Spring 5-15-2018

Investigating the Role of Genomic Variation in Susceptibility to Environmental Chemicals across Populations

Lindsay Adrian Holden
Portland State University

Investigating the Role of Genomic Variation in Susceptibility to Environmental Chemicals across Populations

by

Lindsay Adrian Holden

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Biology

Dissertation Committee:
Kim H. Brown, Chair
Bradley Buckley
Suzanne Estes
Deborah Lutterschmidt
Angela Strecker

Portland State University
2018

Abstract

No two individuals are identical. This is true at the genetic level and at the phenotypic level. One of the traits that varies between populations is toxicant susceptibility: some individuals are sensitive to the effects of environmental chemical exposure, and others are resistant. This body of work aims to address the impact of genomic copy number variants (CNV)—large (>1 Kb) duplications or deletions across the genome—on the toxicant-susceptibility phenotype.

Herein I have characterized copy number variants across three commonly used laboratory strains of zebrafish (*Danio rerio*) and identified mRNA expression phenotypes in the same strains. I found that males and females have only a 14% overlap in differentially expressed mRNA transcripts across three common laboratory strains, congruent with the growing body of work identifying sex- and strain-specific phenotypes in zebrafish. Furthermore, I identified two strain-specific response quantitative trait loci (QTL) that explain about a third of the variation in susceptibility to PCB and tested the response QTL using targeted CRISPR-Cas9 editing of the CNV involved. Overall, this body of work defines CNV and mRNA expression variation across zebrafish strains, identifies CNV causal in the PCB-susceptibility phenotype, and confirms the PCB-susceptibility QTL using targeted genomic editing.

Dedication

To the entire Lady Boss crew: past, present, and future. *Non nobis solum nati sumus*, not unto ourselves alone are we born. Thank you for being my support, guidance, and cheerleaders, especially because I'm not good at sharing my feelings.

Acknowledgments

First and foremost I would like to acknowledge the support and patience of my husband, Adam. Without his kindness and care I would not have been able to survive this journey. Thank you, lovee. Second, I would like to thank my parents for always encouraging me to follow my interests, even if that means that I have a weird job working with monkeys or fish. And an intellectual thank you to my little sister, Kelly, who is blazing a trail down the career track of a professional scientist and inspires me to find ways to continue on my path as a scientist (#sisterswhoscience).

I would like to acknowledge all of the help that I have received while at Portland State University: from the Biology Department staff (we all know you are the real people in charge), the faculty who had meaningful conversations with me about science, career, and life, the undergrads that fed my fish and helped me in the lab, and to my grad student family (especially my past and present lab mates). I thrive most when I feel connected to a community and the Portland State Biology Department grad student community is incredible.

Thank you to my dearest advisory committee—Brad Buckley, Suzanne Estes, Deb Lutterschmidt, and Angela Strecker—for your time and intellectual contributions to making my dissertation excellent. You

made yourselves available to answer my questions and freely contributed your expertise to the guiding and shaping of my dissertation project, analysis, and interpretation. Without your advice, especially in the earlier stages of my PhD, I would not have been able to clarify my vision and the goals of this dissertation project.

Finally, thank you to my advisor, Kim Brown, for allowing me the freedom to develop my dissertation project as opportunities came my way. Experiments are a lot like a river. You can see the path that the river follows and plan your route accordingly, but the path of the river may change and it is out of our control. You can fight the changes and try to pave your way through, or you can accommodate the changes and reroute your path. Part of developing as a scientist is learning when to be rigid and when to bend, so thank you for allowing me to learn these lessons as I go.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Preface

Chapter 3 is published in *Scientific Reports*

Holden, L. A. & Brown, K. H. Baseline mRNA expression differs widely between common laboratory strains of zebrafish. Sci. Rep. 8, 1–10 (2018).

Chapter 4 is under review for publication in *Aquatic Toxicology*

Holden, L.A. and Brown, K.H., Response eQTL analysis of low-dose PCB exposure connects genomic copy number variants to susceptibility.

Chapter 1

An Introduction

Toxicology as a framework

In toxicology, risk is defined as the product of toxicity and exposure, where exposure is comprised of both dose and duration. To determine risk we perform a formal risk assessment, which is a highly regulated process typically pertaining to human and/or environmental health. Currently there are 97 formal guidance documents from the United States Environmental Protection Agency (US EPA) that direct all aspects of human health risk assessment. To determine the human health risk of a compound, regulators follow four main steps of assessment: hazard identification, dose-response assessment, exposure assessment, and risk characterization[1] (Figure 1.1).

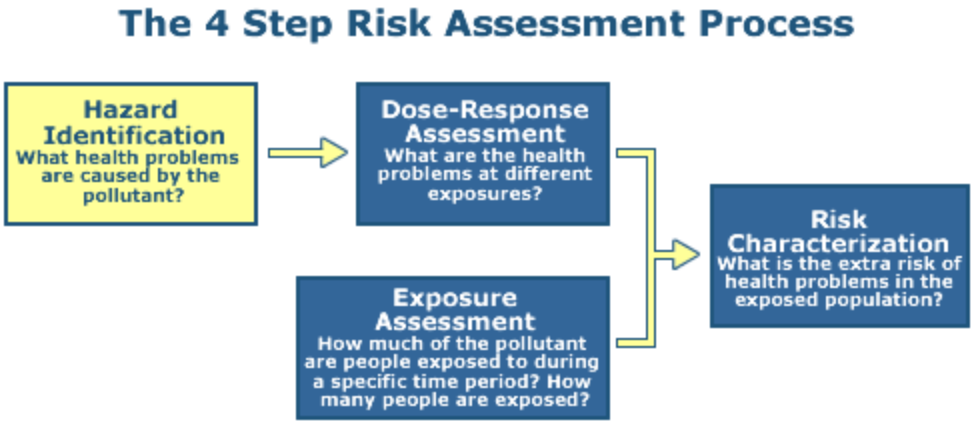


Figure 1.1: The 4 steps of risk assessment. Dose-response and exposure assessment are also collectively known as hazard assessment. Image credit: https://www.epa.gov/risk conducting-human-health-risk-assessment#main-content.

In human health risk assessment, hazard identification is initiated through a literature review process where existing data are assessed for evidence of potential health effects in humans, such as cancer or death. The two key components of hazard identification, toxicokinetics and toxicodynamics, include assessment of compound distribution (i.e., where it goes in the body), compound metabolism or elimination (i.e., how long it stays), and the effects that the chemical has on the body. The US EPA focuses its hazard identification for potential carcinogens on mode of action analysis, where the key chemical, molecular, cellular, and organismal events are delineated and the "weight of evidence" of adverse outcomes at any of the key events resulting in descriptors of the compounds ability to induce carcinogenic effects in humans.

If a compound is deemed potentially hazardous, it is then assessed for a dose-response relationship. Generally, as the dose increases the biological response also increases, but there is a lower limit at which adverse effects are not observed. This theory was originally postulated by Paracelsus in the 15th century, who stated, "solely the dose determines that a thing is not a poison"[2], and manifests in modern toxicology as two dose-response criterial: NOAEL and LOAEL. The NOAEL is the no observable adverse effect level, where no adverse effects are observed at a known concentration and the LOAEL is the lowest observable adverse effect level. Similarly, the benchmark dose is another measurement that involves modeling NOAEL and LOAEL data to predict a single point-of-departure value where the dose induces a response[3].

Exposure assessment goes hand-in-hand with dose-response assessment. At this step the extent of exposure is assessed: who is exposed, at what interface is exposure occurring (skin, lungs, eyes, etc.), and what is the duration of the exposure? Oftentimes it is quite difficult to clearly delineate answers to these questions in humans, so epidemiologic analyses and extrapolations from body burden studies are used to estimate the dose experienced during exposure[4].

The final step in human health risk assessment is risk characterization. This step integrates the conclusions from hazard

identification and hazard assessment (dose-response plus exposure) into an overarching conclusion about the risk of the compound. The final risk characterization is used by regulators and policy makers to direct public health using data that support the extent and direction of human health outcomes following exposure[5]. Risk characterization ultimately shapes policy that directs the prioritization of legacy chemical clean-up (e.g., PCB) and the introduction of new chemicals in commercial application.

One of the most crucial aspects of risk assessment is setting a reference dose for safe levels of exposure. Reference doses are calculated using a point-of-departure estimate, such as NOAEL, LOAEL, or benchmark dose, and applying uncertainty factors and modifying factors (Equation 1.1).

Equation 1.1: $Reference\ dose = \frac{NOAEL}{uncertainty\ factor}$

Uncertainty factors are determined on a case-by-case basis, but consist of four main components with values ranging from 0.1-10: human variability, animal to human extrapolation, sub-chronic data (i.e., acute), and point-of-departure uncertainty. Additionally a modifying factor of up to a value of 10 can be assigned for the level of completeness in the dataset used in the hazard assessment. Let's look

at an example using an acute exposure study in rats to determine a reference dose. In this study the authors determined a LOAEL of 0.1 mg/kg/day for their endpoint. To calculate a reference dose we would take the LOAEL and divide it by the product of the uncertainty factors (Equation 1.2).

Equation 1.2:

$$Reference\ dose = \frac{LOAEL}{UF_{human} * UF_{interspecies} * UF_{sub-chronic} * UF_{Point\ of\ departure} * MF_{database}}$$

The uncertainty factors for human variation (UF_{human}) and animal to human extrapolation ($UF_{interspecies}$) can both be split into toxicodynamic and toxicokinetic subsets, but in this case we don't have any information on either uncertainty factor, so their values will both be assigned the maximum value of 10. Because our rat study was not a chronic study we need to include a sub-chronic uncertainty factor of 10 ($UF_{sub\text{-}chronic}$) and because we used a LOAEL value and not a NOAEL value (i.e., there was no NOAEL identified in our rat study) we also need to include a point-of-departure uncertainty factor of 10 ($UF_{point\ of\ departure}$). Finally, this is the only study that has been performed on our compound of interest, so the breadth of data is extremely lacking and we need to assign a modifying factor of 10 ($MF_{database}$). In this case our final

reference dose would be 1 ng/kg/day in humans (see worked example). So at 1 ng/kg/day we would not expect to see any adverse health outcomes in humans. This is an extremely conservative estimate and may not accurately reflect the true biology and toxicology of the compound.

Worked example:

$$Reference\ dose = \frac{0.1\frac{mg}{kg}/day}{10*10*10*10*10} = \frac{0.1\frac{mg}{kg}/day}{100{,}000} = 1\frac{ng}{kg}/day$$

One of the areas in which we can refine our uncertainties when determining reference dose is human variation. Toxicodynamic and toxicokinetic properties are both influenced by genetic variation. A study investigating the toxicity and efficacy of methotrexate for the treatment of psoriasis found genetic variants that caused changes in the number of receptors for and transporters of the drug, which directly affected the dynamics of the system[6]. Genetic polymorphisms in many isoforms of cytochrome P450, a family of metabolic enzymes responsible for xenobiotic metabolism, have been identified and linked to altered metabolism of many therapeutic compounds[7] (e.g., kinetics). Understanding the normal variation that exists in humans and how it affects the dynamics and kinetics of exposure would be a huge

advancement that could be directly translated into reference dose calculations and the overall risk assessment procedure.

Harnessing human variation

In the last few decades huge strides have been made in human genetics. In 2001 the first draft of the human genome was released[8] and as of 2017 the human reference genome is now in its 38th release (GRCh38) and contains alternate loci representing significant variation in 178 regions[9]. Large scale efforts to identify global genetic variation began with the HapMap Project in 2003, which focused on identifying single nucleotide polymorphisms (SNPs) that were inherited as blocks[10]. The HapMap Project was built on the foundation that many SNPs were observed to exist in linkage disequilibrium and inherited as specific haplotypes. Identification and classification of haplotypes would decrease the number of SNPs required to identify variant regions specific to diverse populations and facilitate discovery of gene-disease associations. In 2005 the first haplotype map of the human genome was released and contained over 1 million SNPs[11]. In 2007 a second generation haplotype map containing 3.1 million SNPs was released[12] and in 2010 the International HapMap Consortium expanded their dataset from 270 individuals from 4 global populations to 1,184 individuals from 11 global populations to sharpen the resolution on rare

variants and included the addition of genomic copy number variants (CNVs)[13].

As understanding of variation in the human genome grew and the limitations of small sample sizes (<300 individuals) to detect rare variants became apparent, a new large-scale genomic variation project began. In 2008 the 1000 Genomes Project aimed to sequence at least 1000 individuals to investigate variants occurring in at least 1% of the population with coverage of genic variants found at 0.1% of the population[14]. In 2015 the 1000 Genomes Project had sequenced 2,504 people across 5 continents and 26 populations and found that the typical genome differs from the human reference genome at 4.1-5.0 million sites and variants differ greatly among populations[15]. With this new level of population variation, the 1000 Genomes Project established that individual genomes contain 2,100-2,500 structural variants that affect 4-5 times as many nucleotides as SNPs and short insertion-deletions (indels). Individuals harbor 18.4 Mb of structural variants per diploid genome (8.9 Mb per haploid genome), largely comprised of multiallelic CNV and biallelic deletions[16].

CNV have shaped human diversity on the evolutionary scale by imparting selective advantages or disadvantages[17] through alteration of gene expression by direct interaction (overlap with a gene) or indirect regulatory mechanisms[18]. Moreover, most structural variants that alter

gene expression do so through enhancers and other regulatory elements (88.3%), not through direct interaction with gene-coding regions[19]. In human health, CNV are associated with Mendelian diseases (e.g., Charcot-Marie Tooth neuropathy[20] and Williams-Beuren syndrome[21]), complex diseases (e.g., diabetes[22] and psoriasis[23]), and non-pathogenic phenotypes (e.g., salivary amylase production[24]). Additionally, CNV cause pharmacogenomic phenotypes[25] where variable copy number across xenobiotic metabolism genes alter the rate of metabolism (pharmacokinetics) and if CNV interact with transporters or receptors they can alter biological activity (pharmacodynamics).

In an effort to understand the effect of human variation on toxicity and refine the human variation uncertainty factor in reference dose determination, 179 chemicals were screened for cytotoxicity in lymphoblastoid cell lines from 1,086 individuals from 9 populations across 5 continents sequenced by the 1000 Genomes Project[26]. In this study about half of the tested compounds had a range of toxicity that would be captured by the $10^{½}$ uncertainty factor for interindividual toxicodynamic variability when calculating reference doses. A portion of the tested compounds had interindividual variation greater than a factor of 10, indicating that the uncertainty factors are woefully inadequate for some chemicals (Figure 1.2). Unfortunately, one of the weaknesses of lymphoblastoid cell lines is CNV artifacts due to differences in replication

timing relative to primary cell lines[27]. This makes extrapolation of copy number effects between systems extremely difficult.

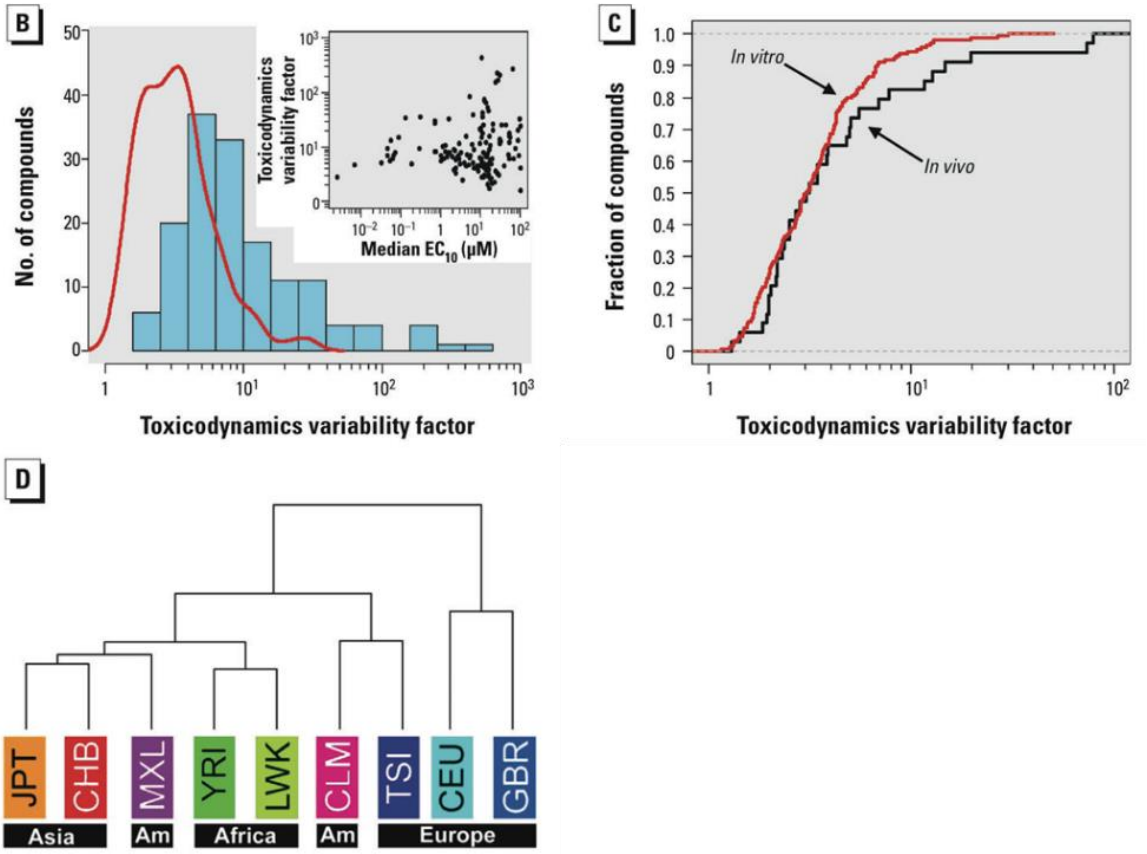


Figure 1.2: B) Histogram of the toxicodynamic variability factor $10^{(q50-q01)}$ for 149 compounds across 1,086 cell lines. The inset shows the relationship between range and median estimated EC_{10} for each chemical. C) Cumulative distribution functions for the in vitro toxicodynamic variability factor shrunken to account for technical variability across 149 compounds and the human in vivo toxicodynamic variability factors across 34 compounds[28]. D) Hierarchical clustering for the 179-length profiles of mean EC_{10}, computed within each population, and shown by continental ancestral origin of the population. AM, Americas. Image and description adapted from work by Abdo et al., 2015[26].

Incorporating human variability is the next great challenge in toxicology. There is clear evidence that the current practice of generic

uncertainty factors in reference dose calculations are inadequate, but our knowledgebase of the driving factors behind interindividual variability is also inadequate. Toxicity clades out by distinct genetic populations (Figure 2c), indicating that toxicodynamic and toxicokinetic phenotypes may be shared by genetically similar groups. It is an extremely complex challenge to study this phenomenon in humans due to uncontrollable confounding factors such as socio-economic and health status. Other systems, including model organisms, may be the answer to delineating the myriad factors involved in interindividual variation and population-based variation, including the role that CNV play in toxicity.

Evidence of resistance to toxic chemicals in wild populations

Atlantic killifish (*Fundulus heteroclitus*) are estuarine fish found along the coast of the Eastern United States. These fish have adapted to local anthropogenic contamination in several locations such as Newark Bay (New Jersey)[29], the Elizabeth River (Virginia)[30], New Bedford Harbor (Massachusetts)[31], and the Hudson River (New York)[32]. In each location, high levels of aryl hydrocarbon mixtures (largely polychlorinated biphenyls, PCBs) are present in the sediments and are generally toxic to resident organisms. At these locations, however, Atlantic killifish have adapted to be resistant to high levels of pollution.

Different mechanisms have been tied to the resistant phenotype in Atlantic killifish. Generally, tolerance is associated with a blockade of the aryl hydrocarbon receptor (AHR) signaling pathway[33]. Aryl hydrocarbons, such as PCBs, impart toxicity through a highly conserved AHR signaling cascade[34] (Figure 1.3). Prior to ligand binding, AHR exists in the cytosol bound to several chaperone proteins such as heat shock protein 90 (HSP90), p23, and AHR-interacting protein (AIP). After binding to a ligand, the AHR complex translocates to the nucleus where it dissociates with the chaperone proteins and binds to AHR nuclear translocator protein (ARNT). The AHR-ARNT complex then binds directly with DNA at xenobiotic response elements (XREs) and induces transcription of a suite of genes, including cytochrome P450 1A (CYP1A) (reviewed in[35]). CYP1A and other cytochrome P450 proteins are responsible for xenobiotic metabolism and both the parent compound (AHR ligand) and its metabolites can exert toxic effects.

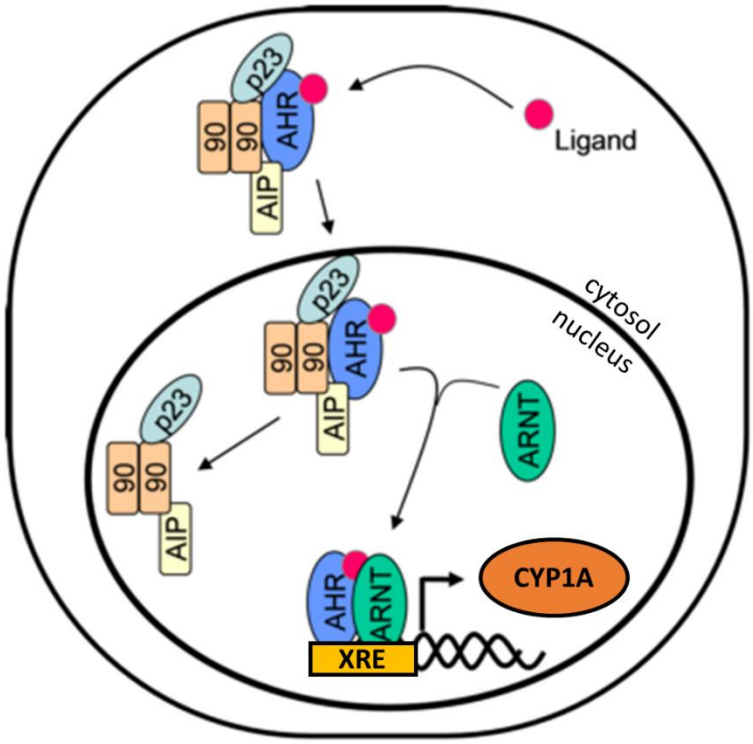


Figure 1.3: The AHR signaling cascade. After binding with a ligand, AHR translocates to the nucleus, directly binds to the DNA at xenobiotic response elements, and induces translation of a suite of genes, including CYP1A. Adapted from Hall, 2014[34].

The genetic component of the toxicant susceptibility phenotype in killifish was tested by a common-garden experiment in which tolerant populations were reared in a clean environment for two generations to isolate the heritable component of tolerance, and then challenged with known concentrations of a specific PCB congener (PCB-126)[36]. The study found that tolerance was heritable for up to two generations indicating that the phenotype is genetically based. Further work identified adaptive

selection of SNPs across AHR genes—which are known mediators of PCB-126 toxicity in killifish[37]—in the New Bedford Harbor population[38].

The Atlantic Tomcod (*Microgadus tomcod*) is another example of local adaptation to highly contaminated sites. Exposure to two aryl hydrocarbons, benzo[a]pyrene and PCB-77, does not induce CYP1A mRNA expression in fish from the contaminated site, but fish from a nearby clean site have robust transcriptional responses to the same exposure[32] (Figure 1.4). The mechanism behind this resistance has been identified as a 6 basepair deletion in the AHR2 gene that results in a two amino acid deletion in the mature protein and is highly penetrant in populations at polluted sites[39].

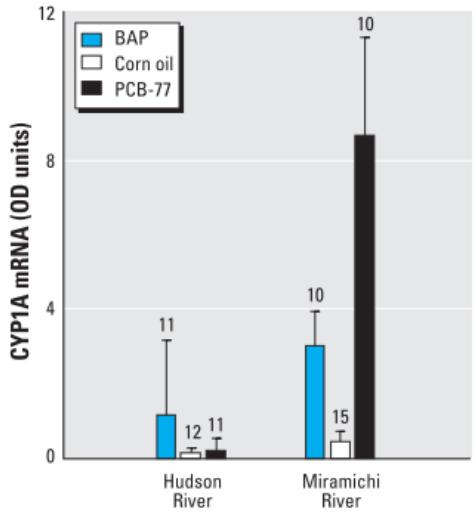


Figure 1.4: CYP1A mRNA expression levels (mean and 95% CIs, expressed in OD units) in juvenile tomcod from a contaminated (Hudson River) and clean (Miramichi River) site injected intraperitoneally with 10 ppm Benzo[a]pyrene, corn oil vehicle, or 10 ppm PCB-77. Numbers above bars represent sample size. Image and description adapted from Yuan et al., 2006[32].

Repeated observations of toxicity-resistance phenotypes in multiple species are strong evidence of the adaptive advantage that these populations have in polluted environments. The genome of the Atlantic killifish was assembled in 2015 and a linkage map was published shortly thereafter, facilitating a deeper understanding of the genetic mechanisms behind the resistance phenotype. The genetic linkage map identified 24 linkage groups (putative chromosomes) and supports a high degree of synteny between killifish and medaka, with slightly less synteny between killifish and zebrafish[40]. Although zebrafish are less syntenic with killifish than medaka, they still share several quantitative trait loci (QTL) identified as the genetic basis for aryl hydrocarbon resistance in killifish[41].

Zebrafish as a model system to study the effects of genetic variation on toxicity

The Atlantic killifish and Atlantic tomcod are excellent examples of repeated evolution of a toxicant-resistance phenotype through convergent genetic mechanisms. However, the complete picture of the genetics driving this phenotype is unclear due to unexplained variance and a lack of genomic tools to discover the cause of the variance. Herein

lies the strength of the well-developed model organism, the zebrafish (*Danio rerio*).

Zebrafish were first used as a model system in the 1940s and following the establishment of several laboratory strains in the 1990s, their popularity has exploded (Figure 1.5). Since the mid-2000s over 1000 studies using zebrafish are published every year (searchable in NCBI PubMed using “zebrafish” or “Danio rerio” keywords) and that number has only continued to grow. In 2017 the number of zebrafish publications hit its current peak of 2940.

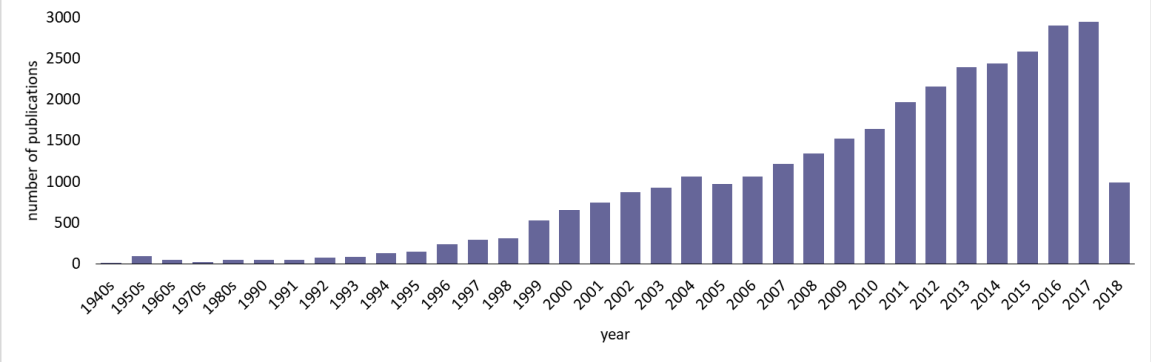


Figure 1.5: Publication count per year, as indexed by NCBI's PubMed using "zebrafish" or "Danio rerio" as keywords.

The rise of zebrafish as a model system is due to several factors. The zebrafish shares basic body design with other vertebrates. Zebrafish reach sexual maturity by 3 months, which facilitates genetic crosses and multi-generational studies. Housing and husbandry are relatively cheap and easy. External fertilization, large clutches of eggs (averaging ~200

embryos per spawning event), clear bodies until juvenile stage, and rapid development (primary organogenesis begins at 10 hours and is complete by 48 hours[42]) make studying early development easier than mammalian systems. The zebrafish also boasts a completely sequenced genome with 71% of zebrafish genes having at least one homolog in the human genome[43]. With a sequenced genome comes a complete set of genetic tools such as morpholinos, CRISPR-Cas9, and transgenic strains.

Additionally, zebrafish have a toxicant-resistance phenotype, similar to the phenotype observed in Atlantic killifish and Atlantic tomcod. Zebrafish larvae exposed to PCB-126 have ranges of developmental toxicity between 9 and 336 ppb across genetically distinct laboratory strains[44] (Figure 1.6). This range of interstrain variation exceeds a factor of 10, such as is used in reference dose assessment, and can serve as a malleable laboratory model of toxicity variation across populations. Although AHR2 was identified as one of the genetic drivers of the resistance phenotype, only 24% of the phenotypic variance could be explained by QTL, leaving a large gap in our understanding of the full effects of genomic variation on toxicant susceptibility.

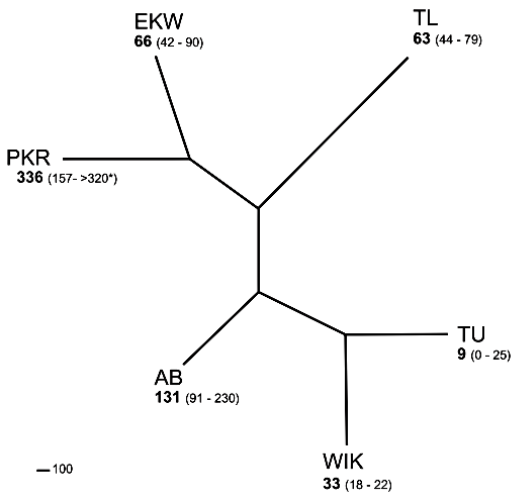


Figure 1.6: Median effective concentration (EC50) of early-life stage toxicity (abnormal looping of the heart, pericardial and yolk-sac edema, reduced heart rate, impaired swim bladder inflation, and craniofacial malformations). Branch lengths reflect genetic distance; numbers in parentheses represent 95% confidence intervals. Image and description adapted from Waits and Nebert, 2011[44].

There is a fair amount of information on the genetic variation between common laboratory strains in zebrafish. Most of the focus has been on SNP variants between strains. In fact, zebrafish strains clade out with high support using SNP markers with over 37.9 million SNPs currently described across the 1.5 Gb zebrafish genome[45]. Beyond SNPs, there is also a detailed analysis of intrastrain CNV across three common laboratory strains and one wild strain of zebrafish. Across strains, there are approximately 1900-3400 CNV, of which only about 500 are shared across all strains, indicating that strains have their own unique set of CNV[46], similar to the case with distinct human populations[16]. These factors prime the zebrafish to be a uniquely helpful tool for basic

research into the genetic mechanisms driving complex phenotypes like toxicant susceptibility.

To test the likelihood that CNV play a role in toxicant susceptibility, I performed a basic assessment of direct and near-neighbor overlap of CNV in zebrafish with a thorough list of 70 genes that are differentially expressed in the toxicant-resistant Atlantic killifish phenotype relative to toxicant-sensitive populations[47]. Using NCBI's blastn alignment tool[48] to assign zebrafish gene homologs (closest sequence match to the nr/nt database with an e-value > 1×10^{-10}), I found 9 genes (12.9%) with directly overlapping CNV and 34 genes (48.6%) with a CNV within 100 Kb up- or downstream in zebrafish (Supplemental data 1.1). These direct and near-neighbor hits indicate that CNV are likely interacting with transcriptional responses that drive the toxicant-susceptibility phenotype.

A brief history of PCBs

Polychlorinated biphenyls are a large class of aryl hydrocarbons comprised of 209 congeners with varying levels of chlorination of a biphenyl molecule (two connected benzene rings, Figure 1.7). PCBs are characteristically hydrophobic and lipophilic, have low vapor pressures, and are resistant to chemical reactions (including degradation)[49]. As the number of chlorine molecules increases, the stability of the compound

increases. These properties make PCBs excellent coolants, flame retardants, and plasticizers, but also result in long half-lives and bioaccumulation in the food web.

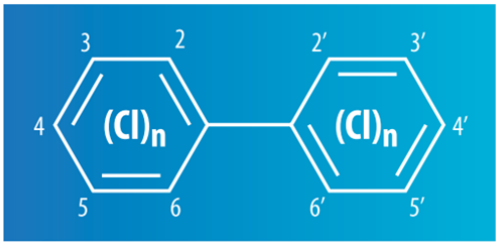


Figure 1.7: Chemical structure of PCB. Image from Crinnion, 2011[50].

There are two general classes of PCB congeners determined by the position and number of chlorines: nonplanar (mono-*ortho*-substituted) and coplanar (non-*ortho*-substituted). Coplanar congeners exert toxic dioxin-like effects by binding and activating the AHR signaling cascade[51,52]. In 1979 the US EPA banned the production and use of PCBs following evidence that bioaccumulation in birds caused brittle shells through aberrant calcium metabolism, a large-scale poisoning incident in Yusho, Japan where over 1000 people consumed food cooked in contaminated cooking oil, and a forced cull of thousands of chickens and eggs that were fed contaminated feed (as reviewed in[53]).

The World Health Organization developed a comparative scale to assess the toxicity of dioxin and dioxin-like compounds called a Toxic Equivalency Factor (TEF). Dioxin (2,3,7,8-TCDD) is set at the maximum TEF of 1 and related compounds are assigned a TEF relative to dioxin after review of published toxicity data[54]. The most toxic PCB listed in the TEF documentation is PCB-126 (3,3',4,4',5-pentachlorobiphenyl) with a TEF of 0.1. Other coplanar (non-*ortho*-substituted) PCBs have TEF values between 0.03-0.0001, while nonplanar (mono-*ortho*-substituted) PCBs have TEF values of 0.00003. PCB-126 is heavily chlorinated (with 5 chlorines) and has a half-life on the order of 20-30 years in adults[55].

The total body burden of PCBs in humans has steadily declined since the ban in 1979, but a national census of human exposure to environmental chemicals by the US Centers for Disease Control and Prevention (CDC) found 34 of 38 PCB congeners in virtually all people tested[56]. The current body burden of PCB-126 in the average American is 100 ppq in the serum (16.3 pg PCB-126/g lipid). *In utero* exposure to PCBs can result in immune deficiency and neurological deficits and dietary exposure (the most common route of exposure) can cause reproductive problems in men and women, hypothyroidism, and increases in the risk of type II diabetes, lung cancer, and liver cancer (as reviewed in[50]).

Because of its highly toxic effects, long-half life, and bioaccumulative properties, PCB-126 serves as a good proxy for the effects of total PCB burdens (from mixtures of multiple PCB congeners with assumed additive effects) in controlled studies. There is extensive documentation of the developmental effects of PCB-126 exposure across multiple fish species, including zebrafish. In zebrafish, developmental toxicity is largely characterized by pericardial and yolksac edema, delay of swim bladder inflation, elongated and/or unlooped heart (abnormal cardiac development), reduced heart rate, and malformation of the spine[57–59]. These physical manifestations of PCB-126 toxicity are easily observable in controlled laboratory studies. The combination of a clear developmental toxicity paradigm for PCB-126 and well-characterized genomic variation make the zebrafish the ideal system to assess the influence of genetic variation on susceptibility to toxicants.

Rationale for chapters

Our current knowledge of CNV in zebrafish was built on within-strain (intrastrain) comparisons. While this allows some comparison of shared CNV across strains, it does not fully capture the interstrain variability in zebrafish. To improve upon this I have assessed three common laboratory strains of zebrafish using a reciprocal comparison study design to maximize identification of interstrain variation. I present

this work in Chapter 2 “An Interrogation of Shared and Unique Copy Number Variants across Genetically Distinct Strains of Zebrafish”. Herein I present a set of 1351 CNV that vary across strains and test a set of ten identified CNV using multiple molecular methods (quantitative PCR (qPCR) and long amplification standard PCR). I also present that CNV that directly overlap a gene or fall within 5 Kb up- or downstream of a gene are likely to cause variation across introns, coding sequence, or untranslated regions at the start or end of genes.

Because we know that there is wide genetic variation across zebrafish strains and I have predicted transcriptional effects of CNV that vary across strains, we can hypothesize that there is also standing variation in gene expression across strains. In Chapter 3 “Baseline mRNA expression differs widely between common laboratory strains of zebrafish” I characterize the normal transcriptional profiles in the same set of commonly used zebrafish strains using microarrays. Surprisingly, I found large differences between males and females, as well as large differences between strains. A total of 421 unique mRNA transcripts were significantly differentially expressed across strains. This study sets the stage for further exploration of phenotypes affected by genetic and transcriptional variation.

In Chapter 4 “Response eQTL analysis of low-dose PCB exposure connects genomic copy number variants to susceptibility” I explore the

relationship between genomic CNV and the PCB-susceptibility phenotype. Using matched CNV genotypes and PCB-induced gene expression phenotypes, I performed an extensive expression QTL analysis to identify CNV drivers of the phenotype. After mapping phenotypes of exposed and unexposed zebrafish to CNV, I found two response QTL (eQTL responsive to PCB exposure) that are strain-specific.

To test the reQTL that I identified as drivers of the PCB-susceptibility phenotype, I performed a functional study using CRISPR-Cas9 to selectively edit the CNV in the two reQTL. In Chapter 5 "Targeted CRISPR-Cas9 Editing of Genomic Copy Number Modulates PCB-Susceptibility Phenotype" I successfully target and edit the CNV regions of both reQTL and show a reversion in phenotype where a PCB-resistant strain becomes extremely sensitive and a PCB-sensitive strain becomes slightly more resistant. This serves as proof-of-principle that CNV play a role in modulation of toxicant susceptibility across populations.

Chapter 2

An Interrogation of Shared and Unique Copy Number Variants across Genetically Distinct Zebrafish Strains

Abstract

Zebrafish (*Danio rerio*) are a widely utilized model system for human disorders, but common laboratory strains have distinct behavioral and physiological differences. Accompanying these known strain differences, commonly used "wildtype" zebrafish strains have both shared and unique suites of single nucleotide polymorphisms and copy number variants (CNV). Despite this, genomic variation is often ignored in study design and the strain used is often not mentioned. The goal of this study is to assess CNV across three common laboratory strains of zebrafish—AB, Tubingen (TU), and WIK—and provide this dataset as a tool for the zebrafish community. Herein we identify 1351 CNV regions within the most recent genome assembly (GRCz11) covering 1.9% of the zebrafish genome (31.7 Mb). CNV were found across all chromosomes and 2200 genes (5121 transcripts) lie within ± 5 Kb of identified CNV, pointing to likely cis regulatory actions of CNV on nearby gene neighbors. We have created a Public Session accessible on the UCSC Genome Browser to view CNV from this study titled

"danRer11 zebrafish CNV across strains" as a tool to view CNV for the zebrafish community.

Introduction

Zebrafish are an important genetic model, but the acknowledgement and incorporation of genomic variation across common lab strains into study design has been slow. It is well established that zebrafish strains contain many shared and unique single nucleotide polymorphisms (SNPs)[45,60–62] and copy number variants (CNVs)[46,63] with several groups diligently working to describe and characterize differences between strains at the genotypic and phenotypic levels. The phenotypic effects of these genetic differences are not well understood and likely manifest as observable, but as-of-yet unidentified, variation between strains. To this end, several studies have described behavioral and physiological differences between strains such as differences in sex-determination[64], fear-related behavior[65], social preference[66], stress[67], susceptibility to toxicants[44], locomotion[68], and growth performance[69].

CNV, which cover 5-10% of the human genome[70], can directly and indirectly affect gene expression via gene dosing[71] and principally act through enhancers and other regulatory elements[19]. CNV can be positively or negatively correlated with gene expression[72] and result in

wide-ranging phenotypic effects like testosterone metabolism[73], the ability to digest starch[24], or complex diseases such as autism and schizophrenia (reviewed in[74,75]). As zebrafish continue to grow as an important model system for basic research, it is imperative to expand our knowledge of the genomic variation within the species. Previous CNV identification in zebrafish focused on within-strain (intrastrain) variation. This study aims to describe zebrafish genomic copy number variation across commonly used laboratory strains (interstrain variation). By providing these data to the global zebrafish community, we hope to highlight the important role that copy number variation has on phenotypes across strains in support of incorporating this information into study design and publication. This is of critical importance to the zebrafish community because clearly defining genomic variation will result in better replication and translation of our research into other model systems, for human health applications, and for application to ecological systems.

Materials and Methods

Animal care and husbandry

All zebrafish husbandry and experimental procedures were performed following protocols approved by Portland State University's Institution Animal Care and Use Committee in accordance with the

National Institutes of Health Guidelines for Care and Use of Laboratory Animals and the Public Health Service Policy on Humane Care and Use of Laboratory Animals. Zebrafish were housed on an Aquaneering semi-recirculating housing system at a density of 5 individuals per liter with 10% daily water changes. Water temperature was maintained at 27.5°C and fish were kept on a 16 hour light, 8 hour dark photoperiod. pH and conductivity were maintained at approximately 7.4 and 1100 uS, respectively. Zebrafish were fed commercial flake food twice daily and supplemented with live *Artemia* and rotifers. This study used 3 strains of zebrafish: AB, Tuebingen (TU), and WIK.

aCGH data analysis

Microarray data were obtained from the National Center for Biotechnology Information (NCBI) BioProject portal (GEO Sample IDs: GSM839719, GSM839720, GSM839721)[46]. Array data are comprised of pooled DNA from 10 individuals each of AB, TU, and WIK strains run in a reciprocal design (AB vs TU, WIK vs TU, AB vs WIK) on custom-designed Aligent Technologies SurePrint GS CGH microarrays. Arrays were designed against the zebrafish Zv8/danRer6 reference genome and had an average probe spacing of 1.4 kb. Copy number variants (CNV) were called using normalized signal intensity files within Nexus Copy Number software (version 5.1; BioDiscovery) and reported as $\log_2$

ratios. Log_2 ratios with a three-probe running average greater than 2 were identified as CNVs[46]. Microarray probe chromosomal locations were updated to the most recent zebrafish reference genome using the LiftOver tool in UCSC Genome Browser[76] (Zv8 → Zv9 → GRCz10 → GRCz11).

qPCR and standard PCR

Quantitative PCR (qPCR) was performed across 10 regions identified as having a CNV using primers designed on Primer3Plus software[77]. DNA was isolated via standard phenol:chloroform extraction for 10 individuals from each strain and assayed across the 10 regions in triplicate on a 364-well plate format using Power SYBR Green Master Mix (Applied Biosystems). Fluorescence was measured on an Applied Biosystems 7900HT Real-Time PCR System. The qPCR cycling protocol included preliminary dissociation (10 minutes at 95°C) and 35 cycles of annealing and extension (95°C for 15 seconds, 60°C for 30 seconds), per manufacturer's protocol. A dissociation melt curve was also obtained to confirm single PCR products. Quantification of PCR product was performed using the ΔΔCt method[78] with an ultra-conserved element (UCE) as a standardized DNA copy number reference sequence[46] and pooled DNA from AB, TU, or WIK strains as reference for each strain. Specifically, ΔCT = target - UCE and $\Delta\Delta CT = \Delta CT_{individual} - \Delta CT_{pool}$.

Pairwise comparisons were made between each strain and fold change was calculated as the inverse log2 of ΔΔCT (or $2^{-\Delta\Delta CT}$).

Standard PCR was performed across 3 CNV regions confirmed by qPCR as an additional confirmation technique and further resolution of the region using Hot Start Taq 2X Master Mix (New England BioLabs) following the manufacturer's protocol. DNA from 3 individuals per strain was extracted using a DNeasy Blood and Tissue kit (Qiagen). PCR products from 3 or 4 regions across each CNV were run on 1% agarose gels with Gel Red nuclear binding stain in 0.5X TBE at 110 volts for 45 minutes and visualized on a digital gel imager. Bands were scored as present/absent and approximate size was noted. All qPCR and standard PCR primers, amplicon sizes, and locations are listed in Supplementary Table 2.1.

Predicting effects of CNV

Consequences of identified CNV were predicted using the Ensembl Variant Effect Predictor[79]. Briefly, effects of CNV were predicted for all RefSeq genes and transcripts falling within a very conservative zone of 5 kb up- or downstream of the CNV location. CNV calls from the GRCz10 assembly were used for this analysis as this is the most recent genome version available for use with the tool.

Results

Across the reciprocal comparisons of pooled DNA from three zebrafish strains, we identified 1941 CNV regions in the Zv8 genome (Supplementary Table 2.2). Stepwise LiftOver to GRCz11 resulted in the loss of CNV calls due to a split across the region or partial deletion in newer versions of the reference genome. The largest loss in CNV calls resulted from the Zv8 to Zv9 LiftOver because of a major genome update. Zv9 to GRCz10 and GRCv10 to GRCz11 LiftOvers also resulted in the loss of some, but fewer, CNV calls. The final count of GRCz11 CNV calls was 1351 (Table 2.1) and the identified CNV regions non-redundantly cover 1.9% of the zebrafish genome (31.7 Mb).

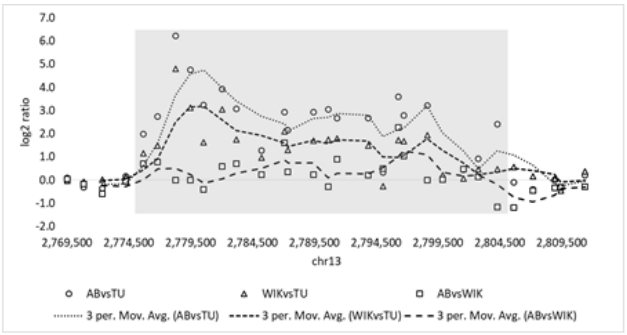
Table 2.1: Summary of copy number count and type across four versions of the zebrafish reference genome.

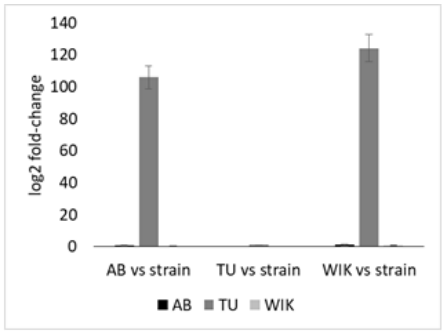
	Zv8	Zv9	GRCz10	GRCz11
High Copy Gain	432	375	355	350
Copy Number Gain	1036	731	631	626
Copy Number Loss	154	128	116	112
Homozygous Copy Loss	319	291	265	263
Partially deleted in new	--	5	4	3
Split in new	--	411	154	13

We chose 10 random CNV regions to confirm by qPCR and further interrogate the putative effects of those CNV on the organism (Figures 2.1 and 2.2). Of the 10 regions, 8 were fully confirmed by qPCR. Three

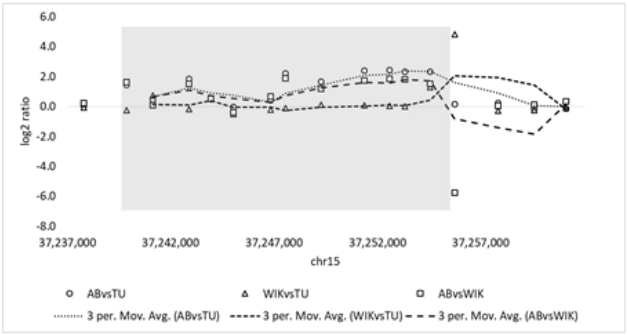
regions showed a loss in WIK (CNV_0311, CNV_0437, CNV_0968), two regions showed a gain in TU (CNV_0222, CNV_0900), two regions showed a loss in AB (CNV_0559, CNV_1736), and one region showed a loss in TU (CNV_0663). The data for one region (CNV_0302) were in disagreement; the array data and CNV call identified a loss in AB while the qPCR data showed a gain in TU. CNV_0572 was predicted to be a loss in WIK from array data, but qPCR failed to show any differences between strains. We further interrogated 3 of these regions by PCR to obtain a high resolution understanding of the loci (Figure 2.3) and were able to confirm the CNV across sub-regions approximately 3.7 Kb in size for CNV_0222, CNV_0311, and CNV_0900. Not all CNV are fully penetrant, as can be seen in CNV_0900.4 which as a gain in WIK across the sub-region (in addition to the gain in TU).

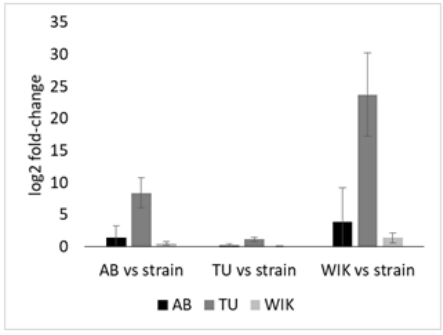
CNV_0222



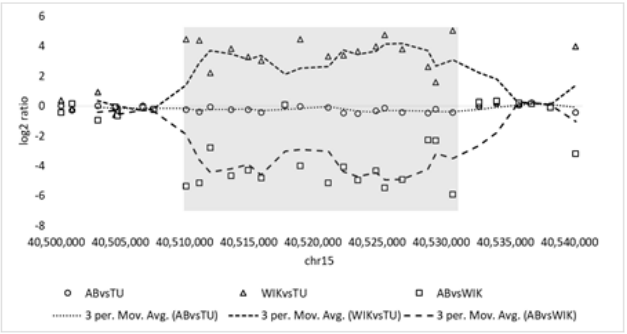


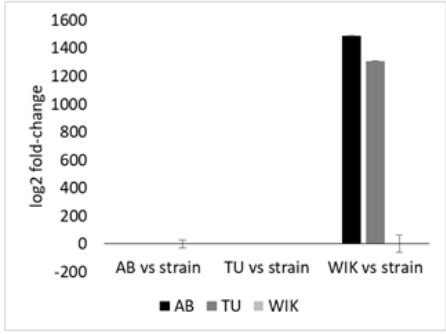
CNV_0302



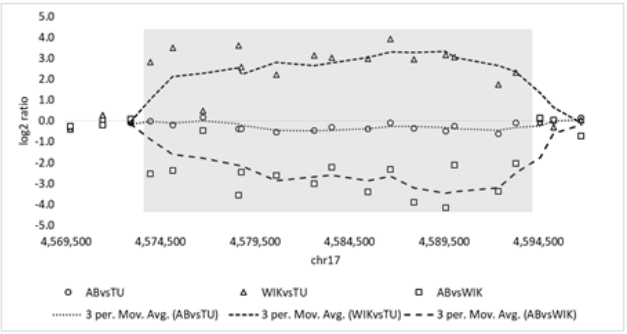


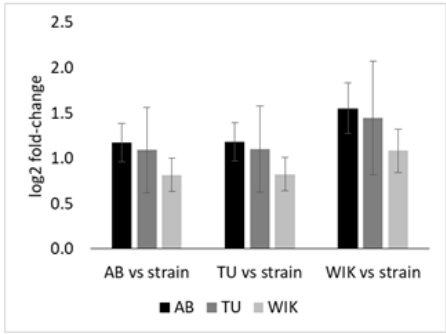
CNV_0311



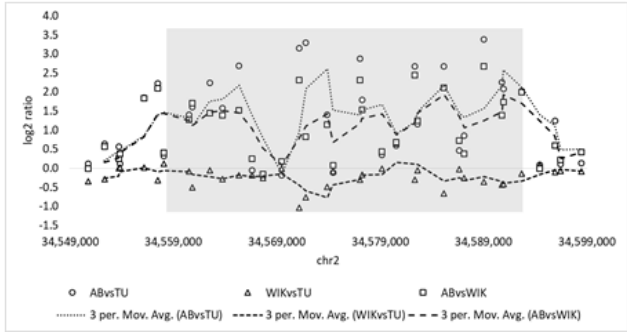


CNV_0437





CNV_0559



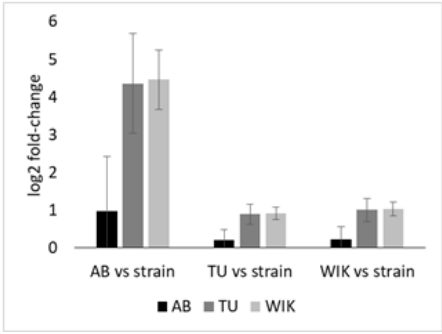
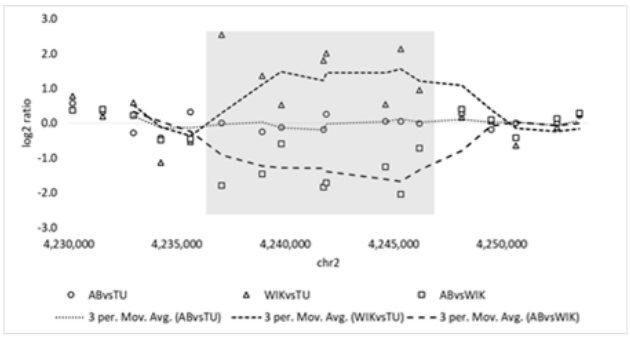
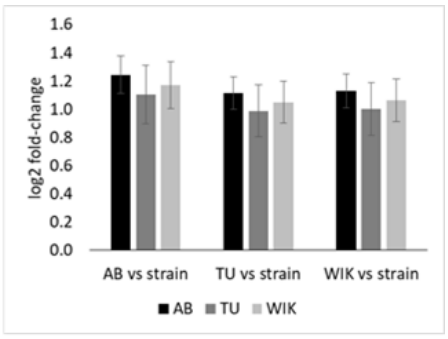


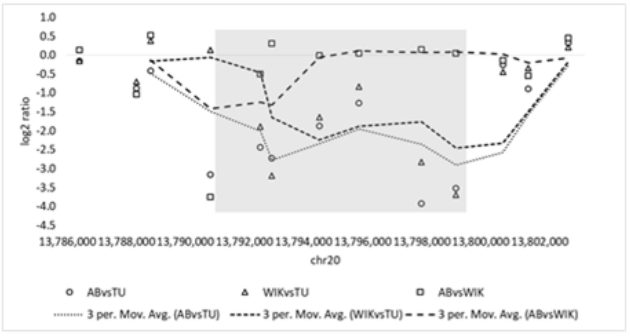
Figure 2.1: log2 ratios of microarray probes across chromosomal locations identified as CNV with corresponding qPCR log2 fold-change values. CNV_0222 = gain in TU, CNV_0302 = disagreement between CNV (loss in AB) and qPCR (gain in TU), CNV_0311 = loss in WIK, CNV_0437 = loss in WIK, CNV_0559 = loss in AB. Grey regions on log2 ratio plots indicate location of CNV. Error bars represent standard deviation and large error bars indicate variation within the strain (i.e., the loss or gain is not fully penetrant).

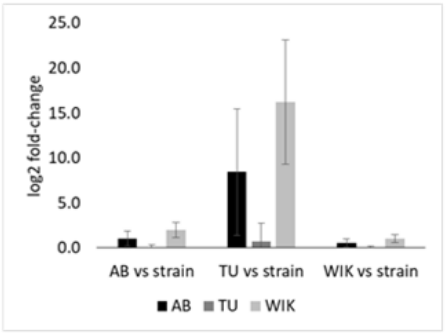
CNV_0572



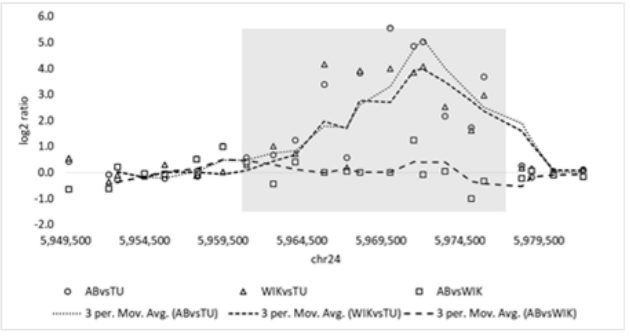


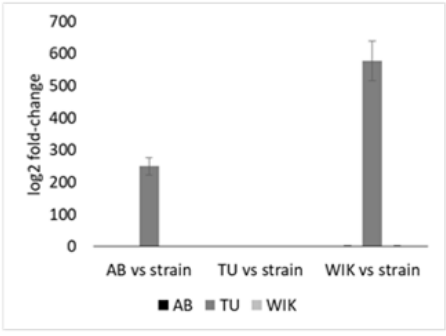
CNV_0663



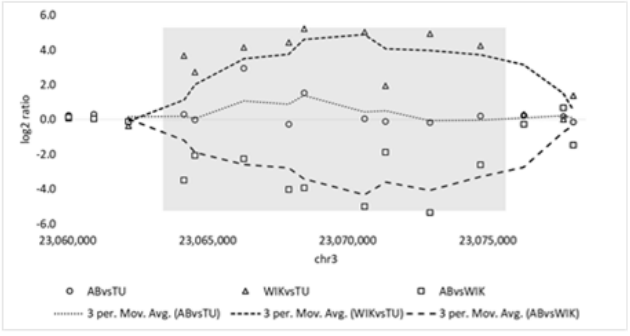


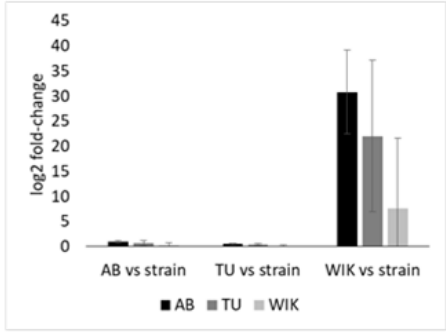
CNV_0900



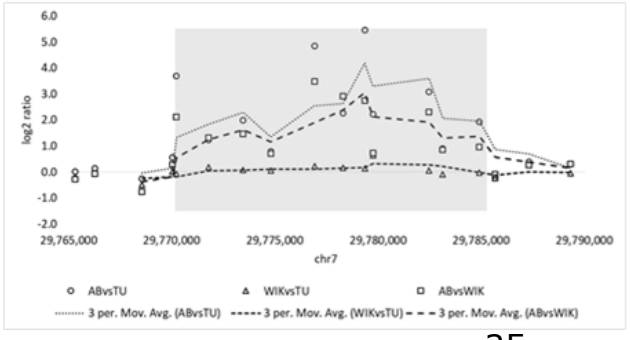


CNV_0968





CNV_1736



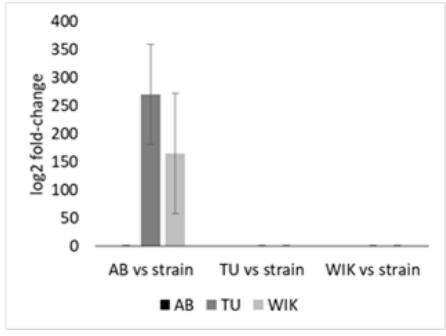


Figure 2.2: log2 ratios of microarray probes across chromosomal locations identified as CNV with corresponding qPCR log2 fold-change values. CNV_0572 = not validated, CNV_0663 = loss in TU, CNV_0900 = gain in TU, CNV_0968 = loss in WIK, CNV_1736 = loss in AB. Grey regions on log2 ratio plots indicate location of CNV. Error bars represent standard deviation and large error bars indicate variation within the strain (i.e., the loss or gain is not fully penetrant).

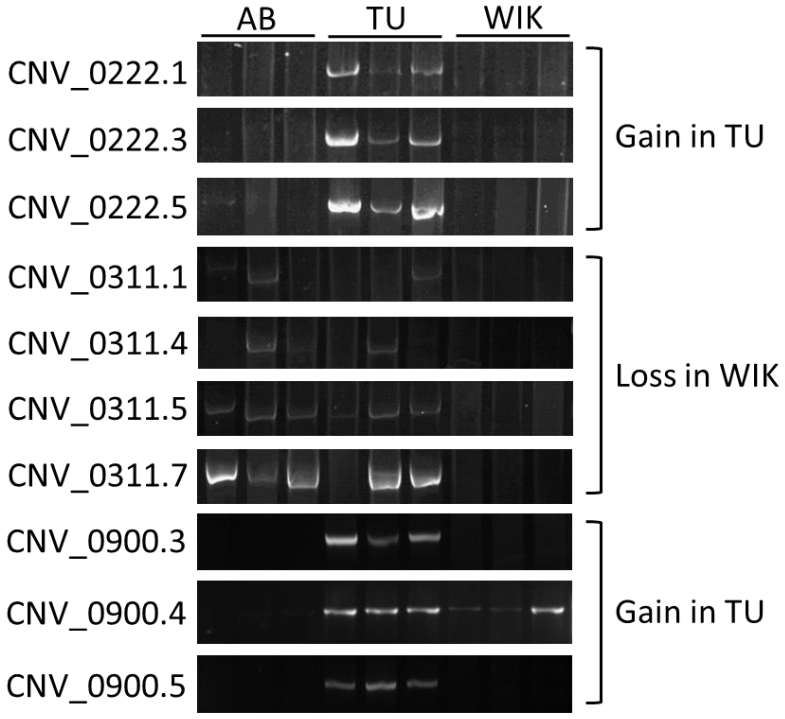


Figure 2.3: Standard PCR across subsets of three CNV regions. Each PCR amplicon was approximately 3.7 kb in length. CNV_0222 is only present in TU, while CNV_0311 and CNV_0900 show some variation across strains, but the array-based CNV call is confirmed by qPCR and PCR data across multiple individuals.

CNV were identified across all chromosomes, with the highest number of calls falling on chromosomes 3, 4, and 7 (Figure 2.4). We have created a Public Session accessible on the UCSC Genome Browser for the zebrafish community to view CNV data from this study titled

"danRer11 zebrafish CNV across strains" (as an example, see Figure 2.5). Using the Ensembl Variant Effect Predictor we queried 1355 CNV (GRCz10) against the RefSeq database and found 2200 genes and 5121 transcripts within ± 5 Kb of identified CNV. Of these genes and transcripts located proximal to CNV regions, 25% resulted in intron variants, 16% resulted in coding sequence variants, and 21% impacted 3' or 5' UTRs (Figure 2.6 and Supplementary Table 2.3).

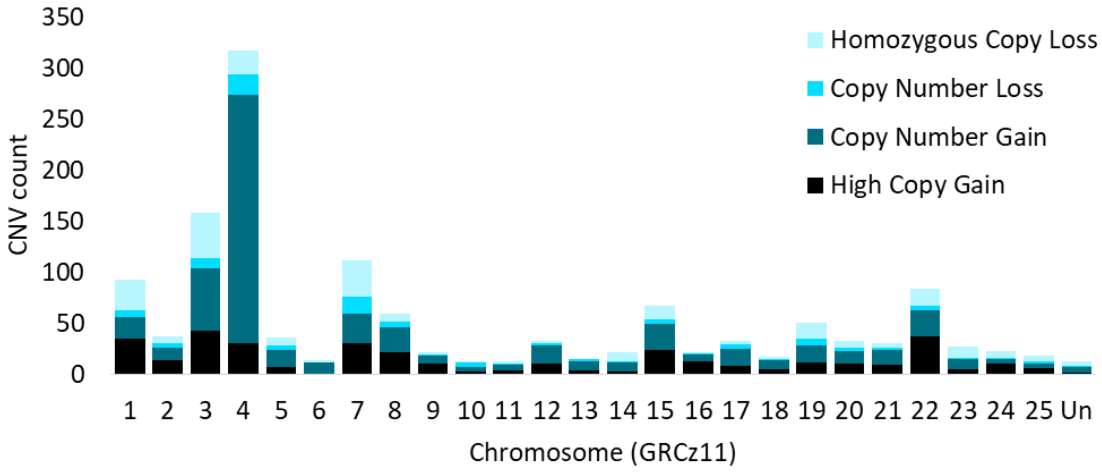


Figure 2.4: Copy number counts across all chromosomes in GRCz11 by type (Homozygous copy loss, copy number loss, copy number gain, or high copy gain.

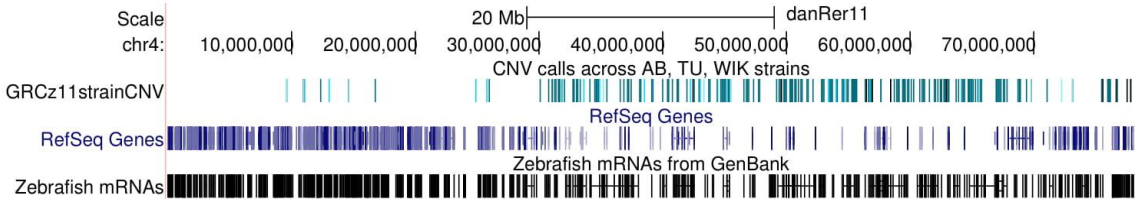


Figure 2.5: UCSC Genome Browser view of chromosome 4 (GRCz11) with CNV locations in teal, RefSeq Genes in blue, and GenBank mRNAs

in black. Freely available at https://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=holdenl&hgS_otherUserSessionName=danRer11%20zebrafish%20CNV%20across%20strains.

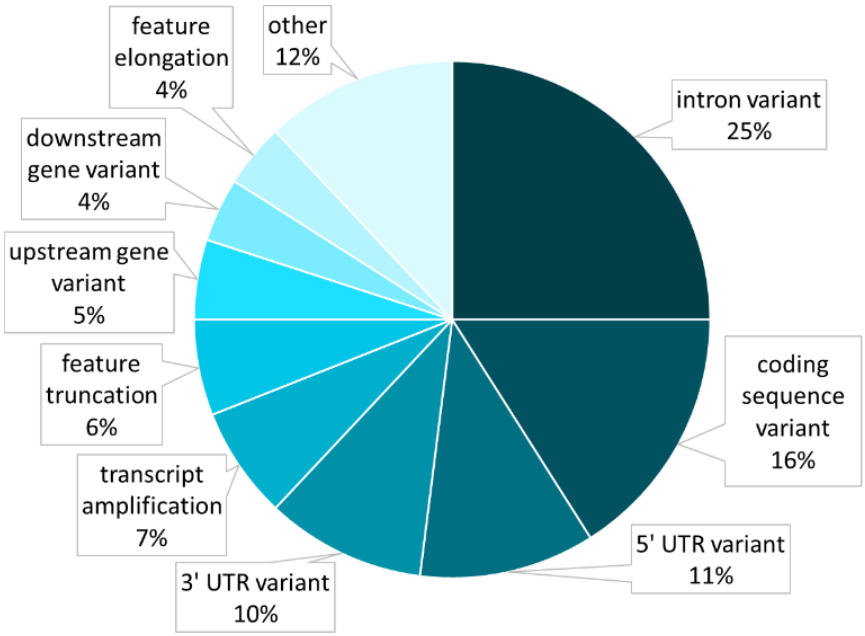


Figure 2.6: Predicted consequences of CNV that occur within 5 kb ± of RefSeq transcripts (GRCz10) using the Ensembl Variant Effects Predictor.

Discussion

In an effort to understand the effects of CNV on common lab strains, we searched for available datasets reflecting structural variation in zebrafish. The current NCBI structural variation database (dbVar, www.ncbi.nlm.nih.gov/dbvar) contains only human and great ape datasets and wholly excludes zebrafish data. The majority of available datasets incorporating CNV and phenotype are from humans, where much work has been done to associate structural variants to disease

phenotypes. The EMBL-EBI Database of Genomic Variants archive (DGVa, www.ebi.ac.uk/dgva) contains 200 genomic structural variant studies, but only a single study utilizes zebrafish (study ID: nstd62). The design of the single zebrafish study in DGVa used within population comparisons to identify strain-specific structural variation, and then used a subtractive model to infer between strain differences[46]. Therefore the majority of the structural variation presented in that study is within-strain (intrastrain) variation. We also found one previous study that looked at copy number variation in the wild zebrafish strain ASWT[80], but the strain is not widely used so the data are not directly applicable to most zebrafish investigators. Interestingly, we found no overlap in their copy number deletion or insertion loci with our dataset, which indicates that CNV are highly strain-specific.

This study performed reciprocal comparisons between three common laboratory strains (AB, TU, and WIK) and found 1351 CNV regions covering 1.9% (31.7 Mb) of the current zebrafish genome assembly (GRCz11). The effects of these CNV on phenotype are not fully known, but 2200 genes comprised of 5121 transcripts fall within ± 5 kb of identified CNV. From studies in humans, we know that CNV can alter gene expression, often by acting through regulatory elements from up to 1 Mb away[72], so the impact of these CNV is probably much larger than the 2200 genes that fall within 5 Kb. With over 5000 transcript

variants within 5 Kb up- or downstream of CNV, we predict that the expression of many of these transcripts likely vary across strains, dependent on copy number status. Previous work that characterized the CNV within strains found CNV cover approximately 15% of the zebrafish genome, but the experimental design focused on within-strain variation[46]. This analysis used pooled sample arrays (n=10 individuals per strain) and likely reduces the level of interstrain variation detected as only the most common variants are likely to be identified. We found 1.9% of the zebrafish genome to be affected by high-confidence CNV that are unique across strains and have made these CNV loci available to the zebrafish community for further exploration.

Acknowledgements

This study was funded by NIEHS K99/R00 ES018892 and start-up funds awarded to KHB. Additional support was provided by the Science, Mathematics, and Research for Transformation (SMART) scholarship, a National Defense Education Program, to LAH.

Chapter 3

Baseline mRNA expression differs widely between common laboratory strains of zebrafish

Published in Scientific Reports, 8, 1–10 (2018).

Abstract

Common strains of wildtype zebrafish (*Danio rerio*) have unique genomic features including SNPs and CNV, but strain information often goes unreported in the literature. As a result, the confounding effects of interstrain variation makes repetition of studies in zebrafish challenging. Here we analyze hepatic mRNA expression patterns between three common zebrafish strains (AB, Tuebingen (TU), and WIK) using Agilent 4x44K gene expression microarrays to establish baseline mRNA expression across strains and between sexes. We observed wide variation in sex-specific gene expression within AB and WIK strains (141 genes in AB and 67 genes in WIK), but no significant variation between sexes within TU. After partitioning the dataset into male and female subsets, we detected 421 unique mRNA transcripts with statistically significant differential expression; 269 mRNA transcripts varied between males, 212 mRNA transcripts varied between females, and 59 mRNA transcripts varied across the three strains, regardless of sex. It is not surprising that mRNA expression profiles differ between sexes and

strains, but it is imperative to characterize the differences. These results highlight the complexity of variation within zebrafish and underscore the value of this model system as a valid representation of normal variation present in other species, including humans.

Introduction

Laboratory strains of zebrafish (*Danio rerio*) have discrete genomic backgrounds; they clade out with very high bootstrap support by distinct SNPs[61] and have unique sets of copy number variant genomic regions[46]. Because of these genomic traits, zebrafish strains may be able to serve as a proxy to incorporate genetic variation into study design, similar to our understanding of the genomic variation in distinct human populations[81]. The human 1000 Genomes Project found that many common genetic variants are shared across populations, but rarer variants are generally only shared by closely related populations[15]. Analogous to distinct human populations, zebrafish strains have unique origin stories and genetic isolation between strains is maintained by strict husbandry practices.

Commonly used zebrafish strains such as AB (ZFIN ID: ZDB-GENO-960809-7), Tuebingen (TU; ZFIN ID: ZDB-GENO-990623-3), and WIK (ZFIN ID: ZDB-GENO-010531-2) have well-documented histories (Figure 3.1) and are easily obtainable for laboratory manipulations. The

AB line began from unknown zebrafish source stocks bought from two pet shops (pet shop A and pet shop B) in Albany, Oregon in the early 1970s[82]. Haploid progeny from AB females were crossed with random AB males for approximately 70 generations until the early 1990s when six diploid progeny stocks (each from a distinct haploid female) were thoroughly intercrossed to produce the modern AB line (sometimes referred to as AB*). The current AB source stock is maintained through large group spawning crosses. The TU strain originated from a composite population of fish purchased from pet shops in 1994 and was maintained as an inbred strain in a lab in Tuebingen, Germany[83,84]. The WIK strain ("Wild India Kolkata") originated from a single pair mating of wild caught fish in 1997[85]. The establishment and maintenance of these different strains has resulted in a similar observable phenotype.

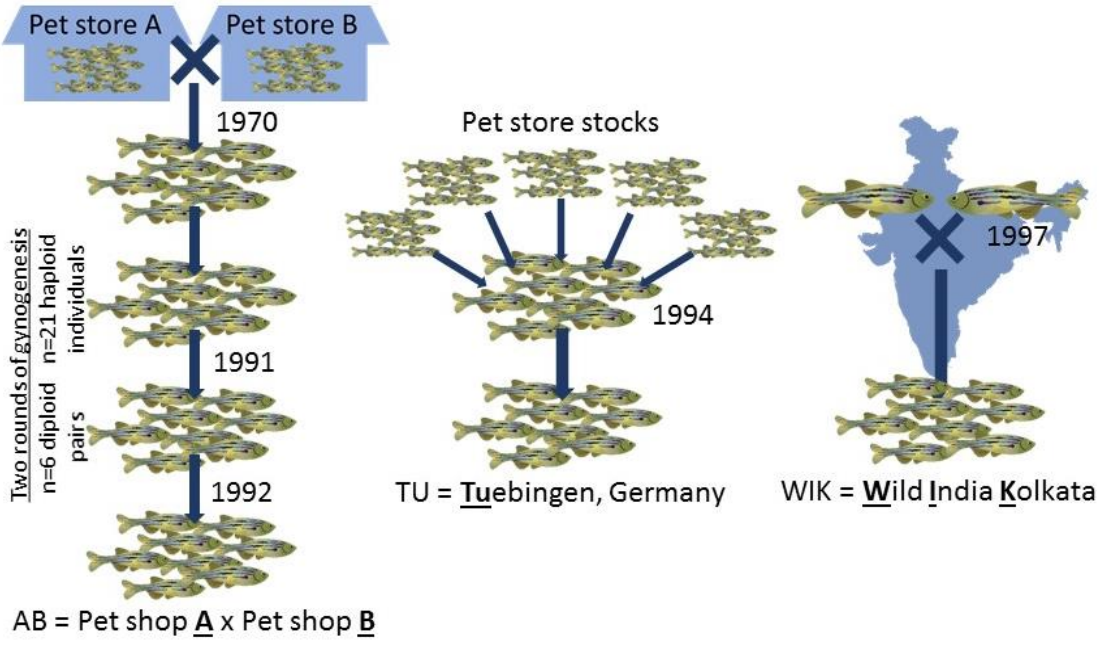


Figure 3.1: History of strain establishment for common laboratory strains of zebrafish. AB, TU, and WIK are three popular zebrafish strains used in genetic, developmental, and toxicological research with very different origin stories.

The high homology between humans and zebrafish—71% of human genes have at least one zebrafish ortholog and 69% of zebrafish genes have at least one human orthologs[43]—makes zebrafish an excellent model to study development, genetics, and toxicology. Unfortunately only 83% of transgenic and 46% of non-transgenic wild type strains of animal models are actually identified in the published literature[86] indicating that strain-based genetic variation is largely overlooked or ignored. Behavioral traits associated with domestication in wild versus lab-reared zebrafish are associated with differential mRNA

expression in the brain[65], indicating that the genetic isolation and population bottleneck inherent during laboratory strain establishments of zebrafish can create distinct characteristics between strains. Sex is an additional factor that drives differential mRNA expression between strains, mostly associated with hormone biosynthesis[87]. The goal of this study is to identify baseline liver mRNA expression variation between different zebrafish strains and between sexes in support of the growing recognition of normal variation between strains and populations[88,89] in an organismal and physiological context to support zebrafish as a strong model for translational research.

Methods

Animal care and husbandry

All zebrafish husbandry and experimental procedures were performed following protocols approved by Portland State University's Institution Animal Care and Use Committee in accordance with the National Institutes of Health Guidelines for Care and Use of Laboratory Animals and the Public Health Service Policy on Humane Care and Use of Laboratory Animals. Zebrafish are housed on an Aquaneering semi-recirculating housing system at a density of 5 individuals per liter with 10% daily water changes. Water temperature is maintained at 27.5°C and fish are kept on a 16 hour light, 8 hour dark photoperiod. pH and

conductivity are maintained at approximately 7.4 and 1100 µS, respectively. Zebrafish are fed commercial flake food twice daily and supplemented with artemia and rotifer live food. AB, TU, and WIK strains are maintained in-house by random single pair breeding. Larvae are screened for developmental abnormalities and 10 individuals from 25 pairs are randomly selected for the succeeding generation. The fish used in this study were second generation adults originally sourced from ZIRC (Eugene, OR) as batches of 100 embryos. All tissues were collected from healthy adults between 12 and 14 months old. At the time of dissection males weighed 331.7 ± 100.4 mg (mean ± SD) and females weighed 346.6 ± 90.7 mg. Male liver weights ranged from 0.002-0.021% of whole body weight and female liver weights ranged from 0.003-0.028% of whole body weight.

Nucleic acid isolation

White muscle and liver tissues were dissected from 3 males and 3 females from AB, TU, and WIK strains (n=6/strain; n=18 total) and disrupted with a mortar and pestle prior to homogenization by passing the samples through a nuclease-free syringe and needle in beta-mercaptoethanol lysis buffer. DNA was extracted on Qiagen DNeasy columns (Qiagen, Valenica, CA, USA) and total RNA was extracted on Qiagen RNeasy columns. Nucleic acid concentrations were determined

on a Nanodrop Spectrophotometer 2000 (Thermo Scientific, Wilminton, DE, USA). Both DNA and RNA exhibited high 260/280 ratios of 1.92 ± 0.04 and 2.10 ± 0.03, respectively (average ± SD), indicating adequate quality for downstream analysis.

mRNA expression arrays

Commercially available 4x44K zebrafish mRNA expression arrays, RNA spike-in kit, and Low Input Quick Amp one-color labeling kit (Agilent) were used following manufacturer's protocols. In brief, cDNA was synthesized from RNA and transcribed into cRNA using Cyanine-3 fluorescent dCTP. Labeled cRNA was purified using a Qiagen RNeasy mini kit per the manufacturer's protocol and quantified on a NanoDrop spectrophotometer. Samples with total cRNA yields greater than 1.65 µg and specific activity greater than 6 pmol Cy3/µg were fragmented, hybridized to array slides at 65°C for 17 hours, washed briefly, and scanned on an Agilent SureScan array scanner using grid file 026437_D_F_20140627 and scan protcol AgilentHD_GX_1Color. Data were extracted from raw TIFF files using FeatureExtraction software (Agilent) and spot brightness values were loaded into R. Raw microarray data files and derived expression values are archived at the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo) under accession number GSE100583.

Data normalization, analysis, and annotation

Data were cleaned by subtracting background fluorescence, normalizing across arrays, and averaging duplicate probes within the limma package[90] using R[91] version 3.3.2. Principal component analysis using the prcomp() function within base R (v.3.3.2) was performed using cleaned data (center = TRUE, scale = TRUE) and illustrated an overlapping, but clear, separation between male and female samples (variance explained by PC1 = 18.3% and variance explained by PC2 = 14.7%; PC1 + PC2 = 33%), so all downstream analysis was performed with male and female datasets separated. Within limma, empirical Bayes fitting of a linear model and pairwise contrasts were applied to AB, TU, and WIK strains separately to test for differences between males and females per strain. These results will be referred to as "sex differences". Similarly, a linear model and pairwise contrasts were applied to males and females separately to test for differences between AB, TU, and WIK per sex. These results will be referred to as "strain differences". Pairwise comparison values for fold change (log-2), average expression (log-10), p-value, and q-value were averaged for each strain and centered on zero to facilitate data interpretation. Significant probes were defined as ≥ 2-fold change in expression and Benjamini-Hochberg[92] adjusted p-value ≤ 0.05 (q-value). Standard Agilent array annotations were applied

to the probes and manually verified across NCBI and Ensembl databases. Conflicting annotations were resolved by direct overlap of mapped probes using UCSC's LiftOver tool as needed. Heatmaps were produced using the gplots heatmap.2 tool in R (v.3.3.2). Ordering of genes within heatmaps was performed using Euclidean distances and complete h clustering without scaling.

Gene ontology analysis

Gene ontology analysis was performed using the Panther Classification Tool[93] developed and maintained by the Gene Ontology Consortium. Ensembl and NCBI's ENTREZ gene ID annotations were assessed for statistical over-representation in the *Danio rerio* database (ZFIN last updated 04/2015) using default settings. GO complete annotations (database released 4/24/2017) for cellular component, biological process, and molecular function were assessed with Bonferroni[94] correction for multiple testing. Genes were considered over-represented at q-value ≤ 0.05 and results are presented as fold enrichment over the *Danio rerio* reference database.

Results

mRNA expression profiles differ between sexes in two of three strains

Analysis of total hepatic mRNA expression arrays detected 149 probes representing 141 genes that are significantly different between AB males and females (Figure 3.2A; Supplementary Dataset 3.1). Of these, 62 probes have a positive fold change indicating an increased expression of the transcript in males relative to females and 87 have a negative fold change indicating an increased expression of the transcript in females relative to males. Gene ontology analysis of 117 gene IDs (82.98%) mapping to *Danio rerio* shows that differences between males and females in the AB strain occur largely at the endoplasmic reticulum membrane (6.40-fold enriched, q-value = 3.01×10^{-3}). The biological processes of response to estradiol (87.14-fold enriched, q-value = 2.03 x 10-2), cellular response to estrogen stimulus (74.36-fold enriched, q-value = 1.00 x 10-9), lipid transport (22.49-fold enriched, q-value = 1.25 x 10-9), small molecule biosynthetic processes (10.46-fold enriched, q-value = 8.01 x 10-4), and monocarboxylic acid metabolic processes (7.88-fold enriched, q-value = 2.98 x 10-2) are statistically over-represented in the dataset and are largely driven by lipid transporter activity (29.05-fold enriched, q-value = 3.99 x 10-10) and oxidoreductase activity (4.79-fold enriched, q-value = 9.63 x 10-4).

Between WIK males and females, 72 probes representing 67 genes are significantly different (Figure 3.2B; Supplementary Dataset 3.2). Of these, 23 probes have a positive fold change indicating an

increased expression of the transcript in males relative to females and 49 have a negative fold change indicating an increased expression of the transcript in females relative to males. Gene ontology analysis of 58 gene IDs (86.57%) mapping to *Danio rerio* shows that differences between males and females in the WIK strain are not restricted to one cellular compartment, but encompass biological processes including response to estradiol (>100-fold enriched, q-value = 2.44 x 10-3), cellular response to estrogen stimulus (>100-fold enriched, q-value = 7.38 x 10-8), hormone biosynthetic processes (>100-fold enriched, q-value = 1.30 x 10-2), and lipid transport (37.80-fold enriched, q-value = 5.23 x 10-10). Similar to AB, these over-represented biological processes in WIK are largely driven by lipid transporter activity (42.61-fold enriched, q-value = 3.51 x 10-8). Interestingly, at our cutoff values of a minimum of 2-fold change in expression and q-value = 0.05, there are no probes that are significantly different between TU males and females. This is most likely due to a wider variation in the TU gene expression dataset.

Overlapping the differentially expressed probe sets from both AB and WIK produces a set of 40 probes mapping to 36 genes that are differentially expressed between males and females, regardless of strain (Figure 3.2C; Supplementary Dataset 3.3). Of these, only 6 probes have a positive fold change indicating an increased expression of the

transcript in males relative to females and 34 have a negative fold change indicating an increased expression of the transcript in females relative to males. Examples of mRNA transcripts conserved across strains include the protein responsible for converting androstenedione to testosterone (*hsd17b3*) in males and an egg yolk precursor (*vtg1-7*) and estrogen receptor (*esr1*), two well-known female-specific transcripts.

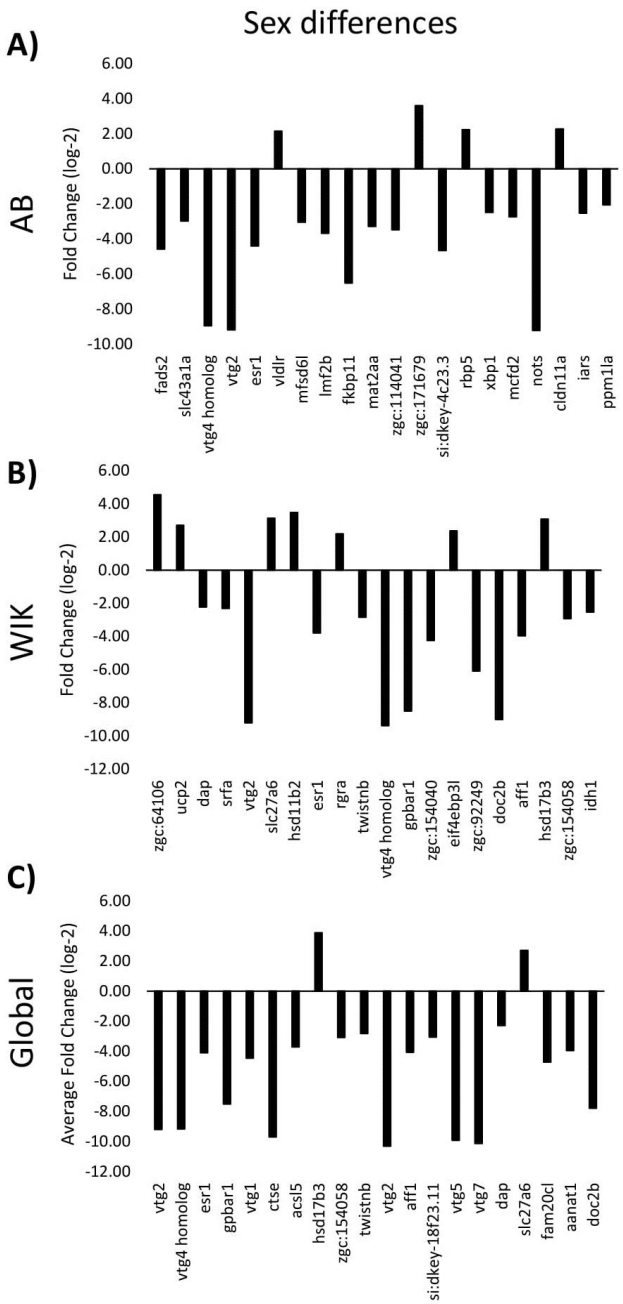


Figure 3.2: Top 20 most significant differentially expressed genes between sexes. Positive fold change values indicate higher mRNA gene expression in males, as compared to females. Negative fold change values indicate higher mRNA gene expression in females, as compared to males. A) In AB the top 20 q-values range from 0.0049 to 0.0062. B) In WIK the top 20 q-values range from 0.0030 to 0.0150. C) Regardless of strain the top 20 q-values range from 0.0081 to 0.0164. Fold change

values were averaged between male and female datasets. See supplementary datasets 1-3 for gene symbol definitions.

mRNA expression profiles differ between strains

Within males, 292 probes representing 269 genes are significantly different between AB, TU, and/or WIK males (Figures 3.3A and 3.4; Supplementary Dataset 3.4). Seventy-three (73) probes varied between TU and WIK (AB = 0 fold-change), 117 probes varied between AB and WIK (TU = 0 fold-change), and 102 probes varied between AB and TU (WIK = 0 fold-change). Within the strains, the percentage of transcripts with significantly increased expression accounted for 49.2-62.5% of the mRNA transcripts, with a mean of 56.3%. Gene ontology analysis of 237 gene IDs (88.10%) mapping to *Danio rerio* shows that differences between AB, TU, and/or WIK males are not restricted to one cellular compartment or molecular function, but are over-represented by the biological process of circadian regulation of gene expression (45.89-fold enriched, q-value = 7.20 x 10-3).

In females, 220 probes representing 212 genes are significantly different between AB, TU, or WIK (Figures 3.3B and 3.4; Supplementary Dataset 3.5). Fifteen (15) probes varied between TU and WIK (AB = 0 fold-change), 80 probes varied between AB and WIK (TU = 0 fold-change), and 125 probes varied between AB and TU (WIK = 0 fold-change). Within the strains, the percentage of transcripts with

significantly increased expression accounted for 57.0-67.5% of the mRNA transcripts, with a mean of 60.5%. Gene ontology analysis of 183 gene IDs (86.32%) mapping to Danio rerio shows that differences between AB, TU, and/or WIK females occur largely at the endoplasmic reticulum membrane (5.32-fold enriched, q-value = 9.58 x 10-4). Biological processes affected include protein targeting to the endoplasmic reticulum (49.52-fold enriched, q-value = 5.06 x 10-3), membrane assembly (31.28-fold enriched, q-value = 3.07 x 10-2), and single-organism metabolic processes (2.50-fold enriched, q-value = 3.27 x 10-3). These over-represented biological processes are largely driven by catalytic activity (1.72-fold enriched, q-value = 4.34 x 10-3).

Overlapping the differentially expressed probe sets from both males and females produces a set of 63 probes representing 59 genes that are differentially expressed between AB, TU, and WIK regardless of sex (Figures 3.3C and 3.4; Supplementary Dataset 3.6). Six (6) probes varied between TU and WIK (AB = 0 fold-change), 29 probes varied between AB and WIK (TU = 0 fold-change), and 28 probes varied between AB and TU (WIK = 0 fold-change). More than 50% of the probes varying between strains, regardless of sex, are attributable to the AB strain alone. Within the strains, the percentage of transcripts with significantly increased expression accounted for 46.7-56.6% of the mRNA transcripts, with a mean of 52.3%. Gene ontology analysis of 52

gene IDs (88.14%) mapping to *Danio rerio* shows no over-representation of any category between AB, TU, and/or WIK, regardless of sex.

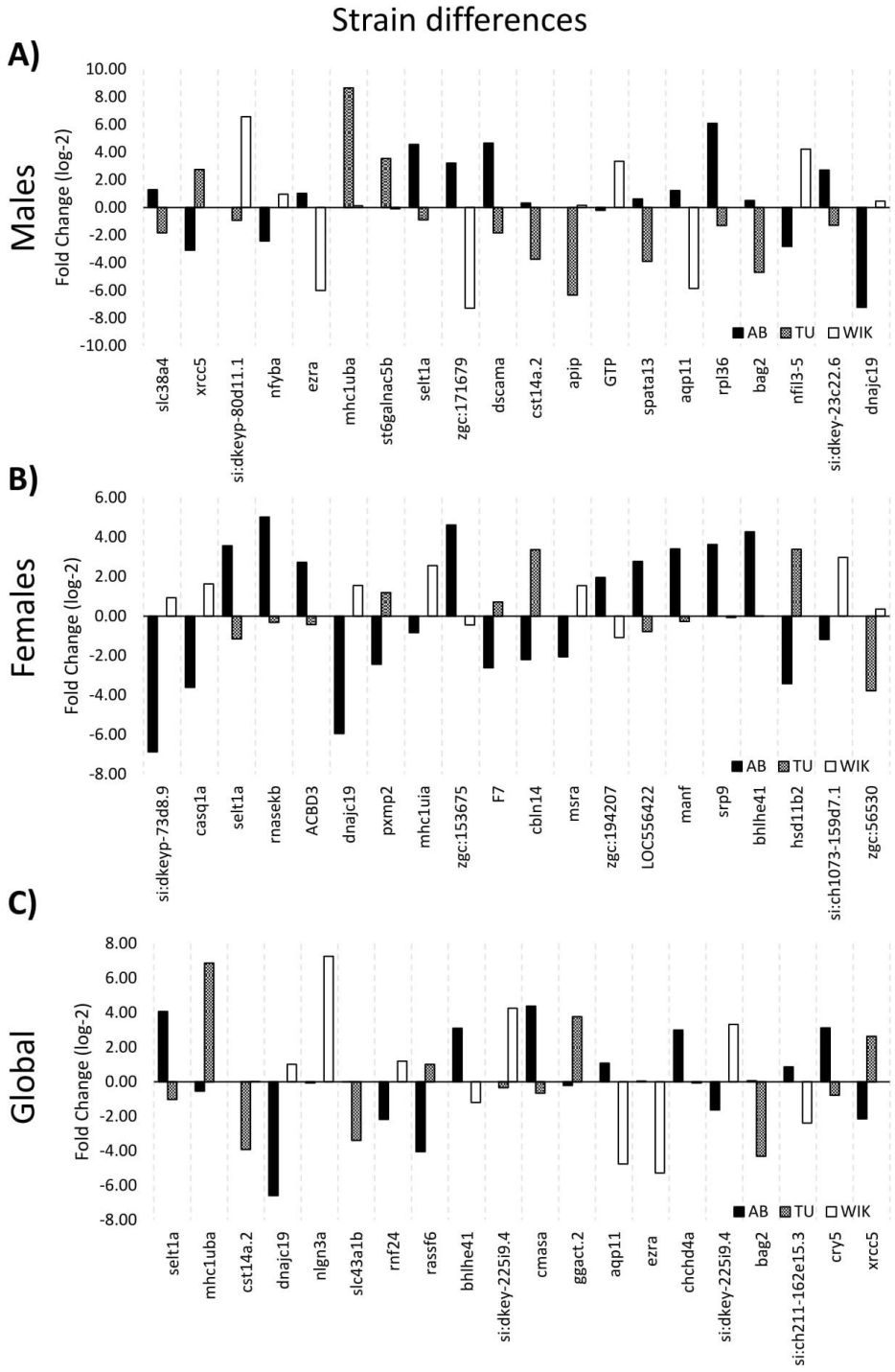


Figure 3.3: Top 20 most significant differentially expressed genes between strains. Positive fold change values indicate an increase in mRNA gene expression and negative fold change values indicate a decrease in mRNA gene expression. AB is represented by black bars, TU is represented by checkered bars, and WIK is represented by white bars. A) In males the top 20 q-values range from 0.0003 to 0.0020. B) In females the top 20 q-values range from 0.0005 to 0.0014. C) Regardless of sex the top 20 q-values range from 0.0007 to 0.0077. See supplementary datasets 4-6 for gene symbol definitions.

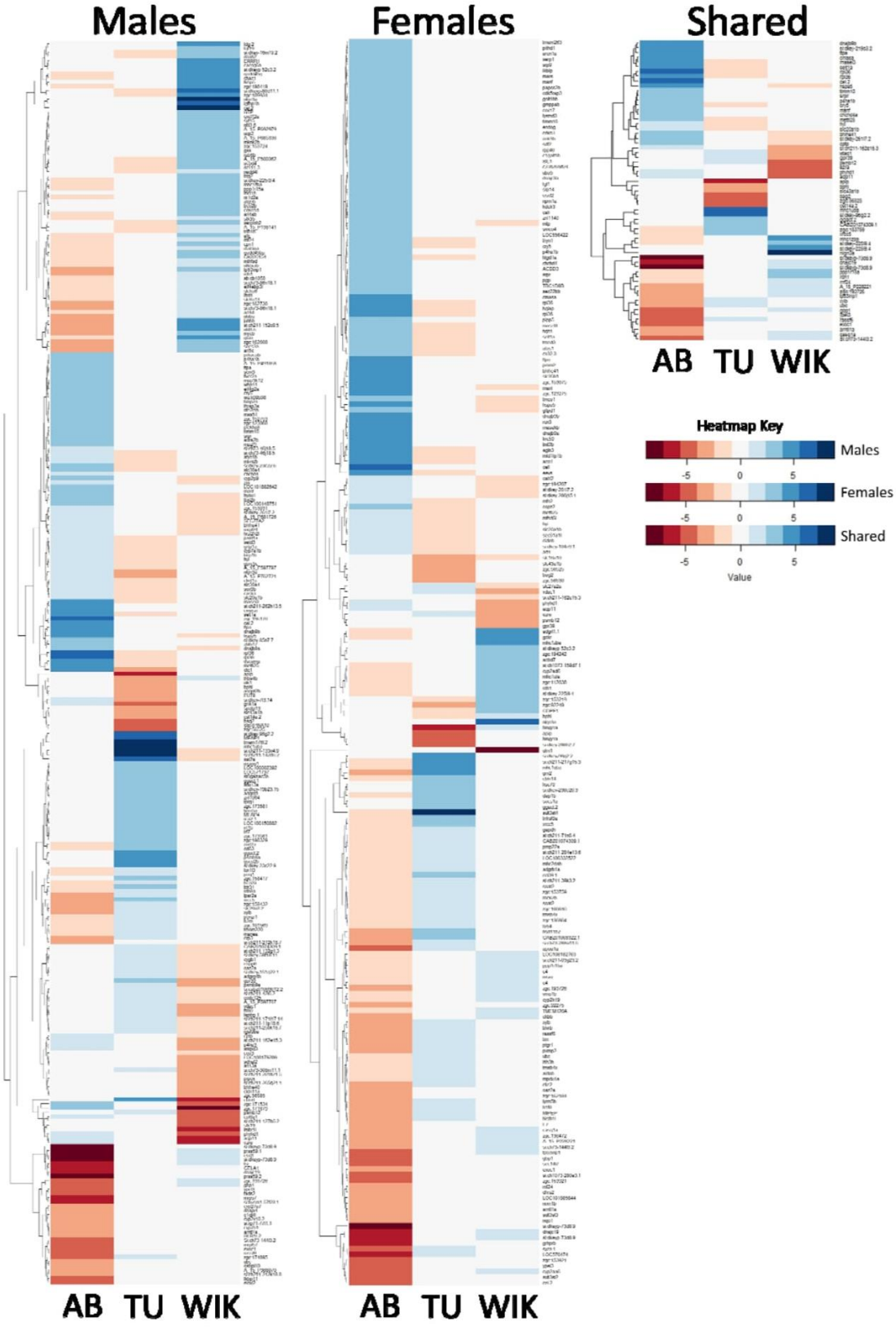

Males
Females
Shared
AB TU WIK
AB TU WIK
AB TU WIK
Heatmap Key
Males
Females
Shared
-5 0 5
Value

Figure 3.4: Differentially expressed mRNA transcript heatmaps. Individual heatmaps for males alone, females alone, and shared between the sexes (global) across AB, TU, and WIK strains. Blue indicates a positive fold change in expression, red indicates a negative fold change in expression. Higher saturation indicates stronger positive or negative fold change.

Discussion

A primary goal of this study was to identify baseline liver mRNA expression variation between different zebrafish strains. We identified large differences between strains, with a majority of differentially expressed mRNA transcripts belonging to AB (Figure 3.5). We hypothesize that this is due to the additional bottleneck of gynogenesis in the early establishment of the AB strain and a resulting decrease in heterozygosity by 34%, as similarly observed in gynogenetic diploid rainbow trout (*Oncorhynchus mykiss*)[95]. Additionally, across all sexes and strains, approximately 59% of probes show an increase in expression versus a decrease. The bottleneck of domestication reduces genetic variation[96], but since there is little to no selection acting on these laboratory strains, we predicted wide variation in expression phenotypes across strains[97] due to the inherent increase in the inbreeding coefficient[98]. Although we have described robust gene expression variation between AB, TU, and WIK, laboratory stocks still have less diversity between strains when compared to wild-caught zebrafish[62].

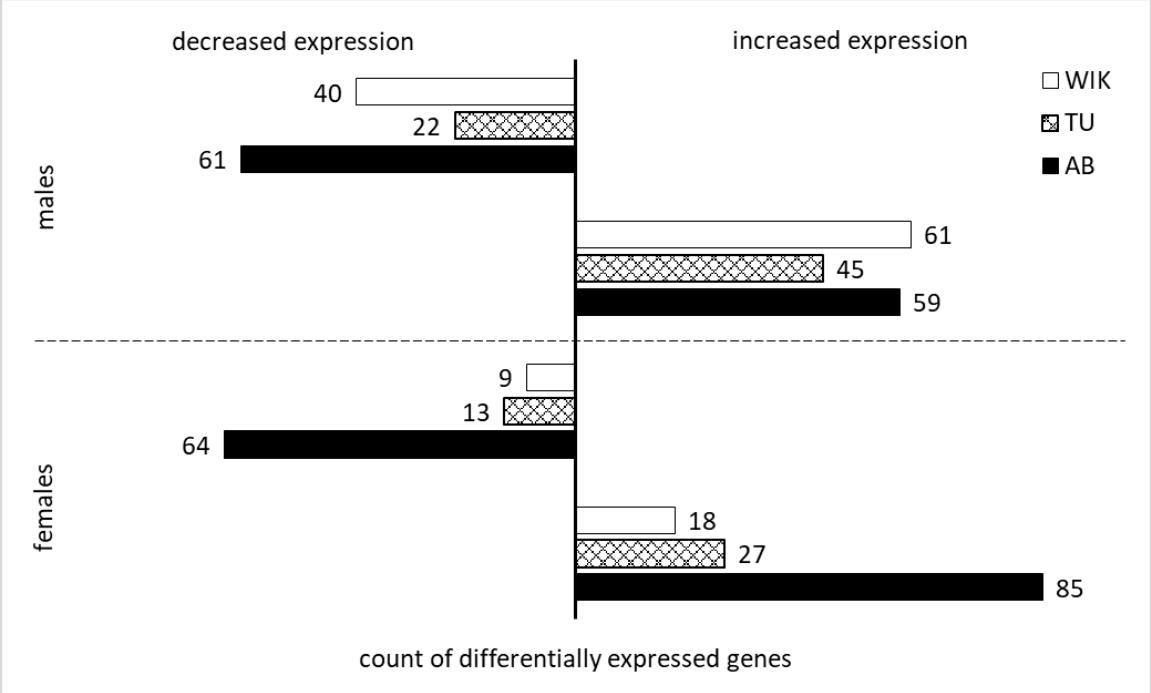


Figure 3.5: Summary chart of highly differentially expressed probe count in males or females across strains. Each bar represents a count of the mRNA transcripts with 2-fold increased (to the right) or decreased (to the left) expression by strain in males (top) and females (bottom) in AB, TU, or WIK. AB is represented by black bars, TU is represented by checkered bars, and WIK is represented by white bars.

Sex and strain both drive mRNA expression profiles in zebrafish

Sex determination in zebrafish has been argued extensively in the last decade, but only recently has a six-strain analysis led to a consensus hypothesis. Our current understanding is that genetic factors on chromosome 4 drive the ZW/ZZ sex-determining mechanism, but ultimate sex determination is sensitive to multiple environmental conditions[99]. Fascinatingly, AB and TU strains appear to have lost sex-specific signal across the sex-associated region in chromosome 4, so

factors defining male or female development in these strains are still unknown. WIK retains the chromosome 4 sex-associated region and has additional regions on chromosome 14 and several unassembled genomic scaffolds that are associated with sex determination. Interestingly, principle component analysis uncovers male and female grouping, as well as a clear separation of AB away from TU and WIK (Figure 3.6). Although sex is a major factor in this dataset, the loss of sex-determining regions in AB and TU do not appear to be driving the difference in mRNA expression between strains. Interstrain variation is most likely due to genetic differences caused by population isolation and bottleneck events during strain establishment. Moreover, we observed a large portion of differentially expressed mRNA transcripts that were specific to the AB strain, probably due to the extreme population bottlenecks and multiple rounds of gynogenesis.

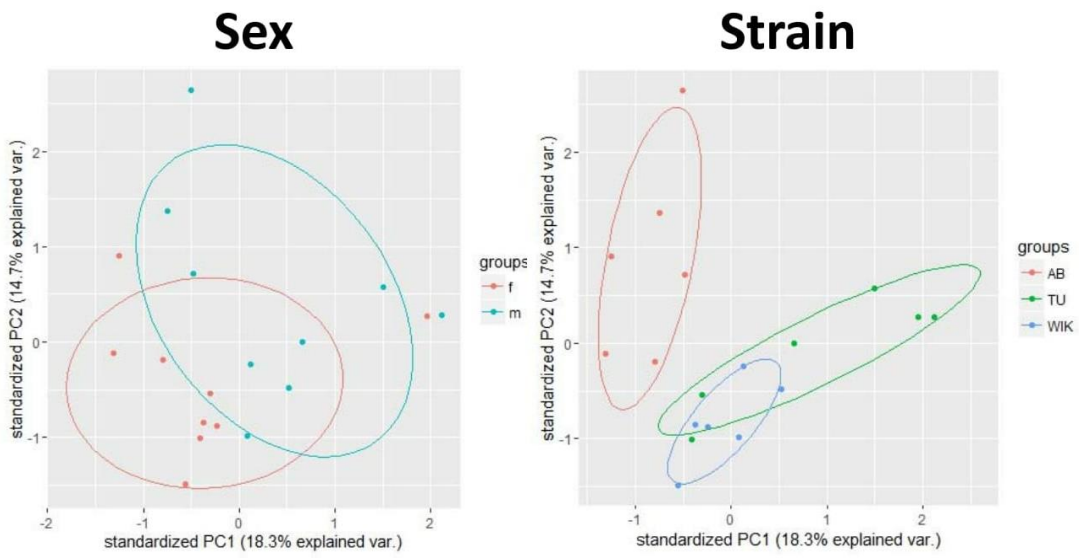


Figure 3.6: Principle component analysis of samples by sex or strain. PC1 and PC2 explain 33% of the total variance in the dataset. Sample identification by sex shows that male and female samples segregate, with the exception of a single female sample. Sample identification by strain shows that the AB strain clearly segregates from the TU and WIK strains.

Lipid transport mRNA transcripts differ between sexes in multiple strains

Genes involved in lipid transport are significantly enriched in differentially expressed mRNA transcripts between males and females in both AB and WIK. Among these are members of the vitellogenin (*vtg1-7*), retinol binding protein (*rbp2a* and *rbp5*), and solute carrier (*slc27a6* and *slc25a48*) families, as well as a transmembrane trafficking protein (*tmed1a*), a kainite glutamate receptor (*grik1a*), and an estrogen receptor (*esr1*). It is important to accurately characterize these differences between strains because lipid transport is critical in chemical

messaging, energy storage, temperature maintenance, and formation of membranes, cholesterol and prostaglandins. Furthermore, vitellogenin is a common marker for endocrine disruption in teleosts; any variation between expression in this endpoint may drastically affect interpretation of pharmacological endpoints including estrogenic activity of xenobiotics.

Circadian rhythm affects mRNA expression in males more than females

Circadian regulation in the zebrafish is directed by light- or dark-induced gene expression in the pineal gland[100]. Although our study design did not control for time of day (AB livers were collected in the morning, while TU and WIK livers were collected in the afternoon), there are only a small number of genes with circadian rhythm annotations. Specifically, we identified 295 annotations to 67 genes by searching AmiGO2[101] annotations for any term including the word "circadian" within zebrafish annotations. In the male dataset there are 6 genes that annotate to circadian rhythm: *arntl1a*, *bhlhe40*, *cry5*, *nfil3-5*, *nfil3-6*, and *nr1d2a*. In the female dataset there are only two genes that annotate to circadian rhythm: *arntl1a* and *cry5*.

To expand our analysis of potential circadian effects on our dataset, we queried circadian rhythm genes annotated in all organisms within AmiGO2. This expanded our list of potential circadian rhythm

genes to 2194, but led to no additional genes for the male dataset, and added only one gene, *F7*, to the list of genes present in female dataset that are known to be influenced by circadian rhythm. *F7* has been observed to be regulated by circadian rhythm in the Norway rat[102] and C57BL/6J mouse[103], but a similar regulation has yet to be identified in zebrafish. Using a comparable approach, we queried a circadian rhythm RNA-seq dataset in *Mus musculus* that assessed gene expression in multiple tissues across time[104]. We found 9 genes—*bhlhe41*, *ptgr1*, *dnaja4*, *fads2*, *fkbp5*, *lmbr1l*, *nedd41*, *slc38a4*, and *stk35*—in the mouse circadian rhythm dataset, but as of yet there is no clear evidence of oscillation in the expression of these genes in the liver of zebrafish. Again in this vein of inquiry, we queried the circadian expression profiles data base (circaDB)[105] against 4 mouse liver microarray studies and found 82 genes that have evidence of circadian regulation (Supplementary Dataset 3.7).

Expression patterns for the two circadian genes that are shared between males and females are conserved, with a decrease in expression of *arntl1a* and an increase in expression of *cry5* in the AB strain as compared to TU and WIK. This can be explained by the timing of liver harvest (AB in AM; TU and WIK in PM). What is fascinating, though, are the other genes affected by circadian rhythms that differed in the WIK strain only. *Bhlhe40* had lower expression in WIK and *nfil3-*

5, *nfil3-6*, and *nr1d2a* had higher expression in WIK. If expression of these genes were solely driven by circadian rhythms, then we would expect to see similar patterns between TU and WIK. Because this relationship is lacking, we hypothesize that there are other genetic factors that regulate the expression of these genes that differ between strains. This is interesting because experimental design accounts for the differences in males, but females seem to be less sensitive, suggesting that males are more sensitive to circadian perturbation than females. This is not unfounded as sex-specific phenotypes related to circadian rhythm have been observed in several animals, including behavioral traits in *Drosophila*[106] and liver metabolism in mice[107]. Most circadian oscillations in gene expression are not conserved across tissues and there are transcriptional "rush hours" prior to dawn and dusk[104]. Our samples were collected starting at 4 hours after dawn and ended 3 hours prior to dusk, which avoids the transcriptional rush hour and minimizes the maximal effects of circadian-driven transcription. Nonetheless, this is a reminder that time of day is a factor that should be considered in zebrafish study design, but that it is not the dominant driver of overall gene expression.

Functional implications of gene expression variation

While this is solely a descriptive study on the standing variation that exists in three strains of zebrafish, there are functional consequences of variable mRNA expression that should be assessed for the continued application of zebrafish as a model system. For example, in this dataset AB males have a greater than expected number of serine-type endopeptidase (GO:0004252) mRNA transcripts: prss59.1, prss59.2, ela2l, try, and cela1. All of these genes have greater than 7-fold lower expression in AB as compared to TU or WIK. Loss of expression of these genes in AB males may indicate a reduction in their ability to break internal amino acid bonds within polypeptide chains. As another example, WIK males have greater than 7-fold higher expression of two presynaptic membrane assembly (GO:0097105) mRNA transcripts: nlgn3a and cel.2. Both of these genes are involved in neuron cell-cell adhesion and neurexin family protein binding. Neuroligin genes, such as *nlgn3a*, are important in zebrafish nervous system development[108]. Disruption of the neurexin pathway at synapses leads to autistic-like behavior in mice[109] and mutation in *nlgn3a* in humans was associated with x-linked Asperger and autism disorders[110]. Moreover, a zebrafish model for autism spectrum disorder displays behavioral differences between strains[66]. *cel.2* is associated with maturity-onset diabetes of the young, type 8, with exocrine disfunction[111]. Because WIK males exhibit higher expression of these

genes, they may be compensating for loss of expression of related genes. A functional follow-up would be to see if neuronal synapses are enriched in WIK males for neurexin receptors or if there are any behavioral or exocrine disruption as compared to AB or TU males.

As a final example, AB have an 8-fold decrease in si:dkeyp-73d8.9 mRNA expression, an unknown transcript, in both males and females. Protein-protein alignment of the predicted amino acid sequence for si:dkeyp-73d8.9 against NCBI's non-redundant protein sequence database indicates that this is most likely a cystatin-like protein. Cystatins are inhibitors of cysteine proteinases and play a role in tumorigenesis, kidney function, and modulation of the immune system[112]. If all AB fish lack expression of this gene, then AB may be a better strain to target for development of mutation strains for model diseases involved in the disruption of the cystatin pathway. Continuation of describing and validating variation within zebrafish is paramount to the expansion of the zebrafish model system. This will further elevate the relevance of zebrafish studies to human health through the incorporation of multiple strains to simulate wide population variances, such as seen in human populations.

Conclusions

Our current understanding of zebrafish as a genetic model is based on the reference genome, which has only included alternate sequence loci as of June 2017[113]. The addition of alternate loci is a pivotal achievement for zebrafish as a model because it allows the interpretation of datasets with wide variance due to underlying structure within the data, such as genetically distinct sub-groups or populations. Our study goes one step further by describing baseline mRNA expression differences between zebrafish strains as a physiological interpretation of established genetic differences between zebrafish strains. We found major differences between strains and sexes including lipid transport and circadian rhythms. In the absence of a practical understanding of intra-population baseline variation, the downstream interpretation of data becomes skewed, reproducibility becomes increasingly challenging, and the application of study results become more abstract. Thus, this study serves as a foundational comparison of the strain-specific variation in mRNA expression in zebrafish and should be used to inform future study designs.

Acknowledgements

This study was funded by NIEHS K99/R00 ES018892 and start-up funds awarded to KHB. Additional support was provided by the Science, Mathematics, and Research for Transformation (SMART) scholarship, a

National Defense Education Program, to LAH. The authors would like to acknowledge Adalid Pelayo for facilitating zebrafish graphics.

Chapter 4

Response eQTL analysis of low-dose PCB exposure connects genomic copy number variants to susceptibility

Under review for publication in Aquatic Toxicology

Abstract

Physiological variation induced by genomic copy number variants (CNV) have received tremendous attention in human disease research, but little work outside of human health has be conducted. Herein we assess variable toxicant susceptibility phenotypes in zebrafish (*Danio rerio*), using microarrays to identify genomic copy number variants associated with induced gene expression differences. A micro-exposure approach (on the ng/L scale) was utilized to uncover secondary sex-specific mechanisms of PCB-126 toxicity without over-inducing the transcriptionally dominate cytochrome P450 family of xenobiotic metabolizing enzymes. We found over 30,500 CNV across all individuals, with approximately 3% of those CNV present at a frequency of ≥ 0.80 per strain. Following 24 hours of 130 ng/L PCB-126 exposure, we found 124 differentially expressed mRNAs in females and 97 differentially expressed mRNAs in males. Using identified CNV with high penetrance and PCB-126 induced mRNA expression data, we identified two sex-specific response eQTL, one each in males and females, for this

phenotype using a linear model analysis. The male response eQTL involves pre-mRNA processing factor 4 (*prpf4*) and the female response eQTL involves dynein cytoplasmic 2 heavy chain 1 (*dync2h1*). The CNV in both response eQTL are gains, but the response of the mRNA in each case differs. In *prpf4*, mRNA expression decreases when fish are exposed to PCB-126 while *dync2h1* mRNA expression increases following exposure. This is the first time that either of these genes have been linked to the PCB-126 susceptibility phenotype and both fall outside the canonical xenobiotic response pathway. Regardless, either or both may be linked tangentially with aryl hydrocarbon nuclear translocator transcriptional response elements.

Introduction

Genetic variation is the cornerstone for building stable populations resilient to environmental and ecological change. In sexually reproducing species, gene flow arising from immigration, emigration, mutation, and drift maintain genetic variation across populations from which natural selection determines "winners and losers". Among most studied species, the primary genetic research efforts have focused on variation arising due to single nucleotide polymorphisms (SNPs) or small insertions and deletions (indels). SNPs and indels can alter phenotypes by causing changes in the rate of transcription or changes in translated

protein sequence. Another, and perhaps more important, source of genetic variation is genomic structural variation, such as copy number variants (CNV) and inversions, which have been shown to modify phenotypes by altering gene expression[114]. Despite their large size and widespread genomic locations, CNV have received much less research attention[115]. A recent comprehensive study using matched DNA and tissue-specific mRNA expression from the Genotype-Tissue Expression (GTEx) consortium found that structural variants have larger effects compared to SNPs and indels, and primarily act via noncoding sequence variants localized to gene enhancer regions and regulatory elements[19].

To interrogate the effect of CNV on complex traits, we utilized a zebrafish (*Danio rerio*) model of toxicant susceptibility. The zebrafish toxicant model is comprised of genetically distinct laboratory strains[4] which exhibit a wide range of susceptibility to PCB-126 (3,3',4,4',5-Pentachlorobiphenyl), having developmental toxicity ranging from 9-336 ppb across six strains[44], and abundant strain-specific CNV[46]. While several SNPs, including one located within the primary ligand-target (aryl hydrocarbon receptor, AHR), have been associated with the PCB-susceptibility phenotype[44], no study to date has addressed the role of CNV for this trait. Toxicant susceptibility and resistance phenotypes are known to be conserved across mammals, birds, fish, frogs, and

invertebrates, but manifest through multiple mechanisms[116], making this trait an interesting target to query for CNV effects.

Animals exposed to PCB-126 have increased oxidative damage via DNA adducts of 7-hydro-8-oxo-2'-deoxyguanosine (8-oxodG)[117], which are generally repaired by the base excision repair pathway[118]. If adducts are not repaired prior to DNA replication, there is an increased likelihood for double strand breaks[119] and errors in double strand break repair is one mechanism by which CNV originate[120]. To circumvent somatic induction of CNV, this study utilizes a micro-dose acute exposure experimental design aimed to induce a xenobiotic response while avoiding *de novo* CNV production, as well as other detrimental effects that have been well-described with PCB-126, including developmental cardiotoxicity[121]. Our micro-dosing paradigm in adult zebrafish aims to elucidate potential alternative mechanisms that may contribute to observed differences in PCB-126 susceptibility outside of this pathway.

Many studies perform exposures in the microgram per liter range (generally, 1-1000 ug/L), but negative effects of PCB-126 have been reported down to 30 ng/L[57]. By targeting exposures well below the common ranges observed in zebrafish studies, we aim to uncover secondary mechanisms outside of the AHR response pathways. Our goal is to avoid the transcriptionally dominant induction of the cytochrome P450 family of xenobiotic metabolizing enzymes[122], while conserving the

real-world relevance of low-dose PCB exposure. The United States Center for Disease Control reports that the geometric mean of PCB-126 in the U.S. population is 16.3 pg/g of lipid (from serum samples)[56], while zebrafish exposed for 48 hours to PCB-126 at 3 ug/L have a body load of 37.2 ug/g of lipid (from whole embryo homogenates)[57]. Taken together, these points indicate that assessment of PCB-126 effects at higher concentrations is inadequate for understanding the effects of human PCB exposure and that our micro-dosing model may be better for elucidating the genetic determinants of PCB-126 susceptibility across populations from the human health perspective.

Finally, due to its lipophilic nature, PCB-126 will preferentially sequester to lipid-rich tissue[123]. Although lipids are largely equivalent between the sexes, males have higher liver and gonad lipid content than females[124] indicating that sex may act as a modifying factor in the variable effects of PCB-126 exposure. Studies in aquatic model organisms have used embryos and larvae due to easily observable developmental toxicity induced by PCB-126 exposure. The gold standard used to identify sex in zebrafish is by visual inspection of gonads because there are currently no reliable genetic markers for sex. Moreover, because zebrafish are juvenile hermaphrodites until sexual differentiation at 6 weeks[125,126] and because identification of sex in the exposed larvae is not possible by genetic means, sex-specific factors

have not been identified in early life stage exposures. For these reasons, we chose to use sexually mature adult zebrafish to assess the role of sex on PCB-126 susceptibility in our micro-dosing experiment.

Materials and Methods

Zebrafish Care and Aquatic Exposures

All procedures were approved by the Portland State University Institutional Animal Care and Utilization Committee in accordance with the Office of Laboratory Animal Welfare's Public Health Service policy. The zebrafish aquatic facility at Portland State University is composed of a recirculating flow-through housing system (Aquaneering). Water temperature is maintained at 28.5°C with average pH of 7.4, average conductivity of 1100 uS, and a 14:10 hour light:dark cycle. Zebrafish are fed flake food (Tetra) twice daily and supplemented with artemia and/or rotifers depending on life stage.

Three strains of zebrafish (AB, Tuebingen (TU), and WIK; n=6/strain/treatment; n_{total}=54) locally maintained for four generations were exposed to three treatments: 1) 130 ng/L 3,3',4,4',5-pentachlorobiphenyl (PCB-126; Ultra Scientific), 2) 20 ppm (v/v) acetone[127] as a vehicle control, or 3) nothing (naïve). We chose 130 ng/L as our exposure dose because it falls within the range of doses known to cause sublethal effects in embryos[57] and to avoid excessive

induction of oxidative DNA lesions[117]. Exposures were performed statically in 4 L glass beakers for 24 hours, zebrafish were fasted for the duration of the exposure, and at the end of the exposure period zebrafish were humanely euthanized for liver and muscle tissue collection. Tissues were flash frozen in liquid nitrogen and individuals were confirmed as male or female by visual inspection of gonadal tissue.

Nucleic Acid Extraction and Quantification

DNA was extracted and isolated from white muscle tissue using a Qiagen DNeasy Blood & Tissue kit per the manufacturer's protocol. RNA was extracted and isolated from liver tissue using a Qiagen RNeasy Mini Kit per the manufacturer's protocol. DNA and RNA concentration was measured using a NanoDrop 2000 (Thermo Fisher Scientific) to obtain 260/280 absorbance ratios for DNA = 1.96 ± 0.04 (mean ± SD) and RNA = 2.09 ± 0.02, indicating that nucleic acid preps were of high quality for downstream analysis. Archived DNA from Casper strain zebrafish[128] was used as a reference pool of DNA for CNV genotyping.

Phenotyping by mRNA Expression Microarray

100 ng extracted total RNA and control RNA (Agilent RNA Spike-In Kit, One-Color) was labeled following the manufacturer's protocol (Agilent RNA Spike-In Kit, One-Color and Agilent Low Input Quick Amp

Labeling Kit). Briefly, cDNA was synthesized then cRNA was synthesized from the cDNA template and concurrently labeled with cyanine 3-CTP (cy3) dye. Labeled cRNA was purified (Qiagen RNeasy Mini Kit) and quantified on a NanoDrop 2000. Only labeled samples with a yield of ≥ 1.65 ug total cRNA and specific activity of ≥ 6 pmol cy3/ug total cRNA were hybridized to arrays.

Labeled samples plus a blocking agent (Agilent) were fragmented for 30 minutes on a 60°C heat block, loaded onto each gasket well, sealed, and hybridized at 65°C for 17 hours at 10 RPM to Agilent 4x44K Zebrafish (V3) Gene Expression microarrays. After incubation, arrays were washed and scanned immediately on an Agilent SureScan microarray scanner. Data were extracted from arrays using Agilent Feature Extraction software and assessed for basic quality control parameters included in the standard Feature Extraction QC report. After passing quality control requirements data were cleaned and loaded into R[91] (v.3.3.2).

Differential expression was determined using the *limma* package[90]. The cyclic loess method was used for inter-array normalization and replicate probes were averaged within arrays. Males and females were separated into sex-specific datasets to avoid the effects of transcriptional differences between sexes[129]. Pairwise comparisons were made between strains in the PCB and naïve groups

after controlling for effects of the vehicle control (PCB minus vehicle). Differential expression was defined as any probe with a Benjamini-Hochberg corrected p-value[92] (q-value) ≤ 0.05.

Genotyping by Copy Number Variant Microarray

1 ug extracted DNA or archived Casper DNA was labeled using the BioPrime Array CGH Genomic Labeling System (Invitrogen) and fluorescent cy3 or cyanine 5-CTP (cy5) dyes (Perkin Elmer) following the manufacturer's protocol. Briefly, DNA was fragmented on a thermocycler at 95°C for 30 seconds to a target size of 200-500 bp, reference DNA (Casper) was labeled with cy3, test strain DNA (AB, TU, or WIK) was labeled with cy5 dye, labeled DNA was cleaned using Amicon Ultra filter columns (Millipore), and quantified on a NanoDrop 2000. Target parameters for labeled samples were a yield of ≥ 2.8 ug labeled DNA (equation 4.1) and specific activity of 20-60 (equation 4.2).

Equation 4.1: $$\frac{[DNA\ ng/uL]*sample\ volume\ (uL)}{1000\ ng/ug} = yield$$

Equation 4.2: $$\frac{[dye\ pmol/uL]}{[DNA\ ng/uL]} * 1000 = specific\ activity$$

Labeled samples were combined, denatured, prehybridized with 50 ug herring sperm to block excessive hybridization across hyper-

repetitive regions, and hybridized on custom designed 1M aCGH arrays designed to span the danRer7/Zv9 zebrafish reference genome (Agilent) at 65°C for 40 hours at 20 RPM. After incubation, arrays were washed and scanned immediately on an Agilent SureScan microarray scanner. Data were extracted from arrays using Agilent Feature Extraction software and assessed for basic quality control parameters included in the standard Feature Extraction QC report. After passing quality control requirements data were cleaned and loaded in to Agilent Genomic Workbench 7.0.

To preprocess arrays, data were passed through a feature filter (DefaultFeatureFilter which removes saturated and non-uniform probes), an aberration filter (minimum number of probes per amplification or deletion = 3, minimum absolute average log ratio for amplifications = 0.33, minimum absolute average log ratio for deletions = 0.5), normalized (legacy center, threshold = 6, bin size = 10, and GC correction with a 2 Kb window size), and intra-array replicates were combined. Copy number variant regions were called using the aberration detection method 2 (ADM-2) algorithms (threshold = 6) within Genomic Workbench. ADM-2 incorporates quality information about each log ratio measurement, identifies all aberrant intervals per sample using high and low log ratios with p-value ≥ 0.05, and determines optimal size of aberrations. The strength of the ADM-2

algorithm is its ability to incorporate noisy data and identify small aberrant regions. CNV calls were processed through a custom perl script (DataMerge2) using a 50% reciprocal overlap across CNV regions to identify a copy number gain or loss and regions of 50% overlap (with shared parents) within CNV[46].

Not all CNV are present in all individuals[16]. Because the PCB-susceptibility phenotype is strain-specific, we assume that CNV that may be involved in this trait are present at a high frequency within each strain. Operating under this assumption, we assessed the strain-specific frequency of CNV and only included high frequency variants (frequency ≥ 0.8) for expression QTL (eQTL) analysis.

eQTL Identification and Response eQTL Confirmation

To identify eQTL we applied a linear model that uses CNV present at a frequency ≥ 0.80 (15/18) in at least one strain as a predictor variable and PCB-induced mRNA expression as the response variable. Specifically, data were loaded into R as four files: CNV genotype per individual, mRNA expression per individual, start/stop sites for CNV, and transcription start site for differentially expressed genes,. eQTL were called using the R package matrixEQTL[130]. This study utilized a linear model within matrixEQTL and set *cis* distances as 2 Mb up- or down-stream of identified loci. Significance thresholds for *cis* and *trans* eQTL

were defined using a conservative Bonferroni-corrected p-value[94] based on the number of comparisons in each group (p-value = 10^{-4} and 10^{-8}, respectively).

Because our linear model only incorporated mRNA expression from PCB-exposed individuals, we did additional data visualization to confirm that the transcriptional response differed with CNV status (as identified by the eQTL association) and differed in PCB-exposed individuals as compared to naïve or vehicle-exposed individuals. To confirm the PCB-specific response of statistically significant eQTL we plotted the eQTL we had identified by treatment, strain, and CNV status to look for treatment-specific differences in gene expression[131].

Results

PCB-mediated differential gene expression among strains

Baseline comparison between strains revealed differential expression of 1539 probes in males and 2610 probes in females. Assessment of the effect of PCB exposure resulted in statistically significant (BH-corrected p-value < 0.05) up- or down-regulation of expression of 249 probes in males and 831 probes in females. Of these differentially expressed probes, only 97 in males (38.96%) and 124 in females (14.92%) remained after removing probes that were significantly different between strains before treatment or significantly

induced by vehicle treatment alone. (Figure 4.1, Supplemental Data 4.1).

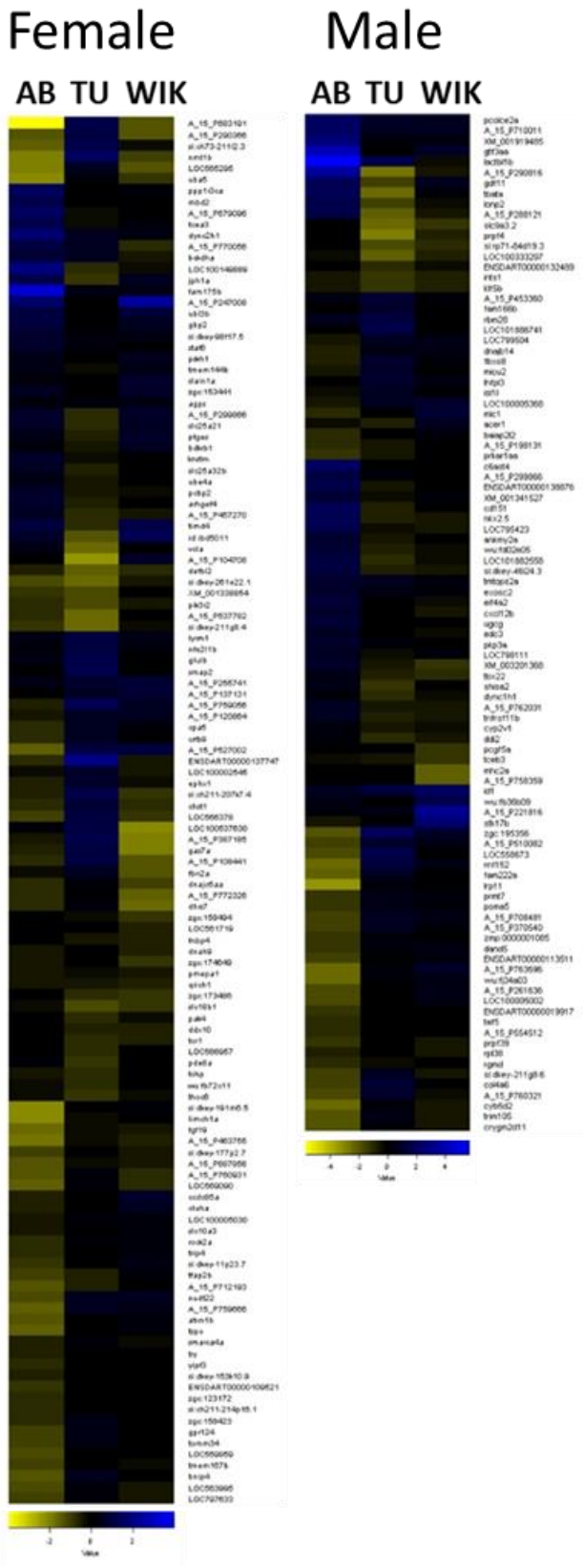

Female
AB TU WIK
Male
AB TU WIK

Figure 4.1: Heatmaps of female and male samples indicating relative mRNA expression induced by 24 hours of PCB-126 exposure. There are 124 probes with significantly differential expression in females and 97 probes in males. Yellow indicates a decrease in expression relative to naïve controls, blue indicates an increase in expression relative to naïve controls, and black indicates no change. Probes are organized by similarity of expression pattern, as determined by Euclidean distance and complete linkage.

CNV Frequency within Strains

Although microarrays can identify CNV, they are unable to determine if the variant regions are somatic (*de novo*) or within the germline. Because this study aims to identify germline CNV that are unique across strains, but conserved within strains, we assessed CNV for frequency by strain to identify targets for eQTL analysis (Table 4.1). Distribution of CNV throughout strains varied widely, with only 7.91% of CNV (2418/30587) with a strain-specific frequency ≥ 0.50 and only 0.94% of CNV (287/30587) with a strain-specific frequency of 1.00. Of the 2.93% of CNV occurring in at least 15 out of 18 individuals per strain (frequency ≥ 0.80), we identified 200 unique CNV gains and 381 unique CNV losses across AB, TU, and WIK strains relative to the reference strain.

Table 4.1: CNV counts per frequency threshold where frequency = # individuals with CNV/total # of individuals, calculated across strains independently (AB, TU, WIK). CNV gain or loss is relative to Casper reference strain. Total indicates number of CNV present in single or multiple strains, but only counted once.

frequency	0.50		0.80		1.00	
CNV	gain	loss	gain	loss	gain	loss
AB	322	618	65	164	16	34
TU	1220	419	423	94	181	25
WIK	279	120	127	141	24	30
total	1414	1004	514	381	200	87

Cis and trans eQTL driving PCB-susceptibility phenotype

Two *cis* eQTL were identified in males and one *cis* eQTL was identified in females that associate a CNV with a gene that is differentially expressed between strains (Figure 4.2). The first male eQTL, prpf4+CNV_155, associated expression of pre-mRNA processing factor 4 (*prpf4*) mRNA with CNV_155 (chr5:60447537-60600000, danRer7/Zv9) with an effect size of -0.32 and a false discovery rate of 0.0004. The second male eQTL, XM_001919485+CNV_717, associated expression of XM_001919485 (filamin-a, *flnb*) mRNA with CNV_717 (chr11:43264797-43262726, danRer7/Zv9) with an effect size of 0.22 and a false discovery rate of 0.002. Finally, the female eQTL, dync2h1+CNV_343, associated expression of dynein cytoplasmic 2 heavy chain 1 (*dync2h1*) mRNA with CNV_343 (chr15:43913087-43913146, danRer7/Zv9) with an effect size of 0.37 and a false discovery rate of 0.005. Relative expression levels and CNV status across each eQTL are indicated in Table 4.2. Of these, the eQTL containing CNV_717 did not show any expression differences in

response to PCB-exposure and was therefore not deemed a true response eQTL (reQTL).

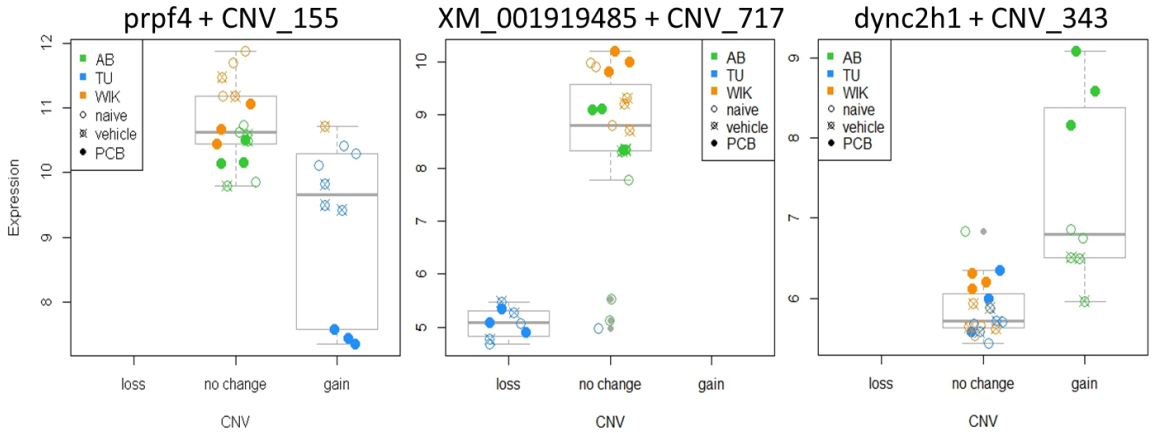


Figure 4.2: eQTL plots showing CNV status on the x-axis (loss, no change, or gain) and mRNA expression level on the y-axis for three statistically significant eQTL. Colors indicate strain type (AB = green, TU = blue, WIK = orange); symbol indicates treatment (empty circle = naïve, circle with x = vehicle control, full circle = PCB-126).

Table 4.2: cis eQTL including gene expression and copy number status. Gene expression values are relative across strains. Copy number status indicates direction of copy number distribution as compared to the reference strain, Casper. P-values are unadjusted.

	Gene expression			Copy number status			
eQTL	AB	TU	WIK	AB	TU	WIK	p-value
prpf4 +CNV_155	-0.14	-2.82	-0.86	no change	gain	no change	1.88E-06
XM_001919485 +CNV_717	2.72	0.28	0.44	no change	loss	no change	1.95E-05
dync2h1 +CNV_343	1.80	0.53	0.46	gain	no change	no change	1.66E-05

Further inspection of reQTL prpf4+CNV_155 placed the CNV region 1.56 Mb downstream of *prpf4* (Figure 4.3) and identified a genic

region within the CNV (*tmem136b*). There was no significant differential expression of *tmem136b*, either in naïve controls or PCB-126-treated zebrafish. Additionally, we identified 20 xenobiotic response elements (XRE, defined as KNGCGTC[29]) within CNV_155. Associated regions in reQTL dync2h1+CNV_343 were much closer; CNV_343 was only 0.67 Mb upstream of *dync2h1*. No genic regions and only a single XRE are located within the CNV region.

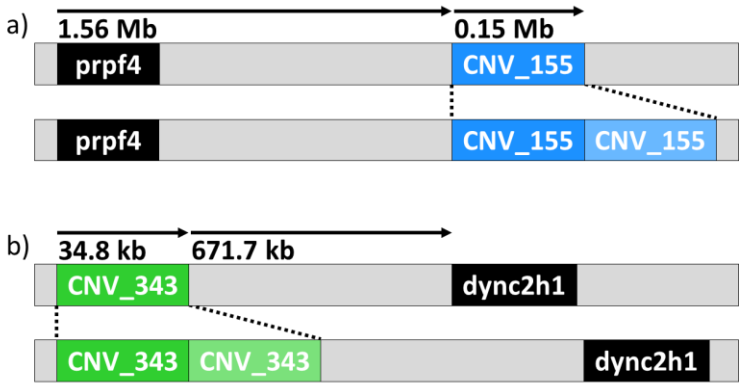


Figure 4.3: Two PCB-sensitivity reQTL identified as a) prpf4+CNV_155 (Chr5:60447537-60600000, Zv9/danRer7) where there is a copy number gain in TU males and associated reduction in prpf4 mRNA expression following PCB exposure and b) dync2h1+CNV_343 (Chr15:43189812-43224569, Zv9/danRer7) where there is a copy number gain in AB females and associated increase in dync2h1 mRNA expression following PCB exposure.

Discussion

To our knowledge, this is the first study using a micro-dosing approach to identify sex-specific mechanisms in the conserved PCB-susceptibility phenotype. Previous studies have over-looked sex-specific

mechanisms, but the reQTL identified in this study point to two mechanisms responsible for the variability in PCB-126 susceptibility across populations that are specific to either males or females. We found prpf4+CNV_155 to be the male-specific mechanism responsive to PCB exposure. *prpf4* facilitates proper spindle assembly during mitosis by recruiting checkpoint proteins at the kinetochore and loss of *prpf4* results in increased aneuploidy and improper mitosis[132]. *prpf4* is a member of the spliceosome[133], but a study in patients with retinitis pigmentosa found that approximately 50% of *prpf4* protein was not associated with the spliceosome complex[134], pointing to other unknown functions of this gene.

Additionally, *prpf4* is associated with heat stress in catfish[135] and ischemic stress in male rats[136], so being involved in an exposure-induced stress mechanism is a reasonable conclusion. An increase in copy number of a region enriched for XRE, such as CNV_155, should theoretically increase recruitment of transcription factor to that genomic location, but male TU zebrafish have a significantly lower expression of *prpf4* mRNA when exposed to PCB-126 and are more susceptible to the toxic effects of PCB than the other strains tested. We hypothesize that PCB-induced levels of *prpf4* mRNA are reduced in TU due to either an overactive post-transcriptional or post-translational regulatory mechanism or by physical blockage of the transcription of *prpf4* due to

recruitment of XRE binding factors. Loss of *prpf4* may affect PCB-126-induced cellular stress through alternate splicing or by a currently unknown surrogate function, and therefore result in higher sensitivity to PCBs.

The gene associated with the female-specific reQTL identified in this study, *dync2h1*, is known to be involved in both aminoglycoside[137] and temozolomide[138] resistance. Additional support for the role of *dync2h1* in PCB-126 resistance is that it is localized to the golgi apparatus[139] where lipids are packaged into vesicles[140]. In a 30-day zebrafish dosing study at 150 ng/L, the bioconcentration factor (uptake rate constant/clearance rate constant) of PCB-126 per animal wet weight was $10^{5.81}$ [141], illustrating that the bioaccumulation potential for PCB-126 is high due to lipid partitioning.

AB female zebrafish have an increased copy number of CNV_343 and a concomitant increase in expression of *dync2h1* following exposure to PCB-126. AB zebrafish are also more resistant to PCB-126 than the other strains tested. Because PCBs are lipophilic compounds, differences in packaging and transport of lipids may alter the deposition, metabolism, and excretion of PCBs. This suggests that increased packaging and transport of PCBs drives PCB-resistance in AB zebrafish.

Interestingly, both *prpf4* and *dync2h1* have aryl hydrocarbon nuclear translocator (ARNT) transcription factor binding sites within

their enhancer regions (as identified by GeneHancer[142]). ARNT is a well-described member of the canonical aryl hydrocarbon receptor (AHR) PCB-response pathway responsible for transporting ligand-activated receptors to the nucleus to directly act as transcription factors[143,144]. So although these two genes have not been previously implicated in PCB-mediated toxicity, they may be linked to the AHR-mediated PCB-exposure response mechanism through interaction with ARNT.

Conclusion

This study identified two novel reQTL associated with population-specific PCB-sensitivity using microarray-based CNV and mRNA expression data. We found two sex- and strain-specific expressed targets, namely *prpf4* in males of the TU strain and *dync2h1* in females of the AB strain, proximal to CNV regions that each explain approximately a third of the variation of the observed expression phenotype. Genetic mechanisms, such as SNPs across the AHR family in the Atlantic killifish[38], have been shown to drive population-specific resistance to PCBs, but this is the first time that genomic copy number has been implicated with this trait. Additionally, micro-dose exposure to PCB-126 allowed us to interrogate secondary mechanisms outside of the traditional AHR-mediated xenobiotic metabolism pathway and identify

novel targets that may act in concert with or independent from traditional response pathways.

Acknowledgements

The authors would like to acknowledge the help of the Oregon Health & Science University Knight Diagnostic Cytogenetics Laboratory, specifically Steve Moore and Lora Lucas, for coordinating the use of their microarray scanner. The authors would also like to thank Holly Paddock of ZFIN and Chris Lawrence of Boston Children's Hospital for help tracking down the Casper strain origins, Andrey Shabalin for recommendations on how to apply matrixEQTL to CNV data, and Spencer Torene for writing the DataMerger2 perl script. This study was funded by NIEHS K99/R00 ES018892 and start-up funds awarded to KHB. Additional support was provided by the Science, Mathematics, and Research for Transformation (SMART) scholarship, a National Defense Education Program, to LAH.

Chapter 5

Targeted CRISPR-Cas9 Editing of Genomic Copy Number Modulates PCB-Susceptibility Phenotype

Abstract

PCBs are ubiquitous legacy chemicals that cause health effects in humans and wildlife. There are several examples of PCB-resistance between wild populations living in highly contaminated environments versus clean environments, illustrating an adaptive advantage for populations that can maintain health in contaminated environments. Studies in natural populations that have evolved PCB-resistance identified multiple mechanisms—including SNPs, indels, and fusions—across the aryl hydrocarbon receptor (AHR) as one of the drivers of the PCB-susceptibility phenotype. Zebrafish also have variable sensitivity to PCB-126 across common laboratory strains making them a unique model system to test the genomic drivers of the PCB-susceptibility phenotype. A recent study identified genomic copy number across non-AHR regions to associate with the PCB-susceptibility phenotype. Herein we test the effect of copy number on PBC susceptibility by targeting duplicated genomic regions in two strains of zebrafish with variable PCB susceptibility (AB strain = resistant, Tuebingen (TU) strain = sensitive) using CRISPR-Cas9 mutagenesis. Because the CNV in AB zebrafish

appears to provide protection from the toxic effects of PCB-126, we expect that CRISPR-mutant (crispant) AB zebrafish will show higher sensitivity to PCB-126. Conversely, the CNV in TU zebrafish associates with increased sensitivity, so we expect crispant TU to have higher resistance to PCB-126. To test this we injected 1-cell stage embryos with CRISPR-Cas9 ribonucleic protein complexes targeted to strain-specific copy number variable regions (CNV_343 in AB and CNV_155 in TU), exposed embryos to variable concentrations of PCB-126, and assessed developmental toxicity by heart rate, edema, and morphology. In support of our hypothesis, crispant AB zebrafish showed a reduction of EC_{50} values by a factor of 10, from 627.7 to 67.9 ppb PCB-126, as compared to sham-injected controls. Crispant TU zebrafish had slightly increased EC_{50} values, from 35.3 to 47.0 ppb PCB-126, as compared to sham-injected controls. This study shows clear evidence of CNV as drivers of the PCB-susceptibility phenotype and is a first step towards inclusion of genomic CNV into modeling who or what will be susceptible to PCB-exposure.

Introduction

Polychlorinated biphenyls (PCBs) are aromatic hydrocarbons that were largely used as insulators, coolants, and plasticizers since their commercial inception in the 1930s. In 1979 the United States

Environmental Protection Agency banned the manufacture of PCBs due to evidence of bioaccumulation, ecological toxicity, observed health risks following occupational exposures, and a large-scale incident in Yusho, Japan where over 1000 people were poisoned following consumption of contaminated cooking oil (reviewed in [53]). Although manufacture of PCBs was banned in 1979, PCBs are still present in human serum at concentrations of 16.3 ± 1.6 pg/g or ppt (geometric mean ± 95% confidence interval) on a lipid-weight basis as measured by a 2000-2004 population survey in the United States[56]. Current exposure to PCBs is mostly through consumption of PCB-contaminated fish, dairy, and meat or exposure to PCB-contaminated air or water[145] which can result in various health effects such as decreased neonatal thymus size[146], reduced male fertility[147], and liver damage[148].

In nature, several examples of the rapid evolution of a PCB-resistance phenotype have been observed in species living in highly contaminated environments, such as the Atlantic killifish (*Fundulus heteroclitus*) in Superfund sites in New Bedford Harbor[36] and the Hudson River[32]. Zebrafish (*Danio rerio*) also exhibit PCB-resistant phenotypes that are graded across common laboratory strains; some strains are resistant to PCB exposure and others are extremely sensistive[44]. Reports of the genetic mechanisms behind PCB resistance have identified alterations to the canonical target of PCB, the aryl

hydrocarbon receptor (AHR)[39,44,47], by single nucleotide polymorphisms (SNPs), short insertions or deletions (indels), or splicing of AHR paralogs as the major mechanisms involved in this phenotype, but a substantial percentage of the phenotypic variation is not ascribable to short nucleic sequence variations in AHR.

An additional type of genomic variation that has been largely overlooked is copy number variation. Copy number variants (CNV) are large genomic duplications or deletions, generally on the order of 1 kb to 1 Mb. CNV span more of than genome than SNPs[15], act mainly through direct interaction with regulatory elements, and have a larger effect size than SNPs and indels when altering gene expression[19]. Zebrafish are an excellent model to use for studies on the effects of CNV because of their well-described strain-specific CNV[46], such as in the common laboratory strains AB and Tubingen (TU). The goal of this study is to test the effect of copy number on the PCB-susceptibility phenotype to further understand genomic drivers of this trait.

Herein we used CRISPR-Cas9 to target two strain-specific CNV that associate with PCB-susceptibility in zebrafish as identified in a recent reQTL study[149]. CNV_155 and CNV_343 are duplications that associate with changes in expression of *dync1h2* or *prpf4* following exposure to PCB-126 (3,3',4,4',5-Pentachlorobiphenyl), respectively. We hypothesize that reducing copy number at these loci will ablate

changes in gene expression and "recover" the PCB-susceptibility phenotype. CNV_155 appears to drive PCB-sensitivity in the TU strain, which manifests as a very low tolerance to PCB (developmental toxicity EC_{50} = 9 ppb PCB-126[44]). Conversely, CNV_343 appears to drive PCB-sensitivity in the AB strain, which manifests as a higher tolerance (developmental toxicity EC_{50} = 131 ppb PCB-126[44]). We predict that CRISPR-Cas9 targeting to these CNV regions will result in multiple cut sites (Figure 5.1), loss of the duplicated sequence, and cause a reversion to the mean phenotype. We expect CRISPR-mutant (crispant) TU fish to exhibit higher PCB resistance as identified by developmental toxicity EC_{50} values higher than controls and crispant AB fish to exhibit lower PCB tolerance as identified by developmental toxicity EC_{50} values lower than controls.

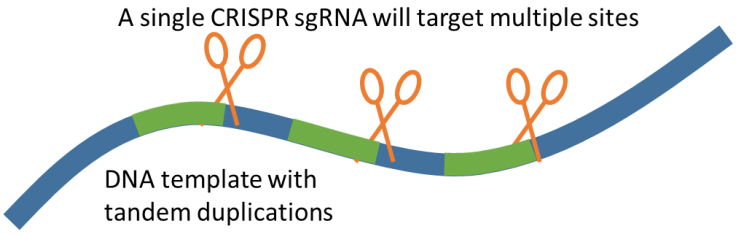


Figure 5.1: CRIPSR-Cas9 targeted at copy number duplicated sites will target multiple loci and produce multiple double strand breaks.

Methods

Zebrafish Care and Husbandry

All procedures were approved by the Portland State University Institutional Animal Care and Utilization Committee in accordance with the Office of Laboratory Animal Welfare's Public Health Service policy. The zebrafish aquatic facility at Portland State University is a recirculating flow-through housing system (Aquaneering). Water temperature is maintained at 28.5°C with average pH of 7.4, average conductivity of 1100 uS, and a 14:10 hour light:dark cycle. Zebrafish are fed flake food (Tetra) supplemented with artemia and rotifers twice daily.

CRISPR target design and microinjection

Embryos from three strains of zebrafish (AB, Tuebingen (TU), and WIK) locally maintained for four generations were pair-mated. Embryos were collected within 15 minutes of fertilization and 1-cell stage embryos were microinjected with 2 nL of ribonucleoprotein (RNP) complex[150]. RNP complex was formed by combining 1 µL [750 ng/µL] sgRNA (Synthego) and 1 µL [3.84 mg/mL] Cas9 (Integrated DNA Technologies) in 10 µL of embryo media (15 mM $CaCl_2 \cdot 2H_2O$, 250 mM NaCl, 10 mM KCl, and 15 mM $MgSO_4$ in ultrapure water) containing 10 ug/mL fluorescein salt and incubating at room temperature for 10 minutes prior to loading injection needles. sgRNA was composed of a 20 bp guide sequence, 3 bp PAM sequence, and an 80-mer SpCas9 scaffold

designed by Synthego (Table 5.1). Injection needles were pulled from thin-wall single-barrel borosilicate glass capillaries with internal filament (ID=0.75 mm, OD=1mm; World Precision Instruments) using a P-80/PC Flaming/Brown Micropipette Puller (Sutter Instrument Co.). Microinjection needle pull settings were as follows: temp=743, pull=60, vel=70, time=200. 10 ug/mL fluorescein salt in embryo media was used as a sham injection control. CRISPR targets were designed using CHOPCHOP[151,152] to fall across identified CNV regions associated with phenotypically variable trait response eQTLs in two strains: CNV_343 in the AB strain and CNV_155 in the TU strain[149]. Successful injection was defined as visible fluorescence in the animal pole of 3-4 hours post fertilization (hpf) embryos and embryos were held in a 28.5°C incubator until phenotype testing at 24 hpf.

Table 5.1: CNV locations and sgRNA target sequence (GRCz10) with PAM region underlined.

Target	CNV location	sgRNA target sequence	sgRNA location
CNV_155	chr5:58079931-58232394	CTGTATACCATTCCCATATTGGG	chr5:58131493-58131515
CNV_343	chr15:44184417-44219174	GCCCATTTAGCACAGGTATTCGG	chr15:44187063-44187257

CRISPR efficiency

For CNV_343 editing, DNA from 10 injected and 3 uninjected AB embryos was extracted using a modified alkaline lysis HotSHOT method

optimized for zebrafish tissues[153]. For CNV_155 editing, DNA from 3 injected, 1 uninjected, and 3 sham injected TU embryos was extracted using the same method. Primers targeting CNV within identified eQTL were designed with the Primer3Plus tool[77] (Table 5.2) to flank the CRISPR site (designated as "OUT" primers) or to directly anneal to the CRISPR target site (designated as "ON"). DNA was assayed in triplicate on a 96-well plate using Brilliant III SYBR Master Mix with ROX (Agilent). The PCR cycling protocol included preliminary dissociation (3 minutes at 95°C) and 40 cycles of annealing and extension (95°C for 5 seconds, 60°C for 20 seconds), per manufacturer's protocol. A dissociation melt curve was also obtained to confirm single qPCR products. Fluorescence was measured with a Stratagene Mx3005P (Agilent) and data were extracted using the Stratagene MxPro data analysis software. Ct values were extracted by treating replicates as individuals and using the adaptive baseline plus moving average algorithm enhancements. Ratios of Ct values for ON primers versus OUT primers (ON:OUT ratio) were used to assess CRISPR efficiency by comparing ON:OUT ratios for injected vs control fish (Equation 5.1)[154].

Equation 5.1: $efficiency = ON{:}OUT_{injected} - ON{:}OUT_{control}$

Table 5.2: qPCR primers for validation of CNV regions identified as eQTL associated with PCB-induced gene expression.

Target	Forward primer	Reverse primer
CNV_155 OUT	AAAAAGGACTGCCGCCAC	AAATGGCAACAAAACAAACAGA
CNV_155 ON	CTGTATACCATTCCCATATT	AAATGGCAACAAAACAAACAGA
CNV_343 OUT	TGGTCCTCCGGAATGGTTTG	TGAATCAGTGACGGTTGGGG
CNV_343 ON	GCCCATTTAGCACAGGTATT	TGAATCAGTGACGGTTGGGG

CRISPR-induced phenotype assessment via PCB exposure

Exposures were performed statically in 96-well plates beginning at 24 hpf through 120 hpf. Each 96-well plate contained a single strain of zebrafish and all exposure treatments, each well contained a single embryo, and plates were maintained in a 28.5°C incubator for the duration of the assay. Randomly assigned exposure treatments (n=12/treatment) included media alone, vehicle (0.3% acetone, v,v), or PCB-126 (3,3',4,4',5-pentachlorobiphenyl) at 1, 5, 25, 125, 625, or 3125 ppb for 24 hours (24-48 hpf). Embryos were exposed from 24 to 48 hpf, exposure media was replaced with clean embryo media at 48 hpf and fifty percent media changes were performed daily thereafter. Larvae were observed for developmental toxicity endpoints daily on an inverted microscope at 40X on a Leica DM IRB inverted microscope with a Leica DFC 450 C digital camera until 120 hpf. Sample sizes for all treatment groups were n=12, with the exception of AB sham-injected and TU CRISPR-injected groups at n=4.

Observations on pericardial edema, yolksac edema, reduced (non-visible) blood flow to the body, and death were performed at 48, 72, 96, and 120 hpf. Additionally, heartrate was measured at 120 hpf. Edema was classified as a 0, 1, 2, 3, or 4 dependent upon the degree of edema where 0 = no edema, 1 = yolksac or pericardial edema, 2 = the presence of two of yolksac or pericardial edema or reduced body blood flow, 3 = yolksac and pericardial edema and reduced body blood flow, 4 = death following any observation of edema at 48, 72, 96, and 120 hpf. Abnormal morphology was scored at 48, 72, 96, and 120 hpf as the presence of curved tail, scoliosis, short body, uneven body symmetry, death, or unhatched at 120 hpf. Relative EC_{50} values were calculated using a probit regression model via the AAT Bioquest EC_{50} calculator (www.aatbio.com/tools/ec50-calculator) from combined heart rate, abnormal morphology, and edema scores at 120 hpf.

Statistical Analyses

Statistical significance of PCB-126 dose-response was determined for heart rate, abnormal morphology, and edema score and compared across groups (uninjected, sham-injected, and CRISRP-injected) within AB and TU separately. Significance of heart rate dose-response was determined using a one-way ANOVA within the base stats package in R using the anova() command with type I sum of squares, followed by a

pairwise T-test with Benjamini-Hochberg correction for multiple comparisons[92]. Significance of abnormal morphology was determined using a Friedman rank sum test for nonparametric statistics on nominal repeated-measures ratio data[155]. Significance of edema score was determined using a one-way repeated measures ANOVA in the base stats package in R using the aov() command followed by a post-hoc Tukey test for pairwise comparisons using least-squares means in the lsmeans package[156] in R.

Results

Injection survival rates and CRISPR efficiency

This experiment is comprised of 2569 embryos across two strains and three injection treatments (Table 5.3). Overall survival of uninjected embryos was 87.9% in AB and 23.7% in TU, which reflects the overall reproductive strategies of the two strains: AB tended to have smaller clutch sizes with higher survival at 24 hpf while TU tended to have large clutch sizes with decreased survival at 24 hpf. 562 AB embryos and 253 TU embryos were subjected to PCB-susceptibility phenotyping. The survival rate of injected embryos was 9.11% less in AB and 9.31% less in TU compared to strain-matched uninjected controls. Addition of CRISPR RNP to the injection mixture resulted in an additional decrease of survival by 42.15% in AB and 10.03% in TU. Average CRISPR

efficiency for CNV_343 RNP was 0.49 (max = 0.85, min = 0.06) and average CRISPR efficiency for CNV_155 RNP was 0.13 (max = 0.60, min = -0.31).

Table 5.3: CRISPR injection survival rates. *Confirmed injection values for naïve AB and TU embryos represent uninjected, fertilized embryos.

strain	injection	spawn pairs	confirmed injections	24 hr survival	survival rate
AB	naïve	2	471*	414	87.90%
AB	sham	2	66	52	78.79%
AB	CNV_343	8	262	96	36.64%
TU	naïve	2	523*	124	23.71%
TU	sham	7	743	107	14.40%
TU	CNV_155	9	504	22	4.37%

Efficacy of PCB-susceptibility phenotype modulation

Heart rate decreased relative to naïve or vehicle-exposed embryos following exposure to PCB-126 at 625 and 3126 ppb in both uninjected and sham-injected AB embryos (BH-adjusted p-value < 0.001; Figure 5.2A). In CRISPR-injected AB embryos, heart rate decreased following exposure to 125, 625, and 3125 ppb PCB-126 relative to controls (BH-adjusted p-value < 0.025). Heart rate decreased relative to naïve or vehicle-exposed embryos following exposure to PCB-126 at 25, 125, 625, and 3126 ppb in both uninjected and sham-injected TU embryos (BH-adjusted p-value < 0.033; Figure 5.2B). In CRISPR-injected TU embryos, there was no difference in heart rate at any exposure level

tested relative to controls (BH-adjusted p-value = 0.6). See Supplemental Tables 5.1 and 5.2 for all pairwise p-values for AB and TU heart rate, respectively.

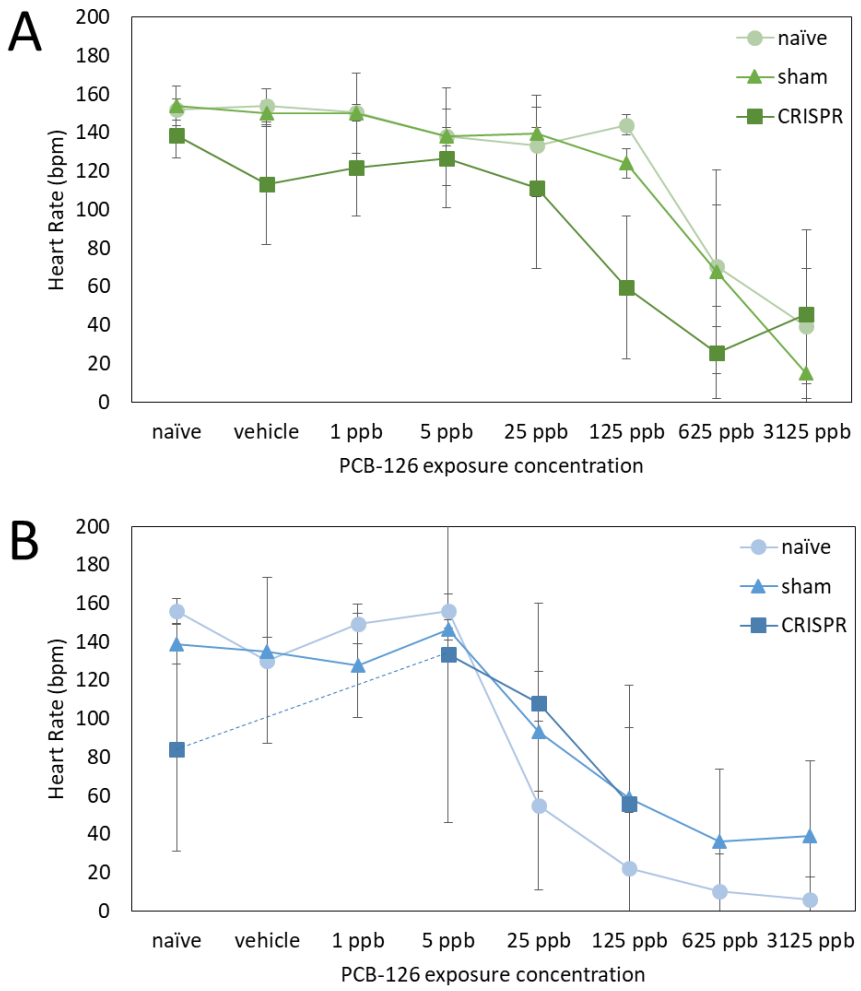


Figure 5.2: Average heart rate at 120 hpf in beats per minute (bpm). Error bars represent 95% confidence intervals. A) AB embryos with CRISPR target to CNV_343, B) TU embryos with CRISPR target to CNV_155. Injection is indicated by color and symbol where shades of green represent strain AB, shades of blue indicate strain TU, circles indicate uninjected embryos, triangles indicate sham-injected embryos, and squares indicate CRISPR-injected embryos.

Abnormal morphology increased relative to naïve or vehicle-exposed embryos following exposure to PCB-126 at 3125 ppb in both uninjected and sham-injected AB embryos (p-value < 0.04). In the uninjected controls, 75% of AB embryos exposed to 3125 ppb PCB-126 showed abnormal morphology while all other treatment groups were no different from the controls (Figure 5.3A). This pattern was conserved in sham-injected AB embryos (Figure 5.3C). In CRISPR-injected AB embryos, abnormal morphology increased following exposure to 25, 125, 625, and 3125 ppb PCB-126 relative to controls (p-value = 0.0020). CRISPR-injected embryos at 25 ppb, 125 ppb, 625 ppb, and 3125 ppb PCB-126 all had higher percentages of abnormal morphology (Figure 5.3E) relative to naïve or vehicle-exposed controls indicating efficacy of CNV_343-targeted RNP to shift the dose-response curve of abnormal morphology to lower doses than controls.

Abnormal morphology increased relative to naïve or vehicle-exposed embryos following exposure to PCB-126 at 25, 125, 625, and 3125 ppb in both uninjected TU embryos, but this was not deemed significant in a Friedman rank sum test. Abnormal morphology in uninjected TU was observed in 64% of 25 ppb PCB-exposed, 75% of 125 ppb PCB-exposed, and 92% of 625 and 3125 ppb PCB-exposed embryos (Figure 5.3B) following the expected higher PCB-sensitivity phenotype as compared to AB. In sham-injected TU embryos, abnormal

morphology increased following exposure to 1, 25, 125, 625, and 3125 ppb PCB-126 relative to controls (p-value = 0.0440, Figure 5.3D). In CRISPR-injected TU embryos only embryos exposed to 125 ppb PCB-126 had higher abnormal morphology percentages than controls, but due to overall toxicity of the CNV_155 RNP (Table 5.3), 88% abnormal morphology in the control group (Figure 5.3F), and the fact that only 4 exposure groups were tested (naïve, 5 ppb, 25 ppb, and 125 ppb), the data from CRISPR-injected TU are fairly uninformative probably because the CNV_155 RNP was highly detrimental to normal development regardless of PCB exposure.

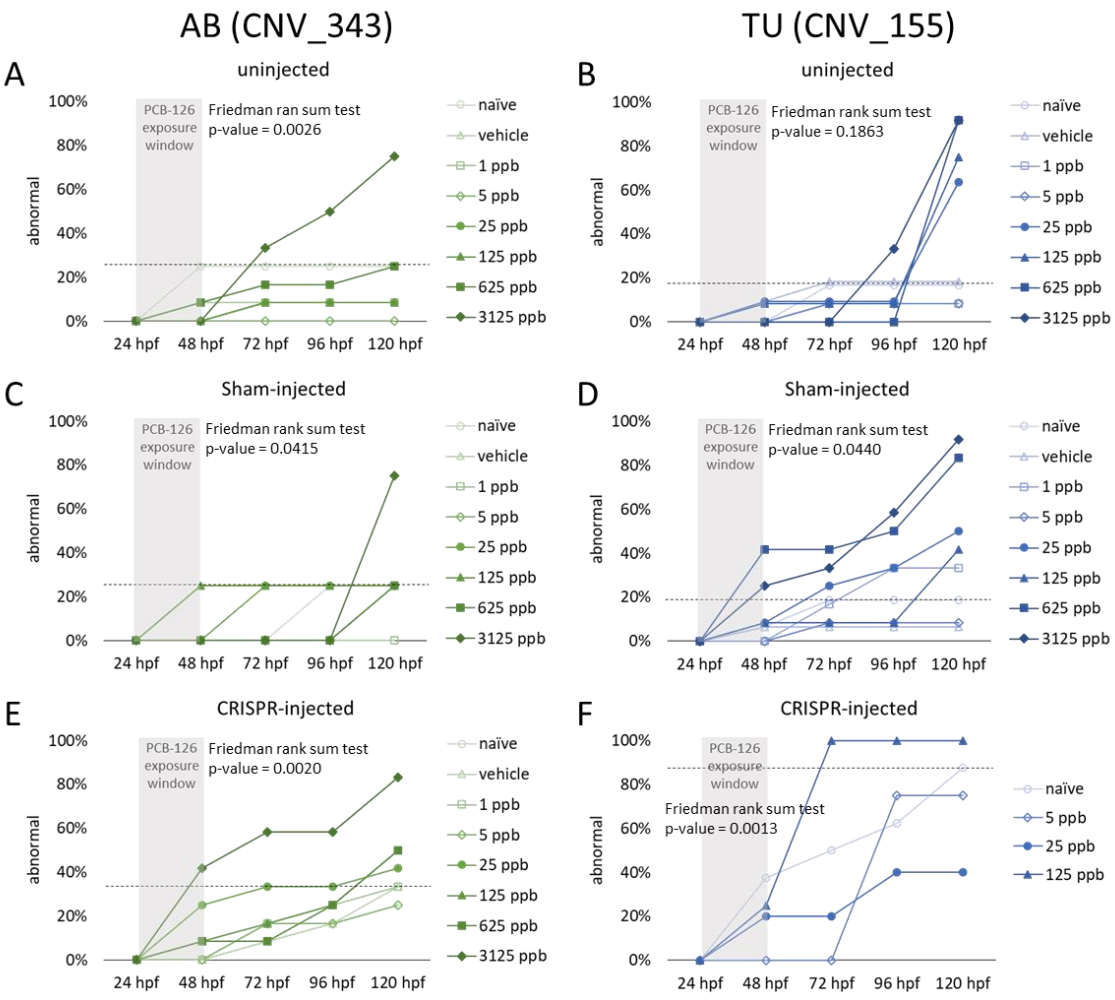


Figure 5.3: Percent of larvae with abnormal morphology. Larvae from strain AB (A,C,E) and TU (B,D,F) were either uninjected (A,B), sham-injected (C,D), or CRISPR-injected at the one-cell stage, exposed to PCB-126 from 24-48 hpf at 1-3125 ppb, and assessed for abnormal morphology at 24 hour intervals until 120 hpf. P-values for significance determined using the nonparametric Friedman rank sum test. Dashed line indicates the highest abnormal morphology percentage for controls in each group.

Average edema score increased relative to naïve or vehicle-exposed embryos following exposure to PCB-126 at 625 and 3125 ppb by 72 hpf in uninjected AB embryos (p-value < 2.2e-16) with average edema scores of 3.2 and 3.5 at 120 hpf, respectively (Figure 5.4A). This

pattern was conserved in sham-injected AB embryos (p-value < 0.0047), with the addition of an increase in edema scores in embryos treated with 125 ppb PCB-126 (p-value = 3.27E-08; Figure 5.4C). In CRISPR-injected AB embryos, average edema score increased following exposure to 125, 625, and 3125 ppb PCB-126 relative to controls by 72 hpf (p-value < 0.0025; Figure 5.4E). Overall this represents a shift to earlier and higher toxicity of PCB-126 in CNV_343-targeted embryos versus sham or uninjected controls. See Supplemental Tables 5.3-5.5 for all pairwise p-values for AB edema scores.

Average edema score increased relative to naïve or vehicle-exposed embryos following exposure to PCB-126 at 25, 125, 625 and 3125 ppb by 72 hpf in uninjected and sham-injected TU embryos (p-value < 0.0003) with average edema scores between 3.6-3.92 in uninjected embryos and between 2.36-3.3 in sham-injected embryos at 120 hpf (Figures 5.4B,D). At 120 hpf edema scores in uninjected embryos were significant in the 5 ppb PCB-exposed embryos (p-value < 3.61-06), but not in sham-injected embryos. In CRISPR-injected TU embryos, average edema score increased following exposure to 125 ppb PCB-126 relative to controls at 96 hpf (p-value < 0.03.8; Figure 5.4F), but due to overall toxicity of the CNV_155 RNP (Table 5.3), high levels of edema in the unexposed group, and the fact that only 4 exposure groups were tested, the data from CRISPR-injected TU are not indicative

of CRISPR-specific effects. See Supplemental Tables 5.6-5.8 for all pairwise p-values for AB edema scores.

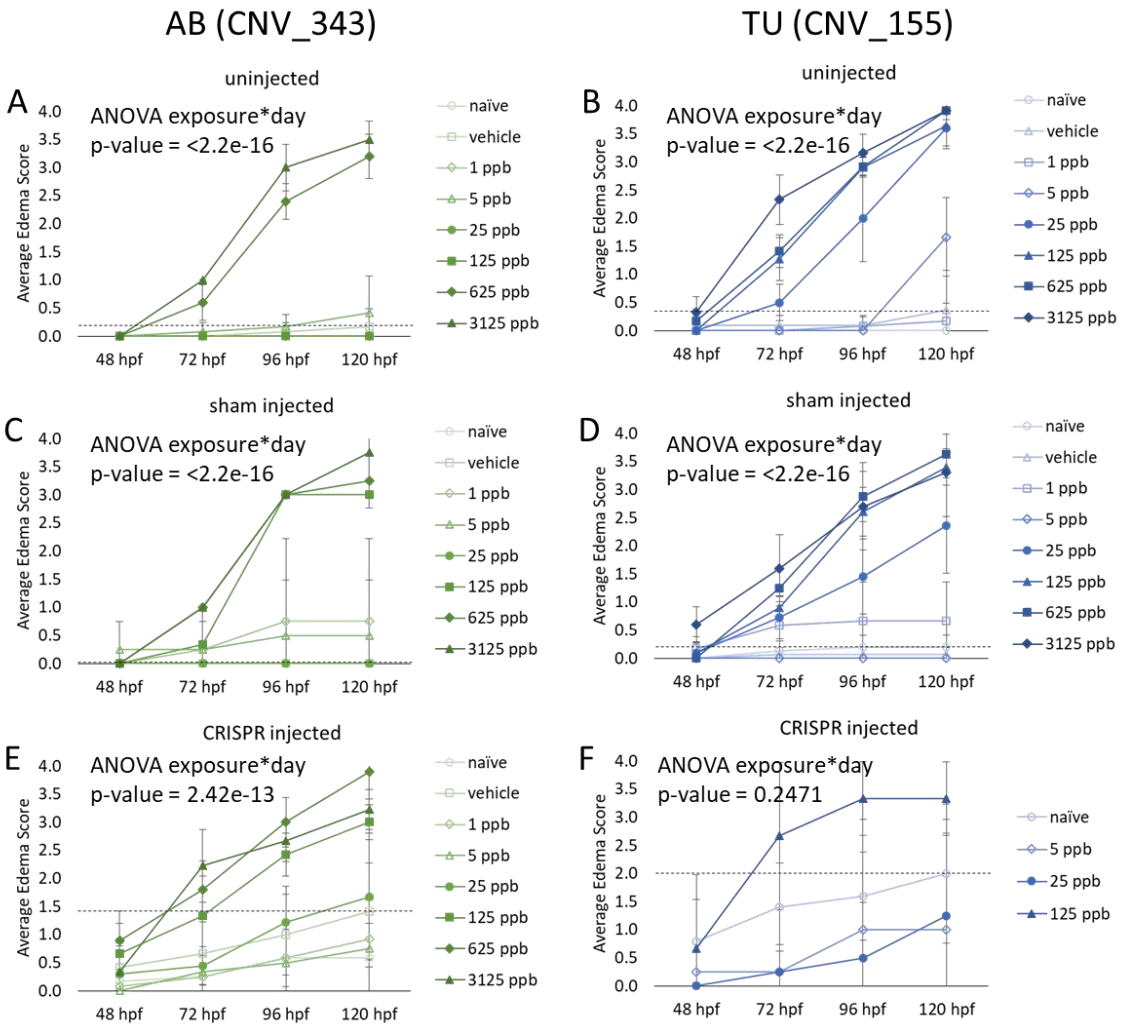


Figure 5.4: Average edema score from 48-120 hpf. Larvae from strain AB (A,C,E) and TU (B,D,F) were either uninjected (A,B), sham-injected (C,D), or CRISPR-injected at the one-cell stage, exposed to PCB-126 from 24-48 hpf at 1-3125 ppb, and assessed for edema at 24 hour intervals until 120 hpf. P-values for significance of the interaction between the exposure level and day to cause changes in average edema score determined using repeated-measures ANOVA. Dashed line indicates the highest average edema score for controls in each group.

By combining heart rate, abnormal morphology, and edema scores from 120 hpf, we calculated EC_{50} values of 512.8 ppb PCB-126 in uninjected AB embryos and 22.5 ppb PCB-126 in uninjected TU embryos (Figure 5.5). Sham injections slightly increased EC_{50} values in both strains, indicating minor toxicity of the injection procedure. CRISPR injections in AB targeting CNV_343 reduced the EC_{50} by a factor of 10 to 67.9 ppb PCB-126 relative to sham controls while CRISPR injections in TU targeting CNV_155 only slightly increased EC_{50} to 47.0 ppb PCB-126 (Table 5.4).

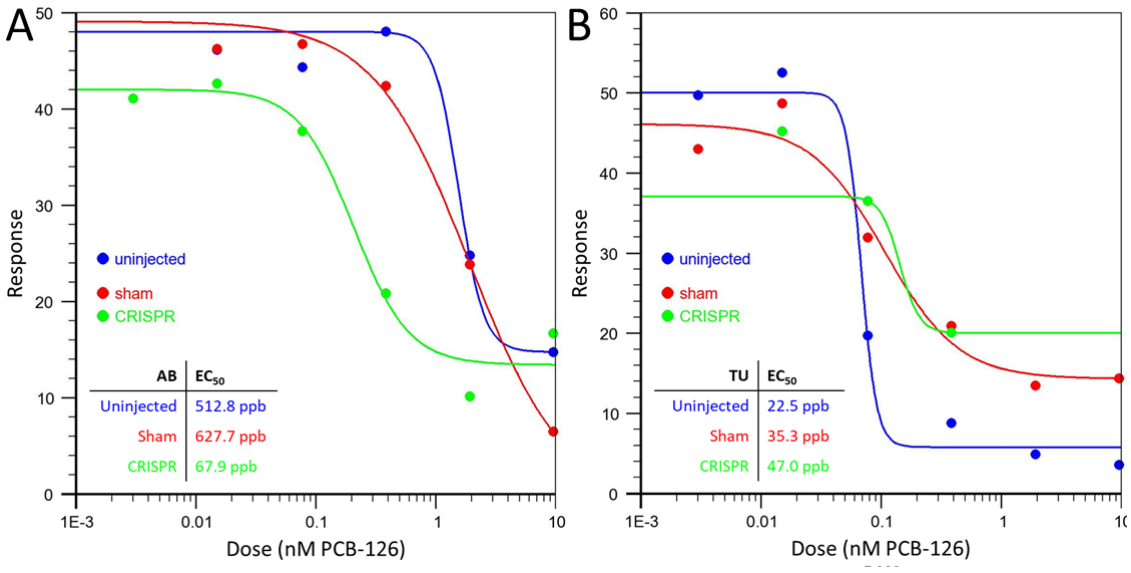


Figure 5.5: Dose-response curves for PCB-126 exposure and calculated EC_{50} values using heart rate, abnormal morphology, and edema scores at 120 hpf in AB (A) and TU (B) embryos. CRISPR target in AB = CNV_343; CRISPR target in TU = CNV_155.

Table 5.4: Calculated EC_{50} values for PCB-126 in nM and ppb with dose-response equation.

	EC_{50} (nM PCB-126)	EC_{50} (ppb PCB-126)	Dose-response curve
AB uninjected	1.571	512.8	$y = 14.739 + \frac{48.011-14.739}{1+\left(\frac{x}{1.571}\right)^{4.214}}$
AB sham-injected	1.923	627.7	$y = -1.024 + \frac{49.078+1.024}{1+\left(\frac{x}{1.923}\right)^{1.086}}$
AB CRISPR-injected	0.208	67.9	$y = 13.403 + \frac{42.005-13.403}{1+\left(\frac{x}{0.208}\right)^{1.892}}$
TU uninjected	0.069	22.5	$y = 5.783 + \frac{50.040-5.783}{1+\left(\frac{x}{0.069}\right)^{6.735}}$
TU sham-injected	0.108	35.3	$y = 14.306 + \frac{46.115-14.306}{1+\left(\frac{x}{0.108}\right)^{1.412}}$
TU CRISPR-injected	0.144	47.0	$y = 20.020 + \frac{37.107-20.020}{1+\left(\frac{x}{0.144}\right)^{5.376}}$

Discussion

By targeting CNV_343, a region that is duplicated in the AB strain and associated with the PCB-susceptibility phenotype, we were successfully able to reduce the overall sensitivity of crispant AB zebrafish to PCB-126. We observed AB zebrafish to have an EC_{50} of 512.8 ppb PCB-126. Sham injections increased the EC_{50} slightly to 627.7 ppb PCB-126, but this may be a false increase due to small sample sizes in the AB sham injection groups (n=4/exposure). Injection of CNV_343 RNP resulted in a reduction of EC_{50} by a factor of 10 to 67.9 ppb PCB-126 in crispant AB zebrafish, which supports our hypothesis that targeting the duplicated genomic region with CRISPR-Cas9 would result in lower tolerance of PCB-126. We were unable to fully replicate this in TU zebrafish using CNV_155 as a target, most likely due to the highly

lethal effects of the CNV_155 RNP and an overall smaller sample size, but the resulting trend towards an increased EC_{50} is in agreement with our predicted results for CNV_155 crispants. Further optimization of sgRNA:Cas9 ratios and total RNP concentration may ameliorate this mortality[154].

CRISPR-Cas9 genome editing is rapidly becoming a premier tool in genetic manipulation. Precise targeting is more and more commonly used to knock down gene expression in a wide variety of model organisms and cell cultures. CNV have proven to be a nuisance in large CRISPR screens by generating false-positives due to multiple target cut sites[157–159], but we have utilized this phenomenon to specifically target a single duplicated region to remove all or some of the duplicated sequence (Figure 5.1). Our data are not able to reveal the mechanistic drivers of CRISPR-Cas9 CNV targeted response, but we hypothesize that CRISPR-Cas9 either induced multiple indels or SNPs at CNV target sites or simultaneous double-strand breaks resulted in complete excision of larger genomic regions. Nevertheless, we were able to experimentally ablate the effects of a non-genic CNV and alter a phenotype.

Conclusion

To our knowledge, this is the first use of CRISPR-Cas9 technology to target a CNV-driven phenotype. We were able to show clear and

statistically significant changes in developmental toxicity following exposure to PCB-126 in crispant AB zebrafish, and to a lesser extent, changes in crispant TU zebrafish. This study serves as a proof-of-principle that CRISPR can be successfully used to target CNV and shows that the PCB-susceptibility phenotype is not solely driven by modifications to AHR, but is also driven by non-canonical endpoints. Furthermore, this study highlights the importance of understanding variation across populations. Genetic variation in human populations directly affect how we model the effects and prioritize clean-up of environmental toxicants. By confirming the interaction of CNV on PCB-susceptibility, we hope to continue to advance our understanding of the complex relationship between genetic variation and toxicant susceptibility.

Acknowledgments

The authors acknowledge Dr. Adam Miller at the University of Oregon and Dr. Carrie Hanna at the Oregon National Primate Research Center for advice on the CRISPR-Cas9 system. We thank Drs. Jason Podrabsky, Suzanne Estes, Brad Buckley, and Randy Zelick for sharing reagents and equipment, Drs. Amie LT Romney and Claire L Riggs for microinjection training and advice, and Jobe Ritchie for perfecting the microinjection needle pulling protocol.

Chapter 6

A Summary of Findings

In the first chapter I introduced the four step risk assessment process: hazard identification, dose-response assessment, exposure assessment, and risk characterization. As part of the dose-response assessment, regulators develop a reference dose for safe levels of exposure using toxicity data from existing studies and the incorporation of several uncertainty factors. One of the weaknesses of this approach is the assumptions made while incorporating uncertainty factors. This dissertation aims to address the human variation uncertainty factor by investigating the effects of genomic CNV, a vastly understudied type of variation, on toxicant susceptibility. By fully understanding the genomic drivers of toxicant susceptibility, we can better predict the impacts of intentional (pharmacologic) and unintentional (occupational or environmental) exposures and incorporate our understanding into clean-up and prevention for sensitive populations.

I begin by expanding our working knowledge of CNV in zebrafish. Prior knowledge of CNV in zebrafish focused on intrastrain (within-strain) variation, but did not focus on interstrain (across-strain) variation. Expanding our knowledge of genomic differences across zebrafish strains is critical for application mechanisms that drive human

variation. By using a reciprocal comparison study design, I assessed CNV across three common strains of zebrafish: AB, Tuebingen, and WIK. I identified 1351 CNV that cover 1.9% of the most recent genome assembly (GRCz11), created a publically available track in the UCSC Genome Browser for CNV visualization and exploration, and predicted the transcriptional effects of the CNV. Using an extremely conservative ± 5 Kb window, I identified 2200 genes that are likely to be affected (directly via overlap, or indirectly via transcriptional regulation) by CNV, illustrating the large impact of CNV on transcriptional variation across populations.

Knowledge of genomic variation across strains is not useful in isolation. To move our understanding forward, I characterized the extent of mRNA expression variation across the same strains that I had assessed for CNV. One main finding from this work was that male and female mRNA expression in the liver is vastly different. Moving forward, I recommend that all analysis using adult zebrafish be partitioned into male and female datasets, as the overlap of differentially expressed mRNA between the sexes is only 14%. I found 269 mRNA transcript in males and 212 mRNA transcripts in females that differed significantly across strains. Lipid transport was over-represented in the differentially expressed mRNA datasets, indicating that strains may use different mechanisms for transport and storage of lipids. This has important

consequences in the partitioning, sequestration, and transport of lipophilic compounds, including PCBs.

To evaluate the interactions between CNV and mRNA expression, I assessed the eQTL of the toxicant-susceptibility phenotype that varies across strains. To do this I exposed adult zebrafish to micro-doses of PCB-126 (130 ppt = 130 ng/L = 0.4 pM) and then identified PCB-induced mRNA expression. Using paired CNV and mRNA expression data from 54 individuals, I was able to identify three statistically significant eQTL. I then mapped exposure status (unexposed, vehicle control, or PCB-exposed) across CNV and mRNA expression plots to identify response eQTL (reQTL). Using this technique I narrowed my list of QTL down to two strain-specific mechanisms: one involving *prpf4* in the Tuebingen strain and the other involving *dync2h1* in the AB strain. This is an exciting peek into CNV-based eQTL. As more and more data become available on CNV across species, I expect a burst of discovery around CNV as drivers of both simple and complex phenotypes.

To test my identified reQTL as drivers of the toxicant-susceptibility phenotype, I used targeted CRISPR-Cas9 editing to modify CNV and then assessed developmental toxicity of PCB-126. Both reQTL identified CNV duplications that drove a resistant phenotype in AB and a sensitive phenotype in Tuebingen. By removing or reducing the CNV duplications with CRISPR-Cas9, I hypothesized that I would be able to reverse the

susceptibility phenotype. Specifically, AB crispants would be less resistant (more sensitive) and Tuebingen crispants would be less sensitive (more resistant) to PCB-126 than their non-modified brethren. And indeed, the developmental toxicity of PCB-126—as measured by levels of edema, heart rate, and abnormal morphology—was reversed in both strains. In AB crispants the EC_{50} was reduced by a factor of 10 and in Tuebingen crispants the EC_{50} was increased by 33%, although increases in Tuebingen were less statistically significant due to high levels of toxicity. This experiment was a solid proof-of-principle than CNV are directly involved in the PCB-susceptibility phenotype and that the mechanistic drivers of this phenotype vary across strains.

Overall this work aims to illuminate how genetic variation affects phenotype, especially in relation to toxicant susceptibility, using a model organism to characterize and manipulate a complex phenotype. Broadly, my goal is to improve the risk assessment process by refining the human variation uncertainty factor in hazard assessment. This is not a trivial task. Human variation is vast and complicated. Not only do we vary at millions of nucleotides, we also vary in our transcriptional response networks. To simplify my approach, and to be able to directly modify and test genotype-phenotype interactions, I used the zebrafish as a model system. To use this system I first needed to define the existing variation in CNV genotype and baseline transcription, then perturb the

system with PCBs and identify the transcriptional response. Once completing those monumental tasks, I identified genetic factors that drive the PCB-susceptibility phenotype, modified them using CRISPR-Cas9 genome editing, and tested the resulting phenotype. While I have not yet attained my overarching goal of improving the risk assessment process, I have clearly shown that CNV are important drivers of the toxicant-susceptibility phenotype and can continue in this vein of inquiry as I move forward in my career.

References

1. National Research Council, C. on the I. M. for A. of R. to P. H. *Risk Assessment in the Federal Government*. (1983). doi:10.17226/366
2. Borzelleca, J. F. Paracelsus: Herald of Modern Toxicology. *Toxicol. Sci.* **53,** 2–4 (2000).
3. EPA, U. S. *Benchmark Dose Technical Guidance*. (2012).
4. US EPA. Guidelines for Exposure Assessment. *Risk Assess. Forum* **57,** 22888–22938 (1992).
5. Fowle, J. & Dearfield, K. Science Policy Council Handbook: Risk Characterization. 189 (2000).
6. Warren, R. B. *et al.* Genetic variation in efflux transporters influences outcome to methotrexate therapy in patients with psoriasis. *J. Invest. Dermatol.* **128,** 1925–1929 (2008).
7. Zanger, U. M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.* **138,** 103–141 (2013).
8. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
9. Consortium, G. R. Human Genome Assembly GRCh38.p12. (2017). Available at: https://www.ncbi.nlm.nih.gov/grc/human.
10. Gibbs, R. A. *et al.* The international HapMap project. *Nature* **426,** 789–796 (2003).
11. International HapMap Consortium & The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–320 (2005).
12. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
13. Consortium, T. I. H. 3. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
14. Siva, N. 1000 Genomes Project. *Nat. Biotechnol.* **26,** 256 (2008).
15. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
16. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
17. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy number variation. *Science (80-.).* **349,** 1181, aab3761 (2015).
18. Haraksingh, R. R. & Snyder, M. P. Impacts of variation in the human genome on gene regulation. *J. Mol. Biol.* **425,** 3970–7

(2013).
19. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49,** 692–699 (2017).
20. Abe, A. *et al.* Compound heterozygous PMP22 deletion mutations causing severe Charcot-Marie-Tooth disease type 1. *J. Hum. Genet.* **55,** 771–773 (2010).
21. Bayés, M., Magano, L. F., Rivera, N., Flores, R. & Pérez Jurado, L. A. Mutational Mechanisms of Williams-Beuren Syndrome Deletions. *Am. J. Hum. Genet* **73,** 131–151 (2003).
22. Jeon, J. *et al.* Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus. *BMC Genomics* **11,** 426 (2010).
23. Hollox, E. J. *et al.* Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.* **40,** 23–5 (2008).
24. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39,** 1256–60 (2007).
25. He, Y., Hoskins, J. M. & McLeod, H. L. Copy number variants in pharmacogenetic genes. *Trends Mol. Med.* **17,** 244–251 (2011).
26. Abdo, N. *et al.* Population-based in vitro hazard and concentration-response assessment of chemicals: The 1000 genomes high-throughput screening study. *Environ. Health Perspect.* **123,** 458–466 (2015).
27. Joesch-Cohen, L. M. & Glusman, G. Difference between the genomes of lymphoblastoid cell lines and blood-derived samples. *Adv Genomics Genet* **7,** 1–9 (2017).
28. World Health Organization. *Harmonization Project Document 11 GUIDANCE DOCUMENT ON EVALUATING AND EXPRESSING UNCERTAINTY IN HAZARD CHARACTERIZATION*. (2014). doi:ISBN 978 92 4 150761 5
29. Prince, R. & Cooper, K. R. Comparisons of the effects of 2,3,7,8-tetrachlorodibenzo- p -dioxin on chemically impacted and nonimpacted subpopulations of Fundulus heteroclitus : I TCDD toxicity. *Environ. Toxicol. Chem.* **14,** 579–587 (1995).
30. Meyer, J. N., Nacci, D. E. & Di Giulio, R. T. Cytochrome P4501A (CYP1A) in killifish (Fundulus heteroclitus): Heritability of altered expression and relationship to survival in contaminated sediments. *Toxicol. Sci.* **68,** 69–81 (2002).
31. Nacci, D. *et al.* Adaptations of wild populations of the estuarine fish Fundulus heteroclitus to persistent environmental contaminants. *Mar. Biol.* **134,** 9–17 (1999).
32. Yuan, Z., Courtenay, S., Chambers, R. C. & Wirgin, I. Evidence of Spatially Extensive Resistance to PCBs in an Anadromous Fish of the Hudson River. *Environ. Health Perspect.* **114,** 77–84 (2006).

33. Whitehead, A., Triant, D. A., Champlin, D. & Nacci, D. Comparative transcriptomics implicates mechanisms of evolved pollution tolerance in a killifish population. *Mol. Ecol.* **19,** 5186–5203 (2010).
34. Hall, J. The Aryl-hydrocarbon Receptor (AhR) as a Therapeutic Target in Human Breast Cancer. *J Steroids Horm. Sci* **5,** (2014).
35. Mandal, P. K. Dioxin: a review of its environmental effects and its aryl hydrocarbon receptor biology. *J. Comp. Physiol. B.* **175,** 221–30 (2005).
36. Whitehead, A., Pilcher, W., Champlin, D. & Nacci, D. Common mechanism underlies repeated evolution of extreme pollution tolerance. *Proc. Biol. Sci.* **279,** 427–33 (2012).
37. Clark, B. W., Matson, C. W., Jung, D. & Di Giulio, R. T. AHR2 mediates cardiac teratogenesis of polycyclic aromatic hydrocarbons and PCB-126 in Atlantic killifish (Fundulus heteroclitus). *Aquat. Toxicol.* **99,** 232–240 (2010).
38. Reitzel, A. M. *et al.* Genetic variation at aryl hydrocarbon receptor (AHR) loci in populations of Atlantic killifish (Fundulus heteroclitus) inhabiting polluted and reference habitats. *BMC Evol. Biol.* **14,** 6 (2014).
39. Wirgin, I. *et al.* Mechanistic basis of resistance to PCBs in Atlantic tomcod from the Hudson River. *Science* **331,** 1322–1325 (2011).
40. Waits, E. R. *et al.* Genetic Linkage Map and Comparative Genome Analysis for the Atlantic Killifish (Fundulus heteroclitus). *Open J. Genet. Open J. Ge-netics* **6,** 28–38 (2016).
41. Nacci, D., Proestou, D., Champlin, D., Martinson, J. & Waits, E. R. Genetic basis for rapidly evolved tolerance in the wild: adaptation to toxic pollutants by an estuarine fish species. *Mol. Ecol.* **25,** 5467–5482 (2016).
42. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn. an Off. public* **203,** 253–310 (1995).
43. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496,** 498–503 (2013).
44. Waits, E. R. & Nebert, D. W. Genetic architecture of susceptibility to PCB126-induced developmental cardiotoxicity in zebrafish. *Toxicol. Sci.* **122,** 466–75 (2011).
45. Balik-Meisner, M., Truong, L., Scholl, E. H., Tanguay, R. L. & Reif, D. M. Population genetic diversity in zebrafish lines. *Mamm. Genome* **0,** 1–11 (2018).
46. Brown, K. H. *et al.* Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 529–34 (2012).

47. Reid, N. M. *et al.* The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science (80-.).* **354,** 1305 LP-1308 (2016).
48. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215,** 403–410 (1990).
49. Shiu, W. Y. & Mackay, D. A Critical Review of Aqueous Solubilities, Vapor Pressures, Henry's Law Constants, and Octanol–Water Partition Coefficients of the Polychlorinated Biphenyls. *J. Phys. Chem. Ref. Data* **15,** 911–929 (1986).
50. Crinnion, W. Polychlorinated biphenyls: persistent pollutants with immunological, neurological, and endocrinological consequences. *Altern. Med. Rev. a J. Clin. …* **16,** 5–13 (2011).
51. Danis, B. *et al.* Contrasting effects of coplanar versus non-coplanar PCB congeners on immunomodulation and CYP1A levels (determined using an adapted ELISA method) in the common sea star Asterias rubens L. *Aquat. Toxicol.* **69,** 371–383 (2004).
52. Bemis, J. C., Nazarenko, D. A. & Gasiewicz, T. A. Coplanar polychlorinated biphenyls activate the aryl hydrocarbon receptor in developing tissues of two TCDD-responsive lacZ mouse lines. *Toxicol. Sci.* **87,** 529–536 (2005).
53. US EPA. *Polychlorinated Biphenyls 1929-1979: Final Report*. (1979).
54. Van den Berg, M. The 2005 World Health Organization Reevaluation of Human and Mammalian Toxic Equivalency Factors for Dioxins and Dioxin-Like Compounds. *Toxicol. Sci.* **93,** 223–241 (2006).
55. Ogura, I. Half-life of each dioxin and PCB congener in the human body. *Organohalogen Compd.* **66,** 3329–3337 (2004).
56. US Department of Health and Human Services, C. for D. C. and P. Fourth National Report on Human Exposure to Environmental Chemicals: Updated Tables, January 2017, Volume Two. **2,** (2017).
57. Di Paolo, C. *et al.* Early life exposure to PCB126 results in delayed mortality and growth impairment in the zebrafish larvae. *Aquat. Toxicol.* **169,** 168–178 (2015).
58. Şişman, T., Geyikoğlu, F. & Atamanalp, M. Early life-stage toxicity in zebrafish (Danio rerio) following embryonal exposure to selected polychlorinated biphenyls. *Toxicol. Ind. Health* **23,** 529–536 (2007).
59. Liu, H. *et al.* Developmental toxicity, oxidative stress, and related gene expression induced by dioxin-like PCB 126 in zebrafish (Danio rerio). *Environ. Toxicol.* 1–9 (2014).

doi:10.1002/tox.22044
60. Butler, M. G. *et al.* SNPfisher: tools for probing genetic variation in laboratory-reared zebrafish. *Development* **142,** 1–11 (2015).
61. Guryev, V. *et al.* Genetic variation in the zebrafish. *Genome Res.* **16,** 491–7 (2006).
62. Coe, T. S. *et al.* Genetic variation in strains of zebrafish (Danio rerio) and the implications for ecotoxicology studies. *Ecotoxicology* **18,** 144–50 (2009).
63. Faber-Hammond, J. J. & Brown, K. H. Pseudo- *De Novo* Assembly and Analysis of Unmapped Genome Sequence Reads in Wild Zebrafish Reveal Novel Gene Content. *Zebrafish* **13,** zeb.2015.1154 (2016).
64. Wilson, C. A. *et al.* Wild sex in zebrafish: Loss of the natural sex determinant in domesticated strains. *Genetics* **198,** 1291–1308 (2014).
65. Drew, R. E. *et al.* Brain transcriptome variation among behaviorally distinct strains of zebrafish (Danio rerio). *BMC Genomics* **13,** 323 (2012).
66. Scerbina, T., Chatterjee, D. & Gerlai, R. Dopamine receptor antagonism disrupts social preference in zebrafish: A strain comparison study. *Amino Acids* **43,** 2059–2072 (2012).
67. Wong, R. Y. *et al.* Comparing behavioral responses across multiple assays of stress and anxiety in zebrafish (Danio rerio). *Behaviour* **149,** 1205–1240 (2012).
68. Lange, M. *et al.* Inter-Individual and Inter-Strain Variations in Zebrafish Locomotor Ontogeny. *PLoS One* **8,** (2013).
69. Meyer, B. M., Froehlich, J. M., Galt, N. J. & Biga, P. R. Inbred strains of zebrafish exhibit variation in growth performance and myostatin expression following fasting. *Comp. Biochem. Physiol. - A Mol. Integr. Physiol.* **164,** 1–9 (2013).
70. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16,** 172–183 (2015).
71. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47,** 296–303 (2015).
72. Stranger, B., Forrest, M. & Dunning, M. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (80-.).* **315,** 848–853 (2007).
73. Schulze, J. J. *et al.* Doping test results dependent on genotype of uridine diphospho-glucuronosyl transferase 2B17, the major enzyme for testosterone glucuronidation. *J. Clin. Endocrinol. Metab.* **93,** 2500–2506 (2008).
74. Buchanan, J. A. & Scherer, S. W. Contemplating effects of

genomic structural variation. *Genet. Med.* **10,** 639–647 (2008).
75. Wain, L. V., Armour, J. AL & Tobin, M. D. Genomic copy number variation, human health, and disease. *Lancet* **374,** 340–350 (2009).
76. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** D764-70 (2014).
77. Untergasser, A. *et al.* Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **40,** 1–12 (2012).
78. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402–8 (2001).
79. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 1–14 (2016).
80. Patowary, A. *et al.* A Sequence-Based Variation Map of Zebrafish. *Zebrafish* **10,** 15–20 (2013).
81. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
82. Staff, Z. Mutation Details Curation of Older Features - Wild -Type Line : AB. *ZFIN Historical Data* (2016). Available at: http://zfin.org/ZDB-GENO-960809-7.
83. Haffter, P. *et al.* The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development* **123,** 1–36 (1996).
84. Mullins, M. C., Hammerschmidt, M., Haffter, P. & Nusslein-Volhard, C. Large-scale mutagenesis in the zebrafish: In search of genes controlling development in a vertebrate. *Curr. Biol.* **4,** 189–202 (1994).
85. Rauch, G.-J., Granato, M. & Haffter, P. A polymorphic zebrafish line for genetic mapping using SSLPs on high-percentage agarose gels. *Tech. Tips Online T01208* **2,** 148–150 (1997).
86. Vasilevsky, N. A. *et al.* On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* **1,** e148 (2013).
87. Wong, R. Y., McLeod, M. M. & Godwin, J. Limited sex-biased neural gene expression patterns across strains in Zebrafish (Danio rerio). *BMC Genomics* **15,** 905 (2014).
88. Bowen, M. E., Henke, K., Siegfried, K. R., Warman, M. L. & Harris, M. P. Efficient mapping and cloning of mutations in zebrafish by low-coverage whole-genome sequencing. *Genetics* **190,** 1017–1024 (2012).
89. Gasch, A. P., Payseur, B. A. & Pool, J. E. The Power of Natural Variation for Model Organism Biology. *Trends Genet.* **32,** 147–154 (2016).

90. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47–e47 (2015).
91. Team, R. C. & Computing, R. F. for S. R: A Language and Environment for Statistical Computing. (2016).
92. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).
93. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **45,** D183–D189 (2017).
94. Dunn, O. J. Multiple Comparisons Among Means. *J. Am. Stat. Assoc.* **56,** 52–64 (1961).
95. Leary, R. F., Allendorf, F. W., Knudsen, K. L. & Thorgaard ', G. H. Heterozygosity and developmental stability in gynogenetic diploid and triploid rainbow trout. *Heredity (Edinb).* **54,** 219–225 (1985).
96. Nei, M., Maruyama, T. & Chakraborty, R. The Bottleneck Effect and Genetic Variability in Populations. *Evolution (N. Y).* **29,** 1–10 (1975).
97. Romero, I., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* **13,** 505–516 (2012).
98. Bouzat, J. L. Conservation genetics of population bottlenecks: The role of chance, selection, and history. *Conserv. Genet.* **11,** 463–478 (2010).
99. Wilson, C. A. *et al.* Wild sex in zebrafish: Loss of the natural sex determinant in domesticated strains. *Genetics* **198,** 1291–1308 (2014).
100. Ben-Moshe, Z. *et al.* The light-induced transcriptome of the zebrafish pineal gland reveals complex regulation of the circadian clockwork by light. *Nucleic Acids Res.* **42,** 3750–3767 (2014).
101. Blake, J. A. *et al.* Gene ontology consortium: Going forward. *Nucleic Acids Res.* **43,** D1049–D1056 (2015).
102. Soulban, G. & Labrecque, G. Circadian rythms of blood clotting time and coagulation factors II, VII, and X in rats. *Life Sci.* **45,** 2485–2489 (1989).
103. Bertolucci, C. *et al.* Evidence for an overlapping role of CLOCK and NPAS2 transcription factors in liver circadian oscillators. *Mol Cell Biol* **28,** 3070–3075 (2008).
104. Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proc. Natl. Acad. Sci.* **111,** 16219–16224 (2014).

105. Pizarro, A., Hayer, K., Lahens, N. F. & Hogenesch, J. B. CircaDB: A database of mammalian circadian gene expression profiles. *Nucleic Acids Res.* **41,** 1009–1013 (2013).
106. Helfrich-Förster, C. Differential Control of Morning and Evening Components in the Activity Rhythm of *Drosophila melanogaster* —Sex-Specific Differences Suggest a Different Quality of Activity. *J. Biol. Rhythms* **15,** 135–154 (2000).
107. Bur, I. M. *et al.* The circadian clock components CRY1 and CRY2 are necessary to sustain sex dimorphism in mouse liver metabolism. *J. Biol. Chem.* **284,** 9066–9073 (2009).
108. Davey, C., Tallafuss, A. & Washbourne, P. Differential Expression of Neuroligin Genes in the Nervous System of Zebrafish. 703–714 (2010). doi:10.1002/dvdy.22195
109. Rabaneda, L. G., Robles-Lanuza, E., Nieto-González, J. & Scholl, F. G. Neurexin Dysfunction in Adult Neurons Results in Autistic-like Behavior in Mice. *Cell Rep.* **8,** 338–346 (2014).
110. Jamain, S. *et al.* Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nat. Genet.* **34,** 27–9 (2003).
111. Raeder, H. *et al.* Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nat. Genet.* **38,** 54–62 (2006).
112. Ochieng, J. & Chaudhuri, G. Cystatin Superfamily. *J Heal. Care Poor Underserved* **21,** 51–70 (2010).
113. Consortium, R. G. Zebrafish Genome Overview. (2017). Available at: https://www.ncbi.nlm.nih.gov/grc/zebrafish.
114. Gamazon, E. R., Nicolae, D. L. & Cox, N. J. A study of CNVS as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet.* **7,** (2011).
115. Gamazon, E. R. & Stranger, B. E. The impact of human copy number variation on gene expression. *Brief. Funct. Genomics* **14,** 352–357 (2015).
116. Korkalainen, M. *Structure and Expression of Principal Proteins Involved in Dioxin Signal Transduction and Potentially in Dioxin Sensitivity*. *Publications of the National Public Health Institute KTL A11 / 2005* (2005).
117. Aqil, F. *et al.* Sustained expression of CYPs and DNA adduct accumulation with continuous exposure to PCB126 and PCB153 through a new delivery method: Polymeric implants. *Toxicol. Reports* **1,** 820–833 (2014).
118. Klein, J. C. *et al.* Repair and replication of plasmids with site-specific 8-oxodG and 8-AAFDG residues in normal and repair-deficient human cells. *Nucleic Acids Res.* **20,** 4437–4443 (1992).

119. Ensminger, M. *et al.* DNA breaks and chromosomal aberrations arise when replication meets base excision repair. *J. Cell Biol.* **206,** 29–43 (2014).
120. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–564 (2009).
121. Grimes, A. C. *et al.* PCB126 exposure disrupts zebrafish ventricular and branchial but not early neural crest development. *Toxicol. Sci.* **106,** 193–205 (2008).
122. Nebert, D., Roe, A. & Dieter, M. Role of the aromatic hydrocarbon receptor and [Ah] gene battery in the oxidative stress response, cell cycle control, and apoptosis. *Biochem. Pharmacol.* **59,** 65–85 (2000).
123. Matthews, H. B. PHARMACOKINETICS OF PCBs. *Annu. Rev. Pharmacol. Toxicol.* **24,** 85–103 (1984).
124. Pery, A. R. R. *et al.* A Physiologically Based Toxicokinetic Model for the Zebrafish Danio rerio. *Environ. Sci. Technol.* **48,** 781–790 (2014).
125. TAKAHASHI, H. Juvenile hermaphroditism in the zebrafish, Brachydanio rerio. *Bull. Fac. Fish. Hokkaido Univ.* **28,** 57–65 (1977).
126. Orban, L., Sreenivasan, R. & Olsson, P.-E. Long and winding roads: testis differentiation in zebrafish. *Mol. Cell. Endocrinol.* **312,** 35–41 (2009).
127. Jonsson, M. E., Jenny, M. J., Woodin, B. R., Hahn, M. E. & Stegeman, J. J. Role of AHR2 in the expression of novel cytochrome P450 1 family genes, cell cycle genes, and morphological defects in developing zebra fish exposed to 3,3’,4,4’,5-pentachlorobiphenyl or 2,3,7,8-tetrachlorodibenzo- p -dioxin. *Toxicol. Sci.* **100,** 180–193 (2007).
128. White, R. M. *et al.* Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell Stem Cell* **2,** 183–9 (2008).
129. Holden, L. A. & Brown, K. H. Baseline mRNA expression differs widely between common laboratory strains of zebrafish. *Sci. Rep.* **8,** 1–10 (2018).
130. Shabalin, A. A. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).
131. Kim-Hellmuth, S. *et al.* Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat. Commun.* **8,** (2017).
132. Montembault, E., Dutertre, S., Prigent, C. & Giet, R. PRP4 is a spindle assembly checkpoint protein required for MPS1, MAD1, and MAD2 localization to the kinetochores. *J. Cell Biol.* **179,** 601–

609 (2007).
133. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136,** 701–718 (2009).
134. Tanackovic, G. *et al.* PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. *Hum. Mol. Genet.* **20,** 2116–2130 (2011).
135. Jin, Y. *et al.* A genome-wide association study of heat stress-associated SNPs in catfish. *Anim. Genet.* **48,** 233–236 (2017).
136. Yakubov, E. *et al.* Overexpression of genes in the CA1 hippocampus region of adult rat following episodes of global ischemia. *Mol. Brain Res.* **127,** 10–26 (2004).
137. Stawicki, T. M. *et al.* Cilia-Associated Genes Play Differing Roles in Aminoglycoside-Induced Hair Cell Death in Zebrafish. *G3 Genes|Genomes|GeneticsGenes|Genomes|Genetics* **6,** 2225–2235 (2016).
138. Wang, H. *et al.* Expression of dynein, cytoplasmic 2, heavy chain 1 (DHC2) associated with glioblastoma cell resistance to temozolomide. *Sci. Rep.* **6,** 1–12 (2016).
139. Vaisberg, E. A., Grissom, P. M. & Mclntosh, J. R. Mammalian Cells Express Three Distinct Dynein Heavy Chains That Are Localized to Different Cytoplasmic Organelles. *J. Cell Biol.* **133,** 831–842 (1996).
140. Van Meer, G., Voelker, D. R. & Feigenson, G. W. Membrane lipids: Where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* **9,** 112–124 (2008).
141. Fox, K., Zauke, G. P. & Butte, W. Kinetics of Bioconcentration and Clearance of 28 Polychlorinated Biphenyl Congeners in Zebrafish (Brachydanio rerio). *Ecotoxicol. Environ. Saf.* **28,** 99–109 (1994).
142. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017,** 1–17 (2017).
143. White, S. S. & Birnbaum, L. S. An Overview of the Effects of Dioxins and Dioxin-like Compounds on Vertebrates, as Documented in Human and Ecological Epidemiology. *J Env. Sci Heal. C Env. Carcinog Ecotoxicol Rev* **27,** 197–211 (2009).
144. Hankinson, O. The aryl hydrocarbon receptor complex. *Annu. Rev. Phamacol. Toxicol* **35,** 307–40 (1995).
145. ASTDR. *Toxicological Profile for Polychlorinated Biphenyls (PCBs)*. *Agency for Toxic Substances and Disease Registry* (2000).
146. Hertz-Picciotto, I. *et al.* Prenatal exposures to persistent and non-persistent organic compounds and effects on immune system development. *Basic Clin. Pharmacol. Toxicol.* **102,** 146–54

(2008).
147. Meeker, J. D. & Hauser, R. Exposure to polychlorinated biphenyls (PCBs) and male reproduction. *Syst. Biol. Reprod. Med.* **56,** 122–31 (2010).
148. Yu, M.-L., Guo, Y. L., Hsu, C.-C. & Rogan, W. J. Increased mortality from chronic liver disease and cirrhosis 13 years after the Taiwan 'Yucheng'('oil disease') incident. *Am. J. Ind. Med.* **31,** 172–175 (1997).
149. Holden, L. A. & Brown, K. H. Response eQTL analysis of low-dose PCB exposure connects genomic copy number variants to susceptibility. *Aquat. Toxicol.* under review (2018).
150. Mullins, M. *Egg Microinjection Technique and Morpholinos*. *Zebrafish Course* (2013).
151. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: A CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42,** 401–407 (2014).
152. Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44,** W272–W276 (2016).
153. Meeker, N. D., Hutchinson, S. A., Ho, L. & Trede, N. S. Method for isolation of PCR-ready genomic DNA from zebrafish tissues. *Biotechniques* **43,** 610–614 (2007).
154. Shah, A. N., Davey, C. F., Whitebirch, A. C., Miller, A. C. & Moens, C. B. Rapid Reverse Genetic Screening Using CRISPR in Zebrafish. *Nat. Methods* **12,** 535–540 (2015).
155. Friedman, M. The use of ranks to avoid the assumtion of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32,** 675–701 (1937).
156. Lenth, R. V. Least-Squares Means: The *R* Package **lsmeans**. *J. Stat. Softw.* **69,** (2016).
157. Shen, J. P. & Ideker, T. Correcting CRISPR for copy number. *Nat. Genet.* **49,** 1674–1675 (2017).
158. Sheel, A. & Xue, W. Genomic amplifications cause false positives in CRISPR screens. *Cancer Discov.* **6,** 824–826 (2016).
159. Aguirre, A. J. *et al.* Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.* **6,** 914–929 (2016).

Appendix

Supplemental Files

Chapter 1

Supplemental data 1.1 (xlsx)
Annotation of genes with statistically significant changes in gene expression following PCB exposure in Fundulus heteroclitus plus annotation of blastn homolog in Danio rerio. CNV status is indicated as yes/no for direct overlap or indirect association (±100 Kb) of gene. File size: 19.4 kB

Chapter 2

Supplementary Table 2.1 (xlsx)
qPCR and standard PCR primers, amplicon sizes, and locations. File size: 10.5 kB

Supplementary Table 2.2 (xlsx)
1941 CNV regions in the Zv8 genome and stepwise liftover to Zv9, GRCz10, and GRCz11. File size: 215.2 kB

Supplementary Table 2.3 (txt)
Tabular results of Ensembl Variant Effect Predictor of 1355 CNV (GRCz10) queried against the RefSeq database. File size: 1.8 MB

Chapter 3

Supplementary Dataset 3.1 (xls)
Differentially expressed mRNA transcripts between males and females in AB. File includes probe name, log fold change, average expression, p-value, Bonferroni-corrected p-value, and genomic coordinates (danRer7/Zv9) as reported by Agilent Feature Extract and annotation files. Entrez gene ID, Ensembl ID, gene symbol, and gene name were manually confirmed and harmonized. Gene symbols and names highlighted in pink are differentially expressed in females. Gene symbols and names highlighted in blue are differentially expressed in males. File size: 76.0 kB

Supplementary Dataset 3.2 (xls)
Differentially expressed mRNA transcripts between males and females in WIK. File includes probe name, log fold change, average expression, p-value, Bonferroni-corrected p-value, and genomic coordinates (danRer7/Zv9) as reported by Agilent Feature Extract and annotation files. Entrez gene ID, Ensembl ID, gene symbol, and gene name were

manually confirmed and harmonized. Gene symbols and names highlighted in pink are differentially expressed in females. Gene symbols and names highlighted in blue are differentially expressed in males. File size: 54.0 kB

Supplementary Dataset 3.3 (xls)
Differentially expressed mRNA transcripts between sexes regardless of strain. File includes probe name, log fold change, average expression, p-value, Bonferroni-corrected p-value, and genomic coordinates (danRer7/Zv9) as reported by Agilent Feature Extract and annotation files. Entrez gene ID, Ensembl ID, gene symbol, and gene name were manually confirmed and harmonized. Gene symbols and names highlighted in pink are differentially expressed in females. Gene symbols and names highlighted in blue are differentially expressed in males. Log fold change, average expression, p-values, and Bonferroni-corrected p-values are averaged between AB and WIK differential expression datasets (SupplementaryDataset3.1 and SupplementaryDataset3.2). File size: 46.5 kB

Supplementary Dataset 3.4 (xls)
Differentially expressed mRNA transcripts between AB, TU, and WIK in males. File includes probe name, log fold change, average expression, p-value, Bonferroni-corrected p-value, and genomic coordinates (danRer7/Zv9) as reported by Agilent Feature Extract and annotation files. This file also includes calculated log fold change values for each strain individually by taking the average of the relative log fold change: (AB.TU-AB.WIK)/2 = ABcalc. Calculated strain-specific log fold change values were then centered on zero for each probe (ABcenter). Entrez gene ID, Ensembl ID, gene symbol, and gene name were manually confirmed and harmonized. Several probes annotate to deprecated gene IDs; the few that fall into this category are retained in the file, but identified by strike-through. File size: 184.5 kB

Supplementary Dataset 3.5 (xls)
Differentially expressed mRNA transcripts between AB, TU, and WIK in females. File includes probe name, log fold change, average expression, p-value, Bonferroni-corrected p-value, and genomic coordinates (danRer7/Zv9) as reported by Agilent Feature Extract and annotation files. This file also includes calculated log fold change values for each strain individually by taking the average of the relative log fold change: (AB.TU-AB.WIK)/2 = ABcalc. Calculated strain-specific log fold change values were then centered on zero for each probe (ABcenter). Entrez gene ID, Ensembl ID, gene symbol, and gene name were manually

confirmed and harmonized. Several probes annotate to deprecated gene IDs; the few that fall into this category are retained in the file, but identified by strike-through. File size: 144.0 kB

Supplementary Dataset 3.6 (xls)
Differentially expressed mRNA transcripts between strains regardless of sex. File includes probe name, log fold change, average expression, p-value, Bonferroni-corrected p-value, and genomic coordinates (danRer7/Zv9) as reported by Agilent Feature Extract and annotation files. Log fold change, average expression, p-values, and Bonferroni-corrected p-values are averaged between male and female differential expression datasets (SupplementaryDataset4 and AddtionalFile5).This file also includes calculated log fold change values for each strain individually by taking the average of the relative log fold change: (AB.TU-AB.WIK)/2 = ABcalc. Calculated strain-specific log fold change values were then centered on zero for each probe (ABcenter). Entrez gene ID, Ensembl ID, gene symbol, and gene name were manually confirmed and harmonized. Several probes annotate to deprecated gene IDs; the few that fall into this category are retained in the file, but identified by strike-through. File size: 63.5 kB

Supplementary Dataset 3.7 (xls)
Differentially expressed mRNA transcripts that have corresponding evidence of circadian regulation in 4 mouse liver microarray experiments from the Circadian Expression Profiles Data Base (circaDB, http://circadb.hogeneschlab.org/). File includes a tab for each supplemental dataset (sd1-6) that contains Probeset_ID, Symbol, JTKP, JTKQ, JTKperiod, JTKphase, and Tissue columns. Probeset ID = unique to each microarray expression platform. Symbol = gene symbol. JTKP = JTK_CYCLE p-value. JTKQ = JTK_CYCLE q-value. JTKperiod = period of circadian oscillation, in hours. JTKphase = phase of circadian oscillation, in hours. Tissue = original dataset where mogene_liver = Mouse 1.OST Liver (Affymetrix), liver = Mouse Liver 48 hour Hughes 2009 (Affymetrix), panda_liver = Mouse Liver Panda 2002 (Affymetrix), and WT_liver = Mouse Wild Type Liver (GNF microarray). "Merge" tab combines all circadian-driven genes from sd1-6 tabs with duplicates removed. "Unique" tab lists the gene symbol for the 82 genes described in this dataset. File size: 97.5 kB

Chapter 4
Supplemental Data 4.1 (xlsx)
Differentially expressed mRNA transcripts between PCB-exposed and control fish. File includes probe name, log fold change, average

expression, p-value, Bonferroni-corrected p-value, and genomic coordinates (danRer7/Zv9) as reported by Agilent Feature Extract and annotation files. File size: 51.4 kB

Chapter 5

Supplemental Table 5.1 (xlsx)
Heart rate in AB zebrafish. Pairwise t-test p-values with Benjamini-Hochberg correction for multiple comparisons. File size: 13.6 kB

Supplemental Table 5.2 (xlsx)
Heart rate in TU zebrafish. Pairwise t-test p-values with Benjamini-Hochberg correction for multiple comparisons. File size: 12.4 kB

Supplemental Table 5.3 (xlsx)
Edema score in uninjected AB zebrafish. Pairwise t-test p-values from Tukey honest significant difference test. File size: 16.0 kB

Supplemental Table 5.4 (xlsx)
Edema score in sham-injected AB zebrafish. Pairwise t-test p-values from Tukey honest significant difference test. File size: 15.7 kB

Supplemental Table 5.5 (xlsx)
Edema score in CRISPR-injected AB zebrafish. Pairwise t-test p-values from Tukey honest significant difference test. File size: 17.0 kB

Supplemental Table 5.6 (xlsx)
Edema score in uninjected Tuebingen zebrafish. Pairwise t-test p-values from Tukey honest significant difference test. File size: 16.4 kB

Supplemental Table 5.7 (xlsx)
Edema score in sham-injected Tuebingen zebrafish. Pairwise t-test p-values from Tukey honest significant difference test. File size: 16.9 kB

Supplemental Table 5.8 (xlsx)
Edema score in CRISPR-injected Tuebingen zebrafish. Pairwise t-test p-values from Tukey honest significant difference test. File size: 11.0 kB