R., & Sherlock, R. E. (2005). Ocean science: Giant larvacean houses: Rapid carbon transport to the deep sea floor. *Science, 308*(5728), 1609-1611. doi:10.1126/science.1109104

Rodriguez, M. M. (2014). Congenital Anomalies of the Kidney and the Urinary Tract (CAKUT). *Fetal Pediatr Pathol, 33*(5-6), 293-320. doi:10.3109/15513815.2014.959678

Roodhooft, A. M., Birnholz, J. C., & Holmes, L. B. (1984). Familial Nature of Congenital Absence and Severe Dysgenesis of Both Kidneys. *New England Journal of Medicine, 310*(21), 1341-1345. doi:10.1056/NEJM198405243102101

Rosselot, C., Spraggon, L., Chia, I., Batourina, E., Riccio, P., Lu, B., . . . Mendelsohn, C. (2010). Non-cell-autonomous retinoid signaling is crucial for renal development. *Development, 137*(2), 283-292. doi:10.1242/dev.040287

Roth, M., Obaidat, A., & Hagenbuch, B. (2012). OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies. *Br J Pharmacol, 165*(5), 1260-1287. doi:10.1111/j.1476-5381.2011.01724.x

Roth, W., Kermer, P., Krajewska, M., Welsh, K., Davis, S., Krajewski, S., & Reed, J. C. (2003). Bifunctional apoptosis inhibitor (BAR) protects neurons from diverse cell death pathways. *Cell Death Differ, 10*(10), 1178-1187. doi:10.1038/sj.cdd.4401287

Rozenberg, A., Parida, M., Leese, F., Weiss, L. C., Tollrian, R., & Manak, J. R. (2015). Transcriptional profiling of predator-induced phenotypic plasticity in Daphnia pulex. *Frontiers in Zoology, 12*, 18. doi:10.1186/s12983-015-0109-x

Ruderfer, D. M., Hamamsy, T., Lek, M., Karczewski, K. J., Kavanagh, D., Samocha, K. E., . . . Purcell, S. M. (2016). Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature Genetics, 48*(10), 1107-1111. doi:10.1038/ng.3638

Rush, G. F., Gorski, J. R., Ripple, M. G., Sowinski, J., Bugelski, P., & Hewitt, W. R. (1985). Organic hydroperoxide-induced lipid peroxidation and cell death in isolated hepatocytes. *Toxicol Appl Pharmacol, 78*(3), 473-483.

Sagane, Y., Zech, K., Bouquet, J. M., Schmid, M., Bal, U., & Thompson, E. M. (2010). Functional specialization of cellulose synthase genes of prokaryotic origin in chordate larvaceans. *Development, 137*(9), 1483-1492. doi:10.1242/dev.044503

Salih, M., Gautschi, I., van Bemmelen, M. X., Di Benedetto, M., Brooks, A. S., Lugtenberg, D., . . . Hoorn, E. J. (2017). A Missense Mutation in the Extracellular Domain of alphaENaC Causes Liddle Syndrome. *J Am Soc Nephrol, 28*(11), 3291-3299. doi:10.1681/ASN.2016111163

Salinas, A. E., & Wong, M. G. (1999). Glutathione S-transferases--a review. *Curr Med Chem, 6*(4), 279-309.

Sanna-Cherchi, S., Caridi, G., Weng, P. L., Scolari, F., Perfumo, F., Gharavi, A. G., & Ghiggeri, G. M. (2007). Genetic approaches to human renal agenesis/hypoplasia and dysplasia. *Pediatric Nephrology, 22*(10), 1675-1684. doi:10.1007/s00467-007-0479-1

Sanna-Cherchi, S., Khan, K., Westland, R., Krithivasan, P., Fievet, L., Rasouly, H. M., . . . Gharavi, A. G. (2017). Exome-wide Association Study Identifies GREB1L Mutations in Congenital Kidney Malformations. *Am J Hum Genet, 101*(5), 789-802. doi:10.1016/j.ajhg.2017.09.018

Sanna-Cherchi, S., Ravani, P., Corbani, V., Parodi, S., Haupt, R., Piaggio, G., . . . Ghiggeri, G. M. (2009). Renal outcome in patients with congenital anomalies of the kidney and urinary tract. *Kidney International, 76*(5), 528-533. doi:10.1038/ki.2009.220

Saxén, L., & Sariola, H. (1987). Early organogenesis of the kidney. *Pediatric Nephrology, 1*(3), 385-392. doi:10.1007/BF00849241

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., . . . Ye, J. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 40*(D1), D13-D25. doi:10.1093/nar/gkr1184

Schaack, S. (2008). Daphnia comes of age: An ecological model in the genomic era. *Molecular Ecology, 17*(7), 1634-1635. doi:10.1111/j.1365-294X.2008.03698.x

Schirmer, K., Fischer, B. B., Madureira, D. J., & Pillai, S. (2010). Transcriptomics in ecotoxicology. *Analytical and Bioanalytical Chemistry, 397*(3), 917-923. doi:10.1007/s00216-010-3662-3

Schreuder, M. F., Langemeijer, M. E., Bökenkamp, A., Delemarre-Van De Waal, H. A., & Van Wijk, J. A. E. (2008). Hypertension and microalbuminuria in children with congenital solitary kidneys. *Journal of Paediatrics and Child Health, 44*(6), 363-368. doi:10.1111/j.1440-1754.2008.01315.x

Schulmeister, A., Schmid, M., & Thompson, E. M. (2007). Phosphorylation of the histone H3.3 variant in mitosis and meiosis of the urochordate Oikopleura dioica. *Chromosome Research, 15*(2), 189-201. doi:10.1007/s10577-006-1112-z

Schultz, M. A., Abdel-Mageed, A. B., & Mondal, D. (2010). The nrf1 and nrf2 balance in oxidative stress regulation and androgen signaling in prostate cancer cells. *Cancers (Basel), 2*(2), 1354-1378. doi:10.3390/cancers2021354

Scitable. (2014a). microarray-202.   Retrieved from https://www.nature.com/scitable/definition/microarray-202

Scitable. (2014b). *transcription factors* Retrieved from https://www.nature.com/scitable/definition/general-transcription-factor-transcription-factor-167

Scitable. (2014c). Translation: DNA to mRNA to Protein.   Retrieved from https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., . . . Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science, 305*(5683), 525-528. doi:10.1126/science.1098918

Seo, H. C., Kube, M., Edvardsen, R. B., Jensen, M. F., Beck, A., Spriet, E., . . . Chourrout, D. (2001). Miniature genome in the marine chordate Oikopleura dioica. *Science, 294*(5551), 2506. doi:10.1126/science.294.5551.2506

Shaw, J. R., Colbourne, J. K., Davey, J. C., Glaholt, S. P., Hampton, T. H., Chen, C. Y., . . . Hamilton, J. W. (2007). Gene response profiles for Daphnia pulex exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. *BMC Genomics, 8*, 477. doi:10.1186/1471-2164-8-477

Shaw, J. R., Colbourne, J. K., Davey, J. C., Glaholt, S. P., Hampton, T. H., Chen, C. Y., . . . Hamilton, J. W. (2007). Gene response profiles for Daphnia pulex exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. *BMC Genomics, 8*(1), 477. doi:10.1186/1471-2164-8-477

Shaw, J. R., Dempsey, T. D., Chen, C. Y., Hamilton, J. W., & Folt, C. L. (2006). Comparative toxicity of cadmium, zinc, and mixtures of cadmium and zinc to daphnids. *Environ Toxicol Chem, 25*(1), 182-189.

Shaw, J. R., Pfrender, M. E., Eads, B. D., Klaper, R., Callaghan, A., Sibly, R. M., . . . Colbourne, J. K. (2008). Daphnia as an emerging model for toxicological genomics. In C. Hogstrand & P. Kille (Eds.), *Advances in Experimental Biology* (Vol. 2, pp. 165-328): Elsevier.

Shawlot, W., & Behringer, R. R. (1995). Requirement for liml in head-organizer function. *Nature, 374*(6521), 425-430. doi:10.1038/374425a0

Sheng, Q., Vickers, K., Zhao, S., Wang, J., Samuels, D. C., Koues, O., . . . Guo, Y. (2017). Multi-perspective quality control of Illumina RNA sequencing data analysis. *Brief Funct Genomics, 16*(4), 194-204. doi:10.1093/bfgp/elw035

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res, 29*(1), 308-311.

Shim, G. J., Kis, L. L., Warner, M., & Gustafsson, J. Å. (2004). Autoimmune glomerulonephritis with spontaneous formation of splenic germinal centers in mice lacking the estrogen receptor alpha gene. *Proceedings of the National Academy of Sciences of the United States of America, 101*(6), 1720-1724. doi:10.1073/pnas.0307915100

Sidhu, M., Sharma, M., Bhatia, M., Awasthi, Y. C., & Nath, R. (1993). Effect of chronic cadmium exposure on glutathione S-transferase and glutathione peroxidase activities in rhesus monkey: the role of selenium. *Toxicology, 83*(1-3), 203-213.

Siegmund, O. H., Vallerga, J. V., Tremsin, A. S., McPhate, J., Michalet, X., Colyer, R. A., & Weiss, S. (2011). Microchannel Plate Imaging Photon Counters for Ultraviolet through NIR Detection with High Time Resolution. *Proc SPIE Int Soc Opt Eng, 8033*, 1350904. doi:10.1117/12.884271

Sigrist, C. J., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., . . . Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Res, 41*(Database issue), D344-347. doi:10.1093/nar/gks1067

Sinnott, R., Winters, L., Larson, B., Mytsa, D., Taus, P., Cappell, K. M., & Whitehurst, A. W. (2014). Mechanisms promoting escape from mitotic stress-induced tumor cell death. *Cancer Research, 74*(14), 3857-3869. doi:10.1158/0008-5472.CAN-13-3398

Sirbu, B. M., & Cortez, D. (2013). DNA damage response: three levels of DNA repair regulation. *Cold Spring Harb Perspect Biol, 5*(8), a012724. doi:10.1101/cshperspect.a012724

Skinner, M. A., Safford, S. D., Reeves, J. G., Jackson, M. E., & Freemerman, A. J. (2008). Renal Aplasia in Humans Is Associated with RET Mutations. *American Journal of Human Genetics, 82*(2), 344-351. doi:10.1016/j.ajhg.2007.10.008

Skreb, Y., & Fischer, A. B. (1984). Toxicity of nickel for mammalian cells in culture. *Zentralbl Bakteriol Mikrobiol Hyg B, 178*(5-6), 432-445.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics, 6*. doi:10.1186/1471-2105-6-31

Smith, L. C., Azumi, K., & Nonaka, M. (1999). Complement systems in invertebrates. The ancient alternative and lectin pathways. *Immunopharmacology, 42*(1-3), 107-120. doi:10.1016/S0162-3109(99)00009-0

Smith, M. E., & Morton, D. G. (2011). The Digestive System : Systems of the Body Series.

Sobkowiak, R., & Deckert, J. (2004). The effect of cadmium on cell cycle control in suspension culture cells of soybean. *Acta Physiologiae Plantarum, 26*(3), 335-344. doi:10.1007/s11738-004-0023-x

Soetaert, A., van der Ven, K., Moens, L. N., Vandenbrouck, T., van Remortel, P., & De Coen, W. M. (2007). Daphnia magna and ecotoxicogenomics: Gene expression profiles of the anti-ecdysteroidal fungicide fenarimol using energy-, molting- and life stage-related cDNA libraries. *Chemosphere, 67*(1), 60-71. doi:10.1016/j.chemosphere.2006.09.076

Soetaert, A., Vandenbrouck, T., van der Ven, K., Maras, M., van Remortel, P., Blust, R., & de Coen, W. M. (2007). Molecular responses during cadmium-induced stress in Daphnia magna: Integration of differential gene expression with higher-level effects. *Aquatic Toxicology, 83*(3), 212-222. doi:10.1016/j.aquatox.2007.04.010

Spada, F., Steen, H., Troedsson, C., Kallesøe, T., Spriet, E., Mann, M., & Thompson, E. M. (2001). Molecular Patterning of the Oikoplastic Epithelium of the Larvacean Tunicate Oikopleura dioica. *Journal of Biological Chemistry, 276*(23), 20624-20632. doi:10.1074/jbc.M100438200

Sritunyalucksana, K., Wongsuebsantati, K., Johansson, M. W., & Soderhall, K. (2001). Peroxinectin, a cell adhesive protein associated with the proPO system from the black tiger shrimp, Penaeus monodon. *Dev Comp Immunol, 25*(5-6), 353-363.

Stach, T., Winter, J., Bouquet, J. M., Chourrout, D., & Schnabel, R. (2008). Embryology of a planktonic tunicate reveals traces of sessility. *Proceedings of the National Academy of Sciences of the United States of America, 105*(20), 7229-7234. doi:10.1073/pnas.0710196105

Steinberg, C. E. W. (2012). Stress Ecology Environmental Stress as Ecological Driving Force and Key Player in Evolution.

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S. T., & Cooper, D. N. (2009). The Human Gene Mutation Database: 2008 update. *Genome Medicine, 1*(1), 13-13. doi:10.1186/gm13

Sterling, K. M., Mandal, P. K., Roggenbeck, B. A., Ahearn, S. E., Gerencser, G. A., & Ahearn, G. A. (2007). Heavy metal detoxification in crustacean epithelial lysosomes: role of anions in the compartmentalization process. *J Exp Biol, 210*(Pt 19), 3484-3493. doi:10.1242/jeb.008300

Stibor, H. (2002). The role of yolk protein dynamics and predator kairomones for the life history of Daphnia magna. *Ecology, 83*(2), 362-369.

Stibor, H., & Luning, J. (1994). Predator-induced phenotypic variation in the pattern of growth and reproduction in Daphnia hyalina (Crustacea: Cladocera). *Functional Ecology, 8*(1), 97-101.

Stohs, S. J., & Bagchi, D. (1995). Oxidative mechanisms in the toxicity of metal ions. *Free Radic Biol Med, 18*(2), 321-336.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics, 16*(1), 16-23.

Sutton, D. J., Tchounwou, P. B., Ninashvili, N., & Shen, E. (2002). Mercury Induces Cytotoxicity and Transcriptionally Activates Stress Genes in Human Liver Carcinoma (HepG(2)) Cells. *International Journal of Molecular Sciences, 3*(9), 965-984. doi:10.3390/i3090965

Talbot, J. C., & Amacher, S. L. (2014). A streamlined CRISPR pipeline to reliably generate zebrafish frameshifting alleles. *Zebrafish, 11*(6), 583-585. doi:10.1089/zeb.2014.1047

Tamir, I. (2012). Sequencing contaminants. Retrieved from https://github.com/csf-ngs/fastqc/blob/master/Contaminants/contaminant_list.txt

Tchounwou, P. B., Newsome, C., Williams, J., & Glass, K. (2008). Copper-Induced Cytotoxicity and Transcriptional Activation of Stress Genes in Human Liver Carcinoma (HepG(2)) Cells. *Met Ions Biol Med, 10*, 285-290.

Tchounwou, P. B., Yedjou, C. G., Patlolla, A. K., & Sutton, D. J. (2012). Heavy metal toxicity and the environment. *EXS, 101*, 133-164. doi:10.1007/978-3-7643-8340-4_6

Tellam, R. L., Wijffels, G., & Willadsen, P. (1999). Peritrophic matrix proteins. *Insect Biochem Mol Biol, 29*(2), 87-101.

Thisse, C., & Thisse, B. (2008). High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nature Protocols, 3*(1), 59-69. doi:10.1038/nprot.2007.514

Thompson, E. M., Kallesøe, T., & Spada, F. (2001). Diverse genes expressed in distinct regions of the trunk epithelium define a monolayer cellular template for construction of the oikopleurid house. *Developmental Biology, 238*(2), 260-273. doi:10.1006/dbio.2001.0414

Thomson, S. A., Baldwin, W. S., Wang, Y. H., Kwon, G., & LeBlanc, G. A. (2009). Annotation, phylogenetics, and expression of the nuclear receptors in Daphnia pulex. *BMC Genomics, 10*, 500. doi:10.1186/1471-2164-10-500

Toka, H. R., Toka, O., Hariri, A., & Nguyen, H. T. (2010). Congenital anomalies of kidney and urinary tract. *Seminars in Nephrology, 30*(4), 374-386. doi:10.1016/j.semnephrol.2010.06.004

Tokishita, S. i., Kato, Y., Kobayashi, T., Nakamura, S., Ohta, T., & Yamagata, H. (2006). Organization and repression by juvenile hormone of a vitellogenin gene cluster in the crustacean, Daphnia magna. *Biochemical and Biophysical Research Communications, 345*(1), 362-370. doi:10.1016/j.bbrc.2006.04.102

Tollrian, R. (1995a). Chaoborus crystallinus predation on Daphnia pulex: can induced morphological changes balance effects of body size on vulnerability? *Oecologia, 101*(2), 151-155. doi:10.1007/BF00317278

Tollrian, R. (1995b). Predator-induced morphological defenses: Costs, life history shifts, and maternal effects in Daphnia pulex. *Ecology, 76*(6), 1691-1705. doi:10.2307/1940703

Tollrian, R., & Dodson, S. I. (1999). Inducible defenses in cladocera: Constraints, costs, and multipredator environments. *The Ecology and Evolution of Inducible Defenses*, 177-202.

Tollrian, R., & Harvell, D. (1999). *The Ecology and evolution of inducible defenses*. Princeton, N.Y.: Princeton University Press.

Tollrian, R., & Leese, F. (2010). Ecological genomics: Steps towards unraveling the genetic basis of inducible defenses in Daphnia. *BMC Biology, 8*. doi:10.1186/1741-7007-8-51

Tootle, T. L., & Spradling, A. C. (2008). Drosophila Pxt: a cyclooxygenase-like facilitator of follicle maturation. *Development, 135*(5), 839-847. doi:10.1242/dev.017590

Torres, F., das Gracas, M., Melo, M., & Tosti, A. (2009). Management of contact dermatitis due to nickel allergy: an update. *Clin Cosmet Investig Dermatol, 2,* 39-48.

Torres, M., Gómez-Pardo, E., Dressler, G. R., & Gruss, P. (1995). Pax-2 controls multiple steps of urogenital development. *Development, 121*(12), 4057-4065.

Toyoda, Y., Hagiya, Y., Adachi, T., Hoshijima, K., Kuo, M. T., & Ishikawa, T. (2008). MRP class of human ATP binding cassette (ABC) transporters: historical background and new research directions. *Xenobiotica, 38*(7-8), 833-862. doi:10.1080/00498250701883514

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol, 31*(1), 46-53. doi:10.1038/nbt.2450

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28*(5), 511-515. doi:10.1038/nbt.1621

Troedsson, C., Bouquet, J. M., Aksnes, D. L., & Thompson, E. M. (2002). Resource allocation between somatic growth and reproductive output in the pelagic chordate Oikopleura dioica allows opportunistic response to nutritional variation. *Marine Ecology Progress Series, 243*, 83-91.

Trueba, S. S., Augé, J., Mattei, G., Etchevers, H., Martinovic, J., Czernichow, P., . . . Attié-Bitach, T. (2005). PAX8, TITF1, and FOXE1 gene expression patterns during human development: New insights into human thyroid development and thyroid dysgenesis-associated malformations. *Journal of Clinical Endocrinology and Metabolism, 90*(1), 455-462. doi:10.1210/jc.2004-1358

Truhaut, R. (1977). Ecotoxicology: objectives, principles and perspectives. *Ecotoxicol Environ Saf, 1*(2), 151-173.

Tufail, M., & Takeda, M. (2012). Hemolymph Proteins and Functional Peptides : Recent Advances in Insects and Other Arthropods.

Turgeon, D., Chouinard, S., Belanger, P., Picard, S., Labbe, J. F., Borgeat, P., & Belanger, A. (2003). Glucuronidation of arachidonic and linoleic acid metabolites by human UDP-glucuronosyltransferases. *J Lipid Res, 44*(6), 1182-1191. doi:10.1194/jlr.M300010-JLR200

Uetani, N., & Bouchard, M. (2009). Plumbing in the embryo: Developmental defects of the urinary tracts. *Clinical Genetics, 75*(4), 307-317. doi:10.1111/j.1399-0004.2009.01175.x

Uniprot. (2018). MK09_HUMAN. Retrieved from http://www.uniprot.org/uniprot/P45984

UniProt: the universal protein knowledgebase. (2017). *Nucleic Acids Research, 45*(D1), D158-D169. doi:10.1093/nar/gkw1099

USDOEGRP. (2008). Genomics and Its Impact on Science and Society. Retrieved from https://web.ornl.gov/sci/techresources/Human_Genome/publicat/primer2001/primer11.pdf

Valko, M., Morris, H., & Cronin, M. T. (2005). Metals, toxicity and oxidative stress. *Curr Med Chem, 12*(10), 1161-1208.

Van Esch, H., Groenen, P., Nesbit, M. A., Schuffenhauer, S., Lichtner, P., Vanderlinden, G., . . . Devriendt, K. (2000). GATA3 haplo-insufficiency causes human HDR syndrome. *Nature, 406*(6794), 419-422. doi:10.1038/35019088

Van Hoof, C., & Goris, J. (2003). Phosphatases in apoptosis: to be or not to be, PP2A is in the heart of the question. *Biochim Biophys Acta, 1640*(2-3), 97-104.

Van Kanegan, M. J., Adams, D. G., Wadzinski, B. E., & Strack, S. (2005). Distinct protein phosphatase 2A heterotrimers modulate growth factor signaling to extracellular signal-regulated kinases and Akt. *J Biol Chem, 280*(43), 36029-36036. doi:10.1074/jbc.M506986200

Vandenbrouck, T., Soetaert, A., van der Ven, K., Blust, R., & De Coen, W. (2009). Nickel and binary metal mixture responses in Daphnia magna: Molecular fingerprints and (sub)organismal effects. *Aquatic Toxicology, 92*(1), 18-29. doi:10.1016/j.aquatox.2008.12.012

Vasilyev, A., Liu, Y., Hellman, N., Pathak, N., & Drummond, I. A. (2012). Mechanical stretch and PI3K signaling link cell migration and proliferation to coordinate epithelial tubule morphogenesis in the zebrafish pronephros. *PLoS ONE, 7*(7). doi:10.1371/journal.pone.0039992

Vasilyev, A., Liu, Y., Mudumana, S., Mangos, S., Lam, P. Y., Majumdar, A., . . . Drummond, I. A. (2009). Collective cell migration drives morphogenesis of the kidney nephron. *PLoS Biology, 7*(1). doi:10.1371/journal.pbio.1000009

Veal, E. A., Toone, W. M., Jones, N., & Morgan, B. A. (2002). Distinct roles for glutathione S-transferases in the oxidative stress response in Schizosaccharomyces pombe. *J Biol Chem, 277*(38), 35523-35531. doi:10.1074/jbc.M111548200

Vergilino, R., Markova, S., Ventura, M., Manca, M., & Dufresne, F. (2011). Reticulate evolution of the Daphnia pulex complex as revealed by nuclear markers. *Molecular Ecology, 20*(6), 1191-1207. doi:10.1111/j.1365-294X.2011.05004.x

Viarengo, A., & Nott, J. A. (1993). Mechanisms of heavy metal cation homeostasis in marine invertebrates. *Comparative Biochemistry and Physiology Part C: Comparative Pharmacology, 104*(3), 355-372. doi:https://doi.org/10.1016/0742-8413(93)90001-2

Vila-Petroff, M., Salas, M. A., Said, M., Valverde, C. A., Sapia, L., Portiansky, E., . . . Mattiazzi, A. (2007). CaMKII inhibition protects against necrosis and apoptosis in irreversible ischemia-reperfusion injury. *Cardiovasc Res, 73*(4), 689-698. doi:10.1016/j.cardiores.2006.12.003

Vilar, J., Gilbert, T., Moreau, E., & Merlet-Bénichou, C. (1996). Metanephros organogenesis is highly stimulated by vitamin A derivatives in organ culture. *Kidney International, 49*(5), 1478-1487. doi:10.1038/ki.1996.208

Visser, G. J., Peters, P. H., & Theuvenet, A. P. (1993). Cadmium ion is a non-competitive inhibitor of red cell Ca(2+)-ATPase activity. *Biochim Biophys Acta, 1152*(1), 26-34.

Vivante, A., Mann, N., Yonath, H., Weiss, A. C., Getwan, M., Kaminski, M. M., . . . Hildebrandt, F. (2017). A dominant mutation in nuclear receptor interacting protein 1 causes urinary tract malformations via dysregulation of retinoic acid signaling. *Journal of the American Society of Nephrology, 28*(8), 2364-2376. doi:10.1681/ASN.2016060694

Vize, P. D., Seufert, D. W., Carroll, T. J., & Wallingford, J. B. (1997). Model systems for the study of kidney development: Use of the pronephros in the analysis of organ induction and patterning. *Developmental Biology, 188*(2), 189-204. doi:10.1006/dbio.1997.8629

Voogt, P. A., den Besten, P. J., Kusters, G. C., & Messing, M. W. (1987). Effects of cadmium and zinc on steroid metabolism and steroid level in the sea star Asterias rubens L. *Comp Biochem Physiol C, 86*(1), 83-89.

Vrontou, S., Petrou, P., Meyer, B. I., Galanopoulos, V. K., Imai, K., Yanagi, M., . . . Chalepakis, G. (2003). Fras1 deficiency results in cryptophthalmos, renal agenesis and blebbed phenotype in mice. *Nature Genetics, 34*(2), 209-214. doi:10.1038/ng1168

Wang, D., & Dubois, R. N. (2010). Eicosanoids and cancer. *Nat Rev Cancer, 10*(3), 181-193. doi:10.1038/nrc2809

Wang, X. F., Xing, M. L., Shen, Y., Zhu, X., & Xu, L. H. (2006). Oral administration of Cr(VI) induced oxidative stress, DNA damage and apoptotic cell death in mice. *Toxicology, 228*(1), 16-23. doi:10.1016/j.tox.2006.08.005

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi:10.1038/nrg2484

Wang, Z., Liu, X., Yang, B. Z., & Gelernter, J. (2013). The role and challenges of exome sequencing in studies of human diseases. *Front Genet, 4*, 160. doi:10.3389/fgene.2013.00160

Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet, 5*(4), 276-287. doi:10.1038/nrg1315

Webb, M. (1979). *The chemistry, biochemistry and biology of cadmium : Topics in environmental health 2*. Amsterdam: Elsevier, 1979.

Weber, A., & Van Noordwijk, A. (2002). Swimming behaviour of Daphnia clones: Differentiation through predator infochemicals. *Journal of Plankton Research, 24*(12), 1335-1348. doi:10.1093/plankt/24.12.1335

Weiss, L. C., Kruppert, S., Laforsch, C., & Tollrian, R. (2012). Chaoborus and gasterosteus anti-predator responses in Daphnia pulex are mediated by independent cholinergic and gabaergic neuronal signals. *PLoS ONE, 7*(5). doi:10.1371/journal.pone.0036879

Weller, P. F., Bach, D. S., & Austen, K. F. (1984). Biochemical characterization of human eosinophil Charcot-Leyden crystal protein (lysophospholipase). *J Biol Chem, 259*(24), 15100-15105.

Wells, R. A., Gu, C. H., & dela Paz, J. (2009). ARNT Mediates Chemotherapy Resistance by Modulating the Response to Oxidative Stress in AML Cells. *Blood, 114*(22), 1071-1071.

Westerfield, M., & Zfin. (2000). *The zebrafish book : a guide for the laboratory use of zebrafish Danio (Brachydanio) rerio*. [Eugene, Or.]: ZFIN.

Whitehurst, A. W., Bodemann, B. O., Cardenas, J., Ferguson, D., Girard, L., Peyton, M., . . . White, M. A. (2007). Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature, 446*(7137), 815-819. doi:10.1038/nature05697

Wiley, H. S., & Wallace, R. A. (1981). The Structure of Vitellogenin - Multiple Vitellogenins in Xenopus-Laevis Give Rise to Multiple Forms of the Yolk Proteins. *Journal of Biological Chemistry, 256*(16), 8626-8634.

Wilking, M., Ndiaye, M., Mukhtar, H., & Ahmad, N. (2013). Circadian rhythm connections to oxidative stress: implications for human health. *Antioxid Redox Signal, 19*(2), 192-208. doi:10.1089/ars.2012.4889

Wilkinson, L. (2011). Venneuler: Venn and Euler Diagrams. *Venneular: Venn and Euler Diagrams*.

Wingert, R. A., Selleck, R., Yu, J., Song, H. D., Chen, Z., Song, A., . . . Davidson, A. J. (2007). The cdx genes and retinoic acid control the positioning and segmentation of the zebrafish pronephros. *PLoS Genetics, 3*(10), 1922-1938. doi:10.1371/journal.pgen.0030189

Winkles, J. A., & Alberts, G. F. (2005). Differential regulation of polo-like kinase 1, 2, 3, and 4 gene expression in mammalian cells and tissues. *Oncogene, 24*(2), 260-266. doi:10.1038/sj.onc.1208219

Woolf, A. S., & Hillman, K. A. (2007). Unilateral renal agenesis and the congenital solitary functioning kidney: developmental, genetic and clinical perspectives. *BJU Int, 99*(1), 17-21. doi:10.1111/j.1464-410X.2006.06504.x

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L. S., Natale, D. A., Vinayaka, C. R., . . . Barker, W. C. (2004). PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res, 32*(Database issue), D112-114. doi:10.1093/nar/gkh097

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics, 26*(7), 873-881. doi:10.1093/bioinformatics/btq057

Xu, C., Li, C. Y., & Kong, A. N. (2005). Induction of phase I, II and III drug metabolism/transport by xenobiotics. *Arch Pharm Res, 28*(3), 249-268.

Yalavarthy, R., & Parikh, C. R. (2003). Congenital renal agenesis: A review. *Saudi J Kidney Dis Transpl, 14*(3), 336-341.

Yamamoto, K., Higashiura, A., Suzuki, M., Aritake, K., Urade, Y., & Nakagawa, A. (2017). Molecular structure of a prostaglandin D synthase requiring glutathione from the brown planthopper, Nilaparvata lugens. *Biochem Biophys Res Commun, 492*(2), 166-171. doi:10.1016/j.bbrc.2017.08.032

Yamanaka, K., Urano, Y., Takabe, W., Saito, Y., & Noguchi, N. (2014). Induction of apoptosis and necroptosis by 24(S)-hydroxycholesterol is dependent on activity of acyl-CoA:cholesterol acyltransferase 1. *Cell Death Dis, 5*, e990. doi:10.1038/cddis.2013.524

Yang, H., & Shu, Y. (2015). Cadmium transporters in the kidney and cadmium-induced nephrotoxicity. *International Journal of Molecular Sciences, 16*(1), 1484-1494. doi:10.3390/ijms16011484

Yant, L. J., Ran, Q., Rao, L., Van Remmen, H., Shibatani, T., Belter, J. G., . . . Prolla, T. A. (2003). The selenoprotein GPX4 is essential for mouse development and protects from radiation and oxidative damage insults. *Free Radic Biol Med, 34*(4), 496-502.

Ye, Z., Xu, S., Spitze, K., Asselman, J., Jiang, X., Ackerman, M. S., . . . Lynch, M. (2017). A New Reference Genome Assembly for the Microcrustacean Daphnia pulex. *G3 (Bethesda), 7*(5), 1405-1416. doi:10.1534/g3.116.038638

Yedjou, C. G., & Tchounwou, P. B. (2006). OXIDATIVE STRESS IN HUMAN LEUKEMIA (HL-60), HUMAN LIVER CARCINOMA (HepG2), AND HUMAN (JURKAT-T) CELLS EXPOSED TO ARSENIC TRIOXIDE. *Met Ions Biol Med, 9*, 298-303.

Yennamalli, R. M., Rader, A. J., Kenny, A. J., Wolt, J. D., & Sen, T. Z. (2013). Endoglucanases: insights into thermostability for biofuel applications. *Biotechnology for Biofuels, 6*(1), 136. doi:10.1186/1754-6834-6-136

Yin, M., Laforsch, C., Lohr, J. N., & Wolinska, J. (2011). Predator-induced defense makes daphnia more vulnerable to parasites. *Evolution, 65*(5), 1482-1488. doi:10.1111/j.1558-5646.2011.01240.x

Yokoi, K., Uthus, E. O., & Nielsen, F. H. (2003). Nickel deficiency diminishes sperm quantity and movement in rats. *Biol Trace Elem Res, 93*(1-3), 141-154. doi:10.1385/BTER:93:1-3:141

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology, 11*(2). doi:10.1186/gb-2010-11-2-r14

Zaffagnini, F. (1987). Reproduction in Daphnia. *Daphnia, 45*, 245-284.

Zaffagnini, F., & Zeni, C. (1986). Considerations on Some Cytological and Ultrastructural Observations on Fat Cells of Daphnia (Crustacea, Cladocera). *Bolletino di zoologia, 53*(1), 33-39. doi:10.1080/11250008609355480

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics, 16*(3), 172-183. doi:10.1038/nrg3871

Zhang, M., An, C., Gao, Y., Leak, R. K., Chen, J., & Zhang, F. (2013). Emerging roles of Nrf2 and phase II antioxidant enzymes in neuroprotection. *Prog Neurobiol, 100*, 30-47. doi:10.1016/j.pneurobio.2012.09.003

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, *9*(1). doi:10.1371/journal.pone.0078644

Zheng, X., Xu, Z., Qin, G., Wu, H., & Wei, H. (2017). Cadmium exposure on tissue-specific cadmium accumulation and alteration of hemoglobin expression in the 4th-instar larvae of Propsilocerus akamusi (Tokunaga) under laboratory conditions. *Ecotoxicol Environ Saf*, *144*, 187-192. doi:10.1016/j.ecoenv.2017.06.019

Zou, D., Ma, L., Yu, J., & Zhang, Z. (2015). Biological Databases for Human Research. *Genomics, Proteomics & Bioinformatics*, *13*(1), 55-63. doi:10.1016/j.gpb.2015.01.006